

P

Pacific Islands, Archaeology of

The Pacific Islands, also known as Oceania, have been subdivided traditionally by anthropologists into three main geographic regions: Melanesia, Micronesia, and Polynesia. Following Green (1991), prehistorians now stress the division between *Near Oceania* in the west (including the Bismarck Archipelago and the Solomon Islands), and *Remote Oceania* (which includes all of island Melanesia southeast of the Solomons, along with Polynesia and Micronesia). This latter distinction recognizes the Pleistocene settlement of Near Oceania, whereas the widely dispersed islands of Remote Oceania were discovered and settled only within the past 4,000 years. Archaeological research in Oceania has a long history, but modern efforts emphasizing stratigraphic excavations did not begin until after World War II (Kirch 2000), and have revealed the main chronological sequence for human settlement. This sequence is summarized here, followed by reviews of the development of complex societies in Oceania, and of the human impact on island environments.

1. Early Human Settlement of Near Oceania

The oldest known occupation sites are radiocarbon dated to ca. 36,000 years ago (the late Pleistocene), on the large island of New Guinea and in the adjacent Bismarck Archipelago (Allen 1996). At several times during the Pleistocene, New Guinea was joined to Australia as a single land mass (known as Sahul), and human entry into and expansion throughout this vast Australasian region occurred rapidly. Late Pleistocene sites in the Admiralty Islands, New Ireland, and Buka (Solomons), all would have required open ocean transport, suggesting the presence of some form of watercraft (possibly rafts, bark boats, or dugouts) (Irwin 1992).

Early human colonists in Near Oceania were hunters and gatherers, who exploited tropical rainforests as well as inshore marine resources (see *Hunter-Gatherer Societies, Archaeology of*). Long-distance communication and exchange is indicated by the movement of obsidian between islands. By the early Holocene period (after 8000 BC), there is archaeological evidence for domestication of tree, root, and tuber crops (such as the *Canarium* almond and various aroids) within Near Oceania. Archaeological evidence for the cultivation of swamplands at Kuk in

the Highlands of New Guinea commences as early as 7000 BC (Golson 1988). These archaeological indications confirm the long-held ethnobotanical hypothesis that Near Oceania was one of several independent centers for the origins of tropical horticulture (see *Food Production, Origins of*).

2. Austronesian Expansion and Lapita

During the early Holocene, the southeastern Solomon Islands marked the limit of human expansion. Beginning around 2000 BC, a major expansion or diaspora of people speaking languages belonging to the Austronesian language family commenced (Blust 1995). Their immediate homeland has generally been regarded as comprising the island of Taiwan (and perhaps adjacent areas of mainland China). The ability of early Austronesians to disperse rapidly has been attributed to their invention of the outrigger sailing canoe (Pawley and Ross 1993). The Austronesians were horticulturalists who transported root, tuber, and tree crops via their canoes, along with breeding stocks of domestic pigs, dogs, and chickens.

The Austronesian diaspora rapidly encompassed the major archipelagoes of island Southeast Asia; one branch of Austronesian-speakers expanded along the north coast of New Guinea into the Bismarck Archipelago. This branch is known to linguists as Oceanic, and the Oceanic languages (numbering about 450 modern languages) include most of those spoken throughout the Pacific. The great exception is New Guinea, where roughly 750 non-Austronesian languages are spoken.

Archaeological evidence for the initial Austronesian dispersal into the Pacific comes from both western Micronesia (the Marianas and Palau archipelagoes), and from the Bismarck Archipelago. In western Micronesia, early sites contain red-slipped pottery, some of which is decorated with lime-filled, impressed designs (Rainbird 1994). These sites, along with radiocarbon-dated sediment cores exhibiting signals of human presence (e.g., high influxes of microscopic charcoal resulting from anthropogenic burning) suggest that humans settled the Marianas and Palau no later than 1500 BC, and possibly as early as 2000 BC.

In the Bismarck Archipelago, the initial Austronesian incursion has been correlated with the appearance of a distinctive suite of sites, also containing pottery with lime-infilled decorations, but with motifs

made largely by a technique of dentate-stamping. These sites and the associated artifacts (such as *Tridacna*-shell adzes and *Trochus*-shell fishhooks, as well as ornaments) represent the earliest known phase of the Lapita cultural complex, dating to ca. 1500–1300 BC (Gosden et al., 1989, Spriggs 1997). Early Lapita sites were frequently hamlets or villages consisting of houses elevated on posts or stilts, situated over tidal reef flats or along shorelines. Excavated plant and animal remains indicated a mixed economy with horticulture and marine exploitation. Substantial quantities of obsidian, chert, pottery, shell artifacts, and other materials were exchanged between communities (Kirch 1997).

Correlations among archaeological assemblages, language groups, and human biological populations are often complex, and need not be isomorphic. However, a strong consensus is emerging among scholars in several disciplines that the initial phase of the Lapita cultural complex can be correlated with the Proto Oceanic interstage of the Austronesian language family. Moreover, genetic evidence (such as mtDNA and hemoglobin markers) supports the view that the Lapita phenomenon reflects a substantial population intrusion into the Bismarck Archipelago, deriving out of island Southeast Asia (Hill and Serjeantson 1989). At the same time, the Proto Oceanic speakers undoubtedly had considerable interaction (cultural, linguistic, and genetic) with the indigenous non-Austronesian speaking populations who already occupied the Bismarck region in the mid-Holocene. Thus the Lapita cultural complex is seen as an outcome of cultural processes of intrusion, integration, and innovation.

3. Human Colonization of Remote Oceania

Beginning ca. 1300 BC, the Lapita pottery makers expanded rapidly beyond the Solomons and into the southwestern archipelagoes of Remote Oceania: Vanuatu, the Loyalty Islands, New Caledonia, Fiji, Tonga, and Samoa. Numerous radiocarbon-dated archaeological sites document that Lapita sites in all of these archipelagoes are no later than 900 BC.

We have already noted that the westernmost islands of Micronesia were colonized directly out of island Southeast Asia by Austronesian speakers ca. 2000–1500 BC. Around 2,000 years ago, Oceanic speakers who made plainware pottery (a late form of Lapita) and who used shell adzes, fishhooks, and other implements, founded settlements on several volcanic islands of central Micronesia (Chuuk, Pohnpei, and Kosrae). The atolls of the Marshall Islands were also colonized at this time.

The final stage in the human settlement of the Pacific Islands began after 500 BC, with the Polynesian dispersals eastwards out of Tonga and Samoa. Ancestral Polynesian culture and Proto Polynesian lan-

guage themselves had developed in this Tonga–Samoa region between ca. 900–500 BC, directly out of the founding Lapita cultural complex (Kirch 2000, Kirch and Green, in press). While archaeologists debate the exact chronology and sequence of Polynesian dispersals, most agree that the central Eastern Polynesian archipelagoes (such as the Society Islands, Cook Islands, and Marquesas Islands) were settled first, no later than AD 300 and perhaps some centuries earlier (Rolett 1998). Remote Easter Island was discovered by AD 800–900 (Van Tilburg 1994), while the Hawaiian Islands were also well settled by this date. The twin large, temperate islands of New Zealand were colonized by Polynesians around AD 1200 (Anderson 1989, Davidson 1984). Critical to the success of this unprecedented diaspora was the double-hulled sailing canoe, capable of carrying 40–60 people on voyages lasting one month or longer (Irwin 1992). That the Polynesians reached South America and returned is suggested by preserved remains of the sweet potato (*Ipomoea batatas*), a South American domesticate, in several prehistoric Polynesian sites.

Because it was the last sector of Remote Oceania to be settled, and because its populations represent a single radiation or diaspora, Polynesia constitutes a monophyletic cultural and linguistic group. Thus, Polynesia has often been regarded as an ideal region for testing models of cultural differentiation from a common ancestor (e.g., Kirch 1984, Kirch and Green, in press).

4. Development of Complex Societies

When the eighteenth- to nineteenth-century European voyages of discovery inspired by the Enlightenment encountered Pacific island societies, they frequently encountered large, dense populations organized into complex, hierarchical sociopolitical formations. With populations often numbering into the tens or hundreds of thousands, such societies had two to three decision-making levels, and hereditary leaders who enjoyed elite privileges and status markers. Anthropologists classify such sociopolitical formations as chiefdoms, and indeed, the Polynesian chiefdoms are often considered the archetypal model (see *Chiefdoms, Archaeology of*).

The origins, development, and elaboration of Pacific island chiefdoms have been a major topic of archaeological research (e.g., Davidson 1984, Kirch 1984, Rainbird 1994, Sand 1995, Spriggs 1997). Based on linguistic and archaeological evidence, early Austronesian societies were characterized by some degree of internal ranking (especially between senior and junior branches of a descent line), but were probably heterarchical rather than hierarchical in structure. However, heterarchic competition (in such social arenas as marriage and exchange, as well as competition for land) between social groups provided the

basis for true hierarchy (and eventually, in the largest societies, class stratification) to emerge.

Archaeologists have identified several factors and processes that were significant in the rise of Oceanic chiefdom societies. Many of these were closely linked, and should not be considered unicausal variables. For example, population growth leading to large, high-density populations can be identified as a necessary, but probably not sufficient, cause underlying socio-political complexity (Kirch 2000, Sand 1995). The human populations of the volcanic islands typically reached densities of between 100–250 persons per square kilometer prior to European contact, resulting in intense competition for arable land and other resources. Such conditions encouraged centralized, hierarchic control, as well as providing incentives for militaristic aggression.

A second process linked to population growth and to increased hierarchy was intensification of production, including agricultural systems and other forms of production or resource extraction, as well as economic specialization (e.g., in pottery production and trade). On many islands, large-scale irrigation works or dryland field systems were developed during late prehistory. Although population increases may have initially spurred intensification, once in place such intensive production systems provided a means for surplus extraction by chiefs and other leaders, thus encouraging hierarchy. Often culturally marked as tribute, such surpluses were the economic foundation of an emergent elite, including not only hereditary chiefs, but priests, warriors, craft specialists, and others.

Ideology likewise played a key role in Pacific island societies, with the elite cadres of the larger and most complex societies actively employing ideological control as a means of legitimation. The origins of Oceanic ritual systems can be traced back to common Austronesian concepts of the sacredness of ancestors; these concepts later became elaborated as cults in which the highest chiefs were seen as directly descended from powerful gods, and hence essential to the continued well-being of the society at large. Archaeologically, the rise of elite-dominated ideological systems is particularly reflected in monumental architecture, of which the most impressive examples are the giant statue-bearing temples of Easter Island, and the site of Nan Madol on Pohnpei. Other forms of monumental architecture, however, are ubiquitous throughout Pacific islands chiefdoms. Even when monumental architecture is absent, material signs of ideological control can be quite striking, as in the multiple sacrificial interments associated with the burial of Roy Mata, a chief of Vanuatu (Garanger 1972).

Finally, competition, conflict, and warfare also characterized many of the complex societies of the Pacific, especially following the rise of large and dense populations (see also *Cultural Resource Management (CRM): Conservation of Cultural Heritage*). Arch-

aeologically, warfare is marked by a diversity of kinds of fortifications, such as the pallisaded *pa* volcanic cones and headlands of New Zealand, or the ring-ditch fortified villages of Fiji. Another, more gruesome signal of the levels that inter-societal aggression reached on some islands is cannibalism (or para-cannibalistic treatment of enemies, such as dismembering, roasting, and the nonfunerary discard of skeletal remains). Although some anthropologists have expressed skepticism regarding the accounts of cannibalism in the Pacific by European voyagers, there is now direct archaeological evidence for cannibalistic or para-cannibalistic practices in late prehistory on Easter Island, the Marquesas, New Zealand, Mangaia, and Fiji.

5. Human Impacts to Island Ecosystems

The islands of Remote Oceania, due to isolation and related factors, provide model conditions for studying the effects of human colonization and land use on pristine ecosystems. Interdisciplinary research among archaeologists and natural scientists (see also *Environmental Archaeology*) over the past three decades has amplified our understanding of such human–ecosystem interactions (Kirch and Hunt 1997).

Because of the substantial open-ocean distances isolating them from continents as well as other islands, and the difficulty of dispersal to islands, prior to human arrival oceanic ecosystems were typically characterized by: high species-level endemism, but lower diversity in higher-level (generic and family) taxa; lowered competition; survival of archaic forms; and vulnerability to disturbance from outside agents. Larger vertebrates such as marsupials (wombats, cuscus) and rats, snakes, frogs, and most lizards were restricted primarily to Near Oceania, with only a handful of species declining in numbers eastwards to Fiji and Samoa. (The reef and marine resources of Pacific islands also display a west-to-east decline in species diversity.) Throughout most of Remote Oceania, prehuman vertebrate faunas were dominated by birds (including many flightless forms which had evolved *in situ* from flighted ancestors). Prior to human arrival, these bird populations lacked large vertebrate predators, and presumably also a typical predator avoidance strategy. They must have been extremely easy prey for the first humans to step foot on these islands.

When humans first arrived in Remote Oceania, they typically found the islands to be forested, and inhabited by a range of largely endemic species, dominated by birds, along with invertebrates such as land snails and insects. Oceanic peoples possessed a successful colonization strategy that allowed them to exist on isolated islands, by: (a) transporting in their sailing canoes stocks of horticultural crop plants, along with domestic pigs, dogs, and chickens (rats came along,

presumably as 'stowaways'); (b) clearing areas of rainforest for garden land; and (c) intensively exploiting the abundant natural avifaunal and marine resources.

This colonization strategy had several consequences for island ecosystems, all of which are increasingly well documented through both archaeological and paleoenvironmental indicators. Forest clearance on many islands is signaled in changing pollen spectra from sediment cores, with tree taxa rapidly giving away to ferns and grasses; also characteristic are sharp increases in microscopic charcoal influxes, indicating human-induced burning, in most cases probably associated with shifting cultivation. On some islands, forest clearance led to increased erosion rates, along with alluviation of valley bottoms or along coastal plains. The exploitation of natural resources is particularly evident in the zooarchaeological assemblages from early settlement sites, which are characterized by high numbers of land and seabirds, many of them representing now extinct or extirpated species (Steadman 1995). A dramatic case of avifaunal extinctions on Pacific islands is that of the moa, a group of 13 species of large, flightless birds which became totally extinct in New Zealand during the brief period of Polynesian occupation (Anderson 1989).

Cumulative effects of human actions on islands led to irreversible changes, such as dramatic declines in biodiversity, and the conversion of natural rainforests to intensively managed, anthropogenic landscapes. The consequences for human populations themselves were undoubtedly mixed. The replacement of natural ecosystems with intensive food production systems enabled the growth of large and dense human populations. At the same time, reduction or depletion of natural resources, coupled with the necessity for intensive land use, encouraged highly complex sociopolitical systems which at times competed fiercely for control of land and the means of production.

See also: Australia and New Guinea, Archaeology of; Melanesia: Sociocultural Aspects; Polynesia and Micronesia: Sociocultural Aspects; Southeast Asia, Archaeology of

Bibliography

- Allen J 1996 The pre-Austronesian settlement of island Melanesia: Implications for Lapita archaeology. *Transactions of the American Philosophical Society* **86**(5): 11–27
- Anderson A 1989 *Prodigious Birds: Moas and Moa Hunting in Prehistoric New Zealand*. Cambridge University Press, Cambridge, UK
- Blust R 1995 The prehistory of the Austronesian-speaking peoples: The view from language. *Journal of World Prehistory* **9**: 453–510
- Davidson J M 1984 *The Prehistory of New Zealand*. Longman Paul, Auckland, New Zealand
- Garanger J 1972 *Archéologie des Nouvelles-Hébrides*. Publication de la Société des Océanistes No. 30, Paris

- Golson J 1988 The origins and development of New Guinea agriculture. In: Harris D, Hillman G (eds.) *Foraging and Farming: The Evolution of Plant Exploitation*. Unwin Hyman, London
- Gosden C, Allen J, Ambrose W, Anson D, Golson J, Green R, Kirch P V, Lilley I, Specht J, Spriggs M 1989 Lapita sites of the Bismarck Archipelago. *Antiquity* **63**: 561–586
- Green R C 1991 Near and Remote Oceania: Disestablishing "Melanesia" in culture history. In: Pawley A (ed.) *Man and a Half: Essays in Pacific Anthropology and Ethnobiology in Honour of Ralph Bulmer*. The Polynesian Society, Auckland, New Zealand
- Hill A V S, Serjeantson S W (eds.) 1989 *The Colonization of the Pacific: A Genetic Trail*. Clarendon Press, Oxford, UK
- Irwin G 1992 *The Prehistoric Exploration and Colonisation of the Pacific*. Cambridge University Press, Cambridge, UK
- Kirch P V 1984 *The Evolution of the Polynesian Chiefdoms*. Cambridge University Press, Cambridge, UK
- Kirch P V 1997 *The Lapita Peoples: Ancestors of the Oceanic World*. Blackwell Publishers, Oxford, UK
- Kirch P V 2000 *On the Road of the Winds: An Archaeological History of the Pacific Islands Before European Contact*. University of California Press, Berkeley, CA
- Kirch P V, Green R C, in press, *Hawaiki, Ancestral Polynesia: An Essay in Historical Anthropology*. Cambridge University Press, Cambridge, UK
- Kirch P V, Hunt T L (eds.) 1997 *Historical Ecology in the Pacific Islands: Prehistoric Environmental and Landscape Change*. Yale University Press, New Haven, CT
- Pawley A K, Ross M 1993 Austronesian historical linguistics and culture history. *Annual Review of Anthropology* **22**: 425–59
- Rainbird P 1994 Prehistory in the northwest tropical Pacific: The Caroline, Mariana, and Marshall Islands. *Journal of World Prehistory* **8**: 293–349
- Rolett B V 1998 *Hanamiai: Prehistoric Colonization and Cultural Change in the Marquesas Islands (East Polynesia)*. Yale University Publications in Anthropology No. 84, New Haven, CT
- Sand C 1995 "Le Temps d'Avant": *La Préhistoire de la Nouvelle-Calédonie*. L'Harmattan, Paris
- Spriggs M J T 1997 *The Island Melanesians*. Blackwell, Oxford, UK
- Steadman D W 1995 Prehistoric extinctions of Pacific island birds: Biodiversity meets zooarchaeology. *Science* **267**: 1123–30
- Van Tilburg J A 1994 *Easter Island: Archaeology, Ecology and Culture*. British Museum Press, London

P. V. Kirch

Pain, Health Psychology of

1. Basic Terms

Pain is an adaptive phenomenon that signals impending danger to the body. Pain can, however, also be maladaptive, for example in states of chronic pain, and thus can itself become a significant disorder. Epidemiological studies show that 80 percent of the population experience recurrent pain, with more than 10 percent of the population being permanently disabled by it. Although pain has for a long time been

viewed a purely sensory phenomenon or an epiphenomenon of a medical disorder, this view changed during the last 40 years of the twentieth century. It has been recognized that pain is a psychobiological experience, with emotional aspects being as important as sensory pain characteristics. Thus nociception, the physiological process of a noxious signal being transmitted from the periphery to the brain, has been differentiated from the experience of pain, which encompasses all aspects including psychological, social and cultural factors. The international Association for the Study of Pain (IASP) definition takes this shift from a biomedical to a biobehavioral or psychobiological perspective into account by characterizing pain as an 'unpleasant sensory and emotional experience associated with actual or potential tissue damage or described in terms of such damage' (Merskey 1986). This definition was prompted by the fact that pain may often be experienced in the absence of identifiable objective pathology. A major change in the traditional view of pain was the gate control theory of pain proposed in Melzack and Wall (1965). More important than its physiological aspects, which have not all been confirmed, was the conceptual model behind it. The gate control theory stated not only that pain could be modulated by both afferent and efferent factors at the level of the spinal cord, but also emphasized that pain has sensory-discriminative, motivational-emotional, and cognitive-evaluative components as well, and thus it assigns an important role to psychological factors. A core feature of this model is the integration of peripheral stimuli with psychological variables such as mood, attention, or cultural experience in the perception of pain. Thus, the gate control theory has been instrumental in abolishing the dichotomy of psychogenic versus somatogenic pain—psychological and physiological factors always interact in the production of the experience of pain and are not mutually exclusive causes of pain.

2. Classification of Pain

The common distinction between acute and chronic pain—referring to pain of at least three to six months duration and/or pain that exceeds the normal time for healing in the case of an acute injury—is useful, because chronic pain is often maladaptive and needs special attention. Whereas acute pain such as pain related to medical procedures or childbirth usually leads to anxiety and apprehension, chronic pain tends to be associated with helplessness, depression, and irritability as well as interference with family function, work, or social interaction (Gatchel and Turk 1999). In addition to the large number of workdays that are lost as a consequence of it, chronic pain is the leading cause of invalidity in the age under-50s group and causes enormous costs for the health care system (Melzack and Wall 1994).

The IASP classification of pain is organized in five general categories: (a) the bodily system affected, (b) time characteristics, (c) intensity of pain (d) duration of pain, and (e) its presumed etiology. This classification lacks reliability especially in respect of the last characteristic. In addition, it neglects psychological factors that are only coded with respect to the presence of a psychiatric disorder or within the category 'psychophysiological dysfunction.' *ICD-10* as well as *DSM-IV* still adhere to the distinction between psychogenic-somatogenic or physical vs. somatoform pain, which must be viewed as obsolete given the large number of research findings emphasizing the importance of psychological factors in any type of pain, both acute and chronic (Gatchel and Turk 1999).

The biopsychosocial model of pain views pain as a complex response that can be described on the verbal-subjective, the motor-behavioral and physiological levels. Pain can be based on nociceptive input, but nociception is not a prerequisite of the experience of pain, which can be an exclusively central phenomenon, but always has physiological antecedents and consequences (Flor et al. 1990). The multiaxial classification of pain as proposed by Turk and his colleagues (cf. Turk and Rudy 1988) categorizes pain patients along both the somatic and the psychosocial dimensions. This classification has yielded three subgroups of patients in the psychosocial domain characterized as adaptive copers, interpersonally distressed, and dysfunctional, with significant differences in the course of the illness and response to treatment.

3. The Role of Learning in Chronic Pain

Both associative and nonassociative learning processes as well as social learning have been found to be of fundamental significance for the development of chronic pain. The repeated application of painful stimuli leads to reduced responsivity, i.e., habituation to the painful stimulation. In many states of chronic pain, sensitization rather than habituation occurs due to changes both at the level of the receptor and of the central nervous system (Woolf and Mannion 1999). Sensory information accelerates habituation and reduces activation caused by surprise, insecurity, and threat. This mechanism may underlie the effects reported in a large number of studies that support the positive results of preparatory information prior to acutely painful procedures such as surgery or bone marrow aspiration.

The most influential model of psychological factors in chronic pain was Fordyce's assumption that chronic pain can develop and be maintained due to operant conditioning of pain behaviors, i.e., overt expressions of pain. Fordyce (1976) postulated that acute pain behaviors such as limping or moaning may come under the control of external contingencies of reinforcement, and thus develop into a chronic pain

problem. Positive reinforcement of pain behaviors (e.g., by attention and the expression of sympathy), negative reinforcement of pain behaviors (e.g., by attention and the expression of sympathy), negative reinforcement of pain behaviors (e.g., the reduction of pain through the intake of medication or by the cessation of activity) as well as a lack of reinforcement of healthy behaviors could provoke chronicity in the absence of nociceptive input. Thus, pain behaviors, originally elicited by nociceptive input, may, over time, occur in response to environmental contingencies. This model has generated much research, which has not only confirmed the original assumptions made by Fordyce; it has also been shown that in addition to pain behaviors the subjective experience of pain, as well as physiological responses related to pain, are subject to reinforcement learning.

A special role has been assigned to the 'significant others' of chronic pain patients, who have high reinforcement potential. When solicitous, i.e., pain-reinforcing, spouses were present, several studies found that the patients were more pain-sensitive during acute pain tests than when the spouses were absent. The patients with nonsollicitous spouses did not differ whether the spouse was present or absent. These studies suggest that spouses can serve as discriminative stimuli for the display of pain behaviors by chronic pain patients, including their reports of pain intensity. Health care providers may also serve as discriminative cues influencing patients' responses. Solicitous responses by significant others can also lead to increased physiological pain responses in the patients, whereas the focusing on healthy behaviors by significant others can have positive effects on the pain experience.

Of equal importance is the operant conditioning related to the intake of pain medication. Patients are often told by their physicians or by well-meaning family members that they should not continue to take analgesic medication unless the pain increases to a point where it becomes intolerable (referred to as *prn* from the Latin, 'take as needed'). When pain medication is taken at that stage, both pharmacological and behavioral factors can contribute to the development of the misuse of medication and, in severe cases, even drug dependence. If analgesic medication is taken only at peak pain levels, the effect of the medication is less potent, and patients cycle between high and low levels of medication that facilitate the development of dependence. In addition, medication intake is negatively reinforcing, since the intake of medication ends an aversive state (pain). Subsequently, the pain-reducing behavior (use of analgesics) increases in frequency. Thus, both pharmacotherapists and behavioral psychologists recommend that analgesic medication should be taken not in a pain-contingent but rather in a time-contingent fashion, adapted to the specific pain level of the patient and to the half-life of the drug.

The negative reinforcement of activity levels is an important process in the development of disability. A specific activity, for example walking, is performed until pain sets in, at which point the activity is interrupted and replaced by rest. Subsequently, the pain will be reduced. The reduction of an aversive state—pain—negatively reinforces the cessation of activity. As was the case with the intake of analgesic medication, the cessation of activity has to be made dependent on the amount of activity achieved—quota-based (e.g., number of stairs climbed, distance walked) rather than on the amount of pain. Thus, the pain-reinforcing quality of rest is eliminated. This formulation supports the strategy of encouraging patients to perform activities to meet a specific quota, and not until pain is perceived as overwhelming.

The respondent conditioning model postulates that numerous, formerly neutral, cues can be associated with the experience of pain and can—over time—themselves elicit responses that lead to an increased pain response, and can create the experience of pain in the absence of nociceptive input. To illustrate this process, the patient may have learned to associate increases in muscle tension with all kinds of stimuli that were originally associated with nociceptive stimulation. Thus, sitting, walking, bending, or even thoughts about these movements, may elicit anticipatory anxiety and an increase in muscle tension. This fear of movement or 'kinesophobia' has been discussed as an important factor in the maintenance and exacerbation of chronic pain. Subsequently, patients may display maladaptive responses to any number of stimuli and reduce the frequency of performance of many activities other than those that initially reduced pain. Thus, although the original association between injury and pain results in anxiety regarding movement, over time the anxiety may lead to increased muscle tension and pain even if the nociceptive stimuli are no longer present. In addition, stress situations can increase muscle tension levels and cause sympathetic activation and may, thus, reinforce this process.

Many patients report that an acute pain problem evolved into chronic pain at a time where personal stressors co-occurred with the pain. Stress situations may serve as additional unconditioned stimuli and also as conscious stimulus for muscle tension increases, increased sympathetic activation, and subsequently pain. Nonoccurrence of pain is a powerful reinforcer for reduction of movement. Thus, the original respondent conditioning may be complemented by an operant process, whereby the nociceptive stimuli need no longer be present for the avoidance behavior to occur. People who suffer from acute back pain, regardless of the cause, may adopt specific behaviors (e.g., limping) to avoid pain and they may never obtain 'corrective feedback' because they fail to perform more natural movements. Reduction in physical activity may subsequently result in muscle atrophy and increased disability. In this manner, the physical

abnormalities proposed by biomechanical models of pain may actually be secondary to changes in behavior initiated through learning. Chronic pain patients tend to focus their attention on impending pain and subsequently avoid many types of activity, thus fostering the development of disability and depression. The release of endogenous opioids—the body's own analgesic system—may also be influenced by respondent conditioning as well as brain responses related to the experience of pain. Moreover, fear of pain and subsequent avoidance of activity is one of the best predictors of subsequent chronicity (Asmundson et al. 1999).

Response acquisition through the observation of others and modeling is an essential mechanism of learning new patterns of behavior. Children acquire attitudes about health and health care, and the perception and interpretation of symptoms and physiological processes from their parents and their social environment. They also learn appropriate responses to injury and disease, and thus may be more or less likely to ignore or over-respond to normal bodily sensations they experience. The culturally-acquired perception and interpretation of symptoms determines how people deal with illness; the sight of others in pain is an event that captures attention and may have survival value, may help to avoid experiencing more pain and help to learn what to do about acute pain. Modeling probably plays a part in the phenomenon of 'pain-prone families' families with a significantly elevated occurrence of pain problems (Whitehead et al. 1994). It was, for example, reported that children show the same pain syndromes as their parents currently have, rather than the pain problems their parents had in their own childhood. The large cultural variations in pain expression are also important. In common clinical practice, the acquisition or extinction of pain-related behavior by means of modeling has received little attention. However, there are occasional indications for the role of modeling in treating pain problems in children in burn units and in the treatment of postoperative pain.

Despite the great deal of data available on the modification of experimentally induced pain behavior by means of modeling in normal (healthy) subjects, there are few experimental results concerning chronic pain patients. Nor are there any longitudinal studies of the development of pain syndromes in 'pain-prone families.' Further investigation is necessary for the evidence of modeling as a factor in the development of chronic pain disorders.

4. Cognitive Factors and Pain

Cognitive-behavioral models of chronic pain emphasize that the evaluation of the pain experience by the patient greatly determines the amount of pain that is experienced as well as its negative consequences (Turk

et al. 1983). General assumptions that characterize the cognitive-behavioral perspective are: (a) people are active processors of information, not passive reactors; (b) thoughts (e.g., appraisals, expectancies) can elicit or modify mood, affect physiological processes, influence the environment, and serve as impetus for behavior. Conversely, mood, physiology, environmental factors, and behavior can influence thought processes; (c) behavior is reciprocally determined by the person and by environmental factors; (d) people can learn more adaptive ways of thinking, feeling, and behaving; and (e) people are capable and should be involved as active agents in change of maladaptive thoughts, feelings, and behaviors.

From the cognitive-behavioral perspective, people suffering from chronic pain are viewed as having negative expectations about their own ability to control certain motor skills such as performing specific physical activities (e.g., climbing stairs, lifting objects) that are attributed to one overwhelming factor, namely, a chronic pain syndrome. Moreover, chronic pain patients tend to believe that they have a limited ability to exert any control over their pain. Such negative, maladaptive appraisals about their situation and personal efficacy may reinforce the experience of demoralization, inactivity, and overreaction to noxious stimulation.

A great deal of research has been directed toward identifying cognitive factors that contribute to pain and disability. These have consistently demonstrated that patients' attitudes, beliefs, and expectancies about their plight, themselves, their coping resources, and the health care system affect their reports of pain, activity, disability, and response to treatment.

Pain, when interpreted as signifying ongoing tissue damage or a progressive disease, seems to produce considerably more suffering and behavioral dysfunction than if it is viewed as being the result of a stable problem that is expected to improve. A number of studies using experimental pain stimuli demonstrated that the conviction of personal control can ameliorate the experience of experimentally-induced nociception. Moreover, the type of thoughts employed during exposure to painful stimulation has been related to pain tolerance and pain intensity ratings. Catastrophizing thoughts have been associated with lower pain tolerance and higher ratings of pain intensity. In contrast, coping thoughts have been related to higher pain tolerance and lower pain intensity ratings.

Certain beliefs may lead to maladaptive coping, increased suffering, and greater disability. Patients who believe their pain is likely to persist may be passive in their coping efforts and fail to make use of available strategies, even when in their repertoire, to cope with pain. Patients who consider their pain to be an 'unexplainable mystery,' may negatively evaluate their own ability to control or decrease pain, and are less likely to rate their coping strategies as effective in controlling and decreasing pain.

Once beliefs and expectancies (cognitive schemata) about a disease are formed they become stable and very difficult to modify. Patients tend to avoid experiences that could invalidate their beliefs, and they adapt their behavior in accordance with these beliefs, even in situations where the belief is no longer valid (no corrective feedback is received to discredit this belief). For example, feeling muscular pain following activity may be caused by lack of muscle strength and general deconditioning and not by additional tissue damage.

Self-regulation of pain and its impact depends upon a person's specific ways of dealing with pain, adjusting to pain, and reducing or minimizing pain and distress caused by pain—their coping strategies. Coping is assumed to be manifested by spontaneously-employed purposeful and intentional acts, and it can be assessed in terms of overt and covert behaviors. Overt, behavioral coping strategies include rest, medication, and use of relaxation. Covert coping strategies include various means of distracting oneself from pain, reassuring oneself that the pain will diminish, seeking information, and problem solving. Studies have found active coping strategies (efforts to function in spite of pain, or to distract oneself from pain such as activity, or ignoring it) to be associated with adaptive functioning; and passive coping strategies (depending on others for help in pain control and restricted activities) to be related to greater pain and depression. However, beyond this there is no evidence supporting the greater effectiveness of one active coping strategy compared to any other (Fernandez and Turk 1989). Specific coping strategies need not always be adaptive or maladaptive. It seems more likely that different strategies will be more effective than others for some people at some times, but not necessarily for all people all of the time, or even the same person at different times.

5. Pain and Affective Factors

The affective factors associated with pain include many different emotions, but they are primarily negative in quality. Anxiety and depression have received the greatest amount of attention in chronic pain patients; however, anger has recently received considerable interest as an important emotion in chronic pain patients. Research suggests that from 40 to 50 percent of chronic pain patients suffer from depression (Romano and Turner 1985). There have been extensive and fruitless debates concerning the causal relationship between depression and pain. In the majority of cases, depression appears to be the patients' reaction to their plight. The presence of depression is closely related to the feelings of loss of control and helplessness often associated with pain.

Several investigators have also found a close association between fear of pain and dysfunctional

coping. In addition, high comorbidity between anxiety disorders and pain seems to be present. Muscular hyperreactivity to stress seems to be closely associated with fear of pain.

Anger has been widely observed in individuals with chronic pain. The internalization of angry feelings seems to be strongly related to measures of pain intensity, perceived interference, and reported frequency of pain behaviors. Anger and hostility are closely associated with pain in persons with spinal cord injuries. Frustrations related to persistence of symptoms, limited information on etiology, and repeated treatment failures along with anger toward employers, the insurance, the health care system, family members, and themselves, also contributed to the general dysphoric mood of these patients. The impact of anger and frustration on exacerbation of pain and treatment acceptance has not received adequate attention. It would be reasonable to expect that the presence of anger may serve as an aggravating factor, associated with increasing autonomic arousal, and having the effect of blocking motivation and acceptance of treatment oriented toward rehabilitation and disability management rather than cure, which are often the only treatments available for chronic pain.

6. Biobehavioral Perspective

A biobehavioral perspective of chronic pain needs to consider the factors discussed above and their mutual interrelationships in the explanation of chronic pain. The existence of a physiological disposition or diathesis is one important component. This predisposition is related to a reduced threshold for nociceptive stimulation and can be determined by genetic factors or acquired through early learning experiences. For example, Mogil (1999) showed that large genetic variations in individual pain sensitivity exist. Very impressive evidence for the role of early traumatic experience comes from the work of Anand et al. (1999), who showed that minor noxious experience in neonate rats leads to dramatic alterations (sensitization) in nociceptive processing in the adult organism. A further component of the biobehavioral model is a response stereotype of a particular bodily system, such as exaggerated muscular responses of the lower back muscle to stress and pain that is based both on the diathesis and on aversive experiences present at the time of the development of the response. These aversive stimuli may include personal or work stress or problematic occupational conditions and will lead not only to painful responses but also to avoidance behaviors and associated maladaptive cognitive and affective processes. The cognitive evaluation of these external or internal stimuli is of great importance in the pain response, as discussed above.

7. Memory for Pain

An important maintaining factor in this chronicity process is the development of pain memories. These pain-related memories may be explicit or implicit and may subsequently guide the patient's experience and behaviors (Erskine et al. 1990). For example, pain patients have a tendency to remember preferentially negative and pain-related life events and show a deficit in the retrieval of positive memories. The experience of chronic pain also leads to the development of somatosensory pain memories for example, an expanded representation of the affected body part in the primary somatosensory cortex. This expanded cortical representation is accompanied by increased sensitivity to both painful and nonpainful stimuli, and may be further enhanced by learning processes or attention to painful stimulation. A dramatic example of a learned memory for pain has been found in phantom limb pain patients (Flor et al. 1995). In upper extremity amputees, the magnitude of the phantom limb pain was found to be proportional to the amount of reorganization in the primary somatosensory cortex, namely, the shift of the cortical mouth representation into the area where the amputated limb was formerly represented. The brain obviously maintains a memory of the former input to the deafferented area. Subsequently, stimulation stemming from areas adjacent to the deafferented zone elicit sensations and pain in the now absent limb. Phantom sensations and cortical reorganization are absent in congenital amputees. The focus of the biobehavioral perspective is thus on the patient and not just on the symptoms or the underlying pathology. This focus also requires that the treatment of the patients be tailored not only to medical factors, but that it incorporates psychosocial variables that may often be predominant in states of chronic pain (see *Chronic Pain: Models and Treatment Approaches*).

8. Future Perspectives

At the end of the twentieth century, there has been an enormous increase of research in the neurosciences, where the mutual interactions of molecular and genetic as well as neuronal and behavioral variables have begun to be examined. This integration of the psychosocial and biomedical perspective is much needed and will bring substantial advances in our understanding of the phenomenon of pain. A further important development is the recent finding in neuroscience that the adult brain is enormously plastic and that learning processes as well as training measures have great impact on these plastic changes (Recanzone 2000). This opens the door for new approaches to treatment. It will also aid in delineating and eliminating the factors that lead to resistance to the treatment of chronic pain syndromes.

See also: Chronic Illness, Psychosocial Coping with; Chronic Pain: Models and Treatment Approaches; Pain, Management of; Stress and Coping Theories; Stressful Medical Procedures, Coping with

Bibliography

- Anand K J, Coskun V, Thiruvikraman K V, Nemeroff C B, Plotsky P M 1999 Long-term behavioral effects of repetitive pain in neonatal rat pups. *Physiol. Behav.* **66**(4): 627–37
- Asmundson G J, Norton P J, Norton G R 1999 Beyond pain: The role of fear and avoidance in chronicity. *Clinical Psychology Review* **19**: 97–119
- Erskine A, Morley S, Pearce S 1990 Memory for pain: A review. *Pain* **41**(3): 255–65
- Fernandez E, Turk D C 1989 The utility of cognitive coping strategies for altering pain perception: A meta-analysis. *Pain* **38**(2): 123–35
- Flor H, Birbaumer N, Turk D C 1990 The psychobiology of pain. *Advances in Behav Res Ther.* **12**(2): 47–84
- Flor H, Elbert T, Wienbruch C, Pantev C, Knecht S, Birbaumer N, Larbig W, Taub E 1995 Phantom limb pain as a perceptual correlate of cortical reorganization—following arm amputation. *Nature* **375**: 482–4
- Fordyce, W E 1976 *Behavioral Methods for Chronic Pain and Illness*. Mosby, St. Louis, MO
- Gatchel R J, Turk D C 1999 *Psychosocial Factors in Pain: Critical Perspectives*. Guilford Press, New York
- Melzack R A, Wall P D 1965 Pain mechanisms: A new theory. *Science* **150**: 3699–709
- Melzack R A, Wall P D (eds.) 1994 *The Textbook of Pain*, 3rd edn. Churchill Livingstone, Edinburgh, UK
- Merskey H 1986 Classification of chronic pain: Descriptions of chronic pain syndromes and definitions of pain terms. *Pain* **3**: 1–225
- Mogil J S 1999 The genetic mediation of individual differences in sensitivity to pain and its inhibition. *Proceedings of the National Academy of Sciences of the United States of America* **96**(14): 7744–51
- Recanzone G 2000 Cerebral cortex plasticity: Perception and skill acquisition. In: Gazzaniga M S (ed.) *The New Cognitive Neurosciences*. MIT Press, Boston, MA
- Romano J M, Turner J A 1985 Chronic pain and depression: Does the evidence support a relationship? *Psychology Bulletin* **97**(1): 18–34
- Turk D C, Meichenbaum D, Genest M 1983 *Pain and Behavioral Medicine: A Cognitive–Behavioral Perspective*. Guilford Press, New York
- Turk D C, Rudy T E 1988 Toward an empirically-derived taxonomy of chronic pain patients: integration of psychological assessment data. *Journal of Consulting and Clinical Psychology* **56**: 233–8
- Whitehead W E, Crowell M D, Heller B R, Robinson J C, Schuster M M, Horn S 1994 Modeling and reinforcement of the sick role during childhood predicts adult illness behavior. *Psychosomatic Medicine* **56**: 541–50
- Woolf C J, Mannion R J 1999 Neuropathic pain: Aetiology, symptoms, mechanisms, and management. *Lancet* **353**: 1959–64

H. Flor

Copyright © 2001 Elsevier Science Ltd. 10995
All rights reserved.

Pain, Management of

1. The Physiology of Pain

Pain is defined as 'an unpleasant sensory and emotional experience associated with actual or potential tissue damage, or described in terms of such damage.' It is the most common complaint voiced by patients when they seek medical attention. The ability to perceive pain serves a protective function for the individual. The sensation warns of impending injury and elicits reflex and behavioral responses that keep tissue damage to a minimum (see Livingston 1998, Melzack 1973, Rey 1995).

In many instances, noxious stimuli applied at the skin surface (blunt trauma, burns, a surgical incision) serve to activate pain receptors. These are special nerve cells, termed nociceptors, which convert (transduce) the energy delivered at the periphery to neural impulses. Transduction is the first of four processes allowing the energy that produces noxious stimuli to result eventually in the perception of pain. Once this initial energy is transduced into electrical impulses, nerves carry the stimulus to higher levels of the nervous system. This activity is termed *transmission*, and it allows the noxious stimuli to be conveyed from the skin to the spinal cord as a series of electrical messages. As the signals arrive in the dorsal horn of the spinal cord, they are altered by other impulses within the cord and by descending impulses from higher neural centers. This is termed *modulation*. These modified and processed impulses continue upward towards the brain and allow a person to experience the perception of pain. This is an individual and highly subjective phenomenon. Another person can not reliably measure it. There is no objective appraisal of the sensation that an individual calls pain.

An understanding of the four events is critical to an appreciation of the mechanisms by which pain can be managed. The physician targets therapies to prevent the electrical signals from reaching the cortex and limbic systems of the brain where pain enters a person's consciousness. Thus treatments are aimed at interrupting pain information as they are transduced to nerve impulses; transmitted towards the spinal cord; and modulated within the cord; or by modifying the neural impulses as they enter the brain where the actual perception occurs.

Minimizing the pain a patient perceives has value on a number of fronts: humanitarian, physiologic, functional, and economic. Progress made in this area of medicine was begun by anesthesiologists in the 1980s when they began to organize services to treat surgical inpatients immediately following operations. These activities were logical extensions of the work they already performed during surgery. Acute pain services are now well established divisions of most hospitals in the United States.

Workers in the field of pain management divide pain into three separate categories: acute, chronic non-malignant, and cancer pain. Management varies depending on the etiology. Acute pain is experienced following trauma, elective surgical procedures, or the exacerbation of chronic medical conditions (e.g., sickle cell crisis). Tissue damage initiates a cascade of physiological and behavioral changes. These have the potential to progress into chronically painful states if left unattended.

1.1 Acute Pain

Mechanisms to manage operative and postoperative pain focus on the delivery of medications to specific sites within the body. Local anesthetics, such as lidocaine can be injected under the skin prior to an incision. This is exemplified by the use of a nerve field block performed immediately before an inguinal hernia repair. The medication is injected at the site of the surgery and given time to work. Local anesthetics act by blocking the conduction of impulses down nerve fibers. These nerves, then, cannot transmit to the central nervous system (CNS) the information that tissue damage (the surgical incision) is occurring. Since the CNS never receives the nerve impulses, the patient does not perceive pain.

More sophisticated techniques include the deposition of local anesthetics into areas wherein numerous nerves converge. Because these areas are densely packed with passing nerve fibers, a small amount of anesthetic can effectively block pain messages from a large segment of the body. Examples of these procedures include brachial plexus (an area of convergent nerves from the arm and hand) blockade prior to hand surgery, and epidural (a space just outside the spinal cord) blockade, which is used to blunt the intensity of labor pains prior to childbirth. One problem of the use of local anesthetic is that these agents are generally fairly short-acting. The duration of the blockade can be prolonged by co-injection of agents which decrease blood flow in the immediate area of injection. Less blood flow to the area of blockade means the anesthetic agent is taken away by the bloodstream more slowly, prolonging the presence, and effect, of the local anesthetic. Another method physicians use to prolong local anesthetic action is to pump anesthetic continuously into the needed area through a plastic tube or catheter that has been inserted into that region. This technique allows pain control for very long periods of time. For example, a catheter placed in a position that blocks pain from the chest (thorax) can be used to manage pain related to chest (thoracic) surgery for five days following the operation. A catheter is placed at the appropriate level within the epidural space and taped to the patient's skin. An infusion pump is programmed to deliver local anesthetic, opioid (such as morphine), or a mixture of both agents at an hourly

rate. When the mixture of agents is used, very dilute concentrations of each are employed. Inhibiting the transmission and modulation of nerve impulses with different classes of agent allows for synergy of effect. Excellent control of very strong pain can be achieved with the use of epidural catheter placement and the constant infusion of analgesic mixtures. Sensitivity of the affected area returns quickly when the drugs are stopped.

Unfortunately, local anesthetics block not just pain sensors, but all nerves. Thus, they produce a profound numbness and, in some cases, difficulty in moving muscles. In addition, they are only active for a short time, or require the placement of a catheter near the nerves to be affected. Therefore, other means are necessary to deal with more long-duration pain. For example, systemic medications can be delivered intravenously to blunt the perception of pain. Newer technology in the form of microprocessors and physically small infusion pumps have made the management of acute, post-surgical pain more precise. Patient-controlled analgesia (PCA) pumps provide an infusion of morphine or similar drugs at a constant background rate. This is usually 1 or 2 milligrams of morphine over the course of an hour. Furthermore, a patient can activate the pump to deliver additional morphine by the push of a button. Dosage is often 2 milligrams which can be repeated every sixth minute. This machinery allows the patient substantial control of how much morphine he or she is receiving so that the dose delivered matches his or her needs. The background infusion guarantees that some analgesia is delivered to the bloodstream to compensate for normal clearance of the analgesic by the liver and kidneys. The patient-controlled aspect enables the patient to raise immediately the blood level of the morphine should the perception of pain increase in intensity. This commonly correlates with the patient's degree of physical activity and can often be anticipated. A patient who is comfortable while lying down may feel pain when getting into a chair. In this instance, self-dosing permits the level of opioid to match the need for analgesia on a minute by minute basis. Since the pump must be activated before delivering additional medication, the risk of overdosage is minimized. Patients who are sleepy from a high serum level of morphine will not push the button for additional medication. In some instances, morphine is not the appropriate agent to use as it produces significant side-effects in some patients. In these cases, other opioids such as fentanyl, demerol, or dilaudid can be used. Individual preferences and the reactions to treatment are used to guide further therapy.

Progress has been made in the development of non-steroidal, anti-inflammatory drugs (NSAIDs). Aspirin, acetomenophin, and ibuprofen are well known examples of these drugs. NSAIDs are a class of medications that diminish swelling and decrease the perception of pain. They act both at painful sites and

in the CNS by inhibiting the enzyme cyclo-oxygenase. Intramuscular and intravenous NSAIDs have been developed so that patients in the immediate post-operative period and still unable to consume pills can benefit from the agents. The newest drugs in the class are highly selective for the desired effects and show little of the side effects which have been problems with the long-established NSAIDs. Patients are often offered intravenous doses of an NSAID in the post-operative period if they experience pain breaking through despite continuous epidural infusion of local anesthetics (see above).

New approaches to delivering analgesic drugs into the body are also being developed. Intranasal (nasal spray), pulmonary (inhaler), and transbuccal (under the tongue) routes of opioid delivery are all being researched with the goal of providing immediate pain relief. A transdermal system of the opioid fentanyl delivery is currently available but is rarely used to treat acute pain. The skin directly under the sticky transdermal patch absorbs fentanyl which acts to gradually release the medication over hours or even days. Unfortunately, changing the blood level of a drug is slow when using transdermal patches, and so the risk of drug overdose in this patient population is higher. Transdermal patches are rarely used in the post-operative phase of patient care, but are more commonly used for chronic and cancer-related pains.

1.2 Chronic Pain

Chronic, noncancer related pain affects 75 million Americans. Pain decreases the quality of life of these people and impairs their ability to work productively. Chronic pain is defined as pain that persists for longer than the expected time frame for healing, or pain associated with progressive diseases. Many patients suffer from clinical syndromes for which there are no x-ray, laboratory, or other physical evidence of something wrong, and so diagnoses rely on clinical criteria alone. These syndromes include many kinds of headache, myofascial pain syndromes (for example chronic jaw pain), fibromyalgia, neuropathic pain, low back pain, and central pain syndromes. Little is known about the underlying causes of these syndromes. Patients do, however, present with the complaint of pain. Other patients have disorders that are well described, e.g., osteoarthritis, but for which we have no cure.

Irrespective of the cause, the effects of chronic pain on a patient's life tend to be more pervasive than that of acute pain. Chronic pain affects all of the patient's social relationships, personality, and mood. Typically, people with chronic pain tend to have sleep disturbances, generalized fatigue and a decrease in their overall level of functioning. In many, the perception of pain cannot be eliminated, only lessened. The goal becomes one of managing the pain as well as is

possible while improving the person's involvement and function in his life. To this end, a multidisciplinary approach that focuses on all aspects of the patient's condition seems to have the best results. The core team consists of a pain management specialist along with representatives from pharmacy, physical and occupational therapy, psychology, and internal medicine. The care team tailors the care plan according to the needs of the individual patient. An open discussion of goals is necessary to assess objectively if progress is being made. Often the goals of productive activity and a return to work are stressed.

Treatment of chronic, non-cancer related pain conditions requires the use of analgesic medications. NSAIDs are often the mainstay of therapy. Long-term use of the agents carries the risk of easy bruising, renal dysfunction, and the development of gastric ulcers. As mentioned above, the newer NSAIDs help to limit these side effects. Patients must be warned of these complications and followed closely. The agents are potent analgesics that improve the quality of life for those who suffer from many chronic pains, especially arthritis.

Long-acting opioids can effectively control pain in a subset of chronic pain patients. Patients must be assessed carefully before the start of therapy. The potential for substance abuse or addiction must be considered even though they are not commonly encountered in patients with chronic pain. Guidelines exist which help to clarify the conditions under which the prescribing of opioids is appropriate in the treatment of chronic pain states. Common adverse side effects include constipation, pruritis, sedation, and impaired cognition. Although most of the side effects diminish over time, constipation usually persists and must be treated.

Antidepressants are effective in the treatment of many painful conditions. Their mechanism of action is unclear, but they seem to act by increasing the levels of norepinephrine and serotonin at the nerve synapses of the central nervous system. In addition to their direct effect in diminishing the perception of pain, the agents also are useful as an aid to sleep. Patients take the medication in the late evening and report improvements in their ability to fall asleep and stay asleep throughout the night. The agents have the potential for bothersome side effects which include dry mouth, blood pressure changes, urinary retention, and, rarely, heartbeat irregularities. The drugs are used as a part of an integrated care plan and improve the overall quality of the patient's life.

Anticonvulsants are effective in the treatment of many chronic pain states. The mechanism of action is again unclear. They may suppress the production of nerve impulses in nerve cells that are critical in developing the perception of pain. A newer anticonvulsant, gabapentin, provides symptomatic relief with fewer side effects than older drugs of this class. Patients notice sedation and episodes of dizziness

when they begin taking the drug, but acclimate to it over a matter of days.

More invasive methods for the treatment of chronic painful states have been tried with no proven benefit. Cutting the spinal cord pathways that transmit pain messages to the brain produces temporary (6–12 month) pain relief, but the pain then returns, sometimes worse than ever. The direct injection of local anesthetics into the most painful muscle areas of myofascial pain syndrome seems to provide symptomatic relief for some of the patients. The procedures have not been tested in a randomized or blinded manner, however, and the benefits may be the result of the placebo reaction. Epidural steroid and local anesthetic injections have also been used for those who complain of vague lower back pains. The steroids are known to decrease inflammation, edema, and decrease the synthesis of prostaglandins (substances known to mediate the perception of pain). Again, the patients often report benefits in their symptoms, but there is no clear mechanism to explain the improvement.

1.3 Cancer Pain

The prevalence of cancer pain ranges from 30–50 percent of patients who are undergoing active treatment for solid tumors. The number climbs to 80–90 percent for those who suffer from advanced forms of the disease. Approximately three-quarters of these chronic pain syndromes result as a direct effect of the tumors themselves. The remaining complaints are related to the therapies administered to manage the primary or metastatic aspects of the cancer.

Neoplastic (tumor) invasion of bone or connective tissue is a major cause of severe pain. The spine is the most common site of bony metastasis. Many patients with cancer have back pain because the cancer has spread to this bony area. Continued growth of the tumor causes spinal cord and nerve root compression. Cancerous growth in this area will result in pain, as well as paralysis and the loss of bowel and bladder control for the patient. Obstruction and compression of the bowel produces intense visceral complaints. Patients will present with persistent abdominal pain, constipation, nausea, and vomiting. The tumor will have to be surgically removed, sometimes resulting in nerve injury-induced (neuropathic) pain. Radiation, frequently used to shrink tumors, can cause fibrosis, which may in turn damage peripheral nerve and central nervous system tissue. These injuries frequently cause a chronic neuropathic-type pain, along with tingling and numbness in the area.

The mainstay for management of cancer-related pain is the use of oral analgesics leading up to opioids in progressively larger doses as the disease progresses. The World Health Organization has devised an algorithm for use of oral analgesics that matches the intensity of the pain with the potency of the medication

prescribed. Although the oral route is commonly used, the transdermal approach has gained popularity because of the ease of administration. Each patch is worn for approximately three days before it must be replaced. The system has the advantage of infrequent dosing, but as mentioned above, keeping the right blood level is difficult. Most practitioners will start the patient with a relatively low dose of transdermal fentanyl, but then give the patient morphine pills to use as needed in order to supplement the relief provided by the patch. Individualization is the key to successful treatment. Each patient must be evaluated and a treatment plan formulated based on that individual's needs. In some, radiation and chemotherapy will effectively reduce the tumor mass and diminish pain. In others, there will be little hope of shrinking the tumor and the only recourse is to make these patients as comfortable as possible. This 'end-of-life' strategy is termed palliative care.

The success of continuous infusions of local anesthetics in the operating room and immediately postoperatively has prompted the developments of portable and reprogrammable infusion pumps for use by cancer pain patients. In some cases, these pumps are implanted under the skin. Pumps allow the continuous infusion of strong opioids through catheters to sites directly within the spinal canal, either epidurally or directly onto the spinal cord itself (intrathecally). The opioids act directly on nerve cells in the spinal cord to inhibit pain messages, resulting in profound pain relief. The epidural or intrathecal catheters are tunneled under the skin and attached to a pump that has within it a reservoir to hold the opioid. The machinery is programmed to deliver a preset amount of opioid as needed to provide adequate pain relief. On a regular basis, the reservoir is refilled to allow for continuous pain relief as needed.

The most invasive therapies used for those with advanced stages of cancer involve the use of ablative surgical or anesthetic procedures. As mentioned above, spinal cord incisions can produce profound pain relief. Because pain recurs over the long term, however, this procedure is only used near the end of life. Pure alcohol is toxic to nerves. Thus, another fairly drastic procedure involves the injection of pure alcohol into individual nerves of the spinal cord to destroy those nerves carrying pain messages to the CNS. In these ways, patients near the end of their life can achieve a degree of pain relief that is not attainable with less invasive measures.

Cancer pain must be viewed in the context of a progressively advancing illness. The optimum management allows for the treatment to match the intensity of discomfort. The smallest dose of the most benign medication that provides relief is used first. When the intensity of pain increases, stronger medications and more invasive techniques are used. In this way, a comprehensive range of agents and techniques are used in sequence to provide optimal pain relief.

2. *Novel Approaches to the Relief of Pain*

2.1 *Transplantation*

Recent laboratory studies have shown that transplantation of adrenal medullary tissue or of isolated chromaffin cells into CNS pain modulatory regions can reduce pain sensitivity in both acute and chronic (including neuropathic) models in rats. The analgesia produced by these transplants probably results from the release of both opioid peptides and catecholamines such as adrenaline and noradrenaline since it can be blocked or attenuated by both opiate and adrenergic blockers. Studies indicate that even over long periods there is no apparent development of tolerance.

Positive results have also been obtained in preliminary clinical studies using transplants of human adrenal gland pieces to relieve intractable cancer pain in 25 patients. Pain scores and morphine intake were measured prior to transplantation and at regular intervals following the graft implantation. Biochemical assays were performed on a sample of the patients' cerebrospinal fluid (CSF) throughout the study to assess levels of catecholamine and met-enkephalin (an opiate manufactured by the body) as an indication of graft viability. Sixteen out of 20 patients reported long lasting pain relief from these transplants (more than a year in some patients). Narcotic analgesic (e.g., morphine) intake was eliminated, reduced or at least stabilized in all patients who completed the study. These studies indicate that even over long periods, there is no development of opiate tolerance. Histological data was obtained at autopsy from some patients who had transplants for over a year. Study of the spinal cord revealed intact adrenal medullary tissue fragments distributed over several levels, with no signs of inflammation or rejection. The results of these clinical trials are encouraging over all and suggest that chromaffin cell transplants can provide relief from intractable pain. With appropriate modifications to optimize graft preparations and survival, it is likely that this approach can become an alternative means of controlling pain refractory to traditional pharmacotherapies (see Michalewicz et al. 1999, Lazorthes et al. 2001).

2.2 *Gene Therapy*

Another new approach to chronic pain is that of gene therapy. This method so far is only at the animal testing stage, but appears to be very promising. Different approaches have been taken to change the genetics of nerve cells that are involved in pain so that that function of the cells is disrupted. The first of these involves the injection of 'antisense' on to the spinal cord. Genes are essentially codes that cells read, and which cause the cell to make a particular chemical, as,

for example, a protein. Some of these chemicals are critical to the transmission of pain information from one cell to another. Antisense is essentially the code for one of these chemicals, but in reverse order. This effectively disrupts the cell's ability to read the 'real' gene and so inhibits its ability to make the chemical. If antisense is injected on to surface of the spinal cord, some of it will find its way into the correct pain neurons and so, to a greater or lesser extent, inhibit transmission of pain messages (see Peric et al. 2001).

Applying antisense to spinal cord nerve cells appears to work in animals, but is somewhat nonselective as to which cells it goes into, and is not very efficient in getting it there. For this reason, other researchers have been working on ways to use viruses as 'vectors' to carry genes that may inhibit pain transmission into the nerve cells that carry that information. One method involves putting the gene for naturally-occurring opioids into either adenoviruses (the kinds that causes the common cold) or herpes viruses (the kinds that causes cold sores). These viruses are very efficient at getting these genes into spinal nerve cells and produced long-lasting analgesia in animals. The adenovirus is injected directly into the spinal cord itself, whereas the herpes virus is placed near pain receptors outside the CNS where it is taken up and carried back to the spinal cord. The herpes virus approach has also been used to carry antisense directly into the pain-conducting nerves (see Wilson et al. 1999). All of these approaches present new and exciting approaches to the treatment of chronic pain.

See also: Chronic Illness, Psychosocial Coping with; Chronic Illness: Quality of Life; Chronic Pain: Models and Treatment Approaches; Motivation, Neural Basis of; Pain, Health Psychology of; Pain, Neural Basis of; Somatosensation

Bibliography

- Lazarthes Y, Bès J C, Sol C, Pappas C D 2001 Intrathecal chromaffin cell allograft for cancer pain. In: Burchiel (eds.) *Surgical Management of Pain*. Thieme Medical Publishers Inc., New York, pp. 973–9
- Livingston W K 1998 *Pain and Suffering*. IASP Press, Seattle, WA
- Melzack R 1973 *The Puzzle of Pain*. Basic Books, New York
- Michalewicz P, Laurito C E, Pappas G D, Yeomans D C 1999 Purification of adrenal chromaffin cells increases autinociception efficiency of xenotransplants in the absence of immunosuppression. *Cell Transplant* 8: 103–9
- Peric V, Lu Y, Laurito C E, Yeomans D C 2001 Combined effects of N-type calcium channel blockers and morphine on a-delta US. C fiber mediated nociception. *Anesthesiology and Analgesia* 92: 39–243
- Rey R 1995 *The History of Pain*. Harvard University Press, Cambridge, MA
- Wilson S I, Yeomans D C, Beuder M A, Lu y, Glorioso J 1999 Antihyperalgesia effects of delivery of encephalins to mouse

nociceptive neurons by herpes virus encoding proenkephalins. *Proceeding of the National Academy of Sciences of the USA* 96: 3211–16

D. C. Yeomans, C. E. Laurito and G. D. Pappas

Pain, Neural Basis of

The term 'pain' has several different connotations that are a potential source of confusion in its usage. The most frequent of such confusions consists in an equivocation of first- and third-person aspects of the target phenomenon (Metzinger 2000). There is the consciously experienced subjective qualitative aspect of pain. In this sense, to feel pain indicates a state of suffering. This aspect is only experienced from the first-person perspective and, arguably, is not directly accessible for empirical research operating from an objective, third-person perspective. Third-person aspects of pain are constituted by its neural underpinnings (e.g., the minimally sufficient neural correlate of the subjectively experienced, conscious quality of pain) and the functional properties, that is, the causal role, played by them in generating overt behavior, internal reactions, the formation of memories, etc. In this sense, pain is the percept that results from the activation of a specific sensory system, the nociceptive system, which is activated by external stimuli that threaten the integrity of the body, and by internal states of tissues such as an inflammatory process. As such, pain is a sensory experience similar to the experience of hearing a sound or tasting food as a consequence of the activation of the auditory, gustatory, or olfactory system. A comprehensive theory of pain will have to do justice to both aspects, the third-person and first-person aspects, of pain.

Since the 1970s the following definition by the International Association for the Study of Pain has gained widespread acceptance (Merskey et al. 1979):

- Pain is an unpleasant sensory and emotional experience

- associated with actual or potential tissue damage
- or described in terms of such damage.

This definition emphasizes that pain has several components, including at least a sensory one and an affective one. In most instances, pain is not a generalized feeling (*Allgemeingefühl*) such as hunger and thirst, but a localized sensation. The capacity to locate pain is due to the encoding of spatial information in the nociceptive system. Likewise, fluctuations in the perceived intensity of pain can be related to intensity coding in the nociceptive system. The discriminative power and the precision of stimulus encoding in the nociceptive system have traditionally been underestimated. In addition to this sensory-discriminative component, pain almost inevitably encompasses an affective component, mostly of a negative hedonic

quality, which is not a simple function of sensory intensity or quality. While pain may be a particularly unpleasant sensory experience, affective components are present in other sensory systems as well: a person may like or dislike a certain type of music, and a meal may have a pleasant or unpleasant taste. These affective or emotional components of sensory experience provide powerful drives to direct behavior (e.g., withdrawal, guarding, or avoidance behavior). The unpleasant emotional aspects of pain perception are therefore called the affective-motivational component of pain. The aim of this article is to present the neural basis of both components of pain from the point of view of sensory physiology. In the last section, an outlook will be given as to how sensory physiology may provide hints to understanding the subjective experience of feeling pain.

1. *The Nociceptive System*

Many tissues such as the skin, mucous membranes, muscles, joints, bones, and parts of the viscera (with the notable exception of the brain itself) are innervated by free endings of thin primary afferent nerve fibers (Raja et al. 1999). The adequate stimulus to activate these nerve endings is a strong mechanical or thermal stimulus, or one of a variety of chemicals that are either applied from the outside (as with a nettle sting or the application of hot peppers to mucous membranes) or generated within the tissue (such as bradykinin, serotonin, histamine, prostaglandins, etc.). Because they respond to several stimulus modalities, these nerve endings are said to be polymodal. The common denominator of these stimuli is that they indicate either actual or impending tissue damage. Sherrington summarized this concept in the term 'noxious stimulus,' and the sensory system that processes information on noxious stimuli is called the nociceptive system. More recently it was found that polymodality does not depend on actual tissue damage but may also be a property of the ion channels present in free nerve endings. Many nociceptive nerve endings express the so-called vanilloid receptor channel VR1 (Caterina and Julius 1999), which can be activated by a physical stimulus (heat), and by both exogenous (capsaicin and other hot spices) and endogenous chemicals (anandamide and other derivatives of membrane lipids).

Some primary nociceptive afferents (A-fibers) conduct nerve impulses relatively rapidly (about 20 ms^{-1}), because of the saltatory impulse conduction that is enabled by the presence of a myelin sheath; most of the primary nociceptive afferents are unmyelinated C-fibers and conduct slowly (about 1 ms^{-1}). A-fibers are involved in rapid behavioral responses and in motor reflexes. C-fibers are involved in slower signaling processes, including the perception of a delayed 'second pain' after a single brief stimulus such as a

pinprick, changes in the sensitivity of the nociceptive system (see Sect. 4), control of local blood vessels, and trophic functions (Holzer 1992).

Primary nociceptive afferents make synaptic contact with second-order neurons in the dorsal horn of the spinal cord. Some nociceptive specific neurons have no other input, whereas another class of neurons receives convergent input from touch receptors in the skin. These neurons encode a wide range of mechanical stimulus intensities and are therefore called wide-dynamic-range neurons (Willis 1985). Spinal nociceptive neurons are subject to modulation by descending pathways from the brain stem, which are mostly inhibitory and utilize endogenous opioids, nor-adrenaline, and serotonin as neurotransmitters. The net result of primary afferent inputs, descending modulation, and local interactions in the spinal cord dorsal horn is then projected to spinal motor nuclei in the ventral horn, spinal autonomic nuclei in the lateral horn, to the brain stem and to the contralateral somatosensory thalamus.

Via the spinoreticular and spinomesencephalic tract projection to the brain stem, noxious stimuli may activate autonomic reflex centers of the cardiovascular and respiratory system, the descending inhibition of the spinal cord dorsal horn, and the ascending reticular activating system (ARAS) that projects to the cerebral cortex. These projections are the neural basis for many autonomic reflexes elicited by pain, for the fact that one pain may inhibit another pain, and for the strong arousal reaction to painful stimuli.

Via the spinothalamic tract, the nociceptive pathways reach two groups of nuclei in the thalamus (Craig and Dostrovsky 1998). The lateral group is situated in the ventrobasal complex and is identical to the relay station of the tactile system. These thalamic nuclei project to the primary and secondary somatosensory cortex (SI and SII), and to the insula. The medial group is mostly situated within the white matter of the thalamus (intralaminar nuclei). These thalamic nuclei project to the anterior cingulate gyrus, the basal ganglia, and also nonspecifically to wide parts of the cerebral cortex. Since its original suggestion in the late 1960s (Melzack and Casey 1968), the concept that the lateral system processes the sensory discriminative component, and the medial system the affective-motivational component of pain, is still a valid concept (Treede et al. 1999).

2. *Sensory Discriminative Functions*

The sensory discriminative component of pain may be divided into at least three aspects: stimulus localization, intensity discrimination, and quality discrimination. The human capacity to perceive the location of tissue damage is used in everyday medical practice, when doctors ask the question 'where does it hurt?' Sometimes, tissue damage and perceived pain are in

different locations, as in the case of a myocardial infarction, which may cause pain in the left arm. This phenomenon (referred pain) is due to convergence of visceral and cutaneous afferents in the spinal cord. The activity of the spinal neurons does not reveal the afferent source, and higher centers are assumed to project the resulting sensation along the more common pathway from the skin. Because of the mislocalization of pain originating from damage to the viscera and the precise localization of tactile stimuli at the fingertips, localization of noxious stimuli has traditionally been ascribed to simultaneous activation of the tactile system. Tactile acuity, however, rapidly drops outside the 'foveal' areas of the fingertips and lips. On the back of the hand, sequential tactile stimuli must be about 1 cm apart to be perceived as being given to separate locations. When laser-radiant heat pulses (which do not activate tactile afferents) are given to the back of the hand, the sequential spatial discrimination threshold is also about 1 cm (Schlereth et al. 2001).

These psychophysical findings demonstrate that the nociceptive system in humans provides sufficient spatial information to account for our capacity to know where it hurts. What is the neural basis for this capacity? Receptive fields of nociceptive neurons show somatotopic organization in the spinal cord dorsal horn (Woolf and Fitzgerald 1986), lateral thalamus (Albe-Fessard et al. 1985), and the primary somatosensory cortex (Kenshalo and Willis 1991). Receptive field sizes in SI are smaller than in the spinal cord and thalamus, possibly due to lateral inhibition, and match the pain localization capacity. Thus, all existing evidence favors the view that SI is involved in stimulus localization for the nociceptive system as well as the tactile system. Indirect evidence for this concept is provided by studies in humans that use measures of brain perfusion as parameters of activation (by positron emission tomography: PET). These studies suggest that only tasks with a stimulus-localization component activate SI, whereas studies with fixed stimulus location do not (Treede et al. 1999).

The capacity to encode different intensities of noxious stimuli is one of the criteria for identifying nociceptive neurons (Willis 1985). Intensity coding has been demonstrated for nociceptive neurons in SI and to a certain extent in SII, but also in medial thalamic nuclei. Intensity coding is a poor criterion for identifying neurons involved in sensory discriminative aspects of pain, because the affective-motivational component of pain also depends on stimulus intensity. In a study that demonstrated the different relative unpleasantness of experimental painful stimuli, perceived intensity and unpleasantness were both nevertheless related to stimulus intensity (Rainville et al. 1992).

Pain may have different qualities such as burning, stinging, or aching. Neither the number of subjectively discriminable qualities nor their neural basis are known with sufficient precision. The polymodality of primary nociceptive afferents seems to contradict the

capacity for quality discrimination, but the response profiles to different noxious stimuli differ between different afferents. Quality discrimination in the nociceptive system is therefore likely to be due to a population code, similar to the encoding of the taste qualities in the gustatory system.

Current evidence supports the traditional view that the lateral nociceptive system subserves the sensory discriminative component of pain. Different aspects of this pain component (detection, localization, intensity discrimination, quality) may be processed in parallel by separate pathways. In contrast to traditional views, there is no evidence that the tactile system participates in any of these functions. Thus, nociception is established as a sensory modality within the somatosensory system.

3. Affective Motivational Functions

The affective-motivational component of pain sensation encompasses several aspects that are closely related: the negative hedonic quality and emotional reactions, an increase in the arousal level and stimulus-related selective attention, and the drive to terminate the stimulus causing this sensation. Some of these functions can be considered second-order sensory processing, whereas the latter is a premotor function. The affective-motivational component of pain is classically associated with the medial nociceptive system, which in turn is connected to the limbic system.

One part of the medial nociceptive system, the anterior cingulate cortex, has recently gained much publicity, because virtually all PET studies of acute pain gave evidence of activation in that area, including one study in which pain was elicited as an illusion by the interaction of two nonpainful stimuli (Craig et al. 1996). Electrophysiological recordings in humans also show activity in this area (Treede et al. 1999). The anterior cingulate cortex is a functionally heterogeneous brain area that has been implicated in the integration of affect, cognition, and response selection in addition to aspects of social behavior (for review, see Devinsky et al. 1995). Passive functions (emotion, attention) are represented more frontally, whereas a premotor part of the anterior cingulate cortex is situated more posteriorly, below the supplementary motor area. It is still debated, whether the anterior cingulate cortex may contain a nociceptive-specific area, or whether painful stimuli nonspecifically recruit several parts of the large cingulate gyrus.

The affective-motivational component of pain may also be processed in the insula. The contralateral insula was activated almost as frequently in human PET studies of acute pain as the anterior cingulate cortex (Casey and Minoshima 1997). Microstimulation in a thalamic nucleus that projects to the insula elicited pain with a strong affective component only in those patients who had previously experienced such

pain either due to panic attacks or due to angina pectoris (Lenz et al. 1995). These observations suggest that the insula may be part of a sensory limbic projection pathway for pain sensation. The insula projects to the amygdala (Augustine 1996), which is a part of the limbic system that is associated with emotions.

4. Plasticity of the Nociceptive System

As in all other sensory systems, the repeated presentation of noxious stimuli or the exposure of the nociceptive system to stimuli of long duration may lead to a reduction in the intensity of the evoked response (habituation and adaptation). A unique property of the nociceptive system is that under many circumstances the prior presentation of the adequate stimulus may actually enhance subsequent responses (Treede et al. 1992). This enhancement is called sensitization and it may occur at both the peripheral

terminals of primary nociceptive afferents and within the central nervous system. Sensitization refers to a leftward shift in the function that relates the neural response to stimulus intensity. It is characterized by a drop in threshold and an increase in the response to suprathreshold stimuli. Spontaneous activity may also result from sensitization.

The perceptual correlate of sensitization is hyperalgesia, which is characterized by a drop in pain threshold and an enhanced painfulness of supra-threshold stimuli. Even minor tissue injury elicits a transient state of hyperalgesia, both of the injured tissue itself (primary hyperalgesia) and in surrounding uninjured skin (secondary hyperalgesia). Primary hyperalgesia is characterized by a lowered pain threshold for both mechanical and heat stimuli. Secondary hyperalgesia is characterized by a lowered pain threshold for mechanical stimuli only.

In primary hyperalgesia, the heat pain threshold may drop below the normal body temperature, which then starts to act as an adequate stimulus for the

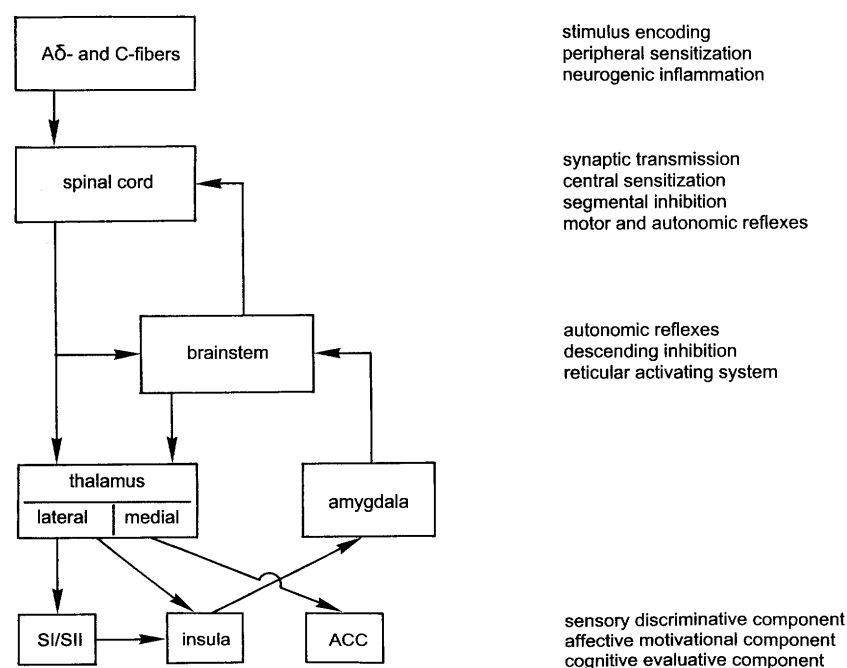


Figure 1

Major components of the nociceptive system. Free endings of A δ - and C-fibers in skin and other tissues are activated by actual or impending tissue damage (noxious stimuli). At the first synaptic relay station in the spinal cord, their signals interact and are strongly modulated by segmental and descending connections. After another synaptic relay station in lateral thalamic nuclei, the signals reach the primary and secondary somatosensory cortex (SI, SII) as a step towards the conscious perception of a noxious stimulus as being painful. Other parts of the cerebral cortex such as the insula and the anterior cingulate cortex (ACC) are reached both in series and in parallel to SI and SII. The distributed processing of nociceptive information is thought to lead to the sensory-discriminative, affective-motivational and cognitive evaluative components of conscious pain perception. Motor and autonomic reactions to noxious stimuli are processed in lower parts of the central nervous system and are mostly independent of conscious pain perception; some of them may still occur after brain death

nociceptive system, leading to ongoing pain. Such pain may, for example, occur in inflammatory conditions and is rationally treated by cooling. The neural basis of primary hyperalgesia to heat stimuli is peripheral sensitization of primary nociceptive afferents by shifts in the temperature sensitivity of vanilloid receptors and related modifications of other ion channels (Cesare et al. 1999).

In secondary hyperalgesia, the mechanical pain threshold may drop in such a way that even gentle stimuli such as movement of cotton wool across the skin is perceived as painful (allodynia). Normal weight bearing on the sole of the feet or contact with clothing then start to act as an adequate stimulus for the nociceptive system, leading to functional deficits by avoidance behavior. Such pain may also occur chronically following lesions of the nervous system (neuropathic pain), and this pain is particularly resistant to current treatment modalities. The neural basis of secondary hyperalgesia to mechanical stimuli is central sensitization of nociceptive neurons in the spinal cord by modifications of the synaptic transmission from A β -fiber mechanoreceptors that normally signal touch and from A-fiber nociceptors (Ziegler et al. 1999).

Central sensitization of the nociceptive system involves the activation of NMDA (N-methyl-D-aspartate) glutamate receptors in the postsynaptic neuron and several additional mechanisms. The intracellular signal pathways are similar to those found for long-term potentiation (LTP) of synaptic efficacy in the hippocampus, which is thought to be related to learning and memory (see *Neural Plasticity and Working Memory, Neural Basis of*). Both LTP and the opposing mechanism long-term depression (LTD) have been observed in slice preparations of the spinal cord using stimulation protocols similar to those that were found to be effective in slice preparations from the hippocampus or neocortex. In the intact organism, the triggering of LTP in the spinal cord is antagonized by the endogenous descending inhibition (Sandkühler 2000). The implications of these mechanisms for the causes of chronic pain are important but not yet fully understood. On the one hand, the intracellular signal pathways of LTP influence gene expression and may thus alter the function of nociceptive neurons for long periods of time. On the other hand, deficiencies in the balance between excitatory and inhibitory influences on the spinal cord may be a decisive factor in pain chronification.

5. Epilogue: Pain and the Brain

If reading the preceding sections have left the impression that the neural basis of actually feeling pain has not been addressed, this impression is correct. It is evident that pain exists only as long as it is being felt by a person. Thus, trying to understand the pain experience as a whole leads into the general mind–

body problem, which is a complex philosophical issue (for a more general account of how personal-level and subpersonal-level descriptions of pain can be combined, see Bieri 1995). This article has intentionally left out the intimate relationship of the subjective pain experience and consciousness. Instead, the preceding sections have described the neural pathways that signal tissue damage and how some of their known properties explain otherwise puzzling phenomena such as the allodynia of neuropathic pain states, where lightly touching the skin may evoke unbearable pain. But how the brain ultimately synthesizes the conscious perception of pain remains a mystery. A few simple examples, however, show that the brain does synthesize this perception:

(a) Blocking the neural pathways between the damaged tissue and the brain (e.g., by a local anesthetic in the dentist's office) creates a situation where there is tissue damage but no pain.

(b) Activation of nociceptive pathways by electrical stimulation within the thalamus may elicit a vivid perception of the pain of a heart attack. In this situation there is pain but no tissue damage.

In other words, a foot cannot hurt by itself, but a brain can perceive a hurting foot. The latter is possible even in the absence of the foot, as observed in phantom limb pain. Pain due to proximal activation of the nociceptive system is called projected pain, because here the normal mechanisms of pain perception are particularly evident: pain sensation is synthesized in the brain, but is then projected into the peripheral receptive field of the active brain areas. Pain due to peripheral activation of nociceptive afferents is called nociceptive pain, but again the pain sensation is synthesized in the brain and projected into the peripheral receptive field, which happens to be the site of tissue damage. This model of pain projection into receptive fields is simplified and does not include the central representation of the body image in brain areas such as the posterior parietal cortex, but this much should be apparent: pain is always in the brain.

See also: Chronic Pain: Models and Treatment Approaches; Pain, Health Psychology of; Pain, Management of; Somatosensation

Bibliography

- Albe-Fessard D, Berkley K J, Kruger L, Ralston H J, Willis W D 1985 Diencephalic mechanisms of pain sensation. *Brain Research Reviews* **9**: 217–96
- Augustine J R 1996 Circuitry and functional aspects of the insular lobe in primates including humans. *Brain Research Reviews* **22**: 229–44
- Bieri P 1995 Pain: A case study for the mind–body problem. In: Bromm B, Desmedt J E (eds.) *Pain and the Brain: From Nociception to Cognition*. Raven Press, New York, pp. 99–110

- Casey K L, Minoshima S 1997 Can pain be imaged? In: Jensen T S, Turner J A, Wiesenfeld-Hallin Z (eds.) *Proceedings of the 8th World Congress on Pain, Progress in Pain Research and Management*. IASP Press, Seattle, WA, pp. 855–66
- Caterina M J, Julius D 1999 Sense and specificity: A molecular identity for nociceptors. *Current Opinion in Neurobiology* **9**: 525–30
- Cesare P, Moriondo A, Vellani V, McNaughton P A 1999 Ion channels gated by heat. *Proceedings of the National Academy of Sciences of the United States of America* **96**: 7658–63
- Craig A D, Dostrovsky J O 1998 Processing of nociceptive information at supraspinal levels. In: Yaksh T L (ed.) *Anesthesia: Biologic Foundations*. Lippincott-Raven, Philadelphia, PA, pp. 625–42
- Craig A D, Reiman E M, Evans A, Bushnell M C 1996 Functional imaging of an illusion of pain. *Nature* **384**: 258–60
- Devinsky O, Morrell M J, Vogt B A 1995 Contributions of anterior cingulate cortex to behaviour. *Brain* **118**: 279–306
- Holzer P 1992 Peptidergic sensory neurons in the control of vascular functions: Mechanisms and significance in the cutaneous and splanchnic vascular beds. *Reviews in Physiology, Biochemistry and Pharmacology* **121**: 49–146
- Kenshalo D R, Willis W D 1991 The role of the cerebral cortex in pain sensation. In: Peters A, Jones E G (eds.) *Cerebral Cortex*. Plenum Press, New York, Vol. 9, pp. 153–212
- Lenz F A, Gracely R H, Romanoski A J, Hope E J, Rowland L H, Dougherty P M 1995 Stimulation in the human somatosensory thalamus can reproduce both the affective and sensory dimensions of previously experienced pain. *Nature Medicine* **1**: 910–13
- Melzack R, Casey K L 1968 Sensory, motivational, and central control determinants of pain. A new conceptual model. In: Kenshalo D R (ed.) *The Skin Senses*. Charles C. Thomas, Springfield, MA, pp. 423–43
- Merskey H, Albe-Fessard D, Bonica J-J, Carmon A, Dubner R, Kerr F-W-L, Lindblom U, Mumford J-M, Nathan P-W, Noordenbos W, Pagni C-A, Renner M-J, Sternbach R-A, Sunderland S 1979 Pain terms: A list with definitions and notes on usage. Recommended by the IASP subcommittee on taxonomy. *Pain* **6**: 249–52
- Metzinger T 2000 The subjectivity of subjective experience: A representationalist analysis of the first-person perspective. In: Metzinger T (ed.) *Neural Correlates of Consciousness: Empirical and Conceptual Questions*. MIT Press, Cambridge, MA, pp. 285–306
- Rainville P, Feine J S, Bushnell M C, Duncan G H 1992 A psychophysical comparison of sensory and affective responses to four modalities of experimental pain. *Somatosensory Motor Research* **9**: 265–77
- Raja S N, Meyer R A, Ringkamp M, Campbell J N 1999 Peripheral neural mechanisms of nociception. In: Wall P D, Melzack R (eds.) *Textbook of Pain*, 4th edn. Churchill Livingstone, Edinburgh, UK, pp. 11–57
- Sandkühler J 2000 Learning and memory in pain pathways. *Pain* **88**: 113–18
- Schlereth T, Magerl W, Treede R-D 2001 Spatial discrimination thresholds for pain and touch in human hairy skin. *Pain* **92**: 187–94
- Treede R-D, Kenshalo D R, Gracely R H, Jones A K P 1999 The cortical representation of pain. *Pain* **79**: 105–11
- Treede R-D, Meyer R A, Raja S N, Campbell J N 1992 Peripheral and central mechanisms of cutaneous hyperalgesia. *Progress in Neurobiology* **38**: 397–421

- Willis W D 1985 *The Pain System: The Neural Basis of Nociceptive Transmission in the Nervous System*. Karger, Basle, Switzerland
- Woolf C J, Fitzgerald M 1986 Somatotopic organization of cutaneous afferent terminals and dorsal horn neuronal receptive fields in the superficial and deep laminae of the rat lumbar spinal cord. *Journal of Comparative Neurology* **251**: 517–31
- Ziegler E A, Magerl W, Meyer R A, Treede R-D 1999 Secondary hyperalgesia to punctate mechanical stimuli: Central sensitization to A-fibre nociceptor input. *Brain* **122**: 2245–57

R.-D. Treede

Paleodemography: Demographic Aspects

1. Introduction

How long humans lived in the past is a question that has challenged researchers for centuries. Once the dominion of philosophers, theologians, and then historians, physical anthropologists have also struggled to answer this question over the last century. For the physical anthropologist, understanding long-term patterns of the human life span aids in placing observed changes within an evolutionary framework, linking both biological and cultural pressures and adaptations to a changing pattern of life expectancy in populations.

Exploration of this question by physical anthropologists has been primarily in the analysis of skeletal remains recovered archaeologically, and has fallen under the purview of a field developed as a subspecialty of physical anthropology—*paleodemography*.

The early days of paleodemography represented an exploration of modern demographic theory applied to ancient populations and the use of the abridged life table as a tool to aid interpretations of age-at-death profiles from cemetery samples. The 1980s marked a pivotal point for paleodemography. While there had been the occasional critique prior to the 1980s, it was not until 1982 that the great debate over the merits of paleodemography began, sparking several years of controversy within the literature (e.g., Bocquet-Appel and Masset 1982, Buikstra and Konigsberg 1985, Wittwer-Backofen 1987). Subsequent to this period, the two critical issues explored most in paleodemography have been (a) the accuracy of aging techniques, and (b) representativeness of samples.

2. Sources of Data

Paleodemographic studies have focused primarily on the reconstruction of human skeletal samples recovered archaeologically. The basic assumption has been that mortality statistics derived from the skeletal sample are sufficient to make inferences about mor-

tality in the once living population. In conjunction with and sometimes ancillary to skeletal data, archaeological evidence for settlement size and distribution have also been used to estimate population growth or to address more broadly-based questions of population structure in the past. Anthropological demography, of contemporary or recent historic hunter-gatherer and foraging populations, can also provide us with models for prehistoric populations. However, the range of fertility and mortality patterns among populations is so wide and overlapping that ethnographic analogy from subsistence base and mobility is extremely problematic. More recently, evidence from genetic studies has begun to be used to make inferences about long-term evolutionary changes in demographic structure in modern human populations over the past 250,000 years or so.

2.1 Skeletal Remains

The accuracy and reliability of age estimation techniques in particular have been central problems in paleodemography, particularly with respect to underestimation of the ages of older adults. A further issue is the necessary separation of techniques used to estimate age from the skeletal remains of adults vs. children. The latter are based on a various criteria related to known rates of development, whereas the former rely on more variable patterns of rates of development and degeneration of the bony tissues. In addition, differential rates of change between the sexes have made most techniques sex-specific. The scope of this article prevents an in-depth review of such methods, and readers are referred to such classic monographs as Acsádi and Nemeskéri (1970) or Krogman and Iscan (1986) for more detailed descriptions.

2.1.1 Estimation of Age. Determination of age from the adult skeleton can be undertaken using a variety of methods, including both quantitative and qualitative measures of continuous and discrete data. Such methods are referred to as skeletal age-indicator techniques. Adult aging techniques examine both macroscopic and microscopic changes in the morphology and structure of bone. Many macroscopic techniques focus on the pattern of age-related degeneration of bone (e.g., pubic symphysis, auricular surface, sternal end of the fourth rib). Others focus on the remodeling of bone as a biomechanical response to microfractures from everyday wear and tear. The physical anthropologist evaluates multiple criteria for each skeleton and then combines all estimates from each age-indicator method into an overall estimated age range for the individual.

Age estimation from the skeletal remains of children is based on the development of dental and skeletal

tissues, including tooth formation and eruption. Estimation of age in the nonadult is much easier and more accurate than in the adult. While there are fewer techniques than are available for adult age estimation, each has a smaller range of error. Overall, estimation of age from a juvenile is accurate within a range of about plus or minus half a year.

2.1.2 Estimation of Sex. It is recognized widely in the anthropological literature that the pelvis or hipbone is the most reliable part of the skeleton for determination of sex. Both metric and nonmetric or morphological techniques have accuracy rates of better than 95 percent for correct sex, while accuracy rates based on other parts of the skeleton are usually lower. While morphological variables are often preferred because of better preservation in skeletal samples, metric variables are considered by many to be more reliable because of their greater precision.

To the regret of many researchers, however, determination of sex from the skeleton has most often been restricted to those who have survived past adolescence and who then manifest changes in the skeleton reflective of sex. While a variety of studies have investigated traits that might be sexually dimorphic in infants and juveniles, only a few have had sufficient levels of accuracy to warrant their application in osteological analyses. More promising, but still restricted by time and money, is the determination of sex by extracting ancient DNA from the bones or teeth of individuals.

2.2 Archaeological Evidence

While paleodemography has been focused primarily within the realms of skeletal biology, other forms of evidence have also been explored to answer demographic questions. In particular, archaeological demography has often shed light on issues of population structure in the distant past. Estimates of population size and growth have been attempted from settlement data by examining features such the size and area of the living site, number of dwellings, the density and distribution of artifacts and food remains, as well as from ethnohistoric estimates of population size. Even when data are available, estimates of population size must often be made through ethnographic analogy, whereby the relationship between population size and material remains observed in modern or historic groups is imposed on the archaeological site.

2.3 Genetic Evidence

Recent studies of DNA have proposed some interesting hypotheses regarding demographic changes in the past that are of interest when exploring the issue

of long-term trends in human prehistory. In particular, advances in nucleotide divergence theory have provided information on human demographic behavior since patterns of gene differences contain information about the demographic history of a species (Harpending et al. 1998). Genetic data have been used specifically to tackle questions of modern human origins, and the evidence implies human demographic expansion from a population of only several thousand about 250,000 years ago.

3. Representativeness

The representativeness of a skeletal series is a crucial factor in paleodemographic studies. While anthropologists do not expect samples to represent the population in absolute numbers, the fundamental assumption governing the analysis and interpretation of skeletal samples is that the *distribution and pattern of any parameter is the same in the skeletal sample as in the living population that contributed to it*. That is, the pattern observed in the skeletal sample can be used as a proxy for the real pattern within the population. However, a variety of factors serve to bias paleodemographic reconstructions relative to the true demography of the once living population. Biological factors related to the fact that skeletal samples are the nonsurvivor portion of a population, cultural factors related to the burial of the body, environmental factors related to postdepositional processes, and methodological factors related to the excavation and analysis of skeletal samples all serve to potentially bias interpretations.

Until the 1990s few attempts had been made to directly assess the validity of interpretations from paleodemographic reconstructions because of the obvious difficulty in obtaining a skeletal sample that is known to be representative of its larger population. With the more recent, detailed studies of historic cemetery skeletal samples, researchers have begun to test the representativeness of their samples by comparing the mortality data derived from the skeletal sample with the documentary mortality data associated with the cemetery from which the sample was drawn. In most cases, clear differences in demographic parameters estimated from the two sources became apparent. As a result, recognizing and controlling for these biases has become an important part of paleodemographic reconstructions.

4. Reconstructing the Demography of Prehistoric Populations

Traditionally, the statistical tool used by paleodemographers has been the abridged life table. Under the assumption of a stationary population, paleodemography uses mean age at death to estimate

expectation of life at birth. Generally, mean age at death is considered approximately equivalent to the inverse of the birth rate in a population, but is independent of both life expectancy and the death rate. However, when the conditions of a stationary population are not met, this calculation simply represents mean age at death.

Traditionally, the abridged life table has been used by paleodemographers to estimate general mortality patterns with fertility being a by-product. However, several estimators of fertility from mortality profiles have been used. Jackes (1992) has noted that comparisons of the age structures across populations are based on the assumption that there is a relationship between juvenile and adult mortality, and that age-at-death data within very broad age categories will carry some information about the fertility rate of the population. However,

in nonstationary populations, age-at-death distributions are extremely sensitive to changes in fertility but not to changes in mortality Thus, if a population is not stationary—and changing populations never are—small variations in fertility have large effects on its age-at-death distribution, while even quite large modifications of mortality have virtually none (Wood et al. 1992, p. 344).

As a result, many researchers have concluded that the age distribution of skeletal samples provides less information about mortality than it does about fertility, a position supported very early on in the demography literature. In fact, the same fertility and mortality schedules can produce different birth and death rates in populations with different age structures.

4.1 Model Life Tables

In the 1970s demographers expressed concern over the paucity of evidence from which to make statements regarding paleodemographic parameters, forcing investigators to extrapolate from models derived from other sources. Subsequently, model life tables from modern demographic studies formed the basis from which anthropological demographers began to develop model life tables for past populations. Skeletal samples can be compared with a model life table and the fit between model and observed mortality distributions can then be assessed statistically.

The use of model life tables in anthropological demography is twofold. First, it provides a means of assessing or compensating for biased and incomplete data, and second, it allows for the estimation of fertility rates and construction of an initial population at risk. In the early 1970s, Weiss (1973) developed a set of model fertility and mortality schedules derived from ethnographic samples of contemporary hunter-gather societies and prehistoric skeletal samples. Many investigators agree that demographic statistics derived

from contemporary non-Western societies represent an effective means of assessing skeletal age profiles of past populations. However, given the variety of conditions under which many contemporary populations live, it is difficult to be certain that ethnographic analogy will always be appropriate. Further, the application of ethnographic estimators to samples for which related sociocultural information is sparse further compounds the problem.

4.2 Hazard Analysis

Although a potentially powerful tool for paleodemographic analyses, model life table fitting techniques are still subject to potential biases resulting from the use of inappropriate model populations. As an alternative, Gage (1988) proposed the use of a hazard model of age-at-death patterns that can be fitted to survivorship, death rate, and age structure data. As Gage has noted, the Siler model of mammalian mortality is useful for studying human mortality because it reduces the number of variables while maintaining interpretive power. The model for survivorship, which can be fitted to the l_x column in a life table using nonlinear techniques to estimate the parameters, is:

$$l_t = \exp \left\{ -\frac{a_1}{b_1}[1 - \exp(-b_1 t)] \right\} \exp(-a_2 t) \times \exp \left\{ \frac{a_3}{b_3}[1 - \exp(b_3 t)] \right\} \quad (1)$$

where t is age, a_1 is the risk of immature mortality at the moment of birth, b_1 is the rate of decline with age of this risk, a_2 is a constant risk factor, a_3 is the risk of senescent mortality at the moment of birth, and b_3 is the rate of increase in senescent mortality with age (the rate of aging).

The instantaneous death rate or force of mortality at age t of this model is

$$h_t = a_1 e^{-b_1 t} + a_2 + a_3 e^{b_3 t} \quad (2)$$

and the proportion of individuals surviving the immature component is $\exp\{-a_1/b_1\}$ (Gage 1988). This technique provides a method of estimating age-specific mortality and fertility directly from anthropological data, and it can smooth demographic data from a variety of populations without imposing a pre-determined age structure (Gage 1988).

Ultimately, the focus of paleodemography has been to refine methods to improve estimates of age at the individual level in order to get some aggregate estimate of the age structure in the population. More recently, however, researchers have begun to try and estimate

the mortality distribution of a sample directly from the distribution of skeletal age-indicator stages scored. While the difference is subtle, it is important in that such techniques attempt to avoid the broad range of error associated with estimates at the individual level. If one is interested in estimating an individual's age, as in the context of forensic anthropology, then such techniques fall into the category of prediction models. However, if one is interested in shape of mortality or the force of mortality within a population then an alternative theoretical approach is required—one that begins with the overall age structure being estimated, and estimates of individual age being derived secondarily. This apparent paradox in paleodemography was noted by Konigsberg and Frankenberg (1994, p. 96):

Thus, for any single skeleton from an archaeological sample, the posterior probability that the individual is a particular age conditional on the observed indicator is proportional to the prior probability that the individual is that particular age times the probability that an individual that particular age would be in the observed indicator state. However, we do not know this prior probability. If we knew the prior probability of death at particular ages, then we would know the age-at-death distribution, which is precisely what we are trying to estimate. This paradox can be solved by using a common statistical method known as maximum likelihood estimation.

While the precise statistical approach to estimating the mortality structure from skeletal age-indicator data may vary, it is clear that the emerging analytical framework on which paleodemography must move forward begins first with an estimate of the age structure, from which estimates of individual age can subsequently be calculated (cf. Hoppa and Vaupel 2001).

5. Summary

The biological basis for human survival and longevity evolved during the tens of thousands of years of human experience under premodern conditions. A deeper understanding of this basis thus depends on knowledge of mortality and fertility patterns over the long stretch of human existence. Furthermore, deeper understanding of the plasticity of survival and longevity depends on knowledge of how different mortality is today from mortality before the modern era. Paleodemography attempts to address this question by integrating evidence from a variety of sources, but focusing primarily on skeletal, archaeological, and even genetic data. Long-standing interpretations have suggested that the bulk of premodern populations had very short life spans with only a few individuals surviving to middle adulthood. However, new methodological approaches are beginning to readdress this position, and future paleodemographic studies will be able to test this hypothesis more reliably.

See also: Anthropology; Demography, History of; Evolutionary Approaches in Archaeology; Human Evolutionary Genetics; Primates, Evolution of

Bibliography

- Acsádi G, Nemeskéri J 1970 *History of Human Lifespan and Mortality*. Akadémiai Kiadó, Budapest
- Bocquet-Appel J-P, Masset C 1982 Farewell to paleodemography. *Journal of Human Evolution* 11: 321–33
- Buikstra J E, Konigsberg L W 1985 Paleodemography: Critiques and controversies. *American Anthropologist* 87(2): 316–34
- Gage T B 1988 Mathematical hazards models of mortality: An alternative to model life tables. *American Journal of Physical Anthropology* 86: 429–41
- Harpending H C, Batzer M A, Gurven M, Jorde L B, Rogers A R, Sherry S T 1998 Genetic traces of ancient demography. *Proceedings of the National Academy of Sciences* 95(4): 1961–7
- Hassan F A 1981 *Demographic Archaeology*. Academic Press, New York
- Hoppa R D, Vaupel J W (eds.) 2001 *Paleodemography: Age Distributions from Skeletal Samples*. Cambridge Studies in Biological and Evolutionary Anthropology. Cambridge University Press, Cambridge, UK
- Jackes M 1992 Palaeodemography: Problems and techniques. In: Saunders S R, Katzenberg M A (eds.) *Skeletal Biology of Past Peoples: Research Methods*. Wiley-Liss, New York, pp. 189–224
- Konigsberg L W, Frankenberg S R 1992 Estimation of age structure in anthropological demography. *American Journal of Physical Anthropology* 89: 235–56
- Konigsberg L W, Frankenberg S R 1994 Palaeodemography: 'Not quite dead.' *Evolutionary Anthropology* 3(3): 92–105
- Krogman W M, Iscan M Y 1986 *The Human Skeleton in Forensic Medicine*, 2nd edn. Charles C. Thomas, Springfield, IL
- Paine R R 1997 *Integrating Archaeological Demography. Multi-disciplinary Approaches to Prehistoric Population*. Center for Archeological Investigations, Occasional Paper No. 24, Southern Illinois University at Carbondale, IL
- Weiss K 1973 Demographic models for anthropology. *American Antiquity* 38(2) Part 2: Memoir 27
- Wittwer-Backofen U 1987 Überblick über den aktuellen Stand paläodemographischer Forschung. *Homo* 38: 151–60
- Wood J W, Milner G R, Harpending H C, Weiss K M 1992 The osteological paradox: Problems of inferring prehistoric health from skeletal samples. *Current Anthropology* 33(4): 343–70

R. D. Hoppa

Panel Surveys: Uses and Applications

A panel survey makes similar measurements on the same sample units (e.g., individuals, firms) at different points in time (Duncan and Kalton 1987). Panel surveys can be distinguished from: (a) cross-sectional surveys, which take measurements on sample units from a population of interest at a single point in time; and (b) repeated surveys, such as the US General

Social Survey (Davis and Smith 1994), in which similar measurements are made on samples from an equivalent population (e.g., defined by identical geographical boundaries or birth cohorts) at different points of time, but without attempting to ensure that any elements are included in more than one round of data collection.

A widely used alternative term for panel survey is longitudinal survey. However, it is useful to use 'longitudinal' only in the context of longitudinal data; such data may be collected by a panel survey, but they could also come from retrospective reports in a cross-sectional survey.

Most panel surveys follow their sample elements for the duration of the survey. Other panel designs include a rotating panel survey, in which sample elements have a restricted panel life; as they leave the panel, new elements are added. Examples of this are the monthly US Current Population Survey (US Bureau of Labor Statistics 2000a) and the Canadian Labour Force Survey (Statistics Canada 2000), in which the sample at each month is composed of a number of rotation groups, each of which is a self-contained sample from the population of interest. One rotation group drops out and one rotation group enters each month. A split panel survey is a combination of a panel and a repeated survey (Kish 1986).

Panel surveys can differ markedly in the interval between rounds of data collection and the overall length of the survey. On the one hand, a consumer panel in market research may involve weekly contacts with respondents to collect their diary records of purchases. On the other hand, there are long-term panel surveys that may collect data only annually or even less frequently, and these surveys may continue for many years. Examples are the Panel Study of Income Dynamics (PSID), which had been collecting data on US households annually (biannually since 1997) since 1968 (Hill 1992), and the British National Child Development Study, which started in 1958 and has collected data on the sample persons on five occasions in the 33 years to 1991 (Institute of Education 2000).

Some panel surveys are concerned with population subgroups that have experienced the same event during the same time period, such as attending eighth grade in a given year, having been born in a particular week, or having been married during the same period. Since these sample universes are defined as cohorts, such studies are often called cohort studies. Examples include the various cohorts defined by age and sex in the US National Longitudinal Surveys (US Bureau of Labor Statistics 2000b). A limitation of cohort studies is that their results relate formally only to the particular cohorts from which the sample was selected. An extension of the approach is the multiple cohort design, in which several cohorts are followed in similar panel surveys. Thus, the National Longitudinal Surveys have begun three youth cohort panel surveys, in

1968 (males) and 1968 (females), 1979, and 1997, while the Monitoring the Future studies have selected samples of US high school seniors each year, and follow each sample in a panel survey (Monitoring the Future 2000).

Despite their lack of formal population representation, still other panel surveys are valuable because of their detailed interviews and often very long durations. Examples include Elder's (1999) 30-year follow-up of individuals who had attended Oakland, California elementary schools and were enrolled in the Oakland Growth Study when in fifth and sixth grades and Furstenberg's (Furstenberg et al. 1987) 30-year panel study of Baltimore teen mothers who had enrolled in a teen parenting program in 1965.

1. Uses of Panel Surveys

To appreciate the analytic advantages of panel surveys, it is important to realize that the time dimension complicates analyses because: (a) characteristics of population elements may change over time, for instance, an individual may be single one month and married the next; and (b) the population itself may change in composition, with both sample units entering into and exiting from the population.

These complications, especially the first, give rise to a variety of objectives for analysis (Kish 1986, Duncan and Kalton 1987), including:

(a) Estimates of population parameters (e.g., monthly rates of unemployment, public opinion regarding gun control) at distinct time points. The resulting time series of such estimates may be the input for further analysis; see, for example, Smith (1994).

(b) Measurement of net change, defined as change at the aggregate level. An example is the change in the UK unemployment rate between two consecutive months or change in US public opinion regarding gun control before and after a gun-related event.

(c) Measurement of various components of individual change, including: (i) gross change—change at the individual level between two time points (e.g., relating the poverty status of sample members in a given year to their poverty status ten years later); (ii) average change, or trend, for each individual over a period of time, with the average serving to smooth out the effects of isolated perturbations and measurement errors; or (iii) instability for each individual (e.g., the stability of locus of control over 25 consecutive weeks from a sample of elderly in Eizenman et al. 1997).

(d) Aggregation of data for individuals over time. An example is the collection of incomes on an annual basis to relate childhood-long economic well-being to children's developmental outcomes (Blau 1999).

(e) Measuring the frequency, timing, and duration of events occurring within a given time period. Examples are the proportion of persons who were ill in

the past two weeks and the average duration of their illnesses, and the complete fertility histories of post-menopausal women.

If the composition of a population is sufficiently stable to be treated as static for the time period of interest, it may be possible to collect the data required to meet all of these objectives from a single cross-sectional survey conducted at the end of the period. Crucial here is that errors of measurement are acceptably small. Unfortunately, respondents' memories are notoriously inadequate when a substantial time is involved or when the characteristics or events have a low degree of salience for the respondents (Bound et al. 2000). For example, respondents are often unable to recall important events such as unemployment (Mathiowetz and Duncan 1984) or hospitalizations (Cannell et al. 1981) within periods as short as several months, and they cannot reliably recall subjective experiences (Moss and Goldstein 1979). Thus, event data often need to be collected at several points throughout the period.

Repeated surveys can meet some of these analytic goals. Their particular strength is that at each round of data collection they routinely select a sample of the population existing at that time. They thus provide a series of cross-sectional estimates that can be used for objective (a) above. They can also be used to measure overall net change (b). The major limitation of a repeated survey is that it does not yield data to satisfy objectives (c) and (d): since elements are not explicitly included in several rounds, none of the components of individual change can be measured and individual data cannot be aggregated across rounds. Data on the frequency and timing of events in a specified period, as in objective (e), can be collected in a repeated survey, but with the measurement error problems noted above.

The major advantage of a panel survey over a repeated survey is its much greater analytical potential. It enables components of individual change to be measured, objective (c), and also the summation of a variable across time, objective (d). By providing high-quality measurements of the frequency, timing, and duration of events, panel surveys provide data useful for the estimating of event-history models (Tuma and Hannan 1984, Allison 1982, Yamaguchi 1991).

A panel survey can be much more efficient than a repeated survey for measuring net change, objective (b). Let the net change be $y_2 - y_1$, where y_1 and y_2 are the means of the variable of interest at times 1 and 2. Then the variance of the net change is $V(y_2 - y_1) = V(y_1) + V(y_2) - 2\rho[V(y_1)V(y_2)]^{1/2}$ where ρ is the correlation between y_1 and y_2 . In a repeated survey with two independent samples, $\rho = 0$, but in a panel survey ρ is the correlation between an individual's y -values on the two occasions, which is often quite high. In this case, a panel survey will yield a much more precise estimate of net change, objective (b), than will a repeated survey of the same size. On the other hand, as

discussed below, panel nonresponse between times 1 and 2 may result in more bias in the estimation of y_2 in a panel than repeated survey.

A panel survey also permits the collection of a wider range of data on each sampled element than is possible with a repeated survey. The amount of data collected from any one respondent on one occasion is limited by the burden imposed; however, different data can be collected on different rounds of a panel survey, thus providing the possibility of accumulating an extremely wide range of variables for each sampled element. In the US Survey of Income and Program Participation (SIPP), for instance, core data on income and program participation are collected on each round, and additional data on different topics (e.g., health care utilization and financing, pension and retirement issues, housing conditions) are collected on specific rounds only (US Bureau of the Census 2000).

Some long-term panels have important intergenerational components, with data on children and their family contexts taken in early waves and measurements of the adult attainments of the children taken in later waves. Examples include population panels such as the National Longitudinal Survey youth panels, the PSID, the British Household Panel Survey (Institute for Economic and Social Research 2000), the German Socio-Economic Panel, and more specialized data collections such as the Oakland Growth Study (Elder 1999).

2. Problems with Panel Surveys

By attempting repeated interviews with the same sample, panel surveys have problems not found in single or repeated cross-sectional designs, the most important of which is panel nonresponse (initial-wave respondents may not respond in later waves). An additional potential problem with panel surveys is panel conditioning, where responses in a given interviewing round may be conditioned by participation in prior rounds of interviews.

Methods developed to cope with nonresponse bias include minimizing nonresponse in panel surveys and developing statistical adjustments for existing nonresponse. Existing panel surveys typically devote sizeable resources to maintaining high response rates, and sometimes are quite successful. For example, the National Longitudinal Survey of Youth conducted interviews in 1991 with 89 percent of the respondents in its initial, 1979, interview (MaCurdy et al. 1998). Losses in the British National Survey of Health and Development amounted to only 12 percent after 26 years (Atkins et al. 1981).

Incentive payments, respondent reports, persuasion letters, using administrative data for tracing, and collecting extensive contact information (e.g., on friends and relatives not living in the household who would know of address and telephone number

changes) help minimize these problems (Freedman et al. 1980, Clarridge et al. 1978, Call et al. 1982).

As with any survey, sizeable nonresponse in a panel survey gives rise to concerns about nonresponse bias. The situation with the first wave of a panel survey corresponds to that with a cross-sectional survey in that very limited information is available on the nonrespondents. The situation with later wave nonresponse in a panel survey is, however, different: in this case a good deal of information is available about later wave nonrespondents from their responses on earlier waves. The earlier wave information can be used to investigate the possibility of nonresponse bias and to develop imputation and weighting nonresponse adjustments that attempt to reduce the bias (Kalton 1986, Lepkowski 1989).

With regard to conditioning, there is ample evidence from several surveys that initial responses in a panel survey differ substantially from those given in subsequent waves (Bailar 1975, 1979, Ghangurde 1982). In the case of the US Current Population Survey, estimates of unemployment from households entering the sample for the first time are almost 10 percent larger than the average over all eight monthly reporting periods. It is not clear whether there is more response bias in the initial or subsequent waves, because the repeated contact with respondents has ambiguous effects on the quality of the data. The crucial question, as yet unanswered for most phenomena reported in surveys, is whether it is merely the reporting of behavior or the behavior itself that is affected by panel membership.

It may be that data collected in subsequent panel waves is less biased, because repeated contact increases the probability that respondents understand the purposes of the study and are thus increasingly motivated to make the effort necessary to give more accurate answers. On the other hand, there is evidence from a validation study (Traugott and Katosh 1979) that extended participation in a panel study on election behavior not only increased the accuracy of responses on voting behavior but may indeed have increased the amount of voting, so that the behavior of the panel was no longer representative of the behavior of the population at large.

It seems unlikely that panel participation has pervasive behavioral effects, especially when changes in the behavior under investigation require more effort than making a trip to the polls. For example, economic behavior such as work effort, saving, commuting, and home ownership are all unlikely to be affected by responses to occasional interviews. Responses to attitudinal questions may be affected by panel membership if participation stimulates interest in the subject matter of the survey.

The limited membership in a rotating panel acts to reduce the problems of panel conditioning and panel loss in comparison with a nonrotating panel survey, and the continual introduction of new samples helps

to maintain an up-to-date sample of a changing population. Rotating panels are used primarily for the estimation of cross-sectional parameters, objective (a), for the estimation of average values of population parameters across a period of time, objective (b), and for measuring net changes, objective (c). A rotating panel survey will generally provide more precise estimates of point of time and, especially, of change parameters than a repeated survey of the same size. Moreover, a rotating panel survey will sometimes have a cost advantage over a repeated survey. This will occur when it is cheaper to conduct a reinterview than an initial interview, as for instance is the case in the US Current Population Survey where initial interviews must be conducted by personal visit whereas reinterviews on some waves may be conducted by telephone (US Bureau of the Census 1978).

The ability of a rotating panel survey to measure components of individual change, objective (c), and to aggregate data for individuals across time, objective (d), is clearly restricted. Since rotating panels are not intended to serve these objectives, they can be designed to avoid the heavy expense of following movers that occurs with nonrotating panel surveys. Thus, for instance, the Current Population Survey employs dwellings, not households or persons, as the sampled units, so that there is no need to follow households or persons moving between panel waves.

In a split panel survey, the panel survey component can be used to measure components of individual change, objective (c), and to aggregate data for individuals over time, objective (d). Its permanent overlap aids in the estimation of net change, objective (b), between *any* two waves whereas the overlap in a rotating panel survey aids only in the estimation of net change between certain prespecified waves.

Both rotating and split panel survey designs provide samples of new entrants to the population and the capacity to use their panel survey components to check on biases from panel conditioning and respondent losses.

3. The Problem of Changing Populations

The composition of almost every population of interest, whether of individuals, families, firms, or housing units, changes over time. Individuals enter the population when they are born, immigrate, or attain the age or other status that is used to define the population of interest, and depart when they die, emigrate, move into institutions such as jails, nursing homes, or the military, or in some other way lose that status. Families are 'born' when children leave their parents and set up their own independent households or when a divorce or separation breaks up one family into two; they 'die' when all members of the original household die or when two households are merged into a single one through marriage or less formal living

arrangements. Over a survey's time span an element may both enter and leave the population, and may do so more than once, as for instance in the case of a person who has several stays in a nursing home during the course of a survey of the noninstitutionalized population. Comparable changes through construction and demolition, or incorporation, merger, and bankruptcy apply to populations of dwellings and firms.

All surveys over time must address the issues raised by these changes in population composition (Duncan and Kalton 1987). In a repeated survey the cross-sectional parameters of objective (a) relate to the population as it exists at the time the sample is drawn for the particular round of the survey. This procedure reflects both the changing values of the variables under study and the changing population over the time period involved. The estimates are thus average values for a form of average population over time.

Panel studies have adopted a variety of strategies for coping with changing populations. The simplest are panel studies of birth cohorts of individuals in the population, where as many members as possible of the original sample are followed throughout the panel period. In effect, these panels ignore all problems of population composition change except death. Disregarding nonresponse and immigration, these studies represent their respective cohorts as they age, while they gradually cease to represent individuals in the original age range. The National Longitudinal Survey cohorts of Older Men and Mature Women are examples. As discussed above, multiple-cohort panel designs avoid this problem by including in subsequent interviewing waves a sample of new entrants into the population.

Panel designs such as that of the PSID, the British Household Panel Survey, and the European Community Household Panel surveys, contain a mechanism for adding to their samples individuals and families that are 'born' into the population, so they have the potential for maintaining representative samples of nonimmigrant individuals and families throughout their duration. For example, the PSID began with a probability sample of dwellings that, in turn, provides a representative sample of subunits within those dwellings: households, families, subfamilies, transfer program reciprocity units, and individuals. The selection probability of each of these units is identical to the selection probability of the dwelling itself. Other panels composed these subsamples with known selection probabilities through other means. With a properly specified set of rules regarding the definition of units, newly formed subunits of interest (including individuals) enter into the sample with known selection probabilities and reflect corresponding changes that are taking place in the population at large, for example, Hill (1992).

Because they originate from existing panel households, newly formed families and individuals are more

‘clustered’ by family lineage in these designs than in repeated cross-sectional samples; however, the statistical inefficiency of this clustering is heavily outweighed by the analytic advantages of being able to link these newly formed units with information on their families of origin. Thus, information on children setting up independent households can be matched to reliably reported information on their parents, information on ex-spouses can be linked together if the marital split occurred during the panel period, etc. Immigration during the panel period is usually ignored although can be a serious problem for a long-term panel survey confined to a city or a local area, but it is a less serious problem for a national panel survey.

4. Examples of What Panel Surveys Have Found

To help convey the analytic advantages of panel surveys, we list in this section some prominent examples of panel-survey based analyses.

4.1 Poverty and Welfare Dynamics

Trends in annual US poverty rates (i.e., the fraction of the population with family cash incomes below a set of ‘poverty thresholds’ that vary with family size) are tracked through cross-sectional estimates based on the Current Population Survey. Net change from one year to the next in these rates rarely amount to as much as a single percentage point (US Bureau of the Census 1999), producing perceptions that ‘the poor’ are an ever-present and little-changing group.

By applying event history methods to the time series of annual poverty measurements for PSID sample members, Bane and Ellwood (1986) characterized the nature and determinants of poverty experiences by the length of their ‘spells’ (i.e., continuous periods of poverty). They find that a clear majority (60 percent) of families who first begin a poverty experience do so for short (i.e., spells of 1 or 2 years) periods of time. On the other hand, a subset of poor families (14 percent) have longer-run (8 or more years) spells, and the remaining one-quarter (26 percent) have medium-length (3–7 years) spells. Thus, there is no single ‘correct’ characterization of poverty—transitory or persistent—since poverty experiences are a mixture of transitory and long-term.

By combining individual spells into simulated lifetime histories, Bane and Ellwood (1994) provide estimates of the likely total number of years of receipt for families just starting to receive cash assistance from the US Aid to Families With Dependent Children program. They find a roughly even distribution of first-time welfare recipients across the three time intervals; roughly one-third (36 percent) have very short welfare experiences, a third (35 percent) medium-length experiences, and the final third (29 percent)

long-term receipt. With welfare, as with poverty, heterogeneity is a key feature, with many families using AFDC as short-term insurance, but a substantial fraction of recipients receiving long-term assistance.

4.2 Intergenerational Transmission of Economic Status

Correlations in earning between fathers and sons is sometimes treated as a useful measure of intergenerational (im)mobility. Solon (1992) argues that there are two important sources of bias in attempts to estimate such correlations: single year proxies for long-run economic status contain transitory fluctuations that, under plausible assumptions, impart a downward bias to the correlation estimates, and samples used in previous studies (e.g., Mormon brothers, white twins who served in the armed forces) tend to be peculiarly homogeneous, also resulting under plausible assumptions in a downward bias to estimates based on homogeneous than population-based samples.

Data from both the NLSY (in Zimmerman 1992) and PSID (in Solon 1992) have been used to gauge the likely biases. For example, Solon et al. (1992) find that increasing the measurement interval for economic status from one to several years causes their estimate of the father-son earning correlation to jump from 0.25, a number consistent with past studies based on single-year earnings measurement and homogeneous samples, to 0.45. Thus, intergenerational economic mobility appears to be substantially lower than had been depicted in studies based on homogeneous samples and single-year measures of income or earnings.

The more specialized panel studies of Elder and Furstenberg provide much more detail on the process of intergenerational mobility. Using 30-year follow-up data on the Oakland Growth Study cohorts, Elder (1999) found in these children of the Great Depression strong associations among economic hardship, parental psychological well-being, and adolescent well-being in intact families.

Fathers who experienced job loss and economic deprivation were more distressed psychologically and prone to explosive, rejecting, and punitive parenting. Preschool-age children in these families, especially boys, were more likely to exhibit problem behaviors, while adolescent girls were more likely to have lower feelings of self-adequacy and to be less goal-oriented. Elder (1999) speculated that the gender and age differences reflected different experiences in families during the deprived times. Adolescent boys reacted to family economic hardship by looking for economic opportunities outside of the home. This time spent taking on responsibility in extra-family contexts reduced the time they spent with their families and may have provided a buffer to the effects of negative family

interactions. Younger children and girls were unable to seek such extra-familial opportunities as readily and therefore did not have access to buffering opportunities.

Furstenberg et al. (1987) take a detailed look at intra- and intergenerational mobility in a panel study of a poor teen mothers participating in a program in a Baltimore hospital. Their 20-year follow-up revealed considerably diversity in the mothers' economic and demographic statuses, with most self-supporting but very few stably married. Most (60 percent) of the daughters of the teen mothers had not themselves become teen mothers and three-quarters enjoyed at least a modest degree of economic success. In contrast, fewer than half of the sons of the teen mothers had attained successful economic trajectories.

4.3 Unemployment and Labor Turnover

Both unemployment and job turnover are most naturally thought about in a 'spells' framework: how long does a spell of unemployment last before employment recommences? How long does a spell of employment with one firm last before the worker leaves to take (or look for) alternative employment? Given the inherently longitudinal nature of these processes, it is not surprising that panel surveys have played an important role in research on these topics. While much research on this topic has used administrative records, a panel survey has the advantages of including those who do not receive unemployment benefits, for comparison to recipients, and it continues observing spells after benefits have been exhausted (i.e., it measures total spells, not just insured spells), so one can observe changes in job-finding rates after benefits are exhausted.

An example is Katz and Meyer's (1990) study of the impact of unemployment insurance on unemployment duration. Katz and Meyer find that probability of recall declines as the duration of a spell increases, for both recipients and nonrecipients. New-job rates, however, are upward-sloping for recipients but relatively flat for nonrecipients. Moreover, there are spikes in both the recall and new-job hazard rates around 26 and 39 weeks (when benefits were typically exhausted in their sample period) for recipients but not for nonrecipients. Thus, not only is there evidence that the approaching end of benefits conditions recipients' job finding, but it also conditions firms' recall policies.

4.4 Antecedents of Life Events

There is great interest in understanding the consequences of important life events such as divorce, widowhood, unemployment, and retirement. Cross-sectional data provide comparisons of, for example, the living standards and labor supply of divorced

women and 'otherwise similar' married women, or the health of unemployed or widowed men with 'otherwise similar' employed or married men. But panel surveys provide valuable post- and, interestingly, pre-event data on the same individuals.

For example, Johnson and Skinner (1986) use data from the PSID to examine work hours of women between seven years before and four years after divorce or separation. Consistent with other research, they find that the average work hours of these women rose by half (from 1024 to 1551 hours) between one year before and four years after the divorce or separation. But Johnson and Skinner (1986) also found a big jump (from 744 to 1024 hours) over the six-year period *prior* to the divorce or separation.

Zick and Smith (1991) analyze the economic antecedents and consequences of widowhood by comparing family-income trajectories of a control group of intact couples with widowed women and men between five years before and five years after the spouse's death. They too find differences prior to the event, with income changes over the five years *prior* to the death accounting for about one-quarter of the control-widow difference at $t+5$ and for fully half of the control-widower difference. A clear message from both of these analyses is that impending life events such as divorce and death produce measurable economic and social changes long before the events actually occur. An implication is that fairly long panels are needed to capture the complete dynamic.

See also: Databases, Core: Demography and Registers; Economic Panel Data; Large-scale Social Survey, Analysis of; Longitudinal Research: Panel Retention; Sample Surveys: Methods; Sample Surveys: Survey Design Issues and Strategies; Sample Surveys: The Field; Survey Research: National Centers; Surveys and Polling: Ethical Aspects

Bibliography

- Allison P 1982 Discrete-time methods for the analysis of event histories. In: Leinhardt S (ed.) *Sociological Methodology* 1982. Jossey-Bass, San Francisco, pp. 61–98
- Atkins E, Cherry N, Douglas J W B, Kieman K E, Wadsworth M E J 1981 The 1946 British birth cohort: an account of the origins, progress, and results of the National Survey of Health and Development. In: Mednick S A, Baert A E (eds.) *Prospective Longitudinal Research*. Oxford University Press, Oxford, UK, pp. 5–30
- Bailar B A 1975 The effects of rotation group bias on estimates from panel surveys. *Journal of the American Statistical Association* 70: 23–30
- Bailar B A 1979 Rotation sample biases and their effects on estimates of change. *Bulletin of the Statistics Institution* 48(2): 385–407
- Bane M J, Ellwood D T 1986 Slipping in and out of poverty: the dynamics of spells. *Journal of Human Resources* 21: 1–23
- Bane M J, Ellwood D T 1994 *Welfare Realities*. Harvard University Press, Cambridge, MA

- Blau D 1999 The effect of income on child development. *Review of Economics and Statistics* **81**(2): 261–76
- Bound J, Brown C, Mathiowetz N 2000 Measurement error in survey data. In: Heckman J, Leamer E (eds.) *Handbook of Econometrics*. North Holland, Amsterdam, Vol. 5
- Call V R A, Otto L B, Spenner K I 1982 *Tracking Respondents: A Multi-Method Approach*. Lexington Books, Lexington, MA
- Cannell C F, Miller P V, Oksenberg L 1981 Research on interviewing techniques. In: Leinhardt S (ed.) *Sociological Methodology*, 1981. Jossey-Bass, San Francisco pp. 389–437
- Claridge B R, Sheehy L L, Hauser T S 1978 Tracing members of a panel: A 17-year follow-up. In: Schuessler K F (ed.) *Sociological Methodology*, 1978. Jossey-Bass, San Francisco, pp. 185–203
- Davis J, Smith T W 1994 *General Social Surveys, 1972–1994: Cumulative Codebook*. National Opinion Research Center, Chicago
- Deutsches Institut für Wirtschaftsforschung 2000 German Socio-Economic Panel Home Page, <http://www.diw-berlin.de/soep/>
- Duncan G, Kalton G 1987 Issues of design and analysis of surveys across time. *International Statistical Review* **55**: 97–117
- Elder G 1999 *Children of the Great Depression: Social Change in Life Experience*, 25th Anniversary Edition. Westview, Boulder, CO
- Eizenman D R, Nesselroade J R, Featherman D L, Rowe J W 1997 Intra-individual variability in perceived control in an older sample: The MacArthur successful aging studies. *Psychology and Aging* **12**: 489–502
- Freedman D, Thornton A, Camburn D 1980 Maintaining response rates in longitudinal studies. *Sociological Methods and Research* **9**: 87–9
- Furstenberg Jr F F, Brooks-Gunn J, Morgan S P 1987 *Adolescent Mothers in Later Life*. Cambridge University Press, New York
- Ghangurde P D 1982 Rotation group bias in the LFS estimates. *Survey Methodology* **8**: 86–101
- Hill M S 1992 *The Panel Study of Income Dynamics*. Sage, Newbury Park, CA
- Institute for Economic and Social Research, University of Essex 2000 British Household Panel Survey Home Page, <http://www.irc.essex.ac.uk/bhps>
- Institute of Education, Center for Longitudinal Studies 2000 British National Child Development Study Home Page, <http://www.cls.ioe.ac.uk/Ncds/nhome.htm>
- Johnson W, Skinner J 1986 Labor supply and marital separation. *American Economic Review* **76**(3): 455–69
- Kalton G 1986 Handling wave nonresponse in panel surveys. *Journal of Official Statistics* **2**: 303–14
- Katz L, Meyer B 1990 The impact of potential duration of unemployment benefits on the duration of unemployment. *Journal of Public Economics* **41**(1): 45–72
- Kish L 1986 Timing of surveys for public policy. *Australian Journal of Statistics* **28**: 1–12
- Lepkowski J 1989 The treatment of wave nonresponse in panel surveys. In: Kasprzyk D, Duncan G, Kalton G, Singh M (eds.) *Panel Survey Design and Analysis*. Wiley, New York pp. 348–74
- MaCurdy T, Mroz T, Gritz R M 1998 An evaluation of the National Longitudinal Survey on Youth. *Journal of Human Resources* **33**(2): 345–436
- Mathiowetz N, Duncan G 1984 Temporal patterns of response errors in retrospective reports of unemployment and occupation. In: *Proc. Survey Res. Meth. Sect.* American Statistical Association, Washington, DC, pp. 652–7
- Monitoring the Future 2000 Monitoring the Future Home Page, <http://monitoringthefuture.org/>
- Moss L, Goldstein H 1979 *The Recall Method in Social Surveys*. University of London Institute of Education, London
- Solon G 1992 Intergenerational income mobility in the United States. *American Economic Review* **82**(3): 393–408
- Smith T W 1994 Is there real opinion change? *International Journal of Public Opinion Research* **6**: 187–203
- Statistics Canada 2000 Guide to the Labour Force Survey Statistics Canada web site, <http://www.statcan.ca/>
- Traugott M, Katosh K 1979 Response validity in surveys of voting behavior. *Public Opinion Quarterly* **79**: 359–77
- Tuma N B, Hannan M T 1984 *Social Dynamics: Models and Methods*. Academic Press, Orlando, FL
- US Bureau of the Census 1978 *The Current Population Survey: Design and Methodology*. Technical Paper 40. US Government Printing Office, Washington, DC
- US Bureau of the Census 1999 Poverty in the United States: 1998 (P60–207). US Government Printing Office, Washington, DC
- US Bureau of the Census 2000 Survey of Income and Program Participation Home Page, <http://www.sipp.census.gov/sipp/sipphome.htm>
- US Bureau of Labor Statistics 2000a Current Population Survey Home Page, <http://www.bls.census.gov/cps/cpsmain.htm>
- US Bureau of Labor Statistics 2000b National Longitudinal Surveys Home Page, <http://stats.bls.gov/nlshome.htm>
- Yamaguchi K 1991 *Event History Analysis*, Applied Social Research Methods, Vol. 28. Sage, Newbury Park, CA
- Zick C, Smith K 1991 Patterns of economic change surrounding the death of a spouse. *Journal of Gerontology* **46**(6): S310–20
- Zimmerman D J 1992 Regression toward mediocrity in economic stature, *American Economic Review* **82**(3): 409–29.

G. J. Duncan

Panic Disorder

Panic disorder (PD) is an anxiety disorder associated with recurrent, unexpected panic attacks, characterized by sudden surges of physical arousal and fear. In addition, most individuals with PD develop agoraphobia to varying degrees. Agoraphobia is a fear of situations in which escape might be difficult or embarrassing, or in which help might be unavailable in the event of a panic attack.

The terms 'PD' and 'panic attack' first appeared in the diagnostic nomenclature with the publication of the third edition of the *Diagnostic and Statistical Manual of Mental Disorders* (DSM-III, American Psychiatric Association 1980). Initially, panic attacks were seen primarily as a feature of agoraphobia. In DSM-III, agoraphobia could be diagnosed with or without panic attacks, and individuals who experienced unexpected panic attacks in the absence of agoraphobia received a diagnosis of PD without agoraphobia.

More recently (e.g., DSM-IV, American Psychiatric Association 1994), the relationship between panic

disorder and agoraphobia has reversed, such that PD is seen as the primary problem, and agoraphobia is viewed as a complication of PD. The disorder 'agoraphobia with panic attacks' has essentially been replaced with 'panic disorder with agoraphobia' to reflect the fact that, in clinical groups, agoraphobia tends to develop as a response to having unexpected panic attacks, rather than the other way around (Craske and Barlow 1988).

1. Definitions and Description

1.1 Panic Attack

DSM-IV (American Psychiatric Association 1994) defines a panic attack as a period of intense fear or discomfort that begins quickly and peaks in ten minutes or less. A panic attack must also be accompanied by at least four out of a list of 13 symptoms, which includes eleven physical sensations (i.e., racing or pounding heart, sweating, trembling, shortness of breath, choking feelings, chest discomfort, nausea, dizziness, feelings of unreality or depersonalization, numbness or tingling sensations, hot flashes or chills) and two cognitive symptoms (i.e., fear of losing control or going crazy, fear of dying).

Panic attacks are common in the general population and are also associated with a broad range of psychological problems. For example, people who experience intense worry may have a panic attack in response to a particular worrisome thought. Individuals who are fearful of a specific situation (e.g., seeing a spider, being in a high place) may experience panic attacks when confronted with these situations. Panic attacks may also occur in the absence of any obvious cue or trigger. In fact, the hallmark symptom of PD is the presence of panic attacks that occur out of the blue.

1.2 Panic Disorder

People who suffer from PD experience recurrent unexpected or uncued panic attacks. In addition, they must experience, for at least one month, one or more of the following: (a) persistent worry about having more panic attacks; (b) frequent worry about the implications of consequences of the attacks (e.g., that an attack will lead to a heart attack, death, fainting, diarrhea, vomiting, embarrassment, or some other catastrophe); or (c) a significant change in behavior related to the attacks (e.g., avoidance of certain situations or activities). Finally, before assigning a diagnosis of PD, the clinician must establish that the panic attacks are not directly related to another psychological problem (e.g., a specific phobia), a medical condition (e.g., hyperthyroidism), or use of a substance (e.g., cocaine, caffeine). Panic disorder may be diagnosed with or without agoraphobia.

1.3 Agoraphobia

A significant proportion of individuals suffering from PD develop agoraphobia, a fear of situations in which escape might be difficult, or in which help might not be available in the event of a panic attack (American Psychiatric Association 1994). Typical situations that are often avoided by people with agoraphobia include crowded places (e.g., concerts, movies, shopping malls), driving, traveling, public transportation, flying, enclosed places, and meetings. For some individuals with agoraphobia, it is the distance from home, rather than the situation itself, that determines the level of fear and avoidance. For example, driving may be difficult, but only if the car ventures more than a few kilometers from home. For many people who suffer from agoraphobia, entering feared situations is easier if accompanied by a safe person, such as a spouse, family member, or close friend. In extreme cases, agoraphobia may be associated with a complete inability to leave the home, even when accompanied.

In addition to the obvious avoidance associated with agoraphobia, individuals with PD often engage in more subtle forms of avoidance. For example, a person might see a movie in a theater, but only if sitting in an aisle seat, near the exit. Or an individual might agree to ride a bus only while engaging in distracting behaviors such as reading or listening to a portable radio. Individuals with PD may also carry objects with them to feel safer, including medication, extra money in case of an emergency, or a portable telephone. Finally, they may engage in overprotective behaviors such as frequently checking their heart rate or blood pressure.

2. Epidemiology, Onset, and Demographics

Antony and Swinson (2000) reviewed the literature concerning age of onset and risk factors for developing PD. PD begins on average when a person is in his or her mid- to late-20s, although onset can occur any time between childhood and late in life. Risk factors include a period of life stress preceding the first attack, being female, a history of childhood separation anxiety, and a history of either substance abuse or dependence (Antony and Swinson 2000).

In the National Comorbidity Survey (a study of over 8,000 Americans in the general population), 3.5 percent of individuals reported symptoms meeting criteria for panic disorder with or without agoraphobia at some time in their life (Kessler et al. 1994). The lifetime prevalence of agoraphobia was 6.7 percent, with about two-thirds of agoraphobics having no prior history of panic attacks or panic disorder (Magee et al. 1996). This finding is in direct contrast to studies based on individuals in clinical settings, where agoraphobia almost never occurs separately from panic disorder (Pollard et al. 1989).

People with PD report impairment across a broad range of life domains including work, recreation, and social functioning (Antony et al. 1998). Furthermore, PD is associated with enormous costs to sufferers and to society, in the form of increased health-care utilization, lost wages, and time off work (e.g., Greenberg et al. 1999). PD is also associated with a higher risk of developing other psychological problems, including other anxiety disorders, depression, certain personality disorders (e.g., dependent, avoidant), and substance-use disorders (for a review, see Antony and Swinson 2000).

3. *Biological Approaches to Understanding Panic Disorder*

3.1 *Biological Models of Panic Disorder*

Most biological approaches to understanding PD have assumed a dysfunction in one or more neurotransmitter systems, including norepinephrine (NE), serotonin (5-HT), and cholecystokinin (CCK). In addition, theorists have hypothesized that genetic and neuroanatomical factors are also involved in the development of PD.

Perhaps the most influential biological model of PD in recent years has been Donald Klein's suffocation alarm model (Klein 1993), which proposes that panic attacks are caused by a pathological misfiring of the suffocation alarm system. According to Klein, humans are equipped with two types of alarm reactions. The first is controlled by the hypothalamic-pituitary-adrenal (HPA) axis and is responsible for dealing with real or perceived emergency situations. It is this alarm system that is activated when an individual is confronted with a feared object or situation.

The second alarm system is the suffocation alarm system, which is mediated by an area of the brain known as the *locus ceruleus*. This alarm system is triggered by increased levels of carbon dioxide in the blood, a sign that individuals might be in a small enclosed area, forced to breathe their own exhaled air, which is rich in carbon dioxide. Klein hypothesized that unexpected panic attacks occur when an individual has a pathologically lowered threshold for the activation of the suffocation alarm system. Presumably, the brains of such individuals are prone to mistakenly conclude that there is not enough air to breathe. As reviewed elsewhere (Antony and Swinson 2000), evidence regarding Klein's model has been mixed.

3.2 *Evidence Regarding Biological Approaches to Panic*

Despite inconsistent evidence for Klein's model, there are many studies supporting the view that biological

processes are involved in the pathophysiology of PD. PD tends to run in families and evidence from twin studies suggests that genetics plays a significant role in the transmission of PD across generations (e.g., Kendler et al. 1992). Findings from brain-imaging studies suggest that the temporal areas of the brain (particularly the right hippocampal region) are involved in the experience of panic attacks and fear (for a review, see Antony and Swinson 2000).

Evidence regarding the role of neurotransmitters in PD have included studies of altered receptor functioning, levels of particular neurotransmitters and related metabolites, and responses to substances that are known to cause changes in neurotransmitter levels. Among neurotransmitters, NE is the one most often implicated in PD and agoraphobia (Sullivan et al. 1999), although other neurotransmitter systems, such as 5-HT (Bell and Nutt 1998) and CCK (Bradwejn 1995) may also play a role. There is evidence that levels of NE metabolites predict panic-related symptoms (Garvey et al. 1990) and of altered NE receptor sensitivity in people suffering from PD (Nutt 1989). In addition, drugs that block the reuptake of NE appear to be effective for preventing panic attacks (Mavissakalian and Perel 1989).

Finally, a large number of studies have shown that certain substances that affect the NE system are more likely to trigger panic attacks in people with PD than in various comparison groups. Infusions of sodium lactate, carbon-dioxide inhalation, and yohimbine injections have all been shown to trigger panic attacks. However, it should be noted that psychological factors such as perceived control over the procedure and the presence of a safe person can influence whether an individual will experience a panic attack following exposure to these substances (for a review of panic induction studies, see Antony and Swinson 2000).

4. *Psychological Approaches to Understanding Panic Disorder*

4.1 *Psychological Models of Panic Disorder*

Attempts to explain PD using a psychological framework have come from a broad range of theoretical perspectives, including cognitive theory, learning theory, and psychodynamic theory. However, the most influential of these models have been those developed by cognitive and behavioral theorists such as David M. Clark (Clark 1986) and David H. Barlow (Barlow 1988).

From a cognitive behavioral perspective, PD is viewed as being maintained by a fear of benign physical sensations. For example, according to David Clark's cognitive model of PD, panic attacks stem from a tendency to misinterpret normal physical sensations (e.g., racing heart, dizziness) as indicating

some immediate danger. As a result of interpreting these sensations as being threatening, the symptoms escalate, leading to further catastrophic interpretations and predictions. Very quickly, the symptoms spiral into a full-blown panic attack, seemingly out of the blue. Clark and other cognitive theorists acknowledge that biological factors may also contribute to the experience of panic. The sensations that initially start the panic cycle may be caused by caffeine, intense emotions, exercise, hormonal changes, fatigue, hunger, hyperventilation, attention to bodily feelings, or other factors. Consistent with the cognitive behavioral view, 70 percent of individuals with PD report that their panic attacks are triggered by a physical sensation (Breitholtz et al. 1998).

4.2 Evidence Regarding Psychological Approaches to Panic

A number of studies have confirmed that people with PD are fearful of the physical sensations associated with panic attacks, compared to people without PD and people with other anxiety disorders (e.g., Taylor et al. 1992). In addition, researchers have demonstrated that people with PD are more likely than others to interpret ambiguous physical sensations as a sign of some impending physical or mental disaster and to believe these thoughts quite strongly (Clark et al. 1997).

Numerous studies have also found that people with PD have information-processing biases that are consistent with their fears. For example, a number of studies have found that PD is associated with a heightened ability mentally to track physical sensations such as heart rate (e.g., van der Does et al. 2000). In addition, results from studies of attention and memory have also been consistent with psychological models of PD. Compared to people without PD, individuals with PD attend more to threat-related information (e.g., McNally et al. 1994) and are more likely than nonanxious individuals to show threat-related biases in memory (e.g., Amir et al. 1996).

In summary, individuals with PD are typically more fearful of and more vigilant for panic-related sensations, compared to people without PD. PD is also associated with a tendency to hold fearful beliefs about panic-related sensations, and to show biases in attention and memory for threat-related information.

5. Treatment of Panic Disorder

In 1998, the American Psychiatric Association published practice guidelines for the treatment of PD. In addition to discussing important issues in the assessment and treatment of this condition, the guidelines reviewed the main approaches that have been shown to be useful for treating PD. Essentially, these

approaches include treatment with medications, and treatment with cognitive-behavior therapy (CBT). A review of studies on the relative and combined effects of these interventions is available elsewhere (e.g., Antony and Swinson 2000). A brief summary of the major findings is provided below.

5.1 Medication Treatments

A large number of controlled clinical trials have established that certain medications are useful for treating PD with and without agoraphobia. The tricyclic antidepressant imipramine is the most extensively studied medication that has been shown to be effective for blocking panic attacks (e.g., Mavissakalian and Perel 1989). More recently, controlled trials have established that the selective serotonin reuptake inhibitors or SSRIs (e.g., fluoxetine, fluvoxamine, sertraline, paroxetine, citalopram) are also useful for treating PD; in fact, the SSRIs are now often considered to be the pharmacological treatment of choice for this condition (for a review, see Antony and Swinson 2000).

Finally, controlled trials have found a number of other antidepressants to be useful, including venlafaxine (a new antidepressant) (Pollack et al. 1996) and phenelzine (a monoamine oxidase inhibitor or MAOI) (Buigues and Vallejo 1987). Studies with high-potency benzodiazepine anxiolytics such as alprazolam have consistently shown these medications to be effective as well (e.g., Ballenger et al. 1988).

Despite the long list of medication options for treating PD, there is little consistent evidence that any one of these medications is more effective than the others. Choosing among medications for a particular individual usually depends on factors such as potential side effects, possible problems during discontinuation (e.g., rebound anxiety), the patient's previous history of treatment response, the cost of the drug, and potential interactions with other medications, substances, or medical illnesses.

5.2 Cognitive-behavior Therapy

Cognitive-behavior therapy for PD includes four main types of strategies: (a) exposure to feared situations such as crowds, public transportation, and being away from home alone (i.e., *in vivo* exposure); (b) exposure to exercises that trigger feared sensations (e.g., overbreathing or spinning to induce feelings of dizziness); (c) cognitive therapy (i.e., systematically considering evidence that contradicts fearful thinking); and (d) relaxation-based strategies (e.g., learning to slow down one's breathing). Most of the time, treatment includes various combinations of these strategies.

A number of comprehensive reviews on CBT for PD have been published (e.g., Antony and Swinson 2000,

Barlow and Brown 1996, McNally 1994). Generally, there is little evidence that it is necessary to include all four treatment components to achieve maximum results. *In vivo* exposure to feared situations appears to be particularly important for treating individuals who are agoraphobic. In addition, relaxation-based strategies are thought to be less important than the exposure-based and cognitive techniques.

Although CBT for PD lasts 10 to 15 sessions, a number of recent studies suggest that gains can be made relatively quickly, with a minimal investment of therapist time. For example, telephone-administered CBT has been used to treat people who do not have easy access to a therapist (Swinson et al. 1995) and a single session of CBT has been shown to prevent the onset of PD in a significant percentage of individuals who present to hospital emergency rooms with a panic attack (Swinson et al. 1992). Finally, a number of studies have found that brief CBT lasting only five sessions (Clark et al. 1999) and CBT delivered in a self-help format (e.g., Hecker et al. 1996) are both useful for helping people to overcome PD (see *Cognitive and Interpersonal Therapy: Psychiatric Aspects*).

5.3 Combination Treatments

Studies comparing medications, CBT, and their combinations have yielded inconsistent results (Antony and Swinson 2000). However, averaging across a large number of studies, it is reasonable to conclude that both medications and CBT are effective for treating PD and there is no clear evidence favoring either antidepressants, high-potency benzodiazepines, CBT, or their combinations for treating panic attacks (e.g., Bakker et al. 1998, van Balkom et al. 1997). However, the combination of antidepressants and *in vivo* exposure appears to be particularly useful for treating agoraphobic avoidance (van Balkom et al. 1997).

It should be noted that most studies on the treatment of PD have focused primarily on the short-term effects of treatment over a period of only two to three months. Few studies have examined the relative effects of medications and CBT after treatment is discontinued. Exceptions include three studies comparing the relative and combined effectiveness of CBT with either imipramine (Barlow et al. 2000), alprazolam (Marks et al. 1993), or fluvoxamine (Sharp et al. 1996). In all three studies, individuals who received CBT (alone or in combination with medication) were more likely to maintain their gains over the long term than individuals who received only medication or placebo treatment.

Taken together, the literature on treating PD suggests that CBT and pharmacotherapy are both effective treatments in the short term. However, there is evidence that CBT is a more effective treatment for ensuring that gains are maintained over the long term.

See also: Anxiety and Anxiety Disorders; Anxiety and Fear, Neural Basis of; Fear Conditioning; Fear: Potentiation of Startle; Fear: Psychological and Neural Aspects; Mental and Behavioral Disorders, Diagnosis and Classification of; Panic, Sociology of;

Bibliography

- American Psychiatric Association 1980 *Diagnostic and Statistical Manual of Mental Disorders*, 3rd edn. (DSM-III). American Psychiatric Association, Washington, DC
- American Psychiatric Association 1994 *Diagnostic and Statistical Manual of Mental Disorders*, 4th edn. (DSM-IV). American Psychiatric Association, Washington, DC
- American Psychiatric Association 1998 Practice guideline for the treatment of patients with panic disorder. *American Journal of Psychiatry* **155**, suppl.: 1–34
- Amir N, McNally R J, Riemann B C, Clements C 1996 Implicit memory bias for threat in panic disorder: Application of the 'white noise' paradigm. *Behaviour Research and Therapy* **34**: 157–62
- Antony M M, Roth D, Swinson R P, Huta V, Devins G M 1998 Illness intrusiveness in individuals with panic disorder, obsessive compulsive disorder, or social phobia. *Journal of Nervous and Mental Disease* **186**: 311–15
- Antony M M, Swinson R P 2000 *Phobic Disorders and Panic in Adults: A Guide to Assessment and Treatment*. American Psychological Association, Washington, DC
- Bakker A, van Balkom A J L M, Spinhoven P, Blaauw B M J W, van Dyck R 1998 Follow-up on the treatment of panic disorder with or without agoraphobia: A quantitative review. *Journal of Nervous and Mental Disease* **186**: 414–19
- Ballenger J C, Burrows G D, DuPont R L, Lesser I M, Noyes R, Pecknold J C, Rifkin A, Swinson R 1988 Alprazolam in panic disorder and agoraphobia: Results from a multicenter trial. *Archives of General Psychiatry* **45**: 413–22
- Barlow D H 1988 *Anxiety and its Disorders: The Nature and Treatment of Anxiety and Panic*. Guilford, New York
- Barlow D H, Brown T A 1996 Psychological treatments for panic disorder and panic disorder with agoraphobia. In: Mavissakalian M R, Prien R F (eds.) *Long-term Treatments of Anxiety Disorders*. American Psychiatric Press, Washington, DC, pp. 221–40
- Barlow D H, Gorman J M, Shear M K, Woods S W 2000 A randomized controlled trial of cognitive-behavioral treatment vs. imipramine and their combination for panic disorder: Primary outcome results. *JAMA—Journal of the American Medical Association* **283**: 2529–36
- Bell C J, Nutt D J 1998 Serotonin and panic. *British Journal of Psychiatry* **172**: 465–71
- Bradwejn J 1995 Cholecystokinin and panic disorder. In: Bradwejn J, Vasar E (eds.) *Cholecystokinin and Anxiety: From Neuron to Behavior*. R. G. Landes, Austin, TX, pp. 73–86
- Breitholtz E, Westling B E, Öst L G 1998 Cognitions in generalized anxiety disorder and panic disorder. *Journal of Anxiety Disorders* **12**: 567–77
- Buigues J, Vallejo J 1987 Therapeutic response to phenelzine in patients with panic disorder and agoraphobia with panic attacks. *Journal of Clinical Psychiatry* **48**: 55–9
- Clark D M 1986 A cognitive approach to panic. *Behaviour Research and Therapy* **24**: 461–70

- Clark D M, Salkovskis P M, Hackmann A, Wells A, Ludgate J, Gelder M 1999 Brief cognitive therapy for panic disorder: A randomized controlled trial. *Journal of Consulting and Clinical Psychology* **67**: 583–89
- Clark D M, Salkovskis P M, Öst L G, Breitholtz E, Koehler K A, Westling B E, Jeavons A, Gelder M 1997 Misinterpretation of body sensations in panic disorder. *Journal of Consulting and Clinical Psychology* **65**: 203–13
- Craske M G, Barlow D H 1988 A review of the relationship between panic and avoidance. *Clinical Psychology Review* **8**: 667–85
- Garvey M, Noyes R, Cook B 1990 Comparison of panic disordered patients with high versus low MHPG. *Journal of Affective Disorders* **20**: 7–12
- Greenberg P E, Sisitsky T, Kessler R C, Finkelstein S N, Berndt E R, Davidson J R T, Ballenger J C, Fyer A J 1999 The economic burden of anxiety disorders in the 1990s. *Journal of Clinical Psychiatry* **60**: 427–35
- Hecker J E, Losee M C, Fritzler B K, Fink C M 1996 Self-directed versus therapist-directed cognitive behavioral treatment for panic disorder. *Journal of Anxiety Disorders* **10**: 253–65
- Kendler K S, Neale M C, Kessler R C, Heath A C, Eaves L J 1992 The genetic epidemiology of phobias in women: The interrelationship of agoraphobia, social phobia, situational phobia, and simple phobia. *Archives of General Psychiatry* **9**: 273–81
- Kessler R C, McGonagle K A, Zhao S, Nelson C B, Hughes M, Eshleman S, Wittchen, H U, Kendler K 1994 Lifetime and 12-month prevalence of DSM-III-R psychiatric disorders in the United States: Results from the National Comorbidity Survey. *Archives of General Psychiatry* **51**: 8–19
- Klein D F 1993 False suffocation alarms, spontaneous panics, and related conditions: An integrative hypothesis. *Archives of General Psychiatry* **50**: 306–17
- Magee W J, Eaton W W, Wittchen H U, McGonagle K A, Kessler R C 1996 Agoraphobia, simple phobia, and social phobia in the national comorbidity survey. *Archives of General Psychiatry* **53**: 159–68
- Marks I M, Swinson R P, Basoglu M, Kuch K, Noshirvani H, O'Sullivan G, Lelliott P T, Kirby M, McNamee G, Sengun S, Wickwire K 1993 Alprazolam and exposure alone and combined in panic disorder with agoraphobia: A controlled study in London and Toronto. *British Journal of Psychiatry* **162**: 776–87
- Mavissakalian M R, Perel J M 1989 Imipramine dose-response relationship in panic disorder with agoraphobia: Preliminary findings. *Archives of General Psychiatry* **46**: 127–31
- McNally R J 1994 *Panic Disorder: A Critical Analysis*. Guilford, New York
- McNally R J, Amir N, Louro C E, Lukach B M, Riemann B C, Calamari J E 1994 Cognitive processing of ideographic emotional information in panic disorder. *Behaviour Research and Therapy* **32**: 119–22
- Nutt D J 1989 Altered central α_2 -adrenoceptor sensitivity in panic disorder. *Archives of General Psychiatry* **46**: 165–9
- Pollack M H, Worthington J J, Otto M W, Maki K M, Smoller J W, Manfro G G, Rudolph R, Rosenbaum J F 1996 Venlafaxine for panic disorder: Results from a double-blind placebo-controlled study. *Psychopharmacology Bulletin* **32**: 667–70
- Pollard C A, Bronson S S, Kenney M R 1989 Prevalence of agoraphobia without panic in clinical settings. *American Journal of Psychiatry* **146**: 559
- Sharp D M, Power K G, Simpson R J, Swanson V, Moodie E, Anstee J A, Ashford J J 1996 Fluvoxamine, placebo, and cognitive behaviour therapy used alone and in combination in the treatment of panic disorder and agoraphobia. *Journal of Anxiety Disorders* **10**: 219–42
- Sullivan G M, Coplan J D, Kent J M, Gorman J M 1999 The Noradrenergic system in pathological anxiety: A focus on panic with relevance to generalized anxiety and phobias. *Biological Psychiatry* **46**: 1205–18
- Swinson R P, Fergus K D, Cox B J, Wickwire K 1995 Efficacy of telephone-administered behavioral therapy for panic disorder with agoraphobia. *Behaviour Research and Therapy* **33**: 465–9
- Swinson R P, Soulios C, Cox B J, Kuch K 1992 Brief treatment of emergency room patients with panic attacks. *American Journal of Psychiatry* **149**: 944–6
- Taylor S, Koch W J, McNally R J 1992 How does anxiety sensitivity vary across the anxiety disorders? *Journal of Anxiety Disorders* **6**: 249–59
- van Balkom A J L M, Bakker A, Spinhoven P, Blaauw B M J W, Smeenk S, Ruesink B 1997 A meta-analysis of the treatment of panic disorder with or without agoraphobia: A comparison of psychopharmacological, cognitive-behavioral, and combination treatments. *Journal of Nervous and Mental Disease* **185**: 510–16
- van der Does A J W, Antony M M, Barsky A J, Ehlers A 2000 Heartbeat perception in panic disorder: A re-analysis. *Behaviour Research and Therapy* **38**: 47–62

M. M. Antony

Panic, Sociology of

The term 'panic' is widely used in everyday speech as well as in the literature of different professional areas and scientific disciplines. This leads to a very wide diversity in the attribution of both the characteristics of, and the conditions that generate, the phenomenon. However, this article confines itself primarily to how sociologists view 'panic.' Thus, it does not consider the substantial but circumscribed literature in economics that focuses solely on 'financial panics,' or another set in psychiatry that deals with 'panic states' from a mental health viewpoint. Although a few sociologists have looked at the first phenomenon under the general rubric of panic, the second has never been of any concern to them.

The focus on sociology is justified for several reasons. From the very origins of the field, sociologists have used the term as part of their professional vocabulary. The first introductory sociology textbook states that 'panic is the crowd in dissolution' (Park and Burgess 1924, p. 876) because it is the opposite of the crowd in the sense of not having any psychological unity. In particular, panic has always been considered part of the subject matter of the sociological specialization of collective behavior, which deals with nontraditional and newly emergent forms of social action (see Sociology of Collective Behavior). Most US textbooks and extended theoretical treatises on collective behavior discuss panic (Miller 1985) with some giving

extensive attention to the phenomenon. Furthermore, such empirical studies of panic as have been undertaken have been mostly by sociologists studying human and group behavior in natural and technological disasters (see *Disasters, Sociology of*). Even psychologists, who next to sociologists have been most interested in panic, primarily cite empirical and theoretical sociological sources (Schultz 1964).

1. Past Approaches

The first extended and systematic treatment of panic by a sociologist is a chapter by LaPiere (1938) in a little-known book on collective behavior. He sees panic as dysfunctional escape behavior generated by fortuitous, ever-varying circumstances, but involving impending danger. Panic is never formally defined except as the antithesis of regimental behavior, which is preplanned collective action for dealing with crises, such as fire drills in schools. Instead what constitutes panic is illustrated by presentations of anecdotal examples from stories of disaster behavior in journalistic and popular sources. While his vague formulations and use of popular sources are typical, LaPiere differs from many later writers in three major ways. He does not generally deal with any covert emotional or feeling state associated with the behavior, and he avoids use of the term 'irrational.' In addition, he distinguishes between collective and individual panic, which is at variance with most treatments, which imply that panic necessarily involves a multiplicity of participants.

As just implied, there is little consensus, historically or currently, on the use of the term. Nevertheless, most discussions about the nature of panic can be grouped into one of two categories. The oldest view, coming out of everyday speech, primarily equates panic with extreme and groundless fear. This is clearly related to the linguistic origins of the word which is derived from the Greek god Pan, who supposedly was able to generate sudden and overwhelming fear in the absence of any actual threat. For many writers taking this position, being seized by such a fear can lead to other irrational reactions although the essence of panic is the emotional manifestation, and not what this might otherwise affect.

Another view visualizes panic as manifesting itself primarily as flight behavior. In this conception, the essence of panic is the overt behavior that is marked by the setting aside of everyday social norms, even the strongest, such as parents abandoning their young children when trying to save themselves in a life-threatening crisis. Often implicit, there is the assumption in this view that such flight behavior will occur only if there is a perception that there is a possibility of escaping the threat. Disaster researchers in particular have emphasized that hope of escape rather than hopelessness is what is involved. People who perceived

themselves as totally trapped, such as in sunken submarines or collapsed coal mines, do not panic because they see no way of getting away from the threat.

Of course it is possible to put the two general formulations together. Smelser (1963, p. 131) defines panic as 'collective flight based on a hysterical belief,' which allows him to talk both about overt escape behavior and the selling behavior of investors in financial panics. But others have noted that the relationship between a terror state and rout behavior is not necessarily a direct one, given that whatever the emotion, it need not inevitably lead to flight of any kind. Apart from agreeing that panic participants are very afraid of some perceived threat, there is little agreement otherwise among students of the phenomenon on the relationship of the emotional state involved to what other, if any, behaviors will be manifested.

2. Current Issues

Apart from differences about the nature of panic, many discussions of panic focus on two other major themes. One such theme, and a very prominent one, assumes that panic behavior is 'irrational' (cf Le Bon [1895] 1960). This is often contrasted with the assumed rationality of most other behavior, where the means-ends relationships are in balance or where the end result is a positive one. This conception has come under sharp criticism, especially from those who have done empirical studies of the behavior of people caught in disasters and fire situations (Quarantelli 1981). These researchers note that when the behavior on such occasions is looked at from the perspective of the social actors involved, the behavior is very meaningful and far from most conceptions of irrationality (Johnson 1985). The argument is that nothing is gained by characterizing the behavior as irrational or along any similar dimension (Wenger 1980). However, it is still very common and typical for social science textbooks that discuss panic to characterize it as irrational.

A second major division among scholars is between those who argue that panic behavior is very contagious, and that human beings are easily swept up into the behavior, and those who strongly disagree with such a conception. Many of the early discussions of panic clearly assume that the participants are overwhelmed by the fearful emotion of others, and will trample over others in their path. In contrast are those researchers who take the position that whatever panic behavior is, it results from meaningful social interaction among the participants. They note that even in those rare cases of extreme instances of panic flight, very few spectators ever get caught up in the behavior.

Several specific studies have particularly influenced all of the previous arguments. One is a study by Cantril (1940) on reactions in the USA to a nationally broadcast radio show supposedly reporting as actual fact an alien invasion from Mars. The study takes the view that those who panicked upon hearing the broadcast lacked 'critical ability.' The behavior was seen as irrational. While even to this day this study is cited as scientific support for this view of panic, the research has come under sharp critical scrutiny. Analysts have noted that even taking the data reported at face value, only a small fraction (12 percent) of the radio audience ever gave even any remote credence to the idea that the broadcast was an actual news story. And the accounts of flight behavior as well as other illogical and bizarre actions reported in the book give the mistaken idea that they were obtained in the survey study done. Actually, they are taken almost unacknowledged from journalistic accounts of the time which reported that supposedly numerous people fled wildly to get away from the alien invaders. A close scrutiny of the actual news report that appeared found that most of them were sensationalized anecdotal stories that these days are typically reported in the tabloid press.

The last point is particularly important because much later Rosengren and his colleagues (1978) studied in a very systematic manner a roughly similar radio show. A Swedish radio station broadcast a fictitious news program about a nuclear plant accident that had created radioactive clouds drifting in southern Sweden. Within an hour, other media were broadcasting reports about widespread panic reactions in the population, with the print media later reporting panic reactions of various kinds on a large scale including escape flight. However, the systematic survey study found less than 10 percent of the audience gave any credence to the broadcast, with less than 1 percent showing any behavioral reaction, and no one who had engaged in flight behavior. The research showed that the supposed panic reaction of the population was almost exclusively a mass media creation. There is very little found by this research that would be supportive of any of the formulations about panic advanced to that time by social scientists.

More recently, Johnson (1988) did intensive studies of persons caught in potential panic-provoking situations such as a fire in a night club and a stampede during a rock music concert, where 160 and 11 persons respectively died. The findings are unambiguous. The great majority of involved persons did not engage in animal-like behavior, contrary to what many early writers on panic suggest occurs. Instead of ruthless competition, the social order did not break down, with cooperative rather than selfish behavior predominating. Contrary to notions of irrationality, there was much evidence for rational responses in the face of the crisis. While strong emotions were experienced, these did not lead to maladaptive behavior. These findings

reinforced the ever-growing viewpoint among many researchers who have studied what seem potential panic situations, that prosocial rather than antisocial behavior predominates even in such contexts.

3. *Future Questions*

Two additional and fundamental but unresolved questions regarding the concept of panic are surfacing. Most scholars seem to agree that whatever 'panic' might mean, the phenomenon is statistically quite rare, usually involves only a handful of persons, and is of short duration. Some researchers have observed that it is very difficult to find clear-cut cases of actual panic in natural and technological disasters (and that they are also extremely rare in the other arena in which they supposedly occur, that is among soldiers in battles during wars). But the term continues to be widely used and persists, despite, the lack of empirical evidence that it happens on any significant scale; it also continues as noted by students of popular culture to be the staple of disaster movies and novels.

Wolfenstein (1957) explicitly questioned why, despite what the research evidence showed, the idea of 'panic' captures the popular imagination and continues to be evoked by scholars of human behavior. Using a psychoanalytical framework, she suggests that the fascination with the term is a psychological fantasy actually useful and functional in a variety of ways for coping with personal crises. Some disaster scholars have thought that a more sociological approach might parallel what Durkheim, a major figure in the origins of sociology, said about crime. If it did not exist, it would be necessary for human societies to create crime, at least symbolically, so as to emphasize the fact that human beings generally adhere to social norms. So perhaps the idea of the possibility of panic is necessary in society to highlight the fact that human beings in contrast react remarkably well in most stressful situations and that the social bonds between and among people usually hold. Supporting this view, as disaster researchers have noted, the mass media in reporting the *absence* of panic are behaving as if the normal expectation is that panic will occur. Whatever the merit of the parallel to crime, clearly there is a need to explain the huge discrepancy between the actual frequency of panic in any sense of the term, and the exaggerated interest in and widespread use of the word, not only in popular culture but also especially among social scientists who have not researched the phenomena.

A second basic question is whether there is still any scientific justification for the continuing use of the concept in any technical sense. Increasingly, students of the phenomenon have questioned if the term 'panic' ought to be continued to be used by students of collective behavior (Mawson 1980, Sime 1981). The logic of the argument is that such behavior as is attempted to be captured under that label can better be

dealt with by using other concepts. In other words, what is often currently described and analyzed as panic behavior can be characterized and explained by other terms. Those taking this view argue that because a word has widespread popular currency, and has been unthinkingly imported into a scientific field, this should not be seen as necessarily giving the term any legitimacy. From this perspective, it is possible that the concept of panic within collective behavior in sociology may disappear as a technical term in the future.

Bibliography

- Cantril H 1940 *The Invasion From Mars: A Study in the Psychology of Panic*. Princeton University Press, Princeton, NJ
- Foreman P 1953 Panic theory. *Sociology and Social Research* 37: 295–304
- Johnson N 1985 Panic and the breakdown of social order: Popular myth, social theory and empirical evidence. *Sociological Focus* 20: 171–83
- Johnson N 1988 Fire in a crowded theater: A descriptive investigation of the emergence of panic. *International Journal of Mass Emergencies and Disasters* 6: 7–26
- Keating J 1982 The myth of panic. *Fire Journal* 77: 57–62
- LaPiere R 1938 *Collective Behavior*. McGraw-Hill, New York
- Le Bon G [1895] 1960 *The Crowd*. Viking, New York
- Lofland J 1981 Collective behavior: The elementary forms. In: Rosenberg S, Turner R (eds.) *Social Psychology: Sociological Perspectives*. Basic Books, New York, pp. 411–46
- Mawson A 1980 Is the concept of panic useful for scientific purposes? In: Leven B, Paulsen R (eds.) *Second International Seminar on Human Behavior in Fire Emergencies*. National Bureau of Standards, Washington, DC, pp. 208–13
- Miller D 1985 *Introduction to Collective Behavior*. Wadsworth, Belmont, CA
- Park R, Burgess E 1924 *Introduction to the Science of Sociology*. University of Chicago Press, Chicago
- Quarantelli E 1954 The nature and conditions of panic. *American Journal of Sociology* 60: 265–75
- Quarantelli E 1957 The behavior of panic participants. *Sociology and Social Research* 41: 187–94
- Quarantelli E 1981 Panic behavior in fire situations: Findings and a model from the English language research literature. In: *Proceedings of the 4th Joint Panel Meeting*. The UJNRR Panel on Fire Research and Safety Building Research Institute, Tokyo
- Rosengren K, Arvidsson P, Stureson D 1978 The Barseback 'panic': A case of media deviance. In: Winick C (ed.) *Deviance and Mass Media*. Sage, Beverly Hills, CA, pp. 131–49
- Schultz D (ed.) 1964 *Panic Behavior: Discussion and Readings*. Random House, New York
- Sime J 1981 The concept of panic. In: Canter D (ed.) *Fires and Human Behavior*. Wiley, New York, pp. 63–81
- Smelser N 1963 *Theory of Collective Behavior*. Free Press, New York
- Strauss A 1944 The literature on panic. *Journal of Abnormal and Social Psychology* 29: 317–28
- Wenger D 1980 Some observations on the nature of panic behavior: A normative orientation. In: Leven B, Paulsen R (eds.) *Second International Seminar on Human Behavior in Fire Emergencies*. National Bureau of Standards, Washington, DC, pp. 214–19

Wolfenstein M 1957 *Disaster: A Psychological Essay*. Free Press, Glencoe, IL

E. L. Quarantelli

Paradigms in the Social Sciences

How is one to explain the diffusion of the concept of 'paradigm,' a word of Greek origin used by some philosophers during the eighteenth century, and which, in the writings of some contemporary scholars, has a pedantic tone? One of the reasons may be the fact that it raises what is, in the history of science, a fundamental question: is scientific progress achieved mostly by steady accretion or mostly by abrupt jumps? This concept was devised by Thomas Kuhn in a particular context, and the circumstances of its formulation poses directly the question to be treated here: Are there paradigms in the social sciences?

1. A Polysemic Concept

Thomas Kuhn explains in the preface to *The Structure of Scientific Revolutions* (Kuhn 1970) that it was during a stay at Palo Alto Center for Advanced Studies, in the company of social scientists, that he was led to formulate the concept of paradigm with the aim of making clear the essential difference between natural sciences and the social sciences. The reason given by Kuhn was the absence of a theoretical consensus in any discipline of social sciences.

I was struck by the number and extent of overt disagreement between social scientists about the nature of scientific problems and methods... Controversies over fundamentals seem endemic among, say psychologists or sociologists... Attempting to discover the source of that difference led me to recognize the role in scientific research of what I have since called paradigms (Kuhn 1970, p. VIII).

For him, the use of the word paradigm in the social sciences is not justified.

As a preliminary it should be noted that many scholars have formulated doubts about the reality of paradigms even in the natural sciences—an issue that cannot be discussed here. Kuhn himself has in the postscript to the second edition of his book reformulated the concept of paradigm, by disentangling this concept from the notion of scientific community, and by admitting that he has used the term in two different senses. On one hand, paradigm stands for a constellation of beliefs, values, techniques and so on shared by the members of a given community; on the other, it denotes a concrete puzzle-solution. He labeled the first meaning 'disciplinary matrix': 'disciplinary, because it refers to the common possession of

the practitioners of a particular discipline; matrix, because it is composed of ordered elements of various sorts' (Kuhn 1970, p. 182). Kuhn makes a distinction between paradigms in the large sense, the disciplinary matrix, and small paradigms called 'exemplars.' This distinction is considered unclear by several sociologists of science.

In the revised version of Kuhn's thesis

any period of scientific development is marked by a large number of overlapping and interpenetrating paradigms, which do not replace each other immediately and do not spring up full-borne, but instead emerge in a long process of intellectual competition... These concessions dilute Kuhn's original message, and in its final version the argument is difficult to distinguish from the average historian's account of the history of science (Blaug 1976, p. 154).

Among the sociologists of science, very few have asked if the concept of paradigm is, or not, applicable to social sciences. Harriet Zuckerman in her substantial *Sociology of Science* (Zuckerman 1988) does not dedicate a single line to this issue, nor do Merton, Price, Collins, Lakatos, or Mullins.

Are there in the social sciences instances of paradigmatic upheavals comparable to those created by Copernicus, Newton, Darwin, or Einstein? Can the theories of Keynes, Freud, Chomsky, or Parsons be described as paradigmatic? In the social sciences, does progress occur through paradigmatic revolutions or through cumulative processes?

For there to be a paradigm, one condition must be met: theories must refer to essential aspects of social reality. However, the more ambitious a theory is, the less it can be directly tested by the data available. In the social sciences there are no 'fundamental discoveries,' as there sometimes are in the natural sciences. Instead, unverifiable theories are constructed, partly because social reality itself changes. Kuhn held that a paradigm shift occurs when two 'incommensurable' paradigms are in competition. Then, each school rejects the premise of the others, and contests their work. Communication breakdown is the sign of an incompatibility. In the social sciences such 'incommensurability' takes the pattern of mutual ignorance.

2. Mutual Ignorance

There are two kinds of mutual ignorance: involuntary and deliberate. The formal disagreement may be based on personal rivalries, on contest about priority, on methodological quarrelling or on ideological hostility.

Within a formal discipline, several major theories may cohabit, but there is a paradigm only when one testable theory alone dominates all other theories and is accepted by the entire scientific community. When

Pasteur discovered the microbe, the theory of spontaneous generation collapsed: contagion became the new paradigm. In the social sciences, however, we see at best a confrontation between several nontestable theories. Most of the time there is not even a confrontation but careful mutual avoidance, superb disregard on all sides. This is relatively easy, owing to the size of scientific disciplines, and their division into schools; it is true for all countries, big or small.

This mutual disregard is an old practice in the social sciences. At the turn of the twentieth century, the great scholars did not communicate, or very little. In the writings of Weber there is no reference to his contemporary Durkheim. Yet Weber was acquainted with Durkheim's journal, *l'Année Sociologique*. For his part, Durkheim, who could read German, makes only one, fleeting reference to Weber. Yet they worked on a number of the same subjects, such as religion.

As Giddens puts it:

Weber certainly knew of the work of Durkheim but there can be little doubt that he was barely influenced at all... The same applies in a reciprocal vein; Durkheim knew of Weber's work, and there is a minor reference to aspects of it in one passage in Durkheim's writings, but he claimed no particular kinship with those of Weber's ideas with which he was familiar (Giddens 1987, p.182).

Durkheim does no more than mention in passing Simmel and Tonnies.

Harshly criticized by Pareto, Durkheim never alluded to Pareto's work. Pareto's judgment of Durkheim's book on suicide was unfavorable. 'Unfortunately,' he wrote, 'its arguments lack rigour.' Weber seems to have been unaware of Pareto's theory on the circulation of elites, and Pareto in his turn says nothing about the Weberian theory of political leadership. There was no exchange between Weber and Freud. Ernst Bloch and Georg Lukacs met regularly with Weber in Heidelberg, but their work shows no sign of his influence. The only one of Weber's contemporaries who referred to him was Karl Jaspers, but he was a philosopher (compare Mommsen and Osterhammel 1987).

Weber and Spengler had not exerted the slightest influence on the ideas of the other. Croce said about Weber: 'I had met Weber in 1909. I had no further contact with him, nor did I follow his scholarly work, from which I had only read his early book *Roman Agrarian History*.' As was noted by Raymond Aron, each of the three great founders of sociology followed a 'solitary path.' There is no room for two theorists like Parsons and Sorokin in the same habitat; the clash between them became inevitable. The Parsonian grand theory had a short life. Robert Dahl and Arnold Rose refute the theory of C. Wright Mills; they do not adopt the stratagem of deliberate ignorance—they contest it openly. On the other hand, C. Wright Mills chooses to

ignore the books published by Berle and Means and by James Burnham, in spite of his own theory.

Mosca had accused Pareto of plagiarism. The imprecision of the language of both elitist theorists, who adopted synonyms with confusing meanings, has generated in Italy in enormous literature with the participation of dozens of Italian scholars debating the priority of the authors of the 'political class' and of 'political elites' (Albertoni 1983). Pareto and Mosca are not mentioned by José Ortega Y Gasset (1883–1955). Schumpeter does not refer to Mosca, Pareto, Michels, or Burnham. Angus Campbell and Paul Lazarsfeld have spent a large part of their parallel life in the study of the same political behavior; they never cite each other. Johan Galtung and Stein Rokkan born in 1930 and 1921 respectively, both citizens of the same small country (Norway), have each contributed a theory of 'center versus periphery.' They have never collaborated, avoiding each other carefully; they coexisted by ignoring each other.

Three lists of major contributions to economics were compiled in 1982 at a symposium in Berlin. The first, prepared by W. Krelle, contained 30 names, the second, by J. Tinbergen, 36 names, and the third, by B. Fritsch and G. Kirchgässner, 44 names. In the first two lists, there were only two names that were the same (including Klein, Nobel prize winner, and Krelle himself?); in the first and the third, there were only nine that were the same. The second and third lists did not have a single name in common (Platt 1986, p. 350). Such widely differing views about leading figures do not say much for the coherence of economics. This is also the opinion of the economist Kenneth Boulding, who speaks of economics as being 'disorientated,' comparing it with the story of the blind man and the elephant. The Nobel prize winner Wasily Leontieff was not more indulgent: 'Year after year, economic theorists produce mathematical models without being able to get any further towards understanding the structure and functioning of the real economic system' (quoted in Platt 1986, p. 350). The monetarist George Stigler objected to six econometricians (Tinbergen, Samuelson, Malinvaud, Moore, Frisch, and Fisher) being described as the authors of 'exceptional contributions' because, he wrote, 'econometrics has no unified core or methodology' and 'has not yet had a major impact on economics' (Platt 1986, p. 342). But three of these six economists have since then won the Nobel prize, as has Stigler himself. 'The existence and unity of a discipline called economics reside in the eye and mind of the beholder. Samuelson, Solow, and Tobin and others reserved their harshest words for neoclassical economists like Friedmann and members of the Chicago School, who ridicule Keynes's consumption function' (Amariglio et al. 1993, pp. 150 and 175). The same remarks can be made with reference to other disciplines, except linguistics.

If scholars belonging to the same generation and working on the same subjects ignore each other,

whether involuntarily or deliberately, how could one perceive the existence of a scientific community able to adhere to a common paradigmatic *Weltanschauung*? Paradoxically, the closer the proximity, the more vivid is the rivalry and the more frequent the mutual ignorance. To the isolation of scholars motivated by competition should be added the weak communication between specialties within the same discipline (see *Specialization and Recombination of Specialties in the Social Sciences*).

The absence of consensus among contemporaries is compensated by the transmission of knowledge from one generation to the next, as revealed by citation patterns, particularly in handbooks and compendia. Such a generational transmission testifies that in social sciences scientific advancement occurs mostly by cumulative knowledge.

3. Cumulative Knowledge

Science consists largely in the adding of successive layers of sediments, which form a patrimony for each discipline and field. In all sciences 'the extent of the innovation that any individual can produce is necessarily limited, for each individual must employ in his research the tools that he acquires from a traditional education, and he cannot in his own lifetime replace them all' (Kuhn 1970, p. 183). Progress does not arise in a vacuum, but develops out of the scientific patrimony. It is difficult to find in the social sciences a virgin domain. Every decade that passes adds layer upon layer to the patrimony, the only recent exception being the 1930s in Europe. New explanations supersede older interpretations. Many examples of cumulative progress can be given. Even giants rely on the patrimony. Karl Marx refers to Adam Smith on 296 of the 1,721 pages of his *Theory of Surplus Value*. In this book he draws on the work of his predecessor in more than one page out of every six. Ralf Dahrendorf cites Marx 160 times (Dahrendorf 1957). Max Weber does not cite Karl Marx, but many of his writings were a reply to the thesis of his predecessor. He once noted that he could not have accomplished crucial parts of his work without the contributions by Marx and Nietzsche.

Darwin had recognized his debt to Malthus; Keynes would not have been able to write one of the most famous books of the twentieth century without the incremental advance achieved by several generations of economists. The 'theory of development' consists of a long chain of accumulated contributions in several disciplines.

The literature on elites is a good example, among others, of cumulative knowledge even in the case of mutual ignorance. Many contributions emphasize the sedimentation of layers of knowledge. The old Mosca–Pareto layer, impregnated with plagiarism, has become consolidated groundwork. Two dozen

important books published since 1950 have built an impressive patrimony. Obsolete theories have served as fertilizers for new complementary theories. One example of cumulative knowledge is the new influence of electoral techniques on party systems. A bibliography, even a very selective one, on this theme could easily list 200 or 300 titles in English, not to mention the many varied observations derived from the direct experience of politicians in numerous countries. From Condorcet, Bachofen, John Stuart Mill, Hare, and Hondt to Hermens, Downs, Duverger, Sartori, Lijphart, the theory is based on the contributions and successive improvements made by a very large number of scholars. (For example, the consequences of proportional representation were described by Bachofen as long ago as 1850).

For each of the dozens of important domains in the social sciences there is a substantial patrimony, the accumulated knowledge of a long chain of scholars. Imagine the following imposture-exercise. Take from the literature a classic article or chapter, for instance Max Weber's seminal chapter on 'Domination and Legitimacy,' falsify the name of the author, make a few cosmetic changes, and offer it to a major journal. Would such a manuscript be accepted today? Would not a paper that ignores the literature of the last decades preclude publication? The patrimony is common property. Although every concept or theory or method has its creators and developers, they do not need to be cited every time the term is used. Even the identity of the originator may disappear into anonymity. We do not, and cannot, remember who used for the first time terms such as 'role,' 'revolution' or 'social mobility.' Given such a patrimony, scholars today can start their research at a much higher level than did their predecessors (Dogan and Pahre 1990, p. 21). A graduate student today knows more than the founders of his or her field—even if they do not have the reasoning capacity of their forebears. That knowledge is largely acquired by cumulation is an obvious fact for all sciences; if it were otherwise, simultaneous discoveries in the natural sciences would not occur, theoretical debates in the social sciences would not be engaged. They are generated when the issue is 'in the air,' when 'time is ripe.' The dispute between Mosca and Pareto about the paternity of ideas in elite studies is similar to the dispute between Newton and Leibnitz about priorities on calculus. In elite studies, for variety of reasons 'the time was ripe' at the end of the nineteenth century. 'Discoveries are like apples on a tree,' wrote Price (Kuhn 1986, p. 60). In the history of social sciences some books have played a crucial role but they are only summits in a chain of mountains, not paradigmatic contributions in the sense given to this word by Thomas Kuhn.

Overarching theories, that is, paradigmatic frameworks, can be built on more solid ground in the natural sciences than in the social sciences, because in the former truth is universal, in the later, contextual.

4. Contextual Diversity vs. Universal Truth

Chemical substances keep indefinitely the same composition and are identical in all latitudes. Not so social phenomena! In contrast with the natural sciences, contextual diversity and social change are two important parameters in all social sciences. Both parameters resist ambitious paradigmatic generalizations. Extrapolating Kuhn's perception of the history of astronomy, physics, or chemistry to social sciences is like looking into a distorting mirror. Many concepts and theories do not have equal pertinence when applied to Western Europe and to tropical Africa. Dependency theory may be a pertinent theory for yesterday's Latin America and for Eastern Europe today, but not for other areas in the world. The comparative method reduces the significance of paradigmatic frameworks. Dozens of theories with universal pretensions neglecting national configurations have been invalidated.

In the natural sciences an experience can be repeated many times in identical conditions anywhere in the world (except when magnetism has an impact). On the contrary, social realities change. For instance, at the time of Malthus, overpopulation in Europe was a realistic hypothesis. Today, birth rates in this continent are not sufficient to replace the outgoing generation. Traditional values concerning religion, nation, family, erotism, or authority have changed at an accelerated rhythm in the last decades of the twentieth century. Theories reflecting the social reality one generation ago have to be replaced in order to explain contemporary trends. History never repeats itself because factors do not recombine in identical ways at all epochs. The scope of social sciences is rarely the discovery of laws with universal validity, but rather the explanation of social diversity. 'What is true in one side of the Pyrénées is not in the other side,' wrote Montaigne four centuries ago. In the social sciences truth is not universal, it is contextual. Social scientists do not make inventions, and rarely make discoveries; what they do best is to observe regularities and exceptions (anomalies). It is for this reason, that the most imaginative strategy in the social sciences is the comparative method, the best surrogate for the experimental method. Social phenomena being contextual, paradigmatic statements in political science, sociology, anthropology, social psychology or social history often appear as ethnocentric and misleading.

Most social scientists who adopt the concept of paradigm apply it to particular domains, not to their entire discipline. For some scholars, elitism is the most powerful paradigm against the Marxist paradigm. In the domain of political economy, rational choice is elevated to the rank of paradigm. Many psychiatrists are mobilized around the Freudian paradigm.

Competing theories contribute to the advancement of knowledge. To explain the fall of the Roman empire, 14 competing theories have been formulated

by great scholars, from Montesquieu to Max Weber, to which can be added a theory proposed by chemistry (saturnism). None is fully comprehensive, but each one enlightens part of the phenomenon. The history of social sciences is not a history of paradigmatic upheavals, but of competing theories, with many being invalidated, but many others constituting the foundations of the contemporary social sciences. Without competing theories, social sciences would not advance. The clash of theories leaves no room for paradigms.

A symptomatic analogy can be drawn between what Kuhn calls in the revised version of his theory, 'scientific community' (adepts of a paradigm) and what some social scientists (Almond, Merton, Smelser, Zuckerman) call 'school.' In the recent history of social sciences there are dozens of schools: monetarist, rational-choice, bio-social, individualist, postmodernist, and so on, and that means that attempts to formulate overarching disciplinary paradigms may appear as 'imperialist views.'

As Merton put it:

Since all virtues can readily become vices merely by being carried to excess, the sociological paradigm can be abused almost as easily as it can be used. It is a temptation to mental indolence. Equipped with his paradigm, the sociologist may shut his eyes to strategic data not expressly called for in the paradigm. He may turn the paradigm from a sociological field-glass into a sociological blinker. (Merton 1957, p. 16).

In some cases a paradigm may assume the features of a dogmatic orientation. No wonder if not so many scholars adopt it: 'The notion of paradigms and paradigmatic revolutions often seem to be taken up only in order to be rejected. The beneficiary of this exercise in several cases is Lakatos, mostly because he appears as a type of moderate Kuhn' (Weingart 1986, p. 267). In the social sciences, theoretical disagreements are beneficial to the advance of knowledge. Nevertheless, the word paradigm has taken root, particularly in sociology, political science, psychology, and normative philosophy. Yet most philosophers of science reject it; most historians are reluctant to make such generalizations; most economists continue to think in terms of assumptions. It may be too late now to try to exclude this word from the lexicon, in spite of the fact that many other expressions are available (conceptual framework, assumption, dominant theory, theoretical breakthrough, grand theory, general model, axiom, and so on). It has become necessary to specify it, or to limit its use to particular domains, such as cognitive science, international relations, or hybrid demography.

See also: History of Science; History of Science: Constructivist Perspectives; Problem Selection in the Social Sciences: Methodology; Quantification in the History of the Social Sciences; Science, Sociology of; Theory: Conceptions in the Social Sciences

Bibliography

- Albertoni E (ed.) 1983 *Governo e governabilità nel sistema politico e giuridico di Gaetano Mosca*. Giuffrè, Milan, Italy
- Amariglio J, Resnick S, Wolff R 1993 Division and difference in the 'Discipline of Economics.' In: Messer-Davidow E, Schunway D, Sylvain D (eds.) *Knowledges: Historical and Critical Studies in Disciplinarity*. University Press of Virginia, Charlottesville, VA, pp. 150–4
- Blaug M 1976 Kuhn versus Lakatos or paradigms versus research programmes in the history of economics. In: Latsis S J (ed.) *Method and Appraisal in Economics*. Cambridge University Press, Cambridge, MA, pp. 149–61
- Dahrendorf R 1957 *Class and Class Conflict in Industrial Society*. Stanford University Press, Stanford, CA
- Dogan M, Pahre R 1990 *Creative Marginality: Innovation at the Intersection of Social Sciences*. Westview Press, Boulder, CO
- Giddens A 1987 Weber and Durkheim: Coincidence and divergence. In: Mommsen W J, Osterhammel J (eds.) *Max Weber and Contemporaries*. Allen and Unwin, London, pp. 182–9
- Kuhn T S 1970 [1962] *The Structure of Scientific Revolution*. University of Chicago Press, Chicago, IL
- Merton R K 1957 *Social Theory and Social Structure*. The Free Press, New York
- Mommsen W J, Osterhammel J 1987 *Max Weber and His Contemporaries*. Allen and Unwin, London
- Platt J 1986 Is everything equal? In: Deutsch K W, Markovits A S, Platt I (eds.) *Advances in the Social Sciences*. University Press of America, New York
- Price D de S 1986 [1963] *Little Science, Big Science and Beyond*. Columbia University Press, New York
- Weingart P 1986 T.S. Kuhn: Revolutionary or agent provocateur? In: Deutsch K W, Markovits A S, Platt J (eds.) *Advances in the Social Sciences*. University Press of America, AS, pp. 265–75
- Zuckerman H 1988 The sociology of science. In: Smelser N (ed.) *Handbook of Sociology*. Newbury Park, CA, pp. 511–99

M. Dogan

Para-lawyers: Other Legal Occupations

1. Definition

Para-lawyers are persons performing 'law work' who themselves are not lawyers. Law work in turn refers to those activities requiring at least some of the specialized legal knowledge and skill of the type associated with the work of lawyers. Thus, para-lawyers are 'law workers' who lack the full credentials and/or status associated with the legal profession but who provide services that draw on expertise in law and/or legal processes. This negative definition reflects the complexity of defining the practice of law and what constitutes membership in a legal profession.

Drawing the line between lawyers and para-lawyers varies from country to country because of variations in the nature of legal tasks within different legal

systems, and how those tasks are assigned to particular occupations. In many countries, the exact line is not clear. In some countries, the term most equivalent to 'lawyer' is customarily applied to an occupation engaged in a very narrow range of activities, while there are other occupations requiring substantial formal legal training that perform work which elsewhere would be the exclusive domain of the legal profession. Deciding whether to include such occupations under the label lawyers or para-lawyers requires drawing an arbitrary line. One example is the legal occupation of 'notary' as it exists in many civil law countries (Malavet 1996). Historically, the occupation of notary arose to provide a neutral person who could be trusted to draw up fair contracts to be executed by one or more persons unable to read or write; today, notaries continue to play a role in the drawing up of and execution of contracts, and apply specialized legal knowledge related to those activities. If the notary's tasks involve anything more than ministerial functions such as witnessing the execution of contracts, it is arbitrary whether or not the notary is part of the legal profession or should be categorized as a para-lawyer.

In addition to problems of drawing a line between lawyers and para-lawyers, it is difficult to draw a line between para-lawyers and 'nonlaw' occupations. For example, many nonlaw occupations require the use of some very specific legal knowledge, typically applied in a very routine way. Examples include drawing up contracts for routine property transactions (in the United States typically done by real estate sales agents), drawing up routine loan contracts (done by bank officials), settling tort-based injury claims (done by nonlawyer representatives of insurance companies), or advising on estate or tax issues (done by financial advisors and accountants). In one sense, all of these occupations could be labeled para-lawyers; for some of the occupations the para-lawyer work is only a part of their activities (e.g., real estate sales agents) while for others (e.g., insurance claims representatives) it is the vast bulk of their work.

Finally, there is no recognized standard of education or specialized training to become a para-lawyer. While there are training courses for occupations variously labeled 'paralegal,' 'legal assistant,' or 'legal executive,' these courses do not provide formal qualifications, and most people enter such occupations without any formalized training (Johnstone and Wenglinisky 1985, pp. 121–3). Other kinds of para-lawyer occupations involve the application of knowledge learned in some other role; examples include former police officers representing persons charged with driving while intoxicated in Ontario, former Internal Revenue Service (IRS) auditors representing or assisting taxpayers in disputes with the IRS in the United States, former insurance company claims inspectors or adjusters representing injury claimants in England, and legal secretaries who take on in-

creasing responsibilities to the point where they are engaged more in law-work than in secretarial work. Thus, there is no clear way to decide whether a person offering a particular set of services, or engaged in a particular set of tasks, should be classified as a para-lawyer; nor is the term para-lawyer in common use in a way that would allow for self-identification.

2. Empirical and Theoretical Issues

Even with these definitional and boundary problems, the concept of para-lawyers, and para-professionals more generally, is useful theoretically and important for practical purposes. It raises the issues of how law-work is and should be structured, and what qualifications are needed to be effective in law work. In addition, there are two other lines of theoretical inquiry that are important vis-a-vis para-lawyers: 'deprofessionalization' and the gendering of occupations.

2.1 Structure of Law Work

Law work readily divides along at least three dimensions: institutional setting, legal advice vs. legal assistance, and routine vs. nonroutine tasks. The degree to which para-lawyers have in the past, and are today, providing legal services varies along these dimensions. As one moves among institutional settings—from office, through tribunal, to lower courts, and finally to the higher courts—the role of para-lawyers tends to decrease. Para-lawyers are most often excluded from law work in upper level trial and appeal courts, and least often from tasks that can be completed in the privacy of an office. For example, in England, lawyers enjoy a monopoly on representation in the higher courts; in contrast, office-based work generally is not limited to lawyers, nor is representation before administrative tribunals. Recent developments are bringing nonlawyers into some courts in England, particularly to handle preliminary matters in criminal cases.

Continuing with the example of England, para-lawyers direct a larger proportion of their effort toward providing advice on legal matters, broadly defined, while lawyers are more likely to engage in active assistance. In part this is because accounting firms and advice bureaus staffed by nonlawyers are major providers of legal advice in England. In other countries, it is likely to be the case that para-lawyers are more involved in advice-giving simply because it is less visible, and hence difficult for regulators to limit or otherwise control.

The more that routines can be established to handle a legal task, the more possible it is for someone with very specific training and/or expertise to handle the task, even if it is highly technical. For example, in the

US, appeals of certain types of government benefit claims involve highly technical issues of both regulations and medical assessment. Such cases tend to be handled by specialists, some of whom are lawyers and some of whom are para-lawyers. On the less technical side, there are many legal processes that are routine but require the completion of various kinds of forms; para-lawyers can both identify the forms needed for various processes and assist consumers in the completion of those forms.

One final issue concerning the structure of legal work and para-lawyers is whether they work as independent service providers or under the supervision of lawyers. Many lawyers employ para-lawyers to complete routine and/or specialized tasks within the lawyers' practices (Johnstone and Wenglinisky 1985); many legal departments inside organizations employ a mix of lawyers and para-lawyers all under the supervision of senior lawyers. Typically, para-lawyers working under the supervision of lawyers is an accepted delegation of responsibility, although there is substantial variation in drawing the line between what a para-lawyer may do and what requires a lawyer. One frequent boundary is the courtroom door: representation in court tends to be limited to lawyers, although that may be less so in the lower courts (the example of drunk driving cases in Ontario noted previously), and is changing under both political and economic pressures (the example of para-lawyers handling certain criminal matters in England, and the role of specialized advocates in certain types of cases such as domestic violence).

2.2 Effectiveness in Delivery of Legal Services

A key question with regard to the work of para-lawyers is how it compares in quality to the same activities carried out by lawyers. While lawyers claim that use of para-lawyers raises threats of poor quality work, efforts to compare systematically the work of lawyers and para-lawyers fail to support this contention (Kritzer 1998). The extant research demonstrates that para-lawyers can be as effective as lawyers in providing specific legal services. Achieving effectiveness involves a combination of specific knowledge or expertise vis-a-vis substantive law, institutional process, and other actors in the system. Para-lawyers who obtain this combination can achieve levels of effectiveness comparable to that of lawyers, and specialized para-lawyers will tend to be more effective than lawyers lacking one or more of the three core elements of knowledge.

The relationship between the service provider and the client can also influence the relative effectiveness of the provider. Specialist para-lawyers who have ongoing relationships with clients may be more effective than even specialist lawyers brought in on an *ad hoc* basis to handle a specific matter; this reflects the

knowledge base created by the ongoing relationship. The payment arrangement between service provider and client may also make a difference. One study that found specialist lawyers to be more effective than specialist para-lawyers in a particular area attributed the difference in significant part to the lawyers being paid on a no-win, no-fee basis while the para-lawyers' compensation was unaffected by the outcome of specific cases.

2.3 Deprofessionalization and Postprofessionalism

'Deprofessionalization' refers to the shift from professional to nonprofessional status. Professional status is conferred on those occupations which maintain a monopoly over a theoretical base of knowledge, have expectation of autonomy in regulation and day to day work, and have among their clientele a belief in the professional having a service ethos. Deprofessionalization refers to the loss of these supposedly unique occupational qualities (Haug 1973, Rothman 1984). Nonprofessionals such as para-lawyers increasingly do work once the province of professionals.

Deprofessionalization reflects several specific developments. First is the combination of the rationalization of knowledge, the rationalization of tasks employing knowledge, and the growth of tools to assist in applying knowledge. The more that tasks can be compartmentalized and the more that knowledge-based tools can be applied, the easier it is to assign the task to a person with limited, specific expertise. The result is the deskilling of knowledge-based occupations. A second development is the loss of governmental support for restrictions limiting the role of para-lawyers. In England solicitors held a monopoly on handling land transfers for many years, but these tasks are now open to licensed non-lawyer specialists; in the US, limitations have been under attack politically at both the national and state level (Commission on Nonlawyer Practice 1995). The third development is the increasing bureaucratization of legal service delivery; once organizational structures become established, there is a drive to rationalize, compartmentalize, and reduce costs. Bureaucratic structures also provide means for supervision and control, which both creates tensions for professionals expecting autonomy and provides a setting to monitor the work of para-lawyers.

A traditional claim by lawyers seeking to secure and maintain restriction on the delivery of legal services is that opening such work to para-lawyers and other nonlawyers jeopardizes the interest of clients both because of lack of competence and the absence of the stringent ethical norms of the professional. As noted in the previous section, systematic research in the United States and elsewhere increasingly challenges this claim. The loss of this argument combined with drives to contain costs through broadening options in

service-oriented markets pushes toward less and less control of 'professional' services.

The concept of deprofessionalization suggests the loss of something. An alternative concept, 'postprofessionalism' (Kritzer 1999) reflects most of the same elements but sees the opening of legal and other professional services to para-professionals such as para-lawyers as part of a natural development. Postprofessionalism builds on the growth of technological tools to access knowledge previously the exclusive realm of persons with professional-level training (Susskind 1996). These developments parallel the industrial revolution when the rationalization of tasks and the development of machines led to the shift from crafts-based production to factory-based production. Postprofessionalism captures the image of a similar shift for people who work with their heads rather than their hands: rationalization of knowledge combined with information technology tools make it possible to produce services using workers with lower levels of training and knowledge.

2.4 Gendering of Occupations

A final issue in the analysis of para-lawyers relates to the gendered nature of occupations. Historically, professions have been male-dominated while many para-professions have been female dominated. Within law-work, women were largely excluded from the professional ranks until about 1970. In the early twentieth century some of the American states formally excluded women from the bar; even where they were not excluded, women were often confined to roles such as legal secretary even if they had the formal credentials of a law degree. Since 1970, most countries have witnessed the opening of the legal profession to women, and the number of women who practice as lawyers has increased sharply.

Nonetheless, para-lawyer roles tend to be dominated by women; this is particularly true of para-lawyer positions that are specifically subordinate to lawyers (Johnstone and Wenglinisky 1985, p. 69). Even for those para-lawyer positions that function independently of lawyers, the gender composition may limit the likelihood of achieving some sort of full professional status as long as they are seen as in competition with the existing legal profession (Hearn 1982). At the same time, efforts to increase occupational status may lead to certain kinds of 'closure' strategies which bear a resemblance to those employed by lawyers: formal licensing, testing requirements, educational qualifications, etc. (Witz 1992, Johnstone and Wenglinisky 1985, pp. 165–6). Whether such efforts ultimately improve status or lock in gendered distinctions is an important theoretical and empirical issue.

A final gender-related issue concerns the professional roles assumed by many of the increasing number of women in the legal profession. A common

observation in the literature on the struggles of women to achieve status and economic parity with men in the legal profession is that women lawyers are disproportionately to be found in law-work positions that are themselves in some sense gendered (e.g., family law practice, as employees rather than as principals, etc.). In some ways, the professional roles occupied by many women lawyers are closer to those of para-lawyers than to the traditional image of the work of male lawyers (Sommerlad and Sanderson 1998). This development may lead to a blurring of lines between lawyers and para-lawyers while preserving patriarchal status structures in the workplace.

3. Future Research and Theory

As theoretical and empirical concepts para-lawyers and para-professionals more generally, are underdeveloped. Current dynamics in the changing occupational structures, such as postprofessionalism, suggest that the para-lawyers should become increasingly important in the study of law work and in the delivery of legal services. The relationship of para-lawyers to the increasing role of information technology is one avenue of potential theory and research; a second is the economics of service delivery; and a third potential area is in the relationship of para-lawyer occupations to broader patterns of changing social structures, particularly the changing impact of gender on social structure.

See also: Counsel, In-house; Lawyers; Lawyers, Regulation of; Legal Professionalism; Professionalization/Professions in History; Professions, Sociology of

Bibliography

- Commission on Nonlawyer Practice 1995 Nonlawyer Activity in Law-Related Situations: A Report with Recommendations. American Bar Association, Chicago
- Haug M R 1973 Deprofessionalization: An alternative hypothesis for the future. *Sociological Review Monograph* 20: 195–217
- Hearn J 1982 Notes on patriarchy, professionalization and the semi-professions. *Sociology* 16: 184–98
- Johnstone Q, Wenglinisky M 1985 *Paralegals: Progress and Prospects of a Satellite Occupation*. Greenwood Press, Westport, CT
- Kritzer H M 1998 *Legal Advocacy: Lawyers and Nonlawyers at Work*. University of Michigan Press, Ann Arbor, MI
- Kritzer H M 1999 The professions are dead, long live the professions: Legal practice in a post-professional world. *Law & Society Review* 33: 713–59
- Malavet P A 1996 Counsel for the situation: The Latin American notary, a historical and comparative model. *Hastings International and Comparative Law Review* 19: 389–488
- Rothman R A 1984 Deprofessionalization: The Case of Law In America. *Work and Occupations* 11: 183–206
- Sommerlad H, Sanderson P 1998 *Gender, Choice and Commitment: Women Solicitors in England and Wales and the Struggle for Equal Status*. Ashgate Publishing, Aldershot, UK

Susskind R 1996 *The Future of Law: Facing the Challenges of Information Technology*. Clarendon Press, Oxford, UK
 Witz A 1992 *Professions and Patriarchy*. Routledge, London

H. M. Kritzer

Parapsychology

According to *The Journal of Parapsychology* the term 'parapsychology' designates 'The branch of science that deals with *psi* communication, i.e., behaviorial or personal exchanges with the environment which are extrasensorimotor—not dependent on the senses and muscles.' The same source defines 'psi' as 'A general term to identify a person's extrasensorimotor communication with the environment.' Psi includes ESP and PK, 'ESP' (extrasensory perception) is defined as 'Experience of, or response to, a target object, state, event, or influence without sensory contact.' ESP includes telepathy (mind to mind communication without normal channels of communication), clairvoyance (extrasensory contact with the material world), and precognition (the knowledge of future events that cannot be inferred with present knowledge). The other component of psi, PK (psychokinesis) is defined as 'The extramotor aspect of psi: a direct (i.e., mental but nonmuscular) influence exerted by the subject on an external physical process, condition, or object.'

J. B. Rhine adapted the term parapsychology from the German word *Parapsychologie* in the 1930s to replace the earlier term *psychical research*. More than a change in terminology was involved. Psychical research covered a broad range of systematic investigations into spiritualistic phenomena, haunted houses, premonitory dreams, visions, and the like. Rhine's program avoided much of this material, especially that which dealt with questions of survival after death. Instead, he emphasized controlled laboratory experiments using normal individuals as subjects. Thus, the change in terminology signaled that parapsychology aspired to become an accepted scientific discipline.

1. Historical Background

1.1 Scientists and Psychic Phenomena

In 1848, the two young sisters Margaret and Katherine Fox initiated modern spiritualism with their demonstrations of rappings and other material phenomena which they claimed were caused by the spirits of dead people. Soon many other individuals, calling themselves mediums, were regularly producing phenomena which they attributed to spirits of the dead. Because science deals with material phenomena, and because

material phenomena were what the spiritualists were offering as support for their paranormal claims, some major scientists became interested in investigating these claims. In 1853, Robert Hare, an important chemist from the University of Pennsylvania, began an investigation of table-tilting. He attended seances where he became convinced that the spirits of the dead were moving the tables.

In 1865, Alfred Russel Wallace, the cofounder with Darwin of the theory of evolution through natural selection, began his investigations of spiritualistic phenomena. He quickly became convinced of the reality of the paranormal and for the remaining 48 years of his life he continued his investigations and outspokenly defended the reality of spiritualistic phenomena. Sir William Crookes, the discoverer of thallium and the inventor of the cathode ray tube, began his investigations of the paranormal in 1869. He concluded, to the dismay of his scientific colleagues, that a psychic force was operating in these seances.

1.2 The Society for Psychical Research

Inspired by the investigations of such eminent scientists as Hare, Wallace, and Crookes, a group of scholars and spiritualists formed The Society for Psychical Research in London in 1882. The philosopher Henry Sidgwick, its first president, stated that the goal was to collect more evidence of the kind that had been reported by those scientists who had investigated mediums. Although Sidgwick believed that the previous investigations of mediums had already provided evidence sufficient to prove the existence of the paranormal, he was aware that the majority of the scientific community were not convinced. Therefore, the aim of the Society was to accumulate enough additional evidence of the same sort so as to force the scientific community to admit that the case for the paranormal had been proved or to accuse the proponents of insanity or gross incompetence.

In his inaugural presidential address, Sidgwick claimed that, in fact, some members of the new society had already collected just the sort of evidence that proved beyond a doubt the existence of telepathy. Sidgwick was referring to the investigations involving the Creery sisters. Unfortunately, the Creery sisters were later discovered to be signaling to each other using a code. Although an attempt was made to claim that such a code could not have accounted for some of the results of their investigations, this body of evidence was subsequently removed from the republished proceedings of the Society for Psychical Research. A similar fate befell the very next set of investigations which the Society originally offered as solid proof for the existence of telepathy. These investigations involved two young men, Smith and Blackburn, who apparently could communicate with each other tele-

pathically. Many years later, Blackburn published a confession of how he and Smith had tricked the investigators. This case was also deleted from later printings of the Society's proceedings (Hyman 1985a).

2. From *Psychical Research* to *Parapsychology*

2.1 J. B. Rhine

Joseph Banks Rhine and his wife Louisa came to Duke University in 1927 and developed a program of research based on card guessing experiments. They termed their program 'parapsychology' and focused their efforts on making this discipline into an experimental science. They replaced the regular deck of playing cards with a deck consisting of five distinct symbols—a circle, a square, three wavy lines, a cross, and a star. Each symbol occurred five times, making a deck consisting of 25 cards. The same set of cards could be used as targets for experiments in telepathy, clairvoyance, and precognition. In addition, the symbols could be presented in a variety of ways.

In a typical clairvoyance condition, for example, the deck would be shuffled and the subject would be asked to guess the symbol of each card at each location in the deck. Since there were five different symbols, the subject had one chance in five of being correct on each guess just by chance. In other words, the subject would be expected to average five correct hits on each run through a deck of 25 cards. In a precognition test, the subject would make his or her guesses before the deck was shuffled. In a telepathy test, a sender would look at the symbol on each card at the time the subject was making his or her guess.

When Rhine published his first monograph, *Extra-Sensory Perception*, in 1934 he had collected data on a total of 91,174 guesses in three years. The average number of hits per run of 25 guesses was 7.1 as compared with the chance expectation of 5. Although the results were barely 2 hits per deck higher than the expected 5, the probability of such a departure from chance, given the huge number of trials, is very low.

The scientific community, at first, took great interest in this outcome. Never before had so much scientific data been amassed in favor of a paranormal claim. After some years, however, skepticism and criticism from the scientific community prevailed. In part, this was because several attempts by independent investigators to replicate these findings failed. In addition, criticisms were aimed at both the statistical procedures as well as the experimental controls. For the most part, the statistical criticisms turned out to be unfounded or insufficient to explain away the results. The methodological critiques were more serious, and gradually parapsychologists improved their procedures for preventing sensory leakage, randomizing the targets, specifying the criteria in advance, and the like.

The paradigm that Rhine initiated under the rubric of parapsychology transformed psychical research into a serious contender for a place among the accepted sciences. It also established the basic procedures for parapsychology for a period covering the years from 1930 to around 1970. During this time the ESP cards were the predominant stimulus for experiments in telepathy, clairvoyance, and precognition. Rhine also initiated the study of psychokinesis (PK) by having subjects attempt to influence the outcomes of dice tosses. The full story of the first couple of decades of this formative period is recounted in Mauskopf and McVaugh (1980).

2.2 S. G. Soal

As already indicated, the problem of replicability haunted the early results of Rhine's card-guessing experiments. Rhine and other US parapsychologists claimed that their results, when they were successful, were equally good for telepathy and clairvoyance trials. The British parapsychologists, however, claimed that their successful experiments only involved telepathy and never clairvoyance. Perhaps the strongest critic of Rhine's early work was the British mathematician, S. G. Soal. During the years 1934–9, Soal managed to amass 128,350 guesses from 160 percipients. He reported that the number of hits was just what would be expected by chance.

A fellow parapsychologist, Whately Carington, eventually convinced Soal to review his data to see if some of his percipients might reveal a 'displacement effect.' Perhaps, he suggested, some of the subjects' guesses were systematically off by one or two targets. Soal, indeed, discovered that two of his 160 subjects showed significant hitting above chance when their guesses were matched not against the intended target but, instead, against the target for the next trial.

Soal realized that finding such displacement patterns in two out of 160 subjects after the fact could be a statistical quirk. He was able to find and persuade one of these subjects, Basil Shackleton, to participate in 11,378 more guesses during the years 1941–3. As was the case in the original experiment, Shackleton's guesses, when matched against the actual target, were consistent with chance. However, when compared with the symbol that came after the target (precognitive hitting), his guesses were well above chance. Soal reported the odds against chance for such a pattern as 10^{35} to 1. Beginning in 1945, Soal was able to gather 37,100 guess from the second subject, Gloria Stewart. Unlike her previous results or those of Shackleton, Mrs. Stewart showed no displacement effect. Instead, her guesses of the intended target were significantly above chance. Soal calculated the odds against chance of this outcome as 10^{79} to 1.

Such strikingly successful results over a period of years, especially from a strong critic of Rhine, caught

the attention of both parapsychologists and skeptics. Parapsychologists hailed the experiments as definitive proof of psi. Critics looked for possible flaws or weaknesses. The early critics managed to find some suspicious patterns, but nothing conclusive. Some proposed fanciful scenarios, involving the collusion among several investigators, to account for the results.

It wasn't until 1978 that Betty Markwick published her findings which convinced most parapsychologists and all the skeptics that Soal had indeed cheated (Hyman 1985a). Ironically, Markwick made her discovery while she was carrying out an elaborate investigation to vindicate Soal's reputation. As part of her plan, she searched tables of logarithms for matches to sequences of targets in Soal's experiments. Because Soal had claimed that he had used Chambers' tables for his random sequences, Markwick reasoned that if she could find matches between Soal's target sequences and sequences in Chambers' tables, this would show that, at least, Soal had not tampered with the target sequences. What she found were partial matches. Many of Soal's sequences had extra insertions within the sequence and these extra insertions usually corresponded with hits. The strong implication was that Soal had deliberately constructed his target sequences with blank spaces inserted systematically. In recording the percipient's guesses, he could surreptitiously insert target numbers in the blank spaces that would match the guess for that trial. Such insertions would suffice to bias the results towards a highly significant, but spurious, outcome.

2.3 *The Post-Rhine Era*

Around 1970, Rhine's specific targets and procedures began to be discarded in favor of new targets and procedures. The striking successes of the early 1930s were no longer in evidence. Results were inconsistent, replicability was still elusive, and the hit rate—even in the few successful outcomes—had become very small. Cynics attributed the declining hit rate to better controls and other methodological and statistical improvements.

Most parapsychologists, however, blamed the weak or nonexistent effects to the austerity of Rhine's procedures. The targets on the ESP cards were meaningless and abstract. The constraint on the subjects' responses to just one of a few choices and the rigidity of the experimental plans all conspired, in the view of many parapsychologists, to dampen the delicate ESP ability.

So parapsychologists began using targets that were emotionally and otherwise meaningful—actual photographs, motion picture clips, actual physical sites, etc. They also eschewed the forced-choice methods of Rhine in favor of free-responding, i.e., the percipients were encouraged to say whatever came to mind and to describe their 'mentation' in whatever words they

chose. In addition, they began to take advantage of new technological breakthroughs, especially video and computer developments. These more natural targets and fewer constraints upon the subjects' responding made problems of control and scoring more complicated, but parapsychologists felt that this was a small price to pay in order to achieve what they call psi-conducive conditions.

3. *The Contemporary Scene*

Contemporary parapsychology has generated a variety of novel approaches. Computers are central to much of the research. A line of research that goes back to the late 1960s involves having subjects attempt to influence psychically the output of random number generators (RNG). Computers are also used to control many of the aspects of conducting experiments such as selecting targets, recording the guesses, reporting the outcomes, etc. Advances in neuroscience are also exploited and many experiments use physiological indices instead of, or in addition to, verbal responses. Among these new approaches, three lines of research predominate. The research on remote viewing began in 1972 at the Stanford Research Institute. Until 1995, much of that research, conducted under the auspices at first of the CIA and later the Defense Intelligence Agency was classified. The work was declassified in 1995 and evaluated by a panel consisting of the statistician, Jessica Utts, and the psychologist, Ray Hyman. Hyman and Utts disagreed about the scientific status of remote viewing. Utts concluded that the evidence clearly supported the existence of psi and Hyman that the scientific claim had not been proven. Both agreed, however, that the results of remote viewing were too unreliable to justify using them for information gathering.

The second line of research, initiated by Helmut Schmidt in 1969, involves having subjects attempt to influence or bias the outputs of random number generators (RNGs). Although many different investigators and laboratories have published research on apparent psychokinesis or PK effects on RNGs, the overwhelming number of trials in this line of research has come from the Princeton Engineering Anomalies Research (PEAR) program at Princeton University under Robert Jahn and his associates. In 1987, Radin and Nelson (Radin 1997) conducted a meta-analysis of the RNG experiments conducted from 1959 through 1987. A meta-analysis is a method for combining or pooling the results from several previously published reports involving similar procedures. They reported that the combined outcome produced an effect with odds against chance of more than a trillion to one. Radin claims that the results of such studies completed since 1987 are still consistent with that meta-analysis.

Despite these claims for the RNG research, the studies have been criticized on a number of grounds. Although the meta-analysis pools the data from all the laboratories, the pooling does not deal with the various and different patterns of outcomes from separate laboratories. For example, the PEAR researchers claim to have found a distinct pattern: The RNG outputs slightly more positive units when the subject is trying to influence it in that direction and slightly more negative units when the subject is trying to bias the output in the negative addition. In the neutral condition, the output is evenly divided but the variability is less than would be expected by a truly normal distribution. When the results from these three conditions are pooled the distribution becomes perfectly normal with a mean of zero. No other laboratory has replicated this interesting pattern. The replicability that Radin and Nelson claim is a rather weak one. In effect, any departure from the chance baseline, no matter the specific pattern, is considered equal to any other departure. The one physicist who has deliberately tried to replicate Jahn's RNG findings, Stanley Jeffers, has failed to do so.

3.1 The Ganzfeld-psi Experiments

The line of research that has had the most influence since the early 1970s and is at the center of parapsychology's claim of a demonstrable psychic anomaly is the ganzfeld-psi program. The March 1985 issue of *The Journal of Parapsychology* was devoted to 'The Ganzfeld Debate.' This debate consisted of a 47-page critique by Ray Hyman of the 42 ganzfeld-psi experiments known to exist at that time. This was followed by a 47-page response by Charles Honorton, one of the pioneers of this type of experiment.

Hyman (1985b) agreed that, taken as group, the overall hit rate in the ganzfeld experiments exceeded what would be expected by chance. However, he identified a number of statistical and methodological flaws characterizing the experiments. These were flaws which parapsychologists recognize as ones that should be avoided. Yet, not one of the experiments in this data base was free of all the flaws. Hyman concluded that, given the various flaws, the alleged significant levels had been vastly overinflated and the set of experiments could not be used to draw any conclusions about the existence of psi.

In his response, Honorton (1985) devised a different method for assigning flaws and concluded, contrary to Hyman, that the flaws, while there, had no impact on the outcome of the experiments. The controversy continued with parapsychologists supporting Honorton's analysis and the skeptics supporting Hyman's. Rather than continue with further responses and rebuttals, Hyman and Honorton collaborated on a joint paper to focus on areas of agreement and to suggest guidelines for future ganzfeld-psi research (Hyman and Honorton 1986).

In 1990 Honorton and his colleagues published an article in the *Journal of Parapsychology* (Honorton et al. 1990) which summarized six years of new ganzfeld-psi experiments supposedly meeting the guidelines devised by Hyman and Honorton. The new set of experiments, mainly controlled by computers, became known as the autoganzfeld experiments. The hit rate was consistent with the previous set of ganzfeld experiments and was highly significant. In 1994, a major psychological periodical, the *Psychological Bulletin*, published the article 'Does psi exist? Replicable evidence for an anomalous process of information transfer' jointly authored by the psychologist Daryl Bem and Charles Honorton (Bem and Honorton 1994) (unfortunately, Honorton died before the article was published). The same issue contained a commentary by Hyman (Hyman 1994) as well as a 'Response to Hyman' by Daryl Bem (Bem 1994).

Hyman questioned the claim by Bem and Honorton that the autoganzfeld findings were consistent with the original ganzfeld data base. He discovered two peculiar and striking patterns in the autoganzfeld data. While Hyman found these two patterns suggestive of an artifact, Bem responded that if it turned out to be a property of psi, parapsychologists should call it the Hyman Effect.

The ganzfeld-psi research continues in many laboratories around the world. Radin (1997) describes a meta-analysis of all the ganzfeld-psi studies through early 1997. He concludes that 'The overall hit rate of 33.2 percent is unlikely with odds against chance beyond a million billion to one' (p. 88). This contrasts sharply with the conclusions from a meta-analysis of 30 ganzfeld experiments by Milton and Wiseman (1999). These experiments were conducted after the autoganzfeld experiments and yielded an effect size consistent with zero. Milton and Wiseman 'conclude that the ganzfeld technique does not at present offer a replicable method for producing ESP in the laboratory.'

4. The Case for Psi

The strongest arguments for the reality of psi have been presented by Jessica Utts (1991) and Dean Radin (1997). Both base their case upon meta-analyses of previously published articles. Utts relies upon meta-analyses of the ganzfeld-psi experiments, early precognition card-guess experiments, RNG studies, early dice-throwing experiments, and studies relating extroversion to ESP performance. Radin includes these as well as several other meta-analyses.

The meta-analyses are used to draw two conclusions. The first is that the pooled effect size emerging from each meta-analysis, although small, is sufficiently large to reject the possibility of being due to chance. The second is that no difference in effect size can be traced to differences attributed to flaws in the ex-

perimental procedures because the effect size is essentially the same for methodologically strong and weak experiments.

It is on this basis that parapsychologists such as Utts and Radin conclude that the evidence for psi is conclusive. Radin argues that the case for psi has been made conclusive several times in the past and that the only mystery is why this is not better known. He implies that the public is unaware of the scientific case for psi because of the attempts by skeptical 'gatekeepers' to prevent the public from knowing about it.

5. Grounds for Caution

Despite such strong claims by Utts and Radin, the skeptics insist that there are good reasons for doubting that the evidence for psi is adequate. The meta-analysis of the most recent ganzfeld studies by Milton and Wiseman (1999) strongly suggested that the previous ganzfeld successes cannot be replicated. Worse, Milton and Wiseman found that most of the claims for correlates of psi performance also did not replicate. Although their meta-analysis is controversial, the controversy itself points to major problems with trying to base the proof of psi on meta-analysis of previously published studies. Critics have pointed out many problems with meta-analyses.

When used properly, meta-analysis can be a useful tool to help generate patterns and hypotheses about a body of data. Such a usage is consistent with what statisticians call exploratory analysis. The mistake the parapsychologists seem to make is to use the same meta-analysis both to find possible patterns and then prove the existence of such patterns. This is like having your cake and eating it too. A meta-analysis is inherently retrospective. The results are best used prospectively to make predictions about specific patterns to be found in new sets of data. So far, parapsychologists have not succeeded in doing this.

6. Prognosis

Although some key parapsychologists argue that their evidence strongly supports the reality of psychic phenomena, the main body of the scientific community remains unconvinced. At best the parapsychologists have demonstrated, by pooling results from previously published studies, that significant departures from the chance baseline occur more often than expected. They have yet to demonstrate, however, that these departures—if real—originate from a common cause. Nor have they demonstrated any lawfulness in these deviations. Worse, they have yet to come close to specifying a set of conditions and procedures that would enable believers and skeptics to reliably observe these so-called psi events. During the 15 years of conducting meta-analyses on parapsychological ex-

periments, the parapsychologists claim that effect sizes have remained relatively stable. However, this cannot be taken as support for the existence of psi. In a progressive science, as the research develops and the sources of error are tamed, the effect sizes should appreciably grow larger with time. Until this happens in parapsychology, the case for psi will make little headway in the scientific community.

See also: Meta-analysis: Overview; Norms in Science; Physicalism and Alternatives

Bibliography

- Bem D J 1994 Response to Hyman. *Psychological Bulletin* **115**: 25–7
- Bem D J, Honorton C 1994 Does psi exist? Replicable evidence for an anomalous process of information transfer. *Psychological Bulletin* **115**: 4–18
- Honorton C 1985 Meta-analysis of psi ganzfeld research: A response to Hyman. *The Journal of Parapsychology* **49**: 51–91
- Honorton C, Berger R E, Vargolis M P, Quant M, Derr P, Schecter E I, Ferrari D C 1990 PSI communication in the ganzfeld experiments with an automated testing system and a comparison with meta-analysis of earlier studies. *The Journal of Parapsychology* **54**: 99–139
- Hyman R 1985a A critical historical overview of parapsychology. In: Kurtz P (ed.) *A Skeptic's Handbook of Parapsychology*. Prometheus Books, Buffalo, NY, pp. 3–96
- Hyman R 1985b The ganzfeld-psi experiment: A critical appraisal. *The Journal of Parapsychology* **49**: 3–49
- Hyman R 1994 Anomaly or artifact? Comments on Bem and Honorton. *Psychological Bulletin* **115**: 19–24
- Hyman R, Honorton C 1986 A joint communiqué: The psi ganzfeld controversy. *The Journal of Parapsychology* **50**: 351–64
- Mauskopf S H, McVaugh M R 1980 *The Elusive Science: Origins of Experimental Psychical Research*. Johns Hopkins University Press, Baltimore, MD
- Milton J, Wiseman R 1999 Does psi exist? Lack of replication of an anomalous process of information transfer. *Psychological Bulletin* **125**: 387–91
- Radin D 1997 *The Conscious Universe: The Scientific Truth of Psychic Phenomena*. HarperEdge, San Francisco
- Utts J 1991 Replication and meta-analysis in parapsychology. *Statistical Science* **6**: 363–403

R. Hyman

Parenthood and Adult Psychological Developments

Extensive debate since the 1950s on the question of whether and how becoming a parent alters one's life has led to three different but not mutually exclusive answers: In terms of individual and marital functioning, the transition to parenthood produces (a) pri-

marily negative changes, (b) few significant negative or positive changes, and (c) an impetus toward psychological development. We conclude that there is truth in all three alternatives.

1. Framing the Questions: 1950–80

Erik Erikson (1959) claimed that people inevitably experience a period of disequilibrium, crisis, and intrapsychic conflict when faced with new and difficult developmental tasks (see *Erikson, Erik Homburger (1902–94)*). Crises are necessary for developmental growth. If they are not resolved, fixations, regressions, and other forms of psychological dysfunction may result. Thus, normative life transitions represent both danger and opportunity.

LeMasters (1957) interviewed 57 married couples who had become parents in the previous five years. He came to what was then a startling conclusion: 83 percent of the couples had experienced an extensive or severe marital crisis within the first year after giving birth. In response, Hobbs and his colleagues conducted a series of studies (e.g., Hobbs and Cole 1976: 728) and argued that ‘initiating parenthood may be slightly difficult, but not sufficiently difficult to warrant calling it a crisis.’

There were four main problems with virtually all studies of the transition to parenthood in the 1960s and 1970s: (a) They contained little description of what actually happens as partners became parents; (b) crisis was always described in negative terms and measured in arbitrary and inadequate ways; (c) couples not having babies were not included as a comparison group; (d) all the studies were retrospective and failed to measure change by following couples from before to after they had become parents.

2. Becoming a Family as a Major Life Transition: Research from 1980s to 2000

2.1 What Happens to Men, Women, and Marriage as Partners become Parents?

Longitudinal prospective studies designed to assess the impact of becoming parents emerged in the 1980s. Studies by Belsky and his colleagues (e.g., Belsky et al. 1989) and the authors of this article (e.g., Cowan et al. 1985) were guided by multidomain models of the family as a system (see also Heinicke 1984). Consistency over time and change in average level of function were examined in five family domains: each parent’s (a) self-concept and sense of well-being; (b) quality of relationship with their parents; (c) quality of relationship with the baby; (d) marital quality; and (e) balance of life stress and social support. Although some measures of adult functioning and life stress remain stable from late pregnancy through the first year or two of parenthood, various studies find both

quantitative and qualitative change in each of these five family domains.

New parents take on new identities and experience shifts in their self-perceptions and their role in the world. Whether or not old intrapsychic conflicts re-emerge, new parents find themselves renegotiating relationships with their parents. And now, two new family relationships take center stage—between mother and child and father and child. Friendship networks and work arrangements change for both men and women. There is consistent evidence of significant qualitative and quantitative changes in the marriages of heterosexual couples. (Longitudinal studies of gay and lesbian couples making the transition to parenthood have yet to be published.) The division of family labor becomes more traditional, with women doing more housework than they had before, and far more of the care of the children than either partner expected. Levels of marital conflict increase over time, especially about ‘who does what?’ (Cowan and Cowan 2000). Not surprisingly, then, satisfaction with marriage declines over the early childrearing years.

2.2 Is Becoming a Parent a Major Life Transition?

We describe transitions as longterm processes that result in a qualitative reorganization of both inner life and external behavior. A major life transition involves a qualitative shift from the inside looking out (how individuals understand and feel about themselves and the world) and from the outside looking in (reorganization of the individual’s or family’s personal competence, role arrangements, and relationships with significant others). Because systematic longitudinal studies find both quantitative and qualitative shifts in partners’ sense of self and relationships—within the family and with friends and co-workers—we have no hesitation in concluding that becoming a family constitutes a major life transition.

3. Does the Transition to Parenthood Elevate the Risk for Individual or Marital Distress?

3.1 Negative Change over Time

Is the disequilibrium experienced by couples becoming parents accompanied by increased levels of individual and marital distress? The evidence supports a qualified affirmative answer. For example, women are at increased risk for depression during the transition to parenthood (e.g., Campbell et al. 1992). Approximately 10 percent of women develop clinical depression serious enough to interfere with their daily functioning in the postpartum period. There are no epidemiological data documenting the incidence of postpartum depression or other psychological distress for new fathers.

Of more than 20 longitudinal studies conducted in various parts of the USA, Germany, the UK, and Israel, only three failed to find a statistically significant decline in marital satisfaction after the birth of a first child. In two of the three studies, the couples having babies had married recently, and none of the three included first-time parents in long-term marriages. While our own study included a control group of couples not having a baby and found smaller amounts of negative change in the no-baby couples over two years, after five years the childless couples had a much higher divorce rate. Thus, in the short run, the childless couples were doing well, but over time their relationships were vulnerable, often to issues concerning whether and when to have children. Although simply adding up studies with findings supporting one side or another has drawbacks, we feel relatively assured in concluding that there is a statistically significant risk that many marriages will suffer increases in conflict and declines in satisfaction in the early years of parenthood.

It is important to note that (a) these studies report group trends, (b) not all couples show a decline in marital satisfaction (about 20 percent do not), and (c) the average decline over the first year or two of parenthood is quite small. Nevertheless, because marital satisfaction tends to decline further over time, even small increases in marital conflict or dissatisfaction during the family-making period may be important markers of potential family distress.

3.2 Consistency over Time

Despite significant shifts in average level of functioning from before to after the birth of a first child in the five domains of family life assessed in studies cited here, there is remarkable consistency over time in rank order of well-being or distress in new parents. A baby's arrival is unlikely to destroy well-functioning marriages or generate closer, more satisfying relationships between already troubled partners. The high level of cross-time consistency means that even if the transition to parenthood does not elevate the risk of marital distress, when men, women, and marriage are not doing well before the baby is born, they are likely to be distressed in the early family-making years. In terms of prevention planning, these findings indicate that it is possible to identify individuals and couples who are at heightened risk for later distress and dysfunctional behavior patterns on the basis of assessments made before their babies are born.

3.3 Single Parents

Surprisingly, the transition to parenthood has not been well studied in samples described as high risk by

virtue of single parenthood, poverty, or neighborhood violence. Although preventive interventions have been offered to single mothers by home visitors (cf. Olds et al. 1998), the conception of what is needed in these interventions does not take into account that many of the high risk factors operate as additional or amplifying risks, over and above those generally associated with becoming a parent.

4. Can the Transition to Parenthood Promote Adult Development?

Given some of the negative changes and marital distress reported following the birth of a first child, it may seem surprising to ask whether becoming a parent can promote adult development. We suggest, as Erikson (1959) argued, that it is precisely this kind of stress that ultimately can lead to positive developmental change (see Erikson, Erik Homburger (1902–94); *Adult Development, Psychology of*).

Development in adulthood has sometimes been defined as passing normative family life milestones (e.g., entrance to adulthood, courtship, marriage, familymaking) or entering new developmental stages. However, there are neither invariant milestones that precede parenthood nor a convincing body of theory or research to suggest that there are hierarchically integrated stage-like organizations of cognition, personality, and relationships associated with becoming a family.

Elsewhere we have advanced an alternative view (P. Cowan 1988). At the beginning of the 1960s, Allport's functional theory of personality (1961) provided a description of maturity as a general criterion of personality development in terms of six characteristics: (a) extension of the sense of self; (b) warmth in relating to others; (c) emotional security/self-acceptance; (d) realistic perceptions and skills; (e) self-insight and humor; and (f) the establishment of a unifying philosophy of life. Change in the direction of maturity occurs when a person's level of functioning can be described as having become more differentiated, integrated, and adaptive.

We are not aware of systematic longitudinal studies of parenthood as a developmental phase or stage from any of the theoretical perspectives outlined here. Anecdotal evidence from participants in our studies of several hundred couples suggests that babies do stimulate shifts in their parents in the direction of increased maturity (see P. Cowan 1988). Most new parents describe complex changes in their identities. Others describe their relationships with their baby as having 'opened them up' to deep feelings that challenge them to become more effective problem solvers. We do not know yet whether these shifts occur together in the same person, or which factors lead some new parents to become more mature.

5. Implications for Intervention

We have been disappointed to find that few service providers and researchers have created and systematically evaluated interventions for expectant parents designed to (a) reduce or prevent the negative outcomes associated with this major life transition, and (b) increase parents' ability to maintain strong and productive family relationships. Early on, Shereshefsky and Yarrow (1973) provided individual counseling for some expectant mothers. Clulow (1982) conducted monthly meetings for expectant couples, but not in a controlled study with a no-intervention comparison group. Our own intervention (Cowan and Cowan 1999) based on random assignment to a no-intervention comparison or to groups of couples meeting weekly with mental health professionals for three months prepartum and three months postpartum had initially promising results. It staved off declines in marital satisfaction from 6 to 66 months postpartum, and all intervention couples remained in intact marriages until their children were three, while 15 percent of the no-treatment controls had separated or divorced. Because of increasing evidence that marital function plays an important role in shaping parent-child relationships and affecting children's early development (Cummings and Davies 1994), we believe that preventive interventions for partners becoming parents have the potential to promote more positive trajectories of adult development and family life quality.

6. Directions for Future Research

More research is needed to understand individual differences in pathways through the transition to parenthood and why some couples show increased satisfaction and adaptation while others decline. The question of whether and how the transition to parenthood produces development in adulthood is wide open for further research. Systematic studies of interventions to facilitate individual and couple functioning as partners become parents could provide important information for strengthening new families, at the same time testing hypotheses about the mechanisms involved in producing positive change during this major life transition.

See also: Adult Development, Psychology of; Adulthood: Developmental Tasks and Critical Life Events; Fatherhood; Marriage and the Dual-career Family; Cultural Concerns

Bibliography

- Allport G W 1961 *Pattern and Growth in Personality*. Holt, Rinehart and Winston, New York
Belsky J, Rovine M, Fish M 1989 The developing family system. In: Gunnar M R, Thelen E (eds.) *Systems and Development*. Erlbaum, Hillsdale, NJ, pp. 119–66

- Campbell S B, Cohn J F, Flanagan C, Popper S, Myers T 1992 Course and correlates of postpartum depression during the transition to parenthood. *Development and Psychopathology* 4: 29–47
Caplan G 1964 *Principles of Preventive Psychiatry*. Basic, New York
Clulow C F 1982 *To Have and to Hold: Marriage, the First Baby and Preparing Couples for Parenthood*. Aberdeen University Press, Aberdeen, UK
Cowan C P, Cowan P A 1999 *When Partners Become Parents: The Big Life Change for Couples*. Erlbaum, Mahwah, NJ
Cowan C P, Cowan P A, Heming G, Garrett E, Coysh W S, Curtis-Boles H, Boles A J 1985 Transitions to parenthood: His, hers, and theirs. *Journal of Family Issues* 6: 451–81
Cowan P A 1988 Becoming a father: A time of change, an opportunity for development. In: Bronstein P, Cowan C P (eds.) *Fatherhood Today: Men's Changing Role in the Family*. Wiley, New York, pp. 13–35
Cummings E M, Davies P 1994 *Children and Marital Conflict: The Impact of Family Dispute and Resolution*. Guilford, New York
Erikson E 1959 Identity and the life cycle. *Psychological Issues* 1: 1–171
Heinicke C M 1984 Impact of prebirth parent personality and marital functioning on family development: A framework and suggestions for further study. *Developmental Psychology* 20: 1044–53
Hobbs D, Cole S 1976 Transition to parenthood: A decade replication. *Journal of Marriage and the Family* 38: 723–31
LeMasters E E 1957 Parenthood as crisis. *Marriage and Family Living* 19: 352–5
Olds D et al. 1998 The promise of home visitation: Results of two randomized trials. *Journal of Community Psychology* 26: 5–21
Shereshefsky P M, Yarrow L J 1973 *Psychological Aspects of a First Pregnancy*. Raven, New York

P. A. Cowan and C. P. Cowan

Parenting: Attitudes and Beliefs

Parenting has been investigated extensively by developmental psychologists ever since Freud drew formal attention to its critical significance for children's social, emotional, and intellectual functioning. In studying parenting, researchers have come to believe that parents' attitudes and beliefs are centrally important in directing their treatment of children. Therefore, much effort has been focused on looking at links between how parents think about children and child rearing—both generally and in specific situations—and child outcomes. In this article, global and specific attitudes and beliefs are discussed, with an emphasis on the challenges provided by their measurement.

1. What are Attitudes and Beliefs?

Attitudes include beliefs or opinions, an evaluation of those beliefs, and an intention to act in accord with them (Ajzen and Fishbein 1980). People with a

positive attitude to exercise, for example, may reveal their positive feelings by agreeing with others who endorse the same position, by exercising more themselves, by admiring others who exercise, or by encouraging others to exercise. Thus attitudes provide a link, albeit not a perfect one, between thinking and action.

2. Global Attitudes and Beliefs

Dozens of measures of parental attitudes and beliefs have been devised, as investigators have tried to understand the relation between parents' thoughts and actions and children's resulting development. Holden and Edwards (1989) and Holden (1995) have noted the basic assumptions in this approach as well as the fact that they are sometimes violated. These include the premise that parents have pre-existing attitudes, although this is not always the case; that attitudes do not change over time, although they can certainly be modified by education and experience with children; and that parents are not ambivalent with respect to them, although they often are. When these assumptions are violated, relations between attitudes, parents' behavior, and children's development will be less robust.

2.1 Approaches and Findings

Attitudes are assessed either through interviews or questionnaires, with the latter more frequently employed. The majority of attitude questionnaires have focused on parents' views about different child-rearing practices, with a smaller number addressing views of children (Holden 1995). Questionnaires addressing ideas about child-rearing practices have focused on two major features: the quality of the parent's relationship with the child and the extent to which the parent attempts to control the child's behavior. Questions therefore deal with issues of warmth, acceptance, responsiveness, coldness, rejection, and hostility on the one hand, and permissiveness, firm control, punitiveness, and restrictiveness on the other. Typical statements, with which parents indicate their degree of agreement or disagreement, are: 'Children should always treat adults with respect' and 'I enjoy being with my child.'

Researchers have looked at relations between these attitudes and a variety of features of children's development. Positive attitudes toward warmth and reasonable firm control are moderately predictive of outcomes such as high academic achievement, high self-esteem, and a strong moral orientation, while endorsement of restrictiveness and lack of acceptance relate moderately to negative outcomes (Maccoby and Martin 1983). It should be noted that most of the research has been conducted with Anglo-European middle-class mothers and their children and that it is

not clear the same pattern of relations holds for lower socioeconomic classes, in other cultural contexts, or with fathers (Parke and Buriel 1998).

2.2 Limitations of Questionnaire Measures and Needed Improvements

Although the assessment of parenting attitudes has provided useful information for developmental psychologists, dissatisfaction with modest relations between attitudes and child outcomes has fostered attempts to improve existing approaches (Holden 1995). For example, investigators have begun to show a greater concern with the actual relation between attitudes and behavior. Given the basic assumption that attitudes inform behavior, and that it is parenting behavior that determines children's development, there is, of course, little use in studying attitudes if they do not reliably reveal themselves in parenting actions. Reasonable attitude-behavior links are more likely to be found when items assessing attitudes on questionnaires match the nature of the parent's action, e.g., the response to a question about the importance of firmness in getting children to obey is related to observations of reactions when a child fails to comply with a parental request. As well, links are stronger when many examples are added together to provide an aggregated measure of a particular parenting behavior. The fact that relations between attitudes and behavior are stronger under these two conditions underlines two things: asking about general attitudes may give incomplete information when it comes to specific actions, and parenting behavior varies across situations and is determined by a number of variables so that a variety of situations needs to be assessed in order to gain an accurate impression of a particular parent's attitudes.

Holden and Edwards (1989) have raised other methodological concerns. For example, statements on questionnaires are often couched in the third person on the assumption that this may lead to more valid answers because respondents are not placing themselves in a negative light. On the other hand, this approach produces situations in which respondents are not sure whether they are to answer with respect to what might be ideal attitudes rather than what they know to be their own attitudes. Another problem is that attitude measures are based on self-report data that depends on the willingness of the respondent to be honest or even to have access to or conscious awareness of a particular attitude. Psychologists have become increasingly aware that much behavior is determined by implicit or automatic processes, and a variety of techniques have been developed to study the operation of these processes (Bugental and Goodnow 1998). Clearly, however, asking people to state their position on a given topic is not going to be effective if that position is unavailable to conscious introspection.

3. Parental Thinking in Specific Situations

One response to the problems of measuring global attitudes has been to focus on the thoughts or beliefs parents have when they are interacting with their children in specific situations. Thus, rather than ask about general and abstract beliefs and attitudes, researchers have asked about the content of parents' thinking in a particular context. For example, as noted above, inquiries about global attitudes have focused on parents' beliefs about the best child-rearing practices, e.g., authoritarianism, or the extent to which children should obey their parents. More recent inquiries focus on goals, such as obedience, that parents have in a particular situation. Thus the interest in general attitudes has been translated into an interest in goals at a particular point in time, with the recognition that goals of a parent change as a function of context (although, on average, some people may be more inclined to favor some goals more than others). Hastings and Grusec (1998) have shown that, when a child misbehaves, many parents have obedience as their goal and they are punitive. If all one knew was the extent to which those parents had endorsed a general measure of authoritarianism without knowing their specific goal, prediction of their actual behavior would be less accurate.

The study of particular goals reflects the interest of developmental researchers in attitudes toward child-rearing practices. However, interest in parenting cognitions has expanded to a wider range of content including theories about how children develop, expectations about children's abilities, and feelings of self-efficacy with respect to parenting. In the next three sections three additional areas that have received considerable attention will be described.

3.1 Causal Attributions

Parents try to find explanations for why their children have behaved in a particular way, that is, they make 'causal attributions.' Although people make causal attributions in a variety of contexts, parents are probably particularly likely to make them because they need to understand their children so that they can effectively influence them. Dix and Grusec (1985) have outlined some of the features of parents' thinking in this context. In the search for explanation, parents can make internal or dispositional attributions that find the source of action in the child's personality or character. Alternatively, they can make external attributions that locate the source of action in the external situation or environment. When a negative action is attributed to dispositional factors it is most often seen as intentional and under the child's control. In this case parents have been shown to react punitively, possibly both because of their accompanying anger and because a Western ethical system dictates that intentional misdeeds should be punished. Parents who

make external attributions, believing their child was tired or provoked or did not know any better, are likely to respond in a more benign way, e.g., by reasoning in an attempt to guide their child's future actions. When the attributions are accurate, parental behavior is likely to be appropriate for modifying the child's actions. When the attribution is inaccurate, however, parenting will be ineffective, given that children who lack knowledge, for example, are merely punished and do not learn what correct actions are, or children who have knowledge do not experience the negative consequences of their actions.

3.2 Relative Control

The impact of parental thinking has also been mapped in the domain of perceptions of relative control between parents and children. In her research program, Bugental (Bugental 1992, Bugental and Goodnow 1998) has shown how parents' beliefs about such control are elicited in difficult interactions with their children. Once elicited, these beliefs have an impact on emotional reactions, behavior, and, subsequently, child outcomes. Parents who perceive their children as having more control of the difficult situation than they do experience physiological arousal that reflects feelings of threat aroused by the apparent reversal of power. Because they feel relatively powerless they resort to coercive and overly aggressive behavior. At the same time they transmit confusing and inconsistent messages to the child, both because their cognitive capacity is limited by the distracting thoughts associated with threat and because they cannot adequately inhibit the negative feelings they are experiencing. In turn, the confusing behavior leads to nonresponsiveness on the child's part that serves to reinforce the parents' negative thoughts and affect.

3.3 Attachment Relationships

Early in their development children learn that their needs for security will either be dealt with satisfactorily, rejected, or responded to inconsistently (Ainsworth et al. 1978) (e.g., see *Attachment Theory: Psychological*). These early experiences form the basis of adult working models of relationships that function throughout the life course and that manifest themselves in parenting behaviors. Thus adults who are secure with respect to the satisfaction of their emotional needs have a model of relationships that enables them to respond sensitively and supportively to their children. Adults whose model includes others as rejecting are cool, remote, and task-focused in their interactions with children, while those with models of other as inconsistent are unpredictable and confusing in their parenting actions (Crowell and Feldman 1991).

4. Automatic vs. Conscious Thinking

The earlier discussion of global attitudes noted that some of these attitudes are not open to conscious awareness. The same distinction has been made in the area of parenting cognitions, with some thoughts seen as automatic and implicit and others as more open to conscious reflection and awareness. Bugental has argued, for example, that the experience of a difficult child activates a high threat schema that operates at a preconscious level and places parents in a state of perceptual readiness to react to certain events in the environment. Thus the stress produced by a problematic interaction with a child is filtered through this schema, which in turn influences both emotional feelings and more conscious or deliberate ideation. Mental representations of childhood attachment experiences are also considered to operate outside conscious awareness (Bretherton 1985). Parents' goals and attributions about children's behavior may be more open to conscious awareness and reflection—they frequently change in response to changes in the external situation. On the other hand, some parents are not able to use situational information to change their attributions for child misbehavior (Milner and Foody 1994), an indication that even the same kinds of cognitions may be sometimes more and sometimes less automatic.

5. Future Directions

The use of global parenting attitudes and beliefs and specific cognitions to understand parent-child relationships will continue to be a significant part of the arsenal of developmental researchers. Measurement issues remain of paramount importance. Strong linkages between measured attitudes and behavior still need to be demonstrated. A fundamental question has to do with how to measure attitudes and beliefs that are automatic in nature and, as such, cannot easily be articulated by respondents even if they are perfectly willing to be open and honest in their answers. The move must be away from straightforward transparent questions to more subtle measures where 'correct' or socially acceptable answers are not so self-evident. Bugental's measure of relative control (Bugental 1992) and the Adult Attachment Interview which assesses mental representations of attachment relationships (Main and Goldwyn 1991) are examples of such measures, and they have proved to be accurate predictors of a variety of parenting behaviors and child outcomes. Another important question has to do with the origins of various parenting belief systems. Some presumably are part of the cultural context in which parents find themselves. Others come from experience with one's own children. And others emerge as a result of a parent's own experiences as a child; it is these latter, presumably, that are most automatic and most difficult to alter when alteration seems desirable.

Finally, linkages among the many sets of beliefs that have taken up the attention of researchers need to be made, as well as explorations of the distinctive ways in which each contributes to the parenting process.

See also: Attitudes and Behavior; Family and Gender; Fatherhood

Bibliography

- Ainsworth M D S, Blehar M C, Waters E, Wall S 1978 *Patterns of Attachment: A Psychological Study of the Strange Situation*. Lawrence Erlbaum Associates, Hillsdale, NJ
- Ajzen I, Fishbein M 1980 *Understanding Attitudes and Predicting Social Behavior*. Prentice-Hall, Englewood Cliffs, NJ
- Bretherton I 1985 Attachment theory: Retrospect and prospect. In: Bretherton I, Waters E (eds.) Growing points of attachment theory and research. *Monographs of the Society for Research in Child Development*. 50 Serial No. 209, pp. 3–35
- Bugental D B 1992 Affective and cognitive processes within threat-oriented family systems. In: Siegel I E, McGillicuddy-DeLisi A V, Goodnow J J (eds.) *Parental Belief Systems: The Psychological Consequences for Children*, 2nd edn. Lawrence Erlbaum Associates, Hillsdale, NJ
- Bugental D B, Goodnow J J 1998 Socialization processes. In: Damon W, Eisenberg N (eds.) *Handbook of Child Psychology Vol. 3 Social, Emotional, and Personality Development*. J. Wiley, New York, pp. 389–462
- Crowell J A, Feldman S S 1991 Mothers' working models of attachment relationship and mother and child behavior during separation and reunion. *Developmental Psychology* **27**: 597–605
- Dix T, Grusec J E 1985 Parent attribution processes in child socialization. In: Sigel I (ed.) *Parent Belief Systems: Their Psychological Consequences for Children*. Lawrence Erlbaum Associates, Hillsdale, NJ, pp. 201–33
- Hastings P D, Grusec J E 1998 Parenting goals as organizers of responses to parent-child disagreement. *Developmental Psychology* **34**: 465–79
- Holden G W 1995 Parental attitudes toward childrearing. In: Bornstein M (ed.) *Handbook of Parenting Vol. 3 Status and Social Conditions of Parenting*. Lawrence Erlbaum Associates, Mahwah, NJ, pp. 359–92
- Holden G W, Edwards L A 1989 Parental attitudes toward childrearing: Instruments, issues, and implications. *Psychological Bulletin* **106**: 29–58
- Maccoby E E, Martin J A 1983 Socialization in the context of the family: Parent-child interaction. In: Hetherington E M (ed.) *Handbook of Child Psychology Vol. 4 Socialization, Personality and Social Development*. Wiley, New York, pp. 1–102
- Main M, Goldwyn R 1991 Adult attachment rating and classification systems. Unpublished manuscript, University of California at Berkeley
- Milner J S, Foody R 1994 Impact of mitigating information on attributions for positive and negative children's behavior by adults at low-risk and high-risk or child abusive behavior. *Journal of Social and Clinical Psychology* **13**: 335–51
- Parke R D, Buriel R 1998 Socialization in the family: Ethnic and ecological perspectives. In: Eisenberg N (ed.) *Handbook of Child Psychology Vol. 3 Social, Emotional and Personality Development*. Wiley, New York, pp. 463–552

J. E. Grusec

Copyright © 2001 Elsevier Science Ltd.

11041

All rights reserved.

Parenting in Ethnic Minority Families: United States

'Parenting' refers to the training or raising of children, focusing on parents' responsibility to guide children's development toward behaviors determined as functional in the surrounding culture. Parenting practices are shaped by the contextual circumstances in which the parenting occurs. One such contextual circumstance is ethnic minority status. Ethnic minority status in the USA usually refers to people of color, those with physical characteristics differentiating them from the majority white American. Ethnic minority groups include African Americans, Asian Americans, Mexican Americans, and American Indians. The implications of ethnic minority status have been influenced greatly by the difficulties such designated groups have experienced in attempting to integrate and assimilate into the mainstream of US society.

Past conditions of involuntary immigration and slavery, legal discrimination and segregation, exclusion and restrictions, and forced removal and internment to reservations have influenced the attitudes, perceptions, and behavioral patterns of the respective ethnic minorities. Interestingly, it does not matter whether ethnic minority groups were volunteer or forced immigrants or whether they were an indigenous people, their historical struggles of assimilating into US society was qualitatively different from other noncolor ethnic Americans (Ogbu 1981, Wilson 1989). This entry considers the evolved parenting practices of the four ethnic American groups, focusing on African-American, Chinese-American, Mexican-American, and American Indian practices, illustrating their marked distinction from the dominant culture as well as variations between the ethnic groups.

1. Parenting Behaviors and Minority Families

Historically, past research on parenting has described practices in minority families in terms of their contrast to practices of the majority. In comparisons, dominant cultural norms and values have been used to evaluate the practices of minority parents, deeming them changeworthy. Comparing practices of ethnic minorities with those of the majority remains an important topic given the stigma and social implications such comparisons elicit for ethnic minorities in the USA. Closer examination of parenting in each minority group suggests a link between the differing historical experiences of the groups and their evolved parenting practices. Analyses of adolescent adjustment indicate the adaptive nature and functionality of parenting among ethnic minorities.

Two basic parental behaviors, warmth and control, presume to underlie most categories of parenting practices. Baumrind (1991) formed four patterns of

parental authority based on differing levels of parental warmth and control. The patterns emerging from Baumrind's studies were authoritarianism, authoritativeness, permissiveness, and neglectful/rejecting. Authoritarian parents displayed high levels of parental control but low levels of warmth. Authoritative parents were high in both parental warmth and control. Permissive parents were highly responsive (i.e., warm) to their children, but did not place many demands (i.e., control) on them (Baumrind 1991). Neglectful/rejecting parents were low in both parental control and warmth.

Because authoritarian, authoritative, and permissive parenting represent a central concern of this entry, additional comments are warranted. Specifically, authoritarian parents are described as using absolute standards in order to shape and control the attitudes and behaviors of their children. They place high regard on such values as respect for authority, hard work, and maintenance of order; simply put, the parents' word is not to be questioned. Authoritative parents are best described as applying firm parental authority while engaging in and encouraging parent-child interaction with respect to socialization. Further, authoritative parents address their child's affective needs. Permissive parents do not maintain and enforce clear behavioral expectations but are highly attentive to the child's affective needs (Baumrind 1991).

Since parenting practices are assumed to be flexible and contextually sensitive to situational demands, comparative claims will not be offered regarding the advantages of any parenting style. Parents presumably incorporate their own beliefs about how the world works into their child-rearing practices. Such beliefs, developed by parental experiences, naturally have some bearing on what parents believe their children are likely to face in their own lives. Hence, a basis of understanding parenting practices is the notion that many contextual differences will shape the child-rearing practices. Each parenting style is manifested in attitudinal and behavioral attributes that vary both between different ethnic groups and within ethnic groups (see *Cultural Variations in Interpersonal Relationships*).

2. Socialization in African-American Families

Wilson (1989) asserts that an important social context of African-American family development involves the involuntary migration to the Americas followed by a protracted period of African-American enslavement and, after emancipation, segregation and inferior status base on race and struggle for civil rights and political freedom. The African-American family evolved into an extensive familial structure that was based on mutual support and obligation.

Two types of studies have explored the child-rearing practices common to African-Americans. Compara-

tive studies have focused on the differences between African-American child-rearing practices and those common to other ethnic groups (Rickel et al. 1988, Baumrind 1991, Bartz and Levine 1978), whereas within-group approaches focus on the variance in child-rearing practices among African-Americans only (Hill 1995, Kelley et al. 1993, Luster and McAdoo 1994). Much of the comparative research on parenting describes African-American parenting in terms of its relation to the dominant group's child-rearing practices and ideals (e.g., Baumrind 1991). In contrast, the within-group research has described the diversity of attributes and behaviors within African-American families (see *Socialization in Adolescence*).

Comparative studies on socialization have generally found African-American parenting styles to be authoritarian and more restrictive than European-American styles (Baumrind 1991, Rickel et al. 1988, Kelley et al. 1993). Also, African-American parents appear to employ gender-specific socialization practices. Specifically, African-American boys are more likely to be expected to act maturely and responsibly than are European-American boys. Parents of African-American girls have been shown to exert high levels of control. African-American fathers discourage infantile behaviors in their daughters. African-American mothers display firm enforcement and rejecting behaviors, and discourage nonconformity in their daughters (see *Poverty and Child Development*).

Bartz and Levine (1978) examined specific parenting practices in low-income African-American, European-American, and Mexican-American families. African-American parents exercised more control over their child's behavior, were more likely to press for acceleration of development, were more concerned about their children wasting time, and used supportive behaviors more frequently than European-American and Mexican-American parents. African-American parents also reported doing more to enforce their rules and expectations. Generally, African-American parents value strictness and expect their children to gain a sense of responsibility. They encouraged decision making, expressed loving concern, and closely monitored their children's behavior to assure goals such as obedience and high achievement (Bartz and Levine 1978).

Within-group ethnic studies generally focus on the specific characteristics that influence the development of a particular parenting style (Kelley et al. 1993). Generally, it has been found that mothers with a higher family income and a greater number of children report using more parental warmth and firm discipline and lower levels of control. In addition, discipline has been shown to correlate negatively to the number of adults in a family (Wilson 1989).

Inasmuch as authoritarian practices appear to be common among African-American families, there are important familial caveats that are related to familial socioeconomic levels, maternal resources, and neigh-

borhood. Namely, older, educated, middle-class, married mothers employed more authoritative practices than do young, single mothers with less education. Conversely, young, single, less educated mothers were more likely to emphasize obedience, use physical punishment, and employ a more parent-oriented approach (Kelley et al. 1993). Also, mothers who perceived their neighborhoods as dangerous were likely to use harsh discipline and authoritarian practices.

In addition, Luster and McAdoo (1994) found that adolescents with the highest level of cognitive competence came from homes with older, more educated and high esteem mothers, a supportive environment, and a small family size. Thus, there is variance of socialization practices among African-American families.

3. *Socialization in Chinese-American Families*

Research on Chinese-American parenting practices is often described in ways similar to the characteristics of the authoritarian style. However, the child-rearing practices may be the outgrowth of the Chinese philosophy of hsaio, or filial piety, which prescribes Chinese children to fulfill the desires of their parents (Lin and Liu 1993, Kelley and Tseng 1992, Chiu 1987). Devotion to one's parents takes precedence over other relationships, including one's obligation to one's own children (Lin and Liu 1993). Parents are very involved in the training of their children, and hold a definite position of authority over them. Strict discipline is highly valued. Parental control, therefore, is apparent as children learn that their parents are 'always right' (Chiu 1987, Kelley and Tseng 1992, Lin and Liu 1993).

Kelley and Tseng (1992) examined the child-rearing practices of Chinese-American and European-American mothers. Specifically, they measured use of external threats and punishment, permissiveness, and emphasis on school-related skills, obedience, concern for others, and self-reliance. Chinese-American mothers scored higher than did European-American mothers on ratings of restrictiveness and control (Chiu 1987, Kelley and Tseng 1992).

4. *Socialization in Mexican-American Families*

Mexican Americans make up the vast majority of Latino peoples living in the USA. Other groups include Central and South Americans, Puerto Ricans and Cuban Americans. Mexican Americans have been incorporated in US society through military conquest of the Southwest and ambivalent immigration policies that at some times encourage immigration and at other times deports citizens of Mexican decent. Bartz and Levine (1978) found Mexican-American parents to exhibit more authoritative behaviors than African-

American parents but less authoritative behaviors than European-American parents. Specifically, Mexican-American parents' reported levels of firm control and warmth were intermediate between those reported by African-American and European-American parents. In examination of Mexican-American and European-American mothers from low and high socioeconomic status (SES), Cardona et al. (2000) indicated that Mexican-American and higher SES mothers reported using higher levels of discipline and lower levels of nurturing behaviors than did European-American and lower SES mothers. Moreover, higher SES Mexican-American mothers reported more frequent use of discipline than did other mothers. Further, Buriel (1997) found generational patterns in the Mexican-American parents of first-, second- and third-generational children. Parents of first- and second-generation children reported authoritative-oriented practices whereas the parents of third-generation children reported authoritarian practices. In addition parents of first- and second-generation children reported higher level of parental permissiveness than did the parents of third generation children. Taken together, it would appear that as Mexican-American parents become more acculturated and integrated into US society, they resort to more authoritarian patterns of parenting.

5. Socialization in American Indian Families

Although American Indians are a diverse group making up 450 distinct groups, one common experience shared by the diverse groups is their history of military conflicts with white Americans and forced removal from the ancestral homelands and placement on reservation. American Indian customs place great value on collectivism, natural harmony, humility, and respect for and importance of family. Because the family network is an important part of American Indian life, child rearing is a shared activity among the network. Through valued collectivism and the family support network, the growing child is placed in the protected environment of close relationship with familial kin (Forehand and Kotchick 1996). Several researchers have indicated that specific parenting practices among American Indians appear permissive in that American Indians report using low levels of control or punishment and high levels of warmth (Cross 1986, Red Horse 1997).

6. Socialization of Children and Adolescents in the Context of Ethnic Differences

Although research maintains that authoritative parenting produces the most optimal child outcomes (Baumrind 1991), numerous studies suggest that the actual positive effects of authoritative practices are correlated with some specific domains of development

only (Steinberg et al. 1995) and are generally more relevant to European-American adolescents than ethnic minority adolescents (Steinberg et al. 1995, Darling and Steinberg 1993) (see *Parenting: Attitudes and Beliefs*).

The differential cultural and contextual experiences of ethnic minority groups are likely to provide an explanation for the differing effects of the authoritarian style on child outcome between ethnic minorities and the majority group. Although authoritarian parenting practices among European-American parents may be viewed as excessively harsh and punitive, and may be indicative of parental rejection and lower level of investment, similar practices among African-Americans may be interpreted as a sign of parental involvement and concern (McLeod et al. 1994). African-American mothers report using authoritarian measures to teach their children about harsh realities in a racist society (McLeod et al. 1994, Hill 1995), for teaching values such as respect, child safety, and school achievement (Kelley et al. 1993), and as a protective measure in high-risk environments (Rickel et al. 1988).

Chao (1994) suggests that Chinese-American parents' use of authoritarian and restrictive practices are more indicative of parental concern, involvement, and caring than parental hostility and aggression. Authoritarian practices are part and parcel of successful and harmonious family and community life (see *Cultural Psychology*).

7. Concluding Remarks and Future Directions

Several implications of the review on minority families' socialization practices are evident. First, although the concept of authoritarianism is generally analogous to parental hostility, aggression, and dominance, the common occurrence of authoritarian and restrictive parenting practices in African-American, Chinese-American, and Mexican-American populations are not necessarily associated with such negative parental attributes. Rather, the use of authoritarian procedures has demonstrated functional benefits for ethnic minority Americans. Second, although common among several of the ethnic minority American groups, the authoritarian and restrictive practices evolved from different contextual concerns. For African Americans, the use of authoritarian practices were developed and maintained in response to harsh environmental realities like low-income, inner-city living environments, and hostile discriminatory practices of US society. For Chinese Americans, traditional cultural beliefs and values of filial piety initiated the practices. For Mexican Americans, value of the strong patriarchal presence in families appears to increase the use of authoritarian practices. Interestingly, American Indian socialization practices are typically permissive, while close-kin networks maintain protectiveness. Last, different parenting practices do have different

meanings, uses, and benefits for minority and majority children (Chiu 1987, Ogbu 1981). Contending that one method embodies positive outcomes for one group does not imply that another group would or would not benefit from that practice. Parenting practices must be examined carefully from the perspective of the people who employ them and the contexts in which those practices will be employed. In order to understand ethnic minority socialization, it is important to examine applied socialization practices within their particular ecological context (see *Socialization, Sociology of*).

It is clear that the US population will continue its racial and cultural diversification. Likewise it is acceded that a diverse US population will display different kinds of socialization practices. Contextual aspects of socialization will continue to play a fundamental role in understanding and determining adaptive socialization processes. Thus, socialization practices must be appreciated for their particular situational meaning, significance, and consequence. Rather than presume peremptory negative and positive connotations of authoritative and authoritarian parenting practices, the merits of each style should be judged according to its contextual contingencies and influence on adolescent adjustment. Also, it is important that we view socialization practices from a cultural perspective. An understanding of the interacting mechanism of culture, family, and socialization may lead to an appreciation of the familial complexities experienced by various ethnic groups. Furthermore, understanding the links between ethnic minority status, socialization practices, and child outcomes may foster appropriately directed family interventions.

See also: Ethnic Identity, Psychology of; Family and Gender; Fatherhood; Gender, Class, Race, and Ethnicity, Social Construction of; Minorities; Parenting: Attitudes and Beliefs

Bibliography

- Bartz K, Levine E 1978 Child-rearing by black parents: A description and comparison to Anglo and Chicano parents. *Journal of Marriage and the Family* **40**: 709–19
- Baumrind D 1991 Effective parenting during the early adolescent transition. In: Cowan P A, Hetherington M (eds.) *Family Transitions*. Lawrence Erlbaum, Hillsdale, NJ, pp. 111–64
- Buriel R 1997 Childrearing orientations in Mexican American families: The influence of generation and sociocultural factors. *Journal of Marriage and the Family* **55**(4): 987–1000
- Cardona P G, Nicholson B C, Fox R A 2000 Parenting among Hispanic and Anglo American mothers with young children. *Journal of Social Psychology* **140**(3): 357–65
- Chao R K 1994 Beyond parental control and authoritarian parenting style: Understanding Chinese parenting through the cultural notion of training. *Child Development* **65**: 1111–19

- Chiu L H 1987 Child-rearing attitudes of Chinese, Chinese-American, and Anglo-American mothers. *International Journal of Psychology* **22**: 409–19
- Cross T 1986 Drawing on cultural tradition in Indian child welfare practice. *Social Casework* **65**(5): 283–9
- Darling N, Steinberg L 1993 Parenting style as a context: An integrative model. *Psychological Bulletin* **113**(3): 487–96
- Forehand R, Kotchick B 1996 Cultural diversity: A wake-up call for parent training. *Behavior Therapy* **21**: 187–206
- Hill N 1995 The relationship between family environment and parenting style: A preliminary study of African American families. *Journal of Black Psychology* **21**(4): 408–23
- Kelley M R, Sanchez-Hucles J, Walker R R 1993 Correlates of disciplinary practices in working- to middle-class African-American mothers. *Merrill-Palmer Quarterly* **39**(2): 252–64
- Kelley M L, Tseng H M 1992 Cultural differences in child-rearing: A comparison of immigrant Chinese and Caucasian American mothers. *Journal of Cross-Cultural Psychology* **23**(4): 444–55
- Lin C, Liu W 1993 Intergenerational relationships among Chinese immigrant families from Taiwan. In: McAdoo H P (ed.) *Family Ethnicity: Strength in Diversity*. Sage, Newbury Park, CA, pp. 271–86
- Luster T, McAdoo H P 1994 Factors related to the achievement and adjustment of young African-American children. *Child Development* **65**: 1080–94
- McLeod J, Kruttschnitt C, Dornfield M 1994 Does parenting explain the effects of structural conditions on children's antisocial behavior? A comparison of Blacks and Whites. *Social Forces* **73**(2): 575–604
- Ogbu J 1981 Origins of human competence: A cultural-ecological perspective. *Child Development* **52**: 413–29
- Red Horse J 1997 Traditional American Indian family systems. *Family Systems Medicine* **15**(3): 243–50
- Rickel A U, Williams D L, Loigman G A 1988 Predictors of maternal child-rearing practices: Implications for intervention. *Journal of Community Psychology* **16**: 32–40
- Steinberg L, Darling N, Fletcher A 1995 Authoritative parenting and adolescent adjustment: An ecological journey. In: Moen P, Elder G H, Luscher K (eds.) *Examining Lives in Context: Perspectives on the Ecology of Human Development*. American Psychological Association, Washington, DC, pp. 423–66
- Wilson M 1989 Child development in the context of the black extended family. *American Psychologist* **44**(2): 380–5

M. N. Wilson and M. D'Amico

Parents and Teachers as Partners in Education

Parents have always been involved in children's preparation for adulthood, but during the twentieth century many responsibilities were taken over by schools. In Western societies a model of teachers and parents as 'partners in education' gradually evolved which currently enjoys widespread academic and political support. In the contemporary climate of globalization it may well be presented as the ideal for

all societies with mass education systems. Yet critics claim it is difficult to implement at best, and may easily serve further to disempower socially disadvantaged sectors of the population.

1. *Evolution of the Partnership Model*

In any society, economic, sociocultural, and political conditions, as well as pedagogical theory, help define the preferred model of home-school relations.

1.1 *Historical Shifts*

During the early decades of mass schooling in Western societies, as teaching methods became more specialized and when teachers were among the most highly educated members of their communities, parents were viewed as ill-equipped to understand what happened in classrooms and liable to do more harm than good if they interfered. In the more widespread affluence of the post-World War II period, parents came under much greater pressure to assist by providing a materially enriched home environment, but were only expected to become significantly engaged in discussion of their child's education if schools identified problems. Parents were not regularly invited beyond the school gate until the 1970s; opportunities were then limited initially to formal discussions about children's progress, or provision of unskilled assistance to teachers, but gradually expanded to include more responsible duties in classrooms and more freedom to engage informally with staff. By the late 1980s research proclaimed major parental involvement *as educators* to be essential to children's success in school, and parents themselves were demanding more say in educational decision-making.

1.2 *Contemporary Partnership Programs*

The language used to strategize home-school connections has reflected these shifts. Current policies espousing 'partnership' have implications that distinguish them in important ways from those based primarily on parental 'involvement' or 'consultation.' The partnership paradigm requires consensus between teachers and parents on the goals and processes of education, and implies a full sharing of information, skills, decision-making, and accountability.

Early partnership programs were typically the initiatives of individual school principals, but their establishment is now commonly assisted by research-based and/or organizational support. In a large-scale project in New Zealand, for example, experienced senior teachers were imported into schools as 'developers' to support and legitimate innovative practice and constructively challenge institutional resistance to

collaborative ways of working (Ramsay et al. 1993). In the USA, the Center for Social Organization of Schools has theorized an extended partnership framework including other family members and the wider community (Epstein et al. 1997). Implementation involves exploring opportunities for six types of engagement: *parenting* (assisting families with childrearing and establishing home environments conducive to learning); *communicating* (developing channels for reporting student progress and school news); *volunteering* (improving involvement of families as volunteers and audiences to support students); *learning at home* (assisting all parents to help children with homework and other curriculum-linked activities); *decision-making* (including families in school governance as well as Parent-Teacher Associations and other committees); *collaborating with the community* (linking students and families with other community agencies, and providing services to the community). This model promotes 'school-like families' and 'family-like schools': parents are expected to reinforce the importance of school, and schools to recognize each child's individuality.

It is recognized that program implementation is often problematic within educational systems that have traditionally operated in a very hierarchical manner. Research suggests that projects are relatively easily established when children are most dependent (of preschool age or with special educational needs) but still difficult to operate successfully in high schools, even though it is now strongly argued that parental participation remains crucial at this level. It is widely reported that training programs continue to pay insufficient attention to preparing teachers and administrators to work in new ways with families (including dealing with attendant challenges to their professional status), and that even the best-planned initiatives rarely fulfill all criteria of partnership or manage to attract all families.

2. *Critique of the Partnership Model*

Despite acknowledged implementation difficulties, the literature nonetheless generally supports the view that 'partnership' remains the ideal to be pursued. However, some repeatedly articulated misgivings critique this premise.

2.1 *Research Support*

Some writers contend that research has *not* provided compelling support for the core claim that the impact of parental involvement is positive, that too much of the corroborative evidence comes from studies of self-selected samples of schools and parents, and that assumptions about the *direction* of causal connections between involvement and children's success have been

insufficiently interrogated. It is also argued that researchers have failed to investigate seriously some of the potentially *negative* consequences, for example, the impact of highly engaged parents whose motivation may be rooted in an antischool stance, the 'dark side' to the intense involvement of some ambitious or 'pushy' middle-class parents, and the risks associated with seeing increased parental participation as the universal panacea for schools' pedagogical or financial problems. Possibly there has been a 'conspiracy' to avoid examining the new buzzword of partnership too closely in case it fell apart (Bastiani 1993).

2.2 Partnerships and Power Relations

Whereas partnership protagonists such as Epstein argue that parents are more concerned about 'information,' 'communication,' and 'participation' than about 'power' or 'control,' others claim the language of collaboration and partnership helps to silence the naming of important power asymmetries in home-school relations. 'Participation' may therefore mean that parents play a more active role, but are given no genuine opportunity to challenge the school's authority to define the nature of that role (Fine 1993). It may deflect attention from gender disparities, such as mothers' under-representation in positions of governance or schools' reported reluctance to utilize fathers in classroom support roles.

Furthermore, parents and teachers do not necessarily mean the same thing when they use terms such as 'helping children succeed.' Discussing the enduring documentation of socioeconomic class (SES) differences in parental involvement, Lareau (Lareau 1989, Lareau and Horvat 1999) suggests that these differences have several roots, including the fact that whereas middle-class parents typically embrace opportunities to scrutinize and question school practices, low-SES parents often prefer a relationship characterized by *separateness*, expecting teachers to do their job and seeing themselves as helping best by deferring to professional expertise. Her analysis also incorporates Bourdieu's (1990) notion of *cultural capital*: high-SES children enter schools with more cultural capital potentially available for activation because schools tend to use the linguistic structures, authority patterns, and types of curricula with which they are more familiar. There may be negative consequences for parents' dignity and authority in the home if they cannot (or prefer not to) perform their designated roles as educators in accordance with these terms (see *School Achievement: Cognitive and Motivational Determinants*). Low-SES parents may feel less confident or comfortable visiting schools, and may have more reason than high-SES parents to fear that closer links with schools could result in greater family surveillance by government agencies. Middle-class domination of that aspect of 'partnership' quintessentially disting-

uishing it from other models—involvement of parents in all aspects of school governance—may function to sustain rather than challenge societal inequalities. While some projects work hard to minimize such impediments, or avoid them through being established in more economically homogeneous communities, the generic Western model arguably privileges the ways in which *well-educated* parents wish to, and can, engage with schools.

Most of these concerns have also been addressed with reference to multiculturalism, neocolonialism, and racism (e.g., Ramsay et al. 1993, Limerick and Nielsen 1994, Lareau and Horvat 1999) (see *Minorities; Multicultural Education*). Whether or not imbalances of power or influence fundamentally violate the spirit of partnership remains a key feature of current debate.

2.3 Political Intervention

In Western societies, partnership models enjoy widespread political support, provoking debate about the extent to which schools should be legally mandated to develop partnership programs and parents required to participate. Legislation might be the only way to reduce significantly existing societal inequalities *or* it could equally well serve to consolidate or amplify them. Partnership initiatives may be compromised by the effects of other political interventions in education, such as imposition of a 'market model' that positions parents as selfish 'consumers' of school services. Moreover, no less than other paradigms, the partnership model should be recognized as offering national or local governments opportunities to absolve themselves from the responsibilities of maintaining struggling public school systems (Fine 1993). While the market model seeks to lay blame primarily at the door of schools, the 'family-is-school' approach could arguably enable school failure to be more easily read as *family* failure.

3. Future Trends

To some extent, future developments will be driven by research findings that resolve the implementation problems outlined above. They will also be influenced by how communities prioritize their objectives. While the pre-eminent objective may continue to be enhancement of parents' ability to promote their own children's school success (however defined), some communities or cultures may prefer a more collective approach that encourages all adults to work on behalf of all children. Other programs may focus less exclusively on children, and place more emphasis on helping parents improve their quality of life and on the broader rejuvenation of community spirit. If educational resources are limited, families may play a critical role in the development of curriculum

materials; in nations where the legacy of colonialism still positions schools as somewhat alien institutions, schemes may be focused on integrating schools more fully into the society. Whatever the objectives, a partnership ideology suggests that programs are most likely to succeed when goals are clearly articulated and informed consensus has been reached regarding the strategies employed to achieve them. As always, however, developments will be directly or indirectly influenced by external factors such as funding support and other aspects of government policy.

'Partnership' is clearly the current model of choice for white Western middle-class communities, and has many features with the potential to inform good educational practice in a wide variety of cultural settings. Nevertheless, certain conditions must be met in order that partnership programs offer all families an equitable chance of reaping the rewards of participation. Furthermore, inasmuch as technological advances may mean that children in future spend much less time either physically located in schools or pedagogically locked into the programs of individual education systems, current ideas about 'consumers,' 'providers,' and 'partnerships' may possibly soon require radical re-examination.

See also: Families as Educational Settings; Family and Schooling; Family as Institution; Family Systems in Europe; Parenting: Attitudes and Beliefs; School Effectiveness Research; School Management; Teacher Expertise; Teaching as a Profession

Bibliography

- Bastiani J 1993 Parents as partners: Genuine progress or empty rhetoric? In: Munn P (ed.) *Parents and Schools: Customers, Managers or Partners?* Routledge, London, pp. 101–16
- Berger E H 1987 *Parents as Partners in Education: The School and Home Working Together*, 2nd edn. Merrill, Columbus, OH
- Bourdieu P 1990 *In Other Words: Essays Towards a Reflexive Sociology*. Polity Press, Cambridge, UK
- Epstein J L, Coates L, Salinas K C, Sanders M G, Simon B S 1997 *School, Family, and Community Partnerships: Your Handbook for Action*. Corwin Press, Thousand Oaks, CA
- Fine M 1993 [Ap]parent involvement: Reflections on parents, power, and urban public schools. *Teachers College Record* **94**: 682–710
- Lareau A 1989 *Home Advantage: Social Class and Parental Intervention in Elementary Education*. Falmer Press, London
- Lareau A, Horvat E McN 1999 Moments of social inclusion and exclusion: Race, class, and cultural capital in family–school relationships. *Sociology of Education* **72**: 37–53
- Limerick B, Nielsen H (eds.) 1994 *School and Community Relations: Participation, Policy and Practice*. Harcourt Brace, Sydney, Australia
- Ramsay P, Hawk K, Harold B, Marriott R, Poskitt J 1993 *Developing Partnerships: Collaboration Between Teachers and Parents*. Learning Media, Ministry of Education, Wellington, New Zealand

M. A. Payne

Pareto, Vilfredo (1848–1923)

1. Life

Fritz Wilfried Pareto was born on July 15, 1848 in Paris, France, the son of an Italian political émigré, Raffaele Pareto (1812–82). Raffaele's father, Raffaele Benedetto Pareto (1768–1831), had been made a marchese di Parigi—or, as some have claimed, even a pair de France—by Napoleon, in recognition of his collaboration in erecting the Cisalpine Republic. The Pareto family was rooted in the trade bourgeoisie of Genoa, Italy, where it had been registered in the *Libro d'oro*.

In addition to her son, whose given name was Italianized and shortened to Vilfredo, Pareto's mother, Marie Métenier (1816–89), also had two daughters. Due to a political amnesty granted by the House of Savoy, probably as early as 1858, the Pareto family was eventually able to move back to Italy. Pareto's father started to work as a French instructor for Genoa's *Scuola di Marina*. Vilfredo attended the same school, the beginning of his Italian education. When, later, Pareto's father was offered a better position in Turin, Vilfredo followed him to the capital of the kingdom of Savoy. Here, in 1867, Vilfredo received his *Licenza in scienze matematiche e fisiche*.

He then enrolled at the *Scuola di Applicazione per Ingeneri*, where, in 1870, he received his doctorate. Much later, Pareto still considered his dissertation—on 'Principi fondamentali della teoria di elasticità dei corpi solidi'—important enough to include it in his *Collected Essays on Economic Theory*. He began his professional career in Florence, working for the *Centrale della S. A. delle Strade Ferrate Romane*. From there, he went on to a management position with the *Società dell'Industria di Ferro*, traveling widely in Germany, Austria, France, England, and Switzerland.

His career left him with a sound knowledge of practical business matters. In 1882, the year his father died, Pareto chose professional retirement and settled in Florence. At around this time, he made an unsuccessful attempt at running for political office on a platform promoting his ideal of a liberal market economy. Pareto became resolved to dedicate his remaining life to his scientific pursuits. He joined Italy's oldest learned society, the highly respected *R. Accademia Economico-Agraria*. He also became a member of a second society, *Adamo Smith*, where he gave several lectures on scholarly topics and economic policy, which he also published. Moving in the upper-class circles of Florence—which then exerted a leading cultural influence throughout Italy—Pareto made a number of influential friends: the Lord Mayor Peruzzi and his wife; the politician S. Sonnino; P. Villari, who wrote a famous biography of Machiavelli; Colli, the author of *Pinocchio*; the anthropologist Mantegazza; such renowned philologists as D. Comparetti, A.

Franchetti, and A. Linaker, as well as several others. It was at this time that he married Alessandra (Dina) Bakounine (1860–1940), a Russian (no relation to the anarchist of the same name).

When Pareto met Italy's leading economist, Maffeo Pantaleoni, they soon became friends. Pantaleoni supported Pareto, as a successor had to be appointed for the Lausanne University chair held by Léon Walras, who had had to vacate his position due to his worsening mental illness. As an academic teacher, Pareto had considerable success. However, he soon asked to be relieved of his teaching duties. An unexpected inheritance set him free to concentrate on his preference for research and scholarship. Pareto's *Feste Giubilari* was published in 1917 to great critical acclaim. While this represented a personal triumph for him, Pareto continued to view himself as a victim of Italy's politicians and the *Camorra universitaria italiana*. When the Fascists took over Italy's government, various honors were offered to him; he either declined them explicitly or did not respond to them at all.

Only once did Pareto return to Italy and then only because he was forced to visit the Rome family record registration office in person. To get a divorce and be able to remarry, he needed his birth certificate. Dina, his first wife, had left him. In the summer of 1923, he married Jeanne (Jane) Régis (1879–1948), his second wife. He had met her through a newspaper advertisement in Paris. When Pareto died two months later, on August 23, 1923, he no longer possessed an Italian passport. To be able to marry Jane, he had been forced to turn in his passport at the Republic of Fiume. Conforming to his last wishes, he was buried in Céligny, his last place of residence, in a modest grave listing only the years of his birth and death and his *nom de guerre*.

2. Works

From his beginnings as a researcher, Pareto placed immense importance on the question of method. He had named his approach the logical-empirical method. In his *Manuale d'economia politica* (1994), first published in 1906, he devoted no less than 55 paragraphs to it. In his *Trattato di sociologia generale* (1983), that number rose to 144 paragraphs. 'My only intention,' Pareto wrote, 'is to describe facts and to draw logical conclusions from them' (Pareto 1983, 55). Facts were to be weighted according to relevance, not quantity. Facts based on experience were considered valid only 'as far as experience does not contradict observation' (Pareto 1983, 6).

Pareto considered this method applicable not only to the field of economics but to the social sciences as well and—implicitly or explicitly—to every field of scholarship (Pareto 1994, Vol. 1, 14). However, every inductive approach will always contain an implicit

assumption. In the sciences, this would be the supposedly homogeneous character underlying all of nature. Thus, Pareto's 'survey' is strongly similar to J. A. Schumpeter's concept of 'vision.' According to Pareto, it is quite feasible to differentiate between economic and social phenomena initially. To regard them separately does not violate the dictates of empiricism. But 'both parts then have to be integrated with each other. Only a full view of the whole thing possesses any empirical reality' (Pareto 1994).

At the heart of Pareto's *Cours d'économie politique* lies his theory of economic equilibrium. He refers to two sets of mathematical equations to describe two 'economic systems.' One system applies to barter and exchange, the other to production. If combined, both sets of equations attain validity only in a state of equilibrium (Pareto 1994, App. 131–3). 'The most interesting case,' according to Pareto, is described as a situation where the number of equations is less than the unknowns contained in the equations: 'The system is then flexible and can balance itself' (Pareto 1994, App. 131–3). 'Mathematics,' Pareto insisted, 'should be used only if it allows for an expression of certain relations between facts which otherwise could not have been expressed or which could not have been formulated using standard language' (Pareto, 559, n. 4, cf. Pareto 1983, 1732).

To describe different states of economic equilibrium, Pareto coined two terms: *ophelemité* and *utilité*. When Irving Fisher wrote a review of Pareto's work, Pareto tried in vain to discourage Fisher from confusing his terms with the terminology of Walras and Edgeworth. Fisher, Pareto argued, did not understand enough French and even misunderstood some of Pareto's mathematics (Pareto 1960, Vol. 3, p. 408).

In 1902–3, Pareto published the two volumes of *Les systèmes socialistes*, an incisive analysis of historical and present-day socialist theories, including a discussion of Werner Sombart (he had treated Karl Blücher in a similar way in his earlier *Cours*). He paid due reverence to Karl Marx as a sociologist. Economically, however, he did not give Marx much credit. Theoretically, a socialist economy might be able to solve its problems as well as any nonsocialist economy. Earlier, Pareto had already integrated socialist society and collective capital with his overall theory of economics (Pareto, Vol. 2, 714–24). Indeed, his followers, E. Barone and O. Lange, later based their own socialist theories of economics on this insight. Pareto argued against socialism not on economical but on social and anthropological grounds. He concluded that all socialist systems were eventually based on untenable social and anthropological contradictions.

In his *Manuale d'economia politica* (1994), Pareto presented a reworked version of his theory of economics with special emphasis on the various states of economic equilibrium. Typically, he differentiated between them in his famous dictum:

We can say that the members of a community, in a certain situation, enjoy the optimum of *ophelemité* if it has become impossible to distance oneself from this situation even to a small degree without benefit or damage to all members of this community; even the smallest step then necessary leads to profit for some members of the community or loss for others (Pareto 1994, 33).

This concept has been named the Pareto optimum. Later, it served as the foundation for welfare economics. It led to several attempts at an improved theory of income distribution.

Pareto always insisted that to understand his *Manuale*, one needed to know his *Cours*. In a typical misreading, J. R. Hicks complained that in the *Manuale*—by now a classic on the theory of consumption—the problems of capital and interest were hardly noted (Hicks 1950), ignoring their extensive treatment in the *Cours*.

It was Pareto's intention to show 'by examples rather than declarations to prove the relation between economic phenomena and other phenomena of society' (Pareto 1994, Foreword). Thus, it becomes clear how deplorable was the omission, in the French edition of the *Manuale* (Pareto 1981) of Pareto's Foreword or *Proemio*. Still, improvements had been made to other parts of the French edition, including Pareto's mathematics. He was of the opinion that 'pure' theory was applicable only to the flow of goods and trade within a 'purely' competitive economy. Yet, the reality of monopolies, cartels, and trusts asked for a different description of modern markets and their mechanisms. Eventually, he discarded his entire theory of economic equilibrium as a 'dead end,' in spite of his groundbreaking anticipation of what later essentially came to be called the theory of monopolistic competition (cf. the French edition of the *Manuale*, Vol. 3, 40–8, 160–64, Vol. 5, 8–9, 71; App. 141). With reference to the effects of economical cycles, Pareto already differentiated between their objective and subjective sides. In this, he could claim no less a predecessor than Montesquieu.

Pareto's final great work, the *Trattato*, was preceded by the brief *Mythe vertueuse et la littérature immorale* (Pareto 1911), an unforgiving settling of scores with any kind of literary censorship. Taken to its logical conclusion, Pareto argued, censorship would have to call for all of world-class literature to be censored—starting with the Bible.

In the two volumes of his *Trattato di sociologia generale* (1983), Pareto undertook the immense task of integrating his economic and social analyses. It had long been clear to Pareto that for any model of economic action fully to grasp reality it had to encompass a theory of society. Next to those actions described as logical, recognizably illogical actions had to be equally accounted for. Such actions called illogical by Pareto are regularly cloaked by pseudo-logical explanations. On this insight, Pareto erected an entire sociology of knowledge (cf. Eisermann 1987

p. 170ff.). Logical actions, to Pareto, were not synonymous with rational actions, however. For any act to be described as rational, it would have been necessary actually to look into the mind of the social actor. Pareto did not think such introspection was feasible. Rather, he conceptualized such actions as residual social facts. Thus, Pareto arrived at an inventory of six residues, in which instinct of combinations and persistence of aggregates were the most important ones.

He distinguished between two groups of economic actors: 'rentiers' and 'speculators.' Rentiers in general are secretive, cautious, timid souls, mistrustful of all adventure, not only of dangerous ventures but also of any venture to have even the remotest semblance of not being altogether safe (Pareto 1983, 2313). Pareto's rentiers are ruled by aggregates of persistence and his speculators by their instinct of combination. Speculators

are usually expansive personalities, ready to take up with anything new, eager for economic activity. They rejoice in dangerous economic ventures and are in the watch for them, but they work underground and know how to win and hold the substance of power, leaving the outward forms to others (Pareto 1983, 2313).

Accordingly, Pareto expanded his analysis of the economic equilibrium into an analysis of the social equilibrium (Pareto 1983, 2067–87). The crucial movement inside the system of society is construed as *circolazione delle classe scelte*, class circulation. Society is then ruled by fractions, or minorities, which are subject to a continuous turnover, or exchange (2025–46 and throughout).

Pareto always insisted that his 'pure' theory simply was an integral part of his overall economics, and that his economics necessarily formed part of his sociology. He claimed to have introduced the principle of relativity to the social sciences. He took equal pride in his anticipatory powers of foresight. Like his contemporary J. M. Keynes, he foresaw the catastrophic repercussions of the Treaty of Versailles—for purely economic reasons. As early as 1917, he asked, 'whether the Pax americana will turn out to be as productive as the onetime Pax romana?' (Pareto 1960, Vol. 3, p. 240).

See also: Economic Sociology; Economics, History of; Elites: Sociological Aspects; Mobility: Social; Political Sociology; Rationality in Society; Sociology, History of; Sociology: Overview; Welfare

Bibliography

- Bobbio N 1969 *Saggi sulla scienza politica in Italia*. Laterza, Bari, Italy
 Busino G 1989 *L'Italia di Vilfredo Pareto*, 2 Vols. Banca commerciale italiana, Milan

- Convegno Internazionale V. P. 1975 Accademia Nazionale dei Lincei. Rome
- Demaria G 1949 L'oeuvre économique de Vilfredo Pareto. *Revue de l'économie politique* 59: 517–544
- Eisermann G 1961 *Vilfredo Pareto als Nationalökonom und Soziologe*. Mohr, Tübingen, Germany
- Eisermann G 1987 *Vilfredo Pareto*. Mohr, Tübingen, Germany
- Eisermann G 1989 *Max Weber und Vilfredo Pareto*, Dialog und Konfrontation. Mohr, Tübingen, Germany
- Eisermann G, Malinvaud E 1992 *Vilfredo Pareto und sein 'Manuale'*. Verlag Wirtschaft und Finanzen, Düsseldorf, Germany
- Freund J 1974 *Pareto*. Seghers, Paris
- Hicks 1950 *Value and Capital*, 3rd edn. Oxford University Press
- Kirman A P 1987 Pareto as economist. In: Eatwell J, Milgate M, Newman P (eds.) *The New Palgrave* Vol. 3, Macmillan Press, London, pp. 804–9
- Pareto V 1911 *Le mythe vertueuse et la littérature immorale*. Rivière, Paris
- Pareto V 1920 *I sistemi socialistici*. Turin, Italy
- Pareto V 1952 *Scritti teorici*. Edizioni di storia e letteratura Milan
- Pareto V 1960 *Lettere a Maffeo Pantaleoni*, 3 Vols. Edizioni di storia e letteratura Rome
- Pareto V 1963 *The Mind and Society*, 2nd edn. Dover, New York
- Pareto V 1964–89 *Oeuvres complètes*, 30 Vols. Librairie Droz, Geneva, Switzerland
- Pareto V 1965 *Cronache italiane*. Morcelliana, Brescia, Italy
- Pareto V 1966 *Scritti sociologici*. UTET Turin, Italy
- Pareto V 1981 *Manuel d'économie politique*. Librairie Droz, Geneva
- Pareto V 1920 *I sistemi socialistici*. Turin, Italy
- Pareto V 1983 *Trattato di sociologia generale*, 4th edn. AMS Press, New York
- Pareto V 1994 *Manuale di economia politica*, 4th edn. Tesi, Pordenone, Italy
- Parsons T 1968 Vilfredo Pareto. In: Sills D I (ed.) *International Encyclopedia of the Social Sciences*. Vol. 11, Macmillan, pp. 411–16
- Schneider E 1961 Vilfredo Pareto in the light of his letters to M. Pantaleoni, *Banca Naz. Lavoro Quarterly Review* 58
- Schumpeter J 1951 Vilfredo Pareto. In: *Ten Great Economists*. Oxford University Press, New York
- Valade B 1990 *Pareto*. Presses Universitaires de France, Paris

G. Eisermann

Parietal Lobe

The parietal lobe is the part of the cerebral cortex lying between the occipital and frontal lobes, and above the temporal lobe (see Fig. 1). It is named for its proximity to the parietal bone, the wall of the skull (from the Latin *paries*, wall). It has five major functions: the primary analysis of somatic sensation (touch, the position of the limbs, temperature); the analysis of

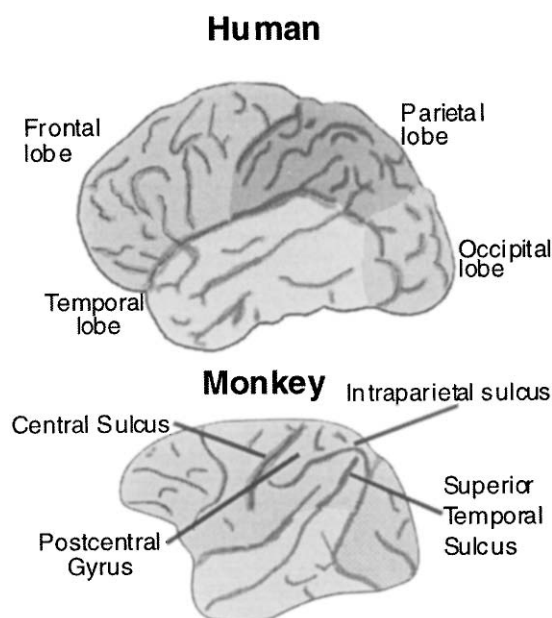


Figure 1

The four major lobes of the human and Rhesus monkey brains. The lobes are labeled in the human brain. In both the monkey and the human brain, the surface area of the cerebral cortex is so great that the cortical surface must be folded so that some tissue is on the gyral surface, for example the postcentral gyrus, and other tissue is hidden deep inside clefts called sulci. The major sulci of the parietal lobe are labeled in the monkey brain, as well as the postcentral gyrus. The lobes in the monkey brain have the same shading as the lobes in the human brain

space using all sensory modalities; the specification of spatial targets for the motor system; the generation of attention; and the analysis of visual motion. Our first understanding of parietal function arose from the analysis of the deficits of patients with lesions of the parietal lobe (Critchley 1953). Subsequent insight into the mechanisms of parietal function came from neuro-anatomical and neurophysiological studies.

The importance of the parietal lobe was apparent to neurologists such as Hughlings Jackson, Dejerine, Head, and Holmes early in the twentieth century. They discovered that humans with parietal lesions had difficulty with tactile perception on the side of the body opposite the lesion (the contralateral side): they were not completely unable to feel stimuli, as they might be after a lesion of the spinal cord, but they had relative deficits: they might be able to feel touch, but not pain or temperature; they might perceive touch but have difficulty localizing the stimulus; they might have difficulty knowing the position of a limb in space; they might have difficulty identifying an object placed

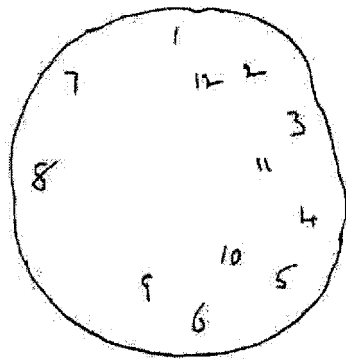


Figure 2

Drawing of clock by a patient with a right parietal lesion. Note that the numbers are displaced to the right, as if the patient is neglecting the left side of the drawing, and that the relationships of the numbers to each other are also disturbed, for example the 10 is to the right of the 9 (reproduced with permission of M. Critchley from *The Parietal Lobes*, 1953)

in their hand without looking at it—for example, they would not be able to reach into a pocket and pull out a key, leaving coins behind. Stimuli were often perceived to be less strong on the affected side than on the normal side of the body, and comparative discriminations were coarser. Sometimes patients who can appreciate a stimulus on an affected limb when it is the only stimulus touching them cannot appreciate it when two stimuli are applied simultaneously, one on the affected limb and the other on the normal limb. The most dramatic parietal deficit is anosognosia—patients, especially those with right parietal deficits, are entirely unaware of the limbs on the contralateral side: they think that a leg in bed with them belongs to someone else.

Patients with parietal lesions do not have deficits limited to the tactile or somatosensory domain: they have striking visual deficits as well. Patients with right parietal damage tend to neglect the left visual field, the left side of space relative to the midline of their bodies, and the left side of objects (see Fig. 2). When they draw a clock they cram most of the numbers on to the right side of the clock, neglecting the left; when they attempt to bisect a line they place the midline near the right end of the line rather than the middle. After this florid neglect has recovered, right parietal patients often still neglect a stimulus in the affected visual field when another one is presented simultaneously in the normal visual field. This deficit is interpreted as a deficit in visual attention: patients with parietal lesions are less efficient at shifting their attention into the affected field than into the normal field.

Patients with parietal lesions often exhibit optic ataxia, the inability to coordinate their movements using visual cues. They have difficulty knowing where

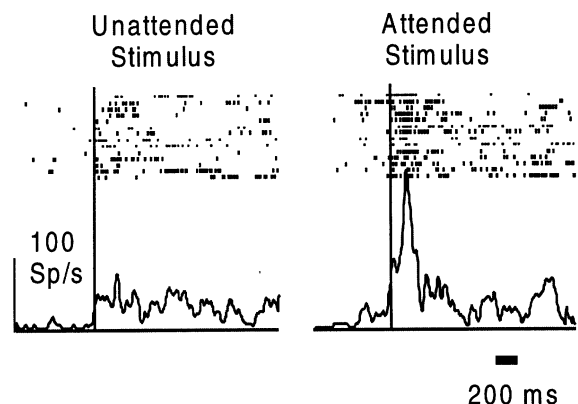


Figure 3

Enhancement of the response of a neuron in the lateral intraparietal area to an attended stimulus. The upper display is a raster diagram: in this diagram each dot represents a single action potential. Each line is a single presentation of the stimulus, which is accomplished by having the monkey make a saccade that brings the stimulus to the proper part of the retina. The graph underneath the raster display is the average firing rate of the neuron as function of time. The vertical line represents when the stimulus arrives in the part of the retina to which the cell responds. The cell responds immediately after the saccade. When the stimulus has suddenly appeared 400ms before the eye movement, and has the attentional effect of an abruptly appearing object, the response of the cell to the appearance of the stimulus at the proper part of the retina (attended stimulus, left) is much greater than the response to the stimulus when it has been present in the environment for a long time, and is therefore not attention worthy (unattended stimulus, right) (adapted from Kusunoki et al. 2000)

objects are relative to their body. They misreach for objects, and make inaccurate eye movements toward them. They also have difficulty transmitting visual information to the motor system. When humans reach for an object, they open their hand to the approximate width of the object before their hand arrives at the object. Thus, people reaching for a can of soda will open their hands wide to accommodate the can; people reaching for a pencil will open theirs narrowly, just to accommodate the pencil. Patients with parietal lesions cannot do this, but reach for all objects with a stereotyped hand posture. The brain mechanisms that drive limb movement have no access to the detailed visual aspects of a target, even though they might know the spatial location of the target. A remarkable aspect of parietal deficits is that, although patients do not know where an object is, they have no difficulty telling what that object is—they know, for example, that the examiner is holding a wristwatch in front of

them, but cannot find it to reach for it. They exhibit constructional apraxia—when asked to duplicate a block design they are unable to place the blocks in proper relationship to each other, but instead leave them in an unordered jumble even though they choose the proper blocks. Patients with lesions of the parietal cortex on both sides of the brain have difficulty shifting their attention away from the current object. The result of this deficit is that they only describe one in the world at a time—they may see a dog's nose but not its ears, for example. They cannot use eye movements to explore the visual world. Patients with small lesions of the parietal cortex occasionally exhibit a specific deficit for the analysis of visual motion and the transmission of this information to the motor system. They have deficits both with the motor aspects of visual motion, for example, they cannot follow a moving object with their eyes; and they have difficulty with the perception of motion—knowing the speed of an object, or judging the predominant motion of a moving random dot pattern.

Space is not represented in a unitary fashion in the parietal cortex. Instead it seems to be organized in concentric spheres, or workspaces, that correspond to the range of action of certain motor systems. Patients with small lesions of the parietal cortex may have these different workspaces affected differentially. Thus a patient who can localize a nearby stimulus by reaching out and touching it with an arm may not be able to localize a stimulus further away in the same direction by touching it with the beam of a laser pointer. These are not merely deficits in movement—the patients show perceptual distortion, such as an inability to bisect a line, in the affected workspace.

Recent studies in neuroanatomy and neurophysiology, especially those done with the Rhesus monkey, have been able to illuminate the brain mechanisms whose destruction causes the dramatic clinical symptoms of the parietal lobe. The brain is a network of information-processing cells called neurons. Each neuron serves as a node in this network, receiving information from other neurons, often in different areas of the brain, and transmitting this information to other neurons. By studying the cytology of the parietal lobe and its connections with other parts of the brain, neuroanatomists have been able to divide it into a number of functional subdivisions. Neurons transmit information by emitting a series of electrical impulses called action potentials. If a neuron responds to a stimulus it issues a burst of impulses in response to the appearance of the stimulus. By monitoring the electrical activity of neurons in the parietal lobe of monkeys that are awake and performing behavioral tasks, neurophysiologists have begun to understand the signal processing that enables the parietal lobe to perform its function in the generation of human behavior.

Physiological and anatomical studies have subdivided the parietal lobe into a number of distinct

areas which perform the various different functions whose deficits appear in patients with parietal lesions. The most anterior part of the parietal lobe is the primary cortical area for somatic sensation. Most of the posterior parietal cortex, starting with the posterolateral bank of the intraparietal sulcus is concerned with space and attention. The most posterior parietal area, in the superior temporal sulcus, is important in the analysis of visual motion.

The primary somatosensory cortex consists of four subdivisions, cytoarchitectonic areas 3a and 3b, located in the posterior bank of the central sulcus, and areas 1 and 2 located on the surface of the postcentral gyrus, the brain surface between the central and intraparietal sulci. Each of these subdivisions receives projections from the somatosensory parts of the thalamus, the way station between the spinal cord and brainstem, and the cortex. Areas 3a and 2 are responsible for deep bone and muscle sensation, including the muscle spindles that measure muscle length, and joint sensors that describe limb position. Areas 3b and 1 are responsible for sensation from the skin. These areas project backward to area 5, on the postcentral gyrus and the anterior bank of the intraparietal sulcus. This area integrates the various primary sensory signals, and transmits somatosensory signals to the premotor cortex so the information can be used for not only the perception of somatic sensation but also the coordination of movement. Single neurons in the anterior parietal cortex describe the properties of tactile stimuli in a very precise manner: for example, Mountcastle showed that neurons responsive to vibration describe the intensity and frequency of that vibration so accurately that a computer could construct the nature of the stimulus by decoding the neural pulse train.

The posterior parietal cortex is subdivided into a number of areas. Those located in the intraparietal sulcus associate somatic sensation, vision, and audition for the analysis of space, and the description of this world for action. Many of the areas in the intraparietal sulcus describe particular workspaces (Colby and Goldberg 1999). The physiological properties of these areas have been studied extensively in the Rhesus monkey.

Perhaps the best-studied area of the monkey parietal cortex is the lateral intraparietal area (LIP). This area describes distant space, the space that primates explore with their eyes. It projects to the frontal eye fields and the intermediate layers of the superior colliculus, brain areas known to be important in the generation of the rapid eye movements (saccades) that human and non-human primates use to explore their environment. It also projects to areas in the temporal lobe, TE and TEO in the Von Bonin and Bailey nomenclature, which are important in the analysis of pattern. An important connection between the generation of saccadic eye movements and the analysis of the visual world is the process of attention, the selection of

objects in the world for special analysis. In general, human and nonhuman primates pay attention to the targets for their eye movements, and saccades and attention are closely linked, but they can be dissociated in certain cases—for example juvenile Rhesus monkeys certainly pay attention to the alpha male of their troop, but they never look directly at him. Animals attend to stimuli either because the stimulus is relevant to the current task, or because the stimulus is naturally attention worthy, stimuli that are, in the words of William James, ‘big things, bright things, moving things, or ... blood’ (James 1890). For example, stimuli that appear abruptly in the visual field inevitably draw visual attention.

Neurons in LIP respond both to saccade targets and to attended stimuli that are not necessarily the targets for saccades. These neurons respond to visual stimuli, and their activity is modulated by whether the animal is paying attention to the stimulus. Neurons in LIP respond more to attended than to nonattended stimuli, independently of whether those stimuli are attended to because of their importance to the task or because they are in the class of stimuli that inevitably draw attention. Presumably it is the absence of this attentional modulation system that impairs the ability of parietal patients to attend to stimuli in the world contralateral to the lesion. Neurons in LIP also respond to sounds when the monkey attends to them (Andersen et al. 1997). A classic task in which to study the activity of neurons is the delayed saccade task.

In this task a monkey looks at a spot of light and another one flashes briefly in the visual field. A few seconds later the original fixation point disappears, and the monkey makes a saccade to the spatial location of the now vanished saccade target. Neurons in LIP respond to the stimulus, and then continue responding even though the stimulus has disappeared. This activity has been interpreted as either maintaining the memory of the saccade target, or focusing attention at the saccade target. It is probable that this area of the brain supports both functions: focusing attention on spatial locations, and providing the oculomotor system with information about the location of potential saccade targets. LIP serves as a model for the other, less well studied areas of the parietal lobe: neurons specify possible targets for the action appropriate for the workspace, as well as responding to attended objects in that workspace.

The most posterior part of the parietal cortex of the monkey lies in the superior temporal sulcus. This area is important for the analysis of visual motion. Neurons in the middle temporal area and the middle superior temporal area are specialized for the analysis of visual motion, and respond selectively to stimuli moving with certain speeds and direction. Electrical stimulation of these areas can affect monkeys’ perception of motion, and lesions in these areas render monkeys unable to follow moving stimuli with their eyes, or to make judgments about the quality of visual motion.

See also: Navigation in Spatial Environments, Neural Basis of; Perception of Extrapersonal Space: Psychological and Neural Aspects; Somatosensation; Topographic Maps in the Brain; Visual System in the Brain

Bibliography

- Andersen R A, Snyder L H, Bradley D C, Xing J 1997 Multi-model representation of space in the posterior parietal cortex and its use in planning movements. *Annual Review of Neuroscience* **20**: 303–30
- Colby C L, Goldberg M E 1999 Space and attention in parietal cortex. *Annual Review of Neuroscience* **2**: 319–49
- Critchley M 1953 *The Parietal Lobes*. Hafner, New York
- James W 1890 *The Principles of Psychology*. Holt, New York
- Kusunoki M, Gottlieb J, Goldberg M E 2000 The lateral intraparietal area as a salience map: the representation of abrupt onset, stimulus notion and task relevance. *Vision Research* **40**: 1459–68

M. E. Goldberg

Parliamentary Government

1. Development of Parliamentary Regimes and Democracy

Parliamentary democracy is the product of the twentieth century and in many countries developed only after 1918. There were some attempts to create such a system during the European revolutions of 1848 without leading to consolidated systems. What Huntington has called the first long wave of democratization (1828–1926) in the light of research on parliamentary systems was only parliamentarization. Parliamentary government is characterized by a government *responsible to the majority in parliament*. Democracy was added to the parliamentary regimes only later. The minimal criterion of democracy is *universal suffrage*. In most countries with a regime of constitutional monarchy, such as Britain, Scandinavia, or the Benelux countries, democracy was introduced later than parliamentary regime. Only in some rather authoritarian monarchies was democratic universal suffrage introduced or preserved without granting a parliamentary regime with responsible government (France 1851 in the Second Empire, Germany 1871 in the Empire of Bismarck) (see Table 1).

Only in a few countries (France, Finland, Sweden) was the process of parliamentarization and democratization synchronized. A continuously democratic parliamentary regime existed only in France. Even Britain—the mother of ‘parliamentary government’—in the nineteenth century could not be called a

Table 1
Parliamentary democracy: the merger of parliamentarization and democratization

	Continuity of a system of estates transformed into a constitutional regime					New regimes					Parliamentarization of neo-absolutist regimes				
	The					Italy-					Piedmont				
	UK	Sweden	Norway	Finland	Netherlands	Luxembourg	Belgium	France	France	France	Spain	Germany	Austria	Denmark	
General male suffrage	1918	1921	1898	1906	1917	1919	1919	1848	1875	1869	1871	1907	1915		
Time of creation of a government responsible to parliament	1688ff 1832-35 final conflict	1917	1884	1906	1868	1868ff	1833ff	1821ff 1837 1848-51 1871ff		1907	1919	1918	1901		

'parliamentary democracy' (cf. Table 1). There was a divergence of *parliamentarization* (mostly implemented by liberal majorities which were reluctant to add universal suffrage to the system) and *democratization* (frequently introduced under the pressure of radicals and socialists). These groups were initially unwilling to accept the checks and balance of the liberal parliamentary regime and advocated a democratic regime with many features of plebiscitarian instead of representative democracy. Only when they came to power did they slowly accept the rules of the game which were mostly developed in former constitutional monarchies.

Democratization in some countries has ruined the parliamentary system (Italy, Austria, Germany, Spain) when counter-regime forces gained power. In other established parliamentary systems the regime in the 1930s came to the brink of a collapse (Belgium, France) under the pressure of right-wing extremism. Only after 1945 and the disastrous experiences with authoritarian or even totalitarian dictatorship, were parliamentary government and democracy combined in peaceful co-existence. Some countries (such as the Federal Republic of Germany) even then, however, showed a tendency to emphasize the legal state and *representative parliamentary government* much stronger than *democratic participation*. The theory of constitutional and representative government did not easily accept parliamentary democracy but interpreted the preponderance of the parliamentary majority in terms of a rather mechanistic system of checks and balances and 'mixed government.' Because responsible cabinet government developed first in Britain and Belgium, the jurists for quite a time used the term '*Coburg government style*.' Only since the 1830s was the term 'parliamentary government' accepted in Britain and France and the regime was recognized not only as a mode of governance but as a *regime sui generis* which made the old typologies of regime types according to the number of rulers rather obsolete.

2. Types and Criteria of Parliamentary Democracy

There are many types of parliamentary government, but most of them have developed common *institutional* criteria.

(a) *Compatibility of parliamentary mandate and ministerial office* (exceptions are the French Fifth Republic, Luxembourg and The Netherlands) in order to guarantee close cooperation between the parliamentary majority and the executive.

(b) Prime ministers are normally *members of parliament*.

(c) The government has to have the confidence of the parliamentary majority. *Ministerial responsibility* is not defined in terms of legal responsibility as in

dualistic constitutional monarchies. The government has to resign in the event of a vote of no-confidence unless it chooses to ask the head of the state to dissolve parliament in order to give the electorate the opportunity to resolve the conflict.

(d) Parliament controls the *government by raising* questions, exercises the right to interpellate and set up committees of enquiry, which facilitates the decision about whether the strongest sanction or a vote of no-confidence should be used. This right is shared by dualistic presidential systems, apart from the possibility of toppling the head of government and his or her cabinet.

(e) Some analysts have postulated as an essential feature a formal vote of the government at its first meeting with parliament (*investiture*), as in the French system under the Third and Fourth Republic and in Italy, or even demanded the *formal election of the prime minister*, which was developed in the German Basic Law of 1949. Most parliamentary regimes do not accept this condition.

(f) More widespread is acceptance that *parliament* should be *dissolved* if the prime minister has lost the confidence of the parliamentary majority. Sometimes this has not been instituted, for example in Norway and Israel. In other cases the legal possibility of dissolution has become inapplicable because it has been abused by unparliamentary presidents, as in the Third French Republic (when President MacMahon dissolved the Chamber of Deputies in 1877).

Besides the minimal institutional criteria, certain social structural features are essential for the consolidation of parliamentary government:

(a) *Organized parties* to facilitate the building of parliamentary majorities.

(b) Party-building to facilitate the development of *cabinet solidarity*.

(c) Development of the *office of prime minister*. In early systems where the estates dominated, and even in certain absolutist monarchies, the power centers—the estates or the crown—impeded the activities of first minister. A certain *hierarchization of the ministerial council* also strengthens cabinet solidarity.

(d) Development of a *political culture* favorable to appropriate parliamentary behavior and alternating government.

Not all these institutional criteria can be found in every type of parliamentary system. Therefore subtypes have been proposed. Most ideological is the distinction between *authentic British parliamentarism* and the allegedly *inauthentic parliamentary system of the French type*—close to revolutionary government. Latecomers to parliamentarization, such as Germany (1918) and Sweden (1917), have often denounced the dangers of the French type, especially the German school of Redslob.

Oddly enough, France itself has reacted strongly against the shortcomings of the parliamentary system of the Third and Fourth Republics. De Gaulle's

constitutional ideas, close to those of Carré de Malberg and later developed by Michel Debré were first pronounced in 1946 in the famous speech of Bayeux but failed to obtain majority when the Fourth French Republic was developed in 1946/47. But De Gaulle got his revenge when the Fourth Republic collapsed and he was able to shape the institutions of the Fifth Republic according to his ideas. A new system was born that since Duverger has been called the *semipresidential system*.

In spite of its name this was clearly a subtype of parliamentary government though the double responsibility before parliament and the head of the state—which existed in constitutional monarchies such as that of Louis Philippe—re-emerged. But in the long run the dominance of parliament in executive-legislative relations is re-established. De Gaulle's new system was not new at all: it had developed as a Republican type of constitutional regime from 1848 (the Second French Republic, 1848–51) and re-emerged in the Weimar Republic (1919) and Finland. Semipresidential systems are not a mixed type, not even the 'three-quarter presidential system' that Yeltsin tried to develop. The backlash of a humiliated legislature was felt in Russia in 1998 when Yeltsin was forced to renounce Chernomyrdin as prime minister and accept the parliamentary candidate Primakov.

3. Functions and Performance in Parliamentary Democracy

Parliaments were more powerful in times of parliamentary sovereignty without universal manhood suffrage than in modern parliamentary democracies. The main actor in parliamentary democracy is no longer parliament or government, but the two powers are linked together. This linkage has strengthened the prime minister in most parliamentary regimes, unless high party fragmentation and consociational practices of conflict resolution have weakened the head of the government. Parliamentary democracy seems to be misnamed because of much discussion on the 'decline of parliament.' Therefore, *cabinet government*, *prime ministerial government*, or *chancellor's democracy* became more popular to characterize the bigger parliamentary regimes.

Parliamentary systems have to perform various functions. There is no general decline of parliament but rather a change of the main functions of the system. *Parliamentary control* is weakened in systems with close relations of parliament and government, interlaced by party government. The *controlling functions* declined or shifted to other institutions from constitutional courts and courts of account to the mass media. *Representation and articulation of interests* in the systems has changed over time. The *legislative function* is still important because more

social problems are regulated in modern democracies than in old constitutional regimes. The *recruiting function* of the system, however, is strengthened after democratization of the parliamentary regimes.

Parliamentarians have become less removed from their voters in terms of *social structure*. What has been called *virtual representation* in early parliaments for those not eligible to vote in *postparliamentarism* is true of the social ties between deputies and voters. But people have adapted to this development and demand less strongly to be represented by members of the same region, the same sex, the same profession, or the same social class. The influence of voters on the selection of representatives is still mediated by the parties and their formal members. Primary elections can undermine the selecting function of the parties among other disadvantages, such as a lower voter turnout in the general election. Modern democracies are not ruled according to President Lincoln's Gettysburg formula of government *of* the people, *by* the people, and *for* the people. The distance of the representatives is compensated for by increasing *responsiveness* and sensibility to the wishes of the voters. Government *for* the people in the modern democracy of fluctuation of mass public moods has increased rather than decreased.

Nonetheless the deparliamentarization of democracy is lamented. This has a *domestic political* side. At the stage of policy formulation, the practice of expert committees working together with the government is growing stronger. In the decision-making arena, the government is the most important initiator of legislation. Informal initiatives behind the formal initiatives of advisory boards, constitutional courts, or international organizations are further undermining the part played by parliamentarians in legislation. The penetration of interest groups into the work of ministers and parliament has shifted the preparatory work to a round tables of extraparlimentary committees and working teams. But parliament is not just a rubber stamp to ratify the decisions of networks outside parliament. Parliaments, formally the institutional seat of the people's sovereignty, represent the framework for the coordination of networks of actors from parliamentarians, party strategists, interest groups, ministry officials and in federate states the federate units. The cosy triangles of early American parliamentary studies, composed of parliamentarians, bureaucrats, and interest groups, have become unc cosy squares, since party steering was rediscovered in the 'legislative Leviathan.' In federal states such as Germany, deparliamentarization at the Länder level has led to greater participation of the Länder in the national decision-making process via intergovernmental decision-making. This further actor has been added, making an unc cosy pentagon. The Maastricht Treaty has increased the number of areas that are subject to EU rules and legislation. With growing competence at the EU level, lobbying at that level is

likewise growing. Meanwhile, national parliaments—more than the executive—have been downgraded to the position of lobbyists. Paradoxically the *democratization* of the European Union has led to *deparliamentarization* at the national level. Until 1979, Members of the European Parliament were tied to the national parliaments, but since the introduction of direct elections this has no longer been the case. However, the European Parliament does not yet pose a real threat to the national parliaments as it can only draw up laws when asked to do so by the Commission. Formally, EU legal acts are passed by the Council of Ministers and not by the European Parliament.

There are three threats to parliamentary democracy:

(a) In the Process of *regionalization*, regional assemblies will try to strengthen themselves via regional politics at the cost of national parliaments.

(b) In the process of *Europeanization*, more and more domains of the national parliaments will be regulated by the EU. This process will be supported by the activities of the European Court and the gradual standardization of the national legal systems.

(c) In the process of *globalization*, a consolation at the national level is the fact that worldwide organizations such as the World Trade Organisation and GATT are starting to retaliate to the EU for what it did to the national decision-makers because the room for maneuver of the European organizations is increasingly limited by the global level, although still strongly limited in some arenas.

See also: Democracy; Democracy, History of; Democratic Theory; European Integration; Globalization: Political Aspects; Parliaments, History of; Political Culture; Political Representation; Regional Government; Representation: History of the Problem

Bibliography

- Blondel J, Müller-Rommel F (eds.) 1988 *Cabinets in Western Europe*. Macmillan, Basingstoke, UK
- Budge I, Keman H J 1995 *Parties and Democracies*. Oxford University Press, Oxford, UK
- Döring H (ed.) 1995 *Parliaments and Majority Rule in Western Europe*. St. Martin's Press, New York
- Laver M, Shepsle K A (eds.) 1994 *Cabinet Ministers and Parliamentary Government*. Cambridge University Press, Cambridge, UK
- Norton P 1993 *Does Parliament Matter?* Harvester Wheatsheaf, New York
- Schütt-Wetschky E 1984 *Grundtypen Parlamentarischer Demokratie*. K. Alber, Freiburg, Germany
- Shugart M S, Carey J M 1992 *Presidents and Assemblies*. Cambridge University Press, Cambridge, UK
- Strom K 1990 *Minority Government and Majority Rule*. Cambridge University Press, Cambridge, UK

von Beyme K 2000 *Parliamentary Democracy. Democratization, Destabilization, Reconsolidation 1789–1999*. Macmillan, Basingstoke, UK

K. von Beyme

Parliaments, History of

Nowadays, a Parliament is a central institution of the democratic state under the rule of law. The term parliament comes from French and describes a forum in which social understandings are reached by verbal negotiation, through the act of speaking. A Parliament is a representative institution, elected by direct universal suffrage. It plays an indispensable role in the workings of the State, as an active participant in the coalescing of political will, in public management, and essentially through conduct of the legislative process and concentration of the ruling elite. The decisive function of policy legitimization performed by parliaments is central to representative democracy which is the predominant system of political organization in contemporary states. Experience of direct democracy, which dispenses with the representative parliamentary system or acts in parallel to it, is generally limited to referendums, Switzerland being among the few countries that practice direct democracy on a regular basis. A Parliament is an assembly of public delegates who receive a mandate for a limited period, are voted to office through elections based upon a pluralistic political-party system, and serve as a channel for the expression of the opinions of society.

The approaches of political scientists and historians to the study of parliaments are closely related. Most contemporary researchers apply the insights of both disciplines in their efforts to analyze and explain the institutional structures and workings of parliaments. Comparative studies of the trajectory of societies, of the formation of states, and of the composition of their social and economic groups are the components of research on representation, participation, and decision-making processes conducted through and by means of parliaments. The achievement that parliamentary institutions represent, in terms of the development of democracy within society throughout the nineteenth and twentieth centuries, provides the social sciences with ample scope for research. Political science, sociology, and history continually delve into the origins, composition, work methods, social relevance, political prominence, and social roles of parliaments.

1. Functions and Types

The functions of parliaments have differed according to the system of government. Various terms may

designate a parliamentary institution according to the different systems, such as: parliament, court, assembly, council. Under a parliamentary system, executive power emanates from the legislative branch invested in parliament, and governments hold office and remain in power so long as they enjoy the confidence of the parliament. Under a presidential system, parliament serves as a counterpoint to the political legitimacy bestowed upon the president through direct presidential elections. Parliaments also exist in nondemocratic regimes, such as the fascist dictatorships and 'democratic peoples' republics' of the twentieth century. Under such systems they tend to be mere puppets or facades since they exercise no real power but rather serve as rubber stamps to endorse the power of the government, devoid of the legitimacy bestowed by a pluralistic party system and in the absence of representative electoral processes.

Depending on the process leading to their formation, parliaments can be classified according to three criteria: composition, action, and structure. Composition is subject to two variants: the first of these relates to criteria based on noble birth or propertied status, whereas the second is based on universal suffrage. Functionally speaking, contemporary parliaments can also be classified as bicameral or unicameral, depending on the constitutional framework of the State, which may have opted for a single chamber or for two chambers. In any case, the chamber (or at least one of the chambers) ought to be elected directly by the people in free elections.

In functional terms, parliaments may be classified as 'active' or 'deliberative.' Active parliaments are those that, within a parliamentary system, have the power to appoint and support the government, as is the case in most modern parliamentary democracies. The so-called deliberative parliaments are those that have no direct responsibility for the business of government, and include the pro-forma parliaments tolerated by authoritarian regimes. Under a presidential system, as a counterpoint to the legitimacy of the 'working' executive branch, the parliament assumes a predominantly deliberative role. These characteristics apply equally to unicameral and bicameral parliaments, in both federative and centralized states.

It is the decisive role performed by parliaments in the process of forming governments and providing them with support that has given currency to the term Parliamentary System to describe the form of government in which empowerment and support of the Executive Branch are dependent on a vote of confidence from the parliament. This definition applies to the form of government a term that expresses the legitimacy that stems from the legislative branch in its executive form. Under a presidential system in which the Executive is the predominant Branch of Government, as is the case in the United States, the Head of State is voted to office by direct elections and consequently his legitimacy competes with that of the

parliament. In both parliamentary and presidential systems, the parliament is said to embody the Legislative Branch and its political legitimacy stems from the ballot.

2. Origins

Modern parliaments stem from two roots. The first of these is the corporative assemblies of the Estates that were summoned under the *Ancien Regime* in Europe, and the second, the broad-based popular representation originating from revolutionary movements. The first model of corporative representation brought together the three Estates—namely, the nobility, the clergy, and the ‘third estate’ which evolved gradually into a form of parliamentary representation as sovereign rulers required assent to levy taxes and make war. This first version of the dependence of the sovereign ruler on an institution of a parliamentary nature did not lead directly to democracy nor to recognition of popular sovereignty. The origin of these lies in the second root of modern parliaments—the popular revolutionary root—from which stem: the origins of the process of definition of popular sovereignty, the concept of citizenship, the legitimacy of power based on representation and the delegation of sovereign power, the rule of impersonality of the law within a context of absolute civil equality among citizens, the extension of such rights to all adult citizens within society and the adoption and consolidation of liberal democracy, the expansion of suffrage and political parties, the constitution of government and of its control (under both the parliamentary and presidential systems), and the expression of popular will through the ballot. Both roots reflect anti-absolutist sentiments and it is from them that a representative democratic model, with elections at predetermined intervals, developed.

Both these origins of parliamentary power are clearly preserved and apparent in the structure and functioning of the British parliament. The House of Commons, based upon the weight it amassed in the civil war of 1642–9 and the upon manifestation of its social powers in the Glorious Revolution of 1688–9, assumed its role of expressing the collective will as a source of power detrimental to that of the Crown. The result of the revolution, which established the right of parliament to chose a new sovereign, is reflected in the jusnaturalist and contractual arguments presented in Locke’s *Two Treatises on Civil Government* (1690)—one of the principal theoretical works in support of the legitimacy and controlling powers of parliament.

The House of Lords, with its elitist composition, preserves the corporativist tradition of the Estates. Modern French tradition, established with the Jeu de Paume Oath of June 20, 1789 at the advent of the French Revolution, signals a break with the system of Estates General of the *Ancien Regime* and use of the

term Parliament to designate the judicial and notarial functions, as it had since the twelfth century. The parliamentary assembly installed in the very earliest days of the revolution enshrined the principal of popular sovereignty. In both the British and the French cases, affirmation of the principle of popular sovereignty arose from revolutionary movements, whereby the political absolutism of European states was substituted by constitutionalism. The role performed by a representative system in the formation of the United States is unique. The affirmation of the will of the Americans against unconditional rule by the British Crown led to the establishment of a community-based version of the decision-making process, similar to the community practices of the Protestant congregations, which served as a collegiate political model for the provincial and continental congresses, in which no form of corporative representation was present.

3. Evolution

At the onset of the nineteenth century there were three types of parliaments: mixed-popular and aristocratic-bicameral under the English model; the French unicameral, popular model; and the American bicameral, popular and federalist model. In Western Europe and in the Latin-American republics, an era of constitutionalization began. Both monarchies and republics adopted political systems dependent on representative legitimacy. Under parliamentary systems (in which the Government is dependent on investiture by the Legislative Branch), and under presidential systems (in which the Executive is chosen by direct elections), parliaments assumed increasing importance in political life. Parliament became the forum for representation and expression of social and economic forces, where the repercussions of rights claimed took on resonance and the extension of political and civil rights was played out.

The role of parliaments evolved during the course of the nineteenth century from being initially consultative to being deliberative. The legitimacy of parliaments became independent of the legitimacy of the sovereign upon emerging from the electoral process. As a consequence of electoral and census systems which, throughout the 1800s, maintained cultural and economic distinctions among citizens, despite their all being held (theoretically) equal before the law, parliaments maintained an indirect corporative dimension until the institution of universal suffrage. Such distinctions were based on the systematic exclusion of women and the preference for the rich, the well-born, and the literate.

The itinerary of consolidation of the power of the Legislative Branch was implemented in different ways under the North American presidential, the British

parliamentary, and the French and German parliamentary models. The United States Congress reflects the federative structure of the State and the fact that the United States spent the nineteenth century engaged in a process of expanding its frontiers and acquiring new territory. Furthermore, the political model adopted by the American Congress implies that each state of the federation, through actions taken by its own political spheres, has joined the Union. The principle of legally joining a federation of states, together with that of popular sovereignty, is enshrined in the bicameral parliamentary systems predominant in federative states (such as the United States, Canada, Australia, Brazil, India, Austria, Mexico, Argentina, Germany, Japan, and Indonesia). In centralized states, bicameralism is the result of the continuity of mixed forms of representation, with roots stemming back to the Estates and to popular sovereignty. Both the British and the French parliamentary systems provide examples of such evolution. In both cases, tradition has been partially preserved through the upper house (the House of Lords and the Senate), whereas popular representation is ensured by the Lower House (the House of Commons and the National Assembly). This model was also adopted by the Scandinavian monarchies and also, at specific times, by such states as the German Empire (1871–1918) and the multinational Austrian–Hungarian Empire (1848–1918).

Parliaments emerged upon the European scene as a counterpoint to monarchic autocracy. In their contemporary form as ‘parliaments of citizens,’ they emerged from an extensive line of development going back to the principle of egalitarian representation first put forward in the eighteenth century. The parliamentary model of the late twentieth century is the fruit of the maturing of philosophical concepts with respect to human rights and the separation of powers. As the twenty-first century dawns, there is widespread social awareness that there can be no exercise of power without political representation, and that this resides in the legislative bodies. The transformation of political regimes into fully-fledged democracies has led to an expansion and diversification of the roles of parliaments. To the role of authorizing tax collection has been added deliberation of public budgets, supervision of government actions, and full participation in decisions regarding public policies.

The reform of electoral systems and extension of suffrage have placed on the stage of public and social life the professional politician, a party member who represents specific interests. The principal interest that became enshrined over the length of the nineteenth century—from the French Revolution to the end of World War I—was to force the acceptance of, respect for, and practice of popular sovereignty as the source of the legitimacy of all power.

Central to an understanding of the evolution of parliaments is the concept of popular sovereignty. It was the new element that entered the political scene in

the eighteenth century, whereby Monarchic authority—though still enjoying wide support—was no longer accepted as being absolute, nor independent of social support. It found expression in parliamentary democracy which essentially signifies a system in which government is made up of a nationally-elected assembly to which it is accountable. Acceptance of this concept was consolidated during the course of the nineteenth century when the ideal of democracy became associated with the parliamentary system of government.

‘Active’ parliaments—the importance of which has increased remarkably since 1945 and is still increasing at the onset of the twenty-first century—have taken on a direct responsibility for furthering the public interest. This is in line with a historic trend, initiated in the 1790s, of delegating sovereignty to elected representative bodies. In such parliaments, the house assumes not only the role of deliberating themes of interest to society but also of making decisions relating to public policies stemming from them. Under presidential systems, as is the case of the United States, the powers of the Executive are legitimized directly by popular vote, independently of parliament. Although this is true in terms of the origin of legitimacy (the ballot), it does not mean that the concrete exercise of power is free of controls. On the other hand, the extension of suffrage and the continuous rate of development of political participation since the 1830s have brought about changes in the exercise of power which have not excluded the presidential systems. Though it may even be possible to govern without parliament—against parliament it is not possible to govern. Though a strong presidential government might tend to ignore collective assemblies, nonetheless, the democratic framework of the State serves to counterbalance the relationship. Parliamentary democracy is an achievement of the Western European States and, increasingly of a number of Latin American and former communist Eastern European States. It is also the result achieved by certain nations of the British Commonwealth which, since becoming independent of Britain, have incorporated the parliamentary system of government. Such is the case in Canada, Australia, India, and New Zealand. Democratic and parliamentary systems are rarer and less stable in Africa, the Middle East, and the Far East.

The expansion of suffrage, resulting in mass democracy and recourse to regular popular participation (formalized through political parties or indirectly through lobbies or special interest groups), have caused parliaments to assume the roles of formulation and assessment of public policies (and of decision-making under parliamentary systems of government), particularly in the financial area (through the preparation of the budget and the exercise of control over its execution). The rate at which parliamentary institutions have assumed a leading role in the political process varies from one region to another. In periods

of economic expansion or of war, the active role of parliaments may go into retreat, as occurred in the *Belle Epoque* and during World Wars I and II. Under authoritarian regimes, of the left or of the right, it is not uncommon for parliaments to be merely irrelevant decorative trimmings to the institutional architecture (as was the case in Fascist Italy, Nazi Germany, in the communist countries, and under the military dictatorships of Latin America in the twentieth century). Beginning in the 1990s, successful experiences of systems of government centered upon parliaments have been adopted in various regions of the world. In Eastern European countries, the West European parliamentary model of political organization has been widely adopted, though it runs up against the tendency to fall back on a powerful Head of State with a presidential profile that is a legacy of the centralizing tendency of a communist past which lasted from 1945 to 1990. In Latin America, with the restoration of democracy that began in the 1980s, parliaments have taken on importance in a context clearly dominated by the presidential models of the North American type. In Asia, Japan, and India active parliaments have been the rule since the late 1940s. In Africa, although practically all countries have parliamentary institutions, the parliamentary model, whether active or deliberative, remains incipient and the role of parliaments largely irrelevant.

4. Future Prospects

Issues discussed when examining the history of parliaments fall largely into three groups: the external environment; domestic characteristics; and the political characteristics of parliaments. The external environment includes efforts to articulate four trends: the constitutional structure of parliaments, the administrative structure, the (political or electoral) system upon which they are based, and the interest groups with which they must interact. Their domestic characteristics comprise five factors that are decisive for the process of constituting parliaments: their members (their social and professional background, political experience and aspirations, and their degree of commitment to the legislative branch); parliamentary parties, committees, and houses as operational institutions, relations between members of parliament and their constituents as a reflection of social, economic, and cultural structures. The characteristics of their politics open up the possibility of studying parliaments as forums where the repercussions of short-, medium-, and long-term structures are felt and the way in which they reflect the dynamics of political decisions.

There are four principal trends in the historical research on modern parliaments: recovery and systemization of sources; sociological analysis of the

composition of parliamentary representation; a physiological examination of parliamentary speeches; analysis of the role of parliaments in public administration (the formation and control of governments, monitoring and assessment of public policies, and taking of measurements among the voters and public institutions). A great portion of these trends is typically European. In Great Britain and in Germany, for example, recovery of sources is an important approach when placing parliaments within a historical context. The same is true in Brazil, Canada, and Australia. In all these cases, this monumental work has been linked directly to the consolidation of their respective identities over time and to the legitimization of the democratic structures of the State. However, as also occurs in other walks of life, historical research trends evolve.

Not so long ago it was possible to write on the history of parliaments by concentrating solely upon formal texts. Causes and developments would be carefully taken into account, accounts taken down in writing in accordance with the level of detail required, and the consequences for political parties and other significant players in the political processes analyzed. In many of the older texts, an implicit (and sometimes explicit) premise recurring throughout the explanation of the causes and consequences of parliamentary life was that it was possible to chart their progress. History reveals the emergence of parliamentary democracy, as legislation expands the electorate (in accordance with the results of the census, especially in the nineteenth century), to a certain degree making the vote more equal and elections more 'honest.' It was perfectly possible to explore democracy through reference to certain key political dates in each country. Thus the British system came to be regarded as a paradigm by others, as it ensures mechanisms which guarantee the peaceful transfer of power.

Such emphases have been radically altered. First, historical research, in a more skeptical and less self-confident age, no longer seeks to chart the progress from a smaller to a larger state, not least because society at the end of the twentieth century has become much more ambivalent about the building of progress. Second, because much more emphasis had been devoted to continuity than to change (especially with respect to representative democracy). Third, research efforts (polling) at the constituency level have greatly increased. It is hardly surprising that such work produces revelations that do not support the progressive view. It could be argued that the nature of the political process has undergone changes for the worse, from the viewpoint of the lower classes. The way in which governments manipulate electoral laws and the rules defining constituencies and how majorities are defined (for example, proportional systems, simple majorities or mixed systems) have become relevant research themes in the field of historical research, the results of which are only now coming to light.

Furthermore, historians have begun to question the importance of the vote for participation in the political process. Indeed, a well-substantiated demand addressed to the administration is likely to be more effective than heated political speeches supported by a few votes in a legislative body. The parallel with nineteenth-century Brazilian politics (and the process of overthrowing dictatorships in the twentieth century, particularly in Latin America) are cases in point. As O'Gorman and Vernon showed for Great Britain, lack of voter participation has not been an obstacle to political participation. Members of Parliament must take a broad spectrum of interests into account if they wish to remain on good terms with their constituents. The right to vote does not necessarily bring with it access to political power. The suffragettes of the early twentieth century produced brilliant political propaganda, showing that educated, intelligent women with an interest in public affairs were denied access to the vote, whereas ignorant, vulgar, and drunken men were free to participate in it. In reality, such campaigns had a reach far beyond the mere exercise of the ballot: they were an affirmation of the importance of women in a stable political system.

Interest in electoral policies has been reinforced by detailed studies of elections, supported by long-term statistics. Such statistics have used materials that historians immediately recognize as being prime raw material for such studies. Voter registration records prior to the introduction of the secret ballot in Great Britain in 1872, for example, show who voted for whom. Such surveys make it possible to monitor changes in voting habits and to identify which candidates or parties lost favor with different social groups. The 'social geography' of elections and the 'sociological premises' of a given society thus become important instruments for historical analysis. Historical research brings together both the quantitative and qualitative techniques of the social sciences to obtain relevant results for the analysis of parliaments.

Political studies relating directly to social and occupational structures also touch upon two other lines of work carried out over the past 40 years. 'History from the bottom up' was in vogue in the 1960s and 1970s, when many researchers began to devote themselves increasingly to social structures. Attempts to discover the origins of the 'working class' and, somewhat later, of the 'middle class' involved deep social analyses. Historians more sensitive to issues of social class admit that political awareness plays a part in the issue. Indeed, class consciousness is related directly to the exercise and pursuit of power and in turn requires the existence of political structures that need to be changed. Nonetheless, in Europe, as of the early nineteenth century, debate on parliamentary reform became enmeshed in discussions on social class and on changes in the 'political culture.'

Another important intellectual development has been emergence of the so-called 'linguistic turn' which

began to influence historians in the 1980s. Detailed studies of language as a vehicle for conveying ideas, conducted by historians, have been inspired by radically new ideas borrowed from literary criticism. The relevance, and even the validity, of such an approach has engendered much debate among historians versed in the empirical tradition. Many are reluctant to accept that language can build its own realities, rather than having been invented to serve economic, social, or cultural needs. Unquestionably, extreme versions of poststructuralism, which claim that a text has no validity except in its own context, undermine any type of historical investigation in that they do not accept the notion of relative importance and suggest that change cannot be explained by intentional 'privilege' attributed to particular parts of historical documents. The 'linguistic turn' causes historians to be even more strict in the control of their premises on parliamentary discourse as a historical source. Language may effectively be used to mold ideas. Studies on demagoguery, populism, and '*caudilismo*,' and their influence upon the formation of parliaments in such countries as Brazil, Mexico, and Argentina, are pioneers in this field. The study of the language used by the Chartist reformers of parliamentary democracy in Great Britain, for example, shows that their real goal was political change.

A third theme relating to the expansion of the democratic process is treated under the term 'progress.' Research shows clearly that, when it comes to political change, progress is neither simple nor linear. Parliamentary history is related to, but not defined by, the history of democratization. 'Democratization' of itself is a highly problematic and contested concept. Nonvoters are not necessarily excluded from politics. No parliamentary reform ever 'grants power' to another social group—at any rate, not directly. Women do not all of a sudden hold power simply as a consequence of a scrap of paper called a 'law.' Current research trends show that the law is more likely to ratify a political order already in place than to introduce sweeping change. The recent development of supranational parliaments (the European Union, Mercosur) provides an entirely new field, into which investigative research is about to begin.

Historians have become more involved with continuities. Much of the historical research carried out in the 1980s and 1990s was accompanied by warnings about the dangers of economic reductionism, or of any other type of reductionism. However, the determination of historians to attempt to see things 'within their context' has contributed to the clarification of a series of relevant conclusions on the subject of parliamentary reform. They are now more aware that neither the 'top-down' explanations based upon the higher policies of central parliaments, nor the 'bottom-up' approaches that concentrate on extraparlimentary forms of pressure, are of themselves sufficient. The emphasis has swung to the need to understand the

intricate interrelations of the complex factors which created the political culture of the twentieth century and, above all, World War II. The changing methods and interpretations nonetheless confirm the central role of parliaments in this culture.

See also: Citizenship, Historical Development of; Democracy, History of; Parliamentary Government; Political History (History of Politics); Political Parties, History of; Public Sphere: Nineteenth- and Twentieth-century History

Bibliography

- Adamietz H 1978 *Parlamentsgeschichte*. Hanschild, Location
- Bélanger R, Jones R, Vallières M 1994 *Les grands débats parlementaires*. Publisher, Laval, Quebec
- Bisset A 1882 *A Short History of English Parliament*. Publisher, London
- Brooks C 1999 *Governors & Government. Political and Public History of Early Modern England 1550–1850*. Arnold, London
- Chacon V 1997 *História institucional do Senado*. Publisher, Brasília
- Copeland G W, Patterson S C (eds.) 1994 *Parliaments in the Modern World: Changing Institutions*. University of Michigan Press, Ann Arbor, MI
- Das Österreichische Parlament* 1984 Vienna
- Franks C E 1987 *The Parliament of Canada*. University of Toronto Press, Toronto, ON
- Fraser A, Mason R H P, Mitchell P 1995 *Japan's Early Parliaments 1890–1905*. Routledge, New York
- International Commission for the History of Representation and Parliamentary Institutions 1970 *Parliaments, Estates and Representation*. International Commission for the History of Representation and Parliamentary Institutions, Ashgate, London
- Interparliamentary Union 1983 *Les Parlements dans le monde*. Interparliamentary Union, Paris
- Kluxen K 1983 *Geschichte und Problematik des Parlamentarismus*. Frankfurt
- Kühne T 1998a *Parlamentsgeschichte in Deutschland: Probleme, Erträge, Perspektiven*. (History of Parliamentarianism in Germany) *Geschichte und Gesellschaft* 24(2): 323–38
- Kühne T 1998b *Parlamentsgeschichte in Deutschland: Probleme, Erträge, Perspektiven einer Gesamtdarstellung. Geschichte und Gesellschaft* 24(2): 323–38
- Kurian G Th (ed.) 1998 *World Encyclopedia of Parliaments and Legislatures*. Congressional Quarterly, Washington, DC
- Laundy P 1989 *Parliaments in the Modern World*. Gower, Dartmouth, UK
- Loewenberg G, Patterson S C 1988 *Comparing Legislatures*. University Press of America, Lanham, MD
- Olivier-Martin F 1997 *L'absolutisme français: suivi de Les parlements contre l'absolutisme traditionnel*. Librairie générale de droit et de jurisprudence, Paris
- Ritter G A (ed.) 1974 *Gesellschaft, Parlament und Regierung: zur Geschichte des Parlamentarismus in Deutschland*. Kommission für Geschichte des Parlamentarismus und der politischen Parteien, Düsseldorf

The History of Parliament 1998 (CD-Rom Edition: House of Commons). Cambridge, UK

The Journal of Legislative Studies 1995 Frank Cass, London

VV. AA. 1963–1984 *Storia del Parlamento Italiano*. S. F. Flaccovio, Palermo

E. C. de Rezende Martins

Parsons, Talcott (1902–79)

From the late 1940s to the mid-1960s, Talcott Parsons was the single most influential social theorist in the world. A developer of what is now popularly called 'action theory,' 'functionalism,' and 'structural-functionalism,' Parsons spent his entire career at Harvard University, which helped considerably in institutionalizing his ideas and also in providing him access to talented graduate students. By the 1950s, his publications became a major part of what was literally a 'common curriculum' for graduate students in all major sociology departments—as well as in many of the best political science departments. By being institutionalized in this way, Parsons' publications elevated the rigor of American graduate training in sociology in particular. More than anyone else, he also defined the 'classic' theoretical tradition for sociology as a discipline and then added a more abstract, arguably more conceptually sophisticated contemporary rival to it. With this, he raised the bar for social theory worldwide, from Germany, France, and Great Britain to Japan and even the former Soviet Union. A strong case can be made today that every major social theory since the mid-1960s has been developed in direct or indirect dialogue with Parsons' functionalism (see Habermas 1981 for a related statement).

Having become literally an icon in the discipline in the first three decades of his career, in the last two decades he attracted a considerable share of the iconoclasm that more generally characterized the student movement and academia of the late 1960s and early 1970s. Parsons was subject across two decades to far greater criticism than that directed to any other theorist of his generation or since. Given the tenor of the times, most criticisms were rhetorical, often personal. Today, they leave a strong impression that few critics had bothered to read his works with the care typically afforded to basic empirical studies, let alone with the rigor and dispassion that any complicated social theory demands. Yet, the cumulative effect of the broadsides he received in the late 1960s and early 1970s was to leave the collective memory of the discipline with an understanding of Parsonian functionalism that is difficult to reconcile with often-repeated, explicitly-stated positions in his publications. What Parsons endeavored to accomplish in his day, providing the social sciences with a common language, and how he went about accomplishing it,

developing a ‘conceptual framework’ comprising analytical distinctions, now eludes the discipline’s collective memory.

Talcott Parsons was born on December 13, 1902 in Colorado Springs, the youngest of six children of Edward S. Parsons, a Congregational minister and college administrator, and Mary Augusta Ingersoll Parsons, a housewife and active suffragist (on biographical details, I rely primarily on Nielsen 1991). Both parents were New England WASPs who traced their ancestry to mid-eighteenth century settlers, Mary to theologian Jonathan Edwards. A graduate of Yale Divinity School, Edward Parsons holds fundamentalist views of Christianity but leftist views in politics, particularly regarding the issue of economic inequality. He is an active participant in the Social Gospel movement that supported organized labor. Both parents inculcate into Talcott and his older siblings the importance of finding a mission in life and then pursuing it assiduously. Many observers comment later on the mature Talcott’s remarkable energy and perseverance—including on the final two days of his life in Munich in 1979. He published over 150 sole-authored articles and 14 books or collections.

In 1917 the Trustees of Colorado College force Edward to resign as Dean after he supports two women who charged the College’s President with sexual harassment. The family then moves to New York City where Talcott enters Horace Mann High School, a laboratory school for boys operated by Teacher’s College of Columbia University. Talcott earns an undergraduate degree at Amherst College (as had his father and two older brothers) from 1920–4. At the time Amherst was also an experimental, progressive school. Parsons at first focuses on biology and chemistry but in his junior year becomes ‘converted’ to the social sciences under the influence of an unorthodox institutional economist, Walton Hamilton.

Parsons’ graduate training was remarkably brief: one academic year (1924–5) at the London School of Economics as a nondegree candidate, then another (1925–6) as a degree candidate at Heidelberg University. At LSE, Parsons was influenced by Morris Ginsberg, L. T. Hobhouse, and the methodological functionalism of anthropologist Bronislaw Malinowski—as well as by fellow-student E. E. Evans-Pritchard. Parsons also met his future wife and mother of three children, Helen Banerott Walker, a Bryn Mawr College student studying banking. Her family of conservative white- and blue-collar workers consider Talcott a leftist, a ‘pinky.’ At Heidelberg, Parsons first learnt of the recently deceased German social theorist Max Weber and discussed his work with his wife Marianne, his younger brother Alfred, and Alexander von Schelting. Taking coursework from Edgar Salin and Emil Lederer (on economics), Karl Mannheim (a friend of Georg Lukacs) and Karl Jaspers (on Kant), Parsons selected how sociology and

economics differ in portraying ‘capitalism’ as his dissertation topic.

During his year at Heidelberg Parsons was offered a one-year teaching appointment at Amherst in the Economics Department which also allowed him to teach an independent course in sociology. In 1927 Parsons was offered an instructorship in economics at Harvard, thus beginning a life-long career there. Harvard’s Department of Sociology was a relative late-comer in the discipline, beginning in the fall 1931 under the chairmanship of Pitrim Sorokin, a 42-year old emigre from Russia recruited from University of Minnesota. Parsons, then 29, was appointed as an instructor. Disagreements and misunderstanding between the two theorists became common knowledge in the Department and across the University. Yet, throughout Parsons’ career, both critics and proponents often commented on his low-key demeanor when grappling with ideas and interacting with colleagues and students. It is difficult to find accounts of him being particularly effusive, let alone dominating. A short, stocky, already balding young man who sports a moustache and always smokes cigarettes, critics also comment often on his charm when interacting with other major national and international academicians at Harvard, in the American government, and overseas.

During his earliest years at Harvard, Parsons was impressed with philosopher Alfred North Whitehead and his notion of the ‘fallacy of misplaced concreteness.’ Consistent with Kant, Whitehead held that scientists apprehend ‘reality’ or scientific ‘truth’ only through analytical distinctions, not more immediately—whether experientially or directly through empirical findings. A few years later in correspondence with phenomenologist Alfred Schutz from 1940–1, Parsons defended the notion of ‘analytical realism’—as opposed to ontological realism or empirical realism—against Schutz’s position that social scientists can somehow gain more direct access to the ‘life-world’ of their subjects of study (Grathoff 1978).

By contrast to Schutz and then also to ethnographers at University of Chicago and American empirical researchers more generally, Parsons endeavored across his 50-year career to identify the most irreducible analytical distinctions unique to sociology, the scientific study of social life. In pursuing this project, Parsons arrived at three successive ‘theoretical syntheses’ and along the way trains four remarkably talented and productive cohorts of graduate students. All three of Parsons’ theories, which he called ‘frameworks of concepts,’ involve a ‘functionalist’ approach to the study of social life. This is an approach much maligned today, and yet a basic premise of functionalism is hardly controversial. Any social scientist who poses the following question is operating within a functionalist approach: What is the relationship, if any, between the substantive area of social life I am studying and the direction of social change? Here,

then, is a common ground shared by critical theorists and Parsonian functionalists but not by more historicist researchers, those who study particular events or periods in isolation and are loathe to identify any direction of social change.

While exceedingly abstract and complex, there are two ways to grasp the significance of Parsons' publications across his career. In general, he endeavored to account for the place and purpose of norms in maintaining social order both historically and cross-nationally. More particularly, he endeavored to account for the place and larger significance of professions, a unique set of occupations, in contemporary societies. Professions are a pivotal subject for sociological inquiry because they are important nongovernmental bodies in all modern societies. Yet, their prestige or status in any society hinges as much on whether they are valued culturally, supported by generally recognized social norms, and centrally situated in a social structure as on whether practicing professionals compete effectively in self-regulating markets. In all modern societies, professionals are simultaneously driven by economic pressures to maximize profits, like other workers, and also by normative pressures to conform to certain extra-economic standards of conduct that typically do not bind most other workers. Parsons was interested in accounting at a theoretical level for the rise, evolution, and institutionalization of these extra-economic restraints on self-interested behavior, and then also to explore whether and how they contribute to social order.

In his earliest publications, from the 1920s to the mid-1940s, Parsons approached professions by first distinguishing their behavior from that of economic enterprises, on one side, and religious organizations, on the other. Corporate managers and shareholders tend to act in strictly utilitarian or market-mimicking ways as they endeavor to maximize either growth or profit, a quantitative end that can be recognized interpersonally. Clerics and religious believers, however, tend to exhibit ritualistic fidelity to norms as they—presumably—seek spiritual salvation. This is a qualitative end that is transcendental or metaphysical; as such, its attainment is a matter of faith, not something that can be recognized interpersonally. Professionals, Parsons held, also exhibit fidelity to norms, but more 'voluntaristically.' They do not act in strictly market-mimicking ways, but they also do not conform to norms ritualistically with any transcendental end in view. Rather, they exhibit fidelity to norms as a means to attain qualitative ends that are worldly or empirical, and thus capable of being recognized interpersonally—such as their patients' physical or mental health, or their clients' legal innocence, or scientific truth.

Parsons came to appreciate the significance of 'voluntaristic action' after a careful reading of two economic theorists, Alfred Marshall and Vilfredo Pareto, and two social theorists, Emile Durkheim and

Max Weber. This reading forms the core of *The Structure of Social Action* (1937) which, along with Parsons's 1930 translation of Weber's *The Protestant Ethic and the Spirit of Capitalism*, established Weber and Durkheim as two of the new discipline's 'classic' theorists. During the 1930s, Parsons trained a first cohort of graduate students in the classics and his own theory of 'voluntaristic action,' including: Robert Merton, Kingsley Davis, John and Mathilda Riley, Edward Devereux, Marion Levy, Wilbert Moore, and Florence Kluckhohn (along with Edward Hartshorne who was killed by a sniper while on active service in postwar Germany).

In spring 1944 Parsons was promoted to full professor and the chairmanship. By January 1946, he transformed the Department of Sociology into the Department of Social Relations, with the explicit aim of linking structural and social psychological approaches in the scientific study of social life. In the same year, Parsons also began formal training in psychoanalysis as a Class C candidate at the Boston Psychoanalytic Institute. He had been lecturing informally on Freud since fall 1938.

As Parsons assumed his leadership role in Harvard's Department, his theory went through a brief mid-career change from the late 1940s to the mid-1950s as he considered findings from Robert Bales's study of small groups. Parsons began thinking about professions (and how to identify norms in social life more generally) in terms of six basic decisions that people performing any role or task either make explicitly or have imposed on them by cultural values, institutionalized norms, or social structures. For example, certain roles (including those performed by professionals) mandate treating others (clients or patients) in universalistic ways. By contrast, other roles (including those performed by parents) encourage more particularistic treatment of others (children). Parsons called these decisions and requirements 'pattern variables.' By his account, researchers can use six such variable-pairs (a) to distinguish professions from other occupations and (b) to identify changes in the behavior of professions both historically and cross-nationally. Parsons' major publications during this period are a collection, *Essays in Sociological Theory Pure and Applied* (1949), and three monographs: *The Social System* (1951), *Towards a General Theory of Action* (1951, with Shils), and *Working Papers in the Theory of Action* (1953, with Robert Bales and Edward Shils).

During this period Parsons trained a second cohort of graduate students in the pattern variables, including: Bernard Barber, Harry Johnson, Robin Williams, Jr., Jesse Pitts, Harold Garfinkel, Francis Sutton, and Robert Bales. More generally, as part of the buoyancy and optimism that pervaded the US after the war, many sociology departments re-evaluate their curricula. The discipline is open collectively to receiving a new theoretical synthesis, and Parsons' notion of pattern variables, while preliminary and as it

turned out fleeting, is a prime candidate. Another is Robert Merton's notion of 'middle range theory,' an idea he first proposed at the annual meeting of the then American Sociological Society in 1947 in opposition to Parsons' 'general theory.'

Finally, beginning slowly in the mid-1950s and then with more rapid developments from the early 1960s to the end of his career, Parsons approached professions within what he called the 'AGIL schema.' This is Parsons' single most significant contribution to social theory. During his transition from the pattern variables to the AGIL schema, Parsons' publications, lectures, and presentations create an unusually kinetic intellectual excitement among collaborators and graduate students. Everyone working in his circle operated with a palpable sense that social theory is in the midst of a fundamental breakthrough, a sense that American sociology has not witnessed since.

This transition period to Parsons' mature social theory is marked in particular by *Economy and Society* (1956, with Neil Smelser), and Parsons' training of a third cohort of graduate students, including: Albert Cohen, Renee Fox, Joseph Berger, Norman Birnbaum, Neil Smelser, James Olds, Jackson Toby, and Miriam Johnson. In addition, many earlier graduate students whose training had been interrupted by wartime service returned to Harvard. Talented undergraduates who come into contact with Parsons during this period include Robert Bellah and Charles Tilly. Also in transition during this period is the leadership of sociology at Harvard. Parsons resigns as chair of the Department of Social Relations in 1956, and within the next two years learns that he has diabetes. Sorokin retires in 1959.

From 1956 to 1961, Lewis Coser and Dennis Wrong in the US, David Lockwood in Great Britain, and Ralf Dahrendorf in Germany challenged Parsons' functionalism in relatively measured tones, but with uneven rigor. They were also shooting at a moving target in that Parsons was developing his mature social theory. The high-point of scholarly criticism during this period comes with a collection of essays by philosophers and sociologists edited by Max Black, to which Parsons responded with a long concluding chapter. In April 1964, Parsons was a central figure at a famous session of the German Sociology Association in Heidelberg devoted to reconsidering Max Weber's contributions. He found himself at odds with Frankfurt School critical theorists, most notably the young Jürgen Habermas and the more established Max Horkheimer and Herbert Marcuse (see Stammer 1965/1972 for a collection of these exchanges).

From this point forward, criticisms of Parsons become harsher, more personal and ideological than analytical and scholarly. Parsons was attacked, particularly in the US, for: his religious background, his WASP ethnic and cultural heritage, his writing style, his amazing productivity, his putatively strategic rather than scholarly decision to make sociology as

rigorous conceptually as economics, his putative mistranslation of Weber, the putative conservative nature of his own family life including his putative responsibility for his daughter's suicide (in June 1964, two weeks after the Heidelberg debate), his putative political conservatism, and his putative general obedience to 'American capitalism' or 'the establishment.'

Parsons unveiled the AGIL schema explicitly in 1963, with major articles on the concepts of power and influence, and then in 1964, with his first statement about 'evolutionary universals' (see the Turner 1999 collection for the 1964 article and references to the others). As fully developed, the AGIL schema isolates analytically those aspects of behavior in any role or position that contribute, respectively, to each of four general social 'functions': economic efficiency ('adaptation'), administrative effectiveness ('goal-attainment'), fidelity to social norms ('integration'), and fidelity to cultural values or a society's most basic institutional arrangements ('latency'). Parsons also proposed that four 'media of interchange' circulate between these analytical subsystems—money, power, influence, and value commitments—thereby bringing a certain orderliness or predictability to the whole. He added notions of 'systems-theory,' 'pattern maintenance,' and 'hierarchy of control' to this basic four-function breakdown of social life.

In 1973, Parsons and Platt published what is arguably his most important single work since 1937, *The American University*. In this work Parsons re-thinks his approach to professions by incorporating two major points into his social theory. First, he proposes that professions and sites of professional practice (such as universities, hospitals, and research institutes) are organized in a 'collegial form,' not in a bureaucratic form or a democratic form. Second, he proposes that professions are distinguished from other occupations by their members' willingness to bear 'fiduciary responsibilities' to clients and the larger society. These two points provide the foundations for an account of the place and purpose of professions in contemporary societies that differs radically from any effort, whether by rational choice theorists or social critics on the left, to treat the professions analytically as interchangeable with other occupations or other corporate entities. By 1973, however, the collective memory of the discipline is so dominated by criticisms of the AGIL schema that this major work, arguably Parsons' single most important substantive contribution to social theory, goes largely unread by critics and proponents alike.

Through the 1960s and 1970s, Parsons trained a fourth and final cohort of graduate students in the AGIL schema and related theoretical developments, including: Victor Lidz, Jan Loubser, Leon Mayhew, Rainer Baum, Dean Gerstein, John Akula, and Willy DeCramer. Parsons also influences undergraduate Jeffrey Alexander. Only in the mid-1970s, a few years before his death, did Parsons begin methodically to

respond to at least a few criticisms of his work, most notably the informed analyses of Bershadsky (1973) and Warner (1978). Bershadsky's *Ideology and Social Knowledge* in particular marks a major sea change in commentary on Parsons' works. Bershadsky is the first to draw attention to Parsons' effort to ground the social sciences against relativism with analytical concepts. More than anyone else Bershadsky appreciates that Parsons pitched his social theory at an unusually abstract level of analysis because he sought literally a common language for the social sciences. Prior to Bershadsky both proponents and critics generally failed to convey the power and sweep of Parsons' social theory in form, its aspiration to unite the social sciences. Parsons' project was essentially to give structuralists and symbolic interactionists the means, the common language, by which to demonstrate to each other the central import of their respective findings.

Parsons demonstrated his capacity to 'translate' others' findings and ideas in his own scholarship. He co-authored, team-taught, or otherwise collaborated closely with a remarkable range of theorists and researchers across disciplines, including: Edward Shils, Joseph Schumpeter, Samuel Stouffer, Florence Kluckhohn, Robert F. Bales, James Olds, Morris Zelditch, Philip Slater, Neil Smelser, Lon Fuller, Kaspar Naegle, Jesse Pitts, Winston White, S. N. Eisenstadt, Robert Bellah, Victor Lidz, Gerald Platt, Edward Laumann, Erwin Scheuch, Andrew Effrat, Mark Gould, and Dean Gerstein. In addition, he engaged in private and public correspondence and debates with Frank Knight, Crane Brinton, Robert Bierstedt, Albert Blumenthal, Alfred Schutz, Chester Barnard, Philip Selznick, Eric Voegelin, Robert Merton, Kenneth Boulding, C. Wright Mills, Robert Bellah, Bennett Berger, David Riesman, Jurgen Habermas, Herbert Marcuse, Max Horkheimer, and Ernest Mayr. Finally, an impressive set of scholars was influenced heavily by Parsons himself, his students, or his theory (some of whom also influence Parsons and his students): Niklas Luhmann, Seymour Martin Lipset, Amitai Etzioni, Robert Marsh, Daniel Bell, Joseph Ben-David, Benjamin Nelson, Gabriel Almond, James S. Coleman, Karl Deutsch, David Apter, Lucian Pye, Sidney Verba, Chalmers Johnson, S. N. Eisenstadt, Ken'ichi Tominaga (Japan's leading postwar sociologist), Lawrence Brownstein, Martin Martel, Adrian Hayes, and Frank Lechner.

Parsons and Helen travelled frequently in the 1970s, including three separate trips to Japan. In early May 1979 they returned to Heidelberg, then proceeded to Munich, on the occasion of the fiftieth anniversary of Talcott's degree from Heidelberg. After Parsons' typical full day of presentations and scholarly exchanges, including a lecture on 'The Declining Significance of Social Class' attended by host Horst Helle as well as Jurgen Habermas, Niklas Luhmann, Richard Munch, and Wolfgang Schluchter, Parsons died of a stroke in the early hours of May 8. At a

memorial service 10 days later in Harvard Chapel, Victor Lidz notes in eulogy: 'No sociologist of recent times has had to endure more bitter criticism than Talcott. Although a passionate man, Talcott bore criticism with equanimity as well as courage. Not daunted by even sweeping attack, he held to his own program of research and writing. Moreover, adhering to principles of value-freedom, he declined to reciprocate ideological criticism. He dealt with criticisms, always, as technical matters within the domain of social scientific theory' (Nielsen 1991).

Parsons' critics inadvertently encouraged the fragmentation of research and theory in sociology and the social sciences that Parsons endeavored to prevent. Today, in the absence of any shared framework of analytical concepts, social scientists routinely talk past each other, as opposed to revealing to each other the general importance of their research findings and theoretical developments. Parsons's 'general theory' has given way not to a finite set of readily identifiable theories of the 'middle range,' as Robert Merton anticipated in 1949, but rather to an ever more finely grained cacophony of research specialties. This trend is then recapitulated, and accelerated, by ongoing changes in graduate training. Graduate students are trained more and more narrowly, and their required survey courses in social theory hardly encourage them to move beyond any research specialty. If there is any 'common language' in the social sciences today, it is that of research methods. But even here there is more internal division among methodologists than they often wish to acknowledge.

In support of Habermas' point that we can expect any new theoretical development in the social sciences to respond directly or indirectly to Parsons' functionalism, French social theorist Pierre Bourdieu in many regards today recapitulates Parsons' AGIL schema. He identifies two poles within what he calls 'fields of power'—the economic field and the field of cultural production—and then distinguishes four fields in between: politics, the higher civil service, the professions, and the university (Bourdieu 1989/1996). The major difference with Parsons is that the French theorist's categories are more directly descriptive, less analytically rigorous. As a result, it is not clear on what basis he is distinguishing fields, why there are six fields of power rather than four or eight, or on what basis researchers may categorize particular groups or activities as components of one field or another.

By the early 1980s, a new generation of social theorists, including members of Parsons' last cohort of graduate and undergraduate students at Harvard, initiated the effort to have Parsons' publications reconsidered on scholarly grounds. Led initially by Munch (1981) in Germany and Alexander (1983, 1985) in the US, today's 'neofunctionalism' finds support from theorists whose political positions range from radical left to liberal reformist to conservative republican (see the collections edited by Hamilton 1992

and Colomy 1990). What they share, with Parsons himself, are two general principles (which Alexander 1988 now explicitly rejects). One is that it is vitally important to the social sciences to continue Parsons' effort to ground social theory on the basic analytical distinctions of the AGIL schema rather than to rely more directly on descriptive concepts (whether empirical generalizations or ideal types). The other principle is that it is incumbent today to demonstrate the empirical potential of Parsonian functionalism, to identify lines of empirical inquiry that this social theory uniquely presents to view (Sciulli 1992). A strong case can be made that Parsons' social theory has been mined only superficially (Takagi in press). An equally strong case can be made that theorists and researchers today can draw fruitfully on Parsons' basic AGIL schema, and then the first set of subdivisions he drew within each 'function,' but that they need not follow Parsons' effort to draw further distinctions within each subdivision (Mouzelis 1995).

Social theorists outside the US today who support the Parsonian tradition, even if critical of certain parts of it, include: Munch, Hans Joas, Uta Gerhardt, Horst Helle, and Harald Wenzel in Germany; Helmut Staubmann in Austria; Jens Kaalhaug Nielsen in Denmark; Pierpaolo Donati, Andrea Maccarini, Matteo Bortolini, and Riccardo Prandini in Italy; Ken'ichi Tominaga, Kiyomitsu Yui, and Kazuyoshi Takagi in Japan; and Nicos Mouzelis, Bryan Turner, and Jeremy Tanner in Great Britain.

See also: Action Theory: Psychological; Action, Theories of Social; Control: Social; Conventions and Norms: Philosophical Aspects; Durkheim, Emile (1858–1917); Educational Institutions and Society; Emergent Properties; Functionalism, History of; Functionalism in Sociology; Influence: Social; Institutions; Norms; Political Economy, History of; Professions in Organizations; Professions, Sociology of; Realism/Neorealism; Realisms and their Opponents: Philosophical Aspects; Shared Belief; Social Change: Types; Sociology, History of; Status and Role: Structural Aspects; Structuralism, Theories of; Structure: Social; Symbolic Interaction: Methodology; System: Social; Theory: Conceptions in the Social Sciences; Theory: Sociological; Value Pluralism; Values, Sociology of; Weber, Max (1864–1920)

Bibliography

- Alexander J C 1983 *Theoretical Logic in Sociology, Vol. 4: The Reconstruction of Classical Thought: Talcott Parsons*. University of California Press, Berkeley, CA
 Alexander J C (ed.) 1985 *Neofunctionalism*. Sage, Beverly Hills, CA
 Alexander J C 1988 *Neofunctionalism and After*. Blackwell, Malden, MA

- Bershadsky H J 1973 *Ideology and Social Knowledge*. Wiley, New York
 Black M (ed.) 1961 *The Social Theories of Talcott Parsons*. Prentice-Hall, Englewood Cliffs, NJ
 Bourdieu P 1989/1996 *The State Nobility: Elite Schools in the Field of Power*. Stanford University Press, Stanford, CA
 Colomy P (ed.) 1990 *Neofunctionalist Sociology*. Elgar, Aldershot, UK
 Grathoff R (ed.) 1978 *The Theory of Social Action: The Correspondence of Alfred Schutz and Talcott Parsons*. Indiana University Press, Bloomington, IN
 Habermas J 1981/1987 *Theory of Communicative Action, Vol. 2. Lifeworld and System: A Critique of Functionalist Reason*. Beacon Press, Boston
 Hamilton P (ed.) 1992 *Talcott Parsons: Critical Assessments*, 4 vols. Routledge, London
 Mouzelis N 1995 *Sociological Theory: What Went Wrong? Diagnosis and Remedies*. Routledge, London
 Munch R 1981 Talcott Parsons and the theory of action, I: The Structure of the Kantian core. *American Journal of Sociology* 86: 709–39
 Nielsen J K 1991 Talcott Parsons' Life, Theory and Political Orientation. Unpublished Manuscript, 5 vols.
 Parsons T 1937/1968 *The Structure of Social Action*, 2 vols. Free Press, New York
 Parsons T, Platt G W 1973 *The American University*. Harvard University Press, Cambridge, MA
 Sciulli D 1992 *Theory of Societal Constitutionalism: Foundations of a Non-Marxist Critical Theory*. Cambridge University Press, New York
 Stammer O (ed.) 1965/1972 *Max Weber and Sociology Today*. Harper and Row, New York
 Takagi K in press *Talcott Parsons and American Intellectual Society*. Cambridge University Press, Cambridge, UK
 Turner B S (ed.) 1999 *The Talcott Parsons Reader*. Blackwell, Malden, MA
 Warner R S 1978 Toward a redefinition of action theory: Paying the cognitive element its due. *American Journal of Sociology* 83: 1317–49

D. Sciulli

Partial Orders

The idea of ordering captures a basic faculty of the human mind: to decide for two given elements from a set of objects which one 'dominates' the other with respect to an attribute of interest, or, alternatively, which one is 'preferred' to the other.

Formally, a partially ordered set (a poset for short) consists of a pair $(P, \leq)$ where P is a nonempty set and $\leq$ a binary relation on P satisfying for all $x, y, z \in P$:

- (a) reflexivity: $x \leq x$,
- (b) asymmetry: $x \leq y$ and $y \leq x$ implies $x = y$,
- (c) transitivity: $x \leq y$ and $y \leq z$ implies $x \leq z$.

Transitivity is, in a sense, the most important property; it is shared by all variants of order relations, though it might be a consequence of other conditions. However, asymmetry also contributes in an essential way to the meaning of a partial order. Reflexivity, on the contrary, is merely a matter of taste or con-

venience. In some cases its opposite, irreflexivity, is postulated. Such an order is called *strict*. If the asymmetry condition is replaced by symmetry, i.e., for all $x, y \in P$ $x \leq y$ implies $y \leq x$, then the resulting structure is an equivalence relation which lacks the interpretation of ‘dominance’ or ‘preference.’ Such a structure is more akin to classification or grouping with respect to similarity.

1. Related Notions and Diagrams

Sometimes asymmetry is simply dropped from the conditions. A relation satisfying reflexivity and transitivity is called a *quasiorder*. The investigation of quasiorders can be split into the respective properties of a partial order and an equivalence relation in the following way: Let $(Q, \lesssim)$ denote a quasiorder. Define $x \sim y$ if $x \lesssim y$ and $y \lesssim x$ both hold true. Then it is readily shown that $\sim$ is an equivalence relation on Q . Let P be the set of equivalence classes and define for $p, q \in P$ the relation $p \leq q$ if $x \lesssim y$ for some $x \in p$ and $y \in q$. Then, one demonstrates two things: first, that the definition of $\leq$ is independent of the particular elements $x \in p$ and $y \in q$; and second, that $(P, \leq)$ is a poset. Thus, the ‘dominance-aspect’ of a quasiorder is contained in $\leq$, and its ‘classification-aspect’ is hidden in $\sim$.

One property of an intuitive meaning of dominance is not captured in the notion of a partial order: the idea that any two elements can be compared with respect to $\leq$. The defining conditions of a poset do not imply connectivity, i.e.,

$$x \leq y \text{ or } y \leq x \text{ for all } x, y \in P.$$

(Note: often this property is called ‘completeness,’ but later in this article another notion of completeness is introduced). If connectivity is added to the properties of a partial order $(P, \leq)$, then $\leq$ is called a linear or total order. A quasiorder satisfying connectivity is called a weak order. Applying the above procedure on a weak order results in a linear order on the equivalence classes of the weak order.

There is a tendency to associate the property of connectivity automatically with an order given in some field of application, e.g., if one rank orders stimuli or subjects. Doing this amounts to representing the order numerically, i.e., to assuming a function

$$f: P \rightarrow \mathbb{R} \text{ such that } x \leq y \text{ if } f(x) \leq f(y). \quad (1)$$

If such a representation is possible, the order is necessarily a weak order. However, in social science, numerical representations are used widely even for orders which cannot easily be justified to be connected. For example, if the items of an achievement test are



Figure 1

The forbidden substructures of a semiorder

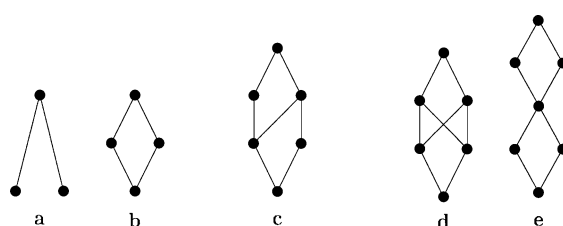


Figure 2

Lattices (b, c, e) and non-lattices (a, d)

‘Rasch-scaled’ (see the discussion following Eqn. (3)), a difficulty coefficient is attached to each item. This numerical scale orders the items weakly and ascribes the often made observation that subjects solve a difficult item correctly and fail at an easier one to influences of chance and error. This interpretation appears so obvious that its proponents tend to overlook the *petitio principii* which it entails, that the difficulty of an item is defined by the numerical parameter, whereas the weak ordering of the items according to a material notion of difficulty is an empirical property which will in many cases not be fulfilled.

For example, an expert judging this might well regard some items as incomparable. In such a situation a partial- (or quasi-)ordering, which allows for incomparable items, is more flexible. Thus, the assignment of numerical values, albeit ubiquitous in psychological tests, restricts drastically the range of possible models. Even partial orders can be too restrictive because transitivity is also a strong condition. The theory of knowledge spaces takes account of this aspect (see *Knowledge Spaces* for a deeper treatment of these issues).

The most convenient way of representing a poset is by its (*Hasse*) *diagram* (see Figs. 1 and 2). To construct it, one first defines $x < y$, in words ‘ y covers x ,’ to mean: $x < z \leq y$ implies $z = y$. Next, one draws a graph in which the vertices correspond to the elements of P , two elements are connected by an edge if one of them covers the other, and furthermore the covering element is placed above the horizontal line determined

by the covered element. Hasse diagrams are easy to draw and are of great help in theoretical considerations as well as in interpreting empirical findings.

2. Special Classes of Orders

There are numerous specializations, that is, additions of further properties, such that the resulting structure is something in between a poset and a linear order. In this section, two classes are considered which have useful numerical representations, although not the strong one of Eqn. (1). The third specialization considered here, namely lattices, is important because it tends to bridge the gap between ordering relations and algebraic operations.

2.1 Semiorders

In psychophysics one has to deal with the phenomenon of subliminal changes in intensity of physical stimuli. One stimulus may be (physically) slightly more intense than another, but a subject does not perceive the difference. For example, one might imagine a coffee-taster confronted with an extended series of cups of coffee, each with one more grain of sugar than its immediate predecessor. He/she will probably not notice the difference in taste between successive cups, but surely the difference between the first and the last cup will be clearly noticeable. In other words, indifference with respect to $\leq$ is not transitive. Similar observations can be made in decision theory and other branches of psychology. One way of describing such facts is to introduce sensory thresholds, thereby creating a scaling problem where one wants to scale both the subjective intensity and the magnitude of the threshold. Luce (1956) considered two conditions which seem plausible in this context (P is the set of stimuli and $x \leq y$ denotes the observation x is less intense than y):

(S1) If $x \leq y$ and $u \leq v$, then $x \leq v$ or $u \leq y$.

(S2) If $x \leq y \leq z$ and $w \in P$, then $w \leq z$ on $x \leq w$.

One easily shows that $(P, \leq)$ is a quasiorder, i.e., without losing generality, one can assume that it is a partial order. Condition (S1) says that it does not contain Fig. 1(a) as a suborder. Similarly, condition (S2) is equivalent to the fact that it does not contain Fig. 1(b) as a suborder.

From the viewpoint of measurement the most important property of semiorders is the possibility of a numerical representation with a constant threshold. More precisely, one can show that for a finite semiorder $(P, \leq)$ there is a function f

$$f: P \rightarrow \mathbb{R}: \text{ such that } x < y \text{ if } f(x) + 1 < f(y). \quad (2)$$

The condition of finiteness is far too restrictive for Eqn. (2), but without further restrictions it does not obtain. The height of the threshold is arbitrary, that is, the number one in Eqn. (2) can be replaced by any other positive number (with a rescaled f). More on this kind of representation theorem can be found in Roberts (1979, Chap. 6.1) or Suppes et al. (1989, Chap. 16). Clearly, the converse of this theorem is true: a structure with a representation, Eqn. (2) is necessarily a semiorder.

Although the semiorder concept seems to have its origin in the psychological context of the possibility of measurement of subjective intensity described above, it is now a well-established notion in the mathematical theory of posets and several remarkable results are connected with it. Some of them are mentioned in the following text.

2.2 Interval Orders

A structure satisfying only condition (S1) but not necessarily (S2) is called an interval order. This notion derives from the numerical representation which is characteristic of this kind of order. Each element can be represented by an interval on the real line and $x < y$ or $y < x$ is tantamount to the nonoverlapping of the respective intervals. Put another way, a scale value and a threshold are attached to each element, but the height of the threshold can vary in contrast to Eqn. (2). The reader will find many interesting properties and proofs in Fishburn (1985).

Historically, it is of interest that, although the representation theorem is due to Fishburn (1970), the investigation of this structure goes back to papers of Norbert Wiener, early, in the twentieth century. What is even more remarkable is that in this work he mentioned possible applications in experimental psychology (see Fishburn and Monjardet 1992).

The theory of interval orders has applications in archeology in connection with the seriation problem. In this case P is a set of artifacts, each element of which was some time in use. From excavations in various historical sites it is known for any two artifacts whether the period of use of one follows that of another with or without overlap. If this relation satisfies (S1) then, by the interval order representation theorem, it follows that it is possible to assign a time interval to each artifact, preserving the empirical observations.

A related notion is that of a biorder. The primitives of a biorder consist of two sets P and Q and a relation ρ between P and Q , i.e., a subset $\rho \subseteq P \times Q$ satisfying for all $x, y \in P$ and $u, v \in Q$

$$x\rho u, y\rho v, \text{ implies } x\rho v \text{ or } y\rho u. \quad (3)$$

To understand this condition, it is helpful to think of P as a set of subjects and of Q as the items of an

achievement test. xpu means ‘subject x fails to solve item u .’ The premise of Eqn. (3) says that x fails u and y fails u : if one assumes that x solves v correctly, then y has to fail u . With this interpretation, the biorder concept is the basis of Guttman scaling. Probabilistic modifications of it underly Rasch scaling, which is well-established in the theory of psychological tests. Thus, with the biorder condition Eqn. (3), a testable property is formulated which, when satisfied, justifies a numerical difficulty value for items because a numerical representation of a biorder consists of two functions $f: P \rightarrow \mathbb{R}$ and $g: Q \rightarrow \mathbb{R}$ such that:

$$xpu \text{ if } f(x) < g(y).$$

This, in turn, implies Eqn. (3). In the terminology of the above interpretation of a biorder $f(x)$ is the ability of subject x and $g(u)$ is the difficulty of item u . If in a biorder, the two sets coincide, $P = Q$, then Eqn. (3) reduces to (S1), i.e., the structure is an interval order.

2.3 Lattices

In a partial order $(P, \leq)$ subsets of P may or may not have upper or lower bounds, where upper bounds are elements which dominate all elements of the subset and lower bounds are elements which are dominated by all its elements. If each two-element subset has a unique smallest upper bound—called the supremum—and a unique greatest lower bound—called the infimum—then the structure is called a lattice. The infimum of x and y is denoted by $x \wedge y$ and the supremum by $x \vee y$. Examples of lattices and non-lattices are given in Fig. 2.

Further examples are the power set of a set ordered by set inclusion, any subset of the reals naturally ordered, and the natural numbers $\{1, 2, 3, \dots\}$ ordered by divisibility, i.e., $m \leq n$ if m divides n . In the last example, the least common multiple and the greatest common divisor are the supremum and the infimum, respectively. It follows easily by induction that any finite subset of a lattice P has an infimum and a supremum. However, infinite subsets do not necessarily have this property. If all subsets of P have it, then the lattice is called complete. Thus, finite lattices are always complete, but the rationals with the natural ordering are not. The question of completion of lattices similar to the construction of the reals by Dedekind cuts is discussed in Sect. 5.

There is a more algebraic way of characterizing lattices, where infimum and supremum are regarded as binary operations on P :

Theorem 1 Let $(P, \leq)$ be a lattice. Then for all $x, y, z \in P$:

$$(a) \quad x \vee (y \wedge z) = (x \vee y) \wedge z \text{ and } x \wedge (y \vee z) = (x \wedge y) \vee z$$

$$(b) \quad x \vee y = y \vee x \text{ and } x \wedge y = y \wedge x$$

$$(c) \quad x \wedge (x \vee y) = x \text{ and } x \vee (x \wedge y) = x.$$

Conversely, if a structure $(P, \vee, \wedge)$ is given with operations $\vee, \wedge$ satisfying the conditions (a), (b), and (c), then $\leq$ is defined by

$$x \leq y, \text{ if } x \wedge y = x$$

is a partial order on P such that $(P, \leq)$ is a lattice.

Lattices play an important role in pure and applied mathematics. John von Neumann, one of founders of lattice theory, developed his formulation of quantum mechanics in terms of lattices. Today they are used in many branches of mathematics—for instance, in combinatorics, probability theory, logic, and functional analysis. The wide applicability of lattices in technical and social science contexts is best demonstrated by the development of ‘concept analysis’ by Rudolf Wille (see Ganter and Wille (1996).

3. Extension of Orders; Szpilrajn’s Theorem

Is it possible to extend a partial order by adding (x, y) or (y, x) to $\leq$, when x and y are incomparable in $\leq$ and at the same time keeping the extended relation a partial order? Is it possible to extend $\leq$ to a linear order in this way? Both questions are answered in the affirmative by the following theorem.

Theorem 2 (Szpilrajn’s Theorem) Let $(P, \leq)$ be a poset and $x, y \in P$ such that neither $x \leq y$ nor $y \leq x$. Then there are two linear orders $\leq'$ and $\leq''$, each of which contains $\leq$ as a subset, and such that $x \leq' y$ and $y \leq'' x$.

The proof of this theorem is not difficult for finite P , but entails a set theoretic twist in the infinite case. If x and y are incomparable in $\leq$ then one defines $x \leq' y$, adds all pairs to $\leq'$ which this choice and transitivity forces and goes ahead until no incomparable pairs are left over. The latter step of this argument is a bit of a problem for an infinite P .

Sometimes one can regard a partial order as ensuing from incomplete data collection, incomplete in so far as a linear order can be expected, but the available data give only a part of the full order information. Then the question of a linear extension is relevant, in particular, one can ask how many additional pairwise comparisons must be performed to obtain the full linear order, which is the most parsimonious way to find such an extension.

For these problems the number of linear extensions is of importance. It was shown by Fishburn and Trotter (1992) that for all partial orders with fixed $|P| = n$ and k pairs in $\leq$ those with a maximal number of linear extensions are semiorders. For $k \leq n$ these semiorders are characterized, but for larger K there seem to remain several unsolved problems.

It can be shown that a linear extension can be found with at most K questions of the form ‘is x larger than

y , where K is of the magnitude of the logarithm of the number of linear extensions (see Trotter 1995 for details). In connection with this problem it seems that a pair x, y for which many extensions with $x \leq y$, as well as many with $y \leq x$ exist is best to investigate first because, whatever the answer, it considerably reduces the number of data which remain to be collected. In this context the so called 1/3–2/3 conjecture was made. It says that each poset which is not a linear order contains a pair (x, y) for which the relative frequency of extensions with $x \leq y$ is between 1/3 and 2/3. This conjecture seems still open. However, bounds are known, e.g., the relative frequency is between 3/11 and 8/11. The proof of this result draws on a geometric inequality and is extremely beautiful. For semiorders the 1/3–2/3 conjecture is true. More on this topic and some of the proofs can be found in Trotter (1995).

4. Order Dimension

A corollary of Szpilrajn's Theorem is that each partial order is the intersection of its linear extensions. One can ask for the minimal number of linear extensions with this property and regard this number as a measure of complexity of the order. It is called the dimension of $(P, \leq)$.

All orders on three, four, and five elements are two-dimensional (provided they are not linear). There are two orders with six elements of dimension three. One can group the posets in classes of increasing dimension. The semiorders have a dimension less than or equal to three, but the interval orders can have arbitrary large dimension.

One classical result on the dimension is that it is always smaller than or equal to the maximal number of pairwise incomparable elements. For $|P| \geq 4$ ($|P|$ denotes the number of elements in P) the dimension is bounded above by $|P|/2$ and this is the best possible solution. It is an active area of research to find inequalities such as these which relate the dimension to other characteristics of the poset (see Trotter (1992)).

There are other variants of the dimension concept: one can replace linear orders by semiorders (and obtain the semiorder dimension) or by interval orders or by any other class of partial orders. For biorders, see Doignon et al. (1984) and Koppen (1987). These two papers attempt to link this dimension concept to techniques of extracting factors to explain data. In this way, an ordinal factor analysis is outlined. Factor analytic procedures used to be widely spread in social science research, but until now they seem to lack a sound axiomatic foundation. The order theoretic dimension concept can provide a weaker concept of a factor which is, on the other hand, much easier to justify, or empirically to test for adequacy.

5. Embeddability and Completion of Orders

Sometimes the nonexistence of suprema and infima is a nuisance; just think of the inconvenience caused when a sequence of rational numbers 'converges,' but the limit does not exist (in the rationals). One tries to avoid this situation by embedding a given poset into a complete lattice which contains the original order as a suborder. In contrast to the techniques in Sect. 3, one adds new elements to P instead of new pairs to $\leq$. Thus, the aim is to find a superstructure which resembles the given order as closely as possible, but which can be handled more conveniently because of the existence of suprema and infima. The prototype of this kind of extension is the time-honored construction of the reals from the rational numbers.

To begin with, some notation must be introduced. For $X \subseteq P$ let:

$$X^\vee := \{y \in P: y \leq x \text{ for all } x \in X\}$$

$$X^\Delta := \{y \in P; x \leq y \text{ for all } x \in X\}.$$

Thus $X^\vee$ is the set of lower bounds of X and X^Δ the set of upper bounds of X . The sets $\{x\}^\vee$ and $\{x\}^\Delta$ for $x \in P$ are abbreviated by $x^\vee$ and x^Δ , respectively. One simple way of embedding consists of considering

$$\nabla_{(P, \leq)} := \{x^\vee; x \in P\}$$

the so called principal ideal representation. Note that $(\nabla_{(P, \leq)}, \subseteq)$ is a suborder of the complete lattice of the power set, isomorphic to $(P, \leq)$. However, this embedding is not sufficient for the purposes of this section; the superstructure carries too little information from the original one. Therefore, a more parsimonious extension must be constructed.

To this end, a pair (X, Y) is called a MacNeille cut if and only if $X = Y^\vee$ and $Y = X^\Delta$. The set of all MacNeille cuts is a better candidate for an embedding. Indeed, one can prove:

Theorem 3 Denote by $C_{(P, \leq)}$ the set of all MacNeille cuts of a partial order $(P, \leq)$. Define for two cuts (X, Y) and (X', Y') the relation

$$(X, Y) \sqsubseteq (X', Y') \text{ if } X \subseteq X'.$$

Then $(C_{(P, \leq)}, \sqsubseteq)$ is a complete lattice. Moreover, it is the unique (up to isomorphism) smallest complete lattice which contains the original poset $(P, \leq)$ as a supremum dense subset, i.e., each element of $C_{(P, \leq)}$ is supremum of elements of $(P, \leq)$.

A few remarks should illustrate this theorem. First, the set corresponding to $(P, \leq)$ in $(C_{(P, \leq)}, \sqsubseteq)$ is the set of cuts $(x^\vee, x^\Delta)$. Second, if this theorem is applied to the rationals with the usual (linear) order, then the reals with added $\{-\infty\}$ and $\{\infty\}$ result as the MacNeille completion. The element $-\infty$ is the cut $(\emptyset, P)$ while ∞ is the cut $(P, \emptyset)$. Third, although the notion of a complete lattice only bears meaning for

infinite base sets P , the MacNeille completion makes sense also for finite posets. Figure 2(b) shows the MacNeille completion of (a) and Fig. 2(e) of (d). Finally, it is worth mentioning that the mappings $X \rightarrow X^\vee$ and $X \rightarrow X^\Delta$ form a pair of functions that are known as a Galois connection. This construction is often used when dealing with posets; it is applicable in other contexts, for example, formal concept analysis (see Ganter and Wille (1996) is based on a Galois connection). In this respect, concept lattices are a generalization of a MacNeille completion.

See also: Majorization and Stochastic Orders; Measurement, Representational Theory of; Order Statistics; Ordered Relational Structures

Bibliography

- Doignon J-C, Ducamp A, Falmagne J-C 1984 On realizable biorders and the biorder dimension of a relation. *Journal of Mathematical Psychology* **28**: 73–109
- Fishburn P C 1970 Intransitive indifference with unequal indifference intervals. *Journal of Mathematical Psychology* **7**: 144–9
- Fishburn P C 1985 *Interval Orders and Interval Graphs*. Wiley, New York
- Fishburn P C, Monjardet B 1992 Norbert Wiener on the theory of measurement (1914, 1915, 1921). *Journal of Mathematical Psychology* **36**: 165–84
- Fishburn P C, Trotter W T 1992 Linear extensions of semiorders: A maximization problem. *Discrete Mathematics* **103**: 25–40
- Ganter B, Wille R 1996 *Formale Begriffsanalyse*. Springer, Berlin
- Koppen M G M 1987 On finding the bidimension of a relation. *Journal of Mathematical Psychology* **31**: 155–78
- Luce R D 1956 Semiorders and a theory of utility discrimination. *Econometrica* **24**: 178–91
- Roberts F S 1979 *Measurement Theory with Applications to Decisionmaking, Utility, and the Social Sciences*. Addison-Wesley, Reading, MA
- Suppes P, Krantz D H, Luce R D, Tversky A 1989 *Foundations of Measurement, Vol. 2: Geometrical, Threshold, and Probabilistic Representations*. Academic Press, New York
- Trotter W T 1992 *Combinatorics and Partially Ordered Sets: Dimension Theory*. Johns Hopkins University Press, Baltimore, MD
- Trotter W T 1995 Partially ordered sets. In: Graham R L, Grötschel M, Lovász L (eds.) *Handbook of Combinatorics*. Elsevier, Amsterdam, pp. 433–80

R. Suck

Participant Observation

Participant observation has been part of the arsenal of methodological techniques employed by social scientists since around 1900, notably within the discipline of sociology. Participant observation involves the active engagement of the researcher with the members

of the community that he or she wishes to study, typically as an equal member of the group.

This methodology falls under the general rubric of field research. It is important to recognize that there exist alternative styles of ‘field research’: a range of possibilities that this chapter will not cover. However, in contrast to the manner in which observational research has traditionally been conducted in anthropology, in sociology much ethnography is based on the assumption that the researcher is at least nominally a participant in the group being studied. Yet, the line that separates participant observation and other forms of ethnographic observation is uncertain and hazy, and for some scholars the two terms have been used as synonyms.

Since the mid-1970s there has been a notable increase in the frequency and the legitimacy of the variety of field methodologies within the social sciences. These qualitative, interpretive strategies, once considered ‘unscientific,’ have demonstrated their utility in a number of disciplines. That this occurred during a period in which social scientific funding was being decreased is surely no coincidence. Although at the beginning of the twenty-first century qualitative field methods do not have the same standing as quantitative methods (especially in psychology and economics), their growth is astonishing.

1. Advantages

Several distinct advantages justify participant observation; among these are the benefits of rich data, validity, *Verstehen*, and economy.

1.1 Richness

This methodology, in contrast to most methods that do not involve personal witnessing, provides for rich and detailed data. To investigate action, direct observation is essential. When this is coupled with the questioning of actors about their choices, tied to some measure of active involvement, the value of participant observation is evident. The form of data collection permits an intense depiction that produces a fullness of understanding.

1.2 Validity

A second benefit is analytical validity. Because the observations are of behavior *in situ*, the researcher can rely upon the claim that the findings are close to the ‘proper’ depiction of the scene. If the conclusions drawn from participant observation research are not always reliable, in that other investigators might reach different conclusions, they do reflect the belief that some aspect of the world as it ‘truly’ is, is being described.

1.3 Interpretive Understanding

Participant observation supports the demand of Max Weber to produce research that is characterized by *Verstehen* or a personal understanding. In this way participant observation with its emphasis on both participation and observation adds to research knowledge. By directly involving the researcher in the activity, one can understand on an immediate level the dynamics and motivations of behavior. The observer becomes other than an outsider. While research projects differ significantly on this dimension, involvement has advantages for a *verstehende* analysis that other approaches can not match.

1.4 Economy

Participant observation research is typically inexpensive. While this research strategy is surely labor-intensive, it is not capital-intensive. In many cases the researcher is the only member of the project, and can set the terms of his or her own involvement. Increasingly, however, team-research projects exist as well, and in these cases wages and benefits for hired observers may involve considerable costs. Yet, in the standard participant observation project the key resources consist of a technology for inscribing observations. Perhaps for reasons of cost, connected to the time and energy involved, participant observation has proven to be a technique most typically engaged in by scholars at the early stages of their academic careers.

2. Disadvantages

Just as there are significant advantages to this methodology, disadvantages are evident. Problems relate to proof, generalizability, bias, and time commitments.

2.1 Proof

Participant observation relies upon a single case study: the examination of one place. This raises questions about the nature of proof, or, put another way, about reliability. Will two researchers examining the same or similar social scenes reach the same conclusions? Often because of different perspectives upon entering the field and different experiences within the field, findings are sharply distinct. While the observations and interpretations of those observations may be compelling, one can reasonably wonder whether any set of conclusions is definitive.

2.2 Generalizability

Even if we accept the legitimacy of analyzing one scene, on what grounds can we generalize beyond that

setting? How far can our conclusions be pushed? Participant observation research has a problem in this regard because of the absence of scientific control that characterizes experimental research and produces confidence in the claim that important variables of social life have been adequately captured. As a result, the extent to which generalizability is legitimate is problematic in participant observation. Participant observers need to present a theoretical model that helps readers to judge the legitimacy of their broader claims in light of the audience's own experiences.

2.3 Bias

A strength of participant observation methodology is that the researcher's insight and perspective is taken into account, but this strength has a downside. One cannot adequately distinguish between perspective and bias. The background that the researcher brings to a scene can be distinctively different from other researchers, and, for that matter, from the perspectives of the participants in the setting. To the extent that the researcher's perspectives differ significantly from the perspectives of the participants—possible because of the generally progressive values and upper middle class status of academics—the understanding of a particular scene may be systematically biased.

2.4 Time

Just as participant observation research is relatively inexpensive, it is also highly labor intensive. This form of research requires that the researcher be present in the observed social scene. One cannot simply fly in and out, but must spend sufficient time so that the full range of activities in which participants engage are noted. Much participant observation depends upon chance—what happens to occur at the moment of observation—and, as a result, a significant investment of time is needed. While there is no definitive rule for the proper length of time necessary for observation, most projects require months, if not years, to complete. This, coupled with a modest requirement for capital equipment support means that, as noted, this methodology is particularly appropriate for younger scholars. This reality can mean that participant observation studies often do not have the depth of theoretical understanding that more likely characterizes the work of senior scholars.

3. The Growth and Development of Participant Observation

Although participant observation methodology has deep roots in ethnographic traditions (including in traveler's tales and missionary tales), participant

observation has its origins in the discipline of sociology at the turn of the twentieth century, particularly with scholars affiliated with the Department of Sociology at the University of Chicago. Notable in this early stage was the participant observation studies of Annie Marion MacLean, a student at Chicago and later professor at Adelphi University. MacLean worked and observed in department stores, sweatshops, and coal-fields, in order to gain an appreciation of conditions in these scenes. By the 1920s participant observation was well established in Chicago, as evidenced in studies by Frances Donovan on the lives of waitresses and Frederic Thrasher's observations of gang life. Although members of the 'Chicago school' of sociology employed numerous methodological techniques, the group was probably best known for their observational research and life stories. By mid-century participant observation had expanded well beyond the bounds of Chicago. Works such as William Foote Whyte's *Street Corner Society* (1943), August Hollingshead's *Elmtown's Youth* (1949), Arthur Vidich and Joseph Bensman's *Small Town in Mass Society* (1958), and Herbert Gans's *The Urban Villagers* (1962), collectively demonstrated that through involvement in a community, the researcher could gain a perspective that was both empirically and theoretically powerful. The fact that many of these research projects covered life in city environments led some to describe this approach as 'urban ethnography'. Another stream of 'participant observation' research was grounded in the study of deviant groups and activities. Researchers in such domains have the dilemma of determining the degree to which participation should bolster their observation. Exemplifying such research are studies of the mentally ill, gang members, and drug dealers, notably the classic works by Erving Goffman *Asylums* (1961), and Elliott Liebow, *Tally's Corner* (1967). These two themes—urban life and social deviance—have been integral to participant observation, and despite reduced interest in the late 1980s and early 1990s, scholars, such as Elijah Anderson, Mitchell Duneier, David Snow, and Carol Stack have returned to these themes, creating a 'new urban ethnography' that focuses on issues of urban poverty and racial injustice.

Not all researchers focus on 'problem domains.' Beginning with the inspiration of the Hawthorne studies, and later the research of Donald Roy and Melville Dalton, increasing attention has been paid to observations of work places. The participant observation of organizational life has developed rapidly, creating an active investigation of organizational domains, both in traditional social science departments and within schools of management. Factories, airlines, restaurants, and high tech engineering firms have all been the focus of research. Yet, all social worlds in which individuals with a common interest meet can be analyzed through participant observation, whether clubs, sports teams, or bars.

4. Comparison with Other Field Methods

It is easy to assert that participant observation is synonymous with qualitative research, and, in reality, the lines are blurred. However, participant observation can be contrasted with several other techniques with which it shares some commonalities. Some scholars engage in 'ethnographic immersion.' In such investigations, the researcher becomes a full member of the group, and, in effect, studies his or her own group. This has the advantage that the researcher comes by his or her 'member's knowledge' naturally, but simultaneously the immersion may diminish the ability to examine the group life from the critical perspective that being marginal sometimes brings. Most participant observation maintains a distinction between the researcher and the full member of the group studied. Typically members are told that they are being observed—an outsider, who, for purposes of the study, may participate actively in the life of the group, even though she or he is not a full member.

A second technique, related to ethnographic immersion, has been labeled 'auto-ethnography,' akin to the methodology of introspectionism of the early years of the twentieth century. In these studies, the researcher mines his or her life experiences, hoping to demonstrate the embodied characteristics of social forces. One such example is Carolyn Ellis's emotional and analytical account of her own reaction to the lengthy dying of her significant other. This writing, while occasionally powerful, insightful, and moving, can be faulted for its reliance on subjective experience. Auto-ethnography might be seen as an extreme instance of total immersion, with nothing outside the researcher's experience; given this, such arguments become difficult for others to evaluate. The author proposes a singular account of her emotions that because it is her account leaves no room for alternative interpretations.

A third, related approach is to collect data by means of in-depth interviews. Most participant observation involves questioning of informants. However, some research relies on this questioning to the exclusion of direct observation. This method has the advantage of permitting the researcher to gain multiple perspectives of the same scene and permits the investigation of domains to which it might be impossible to gain direct access. However, in-depth interviews limit the researcher to the presentations of interested parties. The rhetoric in interviews may be systematically different from the private thoughts of the informant and may also differ from their behavior. Observations provide a check on what one is told for reasons of impression management. As always, triangulation of data provides advantages, curbing the weakness of any one methodology.

Fourth, some researchers de-emphasize participation, and are, in practice, pure observers. Some observers make precise statistical measurements of

public behaviors, but this approach is distinct from participant observation. More relevant are those scholars who examine public behavior, particularly behavior within cities. This distinguished tradition of scholarship includes such researchers as William H. Whyte, Erving Goffman, Lyn Lofland, and, more recently, Carol Brooks Gardner. A close reading of street activity, perhaps parallel to certain forms of animal ethology, permits the recognition of patterns of human activity.

Of course, within participant observation, researchers select various roles or strategies. The key dimension on which roles can be differentiated is the extent to which members are involved within group life. Adler and Adler (1987) divide participant observation methods according to the extent of participation: peripheral membership, active membership, and complete membership. In general, the peripheral member observes as an outsider to the group and its culture, the active member is directly engaged in the life of the group, while making it clear to the group that she or he does not fully belong and is conducting research. In the model of the Adlers, the complete observer essentially engages in ethnographic immersion, sometimes informing the group of the research goals.

5. Strategies of Participant Observation

Unlike many techniques of quantitative research and statistical analysis, there are no firm and fast rules that determine how participant observation research should be conducted. Such a reality has proven frustrating to generations of students. Traditionally a novice learned the skills of participant observation through what amounted to an apprenticeship relation with a senior researcher. However, because of the increasing numbers of field researchers, many social science departments, notably in sociology and anthropology, have instituted field methods courses in which some rudimentary 'rules' and strategies of the method are taught.

5.1 Access

Often the first issue that must be confronted in a research project is the question of access. How does one convince a group that they should open their space to an outsider who may criticize the group? This is surely applicable to groups that are widely perceived as deviant (criminal gangs or politically marginal movements), but it also recognizes that any group, striving to maintain boundaries that differentiate it from others, may engage in some activities that members wish to keep hidden from outsiders and which could be viewed negatively. Participant observers must, within the limits of their ethical standards,

convince the group that they can be trusted. Through this process they set 'the rules of engagement.' Such rules, explicit or implicit, differ for each project. Further, this access involves an ongoing temporal process. One is rarely accepted all at once, but often the group maintains a cautious vigilance until they become satisfied that the researcher can be trusted. The participant observer must be sensitive to this personal equation.

5.2 Socialization

Access is only a first step. Once the researcher gains entrance to a setting he or she must determine what behavior means. In effect, the participant observer must become socialized to the environment with its norms, cultural traditions, and jargon. The ease of this process varies as a consequence of how well acquainted the researcher is with the setting prior to the observation, but unless the researcher begins as a full member, a process of acculturation must typically be mastered. At times, the group may become frustrated by the lack of competence of the observer, but more often this process of learning is valuable in methodological terms, permitting the researcher to ask questions that would be denied to others as being too obvious.

5.3 Developing Relationships

Participant observation is a methodology that depends upon the establishment of relationships. It relies upon sociability. One essential skill is the ability to recruit informants from the field. The most important of these are termed 'key informants,' such as Doc in William Foote Whyte's (1943) classic *Street Corner Society*. These individuals serve both as guides for the participant observer and as a means of vouching for the researcher's legitimacy within the setting, providing an insider's stamp of approval. The researcher who lacks the ability to make connections will have difficulty collecting credible research data.

5.4 Field Notes

The researcher does not only need to observe activity, but must subsequently inscribe that material in 'field notes,' documents that serve as the evidentiary basis of the published report. Considerable discussion has addressed what constitutes appropriate field notes and how these notes should be presented (Van Maanen 1988). These notes provide the 'facts' that are observed, including actions, statements, and the feelings of the observer. Participant observers are enjoined to write as much of what transpires as possible, particularly in the early stages of the research, when one is determining what issues are most salient. This empha-

sizes that participant observation is often an inductive methodology, with the field data revealing the important issues, rather than testing pre-established hypotheses.

5.5 Exit Strategies

Much research simply ends. Once completed, the project closes down. Exiting is not as simple in participant observation because the researcher and informants have developed social ties. Leaving the field involves the termination of a set of powerful relationships, and sometimes produces loneliness or even betrayal. As a result, the process of disengagement often involves a longer period, and some connection with the field setting continues. Researchers are often advised—both for practical and ethical reasons—to share material with their informants, either to gain additional feedback or to permit those informants to use the analysis for their own ends. The linkages that are established during the course of the observation can be a continuing source of strength, both for researchers and for informants.

6. Ethical Concerns

Although all forms of research involving human subjects raise ethical issues, the variety of ethical concerns in participant observation are particularly great because of the degree of interpersonal dynamics involved. Four ethical issues have been central to participant observation research: deception, informed consent, confidentiality, and precision of depiction. Although these do not exhaust potential ethical topics, they raise important concerns.

6.1 Deception

Most participant observers agree that it is both pragmatically and ethically necessary to announce to one's informants that one is conducting research. In part, this is because participant observation depends on the personal relations among individuals, who, for that period of time, are status peers. To claim to be something that one is not—to pretend to be a full member of the group—without revealing one's purpose could undermine the entire enterprise if it becomes discovered. Lies are often difficult to maintain for long periods. Further, deception suggests that one's interests take priority over the interests of another: can we justify such trickery? While research on groups that wish to keep their activities hidden—either deviant groups or, occasionally, elite groups—encourage deceptive practices, such strategies are troublesome.

6.2 Informed Consent

Deception involves the active misleading of others, but how should we view the absence of full disclosure? Do group members have the right to know that they are being investigated? For instance, one can easily join a leisure group without announcing that one is taking notes or planning to publish. Is this proper? As noted, good reasons may exist for announcing one's presence, permitting one's informants to serve, in effect, as research assistants. But should researchers be required to gain the 'informed consent' of their informants; further, how much information is necessary for informed consent, particularly in a research methodology that depends on inductive methods, in which the researcher may truly not be aware of what issues will eventually become central? One's topics depend on what one observes. During the past quarter century, American universities have instituted 'ethics' committees (Human Subjects Committees or Institutional Review Boards), designed to protect the 'rights' of research subjects. Frequently researchers are required to gain some measure of informed consent, but how much information is required remains a matter of debate.

6.3 Confidentiality

One element that most (although not all) researchers agree on is that informants' names and clearly identifying information not be included in any published research report, although some who critically examine elites feel that confidentiality is not essential. In general, researchers typically believe that publishing the names or identifiers of particular informants or settings is neither necessary nor warranted. At times this decision poses a problem in that field notes may describe actions or statements that can be recognized, certainly by others within the scene. As a result, participant observers must choose how much information to alter to preserve confidentiality, even when this shades the full 'facts' of the case.

6.4 Precision of Description

Although it is clearly desirable to present data that is precisely accurate—direct quotations and exact descriptions of behavior—participant observation methodology and the limits of human ability make these values unattainable goals (Fine 1993). In part, we are imperfect in inscribing reality, but in part we may recognize that the details of human activity may actually obscure our crucial analytical points. For instance, false starts, errors, and grammatical infelicities mark the speech of informants, like the speech of researchers themselves. Further, so much behavior is happening at any given moment that the researcher can easily become overloaded with trivial observa-

tions. The ethical balance that must be struck is to do justice to the events that are transpiring, while simultaneously making the account as clear as possible to permit communication with readers.

7. Summary

This brief overview has examined central features of the participant observation methodology. In doing so, it de-emphasized both the strategies of analysis (the extended case study method, grounded theory) used by practitioners and the modes of presentation (realist tales, impressionist tales). No method of data collection can exist without related techniques by which data are analyzed and presented.

Throughout the twentieth century, participant observation proved to be a powerful technique through which researchers understood a diverse set of social worlds in a way that did justice to the complexity of the activities of group members and to their own perspectives and understandings. While this methodology benefits by the triangulation with other styles of research to create adequate theoretical understanding, participant observation provides the basis by which researchers can understand the richness of the social settings that they—and others—inhabit.

See also: Field Observational Research in Anthropology and Sociology; Fieldwork: Ethical Aspects; Fieldwork in Social and Cultural Anthropology; Observational Methods in Human Development Research; Observational Studies: Outcome-based Sampling; Observational Studies: Overview

Bibliography

- Adler P A, Adler P 1987 *Membership Roles in Field Research*. Sage, Thousand Oaks, CA
- Fine G A 1993 Ten lies of ethnography. *Journal of Contemporary Ethnography* 22: 267–94
- Gans H 1962 *Urban Villagers*. Free Press, Glencoe, IL
- Goffman E 1961 *Asylums*. Anchor, New York
- Hollingshead A 1949 *Elmtown's Youth*. Wiley, New York
- Liebow E 1967 *Tally's Corner*. Little Brown, Boston
- Van Maanen J 1988 *Tales of the Field*. University of Chicago Press, Chicago
- Vidich A, Bensman J 1958 *Small Town in Mass Society*. Princeton University Press, Princeton, NJ
- Whyte W F 1943 *Street Corner Society*. University of Chicago Press, Chicago

G. A. Fine

Participation: Political

Political participation refers to the activities of the mass public in politics, including, for example, voting

in elections, helping a political campaign, giving money to a candidate or cause, writing or calling officials, petitioning, boycotting, demonstrating, and working with other people on issues. Political participation figures in philosophical discussions of democracy and representative government since ancient times because it provides the means for citizens to communicate their views. Normative interest in the topic remains vigorous (see *Democratic Theory*). This article focuses upon the empirical study of political participation. Scholars look for systematic patterns that explain why some individuals are inactive and why others choose certain activities. The patterns reflect the political context and both reflect and affect the structure of political power.

Participation provides the link from the mass public to the political elite who are presumed—and sometimes shown—to respond (see *Political Representation*).

1. Differing Definitions of Participation

Scholars differ in their definition of political participation. The currently dominant view limits participation to actions that might affect others, as in the following definitions. 'By political participation we mean activity by private citizens designed to influence government decision-making' (Huntington and Nelson 1976, p. 3). 'By *political* participation we refer simply to activity that has the intent or effect of influencing government action—either directly by affecting the making or implementation of public policy or indirectly by influencing the selection of people who make those policies' (Verba et al. 1995, p. 38). The restriction to private citizens is meant to exclude from the concept activity undertaken in their official capacity by those for whom politics and governing are a vocation. A few scholars (cf. Milbrath 1965) include political involvement and activities to support the regime as participation. For these scholars, reading about politics is also political participation, while under the dominant definition it is not (since it does not have direct effects on others).

Another definitional debate in the field concerns the extent to which actions that have the effect of influencing policy without that being the actor's intent also count as political participation. Adopting the more common view, Verba et al. (1995, pp. 38–39) restrict their study to voluntary activity, which they define as follows: 'By voluntary activity we mean participation that is not obligatory—no one is forced to volunteer—and that receives no pay or only token financial compensation.' In contrast, Huntington and Nelson (1976, p. 7) explicitly include not only '*autonomous*' participation but also '*mobilized*' participation, 'defined as' 'activity that is designed by someone other than the actor to influence governmental decision-making'. Under that definition, they would include as a

participant a worker who attends a rally because his boss threatens to demote him otherwise, even if the worker has no intent of affecting government. In practice, it can be difficult empirically to distinguish mobilized from autonomous activity. For answering questions about the impact of participation upon political outcomes, the two are equivalent. The distinction is useful, however, when one focuses on the motives for participation or the impact of participation upon the actor.

2. The Study of Participation

2.1 Participation as Electoral and Hierarchical: Early Studies

The modern empirical study of political participation began developing in tandem with the modern study of electoral behavior using sample survey data. Election outcomes depend not only upon people's preferences but also upon which citizens choose to exercise their preference at the polls. Partially reflecting that link, many of these studies truncated political participation to electoral participation, that is, voter turnout and occasionally some campaign activities (see *Voting: Turnout*). Much empirical work was done in the United States, and much of that was connected to the Michigan Survey Research Center, especially under the leadership of Warren Miller and Philip Converse. These investigators also collaborated with colleagues outside the United States, especially in Europe, producing further studies addressing participation, especially voting. Under the leadership of Stein Rokkan, Scandinavian scholars produced a substantial body of work on participation as part of their more extensive studies of political behavior (see for example Rokkan 1970). These studies painted a clear sociodemographic and psychological picture of voters (and occasionally of participants in campaigns). In contrast to nonparticipants, participants in general had more money and education, were more interested in politics, more partisan, more efficacious, more likely to be men, more involved in organizations, and more rooted in their community. Some, including Rokkan, also paid close attention to the relationship between political behavior, including participation, and the citizen's position in the social cleavage structure.

Milbrath summarized the findings to date in 1965. He included political support as participation as well as electoral activities beyond voting. He promoted the influential organizing idea that participation lay on a hierarchical scale—there was a ladder of activities, with the easiest at the bottom and the most difficult at the top. 'Persons who engage in the topmost behaviors are very likely to perform those lower in rank also' (Milbrath 1965, pp. 17–18). Given the emphasis on electoral activities and given the frame created by the

ladder concept, levels of participation in the United States were perceived as low.

2.2 Large-scale Cross-national Studies

Interest in democratic stability prompted cross-national projects that took political participation as an important variable. For Harry Eckstein, participation provided the link upwards from subordinates to superordinates (Eckstein and Gurr 1975, pp. 60–7). For Almond and Verba in *The Civic Culture* (1963), the participatory behavior of citizens in a country was one important component in characterizing their political culture, but they were just as interested in attitudes and beliefs about hypothetical behavior as in activity. Nonetheless, the data, collected in 1959 and 1960 in five countries (the United States, Mexico, Italy, West Germany, and Great Britain) provided the first major database used for cross-national comparisons of political participation, especially beyond voting turnout. Some of these analyses contributed the important conclusions that persons involved in organizations participate more in politics and that societies with denser organizational structures exhibit higher levels of participation.

Within the decade (1966 to 1971) Verba was in the field again, with an international team of collaborators, studying political activity. Surveys were completed in the United States, Japan, Austria, the Netherlands, India, Nigeria, and the former Yugoslavia. Some of the first products of this research (e.g., Verba and Nie 1972) overturned the conception of participation as a hierarchy of electoral activities. Instead, participation was expanded to include non-electoral activities and reconceptualized as modes of activity in any one or more of which individuals might specialize. Although often these are positively correlated, meaning that a participant in one is more likely to take part in others than a randomly selected person, many people nonetheless partake in one or some to the exclusion of other forms of participation. In particular, substantial numbers of people who avoided electoral politics did engage in nonpartisan participation, such as involvement with community groups.

As reported in Verba and Nie (1972) and Verba et al. (1978), as well as other publications from the project, wealth and education were reaffirmed as important predictors of participation. But the main theoretical point is that the strength of that relationship varies across societies as a function of the strength of other, 'group,' resources. Where these are strong, such as in countries with strong political cleavages tied to ascriptive characteristics, or with sociological or occupational segments with strong political leadership, the relationship of socioeconomic status (SES) to participation was relatively weak. Certain well-off segments of the population were 'inhibited' from

participation while other, possibly disadvantaged, segments were 'mobilized,' thereby together attenuating the relationship between SES and participation. While others had noted similar phenomena, the finding now had wide cross-national grounding.

The Verba and coworkers' studies deliberately ignored political protest, that is, participation in activities outside accepted conventional norms, including some that are illegal. This omission seemed unfortunate by the 1970s in light of the frequency of real-world protest. Out of the study of revolution and political violence (see *Revolution*) came renewed interest in relative deprivation. But many protest activities are not violent (see *Political Protest and Civil Disobedience*).

A group of European and American scholars launched the Political Action project, a cross-national study focused on protest. They conducted surveys in the United States, West Germany, the Netherlands, Great Britain, and Austria in 1973 to 1974, in Italy, Switzerland, and Finland two years later, and a second wave in the first three countries in 1979 to 1980. The results appeared in *Political Action* (Barnes et al. 1979), *Continuities in Political Action* (Jennings et al. 1989) and other publications. One objective of this research was to understand the relationship between protest and conventional participation. The Political Action project clearly showed that enough persons who participated in conventional politics were adding protest to their 'action repertory' to conclude that conventional and protest participation were complements rather than substitutes. The Political Action studies differed from their predecessors in showing that variations in values predict both who will be active and the form of their activity. Specifically, persons with postmaterialist values are more active and more likely to protest than are materialists. Interpretation of the results is hampered slightly by the inclusion of respondents' hypothetical willingness to protest in some of the measures of protest activity.

Other scholars investigated the effect of political institutions on participation. In particular, they showed that cross-national variations in voter turnout rates reflect differences in electoral systems, party competitiveness, registration systems, government structure, and requirements to vote (see *Voting: Compulsory*).

Attention to participation in organizations shades into studies of social movements. Where a political scientist sees a participant in a voluntary organization addressing a specific problem (and thus somehow engaged in the communal mode of participation), a sociologist sees someone active in a social movement. The latter have produced another extensive, related literature.

Approaches and/or theories from these cross-national projects have been influential in many country-specific studies of participation. These studies have contributed much that is new, but their con-

clusions have also tended to support the theories developed in the large projects.

2.3 *Mobilization as a Predictor of Participation, Not a Type, and the Civic Voluntarism Model (CVM)*

The next major theoretical development in the study of participation involved incorporation of the recognition that people are more likely to participate when they are asked to do so (i.e., when they are 'recruited' or 'mobilized') and, moreover, that mobilization follows systematic patterns. Rosenstone and Hansen (1993) developed this idea in their study of participation in the United States from 1956 to 1990, claiming that a decline in mobilization over that period accounts in large part for the observed decline in participation.

Verba et al. (1995) use the related concept of 'recruitment' cross-sectionally in their study of political participation in the United States based on a 1989–1990 survey. It is the third component of the CVM; the first two are 'resources' and 'engagement.' They argue that higher levels of any of these factors lead to greater participation—people participate in politics because they can, because they want to, and because someone asks.

'Resources' are most readily measured by income and education, but these serve as proxies for more fundamental measures. The most interesting components of resources are the 'civic skills,' operationalized by Verba et al. (1995) as writing a letter, making a speech, and attending or planning a meeting where decisions are made in one or more nonpolitical settings, such as on the job, at church, or in a voluntary association. Civic skills turn out to explain the long-observed association between organization membership and participation. 'Engagement' includes interest, efficacy, political information, and partisan attachment, and also specific issue or ideological concerns. Note the connection between this part of the model and the finding that postmaterialists are more participatory. 'Recruitment' includes both requests from individuals and those from persons in authority (e.g., on the job, at church, or in an organization). This latter route provides one path by which group resources operate.

People with greater wealth and education in general have more resources, are more engaged, and are more likely to be recruited, so overall the relationship between socioeconomic status and participation is positive. The CVM model provides a fuller understanding however of why the relationship holds. Moreover, were it applied cross-nationally, it would point to components that need measuring. For example, churches play a role in the United States not likely to be replicated elsewhere.

Cost and benefit analyses inspired by rational actor models influence much contemporary work on participation, including that of Rosenstone and Hansen

(1993) and Verba et al. (1995) (see *Rational Choice in Politics*).

Many of the factors that lead to high levels of political participation are associated with high levels of social capital (see *Social Capital*) notably including substantial organizational involvement. Since social capital is also measured with political participation, the relationship among the concepts and the phenomena need more attention.

3. Conceptual Issues in Research on Political Participation

3.1 Participation Outside Advanced Democratic Countries

Most of the research discussed above has focused on economically-advanced democratic countries, with the occasional addition of other research sites. The restriction to democratic countries reflects the particular role of political participation in a democracy. Participation provides the means by which people control their government, most decisively so in elections which determine who rules. The unique power of the vote to change rulers provides the force for most of the other modes of participation. Campaigning, expressions of opinion, community organizing, and other actions command attention in part because ultimately the opposition can win an election.

In an authoritarian regime, the electoral connection no longer exists. Nonetheless, there is still political participation. Although the people no longer select the rulers, the rulers still need to keep the population reasonably satisfied to forestall revolt, and, to do so, they need information on the people's preferences, and they need to respond to discontent. Several very interesting studies have examined participation in authoritarian regimes, such as Friedgut's study of the USSR (1979) and the work by Shi (1997) on China.

In economically less-developed countries, one frequently finds that government institutions are also less well developed, that the associational networks are weak, or that loyalties to local groups outweigh those to governmental institutions. In these circumstances, much political participation may consist of informal involvement in the local arena or in traditional institutions (see, for example, Dietz 1998).

Issues of cross-national and cross-cultural comparability loom large in all studies of participation, but they are especially salient for research in non-democratic or less-advanced countries. One must rethink what counts as 'political' in these contexts in light of the institutional situation. For example, some of these studies led to the suggestion that participation includes activity directed towards altering the implementation of policy, rather than just its formulation, or that bribery is an important mode. Certain

actions that are nominally the same can serve very different functions in differing contexts. For example, 'voting' may just show regime support in one country but select a government in another. The independent variables of importance also vary. For example, because of the particular circumstances in China, 'education' does not usefully measure resources there (Shi 1997, pp. 145-8). In many countries, other institutions are likely to play the role of churches in the United States as a source of civic skills.

Matters become more complex when an authoritarian regime democratizes. Many current studies examine participation in Eastern Europe and the republics of the former Soviet Union. With the abrupt changes in regime, the process of political socialization (see *Socialization: Political*) will probably also change, providing good conditions in which to observe the effect of political attitudes on participation.

3.2 Leadership

If people participate more when they have resources, are engaged, and are recruited, then elite action can significantly alter the level of participation by affecting engagement and recruitment. Political leaders provide resources to make participation easier (such as a babysitter, a ride to the polls, or information on what to support). Leaders contact people, directly or through proxies, asking them for action. Leaders can increase engagement by framing the public debate in a way that increases the stakes of political competition by communicating this message to their listeners.

The contemporary mass media play a role very similar to leaders. They frame issues and strongly affect (at times, negatively) engagement and recruitment. However, their motivations differ from those of leaders as they seek attention rather than support, and their actions are not targeted.

The political world contains many persons vying for power, each of whom needs large numbers of supporters to achieve power. Only some are successful in recruiting others and, as noted above, theory has not yet explained the circumstances of success. But political participation will increase when many are active and successful, and it will be higher than otherwise expected for persons who are targets of their efforts.

See also: Alliances: Political; Campaigning: Political; Citizen Participation; Cross-pressures: Political; Democratic Theory; Efficacy: Political; Faction: Political; Nonpartisanship: Political; Political Protest and Civil Disobedience; Political Representation; Power: Political; Rational Choice in Politics; Revolution; Social Capital; Social Movements, Sociology of; Socialization: Political; Voluntary Organizations; Voting: Class; Voting: Compulsory; Voting, Sociology of; Voting: Turnout

Bibliography

- Almond G, Verba S 1963 *The Civic Culture*. Princeton University Press, Princeton, NJ
- Barnes S H, Kaase M, Allerbeck K R, Farah B G, Heunks F, Inglehart R, Jennings M K, Klingemann H D, Marsh A, Rosenmayr L 1979 *Political Action: Mass Participation in Five Western Democracies*. Sage, Beverly Hills, CA
- Conway M Margaret 2000 *Political Participation in the United States*. 3rd edn. CQ Press, Washington, DC
- Dietz H 1998 *Urban Poverty, Political Participation, and the State*. University of Pittsburgh Press, Pittsburgh, PA
- Eckstein H, Gurr T R 1975 *Patterns of Authority: A Structural Basis for Political Inquiry*. Wiley, New York
- Friedgut T H 1979 *Political Participation in the USSR*. Princeton University Press, Princeton, NJ
- Huntington S P, Nelson J M 1976 *No Easy Choice*. Harvard University Press, Cambridge, MA
- Jennings M K, van Deth J W, Barnes S H, Fuchs D, Heunks F J, Inglehart R, Kaase M, Klingemann H-D, Thomassen J J A 1989 *Continuities in Political Action*. de Gruyter, Berlin
- Milbrath L W 1965 *Political Participation*. Rand McNally, Chicago
- Rokkan S 1970 *Citizens Elections Parties*. Oslo University Press
- Rosenstone S J, Hansen J M 1993 *Mobilization, Participation, and Democracy in America*. Macmillan, New York
- Shi T 1997 *Political Participation in Beijing*. Harvard University Press, Cambridge, MA
- Verba S, Nie N H 1972 *Participation in America*. Harper & Row, New York
- Verba S, Nie N H, Kim J-O 1978 *Participation and Political Equality*. Cambridge University Press, Cambridge, UK
- Verba S, Schlozman K L, Brady H E 1995 *Voice and Equality*. Harvard University Press, Cambridge, MA

C. J. Uhlaner

Parties: Litigants and Claimants

This article focuses on parties who are subject to legal proceeding in a nonprofessional capacity. Excluded are lawyers, judges, and other officials who make their living from legal work; also excluded are the individuals who serve as witnesses and jury members in trials. The focus is on actors who are the *subject* of legal decisions. They include litigants, claimants, and those subject to investigation and scrutiny by law enforcers in both criminal and administrative capacities.

It is important at the onset to differentiate between those who initiate legal proceedings and defendants. Plaintiffs in civil courts and complainants in criminal courts exercise volition in choosing to take disputes to court. While defendants in criminal and civil court proceedings may indirectly exercise volition, in the sense that the legal proceedings may be a consequence of their own prior decisions, the choice to be involved in the legal proceeding itself is typically not something they would choose on their own.

Disputes and allegations of illegal behavior that end in court are merely the tip of the iceberg (Felstiner et al. 1980–81). On the civil side, people may formally or informally negotiate a settlement ‘in the shadow of law’, or may ‘lump’ an unresolved dispute due to high costs of proceeding, inability to find an advocate who will take the case, or fear of retaliation for making legal claims. In many places, criminal cases are resolved by plea bargains instead of a formal trial. Complainants may also have their cases turned away by gatekeepers as not being sufficiently ‘legal’ to warrant court attention (Yngvesson 1993, Merry 1990).

1. Dimensions of Difference in Parties

The research literature focuses on (a) the distinction between individuals and organizations, and (b) demographic differences related to social power—primarily gender, race, and class. Both lines of research attempt to account for differences in the experiences of various types of parties with legal systems, but in somewhat different ways.

Comparisons of individuals and organizations have been made in studies testing Marc Galanter’s theory of why ‘haves’—parties with greater resources—may have a greater advantage in legal settings (Galanter 1974). Most empirical tests of the theory have examined success rates for different classifications of organizations compared to individuals and each other, using appellate court cases in which there is a written record. In addition, white-collar crime research has also emphasized differential treatment of organizations and individuals, and poses additional explanations for why organizations may experience more advantageous treatment than individuals when they violate the law.

Studies of gender, race, and ethnicity tend to focus more on case processing, and to examine variations among different categories of individuals in far greater detail, with minimal attention to variations among organizations. A variety of methodologies have been employed, including observational studies, analysis of legal cases, and quantitative survey studies. Some studies focus on variations in resources and skills that affect an individual’s chances of negotiating successfully through legal proceedings, while others focus on the social categories and standards used by legal decision-makers in making decisions on how to proceed with cases.

2. Comparisons of Individuals and Organizations

Galanter’s classic 1974 article ‘Why the “haves” come out ahead’ argues that organizations are more likely, compared to individuals, to have structural advantages that increase the likelihood of winning in court. In addition to their greater financial resources and

ability to hire legal talent, large organizations are more likely to be 'repeat players' who are better able to mobilize legally and to use law to meet strategic goals (Galanter 1974). A number of studies of state and federal appellate court decisions have found support for Galanter's argument (Songer et al. 1999). The most consistent finding is that government organizations, particularly federal and state governments, experience a substantial advantage over both business and individuals, while business organizations appear to have a modest advantage over individuals. Studies that distinguish between larger business corporations and smaller business organizations generally find a stronger advantage for big business (Farole 1999, Songer et al. 1999).

Much of the appellate court research has been conducted in the United States, but similar patterns have also been found for other countries governed by the English common law tradition. Moreover, Songer et al. (1999) found considerable stability over time of the basic pattern of results for the US Court of Appeals, suggesting some generality to the pattern of greater organizational success in appellate courts.

A study of Philippine Supreme Court decisions, however, found strikingly different results, with individuals experiencing greater success in winning cases than organizations (Haynie 1994). Haynie speculates the legitimacy of high courts in developing countries may be more precarious. This may lead courts to make more decisions with redistributive effects in a bid to enhance their legitimacy by establishing their independence from other powerful actors in society.

Research on lower courts points to mixed evidence regarding the thesis that organizations are advantaged over individuals. Anthropological studies of the lower courts suggest that organizations face fewer difficulties than individuals in getting their complaints against others into court. Yngvesson (1993) found that the majority of complaints brought by individuals were characterized as 'garbage' cases by court clerks, who often expended considerable effort to keep them out of court (see also Merry 1990). In contrast, complaints filed by police officers—representatives of a local governmental agency—were routinely admitted with minimal review as to their appropriateness for legal action. Similarly, motor vehicle complaints, code violations, and welfare fraud cases filed by government agencies, as well as debt collection cases filed by local businesses, were given a higher priority.

No doubt one of the reasons why complaints filed by the police, other government agencies, and businesses were more readily accepted and processed by the lower courts is that they are more likely to be explicitly linked to legal provisions and couched in the language of law and legal doctrine. This is in part due to the focus of such cases on disputes about money and property, while complaints brought by individuals tend to involve interpersonal disputes and problems with relationships. In addition to the more explicitly

'legal' character of property disputes, the greater training and proficiency of organizational representatives in formulating their complaints in legal terms may also assist getting borderline cases into court. As one police officer stated, while discussing a complaint he filed of 'deliberate and malicious' damage to a tree. *If you really want to, you can find a law to do what you want to do* (Yngvesson 1993, p. 21; emphasis in original).

Organizational studies abound with examples of the strategic use of the law to achieve organizational goals. Examples in the business sector include innovative interpretations of tax laws in tax avoidance schemes (McBarnet 1991) and the strategic use of bankruptcy law (Delaney 1992). In the government sector, prosecutors in the USA have creatively used the RICO (Racketeer-Influenced and Corrupt Organizations) law to expand prosecutions beyond the original target of organized crime to include white-collar crimes such as commercial and financial fraud, tax evasion, embezzlement, and bribery (Friedrichs 1996).

While the extension of RICO laws to white-collar offenses offers an example of legal innovation that increases the advantage of large government organizations over individuals and business organizations, the research literature on white-collar crime finds that, as suspects and defendants in criminal proceedings, organizations are advantaged over individuals. As Sutherland (1983) noted, illegal behaviors by organizations are typically handled by regulatory agencies instead of criminal law enforcement agencies. Sanctions by regulatory agencies are relatively mild, typically involving fines, cease and desist orders, and other civil sanctions. While regulatory agencies may seek criminal sanctions, the criminal prosecution of organizations remains relatively rare.

The allocation of cases of organizational misconduct to civil courts and administrative enforcement, rather than criminal courts, cannot be attributed to a lesser seriousness of organizational misconduct generally. Cost comparisons typically find that illegal behaviors by organizations tend to be more deadly and expensive in the aggregate than crimes committed by individuals (Poveda 1994). Survey research generally finds that the public considers corporate crimes that result in loss of life as serious in nature (Friedrichs 1996). Some authors have also argued that greater alienation and cynicism by the public against public institutions generally is an additional cost of organizational crime (Maier and Short 1995).

Various explanations exist for why 'street crimes,' primarily committed by lower-class individuals, are relegated to the criminal courts, while offenses of organizations are handled primarily by civil courts and administrative enforcement actions. Legal conflict theorists argue that those who have power and money are better able to influence legislators to exclude their behavior from criminal statutes (Green 1997). Even

when laws define misconduct by organizations as criminal, practical difficulties to securing criminal convictions include the diffusion of responsibility within organizations, and the greater complexity and cost of amassing evidence for organizational crimes (Friedrichs 1996).

Moreover, contemporary theories of organizational misconduct point to the influence of organizational culture and processes or organizational decision-making as causes of misconduct that do not conform easily to models of individual choice such as underlie much of the criminal law (Ermann and Rabe 1997). As Vaughan notes in her study of the *Challenger* disaster, 'repetition, seemingly small choices, and the banality of daily decisions in organizational life—indeed, in most social life—can camouflage from the participants a cumulative directionality that too often is discernible only in hindsight' (Vaughan 1996).

Finally, the pragmatics of securing cooperation from organizations to prevent future harms may be a key factor in the allocation of law enforcement activities related to organizational misconduct to regulatory agencies. While US regulatory agencies have historically been more likely to use adversarial, deterrence-oriented tactics such as the imposition of fines and penalties, regulatory agencies in most countries typically focus their energies on using educational and persuasive techniques to secure compliance (Braithwaite 1985).

The primary rationale for cooperative law enforcement strategies is that, while compliance with law for individuals often means refraining from committing illegal acts, compliance for organizations typically involves perspective behaviors, such as buying new equipment, training employees, and undertaking new routines and processes that are both expensive and time-consuming. Utilizing adversarial tactics may be counterproductive if they result in organized resistance by powerful actors (Bardach and Kagan 1982). Braithwaite (1984) frankly notes that a consequence of cooperative enforcement techniques is the introduction of distributive inequities, but argues that the payoff in increased effectiveness in saving lives and reducing human misery is worth the cost.

Counterbalancing the advantages organizations have in avoiding criminal proceedings for instances of illegal behavior, research on the civil courts typically finds that, when cases go to trial and the plaintiffs win, civil court verdicts are typically harsher for organizations than for individuals (Chin and Peterson 1985, Maccoun 1996). The 'deep pockets' hypothesis suggests that organizations are both more likely to be sued and to have higher damages assessed because they usually have more assets than individuals. While 'deep pockets' may explain decisions by both potential plaintiffs and attorneys to pursue cases, jury simulation studies have found little evidence that greater wealth explains the differential size of damage awards. Instead, it appears that juries apply different standards

to organizations than individuals (see Maccoun 1996). However, the greater harshness of civil damage awards against organizations should be placed in context, in that few lawsuits go to trial.

3. Demographic Differences Related to Social Power

Other articles in this encyclopedia discuss race, class, gender, and the law in greater detail than is possible here. The purpose of this article section is to contrast apparent commonalities in what legal parties want from legal decision-makers with the difficulties they may experience in obtaining what they want due to barriers arising from their position in society.

Research on procedural justice suggests a high degree of commonality in what people want from legal proceedings (Lind and Tyler 1988, Tyler 1988). Studies conducted in different countries and in a variety of legal settings have consistently found that perceptions of fair process affect satisfaction with legal experiences as much or more than objective outcomes and perceptions of distributive justice. Moreover, procedural justice perceptions are generalized beyond the specific experience to affect the legitimacy of other legal and political authorities (Lind and Tyler 1988) and also expectations of fair treatment in other legal settings (Stalans 1994).

According to the group-value model of procedural justice, encounters with legal authorities convey profoundly symbolic messages to individuals about their social standing (Lind and Tyler 1988, Tyler and Lind 1992). The concept of standing emphasizes basic concerns individuals have about their membership and inclusion in social groups, about whether they are valued members of the community. Tyler (1988, 1994) argues that a substantial consensus exists on how people define procedural justice, and suggests that fair process is a means by which multicultural societies can maintain their legitimacy amid the diverse goals and interests of different groups within society. Results from other studies, however, caution that procedural justice may be a less important concern than outcomes for people who do not have a sense of identification with the larger collectivity (Huo et al. 1996).

Despite the emphasis of the group-value model of procedural justice on social standing and intergroup relations, surprisingly little procedural justice research explicitly addresses issues of social stratification. Yet many of the most pressing and controversial issues in law and society research involve the ways in which external social structures of power and domination are reproduced in legal settings during the enactment of legal procedures.

Anthropological research indicates that social stratification is intertwined with relational concerns and definitions of 'legal' issues in ways that systematically

privilege some groups over others in actual legal proceedings. To begin, the types of disputes that are often of central concern to women and lower-income individuals involve troubled relationships, including domestic violence. Yngvesson (1993) notes that clerks in New England courts were reluctant to bring cases involving relationship problems before judges. Even divorce lawyers who make their living from the dissolution of troubled relationships tend to be unresponsive and aloof to clients' tales of relational woes (Sarat and Felstiner 1995).

Merry (1990) notes that cases involving relational issues are more likely to be diverted from court into mediation—and that mediation itself and the court employees who manage mediation have a devalued social status in the court system. While mediation explicitly focuses on relationship issues, the application of rules of discourse that discourage conflict during the process of mediation may serve to systematically disadvantage women, compared to more adversarial court proceedings that focus on advocating and protecting legal rights (Conley and O'Barr 1998).

Discomfort with handling troubled relationships is not confined to Western legal systems. Griffiths (1997) reports that, in Botswana, a common ruling by *kgotla* in marital disputes is simply to order the parties to go home and 'live together in harmony and peace.' In Kenya, Swahili women who tell stories of domestic conflict and discord in Islamic Kadi courts violate social norms against revealing family secrets in public (Hirsch 1998). While some African courts encourage the airing of relationship issues, detailed analyses reveal the systematic shaming of women during the process of talk about relationships (Conley and O'Barr 1998).

The devaluation of relationship talk surfaces in Conley and O'Barr's study of US litigants in small claims court. Conley and O'Barr described two distinct types of narrative structures used by litigants. Rule-oriented accounts are ordered sequentially, describing the unfolding of events in a linear fashion while noting violations of specific rules, duties, and obligations by the other party. They articulate cause and effect and assess responsibility. Relational accounts, on the other hand, focus more on social position and general rules of social conduct. They are more likely to be full of details about the life of the speaker and the history of his/her relationship with the other party, including details that are not relevant to the legal issue at hand.

As Conley and O'Barr (1998) note, rule-oriented accounts conform to the logic of law in Western societies. Legal decision-makers readily understand rule-oriented accounts, while they have more difficulty processing relational accounts. It takes more work for judges and lawyers to translate the jumbled and rambling accounts in relational narratives into the rule-oriented framework of law. Many do not make the effort, to the disadvantage of individuals who

employ relational accounts. This gives an advantage to repeat players, to those who can afford to hire legal counsel, and to better-educated, more articulate parties who bring to court greater skills in couching their narratives in forms readily comprehensible to legal decision-makers.

Conley and O'Barr also describe stylistic features of speech that can serve to reproduce structures of social domination (Conley and O'Barr 1998, Ainsworth 1993). Conley and O'Barr note that women and men of lower social status in society are more likely to evidence styles of speech that project deference and uncertainty, thereby undermining their own credibility. 'Powerless' speech includes the use of hedge words, polite forms, tag questions, exaggerated imprecision about quantities, and a rising intonation at the end of declarative statements. Although Lakoff (1975) first identified these speech styles as gendered features of language, studies in legal settings have found that some men also use these forms of speech, to their detriment (Conley and O'Barr 1998, see also Ainsworth 1993).

While dominant social orders tend to reproduce themselves in legal settings, the process is by no means immutable and unchanging (Griffiths 1997, Hirsch 1998). As Conley and O'Barr (1998) note, many individuals of subordinate social status are capable of adopting a rule-oriented discourse strategy and modifying their speech styles. In addition, the existence of a 'chivalry bias' favoring women in criminal cases (Anderson 1976) and the harsher sentences given to high-prestige individuals convicted of white-collar crimes (Weisbrud et al. 1991) highlight that the general social order is not simply reproduced in legal settings, but can be transformed in ways that create reversals of social fortune.

At the same time, it is important to note that some patterns of domination and unequal treatment are more deeply engrained in the fabric of legal interactions and less amenable to modification or manipulation (e.g., Matoesian 1993, 1995). Studies of legal case processing in particular illuminate seemingly ubiquitous processes of social categorization of parties by legal decision-makers according to race, class, and gender in the course of identifying suspects, managing caseloads, evaluating credibility, and making decisions about how hard to work on cases (e.g., Waegel 1981, Frohmann 1997). While social categorization often works to further organizational goals and to increase the speed and efficiency of legal decision-making, it does so at the cost of procedural and distributive inequities that arise when individuals are judged through the prism of presumed group characteristics.

See also: Class and Law; Conflict and Conflict Resolution, Social Psychology of; Conflict Sociology; Courts and Adjudication; Critical Race Theory; Disputes, Social Construction and Transformation of;

Feminist Legal Theory; Gender and the Law; Justice and its Many Faces: Cultural Concerns; Justice, Access to: Legal Representation of the Poor; Justice: Philosophical Aspects; Justice: Social Psychological Perspectives; Litigation; Mediation, Arbitration, and Alternative Dispute Resolution (ADR); Race and the Law

Bibliography

- Ainsworth J E 1993 In a different register: The pragmatics of powerlessness in police interrogation. *Yale Law Journal* **103**: 259–322
- Anderson E A 1976 The 'chivalrous' treatment of the female offender in the arms of the criminal justice system. *Social Problems* **23**: 350–57
- Bardach E, Kagan R 1982 *Going by the Book: The Problem of Regulatory Unreasonableness*. Temple University Press, Philadelphia, PA
- Braithwaite J 1984 *Corporate Crime in the Pharmaceutical Industry*. Routledge and Kegan Paul, London
- Braithwaite J 1985 *To Punish or Persuade? The Enforcement of Coal Mine Safety Laws*. State University of New York Press, Albany, NY
- Chin A, Peterson M 1985 *Deep Pockets, Empty Pockets, Who Wins in Cook County Jury Trials*. RAND Institute for Civil Justice, Santa Monica, CA
- Conley J M, O'Barr W M 1990 *Rules Versus Relationships: The Ethnography of Legal Discourse*. University of Chicago Press, Chicago, IL
- Conley J M, O'Barr W M 1998 *Just Words: Law, Language, and Power*. University of Chicago Press, Chicago, IL
- Delaney K J 1992 *Strategic Bankruptcy: How Corporations and Creditors Use Chapter 11 to their Advantage*. University of California Press, Berkeley, CA
- Ermann M D, Rabe G A 1997 Organizational processes (not rational choices) produce most corporate crime. In: Lofquist W S, Cohen M A, Rabe G A (eds.) *Debating Corporate Crime*. Anderson, Cincinnati, OH, pp. 53–68
- Farole D J Jr. 1999 Re-examining litigant success in state supreme courts. *Law and Society Review* **33**: 1043–58
- Felstiner W, Abel R, Sarat A 1980–81 The emergence and transformation of disputes: Naming, blaming, claiming. *Law and Society Review* **15**: 631–54
- Friedrichs D O 1996 *Trusted Criminals: White Collar Crime in Contemporary Society*. Wadsworth, Belmont, CA
- Frohmann L 1997 Convictability and discordant locales: Reproducing race, class, and gender ideologies in prosecutorial decision-making. *Law and Society Review* **31**: 531–56
- Galanter M 1974 Why the 'haves' come out ahead: Speculations on the limits of legal change. *Law and Society Review* **9**: 95–160
- Green G S 1997 *Occupational Crime*, 2nd edn. Nelson-Hall, Chicago, IL
- Griffiths A 1997 *In the Shadow of Marriage: Gender and Justice in an African Community*. University of Chicago Press, Chicago, IL
- Haynie S L 1994 Resource inequalities and litigation outcomes in the Philippine Supreme Court. *Journal of Politics* **56**: 752–72

- Hirsch S F 1998 *Pronouncing and Persevering: Gender and the Discourses of Disputing in an African Islamic Court*. University of Chicago Press, Chicago, IL
- Huo Y J, Smith H J, Tyler T R, Lind E A 1996 Superordinate identity, subgroup identity, and justice concerns: Is separatism the problem? Is assimilation the answer? *Psychological Science* **7**: 40–45
- Lind E A, Tyler T R 1988 *The Social Psychology of Procedural Justice*. Plenum, New York
- Lakoff R 1975 *Language and Women's Place*. Harper and Row, New York
- Maccoun R J 1996 Differential treatment of corporate defendants by juries. *Law and Society Review* **30**: 121–62
- Matoesian G M 1993 *Reproducing Rape: Domination Through Talk in the Courtroom*. University of Chicago Press, Chicago, IL
- Matoesian G M 1995 Language, law, and society: Policy implications of the Kennedy Smith rape trial. *Law and Society Review* **29**: 669–701
- McBarnet D 1991 Whiter than white-collar crime: Tax, fraud insurance, and the management of stigma. *British Journal of Sociology*
- Meier R F, Short J F Jr. 1995 The consequences of white-collar crime. In: Geis G, Meier R F, Salinger L M (eds.) *White-collar Crime: Classic and Contemporary Views*, 3rd edn. The Free Press, New York, NY, pp. 80–104
- Merry S E 1990 *Getting Justice and Getting Even: Legal Consciousness Among Working-Class Americans*. University of Chicago Press, Chicago, IL
- Poveda T G 1994 *Re-thinking White-collar Crime*. Praeger, Westport, CT
- Sarat A, Felstiner W L F 1995 *Divorce Lawyers and Their Clients*. Oxford University Press, London
- Songer D R, Sheehan R S, Haire S B 1999 Do the 'haves' come out ahead over time? Applying Galanter's framework to decisions of the US Court of Appeals, 1925–1988. *Law and Society Review* **33**: 811–32
- Stalans L J 1994 Formation of procedural beliefs about legal arenas: Do people generalize from loosely related past legal experiences? *Psychology, Crime, and Law* **1**: 1–19
- Sutherland E H 1983 *White Collar Crime: The Uncut Version*. Yale University Press, New Haven, CT
- Tonry M 1995 *Malign Neglect: Race, Crime, and Punishment in the US*. Oxford University Press, New York
- Tyler T R 1988 What is procedural justice? Criteria used by citizens to assess the fairness of legal procedures. *Law and Society Review* **21**: 103–88
- Tyler T R 1994 Governing amid diversity: The effect of fair decisionmaking procedures on the legitimacy of government. *Law and Society Review* **28**: 809–32
- Tyler T R, Lind E A 1992 A relational model of authority in groups. In: Zanna M (ed.) *Advances in Experimental Social Psychology*. Academic Press, NY, Vol. 25, pp. 115–92
- Vaughan D 1996 *The Challenger Launch Decision: Risky Technology, Culture, and Deviance at NASA*. University of Chicago Press, Chicago, IL
- Waegel W B 1981 Case routinization in investigative police work. *Social Problems* **28**: 263–75
- Weisbrud D, Wheeler S, Waring E, Bode N 1991 *Crimes of the Middle Classes: White-collar Offenders in the Federal Courts*. Yale University Press, New Haven, CT
- Yngvesson B 1993 *Virtuous Citizens, Disruptive Subjects: Order and Complaint in a New England Court*. Routledge, New York

K. A. Kinsey

Parties/Movements: Extreme Right

Political parties and movements that current conventional labeling by political scientists and other informed analysts would regard as of the extreme right have emerged in many countries and on many occasions, since at least the early nineteenth century. The label, as currently employed, is necessarily an encompassing one because it is applied incontrovertibly to numerous parties/movements that nonetheless differ from each other in other ways, and also, more contestably, to certain political formations onto which not every recognized specialist is willing to confer that designation. Some analysts have preferred descriptions such as 'neo-populist' or 'right-populist' for movements not universally classified as 'extreme right.'

1. The Term 'Extreme Right'

The term 'extreme right' can be shown to have been used in a political sense, but not with quite its contemporary meaning and connotations, since the mid-nineteenth century and before. However, the term and its cognates are surprisingly recent arrivals in political science's standard vocabulary, having been regularly employed in the discipline only since the late 1950s. Lipset, for example, used the term very occasionally in his *Political Man* (1959/1963), it became conventional usage among European (except German) commentators during the 1960s, and it achieved final recognition in American political science with Lipset and Raab's *The Politics of Unreason* (1970/1971). One reason for the slowness with which the actual term came to be adopted in the USA was the popularity of the alternative, and still used, 'radical right' in order to describe political phenomena that most, but not all, present analysts would now call 'extreme right.' Lipset had also been responsible for the introduction of the term 'radical right' into US political science, and so for its popularity during much of the following decade. It was perhaps taken from the usage in German of *rechtsradikal*. This neologism, although also postwar, can be traced back to its introduction into German in the early 1950s, probably influenced by the fact that Germany's most significant extreme right party was then the later proscribed *Sozialistische Reichspartei* (SRP). This designation, though sometimes also *extrem rechtsradikal*, was widely used by German authors in the 1960s and 1970s, and has only relatively recently been superseded among political scientists by *rechtsextrem/rechtsextremistisch*. The distinction was also a bureaucratically important one in Germany, since a purported and casuistical distinction between *rechtsradikal* and *rechtsextrem* was the basis on which the German Office for the Protection of the Constitution justified varying monitoring and control strat-

egies toward different right-wing formations. By the 1990s extreme right activities in Germany made this distinction increasingly dubious and inoperable, and the latter term now in practice subsumes the former one.

2. A Definition of 'Extreme Right'

A definition of extreme right parties/movements has necessarily to be extensive and complex, for it must be appropriate for such movements as diverse in time and location as the American Anti-Masonic Movement of the 1820s; the Know-Nothing Movement of the 1850s; the French Boulangist Movement of the 1880s; inter-war fascism in Italy and other European countries; Nazism in Germany; McCarthyism in 1950s USA; Poujadism in 1950s France; George Wallace's 1968 American Independent Party (AIP); more recent European examples such as the French *Front National* (FN) and the German *Die Republikaner* (REPs); Latin American paramilitaries in countries such as Nicaragua and Colombia; Vladimir Zhirinovskiy's Liberal Democratic Party in Russia; and even purportedly more mainstream political parties of the present such as Austria's *Freiheitliche Partei Österreichs* (FPÖ) and Switzerland's *Schweizerische Volkspartei* (SVP). The contemporary American militias are usually regarded as extreme-right, but they are idiosyncratic examples, their most obvious European analogue perhaps being the militant neo-Nazi defense groups of Germany and some Scandinavian countries. Although not all analysts would agree on the 'extreme right' designation for every one of these, each has nonetheless attracted many such attributions from historians, political scientists, or informed journalist observers. Commentators from some countries (such as France and The Netherlands) have been more ready than those elsewhere unwaveringly and without collateral justification to label the contemporary examples as 'extreme right.' An inclusive definition of the extreme right has to be based on ideological features such as those below and, although each would not necessarily apply with equal emphasis and significance to all extreme right parties/movements, all political scientists using this concept would understand that many or most would, at leadership or mass-support level, characterize a political phenomenon so designated:

- selective inclusion, especially within the nation and often based on imposed assumptions about ethnic or religious similarity (which may take the form of aggressive nationalism or ethnocentrism);

- selective exclusion, directed against foreigners and/or indigenous or immigrant ethnic minorities, but also against social minorities such as homosexuals, travelers (gypsies), or in the past even the Masons;

- racism, based on biological perspectives about supposed inferiority or on the fabrication of ethnic and cultural boundaries;

anti-Semitism, which may be both specific and directed at contemporary targets, but which may in some cases involve denying, or mitigating the scale of, the Holocaust and glorifying the deeds of the Third Reich;

a preference for authoritarian initiatives by a strong state;

the cult of a leader figure;

a preference for a hierarchical social order;

antisystemic and antipluralist political perspectives;

overt hostility to political opponents, often associated with the use, encouragement, or tolerance of violence;

low tolerance of social change; and

nostalgia for the past.

2.1 Extreme Right Phenomena in the Twentieth Century

It is useful to give a very general, if crude, periodization for extreme right phenomena of the twentieth century, although in each of these periods there was nonetheless a considerable variety of such parties/movements:

1920–45. The era of ‘classic’ anti-Semitic Nazism and fascism.

1945–70. The era of the ‘old wave’ extreme right, often comprising anti-Semitic neo-Nazi/neofascist movements with a yearning for the ‘classic’ era (e.g., the SRP or Italy’s *Movimento Sociale Italiano*, MSI) and nostalgic anticommunist movements (e.g., McCarthyism).

1975–present. The era of the ‘new wave’ extreme right, appealing particularly on a basis of xenophobia and racism against ethnic minority immigrant groups and asylum seekers.

This periodization is not perfect, particularly because of anomalies among the US cases: the AIP, for example, anticipated the issues of the ‘new wave’ extreme right. However, it is a useful heuristic device for discussing theoretical approaches to the analysis of these movements.

3. Theoretical Approaches to the Extreme Right

Fashions in political science towards the analysis of the support of extreme right parties/movements have not been constant. While it is important not to exaggerate these changes, it is true that research in the 1950s and into the 1970s, conducted on the vote for Nazism (the *Nationalsozialistische Deutsche Arbeiterpartei*, NSDAP) and on ‘old wave’ postwar movements, was especially concerned with answering questions about the partisan origins and the social base of these examples of extreme right support. The lack of any individual-level data on voters for the NSDAP led to the use of different combinations of aggregate data in a substantial body of research about

Nazism’s partisan origins and social base. The most recent and comprehensive research in this tradition has argued the case for considerable social variety in the NSDAP’s later support. Postwar movements for whose support there existed survey data were frequently analyzed in the first instance in terms of the social categories to which they particularly appealed. Ideological and attitudinal dispositions towards the extreme right were not ignored, of course (especially not by research drawing on perspectives in *The Authoritarian Personality*, Adorno et al. 1950), but they were not always treated as the principal explicandum; indeed, many authors (e.g., some of those writing about McCarthyism in *The Radical Right*, Bell 1963/1964) simply imputed motive. Researchers working on movements that have emerged in the ‘new wave’ era have been more inclined to recognize that most extreme right movements draw from a variety of social sources and have given greater attention to the attitudinal, psychological, and contextual bases of such support. The disproportionate support for many ‘new wave’ extreme right movements among the self-employed and among younger voters from manual-worker backgrounds (especially males) has been noted (e.g., in the cases of the FPÖ, the FN, and the German *Deutsche Volksunion*, DVU), but it has not been a central plank to theorizing about these movements.

In giving an overview of the theoretical perspectives towards extreme right parties/movements, it is essential to distinguish between the pre-World War II and postwar cases. Many of the interbellum movements played major political roles, especially so, of course, in the cases of Italy and Germany. Even in countries where such movements did not achieve power, they often had substantial impact upon national politics during the 1930s (as in France or The Netherlands). On the other hand, in the postwar period extreme right parties/movements, though not without some important effects upon the content of political agendas, have not usually achieved actual governmental office—significant exceptions being Italy briefly in 1959 and during the mid-1990s, Austria in 2000, as well as the problematic cases of Spain till 1975, and of contemporary Switzerland in the light of the recent metamorphosis of the SVP towards the extreme right.

Most contemporary theoretical analyses of the interbellum extreme right focused on German Nazism and Italian fascism. Some influential perspectives came from self-avowed Marxists and, though incorporating views about the social base of these movements, attributed to them a particular role in the balance of class forces. Perhaps the more subtle of these Marxist perspectives were those by Gramsci, who used a concept of Caesarism, and by Trotsky and Thalheimer, who drew Bonapartist analogies and regarded fascism as a resort by capitalists to perpetuate their interests in a situation of class stalemate between the proletariat and capital. Postwar Marxists have

sometimes resurrected such perspectives in looking at Italian fascism and Nazism and have even sought to apply them to postwar movements, but, to most analysts, these approaches are no longer relevant. Instead, the great majority of political scientists writing on the 'old and new wave' extreme right have sought, albeit in varying ways, to focus on the social situations that may produce extreme right sympathy and behavior.

A widespread feature of numerous theories of both the 'old and new wave' extreme right is to see them as supported by those who in some way have suffered, or see themselves to have suffered, from social change. This theorizing may take the form of predicating status loss (e.g., Hofstadter and Lipset on McCarthyism), or disorientation as a result of social strain (e.g., Parsons), or a reaction against the growth of postmaterialist values (e.g., Ignazi), material loss due to processes of modernization or globalization (e.g., Heitmeyer). Theories with a social-psychological basis, as by those who have applied perspectives derived from *The Authoritarian Personality* and the Frankfurt School to 'old and new wave' movements, also assume a resistance to, and intolerance of, certain types of social change. 'New wave' extreme right movements since the 1970s, because in many cases they have made racism and xenophobia the principal basis of their appeal, have particularly attracted theorizing seeking to explain the presence of such attitudes among individual voters and certain groups of voters.

4. *The Tendency of Extreme Right Parties to Schism*

Extreme right parties/movements frequently exhibit distinctive organizational characteristics in comparison with mainstream political parties. The former have a particular tendency to splits, and contain strong schismogenetic features. This is sometimes because of personal animosities between leaders, who are often psychologically disposed to be authoritarian and are intolerant both of different perspectives and also of rivals. Because of the variety of ideological appeals offered by extreme right parties, their leaderships and supporters often have a mixture of social origins and this too can lead to divisive differences over minutiae of ideology (a characteristic shared with extreme left movements). Extreme right parties, particularly those in the postwar 'new wave' period, have exhibited a remarkable tendency to fracture. In France, Germany, Great Britain, The Netherlands, and Switzerland, to name only some examples, the indigenous extreme right has had a history of division and internal dissension. Indeed, many of the apparent electoral successes in the 1980s and 1990s by extreme right parties foundered on the organizational collapse of the elected legislative grouping, something that has been a

particular feature of the German extreme right parties, the REPs and the DVU.

5. *The Future for Extreme Right Parties/Movements*

It would be scientifically insupportable to suggest that such parties/movements will wither away. A troubled world is likely to provide just too many dynamics to underpin their emergence, especially in certain Third World contexts. In the liberal democracies, however, despite some successes in obtaining municipal control (as the FN has had since 1995) and breakthroughs by parties such as the FPÖ and the SVP (both of which emerged from mainstream political formations), there are few indications in most countries that, on a national level, extreme right parties will emerge on any scale from a political exclusion zone. Support even for the most successful examples has in most cases reached a zenith and elevated levels of one-time support frequently erode. These parties depend on a pariah status for attracting the votes of the politically alienated, and sometimes, as in France, where a party faction has wanted to seek accommodations with mainstream politics, the result has been a damaging split to the greater disadvantage of the faction arguing for such a rapprochement.

See also: Anti-Semitism; Authoritarian Personality: History of the Concept; Authoritarianism; Fundamentalism (New Christian Right); Marxism and Law; McCarthyism; National Socialism and Fascism; Nazi Law; Racism, History of; Racism, Sociology of; Radicalism; Right-wing Movements in the United States; Women and Gender; Xenophobia

Bibliography

- Adorno T W, Frenkel-Brunswick E, Levinson D J, Sanford R N 1950 *The Authoritarian Personality*. Harper, New York
 Bell D (ed.) 1963/1964 *The Radical Right: The New American Right Expanded and Updated*. Anchor, Garden City, NY
 Betz H-G 1994 *Radical Right-wing Populism in Western Europe*. Macmillan, Basingstoke, UK
 Betz H-G, Immerfall S (eds.) 1998 *The New Politics of the Right: Neo-populist Parties and Movements in Established Democracies*. 1st edn. St. Martin's Press, New York
 Hainsworth P (ed.) 2000 *The Politics of the Extreme Right: From the Margins to the Mainstream*. Pinter, London
 Heitmeyer W, Bushse H, Liebe-Freund J, Möller K, Müller J, Ritz H, Siller G, Vossen J 1992 *Die Bielefelder Rechts-extremismus-Studie: Erste Langzeituntersuchung zur politischen Sozialisation männlicher Jugendlicher*. Juventa, Weinheim, Germany
 Ignazi P 1994 *L'Estrema Destra in Europa*. Il Mulino, Bologna, Italy
 Kaplan J, Weinberg L (eds.) 1998 *The Emergence of a Euro-American Radical Right*. Rutgers University Press, New Brunswick, NJ

- Kitschelt H, McGann A J 1995 *The Radical Right in Western Europe: A Comparative Analysis*. University of Michigan Press, Ann Arbor, MI
- Lipset S M 1959/1963 *Political Man: The Social Bases of Politics*. Anchor, Garden City, NY
- Lipset S M, Raab E 1970/1971 *The Politics of Unreason: Right-wing Extremism in America, 1790–1970*. Heinemann, London
- Merkel P H, Weinberg L (eds.) 1997 *The Revival of Right-wing Extremism in the Nineties*. Frank Cass, London
- Pfahl-Traughber A 1994 *Volkes Stimme?: Rechtspopulismus in Europa*. Verlag J H W Dietz Nachfolger, Bonn, Germany
- Scheuch E K, Klingemann H D 1967 Theorie des Rechtsradikalismus in westlichen Industriegesellschaften. In: Ortlieb H-D, Molitor B (eds.) *Hamburger Jahrbuch für Wirtschafts- und Gesellschaftspolitik*. J C B Mohr, Tübingen, Germany, pp. 11–29

C. T. Husbands

Partner Selection across Culture, Psychology of

Although family structures, forms, and functions vary across cultures, marriage is a central universal phenomenon in the lives of most people. Selection of a partner is a key facet of marriage, and its mechanisms vary in consonance with the social psychological construction of the concept and meaning of marriage in a given culture. Every culture has rules or normative frameworks governing the orientation towards partner search and selection.

Interestingly, a range of changes is evident in partner selection within and across cultures, some of which are a consequence of the shifts in the macro contextual conditions, such as socioeconomic development and the women's movement, which have impacted various cultural structures and processes. Importantly, these factors have introduced transformation in social psychological orientations, and created novel modes of interpretation of, and greater flexibility in the adherence to, the traditional cultural rules for partner search and selection.

1. Theoretical Perspectives on Partner Selection

On what basis and how do people select partners for marriage? This question has elicited considerable scientific interest, and transcends disciplinary boundaries. Different disciplines have generated theories and models to study partner selection systems and strategies. In general, however, two major theoretical perspectives are evident: the evolutionary perspective and the social psychological perspective. A recent trend is to merge disciplinary perspectives and approaches in order to comprehend the complexities involved in partner search and achieve a clearer

understanding of the motives that determine the different elements of the process.

1.1 Evolutionary and Social Psychological Perspectives: Towards a Synthesis

Evolutionary theories enable understanding of the ultimate significance of heterosexual relationships, specifically the determinants of partner selection, whereas social psychological perspectives provide insight into the more proximate factors involved.

The evolutionary models of partner selection essentially view individuals as acting on the basis of evolved mechanisms, selected with a focus on maximizing ancestors' genetic fitness or increasing the likelihood of raising healthy offspring (Kenrick et al. 1993). These models contend that gender differences in preferences or selection criteria are cross-culturally universal, and have unfolded through sexual selection pressures over millions of years. As the physical condition of females is necessary for offspring survival, males generally tend to focus upon characteristics indicative of this aspect. On the other hand, females place more value on males' ability to contribute to resources.

A prominent theoretical model in social psychological perspectives on partner selection is the social exchange model, which posits that individuals search for partners who have resource value that is equitable with one's own 'market value.' In other words, there is an exchange of valued traits or resources (Kenrick et al. 1993). Based upon the market value of each trait, which varies cross-culturally, both women and men seek partners who will maximize gains and minimize costs. Social psychological models generally explain gender differences in terms of structural powerlessness of women and traditional sex role socialization patterns. Evolutionary theories, on the other hand, emphasize gender differences in terms of the reproductive investment hypothesis (Buss and Barnes 1986).

Evidence of the universality of certain gender differences in partner selection preferences across cultures supports the role that evolutionary forces play in shaping cultural factors that mediate the gender-based criteria for partner selection (Buss et al. 1990). Nevertheless, predictions related to psychological mechanisms in selection preferences offered by evolutionary models need to be viewed in conjunction with specific cultural-environmental contexts (see *Evolutionary Social Psychology*).

The sexual strategies theory proposes an integrated contextual-evolutionary perspective of partner selection. Based on the premise that human mating is inherently goal-directed, the theory incorporates factors from the larger context that have a bearing on partner selection and states that mating strategies are context-dependent in terms of the social milieu and the duration of the relationship. Also, the principles that govern selection by women and men are different and,

to that extent, the psychological strategies or mechanisms adopted to maximize reproductive success and offspring survival also differ (Buss and Schmitt 1993).

2. Culture as Mediator of Psychological Orientations

Culture is a significant mediating factor in partner selection. The cultural context provides shared schemas of meaning and knowledge systems that serve as a framework for individual experiences. The individual as a self-reflective and interpreting subject constructs and modifies the cultural schemas, thereby engendering transformation in the cultural rule systems for different domains, including partner selection. Thus, the cultural schemas on the one hand serve as a sort of template for individual and collective experiences, and on the other hand the individual constructs and modifies them in the course of cultural history (Eckensberger 1990) (see *Cultural Psychology*).

Social relationships and rule systems are at the core of culture. In cross-cultural psychology, these are conceptualized in terms of two main differentiations of self-concerned or other-concerned individual psychological orientations, related to Western and non-Western contexts, respectively (Markus and Kitayama 1991) (see *Cross-cultural Psychology*). The dichotomy of individualism and collectivism is commonly used to conceptualize the two types of cultures. Individualism emphasizes individual interests, needs, goals, and independence; collectivism refers to an orientation towards collective interests and norms with the aim to maintain social harmony (Hui and Triandis 1986) (see *Personality and Conceptions of the Self*).

The implications of these orientations for the structure and function of the primary social group, that is, the family, are of particular interest, in terms of the impact upon significant life events of individuals, including marriage and its integral process—selection of a partner.

3. Cross-cultural Perspectives

Multicultural studies of partner selection reveal interesting patterns of commonalities and diversities within and across cultures, which can be substantially attributed to changes in macro contexts as a result of industrialization and modernization.

Traditional and modern or collectivist and individualist cultures represent different clusters of qualities preferred in a prospective mate. Traits such as chastity, and domestic characteristics including desire for home and children, and being a good housekeeper are valued in traditional-collectivist cultures such as in China, India, Indonesia, Taiwan, Iran, the Palestinian Arab community, South Africa, and Colombia. The modern-individualist cultures (e.g., the USA, Canada, most of Western Europe)

consider such characteristics as irrelevant in the partner selection criteria or place a comparatively low value on the same. Relatively uniform gender differences across cultures are observed specifically in men's emphasis on physical appearance and value for traditional characteristics such as domestic skills, and women's emphasis on resource potential. Sexual dimorphism is evident most in the collectivist Asian and African cultures (Buss et al. 1990).

In the Western world, partner selection is essentially based on one's individual criteria, largely independent of familial and societal rules. Also evident is the growing utilization of innovative formal searching services. As one focuses on non-Western societies, specifically some selected Asian and African ones, the scenario reflects an interesting melange of 'traditional' and 'modern' (including Western) patterns in the process of partner selection.

The South Asian countries of China and Japan (Jew 1994, Wang 1994) demonstrate the continuance of certain traditional patterns of partner selection, such as arranged marriages, alongside the changes being introduced by Western influences. Individuals who marry are beginning to have more decision-making power. Nevertheless, parental opinion and approval are generally sought, if only as a mark of duty, obligation, and respect for them.

The ongoing transformation across the non-Western societies in particular, reveals the dynamic interface between culture and the self-reflective intentional individual who is engaged in the complex process of cultural reconstruction (see *Action Theory: Psychological*). Traditional normative rule systems provide a common frame for individuals' experiences of partner selection. In response to the larger context of rapid economic and social influences, individuals are engaging in a process of reinterpretation and reconstruction of the existing rule systems. In this context, societies that are predominantly characterized by collectivist perspectives are beginning to incorporate elements of the individualist perspective, in indigenized forms. The contemporary scenario illustrates a 'transitional phase' with varying blends of 'traditional' and 'modern' or 'collectivist' and 'individualist' perspectives and practices of partner selection represented cross-culturally.

4. The Indian Context

In Hindu India, marriage or *vivaha* is an inevitable significant life ritual (*samskara*). Marriage is considered as an alliance between two families, rather than two individuals.

4.1 Cultural Rules for Partner Selection and Emerging Social Psychological Orientations

Caste (*jati*) compatibility is the most important factor in a Hindu marriage. The position of the family in the

caste hierarchy determines the field of selection. This rule is now applied with greater flexibility, and there are incidences of intercaste marriages. A growing emphasis on individual characteristics of the prospective partner is also evident.

Although the arranged marriage remains a largely prevalent strategy of partner selection, the system has been modified to accommodate the preferences of adult children, and the traditional 'unconsented' model has given way to a 'consented' model (Banerjee 1999). The late 1990s also witnessed the growing involvement of external intermediaries, such as matrimonial advertisements and marriage bureaus, in partner selection.

The contemporary Indian context of partner selection reflects the influences of industrialization and modernization, including the global women's movement and the ensuing changes in values and accepted practices. These changes are revealed in the greater assertion of individual choice, the parallel instances of resistance to certain normative rules, and the diluting role of family and kin networks. Although the evolving climate is more conducive to the articulation of individualistic orientations, the value of family interdependence persists.

The Indian scenario depicts the continued resilience of traditional practices, alongside its adaptive features created to 'fit' with the transforming social context. It thus reveals the two-dimensional interdependent-independent or relational-autonomous orientation characteristic of many non-Western contexts (Kagitcibasi 1997).

5. Future Directions

How cultures operate to 'mate' individuals is a significant issue that provides a rich context for studying the interaction among psychological phenomena, their developmental patterns, and the cultural context. The emerging trend of integrating the social psychological or cultural and evolutionary models needs to be strengthened. Furthermore, indigenous paradigms for theory and research must be developed to trace the ongoing transitions and the diversity in developmental models across societies. Importantly, intracultural heterogeneity must be highlighted, especially in multicultural societies.

The validity of cross-cultural trends in specific gender-based selection preferences supported by the evolutionary model needs to be further corroborated. A critical question in this context is whether and to what extent women's education and economic independence actually enhance gender egalitarianism in the realm of partner selection. Importantly, interdisciplinary culture-based perspectives are imperative to develop conceptual frameworks for interpreting the wide-ranging diversity and complexity in the partner selection process.

See also: Action Theory: Psychological; Cross-cultural Psychology; Cultural Psychology; Cultural Variations in Interpersonal Relationships; Evolutionary Social Psychology; Family Theory: Economics of Marriage and Divorce; Love and Intimacy, Psychology of; Marriage; Marriage: Psychological and Experimental Analyses; Personality and Conceptions of the Self; Personality and Marriage

Bibliography

- Banerjee K 1999 Gender stratification and the contemporary marriage market in India. *Journal of Family Issues* **20**: 648–76
- Buss D M, Schmitt D P 1993 Sexual strategies theory: An evolutionary perspective on human mating. *Psychological Review* **100**: 204–32
- Buss D M, Abbott M, Angleitner A et al. 1990 International preferences in selecting mates: A study of 37 cultures. *Journal of Cross-Cultural Psychology* **21**: 5–47
- Buss D M, Barnes M 1986 Preferences in human mate selection. *Journal of Personality and Social Psychology* **50**: 555–70
- Eckensberger L H 1990 From cross-cultural psychology to cultural psychology. *The Quarterly Newsletter of the Laboratory of Comparative Human Cognition*, 1990a **12**: 37–52
- Hui C H, Triandis H C 1986 Individualism–collectivism. A study of cross-cultural researchers. *Journal of Cross-Cultural Psychology* **17**: 225–48
- Jew C 1994 Understanding the Japanese family structure from the Japanese perspective. *Family Perspective* **28**: 303–14
- Kagitcibasi C 1997 Individualism and collectivism. In: Berry J W, Segall M H, Kagitcibasi C (eds.) *Handbook of Cross-Cultural Psychology: Social Behavior and Applications*, 2nd edn., Allyn and Bacon, Boston, Vol. 3 pp. 1–49
- Kenrick D T, Groth G E, Frost M R, Sadalla E K 1993 Integrating evolutionary and social exchange perspectives on relationships: Effects of gender, self-appraisal, and involvement level on mate selection criteria. *Journal of Personality and Social Psychology* **64**: 951–69
- Markus H R, Kitayama S 1991 Culture and self: Implications for cognition, emotion and motivation. *Psychological Review* **98**: 224–53
- Wang G T 1994 Social development and family formation in China. *Family Perspective* **28**: 283–302

S. Kapadia

Partnership Formation and Dissolution in Western Societies

1. Partnership Formation

1.1 Marriage Trends

In the recent past, marriage heralded the start of a first union for most couples in Western societies and the great majority of marriages survived intact until one of the spouses died. There tended to be identifiable stages

in the development of a relationship: courtship, engagement, and ultimately the marriage ceremony that was followed by the couple setting up home together. Nowadays, there is more flexibility in becoming a couple and whether they co-reside.

After World War II, the general trend in marriage behavior in European and other developed countries was towards a younger and more universal marriage pattern which reached its zenith during the 1960s and early part of the 1970s (Festy 1980, Espenshade 1985, Fitch and Ruggles 2000). Since then, marriage rates in most Western countries have declined (UN Demographic Yearbooks and Council of Europe demographic publications provide collections of relevant data). Younger generations have been marrying less, and among those who marry the trend has been to do so at older ages and over a wider range of ages than was common among their recent predecessors. In broad outline in the case of Europe, the decline in marriage rates began in Sweden and Denmark in the late 1960s, and then spread through most of Western Europe in the early part of the 1970s, and became evident in the southern European countries (Spain, Italy, Portugal, and Greece) around the mid-1970s. Since the 1980s, the decline in marriage rates continued in most European countries but at a slower pace. Changes in the mean age at first marriage provide an illustration of the extent of these changes. In 1975, the average ages of first-time brides in most Western nations were clustered in the 22–4 years range, whereas in the mid-1990s they were clustered in the range 26–9 years. The pattern of change was similar for men, but they marry, on average, two years older than women do, an age difference that has been fairly invariable over time.

1.2 Cohabitation

One of the important engines behind the decline in marriage rates and a movement to a later age at marriage is the rise in cohabitation that has occurred, particularly since the beginning of the 1980s in many countries and even earlier in Scandinavia. However, it should be emphasized that men and women living together outside marriage is certainly not new. Prior to the 1970s, it was largely statistically invisible and probably socially invisible outside the local community or social milieu. For example, in some European countries there were subgroups that were probably more prone to cohabitation than others: the very poor; those whose marriages had broken up but who were unable to obtain a divorce, as there was no such legislation, or it was more stringent than nowadays, or it was very expensive to obtain a divorce; certain groups of rural dwellers; and groups ideologically opposed to marriage.

The form of cohabitation that came to the fore during the 1960s in Sweden and Denmark, and during

the 1970s in other northern and Western European countries, North America, and Australia and New Zealand, is new, and has been variously termed 'premarital cohabitation,' 'nubile cohabitation,' and 'cohabitation juvenile'; in this form, young people predominantly in their 20s and early 30s live together either as a prelude to, or as an alternative to, marriage. Additionally, with the growth in divorce, 'postmarital cohabitation' has also become more prevalent, with the divorced cohabiting either in preference to, or as a prelude to, remarriage. Unfortunately, in many data sources, it is difficult to distinguish between the 'premarital' and 'postmarital' forms of cohabitation. However, the increased prevalence of cohabiting unions is likely to lie behind much of the decline in first marriage and remarriage rates that have occurred in many Western countries in the latter decades of the twentieth century.

Data on cohabitation has tended to be scarce and generally emanated from *ad hoc* surveys. This made comparative analyses problematic, as sample sizes, coverage, and definitions could vary. During the 1990s more information from standardized questionnaires became available from Eurostat (the Statistical Office of the European Communities) and from a series of Fertility and Family Surveys carried out in the main in the first half of the 1990s under the auspices of the UN Economic Commission for Europe (ECE) (United Nations 1992).

1.2.1 The incidence of cohabitation. Analysis of data from Eurobarometer Surveys carried out in the 15 member states of the EU in 1996 provides a perspective on the incidence of cohabiting and marital unions across a range of nations. These surveys are primarily opinion surveys covering a range of topics relevant to the EU, but they contain some very basic demographic information on the respondents, including information on marital status in which 'living as married' is one of the categories, the others being the more conventional ones of single, married, divorced, separated, and widowed. Such marital status distributions may not be as accurate as those obtained in dedicated family and fertility surveys, but they probably reflect the relative position of different European countries in these categories (European Commission, 1996).

Figure 1 shows the proportions of women aged 25–29 years in the 15 EU countries who were cohabiting, married, single, or separated/divorced/widowed at the time of the survey in 1996. In these data never-married and postmarital cohabitants cannot be distinguished, but it is reasonable to assume that at these younger ages the former is likely to be the most prevalent. It is clear from Fig. 1 that there is a good deal of diversity across European states in the incidence of cohabitation. Cohabitation is strikingly most common in the Nordic countries of Denmark,

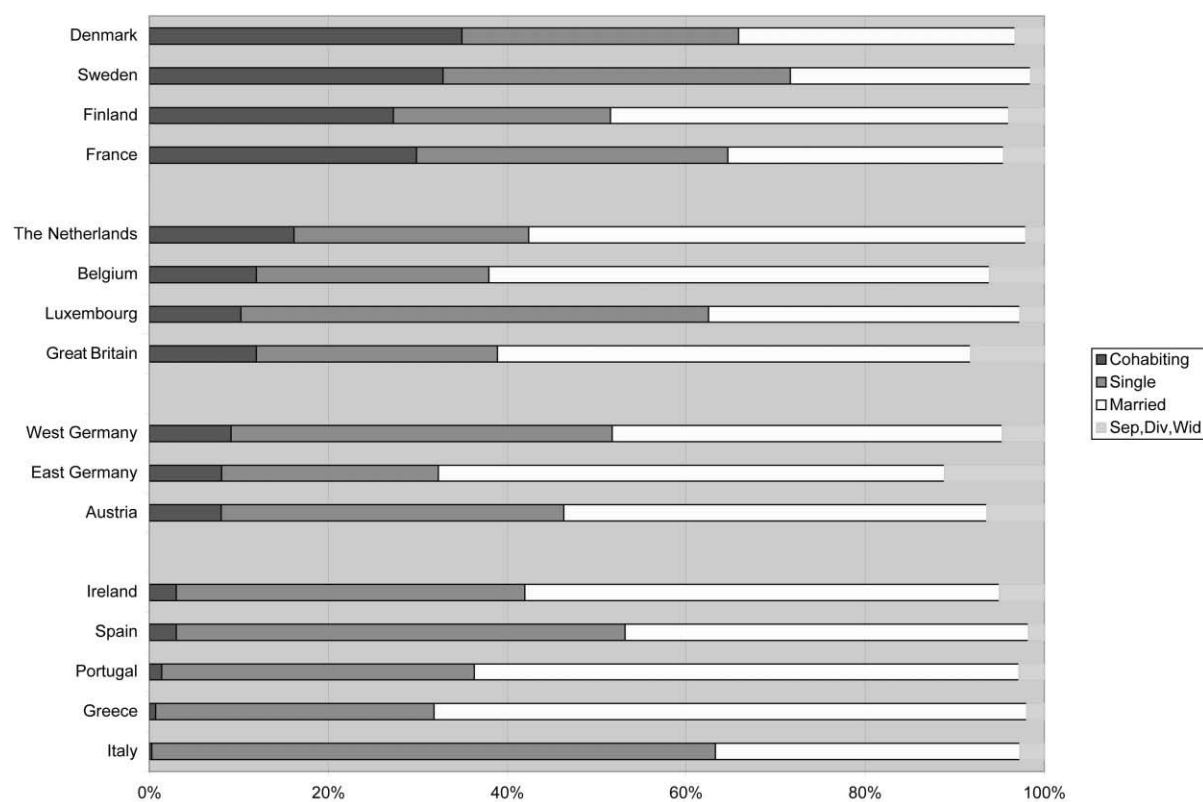


Figure 1
Marital status distribution of women aged 25–29 in 1996

Sweden, and Finland, and France also has relatively high proportions cohabiting. There is also a middle group of countries, which includes The Netherlands, Belgium, Great Britain, Germany, and Austria with mid-levels of cohabitation. Data for the US and Australia suggest that they would fall into this middle group. At the other extreme is the set of southern European countries and Ireland, where cohabitation is seemingly much rarer with only a tiny minority cohabiting.

1.2.2 Type of first partnership. The UN ECE Fertility and Family surveys carried out in the main during the first half of the 1990s included a full partnership history that incorporated dates of marriages and any other co-residential heterosexual intimate relationships. Such histories permit more in-depth examinations of partnership formation and dissolution than can be gleaned from vital registration data or cross-sectional surveys that only include current status information.

In many Western nations there have been large increases in the proportions of couples cohabiting, and nowadays cohabitation rather than marriage marks the formation of a union. Evidence on this can be seen in Table 1, which shows for two recent cohorts of women the proportions who entered their first partnership at marriage. It is clear from these data that the younger women, those aged 25–9, were much less likely to have commenced their first partnership at marriage than the older women. There are marked reductions to be seen in the proportions of women who married directly without cohabiting in most countries; for example, in France 55 percent of the older women but only 21 percent of the younger women married directly, a pattern that is repeated across many of the nations. The main exceptions are Sweden and the southern European countries. In Sweden, cohabiting rather than marrying was already well established among the older women whereas in Italy and Spain there are indications of a rise in cohabitation; but for the majority of women marriage still heralds the start of the first partnership. This is in contrast with the Scandinavian and other Western European nations where it is a minority practice.

Table 1

Percentage of women marrying directly among those who had had a first partnership according to current age group

Country	Age group	
	25–29	35–39
Sweden ^a	8	10
Norway ^a	10	42
Finland	17	31
France	21	55
Austria	21	33
Switzerland	19	31
West Germany	19	45
Great Britain	37	72
Italy	87	92
Spain	81	91

Source: Analysis UN ECE Fertility and Family Surveys and British Household Panel Study, Kiernan 1999

^a Nearest equivalent cohort

It is also the case that in many European countries cohabiting unions have simply replaced the marriages of yesteryear, in that compared with the recent past there has been little change in the proportions of men and women who have formed a residential partnership by their mid-20s, whereas in other countries (most noticeably the southern European states) cohabitation is only part of the story in the decline in marriage rates (Kiernan 1999). Here, young people have been spending longer periods of time as solos than in the recent past; living with their parents (in the main), on their own, or sharing with others (European Commission 1998).

1.2.3 Duration of cohabiting unions. In many developed countries cohabitation has eclipsed marriage as the marker for entry into the first union, but subsequently many of these unions convert into marriages and others dissolve. Life-table estimates of the proportions of cohabitations that had converted into marriages or dissolved five years after union formation for a range of European countries are shown in Table 2. There is some variation in the propensity to marry across nations and age groups. Sweden exhibits the lowest conversion to marriage, with only one in three cohabitations having become marriages within five years of the start of the partnership whereas in most other countries more than one in two cohabitations had converted into marriages by the fifth anniversary of the union. In several countries there are indications of a decline in the propensity to marry over time, most noticeably in Norway and France, whereas in other countries, such as Switzerland, there

Table 2

Proportions (derived from life-table analysis) of first cohabiting unions that had converted to marriages or dissolved by five years of start of union by age of woman

	Married 5 years	Dissolved 5 years
Sweden		
1964 ^a	34	37
1954	44	24
Norway		
1960 ^a	56	35
1950	81	29
Finland		
25–29	60	31
35–39	66	21
France		
25–29	63	31
35–39	78	17
Great Britain		
25–29	58	36
35–39	50	41
Austria		
25–29	54	26
35–39	50	18
Switzerland		
25–29	67	38
35–39	70	26
West Germany		
25–29	57	36
35–39	51	17
East Germany		
25–29	42	27
35–39	26	15

Source: Analysis of UN ECE Fertility and Family Surveys and British Household Panel Study

^a Nearest equivalent cohort

is little sign of change. Turning to the extent to which cohabiting unions dissolve, we see from Table 2 that in most countries among those aged 25–9 years, around one in three unions had dissolved by the fifth anniversary of their start.

1.2.4 Who cohabits? In addition to cross-national variation in union-formation behavior, there is also variation within nations and between subgroups of the population (see Carmichael 1995 for a review). There is now robust evidence that in most nations the younger generations are more prone to cohabit than were older generations, and the more secular mem-

bers of a society and those who had experienced parental divorce during childhood are also the more likely to cohabit. There is also evidence that those residing in metropolitan areas are more prone to cohabit and that being in full-time education tends to inhibit union formation, but the association of level of educational qualifications and employment status with cohabitation is less clear-cut and tends to vary across nations.

1.2.5 A partnership transition? Many developed societies may be going through a transition in the way that men and women become couples or partners (Prinz 1995). Drawing on the experience of Sweden, which is the nation that has gone furthest in these developments, a number of stages can be identified (Hoem and Hoem 1988). To simplify, in the first stage, cohabitation emerges as a deviant or *avant-garde* phenomenon practiced by a small group of the single population, while the great majority of the population marry directly. In the second stage, cohabitation functions as either a prelude to marriage or as a probationary period where the strength of the relationship may be tested prior to committing to marriage, and this is predominantly a childless phase. In the third stage, cohabitation becomes socially acceptable as an alternative to marriage and becoming a parent is no longer restricted to marriage. Finally, in the fourth stage, cohabitation and marriage become indistinguishable with children being born and reared within both, and here the partnership transition could be said to be complete. Sweden, Denmark, and, to a lesser extent, France are countries that have made the transition to this fourth stage. These stages may vary in duration, but once a society has reached a particular stage it is unlikely that there will be a return to an earlier stage. Moreover, once a certain stage has been reached, all the previous types of cohabiting unions can co-exist. Such stages also have parallels at the level of the individual. At any given time, cohabitation may have different meanings for the men and women involved; for example, it may be viewed as an alternative to being single, or as a precursor to marriage, or as a substitute for marriage. Moreover, how a couple perceives their cohabitation may change over time and may also vary between the partners. Dissecting cohabitation in this way highlights the diversity of the phenomenon and suggests that more so than marriage it is a process rather than an event.

In sum, analyses of recently available data on union formation show there to be marked variation in the ways men and women are forming partnerships across developed nations. In the 1990s marriage was still the pre-eminent marker for entry into first union in the southern European countries, whereas in most Western European countries, and in North America, and Australia, and New Zealand, cohabitation has eclipsed

marriage as the marker for first partnership. Notably, in the Nordic countries, long-term cohabitation has become more prevalent. Most Western countries appear to be experiencing changes in the ways that men and women become couples, but the question of whether most countries are on a trajectory to an ultimate destination where marriage and cohabitation are largely indistinguishable or even where cohabitation overtakes marriage as the dominant form of union awaits the future for an answer.

2. Partnership Dissolution

2.1 The Rise of Divorce

The other major development in partnership behavior is the rise of divorce. This has brought to the fore fundamental issues about the roles of men and women in society and the care and support of children. In most Western nations death still terminates the majority of marriages, but over the last few decades marriages have been increasingly dissolved by divorce before death has made any significant inroad and at a stage in the marriage when there are likely to be dependent children. In most Western countries divorce has increased since the mid-1960s, following a period of very stable divorce rates throughout the 1950s and the early part of the 1960s. Figure 2 shows trends since the early 1970s in the extent of divorce as measured by the number of divorces per 1,000 population for a range of Western countries. At the beginning of the 1970s the highest divorce rates were to be found in the US, and in Denmark and Sweden among countries in Europe. Divorce rates increased during the 1980s in most countries, and since then the rates have stabilized in many countries. Between 1960 and the mid-1980s divorce policy was either completely revised or substantially reformed in almost all the Western countries (Phillips 1988). Most countries liberalized their divorce laws, moving from fault-based divorce to no-fault divorce laws whereby fault, responsibility, or offense was no longer attributed by the law to either spouse. Following the liberalization of divorce laws, divorce rates in many countries continued their upward trend, frequently at a faster pace than in the years preceding legislative changes, followed by a period of stabilization from the mid- to late 1980s.

In the mid-1990s there was still a good deal of variation in the level of divorce across Western nations. In 1995 the total divorce rate in the EU was 0.30, whereas in 1970 it had stood at 0.11 (Eurostat 1997). This indicator is an estimate of the mean number of divorces per marriage for a notional cohort subjected to current divorce rates at each marriage duration. It can broadly be interpreted to mean that if the propensity to divorce remained unchanged over time, 30 percent of marriages would end in divorce. If

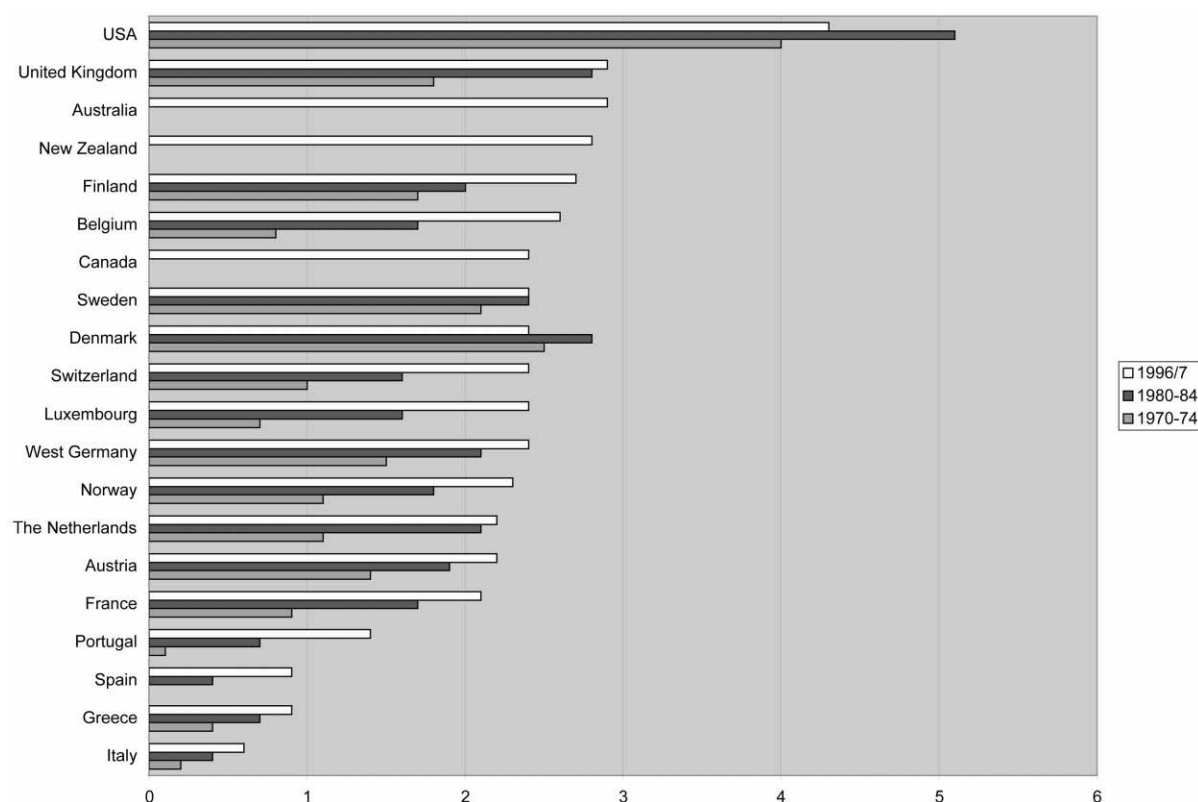


Figure 2

Crude divorce rates: divorces per 1,000 population

current rates were to prevail, nearly one in three marriages in the EU would dissolve. This is a lower level than in Canada (44 percent) and the US, where the total divorce rate has been above 50 percent since the late 1970s. Within Europe there is also a good deal of diversity in the level of divorce. During the 1990s three distinct divorce regions can be distinguished. In the Nordic countries and in England and Wales, the total divorce rate has been consistently above the 0.40 level. In Western European nations (Austria, Switzerland, Germany, France, Belgium, Netherlands, Iceland, and Luxembourg), the indicator lies between 0.30 and 0.40. In the southern European countries (Greece, Portugal, Spain, and Italy) it is below 0.20.

Divorce statistics have invariably underestimated the totality of marital dissolutions, and with the rise of cohabitation even fewer partnership dissolutions are captured by divorce registration data. In countries where divorce is more long-standing and more prevalent, there has been a tendency for marriages to break-up sooner. As a consequence, more couples are likely to be childless, and among couples with children the children are likely to be younger. Rising divorce has

led to a growth in lone-parent families, the residential separation of fathers from their children, and remarried couples and stepfamilies.

2.2 Lone-parent Families

The prevalence of lone parenthood varies considerably between countries, and the proportion of families headed by a lone parent has been increasing just about everywhere, as can be seen in Table 3. As yet, no Western European country has matched the US, where more than one in four families with children are lone-parent families. Various reports made for the EU (for example, Bradshaw et al. (1996) show that the great majority of lone-parent families are headed by a woman (80–90 percent). The largest group of lone parents comprises those who have experienced a marital breakdown, the next largest group comprises widows, and lone mothers who had never been married (but not necessarily never partnered) are the smallest category. In many countries where there have been

Table 3

Lone-parent families as a percentage of all families with dependent children

Country	1980/81 ^a	1990/91 ^a	1996 ^b
United Kingdom	14	16	23
Denmark	18	21	-
Finland	-	-	17
France	8	11	15
Belgium	9	15	15
Germany	10	15	13
Ireland	7	11	13
Portugal	12	12	12
Luxembourg	9	12	11
The Netherlands	8	12	11
Italy	7	6	11
Spain	5	8	8
Greece	4	6	7

Source: Bradshaw et al. 1996

a Eurostat LFS Social Europe 1/94—The European Union and the Family. b Eurostat Statistics in Focus: Population and Social Conditions 1998/12. Data in b from Labour Force Surveys

marked increases in childbearing outside marriage (but again not necessarily outside a partnership, see *Fertility of Single and Cohabiting Women*), this ordering may well have changed, such that never-married women with children may now constitute the second largest group of lone-mother families. Overall, the majority of lone-parent families emanate from the break-up of partnerships, either marital or cohabiting ones.

There is a good deal of evidence that children who experience the break-up of their parents' marriage or nonmarital union are more likely to experience poverty or reduced economic circumstances than children who grow up with both natural parents, but that the depth of poverty varies across nations (Bradshaw et al. 1996). The financial exigencies associated with marital breakdown arise from the precarious economic position of lone mothers, with whom most children live, and the dis-economies of scale associated with the maintenance of two households when fathers live apart from their children. The low incomes of lone-mother families are due to a combination of factors: low earnings from employment, lack of or low levels of child support from the natural father, and inadequate state support.

2.3 Remarriage

Being reared by a lone parent is frequently not a long-term arrangement, as a substantial proportion of divorced persons eventually remarry. Men are even more likely than women to remarry and are also more likely to remarry more quickly after a divorce. As well as being more likely to remarry, divorced men are more likely to cohabit than are divorced women.

Remarriages are also at greater risk of dissolution than are first marriages. After an initial upsurge in remarriage rates in the early years following the enactment of more lenient divorce legislation, which occurred in most European countries, remarriage has taken a downturn due to some extent to postmarital cohabitation becoming more common.

2.4 Children and Divorce

Parental divorce and its aftermath constitute a major factor in the collective make-up of the generation of children born since the 1970s in northern European countries, as they did for the generations born since the 1960s in the US, and from whose experiences much of our knowledge on this topic emanates. This factor is increasingly being experienced by the generation born during the 1980s in many other European countries. Parental separation has been shown to affect the lives of children in both the short and long term (Amato and Keith 1991a, 1991b). Following their parents' separation, children frequently go through a crisis, during which behavior problems at home and at school are more often reported, worries become more prevalent, and anxiety levels increase. After divorce, families may have to move house through necessity rather than choice, which in turn leads to changes in schools, and neighborhood and social networks. Poverty or at least reduced economic circumstances are likely to be a prominent feature of these children's lives.

A number of studies from a range of countries have shown that children who experience the breakup of their parents' marriage subsequently have lower educational attainment, receive lower incomes, and are more likely to be unemployed and to be in less prestigious occupations in adult life than their contemporaries brought up by both parents (Jonsson and Gähler 1997, McLanahan and Sandefur 1994). In the demographic domain, young women who have experienced parental divorce are more likely than their peers to commence sexual relations early, to cohabit or marry at young ages, to bear children in their teens, and to conceive and bear children outside wedlock (Kiernan and Hobcraft 1997, McLanahan and Sandefur 1994). Men and women from disrupted families are also more likely to experience the breakup of their own marriage (Glenn and Kramer 1987, Kiernan and Cherlin 1999).

3. Conclusion

Across many Western nations there have been dramatic changes in partnership behavior, most noticeably the rise of cohabitation and divorce. The general direction of change in most Western countries is similar but there continues to be marked variations

across nations (and probably within nations) in the extent to which these developments have taken hold. Marriage once signaled the start of a lifetime partnership for the great majority of couples, but it is increasingly being replaced by more flexible and contingent partnerships that have no formal commencement and that continue only as long as both partners derive satisfaction from the relationship. A range of sociological and economic explanations has been posited to account for these changes (see *Family Theory: Complementarity of Economic and Social Explanations*).

See also: Lone Mothers in Affluent Nations

Bibliography

- Amato P R, Keith B 1991a Parental divorce and the well-being of children: A meta-analysis. *Psychological Bulletin* **110**: 26–46
- Amato P R, Keith B 1991b Parental divorce and adult well-being: A meta-analysis. *Journal of Marriage and the Family* **53**: 43–58
- Bradshaw J, Kennedy S, Kilkey M, Hutton S, Corden A, Eardley T, Holmes H, Neale J 1996 *The Employment of Lone Parents: A Comparison of Policy in Twenty Countries*. Family Policy Studies Centre/Joseph Rowntree Foundation, London
- Carmichael G 1995 Consensual partnering in the more developed countries. *Journal of the Australian Population Association* **12**: 51–86
- Council of Europe 1999 *Recent Demographic Developments in Europe*. Council of Europe, Strasbourg, France
- Espenshade T J 1985 The recent decline of American marriage. In: Davis K (ed.) *Contemporary Marriage*. Russell Sage Foundation, New York
- European Commission 1996 *Eurobarometer No. 44*. ESRC Data Archive, Essex, UK
- European Commission 1998 *Social Portrait of Europe*. Office for Official Publications of the European Communities, Luxembourg
- Eurostat 1997 *Statistics in Focus: Population and Social Conditions No. 14*. Office for Official Publications of the European Communities, Luxembourg
- Festy P 1980 On the new context of marriage in Western Europe. *Population and Development Review* **6**: 311–5
- Fitch C A, Ruggles S 2000 Historical trends in marriage formation, United States 1850–1990. In: Waite L C, Bachrach C, Hindin M, Thomson E, Thornton A (eds.) *Ties that Bind: Perspectives on Marriage and Cohabitation*. Aldine de Gruyter, Hawthorne, IL
- Glenn N D, Kramer K B 1987 The marriages and divorces of children who divorce. *Journal of Marriage and the Family* **49**: 811–25
- Hoem J, Hoem B 1988 The Swedish family: Aspects of contemporary developments. *Journal of Family Issues* **9**: 397–424
- Jonsson J, Gähler M 1997 Family dissolution, family reconstitution and children's educational careers: Recent evidence from Sweden. *Demography* **34**: 277–93
- Kiernan K 1999 *Cohabitation in Western Europe*. Population Trends, no. 96. Stationery Office, London

- Kiernan K, Cherlin A 1999 Parental divorce and partnership dissolution in adulthood: Evidence from a British cohort study. *Population Studies* **53**: 39–48
- Kiernan K E, Hobcraft J N 1997 Parental divorce during childhood: Age at first intercourse, partnership and parenthood. *Population Studies* **51**: 41–55
- McLanahan S, Sandefur G 1994 *Growing Up with a Single Parent*. Harvard University Press, London
- Phillips R 1988 *Putting Asunder: A History of Divorce in Western Society*. Cambridge University Press, Cambridge, UK
- Prinz C 1995 *Cohabiting, Married, or Single*. Ashgate Publishing, Aldershot, UK
- United Nations 1992 *Questionnaire and Codebook: Fertility and Family Surveys in Countries of the ECE Region*. United Nations, New York
- United Nations (annual) *Demographic YearBook*. United Nations, New York

K. Kiernan

Party Identification

Party identification is one of the central explanatory concepts used in the analysis of voting behavior in mature democracies. It refers to the enduring affective (emotional) attachment a voter feels towards a particular political party which disposes him or her to vote for that party in elections. Republican 'identifiers,' for example, are much more likely than other voters to vote Republican in a range of different elections; Democrat identifiers are similarly more likely to vote Democrat. A distinction is sometimes made between 'strong' and 'weak' identifiers, depending on the strength of the individual's attachment to the party in question. Both of these groups are contrasted with 'floating' or 'independent' voters, who exhibit no obvious attachment to one party over time and who frequently switch their votes between parties at succeeding elections. Party identification is sometimes referred to as 'partisanship.' The two terms are used interchangeably here.

1. Intellectual Context

The concept of party identification was originally developed as an integral part of the Michigan social psychological model of voting behavior. Partly through socialization experiences in childhood and in the workplace, individuals come to feel that their own political beliefs and values are 'closer' to the core ideological and policy positions of one party rather than another. Some individuals will enter adulthood as fully fledged 'identifiers' with a particular party. For others, the process may take longer. But the more frequently that an individual votes for a particular party over time, the more likely it is that s/he will come to 'feel closer to' or 'identify with' that party: s/he will

become a 'party identifier.' This, in turn, will reinforce the individual's tendency to continue to vote for the party, even when other factors might militate against it.

Although party identification is primarily a psychological concept, it also has a direct counterpart in rational choice theory in Downs' concept of the 'standing vote.' According to Downs, voters can develop a habit of voting for one party as a way of reducing the costs of continuously collecting and evaluating political information. Relatively early in their political lives, they make judgements about which party is closest to their own ideological positions and tend subsequently to continue to support that party unless and until some major political trauma causes them to reassess their early judgement. Fiorina's work on voting in the USA goes some way to combining the psychological notion of affective attachment with Down's idea of the standing vote. Fiorina argues that party identification is best seen as a running tally, or cumulation, of voters' retrospective evaluations of the major parties. In short, identification results from a voter's perception that a particular party has outperformed its rival(s) over time; it disposes the voter to support that party in the current election.

2. *Emphases in Current Theory and Research*

The concept of party identification has been deployed in four main ways. The first, and most extensive, use is as a 'nonproximate' cause of vote choice. Voters are obviously swayed by a variety of different factors in deciding how to cast their votes in elections. These include voters' perceptions of any triumphs or disasters that occur in the final weeks of a campaign; their perceptions of the candidates' character; their views of the policy stances of the parties; and their assessments of the parties' performance in office and in opposition.

For advocates of the party identification approach, a crucial additional factor is the extent to which voters are predisposed—independently of any of the above 'proximate' factors—to support one or other of the parties because they 'identify' with it. On this account, the effects of proximate causes of the vote can only be properly estimated if the underlying effects of voters' prior partisan identifications have been fully taken into account. Concomitantly, explanations of voting choices (and election outcomes) that underplay the importance of long-term party identifications risk overstating the significance of the more proximate factors that explain voting choices. In all this, there is no suggestion that party identifiers always vote for the party with which they identify. On the contrary, it is entirely possible for even strong identifiers on occasion to vote against 'their' party because other more proximate factors weigh more heavily in their vote calculations. In US presidential elections, for example,

between five and twenty percent of identifiers typically vote 'against' the party they identify with. For these voters, party policies, party performance or the leadership qualities of the rival candidates (or some other proximate factors) must have weighed more heavily than traditional party loyalties. Party identification theory argues, however, that in subsequent elections 'defecting identifiers' are highly likely, other things being equal, to bring their voting behavior back into line with their partisan identity; they return to their traditional partisanship.

The second usage of the concept of party identification follows from the first. If prior identifications dispose citizens to vote in particular ways, then those identifications are also likely to color the way in which voters interpret political information. Party identification in this sense acts as a framing factor—a sort of perceptual filter—which disposes voters to view *their* party's policies, its leaders, and its record in a more favorable light than might otherwise be the case. It simultaneously disposes voters to view *other* parties' characteristics less favorably. This implies that voting choices are not only affected *directly* by party identification. These choices are also affected *indirectly* because of the 'coloring' effects of identification on voters' perceptions of the parties' competing attractions—perceptions which in turn also influence the way people vote.

A third use of the concept of party identification is made by the political parties themselves in devising electoral strategy. Party strategists are always aware of the need both to mobilize their party's own identifiers and to appeal to those voters who in the past have not been inclined to lend the party their support. In the early stages of a campaign, the main task is to trigger the support of existing partisans; in later stages it is to convert neutrals and those who identify with competing parties. Measures of party identification give strategists some idea of the size of their 'core vote.' This is the minimum level of support that parties can expect to obtain in a general election and against which, once the election outcome is known, their strategists can judge their success in recruiting new supporters. The need to provide participants in the democratic process with accurate information about their performance puts a particular premium on the accuracy of estimates of partisanship levels. For this reason debates about the measurement of party identification (which are discussed below) involve more than just matters of academic interest.

The final usage of the concept of partisanship is as a phenomenon in its own right that requires analysis and explanation. To argue, as party identification theory does in relation to any given election, that a major reason why voters vote Republican or Democrat—or Conservative or Labour—is that they are already Republican or Democrat, identifiers does not, on the face of it, seem to explain very much. Indeed, it borders on the tautological. The notion that 'I vote

Republican because I am a Republican identifier' immediately begs the question as to why I identify with the Republicans in the first place. As implied above, party identification theory's answer to this question is that partisanship develops as a result of long-term exposure to a variety of socialization processes in the individual's family, school, workplace, and social milieu.

Note, however, that socialization experiences of this sort fail to provide an obvious explanation as to why some voters *switch* their identifications (from Republican to Democrat or from Labour to Conservative, for example) midway through their voting lives. Fiorina's notion of identification as a cumulative updating of retrospective assessments of party performance plays a useful explanatory role in this context. Consider a voter A who, as a result of childhood and early adult socialization experiences, identifies with—and accordingly votes for—the Conservatives at election e1. Assume further that the Conservatives appear so incompetent at the subsequent elections e2 and e3 that A switches his vote to Labour on both occasions. In these circumstances, A may also shift his identification to Labour to reflect the fact that his 'running tally' of retrospective performance now favors Labour rather than the Conservatives. Although Fiorina's analysis does not predict the precise point at which voters will switch their identification from one party to another, it offers a plausible explanation as to how, notwithstanding their early political socialization, voters can change their identifications over time.

3. *Methodological Issues: The Problem of Measurement*

A key methodological issue concerns the extent to which the party identification measures developed in the USA in the 1950s can be successfully transplanted to other political contexts. Data from successive US National Election Studies (NES) suggest that roughly three-quarters of US voters are either Republican or Democrat identifiers (the remaining quarter are 'independents') and that substantial proportions of voters retain their partisan identities over time. Given that these findings are consistent with the idea of party identification being both widespread and durable, the concept of partisanship has maintained a prominent position in analyses of electoral outcomes and electoral change. Indeed, most published accounts of US voting behavior include either controls for, or estimates of, the effects of partisanship on the vote.

Not surprisingly, when efforts have been made to extend the theories and methods of the NES to other countries, the concept of party identification has been 'exported' as a central part of the package. An early example was the inclusion of a party identification question in the 1964 British Election Study (BES)

survey. The question was borrowed directly from the NES and it has been employed with only minor variations in question-wording since:

Generally speaking, do you think of yourself as Conservative, Labour, Liberal-Democrat, (Nationalist) or what?

The phrase 'generally speaking' is intended to tap the idea that party identification—as opposed to the more immediate and ephemeral notion of current party preference—endures over time; 'think of yourself' attempts to capture the respondent's political identity.

Most analysts of British electoral politics would endorse the idea that party identification plays an important role in explaining UK voting decisions. However, considerable doubt has also been expressed as to whether the traditional partisanship question measures anything very different from voters' current party political preferences. There are two major reasons for supposing that it does not. First, responses to the standard question outlined above suggest that, since the 1960s, 90–5 percent of British voters have had some sort of party identification. This is clearly a much higher level of 'measured' partisanship than in the USA, which should in turn imply much greater stability over time in British voting patterns. Yet electoral volatility since the 1970s has not been noticeably lower in the UK than in the USA. Second, it is clear from British data that major shifts in party identification levels (in 1979, 1983, and 1997) have coincided exactly with major shifts in voting behavior—implying that both, perhaps, are indicators of current political preferences rather than separate phenomena. This conclusion is supported by monthly time-series data collected by Gallup which show that short-term fluctuations in British partisanship correlate very strongly with short-term changes in party support levels. More formal tests, based on the principles of 'convergent validity' indicate that, in Britain at least, extant measures of party identification are statistically indistinguishable from measures of voting preference. This in turn implies that models of vote choice which include party identification as an explanatory variable are tautologically mis-specified.

Although these difficulties have not been precisely replicated in all of the countries where partisanship measures have been deployed, similar problems have been encountered in a variety of different national contexts. As a result, many analysts outside the USA have expressed concern that in non-American contexts (where the concept of the 'political independent' is less well developed) the standard NES question invites respondents to admit to a partisan identity even when they do not have one. This has led a number of researchers to advocate the use of a 'filter' mechanism in measuring partisanship, such as the following:

I am now going to ask you a question about party politics in [country]. Some people think of themselves as usually being a

supporter of one political party rather than another. Setting aside your current views about the government and opposition parties, do you usually think of yourself as being a supporter of one particular party or not?
If YES, which one is that?

The key feature of the filter is that it allows survey respondents to indicate from the outset that they do *not* usually think of themselves as being supporters of a particular political party. Survey experiments using filters of this kind have been conducted in Canada, the UK, the USA, and elsewhere. The measured levels of party identification fall considerably as a result. In Canada, switching from the traditional question to the filter approach reduces measured party identification levels from 70 percent to 40 percent of the electorate; in the UK, identification falls from around 90 percent to under 50 percent; and in the USA it falls from roughly 70 percent to 60 percent. The effects of using a filter, in short, are very considerable outside the USA, implying that levels of party identification in other countries may have been significantly overestimated in much academic research where the traditional NES question has been imported for local use. This is obviously important in purely descriptive terms but it also suggests significant measurement error problems in any statistical models that (outside the USA) have used party identification measures.

The problems of measuring party identification effectively do not stop here, however. A critical feature of partisanship is that it is, for the individual voter, an enduring feature of his or her political makeup. Partisanship should therefore be quite stable at the individual level over time, which implies that an individual should respond to the party identification question in a consistent way over time. Since 1991, the British Household Panel Survey (BHPS) has been collecting annual data on party identification in the UK using the sort of 'filter' question outlined above. In every wave of the panel, under 50 percent of respondents have indicated that they have a party identification. In the general election year of 1997, for example, 28 percent of the sample were Labour identifiers, 14 percent were Conservative identifiers, and 4 percent Liberal Democrat identifiers. As intimated earlier, these levels are clearly much lower than those recorded on the basis of the traditional NES question wording. But the measurement problem extends much further than this. Party identification is supposed to be a stable and enduring characteristic of individual voters. The BHPS reinterviews the same respondents in successive annual waves of the survey. We can therefore inspect the data to see how far individuals retain the same party identifications over time. As Table 1 shows, even if we consider only a 3-year 'rolling window' (i.e., we consider voters responses in waves 1–3, then waves 2–4, then 3–5, and so on), the levels of consistent identifiers fall even further. For example, in waves 1, 2, and 3 (1991–1993) only

Table 1

Percent of BHPS respondents with consistent and inconsistent party identifications over successive three-year periods, 1991–7

	W1–3	W2–4	W3–5	W4–6	W5–7
Consistent					
Conservative	9.5	8.9	7.6	7.4	6.9
Consistent					
Labour	7.9	9.4	10.5	11.1	12.5
Consistent					
Lib Democrat	1.3	1.1	1.0	0.9	1.0
Inconsistent	52.1	50.2	46.8	43.2	45.8
Non-identifier	29.2	30.4	34.1	37.4	33.8

Source: British Household Panel Survey

9.5 percent of respondents consistently responded as Conservative identifiers and only 7.9 percent consistently as Labour identifiers. In waves 5, 6, and 7 (1995–7), only 6.9 percent were consistent Conservative identifiers and 12.5 percent Labour identifiers. The remainder (excluding the trivial number of consistent Liberal Democrats) were either consistent nonidentifiers (33.8 percent) or else switched identifications across waves (45.8 percent). In short, only around 20 percent of British voters in the 1990s were 'genuine' party identifiers in the sense that the measuring instrument used: (a) allowed them to indicate from the outset that they had no particular attachment to any party; and (b) required them to declare the same general partisan commitment on three (and only three—the figures are even lower if the length of the rolling window is extended!) successive occasions.

Where does this leave us? Party identification is a theoretical construct that makes a great deal of intuitive sense. It tells us that individuals have long-term, enduring attachments to political parties and this idea in turn helps us to understand why individuals often vote for the same party in successive elections. Yet when we try to measure partisanship, things become far less clear. Generations of American scholars have happily employed the concept, confident that the NES question wording effectively captures the enduring nature of the partisan's affective commitment to his or her party. However, data from other countries—and especially from Britain—casts considerable doubt on the use of the concept in other national contexts. In the UK, large numbers of voters—much larger than are revealed by the NES question—appear to have no commitment to any political party. And of those that do, less than 20 percent retain that commitment over a three-year period. In short, if party identification denotes an *enduring* affective attachment to a particular political party, then partisanship in the UK at least is very much a minority predisposition.

Future research which seeks to employ the notion of party identification will need to address two major issues. The first, discussed above, is the need to develop operational empirical measures of the concept that are both durable and have substantive meaning in the specific national contexts in which they used. Critically, these new operational measures must distinguish effectively between voting choice and partisanship. The second issue concerns the impact of the longer-term process of 'partisan dealignment.' It is generally acknowledged that, however partisanship is measured, levels of party identification have gradually fallen in most democracies since the 1970s. Electorates have become increasingly volatile and voters increasingly prepared to behave like 'electoral consumers,' picking and choosing between parties on the basis of the rival policy and leadership packages on offer. In these circumstances, as Fiorina observed, partisan attachments acquired early in voters' political lives are likely to be less important in determining voting choices than the more temporary 'brand loyalties' associated with the perception that 'party X currently appears more competent in managerial terms than party Y.'

See also: Attitude Formation: Function and Structure; Citizen Participation; Identity: Social; Ideology: Political Aspects; Ideology, Sociology of; Participation: Political; Political Culture; Political Discourse; Political Parties, History of; Political Protest and Civil Disobedience; Polling; Public Opinion: Microsociological Aspects; Public Opinion: Political Aspects; Socialization: Political

Bibliography

- Butler D, Stokes D 1974 *Political Change in Britain: The Evolution of Electoral Choice*, 2nd edn. Macmillan, London
- Campbell A, Converse P, Miller W E, Stokes D 1966 *Elections and the Political Order*. Wiley, London
- Fiorina M P 1981 *Retrospective Voting in American National Elections*. Yale University Press, New Haven, CT
- Miller W E, Shanks M 1996 *The New American Voter*. Harvard University Press, Cambridge, MA
- Sinnott R 1998 Party attachment in Europe: methodological critique and substantive implications. *British Journal of Political Science* 28: 627–50

D. Sanders

Party Responsibility

Party responsibility is the attribution of collective responsibility to the political party in power for the performance of government. This contrasts with the attribution of individual responsibility to particular officeholders for their performance in office. Party

responsibility is the core idea of a model, known as responsible party government, of how democratic government should be organized in a modern populous nation.

1. The Party Government Model

The party government model was first developed by the American writers Woodrow Wilson, Frank J. Goodnow, A. Lawrence Lowell, and Henry Jones Ford in the last two decades of the nineteenth century, and served as a base for criticizing the existing American parties and proposing party reform. It lost favor in the Progressive Era and languished until the 1940s, when it was revived by another generation of American writers led by E. E. Schattschneider in *Party Government* (1942) and *The Struggle for Party Government* (1948), with support from Herman Finer in *The Theory and Practice of Modern Government* (1932) and James MacGregor Burns in *The Deadlock of Democracy* (1963). Its visibility was maximized in 1950 with the publication of *Toward a More Responsible Two-party System*, a report of the American Political Science Association's Committee on Political Parties.

The model is based on the proposition that democratic responsibility depends on the possession and exercise of power. If an individual or group does not control the actions of government, a sovereign people cannot effectively hold that individual or group responsible for those actions. Democratic governments in modern populous nations cannot be controlled by individual officeholders, not even US presidents. A member of Congress has only a small share of the total power over government, and he or she can be held responsible only for his or her votes and speeches, not for the performance of the entire government. Even the president cannot control the actions of Congress or the Supreme Court, and so cannot meaningfully be held responsible for governmental performance.

The only entity large enough to control the entire government and public enough to be held responsible for what it does or fails to do is a political party. Therefore, only the collective responsibility of the majority party can establish effective popular control of government.

The model has three basic requirements. (a) The parties must have near-perfect cohesion on all public issues: the members of each party must act together so that they can be held collectively responsible. (b) The voters must vote for a preferred party's candidates as representatives of the collectivity, not as isolated individuals. (c) The winning party must control the entire power of government, so that it can reasonably be held responsible for the government's actions and inactions.

Ideally, the model of responsible party government should work like this: at least two (and, preferably,

only two) unified, disciplined major political parties exist. In a pre-election campaign, each sets forth its proposals for measures to deal with the nation's problems, and the differences between the two sets of proposals are great enough to provide the voters with a meaningful choice. In the election the voters vote for certain candidates for Congress and the presidency or for Parliament, *not* because of their individual eloquence and probity, but because they are members of the party the voter thinks should control the government for the next term. One party receives a majority of the votes and public offices, and takes over complete governmental power and therefore assumes complete responsibility for what government does or fails to do. At the end of its term the voters decide whether the governing party's record is mainly good or bad. If they like it, they vote to return it to power. If they dislike it, they vote to replace it with the other party, which has pointed out the governing party's errors, and stands ready to take its place.

The doctrine's advocates generally present it as a model, a theoretical construct, the merits of which are to be assessed by theoretical reasoning and discourse. They do not wish its merits to be determined by how well or badly any actual governmental system performs. However, it is clear that most of them regard the government and party system of the United Kingdom as close enough to the model to provide a basis for answering questions about how the model would actually work in a nation like the United States.

2. *The Model Used*

From its origins the model of responsible party government has been used mainly to highlight certain perceived failings of the American party system and to prescribe remedies. Three such failings have received the most attention: first, the parties do not stand for anything. As Woodrow Wilson put it,

Provided with parties in abundance, and entertained with very nice professions of political principle, we lack party responsibility. American parties are seldom called to account for any breach of their engagements, how solemnly soever those engagements may have been entered into ... 'Platforms' are built only for conventions to sit on, and fall into decay as of course when conventions adjourn. Such parties as we have, parties with worn-out principles and without definite policies, are unmitigated nuisances. (*Committee or Cabinet Government?* Wilson 1925, p. 109)

Some commentators of Wilson's persuasion said that the American major parties are like bottles bearing different labels but equally empty.

They are uncohesive and therefore irresponsible. Votes in Congress rarely fall along strict party lines, and so it is impossible for voters to tell what each party—as distinct from its individual members—has

done on the issue. Hence, neither party can meaningfully be rewarded for good governmental actions or punished for bad ones. As a result, the American people can choose only between an incumbent officeholder and his or her opponent, neither of whom has much power over—and therefore cannot assume responsibility for—the actions of government. As the American Political Science Association (APSA) committee put it, the American party system consists largely of 'two loose associations of state and local organizations, with very little national machinery and very little internal cohesion ... [both of which are] ill-equipped to organize [their] members in the legislative and executive branches into a government held together and guided by the party program' (American Political Science Association, Committee on Political Parties 1950, p.v).

Some critics of the APSA report pointed out that the American constitutional system is an insurmountable barrier to the achievement of party government. Especially the institutions of separation of powers and federalism make it impossible for one party, by winning a majority of the votes and offices in any single election, to get full power over all governmental machinery and assume responsibility for what it does or fails to do.

Some advocates of party government, however, were not impressed. Schattschneider pointed out that in the past parties have had great success in making changes in the constitutional system without altering a word in the Constitution—for example, in the reshaping of the Electoral College system for electing the president—and he concluded that there are no 'grounds for excessive pessimism about the possibilities of integrating party government with the constitutional system. The greatest difficulties in the way of the development of party government in the United States have been intellectual, not legal. It is not unreasonable to suppose that once a respectable section of the public understands the issue, ways of promoting party government through the Constitution can be found' (Schattschneider 1942, pp. 209–10). The APSA Committee added that when the electorate becomes convinced that responsible parties are the best device for achieving popular control of government, the battle will be largely won, for the voters will insist that each party act together cohesively in a truly responsible manner.

3. *The Model Criticized, the Parties Defended*

From its beginnings the responsible party government model has had its critics. Perhaps the most prominent early opponent was Moisei I. Ostrogorski, a Russian writing in French. He first wrote an analysis of the British party system (especially interesting because it attacked the system often hailed as the working model

of responsible party government). His basic charge was rooted in his belief that the only meaningful form of democratic responsibility is the individual responsibility of each elected representative to his or her constituents. The worst thing about the British system, he said, is that 'the ministers ... easily hide behind the collective [party] title; however incompetent or culpable they may be, it is impossible to punish one of them without punishing all ...' (Ostrogorski 1902, I, p. 716).

His later (1910) volume on American parties charged them with many sins, of which the worst is that they are permanent conspiracies of political bosses interested only in winning elections so as to control government patronage. They have no interest in policy issues, and the voters cannot affect the course of public policy by voting for one party or the other. The only solution is to restore parties to what they were originally intended to be: *ad hoc* assemblies of like-minded citizens concerned with a particular policy issue, which will dissolve when that issue is settled.

The debate about what role parties should play in America and how they should be organized to play that role lay dormant until 1940, when E. Pendleton Herring published *The Politics of Democracy*. Herring argued that the unreformed parties had performed many valuable functions for American government and society, notably overcoming the potential for total gridlock inherent in separation of powers and federalism, stimulating ordinary people's interest and participation in politics, and moderating the ferocity of interest groups' demands by aiming their platform compromises and electoral appeals at the members of as many interest groups as possible rather than standing firmly and clearly for the interests of a restricted number of groups and the total rejection of those of other groups.

Herring's lead was followed by the authors of several works stimulated by *Toward a More Responsible Two-party System*. Political scientists Julius Turner and Austin Ranney argued in the *American Political Science Review* that the doctrine of responsible party government makes sense only in a democracy whose people are firmly committed to the implementation of unlimited majority rule, while in fact the American people have shown that they are more committed to the protection of minority rights. This explains why ordinary Americans have been so resistant to adopting the responsible party government model and its features of party discipline and cohesion and so admiring of individual independence in their officeholders. This same argument was developed at greater length in Ranney and Kendall's textbook *Democracy and the American Party System* (1956).

These defenders of the existing parties did not, however, silence the doctrine of responsible party government. The Schattschneider position was carried on and updated by, among others, the political scientist James MacGregor Burns (1963) and the journalist David Broder (1972). Moreover, since the

1970s both parties in Congress have become noticeably more cohesive and ideologically more distinct than they were in Schattschneider's (or Ostrogorski's) time. On the other hand, the proportion of voters voting split tickets has increased (to about 25 percent in the 1996 elections), and one of its main consequences, divided party government, with one party controlling the presidency and the other controlling one or both houses of Congress—a situation incompatible with responsible party government—has become the rule rather than the exception.

The net result is that while parties and officeholders in some parts of the national government are behaving more as the model requires, the voters' weakening party loyalties and increasing tendency to vote for individual candidates with little regard for their party labels are behaving less so. The dispute between the model's advocates and critics continues, but the party system changes only in some parts and in opposite directions.

See also: Accountability: Political; Electoral Systems; Party Systems; Political Parties; Political Parties, History of

Bibliography

- American Political Science Association, Committee on Political Parties 1950 *Toward a more responsible two-party system. American Political Science Review*, **XLIV** (Suppl.)
- Broder D S 1972 *The Party's Over: The Failure of Politics in America*. 1st edn. Harper & Row, New York
- Burns J M 1963 *The Deadlock of Democracy: Four-party Politics in America*. Prentice-Hall, Englewood Cliffs, NJ
- Croly H 1909 *The Promise of American Life*. Macmillan, New York
- Croly H 1914 *Progressive Democracy*. Macmillan, New York
- Finer H 1932 *The Theory and Practice of Modern Government*. Methuen and Company, London, 2 Vols.
- Ford H J 1898 *The Rise and Growth of American Politics*. Macmillan, New York
- Ford H J 1909 The direct primary. *North American Review* **CXC**: 1–19
- Goodnow F J 1900 *Politics and Administration*. Macmillan, New York
- Goodnow F J 1902 Review of Ostrogorski's *Democracy and the Organization of Political Parties*. *Political Science Quarterly* **XVIII** (June): 332–4
- Herring E P 1940 *The Politics of Democracy: American Parties in Action*. W. W. Norton, New York
- Lowell A L 1896a *Essays on Government*. Houghton Mifflin, Boston
- Lowell A L 1896b *Governments and Parties in Continental Europe*. Houghton Mifflin Company, Boston, 2 Vols
- Lowell A L 1902 The influence of party upon legislation in England and America. Annual Report of the American Historical Association for the Year 1901. Government Printing Office, Washington, DC, I, pp. 321–542
- Ostrogorski M I 1902 *Democracy and the Organization of Political Parties* [trans. Clarke F]. Macmillan, New York, 2 Vols

- Ostrogorski M I 1910 *Democracy and the Party System in the United States*. Macmillan, New York
- Ranney A 1962 *The Doctrine of Responsible Party Government*. University of Illinois Press, Urbana, IL
- Ranney A 1951 Toward a more responsible two-party system: A commentary. *American Political Science Review* 45 (June): 488–99
- Ranney A, Kendall W 1956 *Democracy and the American Party System*. Harcourt Brace, New York
- Schattschneider E E 1942 *Party Government*. Farrar and Rinehart, New York
- Schattschneider E E 1948 *The Struggle for Party Government*. University of Maryland Press, College Park, MD
- Turner J 1951 Responsible parties: A dissent from the floor. *American Political Science Review* XLV (March): 143–52
- Wilson W 1885 *Congressional Government*. Houghton Mifflin, Boston
- Wilson W 1879 Cabinet government in the United States. *International Review* 7(August): 146–63
- Wilson W 1925 Committee or cabinet government? In: Stannard Baker R, Dodd W E (eds.) *The Public Papers of Woodrow Wilson: College and State*. Harper and Brothers, New York, I, pp. 95–129
- Wilson W Leaderless government. *Public Papers* I: 336–59

A. Ranney

Party Systems

A party system is the system of interactions between political parties that results from their mutual competition or cooperation (see Sartori 1976, p. 44). For party systems to exist there must therefore be more than one party involved. One-party states do not have party systems. Beyond the requirement for a plurality of parties, party systems can be said to exist only when these interactions between the parties are patterned, familiar, and reasonably predictable. A plurality of parties is therefore a necessary but not sufficient condition for the existence of a party system. Parties that exist alongside one another, but that do not interact in any patterned or structured fashion, do not make for a party system as such.

While it may seem difficult to conceive of situations in which a plurality of parties exists without at the same time embodying a set of patterned interactions, two types of limiting cases can be identified. The first is where each of the parties involved exists in an entirely self-sufficient world, based on a closed electoral constituency, with neither its survival nor performance being in any way effected by the other parties in the polity. Even in this limiting case, however, the absence of interactions at the electoral level might well be accompanied by the need to interact—through competition or cooperation—at the parliamentary or governmental level. Thus, while no electoral party system would exist, it might well prove possible to speak of a party system in parliament. The

second type of limiting case is where a new multiparty democracy is being established more or less from scratch, such as in the wake of a democratic transition, and where neither the new parties themselves, nor their interactions, have stabilized sufficiently to allow the identification of a distinct pattern. In these situations, a party system may be in the process of being structured or institutionalized but it cannot yet be said to have definitely emerged. In their survey of party systems in Latin America, Mainwaring and Scully (1995) adopt a similar distinction in specifying party systems as either ‘institutionalized’ or ‘inchoate.’ Strictly speaking, however, an inchoate or non-institutionalized party system is a contradiction in terms: to be a system is to be institutionalized.

1. Classifying Party Systems

Although scholars have paid relatively scant attention to what defines a party system as such, tending instead to assume the existence of a party system in all polities where there exists a plurality of parties, they have devoted considerable effort to distinguishing between different types of party system. The most conventional and frequently adopted approach to distinguishing party systems is based simply on the number of parties in competition, and the most common distinction involved here, which goes back to Duverger (1954), is very straightforward—that between two-party systems, on the one hand, and multiparty (i.e., more than two) systems, on the other. This particular classification was also believed originally to reflect a more fundamental distinction between more or less stable and consensual democracies, such as the United Kingdom and the United States, which were seen as typical of the two-party variety, and more or less unstable or conflictual democracies, such as Fourth Republic France, Italy, or Weimar Germany, which were seen as typical of the multiparty type. Although this simple association of numerical categories of party system with levels of political stability and efficacy was later undermined by reference to a host of smaller democracies that were characterized by both a multiplicity of parties and a strong commitment to consensual government, the core distinction between two- and multi-party systems has continued to be widely employed within the literature on comparative politics, although it is sometimes modified by taking into account not only the sheer numbers of parties in competition, but also their relative size. Thus, Blondel (1968), for example, uses the relative size of the parties to distinguish four types of party system: two-party systems, two-and-a-half-party systems, multiparty systems with a dominant party, and multiparty systems without a dominant party.

The most substantial attempt to move away from a primary reliance on the simple numbers of parties in competition was that of Sartori (1976, pp. 117–323),

who combined counting the parties with a measure of the ideological distance that separated them. Sartori's typology was the first that focused directly on the interactions between the parties—the 'mechanics' of the system—and hence on the differential patterns of competition and cooperation. Following this approach, party systems could be classified according to the number of parties in the system, in which there was a distinction between systems with two parties, those with up to some five parties (limited pluralism) and those with some six parties or more (extreme pluralism); and according to the distance that separated the parties lying at either extreme of the ideological spectrum, which would either be small ('moderate') or large ('polarized'). These two criteria were not wholly independent of one another, however, in that Sartori also showed that the format of the system, that is, the number of parties, contained 'mechanical predispositions,' that is, it could help determine the ideological distance, such that extreme pluralism could lead to polarization. When combined, the two criteria resulted in three principal types of party system: two-party systems, characterized by a limited format and a small ideological distance (e.g., the UK); moderate pluralism, characterized by limited or extreme pluralism and a relatively small ideological distance (e.g., Denmark or the Netherlands); and polarized pluralism, characterized by extreme pluralism and a large ideological distance (e.g., Weimar Germany, early postwar Italy, or pre-Pinochet Chile). Sartori also noted the existence of a 'predominant-party system,' in which one particular party, such as, for example, the Congress party in India, or the old Unionist party in Northern Ireland, consistently won a winning majority of parliamentary seats.

2. Party Systems and Structures of Competition

At the core of any party system is the competition for executive office. It is this that structures the party system in the first place, and facilitates its institutionalization. Moreover, albeit often implicitly, the competition for executive office is also the most important criterion employed in many of the various classifications of party systems. Two-party systems, for example, are referred to as such not because there only two parties present themselves to the voters—indeed, this is rarely if ever the case—but rather because only two parties matter when it comes to forming a government, be this in a presidential or a parliamentary system. In two-and-a-half or multiparty systems, by contrast, there are more than two parties that enjoy the potential access to executive office. Even within the more complex classification developed by Sartori, it is the competition for office that proves most important. Sartori's moderate pluralism, for example, involves competition between alternative coalition

governments, whereas polarized pluralism involves a patterns in which a center party or parties is more or less permanently in office, with one or more extreme parties being permanently excluded from government.

It follows from this that party systems might best be understood not only in terms of some all-embracing classification, within which the question of the numbers of parties in competition obviously plays an important role, but also in terms of contrasting patterns of government formation (Mair 1997, pp. 199–223). Three related criteria are important here. The first concerns the ways in which governments alternate with one another, where systems in which there is a regular process of wholesale alternation, in which incumbents are replaced wholly by former nonincumbents, may be separated from those in which alternation is only partially complete. Systems in which two alternative parties or coalitions alternate with one another (the USA, the UK, Fifth Republic France) may therefore be distinguished from those in which one or more parties leaves office, while another remains, albeit often with a new coalition partner (Belgium, the Netherlands). The second, relevant criterion here is familiarity, where systems in which government is always contested by the same parties or sets of parties may be contrasted with those in which patterns of government formation prove more innovative or promiscuous. In postwar Ireland, for example, voters had long become accustomed to facing a choice at election time between, on the one side, the powerful *Fianna Fáil* party, which always governed alone, or, on the other side, a coalition of all of the smaller parties. Once *Fianna Fáil* decided to opt for a coalition strategy, however, in 1989, government formation became more innovative, and the formerly familiar patterns broke down. The third criterion involves the degree of access to the government formation process, and the extent to which new parties can hope to win a place in the executive. New parties have always found it relatively easy to join innovative government coalitions in the Netherlands, for example. In the UK, by contrast, no party other than the Conservatives or Labour has gained access to government since 1945.

Putting these three criteria together enables party systems to be ranged along a notional continuum according to the degree to which the structure of competition is open or closed. At one extreme lie the wholly closed systems, such as the UK or the US, where alternation in government is always wholesale, where government formation processes are wholly familiar, and where new or third parties are always excluded. At the other extreme, it is difficult to speak of a party system at all: there is no discernible pattern in how governments alternate, the potential alternatives themselves are unfamiliar and may never have governed before, and access to office is, in principle, open to all. To travel from this latter extreme to the other is therefore to witness the progressive closure of

the structure of competition, which is simply another way of saying that systemness itself increases. Party systems as systems are strongest and are most institutionalized when the structure of competition is closed. They are weakest, and more or less nonexistent, when this structure is wholly open.

To understand party systems in this way, and hence to treat the systemness of a party system, or its degree of institutionalization, as a variable, also affords an insight into the question of the persistence and change of party systems. Strongly institutionalized party systems will inevitably enjoy a bias towards persistence, not least because the parties at the core of such systems will have a vested interest in maintaining the existing structure of competition. It is for this reason, for example, that Lipset and Rokkan (1967) could speak of the 'freezing' of west European party systems in the wake of full democratization in the 1920s. Less institutionalized systems, on the other hand, will prove more susceptible to change. This approach also affords an insight into the relationship between party systems and individual parties, for it is not just the individual parties that together create the party system, but also the party system that acts to reinforce the role and standing of the parties themselves.

See also: Democracy; Dictatorship; Electoral Systems; Participation: Political; Party Identification; Party Responsibility; Political Parties, History of

Bibliography

- Blondel J 1968 Party systems and patterns of government in western democracies. *Canadian Journal of Political Science* 1: 180–203
- Duverger M 1954 *Political Parties: Their Organization and Activity in the Modern State*. Methuen, London
- Lipset S M, Rokkan S 1967 Cleavage structures, party systems and voter alignments: An introduction. In: Lipset S M, Rokkan S (eds.) *Party Systems and Voter Alignments*. Free Press, New York
- Mainwaring S, Scully T R 1995 Introduction: Party systems in Latin America. In: Mainwaring S, Scully T R (eds.) *Building Democratic Institutions: Party Systems in Latin America*. Stanford University Press, Stanford, CA
- Mair P 1997 *Party System Change: Approaches and Interpretations*. Clarendon Press, Oxford, UK
- Sartori G 1976 *Parties and Party Systems: A Framework for Analysis*. Cambridge University Press, Cambridge, UK
- Wolinetz S B, Brookfield U T (eds.) 1998 *Party Systems*. Ashgate, Aldershot, UK

P. Mair

Pastoralism in Anthropology

Pastoralism refers to a range of subsistence modes based on specialized herding of domestic animals on

free range, and culturally dominated by values associated with such an economic adaptation. Examples are the social formations based on small-stock oriented commercial meat production and long distance trading in the Middle East, on reindeer herding and hunting in the subarctic regions or on the subsistence-oriented cattle and small-stock herding in East Africa. A common synonym to 'pastoralist' is 'nomad.' Since adjustment to animal needs necessitates as well as makes possible a mobile adaptation, the whole pastoral household or parts of the labor force may be permanently or seasonally migrating. The extent of mobility is however in fact highly variable, and dependence on herds a more significant factor than mobility in itself, so that the term pastoralism is preferable. It is useful to consider as pastoralists mainly people for which livestock have a double function as both a means for the production of utilities such as foodstuff, wool, or skins, and as a form of capital with the potential of growth, primarily used for such production (Barth 1964, p. 70). This contrasts them to 'agro-pastoralists' using their herds primarily as an agricultural capital, for example, traditional cultivators in East Africa who invest in cattle to expand the household's agricultural labor pool through bride-wealth and polygyny. Pastoralism is labor-intensive in the sense of requiring a large number of hands for different tasks, and thus also commonly contrasted with, for example, ranching, where fencing makes a much less extensive use of labor possible.

Within the limits of such definitions, 'pastoralism' refers to a wide range of very different situations, regions, and times. Milk-based economies built on cattle/small stock in East Africa and on camels/flocks in the Horn of Africa date 2,000–3,000 years back. Pastoralists exploiting arid regions use species diversification as a risk reducing strategy: different species are subject to different hazards, have supplementary production patterns and require different labor. Contemporary East African pastoralism is based on an emphasis on cow or camel milk production for subsistence purposes, supplemented with small stock rearing to produce milk or meat for market sale or household needs. In the present situation there is often a change towards commercial production of small stock for the meat market, although authorities often encourage the sale of immature cattle to the livestock industry. The basic division of labor in such societies is that young men take care of mobile adult cattle and camels out of lactation, allowing them to profit from valuable but distant pastures. Women, children, and elderly men look after lactating dams and cows, calves, and small stock. These are often managed in a less mobile fashion if conditions allow. In East African cattle cultures, a careful attitude about the reproductive potential of the herd is a prime moral value that reflects on the fundamental worth of the pastoralist as a human being and as a member of the community. Reproductive capital could traditionally

only be disposed of in socially and ritually meaningful contexts such as gift-giving, bride-wealth, or stock friendships. The typical form of social organization among such pastoralists is a confederation of intermarrying clans, led by the property-holding senior men and identifying themselves as having a common ethnicity.

Middle Eastern pastoralism has an emphasis on sheep and goat rearing for the purpose of producing milk and/or meat for domestic consumption or sale. Such pastoralism has existed in the Middle East, around the Mediterranean and in Eastern and Northern Africa for 3,000 to 4,000 years, and is today considered to have developed there as a supplement to horticulture. In the millennium before Christ, camel-saddling technology had developed enough to enable pastoral groups to involve themselves in long-distance caravaneering as well as in extensive camel-back raiding. This brought the breeding of transport and riding animals into the herding operation, and above all involved the desert dwellers in important political and economic relations with sedentary riverine or oasis polities. The spread of Islam reflected the need to find a religious basis of trust when trading out of reach of tight jural or political control. Most shepherding communities of the Middle East are today Muslims, although religious restrictions are often less tight than in urban areas. Small stock rearing provided a subsistence basis for desert or mountain-based family members who did not take part in caravaneering, which was mainly a task for select males. An interesting difference exists between Saudi Arabia and North Africa on one hand and Somalia on the other, in that Somali pastoralists exploit the camel primarily as a milk animal and do not ride it while camel milk is less important where the animal is ridden.

In central Asia other combinations of livestock have enabled other matches of raiding, trade and subsistence dairy production, such as, for example, the traditional salt trade in the Himalayas where yak and sheep are used both for transport and dairy production. Mongolian pastoralism places the strongest emphasis on horses and sheep. Although there are a number of good studies on these communities (e.g., Humphreys and Sneath 1999), they have attracted less contemporary anthropological attention than have East African or Middle Eastern pastoralists. Due to the legacy of the Golden Horde they have instead appealed to the interest of cultural historians.

Reindeer-based pastoralism in the subarctic tundra is the only contemporary form of pastoralism based on specialization in one species. It is considered the most recently developed form of specialized pastoralism and thought to have developed out of hunting only during the last millennium. Herders follow the natural movements of reindeer herds whose primary use is the production of meat. Contemporary Sami herding in Scandinavia is market-oriented, motorized, and generally a large scale operation (Ingold 1980).

Evolutionary anthropologists and philosophers of the eighteenth and nineteenth century concerned themselves with pastoralism because it was assumed to be a transitional stage between hunting and agriculture. Engels, inspired by Morgan, took a particular interest in it, because he saw livestock holdings as the origin both of private property and of male domination. Serious field research among pastoral people was undertaken only in the 1940s and 1950s when monographs on groups such as the Nuer, Dinka, and Somali came to play a decisive role in the development of British structural-functional theorizing about patrilineal descent systems. Within other scholarly traditions, attempts were made to find suitable ways of systematically classifying pastoral economies into subtypes (full nomads, seminomads, etc.) often based on the relative time spent on the move or relative dependence on agriculture. These classificatory systems for a long time dominated what was written on pastoralism but were not theoretically productive.

Substantial progress in the study of pastoralism can be traced to three different but intertwined strands of research that were prominent in the late 1960s and throughout the 1970s. One followed American traditions of cultural ecology and focussed on adaptation. The second was prominent in France and pursued Marxist analyses of modes of production. The third was inspired by Barth's theories of economic action. The conjunction of these interests, followed up by participatory fieldwork, led to a boost in understanding the technical aspects of pastoralism, the complex set of resources required, and the problems of continuity involved in sustaining herd capital over time in arid regions characterized by unpredictable climatic conditions.

As noted, pastoral production involves both a production of utilities and a continuous reproduction of livestock capital. Paine (1971) made a useful distinction between herding and husbandry. The first referred to the care for the wellbeing and growth of animals as such and husbandry is the management of the herd in its totality. The latter takes place through investments, acquisitions, or culling and sales and involves the herd owner in socially highly significant transactions that establish relations of friendship, affinity, economic insurance, spread of risks, and political alliance that are necessary for a socially and economically secure existence as a pastoralist. This implies a long-term scheming which has a very different time perspective from making herding decisions. The latter are made on a daily *ad hoc* basis, taking into consideration the latest information about the nutrition, health, and productive status of the animals as well as the quality and quantity of fodder and water available in accessible pastures, the presence of predators and human enemies, and the immediate availability of labor.

Pastoral economic power has several facets: control over the animals as food producers and as objects of

exchange, and control over the growth potential of one's livestock herd. This makes possible multifaceted property concepts. Animal wealth can be redivided, redistributed, and reorganized in a number of ways, creating diversity in rules and practices of inheritance and other forms of transgenerational property devolution. Senior men often strive to control the labor of younger men by retaining ultimate control of the latter's future herds and by summoning ritual prerogatives such as blessing power among the East African pastoralists or closeness to ancestral spirit owners of the land among the Mongols.

Property rights to grazing and water demonstrate a wide span of variation, depending on the territorial regularities of movement. East African pastoralists exploit land that can be seen as either government property or divine property, as open access or as a commons held by the ethnic group. Transhumant Middle Eastern pastoralists may have very definitive rights to seasonal routes. The Beja clans of Sudan claim tracts of land primarily for defining the political privilege of host-ship. Such differences in land right patterns reflect the degree of predictability in location and extent of seasonal rainfall or river flooding as well as the farming cycles of agricultural neighbors and concomitant fluctuations of markets. Seasonal regularity and predictability structure the possibilities for control of land, grazing, and political power. Pastoralism does not represent any single way of organizing relations of production (cf. Bonte 1981, p. 30). There is a great divergence for example between the quasi-feudal Khirgiz of Afghanistan on the one hand and the loosely organized Turkana of Kenya on the other. The Khirgiz khan distributes stock-loans to thousands of clients while among the Turkana each household strives for its own independence through cattle which are either owned by themselves or borrowed on a more egalitarian basis between stock allies. The issue of the forms of inequality that could be found within pastoral societies and the mechanisms often maintaining a relative degree of egalitarianism has interested many of the researchers referred to. If livestock rearing is considered *per se*, there are many practical problems that are common to pastoralists in different parts of the world, yet the ecological niches they occupy also allow for substantial variation. However, the insufficiency of ecological determinism for understanding pastoral cultures and social conditions has become apparent particularly in relation to such questions of equality and inequality.

Over time, anthropologists looking at pastoralism have come to place more emphasis on the various ways in which pastoral communities were integrated in larger regional economic and political systems, and on the interlinkages between pastoral production and other economic activities. An observer who considers the activities of a pastoral group in a particular place and time is easily misled to disregard the variation that a long-term perspective reveals. Later research has

emphasized that pastoralism is not necessarily a static, 'pure' adaptation, but rather a stage in a flexible series of adaptations to irregular climatic circumstances. It has only existed as a pure subsistence form under particular historical conditions. What is more typical for pastoralists than a monodimensional dependence on one type of resource is flexibility in combining different resources. To survive as pastoralists in the long term, livestock-rearers have had to integrate supplementary pursuits such as venture farming, small scale-irrigation, famine-year hunting, or trade. With a longer-term perspective on some of the ethnic groups of East Africa, a gradual switch between dry-land agriculture, pastoralism, and hunting seems to be no rare occurrence. The apparent degree of specialization in and dependence on pastoralism is thus often spurious. Pastoralists whose work is dominated by care for animals still normally depend on getting parts of their diet from farm produce. Small-scale supplementary cultivation can be added to pastoralism in innumerable ways or these products can be secured from others through barter or cash trade, predation, tribute taking, or labor services. The large variation in possible solutions to this problem in itself precludes ecological determinism as a fruitful point of departure for understanding social organization. Processes of stratification and equalization within the pastoral communities can only be understood when these modes of articulation with surrounding structures are taken into account. Today, an essential issue is to what extent a pastoral group in their combination of assets have access to cash or other resources that can act as a security independent of fluctuations in herd health or climate.

In the 1970s and 1980s, the famines among African pastoralists hit by drought coincided with the international breakthrough of environmentalist discourse. Desertification became a catchy symbol of how humanity was overexploiting its brittle resource base. Hardin (1968) presented his theory of 'the tragedy of the commons,' using as an illuminating but fictive illustration a herd-owner's use of pastures. Maximizing his individual gain, the herd-owner was thought not to pay attention to the collective good. In the end all herd-owners would accordingly lose from their own faulty strategy. Hardin's theory was picked up by policy-makers dealing with pastoral development in East Africa. During the 1980s and 1990s a concern for how policies informed by Hardin's theory affect the situation of pastoralists has turned opposition to Hardin's model into a major inspiration for anthropological research. The debate has centered on the degree to which human-induced desertification is taking place. Are pastoralists to be blamed for overgrazing or to be seen essentially as conservationists? The anthropological critique of Hardin and of evolutionary ecologists is often intermixed with an ecological-technical critique, considering the long term effect of pastoral practices on species diversification

and erosion, emphasizing the resilience of desert vegetation, the relative local impact of global climatic change, and inadequacies in models relating grazing pressure to the subsequent availability of good fodder. More anthropological argumentation emphasizes the necessity to distinguish between the knowledge, values and considerations actually expressed among herdspeople and the effects of their herding practices. A methodological issue, dividing social scientists and evolutionary ecologists, is whether 'self-interest' can be analyzed as a causal variable independent of how the actor interprets his own strategies and choices. Pastoral anthropologists question the relevance of the fictive decision-making situations postulated by model-builders like Hardin to the choices real herd-owners have to make. Pastoral economies are based on constant recombination of the available resources of livestock capital and food, and in order to match productive resources and consumption needs as efficiently as possible. This makes the household a debatable unit of analysis. Anthropologists also emphasize the existence of collective institutions and norms which serve a conservationist function, and argue that external observers often think in terms of quantity and tend to neglect quality issues that pastoralists deem entirely important in relation to the composition of a herd and to fodder, minerals, and water. Neither is the parallel and intertwined rationalities of herding and husbandry appreciated. Caring for continuity involves more than a concern for water and fodder: it requires concern for securing access to territory, to docile, fertile, and reproductive stock and to reliable and knowledgeable labor—all critical resources that cannot be taken for granted but have to be continuously ensured.

While the 1970s gave fresh insights into pastoralism at a generalizable level, the field has been relatively stagnant in the 1980s and 1990s, except for the argumentation in defense against pastoralists blamed for desertification. Plentiful documentation has however been secured of threats to the viability of pastoral subsistence. Official worries over the situation of pastoral nomads are concerned with land degradation, but evidence particularly from East Africa points to problems associated with the extent of accessible pastures rather than their quality. Contemporary pastoralists suffer a successive limitation of critical grazing or water assets through the expansion of towns and of small-holder farming and large-scale mechanized agriculture, through irrigation schemes and game reserves and more brutally, through the expansion of zones of war or insecurity.

Technological changes such as motorized surveillance and transport of herds, easy fencing, mechanical watering, or stationary foddering create forms of management more profitable but less efficient in terms of providing work and food for a large number of people. A concentration of stock ownership takes place parallel to a transition to more intensive small

stock rearing at the expense of cattle or camel herding, something which has been suggested to have negative ecological consequences in the long run. The diversion of farm products once provided by cultivating neighbors into other more competitive markets is another serious problem to modern pastoralists.

See also: African Studies; History; Desertification; Ecological Economics; Ecology, Cultural; Evolutionary Approaches in Archaeology; Land Degradation; Nomads/Nomadism in History

Bibliography

- Barth F 1964 *Nomads of South Persia: The Basseri Tribe of the Khamseh Confederacy*. Allen and Unwin, London
- Bonte P 1981 Marxist theory and anthropological analysis: The study of nomadic pastoralist societies. In: Kahn J S, Llobera J R (eds.) *The Anthropology of Pre-capitalist Societies*. MacMillan, London
- Dyson-Hudson N 1972 *Introduction. Perspectives on Nomadism*. Brill, Leiden, pp. 22–9
- Hardin G 1968 The tragedy of the commons. *Science* **162**
- Humphreys C, Sneath D 1999 *The End of Nomadism? Society, State and the Environment in Inner Asia*. Duke University Press, Durham, NC
- Ingold T 1980 *Hunters, Pastoralists and Ranchers*. Cambridge University Press, Cambridge, UK
- Khazanov A M 1984 *Nomads and the Outside World*. Cambridge University Press, Cambridge, UK
- Lefébure C 1979 Introduction: The specificity of nomadic pastoral societies. In: *Pastoral Production and Society*. Équipe ecologie et anthropologie des sociétés pastorales, Maison des Sciences de l'Homme and Cambridge University Press, Cambridge, UK
- Lewis I M 1961 *A Pastoral Democracy: A Study of Pastoralism and Politics among the Northern Somali of the Horn of Africa*. Oxford University Press, London
- Paine R 1971 Animals as capital: Comparisons among northern nomadic herders and hunters. *American Anthropologist* **44**: 157–72
- Ruttan L M, Borgerhoff Mulder M 1999 Are East African pastoralists truly conservationists? *Current Anthropology* **40**: 621–52
- Spencer P 1998 *The Pastoral Continuum: The Marginalization of Tradition in East Africa*. Clarendon Press, Oxford, UK

G. Dahl

Patient Adherence to Health Care Regimens

1. Overview

Research conducted since the 1970s has demonstrated that, on average, about 40 percent (and in some cases as many as 90 percent) of patients fail to adhere to the

recommendations they have received from their physicians for prevention or treatment of acute or chronic conditions. This is a serious problem known in the medical literature both as 'patient nonadherence' and 'patient noncompliance.' The broadest definition of the terms is that a patient fails to manifest behaviorally the health professional's (usually physician's) intended treatment regimen. It consists of such actions as the patient failing to take antibiotics correctly, forgetting or refusing to take hypertension medication, forgoing important health habits, and persisting in a dangerous and unhealthy lifestyle. As a result of failure to adhere, many patients become more and more seriously ill, they develop infections that are resistant to treatment, physicians alter patients' treatments based on misunderstanding of its initial effects, physicians are misled about the correct diagnosis, and the time and money spent on the medical visit is wasted (for reviews see DiMatteo and DiNicola 1982, Meichenbaum and Turk 1987, Myers and Midence 1998, Shumaker et al. 1998).

2. The Conceptual and Clinical Importance of Adherence

The issue of patient adherence has both conceptual and clinical implications. In the realm of clinical medicine, the decade of the 1990s brought amazing advances in the treatment of both acute and chronic illnesses, vastly improving quality of life for those suffering from them. Even HIV has promising treatments in the form of highly active antiretroviral therapies. Attempts at managing medical costs have enhanced the focus on preventive services. There are concerns, however, about the re-emergence of infectious diseases, as some bacterial infections that were once straightforwardly treated with antibiotics remain resistant even to the most powerful antibiotics. In HIV, viral resistance to the protease inhibitors has emerged just when control of the disease seemed possible. It is now recognized that noncompliance may be one of the major causes of these reductions in the efficacy of medical treatments, the proliferation of drug resistance, and the waste of billions of health care dollars.

In the conceptual realm, the achievement of patient adherence involves the phenomenon of social influence. A health professional essentially tries to influence a patient to change their behavior. Because medical care is delivered through the interpersonal interaction between health professional and patient, it may be best understood as a social psychological phenomenon. Thus, relevant are issues of communication, trust, persuasion, social normative influence, social reinforcement, and social support.

The understanding of compliance/adherence is complicated further by the philosophical issue of patient self-determination, often examined in detail in writings

on medical ethics. It is also affected by issues of information processing, understanding, and retrieval, and the sometime interaction of these issues with that of self-determination. For example, a patient might be noncompliant simply because they never understood the regimen in the first place due to anxiety in the presence of the physician. This patient would follow the regimen had they understood it. Another patient might conveniently 'forget' to follow the treatment in an effort to exert a certain amount of self-determination. Yet another patient might choose 'alternative' medicine therapies, instead of medically recommended ones, because the former may be both easier to understand and provide the patient with a greater sense of control. Finally, by failing to adhere, another patient might be consciously or unconsciously trying to commit suicide, either through clear-headed choice or because of severe depression.

3. Controversy Over the Term Reflects a Philosophical Issue

Historically, the behavior of following a physician's recommendations has been termed 'compliance,' and reviewing the literature from indexed databases, particularly before 1990, has required the use of this term. Some writers (see Haug and Lavin 1983) have argued that the term compliance implies coercion, emphasizing paternalistic control on the part of the physician. Many writers have more recently used the term adherence,' which seems to imply that the patient, in a more self-determining manner, 'sticks to' the regimen if they choose to do so. Still others have argued that the appropriate term is 'cooperation,' which patient and physician may or may not achieve, depending upon their willingness and ability to negotiate effectively with one another. Studies do show that patient adherence is highly dependent upon patient involvement in medical choices and patient commitment to treatment decisions, and so cooperation, as well as participation, negotiation, and partnership are essential. The use of all these terms creates, of course, an indexing nightmare, and so the terms 'adherence' and 'compliance' have remained dominant. Although not always referred to as such in indexing, research on aspects of patient self-determination suggests it to be an essential element in adherence and health behavior change.

Whatever it is to be called, the phenomenon of a patient following through with a health professional's treatment recommendations is not a simple achievement. Given the potential for difficulties at several steps—from the physician conveying information about the regimen, to the patient remembering what was conveyed, to the patient wanting to carry out the behavior, to actually overcoming all the difficulties of doing so—one can easily understand why non-compliance rates are very high.

4. Theories Guiding the Research

Various theoretical models have driven, in some fashion or another, the research on patient adherence (see *Health Behavior: Psychosocial Theories*). They include social cognition approaches that emphasize the patient's thoughts and beliefs as primary influences on adherence behavior. Many theories use a value-expectancy paradigm, in which behavior in response to a threat to health arises from expectations that the action will be self-protective. The Health Belief Model is one such approach, and it emphasizes thoughts regarding risks and benefits of the recommended course of action (Janz and Becker 1984). The Theory of Reasoned Action takes account of beliefs and of social influences, and emphasizes the role of intentions and their tenuous connection to action. The Theory of Planned Behavior adds to this a consideration of perceived behavioral control and perceived barriers to action (Ajzen 1985). There are also attribution theories, in which people are posited to be concerned with explaining the causes of events in terms of their locus of control (internal vs. external), stability, and universality (Wallston et al. 1994).

Some models are based on efficacy beliefs regarding both the ability of the self to carry out the target behavior, and of the action to affect the intended outcome (see *Self-efficacy and Health*). There are likely to be limitations to linear rationality in decision making, and recent modifications of the social cognition approach have tried to take these limitations into account. One of these modifications is the Trans-theoretical, or Stages of Change, Model that proposes that adherence occurs through progressive, nonlinear stages or steps of behavior modification and maintenance (Prochaska and DiClemente 1983). Finally, self-regulatory models view the patient as an active problem solver who strives to achieve a health goal by identifying the health threat, developing and implementing an action plan or coping procedure to deal with the threat, and appraising the outcome of the action (Leventhal and Cameron 1987) (see *Health: Self-regulation*).

No single model is universally valid, or even helpful, in the study of adherence, particularly because the complexity of treatment regimens varies widely from one disease to another. An approach that organizes the research on adherence in terms of a purely biomedical model tends to classify and analyze adherence studies by disease conditions and finds some interesting trends and patterns that might provide clinically relevant insights for the care of specific disease conditions. The difficulty with this approach, however, has been that similarities in psychosocial phenomena across disease conditions are not apparent, and findings that could be applicable to all patients may remain undiscovered. Focusing solely on psychosocial issues, without regard to the uniqueness of disease conditions, may introduce too much variation

into the analysis, making trends difficult to determine. A combined 'biopsychosocial' approach, on the other hand, allows for the analysis of psychosocial factors in the context of the specifics of various diseases and treatment conditions.

5. Limitations in the Literature

Since 1968, there have been over 9,000 papers published about patient adherence (or its alternate name, patient compliance), with a ratio of more than two reviews/opinion pieces to one empirical article (Trostle 1997). Most of the empirical publications have examined only one or two predictors of adherence, either in an observational study or in an intervention to improve adherence/compliance. These studies have demonstrated, not surprisingly, that doing something to help patients comply is better than doing nothing, and that compliance can usually be found to correlate with something about the patient (e.g., motivation), the regimen (e.g., its complexity), or the interaction between patient and health professional (e.g., communication). Studies that have examined several elements of this very complex phenomenon are, on the other hand, rather few and far between. Most chronic illnesses, where adherence is very difficult (e.g., adolescent diabetes), are quite demanding psychologically and behaviorally. Designing and maintaining an effective treatment package requires understanding the entire picture of compliance, from communication of the regimen to its implementation. Unless the phenomenon is fully understood in all its complexity, it is impossible to know when, where, and how it might be best to intervene to help patients to adhere.

6. A Multidimensional Approach to Noncompliance

In examining how difficult and complex adherence to a treatment regimen might be, it is useful to consider a clinical example. Imagine a 48-year old, moderately overweight, borderline hypertensive, male patient who has been told by his physician that he should begin and maintain a program of regular exercise. The goals of this exercise are to reduce his weight, to lower his blood pressure, and improve his level of HDL cholesterol. He understands quite well what he is being asked to do, and he believes that exercise is important and worth the trouble and time it takes. His family and friends agree that the regimen is valuable, and encourage him to follow it. His family members remind and encourage him daily, and have even invested in a treadmill for him to use each evening. In this case, as in general, exercise compliance demands a multidimensional, multifactorial approach because changing the sedentary habits of a lifetime can be very

difficult and often meet with failure. Even among rather seriously at-risk patients, such as who have cardiovascular disease, diabetes, or a previous myocardial infarction, there is a 30–70 percent dropout rate from exercise. In such situations, as with our example patient here, people may have very positive attitudes toward exercise, and even strong networks of social support, but still fail. Maintenance of these very complex behavioral changes over a long period of time requires very careful attention to many factors.

7. Factors That Affect Patient Adherence

As noted above, there tends to be something of a disjuncture between the theoretical models of adherence, most of which are multidimensional, and the empirical findings, which are primarily unidimensional. Nevertheless, the empirical studies tell us some important facts about adherence that are useful both clinically and in terms of how they contribute, albeit in a limited fashion, to the theoretical models. Patient demographics, for example, are thought by many clinicians to be the best predictors of patient adherence. Physicians tend to use easily available information from the clinical encounter to steer their attention to those they believe to be at highest risk for nonadherence. Despite physician beliefs, however, patient age and gender seem to have very little relationship to adherence (except, perhaps, that adolescents tend to have more problems adhering than other age groups). Further, although lower income patients and those with more limited education have somewhat lower levels of adherence than more affluent and educated patients, the effect is quite small, and not nearly as important as that of other variables that can be altered.

Psychological variables, such as attributions, locus of control, and beliefs, while important in the models noted above, have not shown consistent effects on adherence. For example, believing that a disease is severe may in some cases prompt a person to take consistent action to avoid or treat that disease, and in other cases bring about denial and reckless abandonment of the regimen. On the other hand, the effect of another psychological variable—depression—on adherence is potentially noteworthy and should be examined further. Depression may bring about hopelessness and isolation, two phenomena that reduce adherence considerably; but it may be entirely treatable (see *Depression, Hopelessness, Optimism, and Health*).

Adherence has been found consistently to be dependent on the patient's social support system and family/marital environment. Further, a regimen's complexity and adverse effects must not be overlooked by physicians when making a medical recommendation; much research has found that nonadher-

ence is related to long-term regimens with complex and confusing dosages and scheduling. Health professional-patient communication seems to affect adherence as well. Because of limitations in communication during the medical visit, patients may misunderstand what they are to do, and the anxiety of the situation tends to interfere with recall at a later time. The sensitivity and empathy in physician-patient communication often matters as well, because patients tend to be more likely to try to adhere to recommendations given by health professionals that they like and trust. Patients' health beliefs and attitudes, often influenced by their cultural patterns and expectations, may conflict with the regimen, making adherence difficult, and practical barriers, such as side effects and lack of necessary resources, may interfere with the regimen (see *Health and Illness: Mental Representations in Different Cultures*). Finally, studies show that physicians typically do not know whether or not their patients are adherent. They overestimate the degree to which their patients follow their directives, and are typically unable to identify which patients are having adherence problems. Yet, the accurate assessment of adherence is essential because achievement of adherence has been shown to make an important difference in patient outcomes and patient health status.

8. Research Challenges

A major research challenge in the field of adherence involves its measurement. It seems obvious that one good way to measure adherence would be simply to ask the patient. In fact, most adherence research and clinical practices have relied upon the technique of self-report, although such an approach may be biased by self-presentation and by patients' fears of reprimand by the physician. Reports of spouses or other health professionals can be useful, but these reports may vary in accuracy depending upon the opportunity to observe the patient's daily activities. Techniques such as pill counts, patient behavioral diaries, and electronic recording devices have their own drawbacks, including patient manipulation toward the goal of concealing nonadherence. Tests, such as urine or blood assay, may be useful for medication, but tend to reflect only its recent consumption. Physician reports and chart entries are unreliable and are often based on unclear criteria. Because there is no 'gold standard,' convergence of the findings of research is more difficult to achieve. Necessary to the field is the development and implementation of methods for building therapeutic trust so that patients can be forthcoming and frank about their difficulties in following treatment suggestions, and both measurement and adherence itself can be improved (Hays and DiMatteo 1987).

A second issue of concern is that most reviews of adherence have been qualitative. Many writers, in-

cluding this author, have described extensively the factors that influence patient adherence, ranging from the importance of physicians' communication skills to how simple or complex the regimen should be. Although these reviews have attempted to be exhaustive, their conclusions have been biased by several factors including limited sampling of the population of empirical studies, and entrenched, established social psychological (or other) models of behavior guiding their organization. Further, when studies are listed and the effects of different variables on adherence are described, there are invariably conflicting results that cannot be resolved qualitatively. When there is a huge amount of research available, such as on adherence, it is possible to review only a sampling of it, resulting in further bias unless that sampling is random. One good quantitative assessment of a complete set of empirical studies, however, is worth many expert opinions and reviews. Fully understanding all of the complex elements of adherence involves painstakingly organizing, reviewing, and quantitatively summarizing the entire abundant literature on each variable related to adherence employing the research technique of meta-analysis. Nothing short of this will provide the clear answers that are necessary. Such an approach would allow for a better overall understanding of the complex relationships between adherence and characteristics of patients, their lives, their diseases, their regimens, the therapeutic relationship, and the context of medical care delivery.

Finally, conceptual and empirical work is necessary to remedy the common confusion of adherence behaviors with adherence outcomes. A patient might carry out flawlessly every behavioral requirement of care, but the outcome may remain disappointing. Research must pay scrupulous attention to the accurate assessment of behavior, and recognize that physiological measures of outcome should never serve as proxies for adherence. The independent effect of adherence on achieving health outcomes, health status, functional status, psychological and social functioning, and all aspects of health related quality of life needs to be examined (see *Quality of Life: Assessment in Health Settings*).

9. Conclusion

As noted above, despite the complexity and multidimensionality of theoretical models of adherence, nearly all of the research has examined individual factors that correlate with adherence. These empirical investigations have provided some useful answers about adherence, but the theory that could be helpful in guiding our research questions remains disconnected from it. Fully understanding adherence has remained an elusive goal because reviews of the literature have been equivocal. Clear answers are

necessary, however, as some of these questions have strong clinical and policy implications.

This review of the literature has attempted to view it as a whole, to the extent that a qualitative approach can do so, and to point out its strengths, limitations, and promising trends. At this point in time, there are many reviews, and not enough well designed, well executed, multidimensional, longitudinal empirical studies of adherence. Now that we have a compilation of studies in the univariate realm, it is necessary to work toward multifactorial explanatory approaches to adherence. These models need to examine simultaneously the effect of social, psychological, and biological variables on adherence, and to examine their unique interactional effects. Only then will a full and clear picture of this complicated phenomenon emerge.

See also: Attitudes and Behavior; Explanatory Style and Health; Health Behaviors; Health Behaviors, Assessment of

Bibliography

- Ajzen I 1985 From intentions to actions: a theory of planned behavior. In: Kuhl J, Beckmann J (eds.) *Action-Control: From Cognition to Behavior*. Springer-Verlag, Berlin
- DiMatteo M R, DiNicola D D 1982 *Achieving Patient Compliance: The Psychology of the Medical Practitioner's Role*. Pergamon Press, New York
- Haug M R, Lavin B 1983 *Consumerism in Medicine: Challenging Physician Authority*. Sage, Beverly Hills, CA
- Hays R D, DiMatteo M R 1987 Key issues and suggestions for patient compliance assessment: sources of information, focus of measures, and nature of response options. *Journal of Compliance in Health Care* 2: 37–53
- Janz N K, Becker M H 1984 The Health Belief Model: a decade later. *Health Education Quarterly* 11: 1–47
- Leventhal H, Cameron L 1987 Behavioral theories and the problem of compliance. *Patient Education and Counseling* 10: 117–38
- Meichenbaum D, Turk D C 1987 *Facilitating Treatment Adherence: A Practitioner's Guidebook*. Plenum Press, New York
- Myers L B, Midence K (eds.) 1998 *Adherence to Treatment in Medical Conditions*. Harwood Academic, Amsterdam
- Prochaska J O, DiClemente C C 1983 Stages and processes of self-change of smoking: toward an integrative model of change. *Journal of Consulting and Clinical Psychology* 51: 390–95
- Shumaker S A, Schron E B, Ockene J K, McBee W L (eds.) 1998 *The Handbook of Health Behavior Change*, 2nd edn. Springer, New York
- Trostle J A 1997 Patient compliance as an ideology. In: Gochman D S (ed.) *Handbook of Health Behavior Research*. Plenum, New York, pp. 109–22
- Wallston K A, Stein M J, Smith C A 1994 Form C of the MHLC Scales: a condition-specific measure of locus of control. *Journal of Personality Assessment* 63: 534–53

M. R. DiMatteo

Copyright © 2001 Elsevier Science Ltd. 11115
All rights reserved.

Patriotism

‘Patriotism’ is about political allegiance (and, of course, loyalty), commitment, and dedication. In briefest compass, it means love of one’s country or nation and is one of the oldest political virtues. It is rather more emotional than rational in its appeal and demands recognition of what is presumptively a pre-existing duty to that political order or state. One of the best exemplars of patriotism is Stephen Decatur’s well-known toast in 1816:

‘Our Country! In her intercourse with foreign nations may she always be in the right, but our country, right or wrong.’

To which John Quincy Adams replied:

‘My toast would be, may our country always be successful, but whether successful or otherwise, always right.’

Patriotism depends upon an often unarticulated principle that is the political counterpart to the ‘blood is thicker than water’ adage, which reminds people that they should prefer their families before all others. Patriotism is a natural consequence of political membership or citizenship, and it is not clear where—other than membership itself—the duties of patriotism originate or what justifies them other than that membership, for patriotism is not voluntarily assumed as general political obligations are presumed to be. To deny or renounce patriotism or to act contrary to what its proponents deem appropriate is to be disloyal.

There is a Burkean quality to patriotism, both because it projects an almost organic, trans-historical unity among the member of a state, nation, or people such that each individual is inseparable from the past and because it looks to that past and to the achievements of one’s political ancestors rather than to one’s own accomplishments as sources of pride and holds up those achievements as standards by which the successes of the present are to be measured. In this respect, patriotism can be nurturing as well as oppressive, for it defines and constitutes the political member and, in the process, restricts that member’s range of permissible options.

Patriotism is rarely invoked but is in times of stress or trouble. The call for patriots or acts of patriotism is issued when a sacrifice for the presumed good of the people or their state is needed, usually because that good is perceived to be in jeopardy or under attack. In the period since World War II, those attacks have most often been alleged to come from inside a system, from people who in an earlier day would have been labeled ‘traitors’ but are now more conventionally called ‘disloyal’ and even ‘outsiders.’ During World War II, in Europe, according to the *OED*, a patriot was a ‘loyal inhabitant of a country overrun by the enemy, especially a member of a resistance movement.’

But for nearly 150 years prior to that, also according to the *OED*, because the mantle of the patriot had been assumed by persons who were deemed not entitled to it, the term itself was somewhat discredited. Dr Johnson, amplifying his dictum that ‘patriotism is the last refuge of a scoundrel,’ noted that the word was used ‘ironically for a factious disturber of the government.’

Etymologically, ‘patriotism’ springs from the Greek, via Latin, for father and father land, (πατριώτης, from πάτριος of one’s fathers, πατρίς, one’s fatherland; late Latin *patriota*, fellow-countryman [*OED*]), suggesting political membership based on kinship as well as an implicit reliance on family-like bonds to hold a state together. And in the Latinate tongues of Europe, various cognates of *patria* mean ‘fatherland’ or native country. Native too, as well as nation to which it is conceptually and etymologically related, suggests kinship, but these resonances have long since been buried by linguistic evolution. Their value today is primarily as reminders of the emergence of the political order from tribal, familial associations. ‘Fatherland’ is more obviously familial, and its roots are Germanic rather than Latinate. (Interestingly, German has adopted the Latinate forms *der Patriot* and *der Patriotismus* even though it has its own words, *der Vaterlandsfreund*, literally, *friend of the fatherland*, and *die Vaterlandliebe*, *love of the fatherland*.) But patriotism has never had this familial resonance in English, which explains why the term has an air of contentiousness and alarm about it when invoked in English and why its English-speaking history has been checkered. In that largely individualist and voluntarist world, political duty, at least since the seventeenth century, has been viewed as a consequence of intentional commitment and a subject for reason and judgment. In those terms, patriotism has often seemed like an alien concept.

Patriotism is more at home in the conceptual world of republicanism. The republican tradition has always looked upon the state as a closely-knit and relatively homogenous association. Patriotism calls upon the members of this association, when appropriate, to put aside whatever divides them and to rally in support of what they share, a practice that is integral to the communitarian predilections of republican society but stands in need of justification from the perspective of individualism. The call for patriotism—for people to be patriotic, for those who are already patriotic to come forth in the spirit of the *patria*—is an appeal to the emotions, no less than love in its ordinary sense is an invocation of affect rather than reason. And like the affections that bind friends and family members, patriotism works by reminding those at whom it is directed of their ties and of their non-voluntary relationships to other people. Thomas Paine’s ‘sunshine patriot’ is a play on the better known ‘fair weather friend,’ both demeaning of the self-styled supporter who is unwilling to be inconvenienced.

Patriotism is rooted in emotions rather than reason and has to do with feelings of commitment and loyalty to one's nation or state and pride in its history and accomplishments. While it is not necessarily irrational or even unthinking, patriotism does pose as the supreme call on one's commitments, the one that trumps or overrides others with which it may compete.

With the revival of classical republican thinking in English-language political discourse, toward the latter part of the twentieth century, patriotism made an almost grudging return. Republicanism values 'community' and 'civic virtue' with which it seeks to supplant individualism, self-willed obligation, and institutional legitimacy. Further sources of the renewed appeal to patriotism are the collapse of Eastern European communism and the ensuing struggles for national identity and 'liberation.' These linked patriotism to ethnic and religious nationalism (which are akin to those flourishing in Northern Ireland, the Middle East, and on the Indian Subcontinent). Even earlier, the worker and student protest movements of the 1960s and 1970s—especially in France, Germany, the UK, and the USA—and the American civil rights movement were all surrounded by claims of 'disloyalty,' and illegitimacy, which resulted in quests for the proper bearers of the mantle of patriotism. Finally, in the USA, the period of the Cold War was marked by zealous attacks on disloyal 'communist sympathizers,' often in the name of patriotism.

In the West, contemporary proponents of patriotism attempt to harmonize it with the Enlightenment value of 'cosmopolitanism' and struggle to separate it from the virulent nationalism with which it is too easily associated and which it often resembles. These efforts seem destined to fail, for patriotism is particularistic, not universal. It tells people that what they have in common is what deeply and importantly unites them and makes them a nation or a 'people.' This unity overrides their differences. There is, in consequence, a tendency to homogenize those differences into a political blandness that could render society uninteresting and potentially stagnant. Far worse than that, however, patriotism in this homogenizing form prepares the way for the insistence that a greater, underlying good assigns places in the social order. Because of the fundamental sameness of all members, there are no remediations for deprivations. Those who are not sufficiently 'the same' are outsiders who can and should be excluded.

The modern state at the beginning of the twenty-first century is increasingly complex and heterogeneous in ways that traditional patriotism cannot comprehend. Moreover, individual states cannot reject that heterogeneity and continue to exist in a world of international political and economic exchange. States today inevitably function in a world that is at odds with their claims of internal uniformity. That internal coherence—where it is more of an ideal than a fact—often leads to oppression. There are few states

that do not have internal 'minority' peoples who dissent or are excluded from the presumed consensus that undergirds patriotism. While this is not to say that internal political loyalty and cultural and social diversity are incompatible, the responses of the advocates of patriotism to the circumstances that call forth their pleas are antagonistic to cultural variety. One of the hallmarks of modern politics is tolerance and forbearance by states and their members. But tolerance, by its nature, undermines both the spirit and the practice of patriotism. It is far easier—and in many respects more desirable—to give loyalty to those who are regarded as like one's self, which is among the principal reasons that the habits of obedience that patriotism fosters and on which it depends are usually rooted in the family. To bestow that same deference on an 'alien' authority or on one that appears to uphold a different set of values from one's own is often difficult and can require acts of will and judgment that are antithetical to the non-rational, emotional bases of patriotism.

At the same time, however, so long as there are territorial nations, there will be reasons for inculcating loyalty to them; nations necessarily require their members to make sacrifices from time to time, and it is certainly preferable that these sacrifices—these fulfillments of civic duties and responsibilities—be made willingly if not voluntarily and in the belief that they are justified. Individual members must have grounds for accepting the propriety of actions undertaken by and/or in the names of their states; they must feel some dedication and loyalty to their states. The inculcation of those feelings through the process generally known as 'civic education' must be done in ways that preserve the forbearance that toleration requires and do not give rise to the exclusionary and destructively ugly nationalism that often has patriotism as its partner.

See also: Identity Movements; Identity: Social; Nationalism: Contemporary Issues; Nationalism: General; Nationalism, Historical Aspects of: The West; Nationalism, Sociology of; Republicanism: Philosophical Aspects; State, Sociology of the; Xenophobia; Civil Society; Civic Culture

Bibliography

- Canovan M 2000 Patriotism is not enough. *British Journal of Political Science* 30: 413–32
- Dietz M G 1989 Patriotism. In: Ball T, Farr J, Hansen R L (eds.) *Political Innovation and Conceptual Change*. Cambridge University Press, Cambridge, UK, pp. 177–93
- Nussbaum M C 1996 Patriotism and cosmopolitanism. In: Cohen J (ed.) *For Love of Country: Debating the Limits of Patriotism*. Beacon Books, Boston, MA, pp. 1–16

- Rorty R 1999 The unpatriotic academic. In: Rorty R (ed.) *Philosophy and Social Hope*. Reprinted from the *New York Times*, 13 February 1994. Penguin Books, New York, pp. 252–54
- Schaar J H 1973 The case for patriotism. *American Review* 17 (May): 59–99
- Wahlke J C (ed.) 1952 *Loyalty in a Democratic State*. D C Heath, Boston, MA
- Violi M 1995 *For Love of Country: An Essay on Patriotism and Nationalism*. Oxford University Press, Oxford, UK

G. Schochet

Patron–Client Relationships, Anthropology of

Patron–client relations (PCR) involve asymmetric but mutually beneficial, open-ended transactions based on the differential control by individuals or groups over the access and flow of resources in stratified societies. Patrons provide selective access to resources they control, or place themselves or the clients in positions from which they can divert resources in their favor. In exchange, the clients are supposed to provide material and human resources, including their gratitude and loyalty, while boosting the patron's prestige and reputation. The study of PCR has revealed the persisting presence of interpersonal hierarchy in contemporary societies, contributing to the reevaluation of paradigms in social science. Debates continue about their institutional viability and significance in late modernity.

1. Terminology

The terms used originated in ancient Republican Rome, where relationships of *patrocinium* and *clientelae* proliferated during the Republic and into the Empire (later on, they were subsumed under what historians of late medieval times called *bastard feudalism*). These terms found their way into the vernaculars of Mediterranean and Latin American societies. Other terms are used for PCR elsewhere.

2. Background

Whereas in antiquity, PCR formed part of the normative framework of society and could hardly be thought as conflicting with legal institutions and the social ethos, in modern societies PCR are built around such a conflict. While in principle and by law the clients have access to power centers and the ability to convert resources autonomously, this potential is neutralized by the patrons' control over scarce re-

sources, be they land, water sources, employment, loans, schooling opportunities, medical services, public security, or infrastructure. PCR may be rooted in the search for protection, in bureaucrats' providing preferential access to petitioners in return for future favors, in poor people sharing the social visibility of men or rank, while the latter use these links to consolidate their status. Union leaders, political activists, and government officials may use PCR to build a network of supporters, instrumental for gaining higher-level positions, office, and control of resources.

3. The Study of PCR

Expanding in the 1960s, it led (by the late 1970s and 1980s) to general and comparative analyses. In the 1990s in-depth studies renewed interest in PCR in polities under transition to market economy and democracy, re-evaluating some of the premises of earlier approaches, especially regarding the connection between PCR and underdevelopment. The growth of interest has been connected with the spread of research from anthropology to the other social sciences. From early dyadic emphases, research has increasingly revealed a complex range of PCR analyzed in terms of networks, factions and coalitions.

The study of PCR contributed to analytical shifts, as it deals with phenomena for which neither the group corporate model nor formal institutional models provide a satisfactory guide. In parallel, research revealed the complexities of studying PCR, owing to their informal character and to their cross-institutional insertion that requires interdisciplinary knowledge, bridging anthropology and sociology, political science, and economics.

4. A Logic of Social Exchange

Research has identified a specific logic of exchange and reciprocity beneath the wide gamut of PCR. Both the control of material and human resources and the character of trust play a crucial role in it. This logic has been defined in the literature as clientelism. A related term is patronage, used interchangeably at times and alluding to the support granted by the patron to his followers and *protégés*.

This logic implies:

Control of markets. Individuals and institutional actors are able to gather followings by dispensing selective access to valuable benefits and resources. They do so whether acting as patrons (in a strict sense), or through their influence as brokers with those who control the goods and services, or through a combination of both roles, in what has been called patron-brokerage.

Inequality and asymmetry structured though an idiom of reciprocity.

Particularism. A relationship shaped according to the particular traits of the partners rather than on the basis of entitlements or formal roles.

Favoritism. Many people are excluded from PCR or are related indirectly or intermittently. Where clientelism is highly valued, it generates ‘inflationary’ expectations, many of which are disappointed.

The simultaneous exchange of instrumental (e.g., economic and political) and ‘sociational’ or expressive resources and services (e.g., promises of loyalty). A package deal allowing for more than a restricted balanced or market exchange is built-in, identified as connected to generalized exchange. This determines expectations of entering more embracing attachments, often couched in terms of commitment, even though some PCR are very ambivalent in this respect.

PCR undermine group (i.e., class and strata) solidarity among both clients and patrons, especially among clients, and to exacerbate resentment and social antagonisms.

PCR are neither fully contractual nor legal. Most PCR are vulnerable to systemic pressures and are characterized by instability, perpetual contest and resource manipulation.

5. Exchange Strategies

The agent performing as client is not only expected to provide his patron with specific resources but must also accept the patron’s control over access to markets and public goods, and over his ability to convert resources fully. In parallel, the patron’s position is not as solid as it may seem, nor guaranteed. Never fully legitimized, it is vulnerable to attack by social forces committed to universalistic principles, by the competition of other patrons and brokers, potential and actual, and by social forces excluded from PCR. Owing to these constant threats, patrons are compelled to rely on their followers to solidify their position. The patron must also relinquish some short-term gains to project public claims and to bolster images of power and reputation. Sometimes this earns him/her the right to determine the basic rules of the social relationships. In return the client is protected from social or material insecurity, and is provided with goods, services and social advancement. Certain types of trust (focalized rather than generalized) are associated with such dynamics.

Complementary exchange strategies are built, which signal what Vincent Lemieux defined as a ‘double transformation.’ That is, an abdication of autonomy on the client’s part and a relaxation of hierarchical controls on the patron’s part, through which the client’s lack of power becomes dominated power and the latter’s lack of domination becomes dominating authority. These exchange strategies are not only affected by immediate, mostly technical considerations of power and instrumentality, but often encompass

mutual, relatively long-term, compromises based on commitments as the prerequisite for ongoing social relationships.

6. The Context of PCR

Unlike societies in which hereditary ascriptive principles predominate, full-fledged clientelism flourishes where markets are no longer controlled through primordial units and allow for an open flow of resources and opportunities for mobility. (This is what distinguishes clientelism from *feudalism*, as pointed out already by Max Weber in *Economy and Society* (1968, Vol. III, Chaps. 12–13) and later on by historians of Europe). This trend, however, goes hand in hand with a strong tendency toward unequal access to markets and sociopolitical spheres. As they affect distribution and redistribution, PCR remain subject to the dynamics of political economy. Marketing economies, accelerated urbanization and the expansion of the regulatory, extractive, or even sporadic mobilizing activities of central administrations affect PCR. Research has shown that the impact of world economic trends, fluctuations in the price of commodities, the complexity of international trade, banking, and aid—all these affect the pool of patronage resources available to states and other agencies, and influence patterns of control, distribution, and redistribution. Moreover, as these arrangements are not fully legitimized, they remain vulnerable to the challenge of countervailing social forces.

7. The Comparative Analysis of PCR

Until the early 1980s, it was dominated by the contrast between traditional dyadic patronage and modern party-directed clientelism, a distinction that derived from the dominant paradigm of modernization. With a clear-cut developmental emphasis, this approach focused on differences in organizational complexity, deriving other aspects from it in a functionalist manner.

Since the 1980s, comparative analysis has paid attention also to other aspects. First, the way agents configure their role. Second, the styles shaping the relationship, following what Robert Paine called a ‘generative transactional model of patronage.’ Third, symbolic aspects such as images, discourse and trust, so important for the development of PCR. Fourth, the distinction between PCR as institutional *addenda* and clientelism as a major institutional strategy in certain societies and periods. Finally, PCR’s institutional viability in processes of democratization and liberalization.

8. Debates on Institutional Viability

Researchers differ in assessing the institutional viability and significance of PCR in late modernity.

From one perspective, PCR neutralize the system of representation and entitlements, by placing ‘friends’ in strategic positions of power and control. Clienteles are depicted as inimical to the institutionalization of public accountability and contrary to a politics, open to generalization and participation, and to a discourse of rights. Other authors emphasize the pragmatics of social action, stressing that PCR are an important mechanism for obtaining transactional benefits, in resource allocation, and in providing local–regional–national mechanisms of articulation. While PCR run counter to universalistic standards, it is claimed they are sensitive to local sentiment, may solve existential problems, provide access for migrant populations, and serve political entrepreneurs. From this perspective, as long as clienteles maintain some balance of reciprocity, the participation in political and economic markets by the ‘capi-clientele’ (to borrow Mosca’s expression)—be he or she a broker, a patron or a patron-broker—constitutes a means for individuals in their entourage to influence public decisions. In this sense, some literature alludes to PCR as reconciling public and private authority and formal and informal rules of the game.

PCR are criticized, and opposed by social forces and coalitions wishing to curtail its presence alongside bureaucratic universalism and market rationality. In parallel, sectors benefiting from patronage may see it as a pragmatic avenue of controlled freedom, useful for advancing in competitive social, economic, and political domains. This duality reflects a major tension of modern democracies, which are built on citizenship and political equality but leave the economic domain open to inequalities.

This explains the paradoxical development of clientelistic networks under macro-economic adjustment and restructuring. Liberalization, reduction of state intervention in favor of market mechanisms, privatization of services, and curtailment of union power further fragments society and heightens the need for support networks. When available, clientelism remained important throughout institutional revamping in Poland, Russia, Hungary, Brazil, Argentina, and Turkey.

In centralized polities, PCR have been depicted as a ‘lubricant’ to bypass the inefficient and rigid official system, constituting highly trusted informal problem-solving networks that provide material gains and support in situations of low institutional trust. In representative democracies, clientelism can be effective in power competitions, encouraging and rewarding party activists, once power is achieved, for effectively implementing policies. Patronage may become a restricted but legitimate procedure related to high office incumbents’ right to appoint followers to positions of responsibility. In other respects, clientelism remains controversial and open to allegations of corruption, due to its informal particularistic use of public resources. Often, it is ignored as long as possible or is

disguised as friendship, which is more acceptable in terms of the proclaimed ethos of modern equality.

Studies suggest that in post-industrial societies, patronage develops among professional and upper strata, rather than being restricted to the lower classes. Also, PCR are not confined to politics in the narrow sense, but proliferate as well in the arts, the academia, the church, the media, and even in business—whenever we deal with the power of appointment and the granting of access to benefits, goods, services, influence, and honors.

Changes in the perception of PCR are of systemic consequence if they result in the institutionalization of mechanisms through which citizens can press for entitlements without personal mediation. Such changes include: civil service reforms, non-partisan public systems, recognized charters of rights, controls over party fund raising, and non-partisan comptrollers as a prestigious and trustworthy branch of government. The functioning of these institutional mechanisms hinge on public support for a configuration of the public sphere structured around public accountability and formal responsiveness to turn more and more discrete issues into publicly negotiable and politically consequential.

See also: Exchange in Anthropology; Exchange: Social; Feudalism; Social Stratification

Bibliography

- Blok A 1974 *The Mafia of a Sicilian Village*. Blackwell, Oxford, UK
- Caciagli M 1996 *Clientelismo, corrupción y criminalidad organizada*. Centro de Estudios Constitucionales, Madrid
- Clapham C (ed.) 1982 *Private Patronage and Public Power*. Frances Pinter, London
- Dinello N 1999 *Clans for Markets or Clans for Plan: Social Networks in Hungary and Russia*. Woodrow Wilson International Center, Washington, DC
- Eisenstadt S N, Roniger L 1984 *Patrons, Clients, and Friends*. Cambridge University Press, Cambridge, UK
- Gellner E, Waterbury J (eds.) 1977 *Patrons and Clients in Mediterranean Societies*. Duckworth, London
- Graziano L 1984 *Clientelismo e sistema politico. Il caso dell'Italia*. Angeli, Milan
- Roniger L, Güne-Ayata A (eds.) 1994 *Democracy, Clientelism and Civil Society*. Lynne Rienner, Boulder, CO
- Schmidt S W, Guasti L, Landé C H, Scott J C (eds.) 1977 *Friends, Followers and Factions*. University of California Press, Berkeley, CA
- Willerton J P 1992 *Patronage and Politics in the USSR*. Cambridge University Press, Cambridge, UK
- Weber M 1968 *Economy and Society*. Bedminster Press, New York
- Wolf E 1966 Kinship, friendship, and patron–client relationships in complex societies. In: Banton M (ed.) *The Social Anthropology of Complex Societies*. Tavistock, London, pp. 1–22

L. Roniger

Pattern Matching: Methodology

To the degree that products of the social and behavioral sciences merit the term ‘knowledge,’ it is because they are grounded in representations of the social world achieved by matching observations of that world with abstract concepts of it. Pattern matching is essential for attaining knowledge of everyday cultural objects such as food, clothing, and other persons. It is also indispensable for achieving knowledge of inferred entities in science—the size and brightness of distant stars, the existence and magnitude of latent psychological variables such as alienation or intelligence, and the structure of complex social processes such as social mobility, policy implementation, and economic growth. These and other inferred entities are known indirectly and vicariously, through a process of pattern matching (Campbell 1966).

1. The Concept of Pattern Matching

Pattern matching is a ubiquitous feature of knowledge processes in everyday life and in science. The use of pattern matching in everyday knowing, however, is different from pattern matching in mathematics, philosophy, and the social and behavioral sciences.

1.1 Rules of Correspondence

Rules of correspondence are prescriptions that enable the mapping of one set of objects on another. Objects in a domain, *D*, are related to objects in a range, *R*, according to a rule of correspondence such as: Given a set of *n* countries in domain *D*, if the object in *D* is a one-party system, assign the number ‘0’ from range *R*. If it is a multiparty system, assign a ‘1.’ Rules of correspondence perform the same matching function as truth tables in symbolic logic, pattern recognition algorithms in computer science, and the use of *modus operandi* (M.O.) methods in criminology. When a single rule of correspondence is taken to define fully and unequivocally the properties of an event or object—for example, when responses to scale items on an intelligence test are taken to define ‘intelligence’—pattern matching becomes a form of definitional operationism (Campbell 1969, 1988, pp. 31–2).

1.2 The Correspondence Theory of Truth

Until the late 1950s, philosophy of science was dominated by the correspondence theory of truth. The correspondence theory, the core epistemological doctrine logical positivism (see Ayer 1936), asserts that propositions are true if and only if they correspond

with facts. The correspondence theory also requires that factually true propositions are logically validated against formal rules of logic such as *modus ponens* ($p \supset q, p, \therefore q$) and *modus tollens* ($p \supset q, \sim q, \therefore \sim p$). To be verified, however, propositions must match facts (reality, nature). The correspondence version of pattern matching assumes a strict separation between two kinds of propositions—analytic and synthetic, logical and empirical, theoretical and observational—a separation that was abandoned after Quine (1951) and others showed that the two kinds of propositions are interdependent. Because observations are theory dependent, there is no theory-neutral observational language. Theories do not and cannot simply correspond to the ‘facts.’

1.3 Coherence Theories of Truth

The correspondence theory has been replaced by a more complex form of pattern matching, the coherence theory of truth, which has a number of versions (see Alcock 1996). In one version, often called the consensus theory of truth, beliefs are matched against other beliefs, with no requirement that they are tested empirically. Another version, realist coherence (Putnam 1981), requires that two or more empirically tested beliefs are matched. William Whewell’s *consilience theory of induction* is closely related to this (qualified) realist version of the coherence theory. A third version of coherence theory is methodological pragmatism (Rescher 1980). Here, beliefs must satisfy cognitive requirements including completeness, consonance, consistency, and functional efficacy, all designed to achieve optimally plausible knowledge claims. Other versions of coherence theory require the additional condition that the social circumstances under which empirically tested beliefs arise be taken into account. These other versions include ‘social epistemology’ (Fuller 1991) and the ‘sociology of scientific validity’ (Campbell 1994).

1.4 Statistical Estimation and Curve Fitting

Statistical principles, rules, or criteria are applied to achieve an optimal match between a curve and a set of data points or observations. An example is the *least-squares criterion* (the squared distance between observed and predicted values is a minimum or least value), where the match between the pattern supplied by a (linear or nonlinear) curve and a pattern of observations is approximate and probable, not certain as in pattern matching by rules of correspondence. The degree to which a curve and a set of observations match is summarized by coefficients of different kinds. Some of these represent the goodness-of-fit between

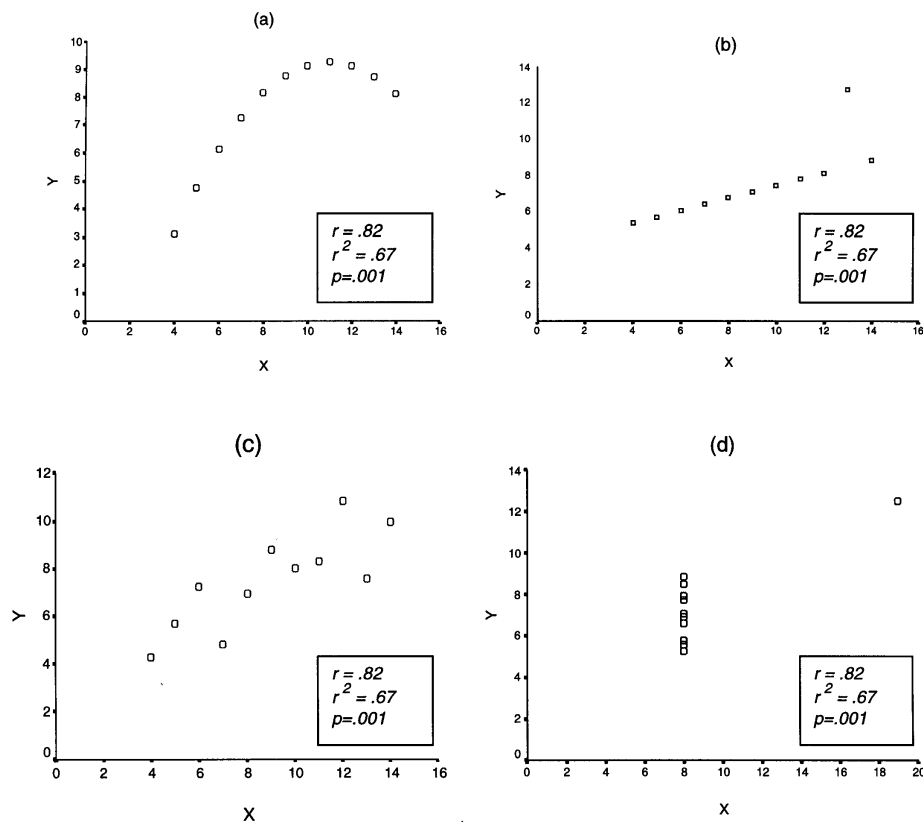


Figure 1
Statistical pattern matching: identical coefficients match different patterns (sources: Tufte 1983, pp. 13–14, Anscombe 1973)

curve and observations, while others represent the magnitude of error in pattern matching. Although coefficients are assessed according to a probability distribution, and given a p -value, the same observations can fit different curves; and the same curve can fit different observations. In such cases, pattern matching is as much a matter of plausible belief as statistical probability.

1.5 Context Mapping

Closely related to curve fitting is context mapping. Knowledge of contexts is essential for understanding elements within them. Context mapping, which is epitomized by the figure-ground relation in Gestalt psychology, also applies to the analysis and interpretation of statistical data. For example, observations in a scatterplot cannot be distinguished—in fact, they all look the same—when they are compared one-by-one, rather than compared as elements of the pattern

of which they are elements. Context mapping is important in statistical analysis. Coefficients that summarize relations among variables are usually misleading or uninterpretable outside the context provided by scatter plots and other *visual methods in statistics* (see Tufte 1983, 1997). Figure 1 illustrates how identical sets of correlation (r), goodness-of-fit (r^2) and probability (p) statistics match different patterns. Statistics are likely to be misinterpreted outside the context provided by scatterplots, each of which has its own visual signature.

2. Theory of Pattern Matching

The first systematic effort to develop a theory and methodology of pattern matching was that of psychologist and philosopher of social science, Egon Brunswik, a professor at the University of California, Berkeley between 1935 and his death in 1955. In this period, he and his close collaborator, Edward Tolman, developed different aspects of the theory of *proba-*

bilistic functionalism. To communicate this complex theory, Brunswik developed, initially for illustrative and pedagogical purposes, a *lens model* that was based on the optics metaphor of a double convex lens. The indispensable procedural complement of probabilistic functionalism and the lens model was his methodology of *representative design*.

2.1 Probabilistic Functionalism

Probabilistic functionalism is a substantive theory and methodology (philosophy of method) that focuses on relations of adaptation and accommodation between the organism and its environment. Probabilistic functionalism investigates the ways that interdependent, intersubstitutable, and uncertain cues about external environments are used to make judgments about these environments. Environments are not fully determined and uniform in their structures and processes of causality; rather they are uncertain, unstable, and *causally textured* (Tolman and Brunswik 1935). Probabilistic functionalism also studies the manner in which knowers learn about their environments by using informational cues or indicators, continuously creating new cues and revising or abandoning old ones. An essential aspect of these cues is that they are interdependent, multiply correlated, and intersubstitutable, features that require a process of knowing that is approximate, indirect, and vicarious. Darwinian in its focus on the adaptation of the organism to its environment, probabilistic functionalism seeks to understand how these two complex systems, the organism and the environment, come to terms with one another through a process of pattern matching.

2.2 The Brunswik Lens Model

Brunswik (1952) developed his 'lens model' to illustrate the probabilistic interrelations between the organism and the environment. The lens model represents processes of perception and judgment in terms of the optics metaphor of light entering (converging on) and exiting (diverging from) a double-convex lens. The lens model, while initially metaphorical in character, was later specified in what has come to be known as the 'lens model equation' (Slovic and Lichtenstein 1971, pp. 656–67):

$$r_a = R_e R_s G + C[(1 - R_e^2)(1 - R_s^2)]$$

The lens model supplies an achievement index, A , which measures the extent to which pattern matching has been achieved (see figure below). The achievement index ($A = r_{Y_e Y_s}$) is the correlation between the stat-

istical properties of an uncertain environment (Y_e) and the statistical properties of a judge's (or knower's) response system (Y_s). A matching index, G , is the correlation between the predicted knower's response (Y'_s) and the predicted criterion value for the environment (Y'_e). The lens model also estimates the degree to which indicators or cues are used in a linear or nonlinear fashion. In the equation above, the coefficient, C , is the correlation of the nonlinear residual variance in the multiple correlation coefficients R_e and R_s .

2.3 Representative Design

The methodology of *representative design* requires that experiments be conducted in realistic settings that are representative of an organism's typical ecology. Representative design is a radical departure from the classical experiment, in which one independent (treatment) variable is manipulated so as to assess its effect on a dependent (criterion) variable, with all other factors held constant through random subject sampling and statistical controls. Instead, experiments were to be conducted in natural settings, in environments that are unstable, dynamic, uncertain, and probabilistic. Such experiments require *situation sampling*, not merely the random sampling of subjects, because only the former optimizes the representative character of experiments. Representative design, in addition to its call to abandon the classical experiment, also required a rejection of the uniformity of nature presumption underlying Mill's Canons. Both ignored, and therefore failed to account for, the effects of the many interrelated contingencies that are causally relevant to experimental outcomes. Significantly, although Brunswik was an active participant in the Unity of Science Movement, and thus to some degree seems to have identified with logical positivism, he rejected the uncritical imitation of physics promoted by Vienna Circle positivists and the bulk of psychologists of his time. Psychology, and by extension other social and behavioral sciences, was practicing 'emulative physicalism' in its attempt 'to copy not only the basic methodological principles but also the specific thematic content of physics, thus nipping in the bud the establishment of (appropriate) methodological directives' (Brunswik 1952, p. 36; quoted in Hammond 1966, p. 55).

3. Methodology Development

Probabilistic functionalism and representative design have influenced the contributions of several highly influential scholars who studied with Brunswik and Tolman at Berkeley. Some of these contributions are principally theoretical, for example, social judgment theory (Hammond 1980) and evolutionary epistem-

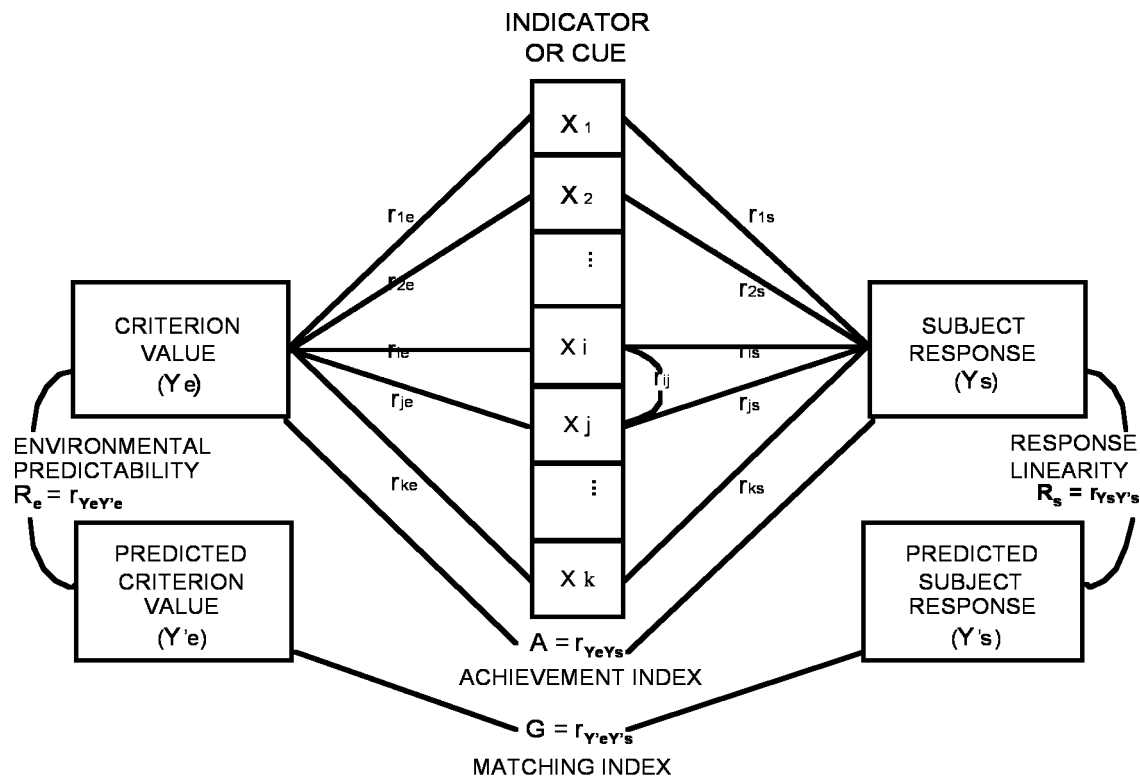


Figure 2

The Brunswik lens model (source: adapted from Slovic and Lichtenstein 1971)

ology (Campbell 1974, 1996). It is in the area of methodology development, however, where many of the most important contributions have been made.

3.1 The Lens Model and Clinical Inference

The lens model, backed by probabilistic functionalism and representative design, has been at the center of an entire research tradition on clinical inference (see Hammond 1980, 1996). Significantly, many of the most important lens-model studies have been applied studies conducted with the aim of improving learning, judgment, and behavior. In this tradition, research problems have turned on the question of how pattern matching can explain differences in achievement (predictive accuracy) among individuals with different levels of experience, including novices as well as experts. The right side of the lens model (Fig. 2) is used to predict a 'distal' environmental variable (e.g., future university enrollments) by regressing individual judgments about the distal variable on a set of interrelated and mutually substitutable informational cues (e.g., unemployment, income per capita, changes in age structure). In turn, the left side of the model focuses on

predictions of that same distal variable derived from a multiple regression analysis in which the same cues (although here they are 'indicators' or 'variables') are predictors.

An important variant of this basic design is one where members of different professional groups—for example, scientists and lawyers in some area of science policy—occupy the right and left sides of the lens model. Either group may be taken as a reference class, and the calculation of the matching index, M , expresses the degree to which their prediction patterns match. In applied contexts, the judgments of each participant (subject) are externalized and made available to other participants. In addition, the way that each participant uses information (e.g., in a linear or nonlinear fashion) is estimated through a process of 'policy capturing.' These and other applications of the lens model have been designed to improve learning and judgments in contexts ranging from R&D planning, police work, gun control, and human factors research. The lens model has a computerized decision support program ('Policy PC') and its theoretical foundation has been redefined from social judgment theory to *cognitive continuum theory* (Hammond 1996).

3.2 Methodological Triangulation

The recognition that nature is not directly observable—that our predicament as knowers is that we must employ many intercorrelated and mutually substitutable proximal cues to infer the properties and behavior of distal objects—means that science and other forms of ‘distal knowing’ involve a process of pattern matching through triangulation (Campbell 1966). A number of important methodologies were developed on the basis of this recognition. One of these is the substitution of definitional operationism with multiple operationism, a substitution that involves triangulation among two or more operational definitions, each of which is seen as approximate, fallible, and independently imperfect. A second development was the expansion of multiple operationism to include theoretical constructs as well as methods for their measurement. Here the multitrait-multimethod matrix reconceptualized *construct validity* as a ‘trait-method unit,’ and introduced the concepts of *convergent validity* (multiple measures of the same construct should converge) and *discriminant validity* (multiple measures of different constructs should diverge) that parallel aspects of Brunswik’s metaphor of a double-convex lens. The 1959 article in which the multitrait-multimethod matrix was first published (Campbell and Fiske 1959) is reputed to be one of the most highly cited in the social and behavioral sciences. A third development is multiple triangulation, including critical multiplism (Cook 1984), which involves the inclusion of multiple theories, methods, measures, observers, observations, and values. These and other forms of methodological triangulation enable ‘strong inference’ in the natural and social sciences (Platt 1964) and affirm that the vast bulk of what is known is based on processes of indirect, vicarious learning.

3.3 Quasi-Experimental Design

The methodology of representative design, as we have seen, rejected the classical experiment on grounds that it is unrepresentative of the usual ecology of in which knowers function. Representative design was carried forward into the applied social sciences by Donald T. Campbell and associates (Campbell and Stanley 1963, Cook and Campbell 1979). Their quasi-experimental designs were contrasted with the classical laboratory experiments in which: an outcome variable is explained by a single independent (treatment) variable (the so-called ‘rule of one’); other possible explanations are ruled out through random selection of subjects; and the experimenter has virtually complete control over all contingencies. Quasi-experimentation, although it may use some of the features of classical experiments (e.g., repeated measures and control groups) should be contrasted with experiments in the analysis of variance tradition of Ronald Fisher, who

envisioned experimenters who ‘having complete mastery can schedule treatments and measurements for optimal statistical efficiency, with the complexity of design emerging only from that goal of efficiency. Insofar as the designs discussed in the present chapter become complex, it is because of the intransigency of the environment: because, that is, of the experimenter’s lack of complete control’ (Campbell and Stanley 1963, p. 1).

Because quasi-experimental designs are intended for research in settings in which numerous contingencies are beyond the control of the experimenter, many rival hypotheses (alternative explanations of the same outcome) can threaten the validity of causal claims. These rival hypotheses are organized in four sets labeled threats to statistical conclusion, internal, external, and construct validity (Cook and Campbell 1979, Chap. 2). Plausible rival hypotheses must be tested and, where possible, eliminated. This process of eliminative induction is a qualified form of Mill’s *joint method of agreement and difference* and Karl Popper’s *falsificationist* program. Quasi-experimentation is part of a wider evolutionary critical-realist epistemology (see Campbell 1974, Cook and Campbell 1979, Shadish et al. 2000) according to which knowers adapt to real-world environments by using overlapping and mutually substitutable informational sources to test and improve their knowledge of indirectly observable (distal) objects and behaviors. Quasi-experimentation is a form of pattern matching.

3.4 Pattern-Matching Case Studies

When quasi-experimental designs are unfeasible or undesirable, several forms of case study analysis are available. Each of these involves pattern matching.

(a) *Theory-Directed Case Study Analysis*. When a well-specified theory is available, a researcher can construct a pattern of testable implications of the theory and match it to a pattern of observations in a single case (Campbell 1975). Using statistical ‘degrees of freedom’ as a metaphor, the theory-directed case study is based on the concept of ‘implications space,’ which is similar to ‘sampling space.’ Testable implications are functional equivalents of degrees of freedom, such that the more implications (like a larger sample) the more confident we are in the validity of the conclusions drawn. But because theories are almost inevitably affected by the culturally acquired frames of reference of researchers, the process of testing implications should be done by at least two ‘ethnographers’ who are foreign to and native to the culture in which the case occurs. The process of triangulation among observers (ethnographers) can be expanded to include two (or more) cases.

(b) *Qualitative Comparative Case Study Analysis*. Two or more cases are compared by first creating a list of conditions that are believed to affect a common

outcome of interest (see Ragin 1999, 2000). The multiplication rule, r^m , is used to calculate the number of possible ordered configurations of r categories, given m conditions. When $r = 2$ and $m = 4$, there are $2^4 = 16$ configurations, each of which may involve causal order. These configurations become the rows in a 'truth table,' and each row configuration is sequentially applied to an outcome with r categories (e.g., successful vs. unsuccessful outcome). Because this method examines possible configurations, and two or more different configurations may explain the same outcome in different cases, the qualitative comparative method should be contrasted with traditional (tabular) multivariate analysis. The qualitative comparative method matches configurable patterns that have been formally structured by means of set theory and Boolean algebra against patterns of observations in case materials.

(c) *Modus Operandi Analysis*. When quasi-experimental research is not possible, *modus operandi* methods (see Scriven 1975) may be appropriate for making causal inferences in specific contexts. *Modus operandi* methods are based on the analogy of a coroner who must distinguish symptoms and properties of causes from the causes themselves. The first step is to assemble a list of probable causes, preferably one that is quasi-exhaustive. The second is to recognize the pattern of causes that constitutes a *modus operandi*—modus refers to the pattern, while operandi refers to specific and 'real' causes. The *modus operandi* of a particular cause is its characteristic causal chain, which represents a configuration of events, properties, and processes. *Modus operandi* analysis has been formalized, partially axiomatized, and advanced as a way to change the orientation of the social and behavioral sciences away from abstract, quantitative, predictive theories toward specific, qualitative, explanatory analyses of causal patterns (Scriven 1974, p. 108).

4. Conclusion

To date, the only thorough and systematic theory of pattern matching is that of probabilistic functionalism and its methodological complement, representative design. Pattern matching is essential for achieving knowledge of external objects and events, of causal regularities that have been repeatedly recognized as the same, of the character of discrete observations formed by viewing them in context, of the fit between a curve and a set of data points, of the extent to which theories cohere, and of the degree to which theory and data correspond. Underlying all pattern matching methodologies is a shared recognition that knowing in science is profoundly indirect, vicarious, and distal.

See also: Case Study: Logic; Case Study: Methods and Analysis; Causation (Theories and Models):

Conceptions in the Social Sciences; Constructivism/Constructionism: Methodology; Explanation: Conceptions in the Social Sciences; Laboratory Experiment: Methodology; Problem Selection in the Social Sciences: Methodology; Qualitative Methods, History of; Triangulation: Methodology

Bibliography

- Alcoff L M 1996 *Real Knowing: New Versions of the Coherence Theory of Truth*. Cornell University Press, Ithaca, NY
- Anscombe F J 1973 Graphs in statistical analysis. *American Statistician* 27: 17–21
- Ayer A J 1936 *Language, Truth, and Logic*. Golanc, London
- Brunswick E 1956 *Perception and the Representative Design of Psychological Experiments*, 2nd edn. University of California Press, Berkeley, CA
- Brunswick E 1952 *The Conceptual Framework of Psychology*. University of Chicago Press, Chicago
- Campbell D T 1959 Methodological suggestions for a comparative psychology of knowledge processes. *Inquiry* 2: 152–82
- Campbell D T 1966 Pattern matching as an essential in distal knowing. In: Hammond K R (ed.) *The Psychology of Egon Brunswik*. Holt, Rinehart, and Winston, New York
- Campbell D T 1974 Evolutionary epistemology. In: Schilpp P A (ed.) *The Philosophy of Karl Popper*. Open Court Press, La Salle, IL
- Campbell D T 1975 "Degrees of Freedom" and the case study. *Comparative Political Studies* 8: 178–93
- Campbell D T 1986 Science's social system of validity-enhancing collective belief change and the problems of the social sciences. In: Fiske D W (ed.)
- Campbell D T 1988 *Methodology and Epistemology for Social Science: Selected Papers*. University of Chicago Press, Chicago
- Campbell D T 1996 From evolutionary epistemology via selection theory to a sociology of scientific validity. *Evolution and Cognition*
- Cook T D, Campbell D T 1979 *Quasi-Experimentation: Design & Analysis Issues for Field Settings*. Houghton Mifflin, Boston
- Hammond K R (ed.) 1966 *The Psychology of Egon Brunswik*. Holt, Rinehart, and Winston, New York
- Hammond K R 1980 *Human Judgment and Decision Making: Theories, Methods, and Procedures*. Praeger, New York
- Hammond K R 1996 *Human Judgment and Social Policy: Irreducible Uncertainty, Inevitable Error, Unavoidable Injustice*. Oxford University Press, New York
- Putnam H 1981 *Reason, Truth, and History*. Cambridge University Press, Cambridge, UK
- Quine W V 1951 Two dogmas of empiricism. *Philosophical Review* 60(1): 20–43
- Ragin C C 1999 Using comparative causal analysis to study causal complexity. *HSR: Health Services Research* 34(5), Part II
- Ragin C 2000 *Fuzzy-Set Social Science*. University of Chicago Press, Chicago
- Rescher N 1980 *Induction: An Essay on the Justification of Inductive Reasoning*. University of Pittsburgh Press, Pittsburgh, PA

- Schweder (eds.) *Metatheory in Social Science: Pluralisms and Subjectivities*. University of Chicago Press, Chicago
- Scriven M 1975 Maximizing the power of causal investigations: The modus operandi method. In: Glass G V (ed.) *Evaluation Studies Review Annual*. Sage Publications, Beverly Hills, CA, Vol. 1
- Shadish W, Cook T, Campbell D T 2000 *Quasi-Experimentation*. Houghton Mifflin, Boston
- Slovic P, Lichtenstein S 1971 Comparison of bayesian and regression approaches to the study of information processing in judgment. *Organizational Behavior and Human Performance* 6(6): 649–744
- Trochim W 1990 Pattern matching and psychological theory. In: Chen R (ed.) *Theory-Driven Evaluation*. Jossey Bass, San Francisco

W. N. Dunn

Pavlov, Ivan Petrovich (1849–1936)

Ivan Petrovich Pavlov is recognized as one of the most renowned of Russian scientists and is clearly the best known Russian physiologist of his era. His work on the physiology of the digestive system was well respected by his colleagues worldwide during his lifetime. This respect earned him the first Nobel Prize awarded to a physiologist and the first to a Russian scientist in 1905. Although the excellence of his work on the physiology of the digestive system is well recognized, Pavlov is even better known for his investigations into what he termed ‘conditioned reflexes.’ This work has had a worldwide impact on the fields of learning and memory, psychology, neuroscience, etc. Pavlov’s life can be divided roughly into three phases. The first, including his early years, was a phase in which Pavlov received his early training in physiology and medicine. A second investigational phase was one in which his work on the physiology of the digestive system was done. The third includes his work, for which he became most famous, on conditioned reflexes.

1. Pavlov’s Early Life

Ivan Petrovich Pavlov was born on September 14, 1849, in the small town of Ryazan, about 100 miles southeast of Moscow. He was the eldest of five children. His father was a priest in one of the poorer parishes. Thus, during Pavlov’s early life he lived in extreme poverty. Nevertheless, the life of Pavlov and his siblings was apparently intellectually rich and their father instilled a love of learning in each of his children. Pavlov entered the ecclesiastical high school of Ryazan, and from there to the local seminary, ostensibly to pursue an ecclesiastical career, as his father and other members of his family had done. Pavlov thus entered school during the Russian Cultural Revolu-

tion of the 1860s and therefore was exposed to the leaders of this revolution, such as Turgeanyev, Doskovsky, Tolstoy, etc. However, he was apparently most heavily influenced by Pisarev, from whom he learned of Darwin’s theory of natural selection and the famous Russian physiologist Sechenov, who had just published *Reflexes of the Brain* in 1896. It is clear that these kinds of influences on Pavlov steered him in the direction of science and consequently he left the ecclesiastical seminary and entered the Mathematics-Physics Department at St. Petersburg State University.

In 1875 Pavlov graduated from the University with a degree in natural sciences. However, due to his interest in physiology he started work as an assistant in the Department of Physiology in the Medical Surgery Academy (now the Military Medical Academy) and in 1879 received a medical diploma. He was then recommended for an appointment with Andre Botkin, who was the most renowned clinician in St. Petersburg at the time. Botkin assigned Pavlov the task of heading his recently established experimental research laboratory. Since Botkin had little interest in experimental work, Pavlov was able to work pretty much independently on problems of his choice. He worked here until 1890. He was heavily influenced by Botkin’s theories on the role that the nervous system plays in the regulation of physiological function, and it was here that Pavlov began his studies of the peripheral nerves of the heart and where in 1883 he successfully defended his doctoral dissertation in this area.

During this period of time he also became acquainted with a small circle of friends including another university student, Seraphima Vasilievna, who was in her last year at the Pedagogical Institute. They were married in 1881. During the first years of their married life they lived in extreme poverty. After a first miscarriage, Seraphima gave birth to a son but he also died the next summer. They did, however, later have four children (three boys and a girl) and, based on Seraphima’s memoirs published after Pavlov’s death, apparently enjoyed a happy family life. In 1883 Pavlov spent two years working abroad in the laboratories of Heidenhain and Ludwig in Germany. On his return to St. Petersburg, he resumed his work in Botkin’s laboratory and in 1890 was appointed Professor in the Pharmacology Department at the Military Medical Academy. Five years later he joined the Physiology Department there, which he headed for the rest of his career. He was also later appointed the Head of the Department of Physiology in the newly established Institute of Experimental Medicine. Pavlov worked in his laboratory there until his death at the age of 87 in Leningrad on February 27, 1936.

2. Pavlov’s Work on the Digestive System

Pavlov did most of his research on the physiology of digestion during his early years at the Institute for

Experimental Medicine. Here he was able to develop his surgical skills with chronic experiments, which he advocated over the acute experiments that were typically done in physiology at that time. Pavlov believed that the study of the entire organism in as normal a state as possible was crucial to understanding physiological systems. Moreover, he was an excellent surgeon and became quite experienced in developing surgical procedures that allowed his dogs to fully recover. In such animals the digestive processes could be studied under normal conditions. As a consequence of this approach to physiology he developed a ‘stomach pouch,’ which involves isolating a portion of the dog’s stomach, so that it can be directly observed by the experimenter. Unlike previous researchers, the nervous input to the isolated portion of the stomach was intact in the ‘Pavlov Pouch.’ Using this preparation Pavlov could observe the digestive juices and enzymes that occurred as a result of normal food intake, or after direct placement of food into the stomach. It was the comparison of these two states which led to Pavlov’s interesting discovery that food placed in the mouth as a normal consequence of eating, produced a significantly larger amount of gastric secretion than when food was placed directly in the stomach. Moreover, when animals were sham fed, viz. when the normal intake of food through the mouth was allowed to take place but the food was diverted externally, a significant amount of gastric secretion occurred, even though the food never reached the stomach.

It was these types of experiments, which led to Pavlov’s pioneering discoveries regarding the control of the digestive system by its autonomic nervous input, primarily the vagus nerve. He showed for example that a major consequence of this neural control involves both gastric and pancreatic secretions before any food actually reaches the stomach, by the sight of the food bowl for example, or other stimuli previously associated with eating. These ‘psychic secretions’ turned out to be of great importance for digestion and led to Pavlov’s major conclusions regarding what came to be known as the doctrine of nervism, in which the major tenet states that the physiological control of bodily systems is primarily through its nervous input. This work culminated in the publication of *Lectures on the Work of the Digestive Glands* in 1897, later translated into English in 1910 (Pavlov 1910). However, it was during this same period of time that two British physiologists, W. Bayliss and E. J. Starling, demonstrated that the secretin of the pancreatic enzymes was primarily due to release of a substance by the intestines during eating, which came to be called secretion. These investigators thus claimed that the doctrine of nervism was invalidated by their discoveries. However, as is typically the case with two opposing theories of biological function, both experimental outcomes were eventually confirmed. The new science of neuroendocrinology thus came to be

established, and it is now widely accepted that nervous control over hormonal output is a ubiquitous aspect of physiological function.

3. Pavlov’s Work on Conditioned Reflexes

As a result of his work on the digestive processes, as noted above, Pavlov also observed that non-food stimuli that became associated with food were able to elicit salivary secretions, and to a lesser extent digestive secretions, even though no food was actually present in the mouth. These ‘psychic secretions’ came to form the basis for the remainder of Pavlov’s work and resulted in his experiments moving in a new direction, which caused an almost complete cessation of work on the digestive system proper. However, Pavlov believed that by studying ‘psychic secretions,’ which he referred to as ‘conditioned reflexes,’ he would be able to demonstrate in an objective manner how the brain controls adaptive behaviors. Pavlov developed a salivary fistula, similar to the stomach pouch, in which the release of salivation in the mouth was directed through a tube to a container outside the mouth. Through this technique he was able to discover the amount and kinds of salivary secretions that were produced by different kinds of sensory stimuli.

Using these techniques Pavlov began the study of a new kind of learning, which has come to be known as ‘classical’ or ‘Pavlovian’ conditioning. Classical conditioning occurs whenever a neutral stimulus acts as a signal for a forthcoming significant event. Thus, for example, in one of Pavlov’s original experiments a pure tone, a metronome, or some other equally neutral stimulus signaled to a dog that it was about to receive an appetitive stimulus, i.e., meat powder (see Pavlov 1927). In Pavlov’s experiments this neutral auditory stimulus was termed the conditioned stimulus (CS) and the meat powder the unconditioned stimulus (UnS). The UnS always elicits what Pavlov referred to as an unconditioned response (UR), viz. salivation in response to the meat powder. The behavior in which Pavlov became most interested, however, was not increased salivation in response to the meat powder, which he had previously studied intensively, but the new learned response to the neutral CS. This response also consisted of salivation, but in this case salivation in response to the initial neutral auditory stimulus, which resulted from its consistently preceding the meat powder over many CS/UnS presentations. This new learned response was referred to as a conditioned response (CR) and was thought to be only temporary and required reinforcement by the UnS for its maintenance. The UnS is thus often referred to as a reinforcer, since it reinforces the new response to the CS.

It should be noted that the CR does not, however, always resemble the UR as is the case with the original salivation experiments, which Pavlov reported. Indeed

often the CR appears to be opposite to the UR. For example, the autonomic changes associated with the contextual cues that signal drug administration, an often-studied type of classical conditioning, are opposite in direction to those produced by the drug itself (Siegel 1979) and the heart rate CR to CSs that signal aversive UnSs consist of bradycardia, whereas the UR to these same aversive stimuli alone consists of tachycardia. There are many other experimental operations that have been studied since the time of Pavlov, which produce similar new responses to an original neutral stimulus that is always followed by either a noxious or appetitive event. Pavlov thus developed a new experimental methodology for studying the role of the brain as it initiates new behaviors enabling animals to adapt to their environmental circumstances.

One of Pavlov's greatest contributions to physiology was to emphasize that the nature of the 'psychic secretions,' which were previously thought to be in the realm of psychology, could be objectively studied by physiologists through the conditioned reflex method. He thus believed that the true road to understanding brain function and therefore human behavior was through the objective physiological techniques, which he had developed. In fact Pavlov went on to consider a range of higher level functions such as thinking, reading, emotional reactivity, etc., in terms of conditioned reflexes. Using the ideas of cortical excitation and inhibition, which he studied extensively using the conditioned reflex methodology, he was thus able to explain many complicated behavioral phenomena, such as the neuroses and other psychiatric disturbances. Although Pavlov's influence on experimental psychology has been extensive and dramatic, as detailed in Sect. 4, many of the details of his theoretical interpretations have since been found to be lacking. Nevertheless the basic operational procedures for determining the effects of a signal on brain processing has been extremely influential in studying brain function.

4. Pavlov's Influence on Modern Psychology

During his lifetime and immediately thereafter, Pavlov had a tremendous influence on physiology and the study of brain function. Many students were drawn to his laboratory during the early part of the twentieth century to study the basic laws governing the activity of the brain, and during this period of time Pavlov received worldwide acclaim and recognition. Pavlov's work was little affected by the Russian Revolution, which was also ongoing during this time. He maintained a skeptical attitude regarding politics and government, but nevertheless his worldwide recognition led the Communists to continue to fund his research at high levels. Thus, the Soviet Union became known for its support of the study of physiology, and

it was during this time that a great center for the study of physiology with many distinguished workers was developed in the Soviet Union, primarily under Pavlov's leadership.

It was his influence outside the Soviet Union, however, that resulted in Pavlov's most notable successes in science, primarily in the field of psychology, which Pavlov had previously rejected. Thus the conditioned reflex methodology was instrumental to the development of the behavioristic movement in psychology in the early part of the twentieth century. John Broadus Watson, one of the major pioneers in the development of behaviorism, utilized the conditioned reflex methodology to explain the entire subject field of psychology, which consisted, according to behaviorists, in the study of overt behavior without reference to subjective phenomena such as cognitions, learning, feelings, etc. The publication of *Psychology as the Behaviorist Views It* (1913) by Watson thus became instrumental in informing American and English scientists of Pavlov's work. The translation of Pavlov's *Conditioned Reflexes* (1927) by Anrep made his work available to English speaking scientists. As a result, experimental psychologists began to explore the new conditioned reflex methods for studying behavior at several academic centers in the USA. As was noted by Babkin (1949), however, much of the work done by the American researchers focused on skeletal reflexes, as opposed to the visceral reflexes, to which Pavlov and his students had previously devoted most of their work.

One of the basic experiments developed by American psychologists was the classical eyeblink-conditioning paradigm. Using this methodology human or animal subjects are presented with a corneal airpuff, which causes reflexive closure of the eyelids. However, when this unconditioned stimulus is preceded by a conditioned stimulus such as a light or pure tone, eventually subjects began to show anticipatory eyeblinks during the conditioned stimulus, which occur even though the corneal airpuff is not presented. This paradigm came to be used extensively in the early 1940s and 1950s as behaviorism began to gather momentum. Its success in the hands of the American psychologists led to the de-emphasis on visceral changes as conditioned responses in the USA. However, Horsley Gantt, who had previously studied with Pavlov, demonstrated in his Johns Hopkins laboratory that both visceral and skeletal responses could be studied in the same organisms, and that the two responses differed greatly in their acquisition and the conditioning parameters required to elicit learning (Gantt 1960). Much of this early work on classical conditioning was, however, overshadowed by the emphasis given to operant conditioning by other behaviorists, such as B. F. Skinner and Clark Hull. Thus, it was only during the early 1960s that classical conditioning became a popular technique for studying behavior. This was due to the development by I.

Gormezano of the classically conditioned nictitating membrane response in the rabbit (Gormezano 1966). The nictitating membrane is a third laterally moving eyelid found in the rabbit and some other mammals (e.g., cat). Using this animal preparation Gormezano and his students were able to demonstrate the parametric circumstances under which this kind of learning takes place and what kinds of visceral changes accompany them, e.g., conditioned changes in heart rate, blood pressure, etc. The popularity of this technique became even greater when integrated with the simultaneously developing field of behavioral neuroscience. Classical conditioning offers several advantages for studying concomitant brain function as well as new learned behaviors. For example, concomitant electrophysiological recording from single neurons in specific parts of the brain have led to the discovery that different brain structures are involved in visceral versus skeletal learning (Thompson 1991). Moreover, more recent brain scanning techniques in humans have demonstrated that even during simple classical eyeblink conditioning activation of several specific areas of the brain, which are known to be involved in learning and memory processes, occurs (e.g., Blaxton et al. 1996).

Thus the use of the conditioned reflex techniques originally developed by Pavlov has come to be one of the major methods used in studying brain-behavior relationships in modern psychology. Pavlov's contribution to this new technology was at a basic level and much of his earlier conclusions regarding the results of his manipulations were, of course, erroneous. However, his contributions we now know were instrumental to the development of modern psychology.

See also: Autonomic Classical and Operant Conditioning; Conditioning and Habit Formation, Psychology of; Experimentation in Psychology, History of; Psychological: Historical and Cultural Perspectives

Bibliography

- Babkin B P 1949 *Pavlov: A Biography*. University of Chicago Press, Chicago, IL
- Blaxton T A, Zeffiro T A, Gabrieli J D E, Bookheimer S Y, Carrillo M C, Theodore W H, Disterhoft J F 1996 Functional mapping of human learning: A positron emission tomography activation study of eyeblink conditioning. *Journal of Neuroscience* **16**: 4032–40
- Gantt W H 1960 Cardiovascular component of the conditional reflex to pain, food, and other stimuli. *Physiological Review* **40**: 266–91
- Gormezano I 1966 Classical conditioning. In: Sidowski J B (ed.) *Experimental Methods and Instrumentation in Psychology*. McGraw Hill, New York
- Pavlov I P 1910 *Lectures on the Work of the Digestive Glands*, 2nd edn. Translated by W H Thompson. Charles Griffin and Co., London

- Pavlov I P 1927 *Conditioned Reflexes*, translated by G V Anrep. Oxford University Press, London
- Sechenov I M 1866 *Refleksy Golovnogo Mozga (Reflexes of the Brain)* 1965. MIT Press, Cambridge, MA
- Siegel S 1979 The role of conditioning in drug tolerance and addiction. In: Keehn J D (ed.) *Psychopathology in Animals*. Academic Press, New York
- Thompson R F 1991 Are memory traces localized or distributed? *Neuropsychologia* **29**: 571–82
- Watson J B 1913 Psychology as the behaviorist views it. *Psychological Review* **20**: 158–77

D. A. Powell

Pay and Compensation, Psychology of

1. The Design of Compensation

In industrialized countries the compensation of managers and employees usually consists of three or four component parts. One part is rooted in the value of the job which is held, providing for the fixed base pay (wage or salary). The second part reflects an assessment of the effort exerted or the performance result, often called the 'performance bonus.' The third part deals with 'secondary labor conditions,' such as health insurance, retirement provisions, etc. Firms in several countries offer still additional benefits—perquisites—to their personnel (or merely their managers), such as car lease, lower interest rates on mortgage loans, etc. However, there are sizable differences between (and often also within) countries regarding the proportional contribution of each of these parts to an individual's total compensation.

Yet the common design in the composition of compensation in most countries is remarkable. This cuts across considerable differences between countries in industrial relations, wealth and welfare, and culture. Why? Probably partly because traditions in trades and industrial sectors, laws and regulations, and the ideas of labor unions and employers' federations are playing a role. And also, in all likelihood, partly because of some widely shared beliefs and attitudes about the differential impact of the various compensation components upon the work behaviors of organization members. This borders on the psychological perspective on pay, which addresses the impact of compensation upon the attitudes and work behaviors of organization members, both individually and as part of teams or larger work units. More particularly, the psychological study of pay and compensation aims to identify determining variables at different levels of analysis (such as pay system, or strength of individual expectations), moderating aspects (such as the nature of the task), and individual, group, or company-wide results (such as performance or productivity).

This perspective determines the design of this article. First, the nature of the four compensation components

is further described, in relation to some characteristic beliefs and attitudes on how they operate. Next, a few psychological theories on pay and compensation will be discussed. Finally, some implications for psychological research on pay and applications in practice are discussed.

2. Four Components and Concomitant Beliefs

Although the background of the compensation components is discussed in relation to what is characteristically expected regarding each of them, I do not imply that a company's management is always engaged in decision making along these lines. Indeed, pay systems and compensation provisions may have been introduced in a firm for reasons of tradition, imitation of other companies, or merely because alternative choices were not considered.

2.1 Base Pay

In most countries base pay constitutes the main part of an organization member's take-home pay. Yet there may be sizable differences in the proportion of base pay between trades and job categories. Traditionally, the pay of sales representatives is determined to a large degree by actually achieved sales figures. Managerial pay is characterized by a proportionally lower base-pay component than nonmanagerial pay, in particular for senior managers (Rynes and Bono 2000). Base or job pay is often set through a system of job evaluation (cf. Thierry and De Jong 1998). Increasingly, international wage and salary surveys are used, as well as more subjective impressions of the going rate on the market. Base-pay levels are usually fixed as long as the term of individual contracts or collective agreements applies.

Often the fixed base-pay level is a reflection of the belief that stability in pay contributes to individual feelings of security. It is also supposed that without the need to worry about their pay, working people will concentrate more upon the content of their tasks. Moreover, a high base-pay level may imply the notion that work motivation is considerably more affected by nonmaterial needs and goals than by material ones. Possibly this results in more intrinsic motivation (cf. Deci and Ryan 1985, Deci et al. 1999). Characteristically, levels of base pay increase each year, usually through an additional 'step' on the wage or salary scale. Although the latter increase may have been awarded to make up for increased costs of living, it often also mirrors the idea that more maturity (i.e., getting older and becoming more experienced) goes along with better performance.

2.2 Results-oriented Pay

This category, also called 'pay for performance,' is host to a variety of systems and forms. Results may be

derived from individual performance appraisal sessions, in which rather abstract characteristics (such as initiative or creativity) are interpreted and considered. Results may also refer to quantitative outcome scores, for instance the number of new bank clients secured and the volume of their investments. Results may relate to individual data, but also to those of teams, units, and whole organizations (such as sales volume or net profit). These data may be provided weekly, monthly, or over a much longer period. The 'pay' is often allocated in the form of a bonus (paid only once), but it may also be given as an extra, 'structural' salary increase. The amount of results-oriented pay may be rather modest (e.g., 3 percent), but it may also exceed the base pay's proportional part of total compensation (see, more extensively, Rynes and Bono 2000, Bartol and Locke 2000, Thierry in press).

What is results-oriented pay supposed to effectuate? A widespread belief holds that it stimulates an individual's work motivation, effort, or performance. Research evidence (Thierry in press) shows that this can be frequently supported, although there are also many instances of failure. Another assumption is that results-oriented pay is perceived by many employees (and managers) as more fair, since it compensates for perceived individual and group differences in effort and/or performance (Van Silfhout 2000). Organizations in which this assumption is endorsed usually assign a particular budget to covering these merit-pay expenses. Making pay dependent upon performance results may be used to cut costs as well. Moreover, it may be used to feed performance data back to employees, and, more generally, for communicating with them.

More than is the case for other compensation components, the phase of preparing and introducing results-oriented pay is rather complicated and time consuming. It may imply that a part of the work flow is being organized more efficiently, that norms and targets are being established, that results are being fed back, that employees and managers are being consulted about the new system, and so forth. In particular, when better performance results have been reached, a point of concern is that these activities may have contributed to these results rather than the linkage between performance and pay.

2.3 Secondary Labor Conditions

There are huge differences between countries in the handling of secondary conditions. Traditionally, companies in the USA offered a very modest package to their members—leaving to them the decision whether or not further to insure themselves—whereas companies in northwestern Europe provide encompassing social security within the framework of collective labor agreements. In the latter case, conditions relate,

for example, to health insurance, disability coverage, retirement benefits, or unemployment provisions. These are intended to help working people to compensate for risks they run in their work and lives. Yet, such social security packages are not very well tailored to the individual organization member's conditions and wishes. Moreover, because of health research and technological innovations, there is a continuing trend to expand the number and the nature of social security provisions.

It is therefore an interesting development that the 'cafeteria plan' is gaining popularity on both sides of the Atlantic (Barber and Bretz 2000). According to this plan (also called 'à la carte' or 'flex-pay') organization members may exchange one or more standard labor conditions for optional ones. In US companies this plan was initiated some 40 years ago, in order to gradually expand social security coverage. Consequently, managers and employees may tailor their social security package to their particular conditions, which may lead to more utility against similar or lower costs for the company. More recent European experience, for instance in the UK and The Netherlands (cf. Thierry 1998), indicates that organizations think the plan equips them better as competitors in the labor market for 'flexible, employable' applicants. They may also express their particular culture and climate in their choice of alternative provisions. Consequently, their members are expected to be more motivated, to work better, to be more satisfied with their pay package, and to have more knowledge of their compensation components. Research evidence shows that only the latter two expectations are frequently met (Thierry 1998, in press).

2.4 Perquisites

Organizations may offer their personnel (or selected categories, such as senior managers) still further labor conditions. Early examples came in the manufacturing sector, in which an allowance for occupational clothing or cloth cleaning was sometimes provided. Current practices include (mobile) phone cost coverage, car leasing, lower mortgage interest, shares on the stock market, options on company shares, and so forth. Some of these options may be available through a cafeteria plan (see Sect. 2.3). Perquisites were provided to make up for particular inconvenient conditions and occupational costs, but current objectives relate to attracting potential applicants on the labor market and, in particular, to keeping qualified personnel within an organization.

2.5 Beliefs in Perspective

The preceding overview of compensation components shows that each component is characterized by rather

distinct beliefs and expectations about how compensation may affect the attitudes and work behaviors of managers and employees. Yet an important issue is whether these beliefs are founded in concepts relevant to psychological theories on pay and compensation. Some of these will be reviewed here (more extensively in Thierry in press), but the various beliefs are to be recognized in almost all objectives of pay (cf. Thierry 1998):

- to *attract* applicants from the labor market;
- to *retain* qualified personnel within an organization;
- to *motivate* managers and employees to perform well;
- to *alert* managers and employees to engage in different performance behaviors in times of change;
- to *balance* inconvenient working conditions with an attractive award;
- to *prevent* industrial conflicts (or to solve these where they have occurred);
- to *cut down* labor costs; and
- to further a distinct, recognizable business unit or group *culture*.

3. Some Psychological Theories on Pay

Most psychological theories on pay have not been explicitly designed for understanding and predicting people's behaviors and attitudes regarding pay. Rather, they have been derived from more general theories, for instance regarding motivation. The theories selected here stem from both 'categories.'

3.1 Expectancy Theory

This theory concerns matters of work motivation (Vroom 1964, Van Eerde and Thierry 1996), in particular the choice process of an employee being faced with the necessity or the opportunity to mark a course of action among alternative activity patterns. Suppose that an employee considers whether to spend all their energy and resources in making the current job a great success, hoping for a promotion in the not too distant future, or whether to apply for a job in another organization, which involves taking several courses in order to become more employable. Expectancy theory holds that three cognitions determine the eventual choice:

- the probability that a particular level of effort leads to a specific level of performance;
- the probability that this performance level causes the attainment of particular outcomes; and
- the attractiveness of those outcomes.

The alternative with the highest subjective utility for the employee will be chosen. Thus, assuming that a job elsewhere is slightly more attractive than a promotion, the employee may nevertheless opt for the internal

promotion, since they estimate the probability of getting access to courses that will increase their employability as much lower than turning the current job into a success.

Lawler (1971) has applied the last two cognitions to the importance of pay. Accordingly, its importance to an employee is a function of:

the extent to which pay is perceived as offering resources for satisfying particular needs or goals; and
the importance of these needs and goals to that employee.

In other words, the more pay is considered to be instrumental in getting more security in life (e.g., through purchasing an insurance policy), in being recognized by other people, and in gaining more independence—to mention some examples—and the more these needs are important to the individual employee, the more important pay has become. Pay is not important in itself: it may be a stronger or weaker vehicle for reaching other outcomes.

One implication bears upon the system of results-oriented pay: in order to be effective, it is necessary that an organization member perceives a clear 'instrumental' link between one or more particular performance levels and specified amounts of pay (e.g., Van Silfhout 2000).

3.2 Equity Theory

Expectancy theory is focused upon the individual human being, engaged in means-to-end relationships. Is that individual affected by one or more others, by smaller or larger social networks in the choice-making process? Yes, according to Ajzen and Fishbein's theory of reasoned action (1980), in which the individual's beliefs about what others think he or she should do are supposed to affect the choices he or she make. Equity theory (Adams 1963, 1965) includes 'social others' indirectly in the balance between contributions and inducements. Adams holds that an individual strives for such a balance *relative* to the balance that individual perceives between contributions and inducements as applying to a referent. A referent may be one or more human beings, a particular group, an organization's policy, and that individual at an earlier or later stage of their career. Contributions relate to long-term investments and short-term costs, such as education, abilities, and effort. Among inducements are task performance, recognition, and compensation. Equity is achieved when the person's perceived contributions-inducements ratio is equal to their perception of the referent's ratio. An unequal balance results in dissonance and inequity.

A typical setting would be an experiment in which the ratio between a person's effort (to perform a task) and pay is compared with the effort-pay ratio as perceived to apply to the referent. A high level of

inequity would be the result when the person perceives their own situation in terms of high effort for low pay, and the referent's ratio as low effort for high pay. Obviously, the person would then be motivated to restore an equitable balance. But how? That is difficult to predict. Available strategies are: effecting change in contributions and/or inducements; changing the referent; cognitively reinterpreting the inequitable situation; or leaving the organization. Harder (1992) showed in a study of baseball and basketball players that under-rewarded players tended to opt for utilitarian, egotistic behaviors that were thought to lead to future higher rewards.

3.3 Reflection Theory

This theory specifies the meanings of pay or compensation. A basic notion is, again, that pay has no significance in itself, but that it acquires meaning to the extent that it conveys information that is vital to an individual's self-identity. Accordingly, pay 'reflects' information about events in other fields. Four categories of meaning are distinguished:

(a) *Motivational properties*. This category is derived from expectancy theory (Sect. 3.1). Pay is meaningful to the extent that it is considered to be instrumental in satisfying a person's motives.

(b) *Relative position*. This bears upon two aspects. One reflects feedback about task performance relative to the targets set. Pay may signal whether particular corrective actions are necessary. The other aspect is social comparative: pay reflects how well a person is doing in comparison with others inside and/or outside the company.

(c) *Control*. Pay may reflect the amount of control the individual exerts on their environment (both the social network and material means) as well as the extent to which they are controlled by the environment.

(d) *Spending*. Here pay reflects the utility of goods and services purchased by the individual, and the ease or hardship with which this occurred.

Scales are available for measuring these meanings. The core of reflection theory is that the more an employee or manager reads meanings into their pay, the more their performance and satisfaction with pay is affected (Thierry 1998, 2001). Obviously, the design of a pay system and the manner in which it is administered are among the variables that influence the meanings pay may have.

4. Conclusion

An important question raised in the introductory section is whether the beliefs and expectations which seem to be basic to the four compensation components are grounded in psychological theories on pay. Just

three theories could be discussed; nonetheless, their tenets are rather representative. Expectancy and reflection theory provide some ground for results-oriented pay, in particular regarding the importance of unambiguous performance leading to pay perceptions. Yet it is not clear how these perceptions should be structured in the case of group performance; possibly equity theory's notion on contribution—inducement ratios should also be considered. The belief embedded in the concept of base pay—that stability and predictability of pay may cause employees and managers to be focused more upon the content of their work—does not take full account of the meanings of pay (put forward in reflection theory) relative to the meanings of nonmaterial work characteristics. This is clearly in need of more research. Continued research on cafeteria plans, including perquisites, is vital, in order to learn whether this compensation component meets its expectations. All in all, however, beliefs and expectations about compensation in practice are not so much missing the point as more outspoken than psychological theory would require them to be.

See also: Money and Finances, Psychology of; Money: Anthropological Aspects; Psychological Climate in the Work Setting

Bibliography

- Adams J C 1963 Toward an understanding of inequity. *Journal of Abnormal and Social Psychology* **67**: 422–36
- Adams J C 1965 Inequity in social exchange. *Advances in Experimental Social Psychology* **2**
- Ajzen I, Fishbein M 1980 *Understanding Attitudes and Predicting Social Behavior*. Prentice-Hall, Englewood Cliffs, NJ
- Barber A E, Bretz R D 2000 Compensation, attraction, and retention. In: Rynes S, Gerhart B (eds.) *Compensation in Organizations: Current Research and Practice*. Jossey-Bass, San Francisco, pp. 32–60
- Bartol K M, Locke E A 2000 Incentives and motivation. In: Rynes S, Gerhart B A (eds.) *Compensation in Organizations: Current Research and Practice*. Jossey-Bass, San Francisco, pp. 104–47
- Deci E L, Ryan R M 1985 *Intrinsic Motivation and Self-determination in Human Behavior*. Plenum Press, New York
- Deci E L, Koestner R, Ryan R M 1999 A meta-analytic review of experiments examining the effects of extrinsic rewards on intrinsic motivation. *Psychological Bulletin* **125**: 627–68
- Harder J W 1992 Play for pay: Effects of inequity in a pay-for-performance context. *Administrative Science Quarterly* **37**: 321–35
- Lawler E E 1971 *Pay and Organizational Effectiveness*. McGraw-Hill, New York
- Rynes S L, Bono J E 2000 Psychological research on determinants of pay. In: Rynes S L, Gerhart B A (eds.) *Compensation in Organizations: Current Research and Practice*. Jossey-Bass, San Francisco, pp. 3–31
- Thierry H 1998 Compensating work. In: Drenth P J D, Thierry H, De Wolff C J (eds.) *Handbook of Work and Organizational Psychology*, 2nd edn. Psychology Press, East Sussex, UK, pp. 291–319

- Thierry H 2001 The reflection theory on motivation. In: Erez M, Kleinbeck U, Thierry H (eds.) *Work Motivation in the Context of a Globalizing Economy*. L. Erlbaum Associates, Mahwah, NJ, pp. 141–58
- Thierry H in press Enhancing performance through pay and reward systems. In: Sonnentag S (ed.) *Psychological Management of Individual Performance*. John Wiley, Chichester, UK
- Thierry H, De Jong J R 1998 Job evaluation. In: Drenth P J D, Thierry H, De Wolff C J (eds.) *Handbook of Work and Organizational Psychology*, 2nd edn. Psychology Press, East Sussex, UK, Vol. 3, pp. 165–83
- Van Eerde W, Thierry H 1996 Vroom's expectancy models and work-related criteria: A meta-analysis. *Journal of Applied Psychology* **81**: 575–86
- Van Silfhout R K 2000 *Inequality in Pay Within Organizations: Normative and Instrumental Perspectives*. PhD thesis, Tilburg University, The Netherlands
- Vroom V H 1964 *Work Motivation*. John Wiley, New York

H. Thierry

Peace

The major cases of violence and peace relate to the way the human condition is cut through by fault lines, dividing humans and nature, normal people from deviants, different genders, generations, races, classes, nations, states. Each border defines at least two categories, self and other. The violence can be direct violence intended by the self to attack the basic needs of the other; or structural violence, also preventable, not intended, usually slower, but at least equally devastating. If the other is dehumanized to a category, we end up with genocide—massive category killing. With two types of violence and eight fault lines we get a 16-fold discourse, as shown in Table 1.

Most of the time and in most places there is 'narrow peace' or an absence of direct violence across most fault lines, but not 'broad peace' or an absence of direct and structural violence. Different cultures view peace differently as descriptive (how is peace possible) or prescriptive (how could peace be possible).

The following is a review of the theories of peace of six macrocultures, or civilizations. Of the six, three are from the Occident, defined here as the cultural space spanned by the Abrahamic religions of the Book, the *kitab*, or Old Testament: Judaism, Christianity, Islam. We start with two Europes, secular Europe—the Europe of antiquity—and then Christian Europe. We then move East to West Asia (I have not designated this as the 'Middle East,' which is a Western European perspective) with Judaism, Islam, and the cradle of Christianity. Continuing eastward we come to South Asia, with what is conveniently called 'Hinduism' and its offspring, Jainism and Buddhism. And further east, in East Asia there are the Chinese and Japanese cultural amalgams with a Confucian core, Buddhism, and then Taoism in China and Shinto in Japan. We

Table 1
Classification of violence

No.	Fault lines	Direct violence	Structural violence
[1]	Nature	slaughter of 'beasts, savages, wilderness'	depletion pollution
[2]	Gender	killing women: abortion, infanticide, witch-burning	patriarchy as prison of women, putting them 'in their place'
[3]	Generation	<i>Privileging</i> abortion euthanasia	<i>middle-aged</i> schools as ghetto 'homes' as ghetto
[4]	Deviance —criminal —mental —somatic	<i>Controlling dangerous persons</i> capital punishment euthanasia euthanasia	institution institution institution
[5]	Race	<i>Controlling dangerous races</i> eradication slavery	colonialism slavery
[6]	Class —military —economic —political —cultural	<i>Controlling dangerous classes</i> elimination	exploitation-body repression-mind alienation-spirit
[7]	Nation Culture Ideology	'genocide' as narrowly defined	the state as a prison of nations run by 'majority'
[8]	State Country Territory	war (killing for food, sacrifice, conquest)	imperialism isolating 'pariah' states

then move into Pacific American African spaces, picking up a Polynesian, an Amerindian, and an African tradition, ending south of Europe.

1. *Europe: Pax and Eirene—The Roman/Greek and Modern Traditions*

Peace can be interpreted narrowly as *absentia belli*, the absence of organized violence between groups defined by the fault lines. International or external peace is the absence of external wars: intercountry, interstate, or international (in the sense of intercultural). Social or internal peace is the absence of internal wars: national, racial, class, or ideological groups challenging central governance or each other. This concept is carried by the Roman *pax*, related to pact, as in *pacta sunt servanda*, 'treaties must be observed.' Peace as a contractual, mutually agreed relation is the source of Western international law.

Another Roman legacy, *si vis pacem, para bellum*, 'if you want peace prepare for war,' is the source of Western military peace theory. Peace is obtained by the balance of power, deterring the aggressor at home with defensive defense, and abroad with offensive

defense. Offensive offense (attack/aggression) and defensive offense (pre-emptive warfare) are not peace concepts.

Aggression is controlled from above by pacts—the Roman Empire was capable of enforcing such pacts—and/or by the balance of power. A better word for this kind of peace is security. And there is a basic dilemma: the military capability used to deter can also be used to attack, even if the motivation is defensive. The result can be an arms race, a cold war, or even an actual war.

A basic problem with the Roman *pax* is its insensitivity to flagrant injustice and inequality, the Roman Empire itself being an example. The Greek *eirene* is closer to 'peace with justice': absence of direct and structural violence, among Greeks. But which part has priority? What if they think justice can only be obtained through war? That idea was picked up by the Marxist tradition as class war, national and international, legitimized as necessary and sufficient to obtain a just, socialist society. The Roman thinking led to the liberal tradition, tolerant of enormous inequalities, nationally and internationally, but strongly against war, both internally and externally. And Roman warfare was picked up by the conservative tradition, extolling still more inequality through

wars, provided they could be won, and even through empire-building. In the year 2000 these views still prevail.

2. *Europe: Agape and Bellum Iustum: The Christian Traditions*

The three main divisions of Christianity (Orthodoxy/Roman Catholicism/Protestantism) and the many smaller sects had Paradise and Hell as archetypes that became models for peace and war, making peace remote and static, and war a highly dynamic hell on earth. Peace is seen as ordained by a God whose law is the only valid law, and valid for all of humanity. But who is God, and how does he relate to human beings in search of peace?

A theological distinction is very useful here, between God as immanent, inside us, making us godlike, sacred, and God as transcendent, above us, saving, choosing some persons and peoples, rejecting and condemning others. We may even talk about soft and hard Christianity, depending on which God-concept is picked up, in what proportion. They do not exclude each other.

Agape (Greek for 'love') can be used as a name for the peace of a soft Christianity based on an immanent conception of God. There is God's love for humankind, through Jesus Christ; the human love for God; and the love of one's fellow beings as being God-loved. The Lord's Supper and the Eucharist are close to this concept: a community of humans, God-enlightened, *in* God. The face-to-face *corpus mysticum* is based on an identity with others so strong that there is a limit to the number of members. This may be a reason why very egalitarian, interactive, direct peace concepts spring out of smaller Christian sects, like the Quakers and the Mennonites.

Then there is the transcendent God, conceived as some male person residing above our planet, and his relation to nonbelievers, pagans, and worse still, to the heretics who have rejected God. This is where hard Christianity enters, administering Hell and torture (the Inquisition) and Holy War to the heretics.

As some kind of peace concept it takes the form of *bellum iustum*, in the Augustine-Aquinas just war tradition:

- (a) Wars must be waged by a lawful authority.
- (b) Wars must be waged for a just cause, to correct injustice.
- (c) Wars must be waged with the right intention, not vengefully.
- (d) Wars must be waged as a last resort and with prospects of success.
- (e) Wars must be waged with proportionality and minimum force.
- (f) Wars must be waged only against combatants.

The first five are *ius ad bellum*; the last two are *ius in bello*.

This can be seen as an effort to limit war even if accepting war as the last resort. But there is no nonviolent alternative, and it may be used to attack any kind of injustice however defined. In addition, it is not biblical but derives from St. Augustine and St. Thomas Aquinas. But this cost-benefit thinking is still with us.

3. *West Asia: Shalom/Sala'am—The Judaic and Islamic Traditions*

Looking at some famous quotes from the basic texts (see Table 2, Christianity is included as a family member for comparison) one conclusion may be that ambiguity is the message. It is neither unconditional peace, nor unconditional war. It is peace under certain conditions and war under certain conditions. The problem is to spell out those conditions. One reading of the Judaic *shalom* and the Arab *sala'am* is peace with justice. Without justice, no peace; hence war for justice is legitimate. The contradiction is mirrored in the quotes shown in Table 2.

If we define justice as absence of structural violence whereby people, nations, and states are repressed, exploited, alienated, but not by an actor, then this may lead to *bellum iustum* as injustice abounds. *Jihad*, however, translates as 'exertion' for the faith. Defending the faith by violence, against Crusades, Zionism, communism, is the fourth stage of *Jihad*.

4. *South Asia: Shanti and Ahimsa—The Hindu/Jainist/Buddhist Traditions*

There is a trend in all three faiths toward unconditional peace by *ahimsa*, nonviolence. The major carrier of this message of all times, Mohandas K. Gandhi (1869–1948) used the formula 'There is no way to peace, peace is the way.' This is a strong stand, ruling out violence as immoral and unproductive. The struggle against violence, direct or structural, is by nonviolence. And, as becomes clear from Gandhi's adulation of the small social unit, the village, peace in the sense of absence of direct violence, cannot be built top-down by heavy international and national hierarchies, in the *pax* and Hobbesian traditions.

This, in turn, is related to *shanti*, inner peace. *Ahimsa* has *shanti* as a necessary condition: with no inner peace, no nonviolence. 'Unprocessed traumas' will be acted out aggressively. If nonviolence does not lead to change of heart in the other, it is for lack of change of heart in self. Nonviolence then turns into self-purification, practiced in the little community of believers, the *sangha*, which is like a monastery.

Gandhi left behind a theory and practice of *satyagraha*, clinging to truth as he called it. Look at the list of nonviolent action: to play a major role in delivering the nonwhite world from white colonialism,

Table 2
Hard and soft peace: quotations from Judaism, Christianity, and Islam

	Judaism	Christianity	Islam
Soft peace	He will decide the disputes of the nations and settle many a people's case ... till swords are beaten into plough-shares and spears into pruninghooks. No nation draws the sword against another. (Isaiah 2:4)	How blest are the peacemakers. God shall call them his sons. (Matt. 5:9)	And whoever saved a human life, should be regarded as though he had saved all mankind. (Qur'an 5:32)
Hard peace	For in the cities within the boundaries of the Promised Land you are to save no one; destroy every living thing. (Deut. 21:16)	Don't imagine that I came to bring peace to the earth. No, rather, a sword. (Matt. 10:34)	Fight for the sake of Allah those that fight against you but do not attack them first. (Qur'an 2:190)

and the white world from its own Cold War, leaving no thirst for bloody revenge behind, is no minor achievement. That gift comes out of South Asia, not the West. The contribution of the West seems to a large extent to have been to create the problems. The burden of proof is on those who teach peace by violence. To assert 'nonviolence does not work' is uninformed, given the amazing successes in the second half of the twentieth century:

(a) the liberation of arrested Jews in Berlin, February 1943

(b) Gandhi's *swaraj* campaign in India; Independence from 1947

(c) Martin Luther King Jr.'s campaign in the US South from 1956

(d) the anti-Vietnam war movement, inside and outside Vietnam

(e) the Buenos Aires Plaza de Mayo mothers against the military

(f) the 'People's Power' movement in the Philippines, 1986

(g) the Children's Power movement in South Africa, from 1986

(h) the *intifada* movement in Occupied Palestine, from 1987, in the beginning mostly nonviolent

(i) the democracy movement Beijing, spring 1989

(j) the Solidarity/DDR movements which ended the Cold War.

harmony would be not only 'absence of violence,' but 'absence of conflict.' The task of conflict resolution has already been carried out. Indications of conflict are swept under the carpet, and the person articulating conflict is frozen out of harmonious society, or prevented from expressing such views.

Take the metaphor of Chinese boxes, or in Russia the *matrushka* dolls, with one box or doll inside the other and so on, till they become extremely small. They all look alike. In modern fractal/chaos theory they talk about 'self-similarity' as something stabilizing. The basic East Asian point is that harmony is produced not by a particular structure, be that a pyramid or wheel or whatever, but by the same structure repeated at the personal, social, and world levels, within and between.

In *Tao Te Ching* we find an example: small, detached countries: 'Though they have armor and weapons nobody displays them—they leave each other in peace while they grow old and die' (No. 80). A modern version would be domestic and global democracy.

6. *Pacific, America, Africa—Ho'o ponopono, Peace Pipe, and Shir*

The following are three examples of peace as non-violent conflict resolution.

5. *East Asia: Ho p'ing/Heiwa—The Chinese/Japanese Traditions*

The standard translation of *ho p'ing* (Chinese) and *heiwa* (Japanese) is 'harmony.' And one reading of

6.2 *Ho'o ponopono—The Polynesian Conflict Resolution Circle*

Ho'o ponopono (setting straight) is a method of conflict resolution that brings together around a table per-

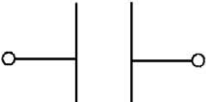
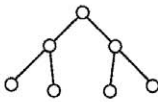
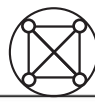
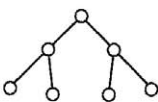
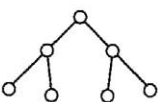
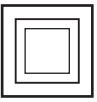
EUROPE, secular: <u>Balance of power</u> <u>Rule of Man, Law</u>		
EUROPE, Christian: Soft: <u>agape</u> Hard: <u>bellum iustum</u>		
WEST ASIA: <u>shalom, Sala'am</u> Soft: the Wheel Hard: the Pyramid		
SOUTH ASIA: <u>The sangha wheel</u> <u>Shanti/Ahimsa</u> both soft		
EAST ASIA: <u>Harmony by the Wheel</u> <u>Harmony by similarity</u> Both soft and hard		
PACIFIC-AMERICAS- AFRICA: <u>Ho'o ponopono</u> <u>peace pipe</u> <u>shir</u> Soft: the Wheel		

Figure 1

Geometric images of peace

petrator, victim, family members, neighbors, friends, and others, with a moderator, the 'wise man,' not from families or neighbors. The process of resolution has four stages.

Everyone is encouraged sincerely to present their version of why it happened, how, and what would be the appropriate reaction, and everybody accepts some part of the responsibility. Sincere apologies are then offered and accepted; forgiveness is demanded and offered. An action program is defined for everybody according to the principle of shared responsibility. And in the end the record of what happened (but not of what has to be done) is then burnt, symbolizing the end of the incident, and the construction of a new relation among all of them.

6.3 The Peace Pipe—The Cheyenne Conflict Resolution Symbol

With the Cheyenne, a zone of peace is created at the top of society, with the peace chiefs being models of

nonviolence to be emulated by the rest. They were not to engage in any quarrels within the tribe regardless of whether their families or children were involved. They were not to engage themselves in any force or violence, even if their son was killed right in front of their tepee. You are to do nothing but take your pipe and smoke. Being a chief becomes a way of life, and the chief's home a sanctuary where others can be safe. The chiefs mediate disputes and don't take sides. Ritual, like smoking the peace pipe together, makes it possible to think with one mind, one heart. Everyone has a right to talk, nobody is interrupted, the talk continues till the end.

6.4 Shir—The Somalian Conflict Resolution Market

A traditional conflict resolution structure that brings together all the mature men in the clans involved in a conflict. Women, children and young hot-blooded warriors are excluded. Men lounge under the thorn trees during the hot, dry day. They chat and drink tea.

At some point, things will jell. The various pieces that make up the main issue for which the *shir* was called will fall into place because a social climate conducive to a solution will have slowly emerged. The result will be proper peace—a peace felt from the inside.

Conflict resolution has here been hitched on to one of the oldest institutions in the history of human beings: the market. The market is based on exchange for mutual benefit: I give, you give (*do ut des*, I give so that you give), and some principle of equal exchange, (*quid pro quo*).

7. Peace as Geometry

Five geometric archetypes for peace are shown in Figure 1: the vertical, big pyramid for the hierarchic rule from above, of human beings or of law; the wheel for the smaller, tight, equal exchange; the symmetry—like the garden of Versailles—for the balance of power; the sun for the nonviolence radiating from inner peace; and the Chinese boxes of harmony based on self-similarity, from micro to macro levels.

Different kinds of peace can also be expressed in terms better known from the history of diplomacy, focusing on interstate and international peace: the Napoleonic peace forged by a prince/lawmaker; the silent peace brought about by groups of ordinary humans, or countries; the Nixon–Brezhnev peace negotiated at a bilateral summit; the Gandhian peace as process, not only outcome; the anonymous peace as built-in harmony. All of these are used for conflict resolution. How conflict is handled is the best peace indicator.

8. Approaches To Peace

Each society, community, family, and person has ways of being peaceful by peaceful means, but they also often have ways of being violent, summarized in their cultures. A major approach to peace concerns is to tilt the cultural balance in favor of peace, letting peace cultures be the dominant cultures. Then, using the survey of peace cultures and the peace as geometry approach we have five ideas in addition to peace culture.

The two Europes and West Asia have one idea in common: the hierarchy, the pyramid. Essentially this is peace from above, whether from a big power, a combination of big powers, or from some central authority. The current term is ‘global governance,’ based on global architecture. Regional versions may produce regional peace among members, excluding nonmembers. The counterpoint to the pyramid, the wheel, found in five cultural spaces, expresses the ideas of peace from below, small is beautiful, inspiring the search for alternative peace structures. Then comes

the *shanti–ahimsa* combination from South Asia: the source of peace is inside you, its expression is non-violence, which makes war abolition a realistic utopia. And we pick up the idea of conflict transformation, with East Asian harmony as outcome, *ho’o ponopono*, peace pipe, and *shir* as examples of processes.

The six approaches divide into two groups: global governance, war abolition, and conflict transformation being more familiar; nonviolence, peace structures, and peace cultures unfamiliar and challenging. Peace in the complex, highly interconnected world of the twenty-first century, bringing conflict parties very close to each other, will depend on the ability to draw upon all six—not easy given that they hail from different parts of the world.

See also: Alliances: Political; Balance of Power, History of; Balance of Power: Political; Conflict Sociology; Diplomacy; International Relations, History of; National Security Studies and War Potential of Nations; Peace and Nonviolence: Anthropological Aspects; Peace Movements; Peace Movements, History of; Religion: Peace, War, and Violence; Violence, History of; Violence: Public; War, Sociology of; Warfare in History

Bibliography

- Bouquet A C, Murty K S 1960 *Studies in the Problems of Peace*. Asia Publishing House, Bombay, India
- Ehrenreich B 1997 *Blood Rites: Origins and History of the Passions of War*. Metropolitan Books, New York
- Galtung J 1980 Social cosmology and the concept of peace. In: Galtung J (ed.) *Essays in Peace Research*. Ejlers, Copenhagen, Vol. 5, pp. 415–36
- Galtung J 1996 *Peace By Peaceful Means*. Sage, London
- Galtung J 1998 *Conflict Transformation by Peaceful Means*. United Nations, Geneva
- Giorgi P P 1999 *The Origins of Violence by Cultural Evolution*. Minerva E&S, Brisbane, Qld, Australia
- Howell S, Willis R (eds.) 1989 *Societies at Peace*. Routledge, London
- Ishida T 1969 Beyond the traditional concepts of peace in different cultures. *Journal of Peace Research* 6: 133–45
- Keeley L H 1996 *War Before Civilization*. Oxford University Press, Oxford
- Prunier G 1998 Somaliland goes it alone. *Current History* May: 225–8
- Shook E V 1985 *Ho’o ponopono*. East-West Center, Honolulu
- Smith-Christopher D L 1998 Indigenous traditions of peace: An interview with Lawrence Hart, Cheyenne Peace Chief. In: Smith-Christopher D (ed.) *Subverting Hatred: The Challenge of Nonviolence in Religious Traditions*. Boston Research Centre for the 21st Century, Cambridge, MA pp. 85–94
- Thompson W S, Jensen K M (eds.) 1991 *Approaches to Peace: An Intellectual Map*. United States Institute of Peace, Washington DC

J. Galtung

Copyright © 2001 Elsevier Science Ltd.
All rights reserved.

11139

Peace and Nonviolence: Anthropological Aspects

1. Definitions

The phrase 'peace studies' refers generically to studies of the dynamics of 'peace,' and specifically to the natural history of 'peaceable' or 'nonviolent' peoples. The two main foci, which overlap significantly, are (a) the analysis of particular conflict-resolution techniques, and (b) the holistic study of peaceable peoples.

1.1 Problems of Definition

The terms 'peace' and 'nonviolence' are relative, not absolute, referring to adaptations which are often transitory rather than to essential natures which persist indefinitely. Although researchers use the words 'nonviolence' and 'peace' as rough synonyms, the former creates methodological problems. How to observe and analyse the absence of violence? It might be useful to reserve this term for ideals such as Gandhi's *ahimsa* or related techniques such as the nonviolent protest of the 1970s 'Peace Movement' of Europe and America, from which peace studies arose and borrowed the term.

Likewise, in most Indo-European languages 'peace' is a null category. When you have no violence, you have peace. Linguistically, violence is the norm. Peaceable peoples are 'nonviolent' (Ashley Montagu) or 'low-conflict' (Marc Ross). It is difficult for students of peace to avoid drifting into the null construction of peace which pervades the cultures within which they live their lives.

By contrast, Semai, a peaceable Malaysian people, use the term *slamad* to refer to a condition of security and serenity which they nurture with the same zeal that Indo-Europeans pursue wealth and happiness. In Semai construals, *slamad* is normal, 'non-*slamad*' the null category. Recent peace studies stress the distinction between 'negative peace' (refraining from violence out of fear) and 'positive peace' (maintaining *slamad* out of love of peace). In this construction, both forms of peace are activities rather than absences of activity, involving conscious and unconscious 'peace-making' rather than voids. Defeat involves an agent's act of surrender as well as another agent's act of conquest. 'Peacemaking' is thus the category of behaviors and social patterns on which peace studies focus.

2. Why Not Study Peace?

2.1 The Ubiquity of Peace

Any survey of the literature will uncover hundreds of articles on violence for every one on peace. Yet the

most violent people spend most of their time in humdrum peaceful activities: Eating, sleeping, talking, daydreaming, scratching, and otherwise doing no harm. Most people in the most violent societies, even in wartime, spend their days in peaceful pursuits (Nordstrom 1997). The life of a soldier in wartime involves more hours of boredom than moments of terror.

No wonder. Violence hurts, and individuals dislike getting hurt. The possibility of counter-violence makes violence self-limiting in a way that peace is not. People prefer hurting those who, for whatever reason, cannot retaliate. Domestic violence is probably the commonest form of violence cross-culturally, simply because perpetrators are relatively safe from retaliation (Gelles and Straus 1988). Even then, when local values justify violence, the way human perpetrators obfuscate their actions by appealing to peace ideology reveals their uneasiness with violence: they are 'maintaining law-and-order,' 'keeping the peace,' 'disciplining students,' or 'teaching them a lesson.'

In evolutionary terms, that makes sense. Other things being equal, organisms that avoid serious fighting with peers are more likely to survive long enough to produce fertile offspring than those that usually win fights but risk counter-violence; i.e., peaceful organisms are fitter. Moreover, violence disrupts cooperative activities that enhance the fitness of social organisms like humans. Such disruption is a common theme in nonWestern peoples' reconciliation ceremonies: if we fight, who will help to harvest the crops? Fear of disruptive violence rationalizes human rulers' stamping out or severely limiting freelance violence, and monopolizing violence for themselves and their state. As a result, throughout human history peace prevails, in most places, most of the time. Peace is normal and normative. Violence is abnormal and disruptive.

2.2 Reasons Not to Study Peace

So why does peace get so little attention, even from social scientists? One reason is the definition of peace as an absence, discussed above. There seem to be several others. One probable reason is evolutionary: Violence threatens fitness. Successful organisms attend to threats. Violence is a problem; peace is unproblematic. So peace remains in the background, and violence is dramatic. Unsurprisingly, the most popular account of paleoanthropic violence is by a dramatist, not a paleoanthropologist, primatologist or prehistorian (Ardrey 1961). The very prevalence of peace makes violence salient: The exception, 'man bites dog,' makes news. Normal conditions do not.

Answers to the question of how to live in peace have potential political consequences, which might affect the lives not only of this generation but of generations to come. In the 1990s, over 2 million children died in wars, 6 million were seriously injured or permanently

disabled, a million orphaned, and 12 million made homeless. Children are the main victims of enslavement and domestic violence.

Awareness that research may be consequential can make it teleological, arriving at conclusions which reflect researchers' political predilections. Hobbes, an early student of violence and peace, deliberately excluded empirical considerations from his work, relying instead on an a priori method which he felt derived from Euclidian geometry, and concluding that the political despotism of his time was a counter-violence needed to control the freelance violence which would otherwise ruin human life. Similarly, peace studies evolved from New Left political activism in the 1970s. A large number of peace theorists are pacifists or members of quietist religious groups like Quakers or Mennonites. Many are explicitly political, arguing that, since they feel peace requires respect for human rights and social justice, anthropology should espouse such causes. Pedagogy is a central concern.

The resulting atmosphere may alienate students who hold traditional social science notions of 'objectivity.' Moreover, this sort of partisanship generates equally political responses by students of violence and others for whom, as Max Weber warned might happen, social science is the handmaiden of the Establishment and the social controls that maintain it in power. Social justice, say the reformers, will produce (positive) peace. No, say the establishmentarians, sometimes marching under the banner of objective social science, you need the hangman and the Bomb (negative peace).

Such politicization makes researching these topics difficult: For example, anthropologists have been reluctant to undertake (and funding agencies to finance) studies on domestic abuse among indigenous peoples, at least partly because the hegemonic liberal ideology frames indigenous peoples as simple and peaceful, and representing them otherwise might exacerbate the difficulties they already face in finding allies among more powerful peoples. In rebuttal, opponents seem to be reviving the nineteenth-century hegemonic representation of indigenous peoples as savage and violent, an equally political representation (e.g., Keeley 1996).

It may be impossible and even undesirable to eliminate political concerns from research that is so 'sensitive.' It would be helpful to make them explicit, as students of peace traditionally do. But scholars may be unaware of the teleological concerns that bias their results. Even researchers, after all, swim in their own culture like fish, mostly unaware of how the medium that supports them also affects how they perceive the world.

A cognate difficulty is that discussions of peaceability often degenerate into speculations about 'human nature,' a Platonic essentialist concept which serves more to terminate research than to stimulate it (but cf. Boehm 2000). If people are 'naturally' violent,

violence needs no explanation; if 'naturally' peaceable, peace requires no further examination. Recent studies such as Keeley's (1996) suggest that violence has been part of human societies since early in prehistory. There may be no society in which violence has never occurred. By the same token, there is no society in which peace never occurs, in which the general rule is a Hobbesian war of each against all; and no period in prehistory or history at which no one lived in peace. One of the most interesting studies of war describes how, in the midst of all-pervasive violence, people continue doggedly to construct peace (Nordstrom 1997). The half-empty glass is half-full (Boehm 2000), and the decision on the question of human nature is more likely to reflect the political sensibilities of the researcher than the empirical data.

3. Styles in the Study of Peace

The two main foci of peace studies reflect the dichotomy between positive and negative notions of peace. Although the sketch given in the previous section of 'establishmentarians' versus 'reformers' is a heuristic oversimplification of impulses which overlap and may coexist within the same person, nevertheless it is not completely inaccurate to see the establishmentarian impulse in the first focus and the reformist in the second.

3.1 Conflict Resolution Studies

The first subfield is, at least implicitly, instrumental or 'applied' anthropology, reflecting governmental interests in international diplomacy and administrative measures to reduce freelance violence, e.g., in the school system. It comprises 'conflict-resolution' or 'mediation' studies of the sort emphasized by the journal *Peace and Conflict Studies*—www.gmu.edu/academic/pcs. In the United States, George Mason University is a center for such studies. Since World War II, conflict resolution studies in anthropology have grown in tandem with their growth in psychology, sociology, and political science. This upsurge seems to be connected with the rise of UN 'peace-keeping' efforts and interventions by the USA and Western Europe in overseas conflicts. More recently, the perceived rise of violence in the USA has led to funding of 'violence prevention programs,' for example, by the Justice Department. Most such programs stress peaceful conflict resolution, sometimes purportedly modeled on 'primitive customs.'

Most conflict resolution theory is a subset of game theory models, in which participants try to maximize benefits to themselves, and minimize their losses. The ideal solution is 'win-win,' in which for each participant the sum of the benefits outweighs the sum of the losses. Ideally, participants arrive at a 'rational'

resolution through reasonable (ideally numerate) discussion, perhaps with the help of mediators from outside, trusted by both parties, experienced in mediation, and dedicated to resolving conflicts peacefully. The 'anthropological' character of conflict resolution studies comes mostly from the attempt, in the face of skepticism from other disciplines, to deploy the concept of culture, and from the occasional choice of 'tribal' peoples as subjects. In general, however, the concerns and analytical techniques (e.g., cost-benefit analysis) are indistinguishable from those of the other social sciences.

Of course, as 'conflict theorists' have insisted since Georg Simmel, conflict is ubiquitous and need not produce violence. Conversely, conflict resolution is neither necessary nor sufficient to produce peace (Fry and Bjorkqvist 1997). Many nonWestern peoples traditionally held interminable meetings explicitly to resolve conflict. During these meetings, everyone got to speak at length until participants were physically and emotionally exhausted, and no one had any unvoiced opinions. The conflict might remain, but angry emotions dissipated and the conferees could reconcile or agree to differ.

Another practical limitation is that most conflict resolution techniques require that participants be equals or equally subject to coercion by a power greater than they are. But, as cost-benefit analysts of domestic violence point out, violence is most likely to occur when unpleasant consequences for the perpetrator are negligible, i.e., when the parties are unequal (Gelles and Straus 1988). Thucydides records a conflict resolution conference between Athenians and Melians. Since the more numerous and better-armed Athenians could (and later did) overwhelm the Melians, the Melian conference 'failed.'

The corollary is that egalitarian societies tend to be peaceable, and peaceable ones tend to be egalitarian (Fry 1999); between nations, 'mutual deterrence' depends on perceived equality. Perhaps social justice in the form of felt equality is also important. Thus, although conflict resolution theorists tend to dismiss other tactics for preventing violence as unreliable (Fry and Bjorkqvist 1997), conflict resolution techniques are most likely to succeed when least likely to be necessary.

3.2 *Peace Studies*

Peace studies are more ethnographic, less technical and more value-laden, finding expression in, for example, *Human Peace and Human Rights*, journal of the Commission on the Study of Peace and Human Rights of the International Union of Anthropological and Ethnological Sciences—rar@syr.edu. Syracuse University is a center for such studies as well as for conflict resolution studies in the USA. This research rests on the empirical observation that conflicts among

some aggregations of people are less likely to erupt into violence than among others. One reason for such peace is a value system that promotes negative peace by abhorring violence (e.g., as stupid, scary, or self-destructive) and promotes positive peace or *slamad*, a value system somewhat like that of the 1970s 'Peace Movement' which inspired peace studies.

But that movement itself grew from the Vietnam War. Peace tends to be valued particularly by people who know violence firsthand. Peaceable values may maintain peace, in the short run, but they gain strength from the threat of violence (DeBenedetti 1978, Dentan in Silverberg and Gray 1992, in Sponsel and Gregor 1994, Nordstrom 1997). And one reason for war is that outsiders threaten *slamad*. In short, attributing peace only to peaceful values is as simplistic as attributing violence to 'innate aggression.' The stipulated variables are insufficient to produce the observed results.

The ethnology of peaceable peoples reveals how complicated is 'peace.' Documentation exists on a fairly large number of relatively peaceable societies, including traditional ethnic groups (e.g., Dentan in Silverberg and Gray 1992, Fry 1999, Howell and Willis 1989), cenobitic pacifists like the Amish or Hutterites, and other voluntary associations such as the Friends, the Rainbow Family, or Alcoholics Anonymous (AA) (1994). Most have formal and informal conflict-resolution techniques, the latter derived from those characteristic of the species as a whole (Boehm 2000, de Waal 1989). Most lack the in-group economic or political hierarchies which impede consensual conflict-resolution, although many are enclaves in hierarchical societies against whom they must maintain identity barriers.

Peaceable values are conducive to peace. Many peaceable peoples like peace better and violence less than more violent peoples do. But many, for example, the Amish and Mbuti, beat children severely. Others, such as the Semai, fantasize warfare against their oppressors, or boast, like the AA, that members can stand up to the challenges of warfare when their country calls. Generally, in-group peace seems more valued than peace with outsiders. And, as the ubiquity of peace may indicate, most people prefer peace to violence except under exceptional circumstances.

'Negative peace' based on fear of consequences seems more salient than 'positive peace' based on idealist values. Many peaceable peoples, particularly cenobitic pacifists, are patriarchal and physically punish disobedience by children. Oppressive regimes can also diminish freelance violence by retaliatory violence such as incarceration, mutilation, or capital punishment.

Avoiding others is yet another way of peacemaking. Cross-culturally, inevitable structural conflicts, such as that between a man and his in-laws over who benefits from his wife's services, may result in formalized respect/avoidance relationships in which one or

the other party expresses extreme deference to, or even avoids social contact with, the other. Avoiding a person with whom one is angry is an effective passive-aggressive communication of the anger and dramatizes the threat of withdrawing cooperation. Many peaceable societies of Southeast Asia arose in the context of a political economy based on slavery and coerced trade, in which the best way to avoid enslavement was to flee. Catholics and Protestants in Northern Ireland, like 'blacks' and 'whites' in the USA, tried less successfully to minimize violence by minimizing mutual contact.

Another way of dealing with inherent conflict is to transform physical violence into 'harmless' symbolic violence. To some extent, sports in the Western world do provide the 'moral equivalent of war' that William James imagined. The 'joking relationships' that occur in many societies allow people to express anxiety and frustration about possible conflicts, especially involving sex, by hostile words and deeds which local ideology defines as 'not serious.' This technique is unreliable: fans battle each other and 'locker room humor' can erupt into violence.

3.3 The Future of Peace Studies

What is needed is meticulous field studies of peoples who seem better at peacemaking than others. Such studies require avoiding essentialist notions of peaceability as a psychosocial deficiency, the inability to be violent under any circumstances. Rethinking definitions of peace and violence seems appropriate (e.g., Dentan 2000). If peaceability is an adaptation rather than an essence or changeless value, then changing circumstances should affect peaceability, so that such studies must take account of history in order to discover what variables are important. Perhaps the most promising future studies are those which make detailed comparisons between (a) peacemaking among specific human groups and among specific other primates (e.g., Cords and Killen 1998), or (b) 'controlled comparisons' between particular peaceable peoples and otherwise similar but more violent peoples, e.g. the studies of peaceable and violent Zapotec towns by O'Neill and Fry (see Fry 1999); the Robarcheks' planned book-length expansion of their chapter on peaceable and violent tropical rainforest swiddeners, the Semai and Waorani (Robarchek and Robarchek 1992); the comparison of peaceable and violent Chicago neighborhoods by Sampson et al. (1997).

3.4 The Use of Peace Studies

Improving conflict resolution techniques is of obvious importance. Are any lessons from peaceable societies of possible use to other peoples? The great student of

primate conflict resolution, Frans de Waal dismisses peaceable human groups as a 'few gentle, nonmartial human societies that have managed to survive in remote corners of the world' (1989, p. 4). But many cruel martial human societies have also perished, because their cruelty and violence was as unsustainable as is peaceability as de Waal implies. The British poet Shelley's poem *Ozymandias* says it best: There is an immense broken statue in the poem, on whose base are the words: 'My name is Ozymandias, King of Kings. Look on my works, ye mighty, and despair!' All around the statue, as far as the eye can see, the desert lies unbroken. Scientists must not let the drama of violent conquest obscure the fact that organisms and societies often survive better by eschewing, preventing, or avoiding violent conflict than by engaging in it. As a long-term adaptation, neither peaceability nor violence is necessarily or essentially 'good' for survival; in the short term, either may work, depending on the circumstances. That is why peaceable people like the Semai may stop being peaceable, and violent people like the Waorani may turn to peace. Both fight and flight are active adaptations, which couple with caring for children and forming alliances ('tend and befriend') as human responses to environmental stress. The dynamics of these adaptations need the same careful study as any other ecological adaptation.

See also: Domestic Violence: Sociological Perspectives; Peace; Peace Movements; Peacemaking in History; Violence, History of; Violence in Anthropology; Violence: Public; War: Causes and Patterns; War, Sociology of

Bibliography

- Ardrey R 1961 *African Genesis: A Personal Investigation into the Animal Origins and Nature of Man*. Dell, New York
- Boehm C 2000 Conflict and the evolution of social control. *Journal of Consciousness Studies* 7: 79–101, 149–183
- Cords M, Killen M 1998 Conflict resolution in human and nonhuman primates. In: Langer J, Killen M (eds.) *Play, Evolution and Development*. Erlbaum, Mahwah, NJ
- DeBenedetti C 1978 *Origins of the Modern American Peace Movement, 1915–1929*. KTO Press, Millwood, NY
- Dentan R K 2000 Ceremonies of innocence and the lineaments of unsatisfied desire. *Bijdragen tot de Taal-, Land- en Volkenkunde* 156: 193–232
- de Waal F 1989 *Peacemaking Among Primates*. Harvard University Press, Cambridge, MA
- Fry D P 1999 Peaceful societies. In: Kurtz L R (ed.) *Encyclopedia of Violence, Peace, and Conflict*. Academic Press, San Diego, CA
- Fry D P, Bjorkqvist K (eds.) 1997 *Cultural Variation in Conflict Resolution: Exploring Alternatives to Violence*. Erlbaum, Mahwah, NJ
- Gelles R, Straus M 1988 *Intimate Violence*. Simon & Schuster, New York
- Howell S, Willis R (eds.) 1989 *Societies at Peace: Anthropological Perspectives*. Routledge & Kegan Paul, London

- Keeley L H 1996 *War Before Civilization: The Myth of the Peaceful Savage*. Oxford University Press, New York
- Montagu A (ed.) 1978 *Learning Nonaggression: The Experience of Non-literate Societies*. Oxford University Press, London
- Nordstrom C 1997 *A Different Kind of War Story*. University of Pennsylvania Press, Philadelphia, PA
- Robarchek C A, Robarchek C J 1992 Cultures of war and peace. In: Silverberg J, Gray J P (eds.) *Aggression and Peacefulness in Humans and Other Primates*. Oxford University Press, New York
- Sampson R J, Raudenbush S W, Earls F 1997 Neighborhoods and violent crime: A multilevel study of collective efficacy. *Science* **227**: 918–24
- Silverberg J, Gray J P (eds.) 1992 *Aggression and Peacefulness in Humans and Other Primates*. Oxford University Press, New York
- Sponsel L E, Gregor T (eds.) 1994 *The Anthropology of Peace and Nonviolence*. Lynne Rienner, Boulder, CO

R. K. Dentan

Peace Movements

‘Peace Movements’ may be used in two ways. On the one hand, a peace movement is a specific coalition of peace organizations that, together with elements of the public, seek to remove a threat of war or to create institutions and cultures that obviate recourse to violence. On the other, it is the organizational infrastructure to do so. Usage is usually clarified by context. Peace organization constituencies are people with shared commitment to common values and traditions, like religious pacifism, or to a program such as world federalism. Such groups form coalitions in order to enlist public support in response to salient issues. If the issue is war or a specific war threat, peace coalitions take the form of antiwar movements. In nearly 200 years of organized peace effort, specific peace movements have affected national policies, international institutions, and popular attitudes. Taken as a whole, they can be viewed as a single, evolving, and increasingly transnational social movement that has interacted with formal analyses of war, peace, and social movements. The conceptualization of peace movements has resulted from the dialectical interaction of the movement’s self-reflection on its experience and subsequent scholarly analysis.

1. Growth and Self-definition, 1815–1939

In western civilization peace has been understood mainly as the absence of war (Chatfield and Ilukhina 1994). Given war’s existence, there were two main ethical alternatives: a dominant ‘just war’ tradition

that legitimated organized violence under specific conditions, and a minority tradition of individual Christian nonresistance that rejected violence altogether. There was no sustained tradition of organized popular effort to check interstate warfare.

That was the situation when small peace societies were formed in the United States and the United Kingdom (1815) by a few local nonconformist Protestants who denounced warfare. In the next 100 years civic associations for societal change became common in western societies, providing the context in which peace organizations grew and defined themselves.

1.1 Early Internationalism

Peace constituencies broadened within an educated elite. British and American leaders solicited support on the Continent, where peace advocates tended to be secular intellectuals and where Richard Cobden’s program of peace through free trade became very influential (Cooper 1991). During the 1840s leaders on both sides of the Atlantic promoted international arbitration. A few went further, a congress of nations, an international court, and even a united Europe. Then, in mid-century, peace advocates were shaken by wars in the Crimea, Germany, Italy, and America.

Gradually the movement was rebuilt in the latter third of the century. American and European leaders promoted treaties of arbitration, forming a program-oriented peace movement propelled by a bias for *internationalism*—the conviction that warfare is irredeemable and that statesmen could render it obsolete by breaking down barriers, building up international law, and seeking practical mutual interests beyond conflicts. The American Peace Society (1928) and the *Ligue internationale et permanente de la paix* (Paris 1863) popularized this approach. The *Ligue internationale de la paix et de la liberté* (Geneva, 1867), however, pressed the view that peace could be secured only by political justice among self-determining, democratic peoples. Concurrently, the First and Second (socialist) Internationals held that peace was contingent upon economic justice—the overthrow of capitalism. In practice, though, as socialists entered the political mainstream they supported programs of arbitration, arms limitation, and anticolonialism—peace as liberal internationalism, which was christened ‘*pacifism*’ (Cooper 1991, p. 237, n. 1).

1.2 An Internationalist Peace Establishment

The movement for liberal internationalism obtained strong leverage from the First Hague Conference in 1899, when the great powers endorsed pacific alternatives to war. By then approximately 100 national and

regional societies were linked by an office in Berne, Switzerland (International Peace Bureau, 1891). Within another decade there emerged (mainly in America) several well-funded associations, notably the Carnegie Endowment for International Peace. They constituted a kind of *peace establishment*, an elite of business, legal, and governmental leaders who promoted the arbitration process. In World War I most of these internationalists proved to be patriots first. The word 'pacifism' became narrowed, especially in Anglo-American usage, to mean the absolute, individual rejection of war. Wartime peace advocates were repressed.

1.3 Peace and Internationalism Redefined

The war that shaped the twentieth-century world reshaped peace movements, which emerged more transnational and politically activist than before.

A few internationalists concluded that peace required some constraints on national sovereignty, that war is irrational, and peace requires 'a systematic effort to institutionalize managerial controls over the fragile interdependence of modern industrial civilization' (DeBenedetti 1978, p. 8). Such *liberal internationalism* was institutionalized in the League of Nations and its efforts for arms control, but also in an international network of civic League of Nations Associations. From wartime pacifism there emerged other transnational associations: the Women's International League for Peace and Freedom (1919), the pacifist International Fellowship of Reconciliation (1919), the linked British and US Quaker service committees (by 1917), and the International War Resisters League (1921).

At the same time, a new constituency of political reformers carried their campaign against corporate evil and systemic injustice into the peace movement. Women fresh from suffrage campaigns, for example, revitalized earlier analyses of war as a problem of social structure. Especially in England and America peace movements acquired a constituency of middle-class women and men who linked peace and justice and for whom peace movements were social change agents.

Following the war these activists informed and mobilized public opinion, engaging in the politics of pressure groups to challenge military spending and influence foreign policy. Peace advocates were vocal also in Germany and France, where they had been suppressed in wartime, and socialists throughout Europe took up the peace cause. Unfortunately, differences over priorities and policy hobbled peace efforts throughout the 1920s. The American movement was a sad case in point: groups promoting the League of Nations, the outlawry of war, disarmament, peace education, and international law failed to

coalesce. Early in the next decade peace advocates organized internationally to support the League's 1932 disarmament conference, but in the face of aggression they soon realigned as rival supporters or opponents of collective security. No cohesive peace movement could be discerned among the bitter contests waged for public opinion.

2. External Assessment: Early Histories

The self-reflection of peace advocates on their experience affected the formal studies that began in the early 1930s with a few general histories. British scholar A. C. F. Beales, for example, interpreted *The History of Peace* (1931) optimistically in light of alternative programs broached when he wrote. A half decade later, Merle Curti's *Peace or War: The American Struggle 1636–1936* (1936) more soberly emphasized the gap between pacific ideals and the realities on which warfare is grounded (notably economic factors). Pacifist Devere Allen's earlier *The Fight for Peace* (1930) interpreted the historic American movement against his own norm of nonviolent activism and war resistance.

3. Nonviolence

Allen's book itself was part of the historic development of secular pacifism into an ideology of proactive, nonviolent change. That shift made the early *Satyagraha* campaigns of Mohandas Gandhi in India important to western contemporaries, and later to American civil rights leaders experimenting with non-violent direct action and civil disobedience. Early sociological interpretations of *Satyagraha* were absorbed into the pacifist literature on nonviolence by the 1930s. Conversely, later experiments with the method in civil rights and antiwar campaigns expanded scholarly analyses of it. Thus nonviolent direct action developed dialectically along practical and theoretical lines (see Powers and Vogeley 1997).

Moreover, the use of nonviolence in the struggle for justice applied the concept of peace as change without violence. Understanding peace as nonviolent social change for justice, rather than only as the absence of war, underlies the now common analytical distinction between positive and negative peace.

4. Cold War Mobilizations

During World War II, a fresh peace coalition formed around the United Nations (UN) ideal. It was grounded in liberal internationalism and it generated public support for a UN organization. Movement-

related nongovernmental organizations (NGOs) were responsible for including human rights under UN purview and for gaining consultative status in the organization. Even world federalism elicited much public support until it, along with initial efforts to contain nuclear arms, was undercut by the Cold War. By 1950 peace and liberal internationalism were victims of global polarization. The older peace organizations remained intact, but they lacked a salient issue around which to build any coalition.

4.1 *The Nuclear Test Ban Movement*

In the mid-1950s the issue of testing thermonuclear weapons on a massive scale aroused a fresh, transnational peace movement. It was grassroots based and was well organized on national levels. It made use of print and film media, mass demonstrations like the annual Aldermaston marches in England, political lobbying, and nonviolent civil disobedience such as sailing into test zones. The movement contributed significantly to a temporary moratorium on testing and to the 1963 Partial Test Ban Treaty.

4.2 *The Antiwar Movement of the Vietnam War Era*

Then the global movement lost coherence. The nuclear threat seemed to diminish and activists faced other issues, most dramatically the US war in Vietnam, which provoked worldwide opposition. The center of the storm was in the United States, where political radicals, pacifists, and liberal internationalists aligned in a loose, tenuous coalition. The whole repertoire of activism was employed: popular education, political pressure and lobbying, electoral politics, and street demonstrations. Although identified with radical counterculture in public opinion, opposition to the war increasingly enlisted mainstream activists and infused congressional politics. Extrication from Vietnam coincided with an exhausted peace movement and apparent détente in the Cold War.

Détente collapsed at the end of the 1970s. In response to a heightened nuclear arms threat, a new peace coalition arose that was transnational, massive, grassroots based, and aligned with environmentalists (most fully the Greens in Germany). Internationally, movement cooperation focused on UN disarmament negotiations, but it was ineffective owing to rival political agendas. Nonetheless, the movement enlisted broad public support, which it focused on the threat of nuclear war, and it reinforced arms control negotiations. It also broadened the political base in several countries and stimulated movement contacts with counterparts in the Soviet bloc.

Meanwhile, an informal coalition of peace and human rights activists in Central and North America effectively stymied overt US intervention on behalf of repressive governments there. An international cam-

paign was mounted against South African apartheid. A small coalition called the Neptune Group helped to broker UN negotiations that led to the 1982 Law of the Sea Treaty. Movement related human rights activists in North and South America parlayed Perez Equeval's 1980 Nobel Peace Prize into leverage for victims of oppression. Throughout the century the Peace Prize Committee had broadened both its understanding of peace and the regions from which winners came; in this it reflected the peace movement itself.

The Cold War collapse in 1989 left peace advocates to redefine their movement. On the centenary of the First Hague Conference, about 4,000 delegates representing NGOs around the world assembled in The Hague to help set a UN agenda for a decade devoted to achieving a culture of peace. They explicitly defined themselves as a transnational peace movement consisting of a global network of NGOs that together reflect a world of growing, interactive social movements.

5. *Scholarly Analysis: History and Sociology*

Scholarly conceptualization of peace movements has tended toward a framework of analysis parallel to the historical phenomenon.

5.1 *Peace Movement History*

In the 1970s historical studies of peace movements flowered into a comprehensive field that expanded for two decades more. In some respects it paralleled growth of resource mobilization theory in the sociology of social movements.

Some historians revisited periods covered earlier and researched them more thoroughly. Others updated the story of peace advocacy. New patterns emerged. Given the massive work of Peter Brock, for instance, religious pacifists could be distinguished from and related to peace activists. Continuities emerged, such the growth of international organization from 1914 to 1945, the process of conceptually linking peace to justice and social change (positive peace), and the links between nonviolent activism against war and for human rights. Drawing on early sociologist Ferdinand Tönnies, 'polity' peace advocates (who stressed institutional arrangements for the sake of order) were distinguished from 'community' ones (who valued popular involvement and the well being of diverse peoples). Charles DeBenedetti incorporated much of the new work in *The Peace Reform in American History* (1980). The title was significant: treating peace as a social reform invited attention to the relationships between organizations, ideologies, and social movements.

Much has been done in the subsequent two decades, both to correct imbalances in the accounts of national movements and to connect them. Several historians rescued the vital roles of women in the movement from

the neglect of male authors, for example (see Alonso 1993). Peace historians formed a transnational network of scholars from Europe, America, Canada, Australia, and Japan, working through the *Arbeitskreis Historische Friedensforschung* (German Working Group on Historical Peace Research) and the Peace History Society. Europeans revised Beales, while an American analyzed the nineteenth century European movement in terms of relationships between patriotism and internationalism and among varying class and political groupings (Cooper 1991). Together they studied the relationship of peace movements and political cultures. US and Russian historians produced a joint history of the concept of peace in western history (Chatfield and Ilukhina 1994).

5.2 The Sociology of Social Movements

While historians were interpreting peace advocacy as a diverse reform movement, sociologists were analyzing social movements with resource mobilization theory: what are the terms on which social movements strengthen and use their resources? Basic works in the field date from the 1970s (see *Social Movements: Resource Mobilization Theory*).

The fully developed method was applied to peace movements in the 1990s, especially to the Freeze campaign of the previous decade but also in a comparative analysis of three US peace coalitions. Most recently, modified resource mobilization analysis has been applied to the dynamics of transnational NGOs and social movements (see Smith et al. 1997). One study demonstrates the inter-relationships of developing nongovernmental and intergovernmental organizations *before* 1945 (Chatfield, Chap. 2 in Smith et al. 1997), and Wittner reifies the transnational social movement model with a longitudinal study in the period *since* then (Wittner 1993, 1997).

Besides filling out the history of specific peace movements, therefore, the future of this field is likely to involve narrative and analysis of more transnational connections among peace and related movements. Future scholarship may well explore specific peace and antiwar movements as dynamic elements in the creation of transnational infrastructures in an increasingly integrated world.

See also: Civil Rights Movement, The; Conflict and Conflict Resolution, Social Psychology of; Conflict Sociology; Peace; Peace and Nonviolence: Anthropological Aspects; Peacemaking in History; Social Change: Types; Social Justice; Social Movements: Resource Mobilization Theory; Social Movements, Sociology of; Theory: Sociological; Voluntary Associations, Sociology of; World Systems Theory; Youth Movements

Bibliography

- Alonso H H 1993 *Peace as a Women's Issue: A History of the US Movement for World Peace and Women's Rights*. Syracuse University Press, Syracuse, NY
- Carter A 1992 *Peace Movements: International Protest and World Politics Since 1945*. Longman, London
- Chatfield C 1992 *The American Peace Movement: Ideals and Activism*. G K Hall, Boston
- Chatfield C, Ilukhina R (eds.) 1994 *Peace/Mire: An Anthology of Historic Alternatives to War*. Syracuse University Press, Syracuse, NY
- Cooper S 1991 *Patriotic Pacifism: Waging War on War in Europe, 1815–1914*. Oxford University Press, New York
- DeBenedetti C 1978 *Origins of the Modern American Peace Movement, 1915–1929*. KTO, Millwood, NY
- Howlett C 1991 *The American Peace Movement: References and Resources*. G K Hal, Boston
- Josephson H (ed.) 1985 *Biographical Dictionary of Modern Peace Leaders*. Greenwood, Westport, CT
- Kuehl W (ed.) 1983 *Biographical Dictionary of Internationalists*. Greenwood, Westport, CT
- Peace and Change: A Journal of Peace Research*. Quarterly. Blackwell, Boston
- Powers R S, Vogeles W B 1997 *Protest Power, and Change: An Encyclopedia of Nonviolent Action from ACT-UP to Women's Suffrage*. Garland, New York
- Smith J, Chatfield C, Pagnucco R 1997 *Transnational Social Movements and Global Politics: Solidarity Beyond the State*. Syracuse University Press, Syracuse, NY
- Wittner L S 1993, 1997 *The Struggle Against the Bomb: Vol. 1, One World or None: A History of the World Disarmament Movement, 1954–1970; Vol. 2, Resisting the Bomb: A History of the World Disarmament Movement, 1954–1970*. Stanford University Press, Stanford, CA

C. Chatfield

Peace Movements, History of

Peace movements are social and political movements appearing since the early nineteenth century as a result of a collective effort to organize pacifism. They can be long-term phenomena aimed at a lasting peace. Alternatively, they can be stimulated by a concrete situation ready to vanish the moment the movement has reached its objective or pass into a latency period to be reactivated only if need be.

1. Peace Movement and Pacifism

Today both terms are frequently used as synonyms. This is even more justified as continental European peace activists adopted the term 'pacifism' the moment it was coined and widely picked up by the sociopolitical language. It proved to be useful as it suggests

a triple meaning i.e., 'movement,' 'theory,' and 'radicalism.' Anglo-American usage often makes a distinction between 'internationalism' and 'pacifism' with 'internationalism' comprising the moderate demands of peace movements in their entirety whereas 'pacifism' means the idea of the greatest possible radicalism in the practice of peace including absolute refusal to serve in the military. Based on British conditions Ceadel offers an elaborate distinction between 'pacifism' and 'pacifism' (Ceadel 1980).

2. *Origins*

The French Revolution first created conditions that allowed peace movements to arise. On the whole, they remained restricted to the European-North-Atlantic sphere far into the twentieth century. Convictions matured within the pre-Revolutionary bourgeois society that there is no rational argument to justify war. The mass experience of war following the French Revolution of 1789 and Napoleonic rule gave the impulse for the organization of an individual peace mentality. The peace movement draw important program elements from a store of peace prospects dating back to Judaic-Graeco-Roman antiquity and early Christianity which manifested themselves in christology based, eschatological expectations of religious movements that included the vision of a non-violent society and in the religious practice of Christian peace sects (Brock 1991a) as well as in literature, the arts, philosophy and political science. Apart from the singularity of Gandhi in India (Brock 1983, Kozhuvanal, in: Brock and Socknat 1999) no indigenous nuclei for peace movements are to be found outside the European-North Atlantic sphere with the exception of Japan where they began appearing at the end of the nineteenth century particularly in connection with the reception of Christian ideas (Powles, in Brock and Socknat 1999), while other moves to organize pacifism starting around 1900 in Australia, New Zealand, Latin America, and even in Japan had either been dependent on or inspired by the Anglo-American model (see Fried 1972).

3. *The Starting Period*

The time spanning from the end of the Napoleonic era to the beginning of the age of imperialism was a take-off period for peace organizations. They emerged from two sources: on the one hand, religious motivation particularly characteristic of Mennonites and Quakers in North-America and England (Brock 1990, Ceadel 1996), which led to the founding of peace societies in New York and Massachusetts in 1815 and London 1816 and on the other hand a predominantly philanthropic rationalistic motivation leading to the foundation of the first peace societies on the continent in Paris in 1821 and Geneva in 1830. The discourse of

both focused on how to ban war as a means of conflict resolution and to get this idea generally accepted nationally and internationally.

3.1 *Difficult Dialectics*

All attempts to come to grips with this problem caused by the revolution of 1789 proved to be difficult. While the social organization to ban war had been possible, peace movements had to face a new type of war that was another by-product of that secular transition. National societies turned nation-states made a concept of war feasible completely different from every preceding one, i.e., it mobilized the nation as a whole with the total annihilation of the adversary as the final goal. The concept of peace based on rationalism arose from bourgeois society yet the same society due to its increasing economic and political dynamics carried an inherent potential for war the dialectics of which constitute the peace movements' field of activity during the nineteenth century until the World War I.

3.2 *Differences and Correspondence*

The general development assumes a different profile in every individual nation-state (van der Linden 1987). The US peace movement owes its origin to a very specific situation (Brock 1968). In Europe's young nation-states Italy and above all the German Empire (national peace societies founded in 1887 and 1892 respectively) the development occurred with considerable delay and had a nature different from that of the British Empire, a well established nation-state, or France a nation-state having aspirations towards Europe and overseas or a multinational state like Austria-Hungary (Austrian peace society founded in 1891). It is of significance whether a country has a stable parliamentary system including minority and nonconformist groups into the national consensus or if a country is characterized by a weak parliamentary structure and/or authoritarian leadership structure as well as a tendency to marginalize dissenting political views (see Ceadel 1987).

Notwithstanding their different national conditions all peace movements of the nineteenth century have a lot in common with regard to their homogenous appearance, their organization, and their means of communication. They use societies and associations typical of the bourgeois society of the nineteenth century. The growing participation of the radical bourgeois women's movement gains importance (see Bortolotti 1985). Communication was established via national and international congresses and a broad spectrum of literature. The result was a well-organized pacifist International, whose cohesion was strengthened by the foundation of the International Peace Bureau in Bern in 1892, and remained almost intact until World War I.

Peace activists were convinced that their concept of the need for peace could be most effectively promoted by individual powers of persuasion. Cognitive learning was believed to be the most effective method for instilling an understanding of the need for peace. That is why great importance was attached to schooling and national education as 'indirect means of peace.' Peace activists believe peace as well as war to be due to an individual act of will. This conviction led to appeals for peace addressed to the rulers and delivered at international congresses (Brussels 1848, Paris 1849, Frankfurt am Main 1850, London 1851, Manchester 1852, Edinburgh 1853). This continued on a universal level with the First Universal Peace Congress in Paris in 1889. The prerequisite for this was social homogeneity within the urban society as represented by merchants, industrialists, bankers, physicians, civil servants, professors, teachers, as Chickering shows for Germany (Chickering 1975). Protestant backgrounds show significantly greater receptiveness to pacifist recruitment than Catholic ones. In countries with a predominantly Catholic majority like Italy the willingness to embrace pacifism is characteristically present in connection with economic prosperity, liberal and/or republican convictions rooted in the Risorgimento, freemasonry, etc.

Eventually the social question vanished from the agenda of the international peace movement despite the considerable contributions of renowned socialists and anarchists in organizing pacifism from the left as early as during the inaugural congress of the International League of Peace and Liberty in Geneva in 1867. This corresponds to the increasing distance between the labor movement and the peace movements. In Germany, the distance is biggest, whereas it is substantially less in France and Italy and almost non-existent in England. The sooner the labor movement had separated from bourgeois democracy while opening up to Marxist theory the more brusquely it emphasized the distance from the peace movements. For the German peace movement cooperation with social democracy remained a top priority, that long went unrealized. In accordance with their bourgeois nature, peace movements maintained regular contact with liberal parties. Wherever political liberalism became diversified like in Germany the peace movement joined left liberalism because of its manifest inclinations towards internationalism, whereas national liberalism accused the peace movement of lacking national identity, and participated in stigmatizing pacifists (Holl 1988b).

Revolutionary wars did not present an argumentative problem to the early peace movements as long as the goal of these wars was national unity (van der Linden 1987). Behind this position was the expectation that lasting peace would set in with the elimination of reactionary governments at the hands of national movements. For the bourgeois left who were active in the early peace movements war for national unification

was legitimate and the last of all wars to be fought. Having found its destiny within the nation-state, a country was expected to fit into the community of nations and to submit to the international legal system in case of conflict. Pacifist arguments also legitimized national defensive warfare necessary to secure the goal reached in the process. S. E. Cooper has coined the term 'patriotic pacifism' for this position (Cooper 1991). The position on revolutionary warfare looked quite different with regard to social revolutionary goals, all the more when socialism began to dissociate itself from peace movements. Ever since bourgeois pacifism's rejection of war has been influenced by the fear that warfare could result in a social coup.

The British peace movement viewed obstruction of international free trade as yet another legitimate reason for war. In connection with the struggle over the Corn Laws economic interests were combined with pacifist propaganda. The international peace movement joined forces with the free trade movement resulting in an organizational symbiosis (Ceadel 1996) and the development of something along the lines of 'free-trade pacifism.'

As a bourgeois reform movement, the international peace movement joined abolitionism leading pacifism, utilitarianism, Quakerism, free-trade policy, and anti-slavery agitation to come together. Faced during the American Civil War with the conflicting aims of the abolition of slavery and maintenance of national unity on the part of the northern states and the defense of slavery and free trade on the Confederate side, the international peace movement leaned in the end towards the union while Quakers on both sides objected to military service (Brock 1968).

3.3 The Methods Propagated

The methods propagated for achieving and securing peace are twofold. As the prevailing state of affairs in international relations was perceived as 'international anarchy,' peace movements proposed a fundamental reform of these relations and suggested the creation of a legal codex. Concepts were discussed for legal procedures of conflict resolution to be enforced by an international tribunal and a court of arbitration with either facultative or compulsory arbitration. The underlying conviction was the call for 'peace through justice' that stimulated the origins of pacifist international law. The inherent idea was to restrain military influence on foreign affairs.

To the extent that the volatile nature of the national movements and the possibility of their exploitation by the cabinets were recognized as threats to peace, the call was made for fraternization and concepts for a league of nations on the basis of a free federation of European nations, republican constitutions, and the autonomy of communities to thus circumvent the state as a basis for social organization. This was propagated

at the Geneva Congress in 1867. In the end, corresponding attempts to organize based on Proudhonism had no chance. Continental peace movements preferred to propagate the 'United States of Europe' instead.

4. *The Age of Imperialism*

From 1901 onwards the supporters of peace movements called themselves 'pacifists' and their program 'pacifism.' Having the advantage of multilingual usability, the term met the requirements of an international movement. Earlier names such as 'friends of peace' finally vanished. The new term signaled an advanced level of theory and the intention of expanding the social base. The creation of the term reflects the success of both socialism and social Darwinism as social theories, and of imperialism. The spread of social Darwinist thought in Europe and in the USA that led public opinion to be more favorably disposed to imperialism and warfare had a negative effect on pacifist campaigning. Owing to its Eurocentrism and its inclination to cultural imperialism the international peace movement itself was occasionally susceptible to the imperialistic zeitgeist. The growing crisis of the global system incited the international peace movement and eminent pacifist authors to develop new concepts and theories (Grossi 1994).

Since the end of the 1860s new peace movements sprang up emphatically propagating peace negotiations by arbitration. Tsar Nicholas II's peace manifesto (1898) and the First Peace Conference in The Hague (1899) proved to be false confirmation of the movement's demand for disarmament. Despite this seeming success the peace movement's societal position remained marginal in Austria and Germany where it faced male prejudice not least because of the participation of the radical bourgeois women's movement and due to the fact that Bertha von Suttner, in whose novel *Lay Down the Arms!* (1889) a woman's concern about world peace is expressed, took a leading position in the peace movement (Holl 1988b). In his oeuvre *War in the Future in its Technical, Economic and Political Aspects* (1898) Jan Bloch predicted, against all social Darwinist and strategic expectations of the time, the disastrous consequences which were confirmed by World War I. A. H. Fried developed a peace theory he termed 'organizational pacifism' which ruled out world war because the growing network of international relations would prevent it for utilitarian reasons. Fried's theory won support from Norman Angell's book *Europe's Optical Illusion* (1909, extended 1910, *The Great Illusion*) that held future wars to be economically futile because of the interdependence of national economic interests.

The ineffectiveness of the peace movement gave way to further pacifist concepts on the eve of World War I. Their aim was to win prominent leaders in academic circles for international understanding. Their target

in Germany was the 'Verband für internationale Verständigung,' founded in 1911 (Chickering 1975). Furthermore there were efforts promoting bilateral international understanding and supported by the international peace movement: on a Protestant church-based level for instance between England and Germany, and on a parliamentary level between French and German MPs. Independent from such phenomena, communities which lived according to the principle of absolute non-violence continued to exist with particularly those groups following Tolstoy's teachings receiving the greatest public attention (Brock 1991b).

5. *World War I*

World War I created a deep rift within the international peace movement and cut the national peace movements' flexibility to a considerable extent. The experience of war, during which traditional peace societies held on to their patriotic pacifism led to the founding of pacifist organizations with a new style and program expressing the radical pacifists' frustration with the insufficient prewar peace societies. This ended the older pacifism's abstinence from domestic policy as the dependence of foreign policy on domestic conditions was scrutinized. In Germany, the 'Bund Neues Vaterland' demanded politics of domestic reform (Holl 1988b). The British 'Union of Democratic Control' called for a new foreign policy without secret diplomacy. In addition resolute pacifism organized in the Netherlands, 'Nederlandse Anti-Oorlog-Raad,' and in the USA, the 'League to Enforce Peace.'

The new pacifism formed a more solid base for international organizations: The international congress of pacifists in the Hague in 1915 led to the short-lived 'Central Organization for a Durable Peace.' The 1915 Hague congress of women war opponents, which was rooted in the radical bourgeois women's movement, ended with the foundation of the 'International Women's Committee for a Lasting Peace' which led to the 'International Women's League for Peace and Freedom' in Zurich in 1919 (see Schott 1997). Inspired neither by the old or new style peace movements nor by Gandhi and Tolstoy's teachings, refusal of military service surfaced for the first time as a mass phenomenon, particularly in Great Britain, where objectors joined the 'No Conscription Fellowship' during World War I.

6. *The Interwar Years*

During the interwar period, peace movements were marked by the coexistence of radical pacifist positions with the moderate orientation of traditional peace organizations and a widening of their social base into the working classes. While radical pacifism expressed

itself in the mass-movement 'No More War,' organizational pacifism lived on in the call for the 'League of Nations' that found outstanding support in Britain. The new peace organizations in France and England differed from the German peace movement in their judgment of the imminent threat to world peace posed by right-wing dictatorships. In France, the 'Ligue internationale des Combattants de la Paix' emerged as the mouthpiece of integral pacifism in the 1930s. It played down the threat of war emanating from Nazi Germany, attacked French defense policy, and demanded strict neutrality towards current military conflicts, a position that could well have led to collaboration during German occupation (Ingram 1991). This attitude corresponded to the British 'Peace Pledge Union,' which began to loose ground only after the obvious failure of the British appeasement policy (Ceadel 1980). With the beginning of national socialist rule, the German peace movement was forced into exile or exposed to persecution (see Holl 1988a).

7. After World War II

The nuclear attacks on Hiroshima and Nagasaki as well as the postwar strategy of nuclear deterrence developed in the course of the East-West confrontation gave new dimensions to the discussion on war and peace. Originating from the USA and directed against the Vietnam war, a radical peace movement staging a new type of mass protest and with the characteristics of the 'New Social Movements' came into being (see Breyman 1998). In accordance with nuclear pacifism, this peace movement argued that nuclear war of aggression, nuclear preventive warfare, and nuclear defensive warfare were virtually indistinguishable from one another. Additionally, it expressed doubt with regard to the peacekeeping effect of the nuclear deterrence strategy and played up the fear of the nuclear self-destruction of humanity. These arguments motivated the movement's opposition to all weapons of mass destruction. In a large number of Western countries in the 1980s, the new peace movement achieved a mass mobilization against the installation of medium-range NATO missiles in Western Europe that in the end proved futile.

See also: Social Movements and Gender

Bibliography

- Bortolotti F D 1985 *La Donna, La Pace, l'Europa*. L'Associazione internazionale delle donne dalle origini alla prima guerra mondiale. Franco Angeli Libri, Milan, Italy
- Breyman S 1998 *Movement Genesis. Social Movement Theory and the 1980s West German Peace Movement*. Westview Press, Boulder, CO
- Brock P 1968 *Pacifism in the United States: From the Colonial Era to the First World War*. Princeton University Press, Princeton, NJ

- Brock P 1983 *The Mahatma and Mother India: Essays on Gandhi's Non-violence and Nationalism*. Navajivan Publishing House, Ahmedabad, India
- Brock P 1990 *The Quaker Peace Testimony 1660 to 1914*. Sessions Book Trust, York, UK
- Brock P 1991a *Freedom from Violence: Sectarian Pacifism from the Middle Ages to the Great War*. University of Toronto Press, Toronto, Canada
- Brock P 1991b *Freedom from War: Nonsectarian Pacifism 1814–1914*. University of Toronto Press, Toronto, Canada
- Brock P, Socknat T P (eds.) 1999 *Challenge to Mars: Essays on Pacifism from 1918 to 1945*. University of Toronto Press, Toronto, Canada
- Ceadel M 1980 *Pacifism in Britain 1914–1945: The Defining of a Faith*. Clarendon Press, Oxford, UK
- Ceadel M 1987 *Thinking about Peace and War*. Oxford University Press, Oxford, UK
- Ceadel M 1996 *The Origins of War Prevention. The British Peace Movement and International Relations, 1730–1854*. Clarendon Press, Oxford, UK
- Chickering R 1975 *Imperial Germany and a World Without War. The Peace Movement and German Society, 1892–1914*. Princeton University Press, Princeton, NJ
- Cooper A H 1996 *Paradoxes of Peace. German Peace Movements Since 1945*. University of Michigan Press, Ann Arbor, MI
- Cooper S E 1991 *Patriotic Pacifism. Waging War on War in Europe 1815–1914*. Oxford University Press, Oxford, UK
- DeBenedetti C, Chatfield C 1990 *An American Ordeal. The Antiwar Movement of the Vietnam War* 1st edn. Syracuse University Press, Syracuse, NY
- Fried A H 1972 *Handbuch der Friedensbewegung, Vols. I and II*. Garland Publishing, New York
- Grossi V 1994 *Le Pacifisme Européen 1889–1914*. Bruylant, Bruxelles, Belgium
- Holl K 1988a German pacifists in exile. In: Chatfield C, van den Dungen P (eds.) *Peace Movements and Political Cultures*. University of Tennessee Press, Knoxville, TN, pp. 165–83
- Holl K 1988b *Pazifismus in Deutschland*. Suhrkamp Verlag, Frankfurt am Main, Germany
- Ingram N 1991 *The Politics of Dissent. Pacifism in France 1919–1939*. Clarendon Press, Oxford, UK
- Liddington J *The Road to Greenham Common. Feminism and Anti Militarism in Britain Since 1820*. Syracuse University Press, Syracuse, NY
- Schott L K 1997 *Reconstructing Women's Thoughts: The Women's International League for Peace and Freedom Before World War II*. Stanford University Press, Stanford, CA
- Vaisse M (ed.) 1993 *Le Pacifisme en Europe des années 1920 aux années 1950*. Bruylant, Bruxelles, Belgium
- van der Linden W H 1987 *The International Peace Movement 1815–1874*. Tilleul Publications, Amsterdam
- Wittner L S 1993 *One World or None. A History of the Nuclear Disarmament Movement Through 1953*. Stanford University Press, Stanford, CA

K. Holl

Peace Promotion, Psychology of

Brutal conflicts such as those in Rwanda, Kosova, and East Timor have boosted awareness of the subjective dimensions of war, genocide, and intergroup conflict

and of the need to address hatreds, fears, and emotional wounds of war in building peace. Within a multidisciplinary perspective, this article examines the psychological origins of war and destructive inter-group conflict and psychological approaches to resolving conflict nonviolently and to building peace with social justice.

1. Historical Context

Psychology became prominent partly through helping the US develop government tests for selecting military recruits in World War I. Through World War II, many psychologists contributed to war efforts, although William James, Edward Tolman, and others worked for peace.

Following World War II, social psychologists such as Morton Deutsch, Herbert Kelman, Otto Klineberg, Gardner Murphy, Charles Osgood, Muzafer Sherif, and Ralph White analyzed sources of intergroup tension and methods of nonviolent conflict resolution, with an emphasis chiefly on the Cold War. The nuclear threat animated large numbers of psychologists worldwide, as researchers such as Robert Jay Lifton, John Mack, and Milton Schwebel examined the psychic impact of living under the threat of nuclear annihilation. Throughout the Cold War, psychological research focused mostly on negative peace, that is, on stopping and preventing war (White 1986), rather than on positive peace, which involves social justice, positive inter-group relations, human rights, and sustainable development.

Following the Cold War, psychologists examined a wider array of peace issues such as women and war, militarism, children and armed conflict, environmental destruction, and postconflict healing.

2. Psychological Sources of War and Destructive Intergroup Conflict

Armed conflicts, including the intrasocietal conflicts that are now the dominant form of war, have complex political, historical, and economic roots. Much destructive conflict originates in a divergence of objective interests, competition over scarce resources such as land and water, and worsening life conditions. The 1994 genocide in Rwanda, for example, had antecedents in land scarcity, social class, and the political and economic privileging of Tutsis under the Belgian colonial regime (Prunier 1995). Many conflicts have their origins in struggles over oil, diamonds, land, and other objective resources.

Armed conflict also has subjective sources such as social identity and ideology. Even when no objective divergence of interest exists, people tend to categorize themselves as belonging to different groups and to

show in-group favoritism in judging people, thereby attaining self-esteem and positive self-identity from identification with the in-group (Tajfel 1982). Typically, they judge others relative to the standards set by the in-group, leading to ethnocentric attitudes and behavior, and they derogate members of the out-group. Identity processes alone do not cause violence, but they set the stage for other subjective processes such as hostile ideologies to generate hatred and fighting. Before the Rwandan genocide, President Habyarimana, a Hutu, used mass media to create an ideology of hatred toward Tutsis. Similarly, Hitler used an ideology of racial superiority and global Jewish conspiracy to fuel anti-Semitism and enable the Holocaust (Staub 1989).

Social identity processes play a key role in inter-ethnic conflicts, often referred to as identity conflicts. Ethnic groups are 'imagined communities' (Anderson 1991) that share distinctive language, religion, social institutions, and origin myths. Ethnic identities exercise powerful emotional influence, as children are taught to love their people, their mother tongue, and their way of life. People learn to define themselves by reference to the larger group (e.g., 'I am Russian'), to experience unity with the group, to be patriotic, and to make sacrifices for the group. The sense of unity and solidarity, integrated with a powerful sense of homeland and a sense that the group requires protection, may fuel nationalism and separatist desire for an independent national state.

Conflict escalation often stems from the interplay of objective and subjective factors. As conflict over resources escalates, negative psychological dynamics come into play, and these become self-perpetuating parts of the conflict process (Rubin et al. 1994). Each group tends to create enemy images—demonizing, exaggerated stereotypes of the 'other' that amplify fear and motivate aggressive action (White 1984). These images bias perceptions, leading adversaries to make negative attributions or inferences about motivations regarding the 'other's' behavior. For example, building military forces could be motivated by desire for security, but an adversary who harbors enemy images tends to perceive such actions as having hostile, offensive intent. Negative perceptions and fears on both sides fuel behaviors that the 'other' regards as threatening, contributing to malignant conflict spirals that tend to become self-fulfilling prophecies. By using public media to propagate enemy images, leaders mobilize society for war and genocide.

In a destructive conflict, damage to the parties' relationship becomes part of the conflict, and the emotional and social wounds encourage fighting. Viewing the other side's actions as transgressions, each party sees itself as a victim, constructs collective beliefs or myths about the traumas inflicted on it, and believes that violence is necessary for self-preservation (Volkan 1997). Further, the parties tend to become highly polarized, and intragroup pressures discourage con-

tact with the other side, which is viewed as having incompatible goals. In protracted conflicts, each side may come to view itself as irreparably opposed to the other side, leading to an ethos of antagonism and a tendency of each group to define itself in part via opposition to the other.

Subjective influences are visible also in the biased decision-making of policy elites. Many leaders' aversion to losses encourages excessive risk-taking and departure from rational decision making. In intense conflicts, leaders may experience powerful fears, which motivate misperceptions regarding their adversaries' motives, strength, and willingness to fight. As intercommunal tensions escalate, leaders may experience a cognitive constriction evidenced in reduced complexity of views of the adversary expressed in speeches. Leaders may also use inappropriate historical analogies to guide thinking about current crises. Under pressure of crises, small leadership groups having a strong *esprit de corps* and a charismatic leader may make flawed decisions guided by premature consensus, a sense of invulnerability, failure to weigh moral concerns, internal suppression of dissent, and poor contingency planning (Janis 1982). In a variety of contexts, obedience to authority is a powerful enabler of armed conflict, human rights abuses, and even genocide (Kelman and Hamilton 1989).

3. Nonviolent Conflict Resolution

Conflict can be constructive when it is managed in ways that promote communication, healthy relationships, and a sense of positive interdependence (Deutsch and Coleman 2000). To handle destructive conflict nonviolently, psychologists have developed or refined numerous tools designed to resolve conflict and to transform or repair damaged relationships (Rubin et al. 1994).

Negotiation, the handling of conflict through communication and bargaining by the parties themselves, has important psychological dimensions. If, as often occurs in official diplomacy between hardened adversaries, opponents view negotiation as a win-lose affair, they may dig into entrenched positions and use damaging, coercive tactics such as threats. Psychologically, a more useful alternative is principled negotiation, which reframes negotiation as a win-win affair. In this approach, parties collaborate to solve their problems and seek not to beat the opponent but to enable all parties to meet their underlying needs, thereby contributing to improved communication, empathy, and relationship. Which approaches to negotiation are likely to succeed depends not only on the situation, the kind of issues at stake, and the amount of fear and relational damage, but also on culture. For example, US negotiators favor direct communication and emphasize what is said, while Asian negotiators often communicate indirectly, re-

lying on context and maintaining group harmony by avoiding direct confrontation.

Strong mistrust, fear of appearing weak, and poor communication can make negotiation infeasible. In this situation, a preferred approach is mediation, in which a third party helps the parties to negotiate a settlement. By proposing alternatives that the parties may not be willing to suggest themselves, mediators enable face-saving. They may also use social influence processes such as 'carrots and sticks,' a mixture of promised rewards for conflict reducing steps and promised punishments for intransigent behavior. Skilled mediators manage the conflict process, separating the parties when they are likely to say or do damaging things.

Official diplomacy, however, is limited since hardened adversaries may be unwilling to meet publicly. Also, public agreements can evoke backlashes, and official treaties may not change the polarization, fear, and hostility present in divided societies and communities. Kelman has pioneered the use of interactive problem-solving workshops (Fisher 1997) as a tool of unofficial diplomacy and relational improvement. Regarding the Israeli-Palestinian conflict, a typical workshop brings together three to six Israelis with an equal number of Palestinians for three days of private dialogue facilitated by social scientists. In analytic, problem-solving discussion, the participants examine the main issues at stake in the conflict, explore their concerns and fears, and identify possible steps and solutions that could reduce the political and psychological barriers on both sides. Empathy, problem solving, and group exploration of 'what-if?' questions are encouraged. Afterwards, participants take their new learning about the other side back into their communities, beginning the wider process of community transformation. By building positive relationships and infusing useful ideas into conflict-torn communities, this method provides a useful complement to official negotiation.

Positive contact between members of polarized, isolated groups can also reduce destructive conflict (Pettigrew 1998). Positive contact is likely to occur in a context of equal status that enables individuals to get to know each other personally and when steps have been taken to gain leaders' support and prevent damaging behavior. To reduce perceptions that the positive individuals from the out-group are exceptions, one may use cognitive strategies that enable generalization. One useful strategy is to build superordinate group identities, as occurred in postapartheid South Africa when blacks and whites forged larger identities as 'South Africans' and championed the national rugby team.

The most powerful form of cross-group contact involves cooperation toward the accomplishment of shared goals that neither group could achieve on its own. Global problems such as nuclear proliferation, environmental destruction, and the HIV/AIDS pan-

demic provide many possibilities for cooperation on shared goals that mitigates and prevents destructive conflict.

4. Building Peace with Social Justice

Beyond war prevention and nonviolent conflict resolution, peace requires social justice, social equity, and sustainable patterns of living. Psychology has analyzed how beliefs, attitudes, and practices support social injustice and '-isms' such as sexism.

In systems of severe oppression and genocide, there is a pattern of moral exclusion in which out-groups are dehumanized and removed from the moral realm, providing a rationalization for killing them. Similarly, sexism is supported by socialization norms and practices that privilege boys, encourage girls to conform to stereotypic roles, and encourage acceptance of existing patterns of patriarchy. Environmental damage through, for example, the production of mass amounts of toxic waste, often leads to environmental racism via 'not in my back yard' attitudes that encourage storage or dumping of toxic materials near impoverished neighborhoods inhabited by minority groups that lack political power. Efforts toward poverty alleviation may be thwarted by tendencies to blame the victim, enshrined in societal beliefs that people are poor because they are lazy or deserve it.

In promoting social justice, useful psychological tools include steps to increase the sense of positive interdependence between groups, media campaigns that heighten salience of justice issues and indicate specific behaviors that can help address the problem, positive modeling of tolerance and restraint by authority figures, and methods of attitude change. In addition, commitment can often be built through the 'foot in the door' method in which agreement to a small request increases the likelihood of subsequent willingness to engage in more effortful action. To change behavior on a wider level, one may use community-based social marketing methods that strategically assess points of leverage for social change, pilot selected interventions for changing attitudes and behavior, and replicate the process on a larger scale (McKenzie-Mohr 2000).

Education for peace is essential in building social justice and peace, as most people are socialized into systems of violence (Raviv et al. 1999). In schools, psychologically informed programs have developed skills of nonviolent conflict resolution, encouraged peer mediation of conflicts, enabled cooperative learning that also reduces excessive competition, and built curricula that undermine hatred and stereotypes. Beyond schools, education for peace has included activities to enable constructive handling of family conflict, cross-conflict dialogues in divided communities, and use of mass media to show positive handling

of conflict, to counteract black-white thinking, and to reduce the excessive emphasis on violence evident in television and the war-toy industry.

5. Future Directions

Since wounds of past violence create emotional vulnerability and enable cycles of violence, building peace requires the healing of emotional wounds and reconciliation. Programs of humanitarian assistance and development increasingly include psychosocial activities aimed at healing the wounds of war through normalizing activities, expressive arts, reintegration of former child soldiers, and use of local cultural networks and resources. A key task for the future is to find the most effective, culturally appropriate ways of rebuilding torn societies. There is great need of research on reconciliation, the repair of damaged relationships, and its connection with processes such as truth-telling, forgiveness, and establishment of justice. As psychology addresses these issues, it will examine a wider array of cultures and develop theories that connect microlevel change in individuals and small groups with macrolevel changes in societal and international structures.

See also: Conflict and Conflict Resolution, Social Psychology of; Dispute Resolution in Economics; Peace; Peace and Nonviolence: Anthropological Aspects; Peace Movements; Peacemaking in History; Religion: Peace, War, and Violence; War: Anthropological Aspects; War: Causes and Patterns; War, Sociology of

Bibliography

- Anderson B 1991 *Imagined Communities—Reflections on the Origin and Spread of Nationalism*. Verso, London
- Christie D J, Wagner R V, Winter D D 2001 *Peace, Conflict and Violence*. Prentice-Hall, Upper Saddle River, NJ
- Deutsch M, Coleman P 2000 *Handbook of Conflict Resolution: Theory and Practice*, 1st edn. Jossey-Bass, San Francisco
- Fisher R 1997 *Interactive Conflict Resolution*. Syracuse University Press, Syracuse, NY
- Janis I 1982 *Groupthink: Psychological Studies of Policy Decisions and Fiascoes*, 2nd edn. Houghton Mifflin, Boston
- Kelman H C, Hamilton V L 1989 *Crimes of Obedience*. Yale University Press, New Haven, CT
- McKenzie-Mohr D 2000 Fostering sustainable behavior through community-based social marketing. *American Journal of Psychology* 55: 531–7
- Pettigrew T F 1998 Intergroup contact theory. *Annual Review of Psychology* 49: 65–85
- Prunier G 1995 *The Rwanda Crisis*. Columbia University Press, New York
- Raviv A, Oppenheimer L, Bar-Tal D 1999 *How Children Understand War and Peace*. Jossey-Bass, San Francisco

- Rubin J Z, Pruitt D G, Kim S H 1994 *Social Conflict*, 2nd edn. McGraw-Hill, New York
- Staub E 1989 *The Roots of Evil*. Cambridge University Press, Cambridge, UK
- Tajfel H 1982 *Social Identity and Intergroup Relations*. Cambridge University Press, New York
- Volkan V 1997 *Bloodlines*, 1st edn. Farrar, Straus and Giroux, New York
- White R K 1984 *Fearful Warriors*. Free Press, New York
- White R K (ed.) 1986 *Psychology and the Prevention of Nuclear War*. New York University Press, New York

M. Wessells

Copyright © 2001 Elsevier Science Ltd.
All rights reserved.

Peacemaking in History

1. Definitions

1.1 Peace and War

War is declared, begun or unleashed, or it breaks out; peace is concluded, settled, agreed or imposed. Ordinary language thus shows an important asymmetry: war is either the outcome of forces neither side can fully control, or the result of one-sided action, while peace has to be established in an effort involving both sides, unless it is simply the result of the dying out of fighting or of the annihilation of one side. For the purpose of this article, *peace* is defined negatively as absence of war and other forms of widespread collective fighting. This static definition does not do justice to the concept of peace as a movement, a process of improvement with stages in between just a momentary absence of war and ideal, eternal peace and harmony. But this positive, dynamic view of peace concerns a different subject; this article deals only with the crucial transition from a state of war to a state of peace. *War* is defined as a state of collective armed violence between politically organized, at least *de facto* and usually *de jure* independent communities, during which special law is valid. The main characteristic of this law is the relaxation of the rules which prohibit killing. War does not presuppose continuous fighting. But if fighting ceases altogether, one can speak of a *de facto* peace. Fighting without application of different law is not war, but anarchy or rebellion. If, later on, the law of war is recognized by both sides, the rebellion or the anarchy has developed into *internal* or *civil war*. Civil war occurs between parties originally forming one political community, while war between independent communities is *external* or *international war*.

With respect to peacemaking, one fundamental aspect of war has to be mentioned: war has inevitably the character of an ordeal, of a 'judgement of God,' as power and not justice decides its outcome (Kant 1968).

1.2 Peacemaking and Pacification

Peacemaking is defined as transition from a state of war to a state of peace by a deliberate action of the parties involved. Departing from a differentiation between three types of peace introduced by Raymond Aron: equilibrium—hegemony—empire (Aron 1962), peacemaking is here placed between two extremes which mark ideal types. *Peacemaking proper* is characterized by coordination, and hence by full legal (although not necessarily material) equality and independence of the parties. If there is a new conflict between them, it will be international war. *Pacification* is characterized by subordination. It implies the legal extinction of one party as an independent entity by its incorporation into the other. Usually, it is connected with an ex post denial of the vanquished's right to conduct war by the victor. If there is a new conflict, it will start as a rebellion and may develop into a civil war. Each concrete case of war termination can be situated between these two extremes. For pacification, an overwhelming victory of one side is a necessary condition, whereas peacemaking proper tends rather to be linked with a stalemate. But the relative strength of the parties is not the only factor deciding the outcome. There are periods in history during which pacification prevails.

2. Peacemaking Before the Twentieth Century

The roots of peacemaking lie beyond the grasp of written sources. Anthropological research as well as written evidence of transactions between literate and nonliterate political communities suggest that wherever there is fighting between politically organized groups, there are also mechanisms to end fighting on more than just a *de facto* basis, which presupposes some reciprocally recognized arrangements, and thus some reliability and stability. Special oral procedures to memorize the contents of the transactions are developed (cf., e.g., the highly elaborate methods of the North American Indians in their dealings with the Europeans in the eighteenth century, which the Europeans had to adopt as long as they could not simply dictate the terms (Franklin 1938)). While the material stipulations can vary considerably according to the way of living of the peoples involved, the most important and the most difficult arrangements concern guarantees for the keeping and carrying out of the stipulations. Whatever the intention of those who conclude the peace, its oral form tends to limit its scope in time and in space. Frequent renewal, especially after the death of the head of one of the contracting parties, is the most important method to maintain the force of the agreements. Wherever writing develops, peace treaties are among the first instruments for which it is used, and they are frequently written on the most durable materials, like

stone or metal. The oldest treaties preserved date from the third millennium BC in the Ancient Near East. They are to be seen in the context of the pacification of vassals, while the first known treaty between great powers is from the thirteenth century BC, between the Hittite and the Egyptian empires. The problem of guarantees is dealt with by oaths and by the invocation of the intervention of deities, while in other cases hostages are given, either unilaterally or reciprocally, or dynastic matrimonial unions link the destinies of the two parties.

In the course of time, the main topics of peacemaking, besides peace itself and its guarantees, turned out to be territorial, financial and economic. Sometimes they were supplemented by clauses dealing with the destiny of certain populations. Peacemaking proper was almost invariably connected with an amnesty for those who had supported the enemy during the war, based on a more general promise of reciprocally forgiving and forgetting what had happened during the war. The contracting parties were to behave as if peace had never been interrupted, thus marking a sharp contrast between war and peace. There was also a strong tendency in all parts of the world to stress the perpetual character of the peace concluded, not so much as an expression of a realistic expectation but rather as a reminder of a general characteristic of peace (Fisch 1979).

Whether peacemaking proper or pacification prevailed in a certain period and in a certain area, depended to a great extent on the general nature of the relationships between political entities. Well-documented genuinely pluralistic systems of a balance of power (which tend to foster peacemaking proper) are rare. The cases which probably came nearest to it were relationships between the Greek city states in the fifth and fourth centuries BC and between the Hellenistic states before the Roman conquest. Even here, there were tendencies for hegemonies to develop as with Athens or Sparta in Greece. Later the Greek system was absorbed by an empire, first Macedonia and then Rome. It is however to be supposed that in areas and periods with many smaller political entities, of which we have few written sources, peacemaking proper was also the rule. In international systems built around powerful empires, on the other hand, there was a strong tendency to pacification, certainly in the formal, although less in the material sense. Great empires, among them the Roman, the Chinese, the Islamic Khalifate, and the Ottoman and the Mogul Empires, frequently showed a reluctance towards peacemaking proper, claiming some kind of overlordship over their neighbors and rivals. They tended not to conclude formally reciprocal treaties with them but to impose unilateral edicts on them, or else they limited the duration of treaties. The Islamic doctrine, for example, allowed no peace but only armistices of up to 10 years with non-Muslims. Peace was not agreed, but granted. However, reality often belied

these pretensions; *de facto* the solution was frequently one between equals, and sometimes the empire was even in the weaker position.

In European history there is a somewhat different interplay between peacemaking proper and pacification. In the Middle Ages there were two trends towards pacification. On the one hand, the Emperor had a claim to supremacy over all the rulers of Christendom. In practice, this claim had hardly any importance; peacemaking among the great monarchs was legally and formally reciprocal. Thus, the basic structure of peacemaking in Europe after major wars was peacemaking proper long before 1648. Much more important with respect to pacification were, on the other hand, feudal relations between lords and vassals on various levels. Vassals tried to acquire full independence, which included the right to be a party in war and thus to become a subject of peacemaking, while lords tried to gain full control over their vassals, which led to attempts at pacification. The result of the century-old struggle was a balance of power-system with fully sovereign entities which, among themselves, practiced only peacemaking proper. In some countries, the ruler (usually a king) had been able to subjugate his vassals by methods of pacification, while in others, especially in the Empire, the vassals had gained full or almost full sovereignty.

Modern European history from the sixteenth century, and even more from the seventeenth century onwards, is thus characterized by a system of legally strictly equal relations. In its framework, particular methods of peacemaking were further developed. During the war, various kinds of peace feelers would be extended. Frequently, mediators played an important part, either as honest brokers or as great powers which tried to influence the settlement according to their own interests (on mediation in general, see Touval and Zartman 1985). The process of peacemaking proper tended to be broken up into three stages: armistice, with cessation of hostilities and, frequently, important material settlements—preliminary treaty, containing the core of the settlement—definitive treaty. This was, however, never a necessary pattern. Preliminary peace treaties played a really significant part only in the 19th century. Frequently, the most important decisions were taken even before hostilities ended (Pillar 1983). One of the most salient features were the peace congresses after general European wars, especially those in 1648, 1713, 1814–15, and 1919. The first of them, in Münster and Osnabrück, which led to the Peace of Westphalia, was an important stage in the development of a European system of legally equal, sovereign states. But it was not, as is frequently claimed, the point of departure for this system, which had much older roots. Such congresses developed general, often multilateral settlements which were likely to be more stable than mere bilateral peace treaties. Their success was remarkable, especially after the Congress of Vienna in 1814/15,

which constituted a kind of joint hegemony, combined with a joint responsibility for the peace, of the great powers. But after all, general peace settlements had as their corollary general wars.

This was, however, but one side of modern European peacemaking. The European states conquered, from the sixteenth to the twentieth century, vast parts of the world. In many areas of Asia and Africa, peacemaking proper between Europeans and non-Europeans was the rule up to the nineteenth century. Especially in America, there were widespread attempts at pacification. They were frequently successful: the extra-European rulers were brought into a legally subordinate position. This happened even in areas where, at first, there had been peacemaking proper, especially in India and Indonesia. The difference was felt by the contemporaries. In Spain, for example, *pacificación* officially replaced *conquista* in 1573, showing the origin of the concept in conquest and not in peace. Yet pacification was often merely a fiction, because the colonial power was too weak to impose its will. Nevertheless, in the long run, pacification tended to become the rule in most colonial dealings.

Thus, at the beginning of the twentieth century, the European experience was one of peacemaking proper in Europe and of frequent pacification outside Europe. For both purposes elaborate and successful techniques had been developed. But this also meant that war was a well-known feature of international life. One tried to avoid it as far as possible; nevertheless it belonged to the accepted experience. This acceptance was to be challenged in the twentieth century.

3. Peacemaking in the Twentieth Century

The First World War began as a traditional European war. The fierceness of its fighting and the destructiveness of modern weapons increased the number of combat casualties beyond that of all previous wars. This led to increased demands for the outlawry of war, that war should no longer be considered as a legitimate means of politics but as a crime, and propaganda on both sides tried to depict the enemy as a criminal.

The Second World War further strengthened the view that war was a crime, while the postwar period brought, with thermonuclear weapons, for the first time in history the possibility of a war destroying mankind altogether, so that peacemaking would be neither necessary nor possible. This did not mean, however, that smaller wars had become impossible.

These developments had consequences for peacemaking. The Covenant of the League of Nations, which was part of the peace treaties after the First World War, prohibited war at least to a certain extent (articles 10; 12–15). The Briand–Kellogg Pact of 1928 brought the renunciation of war as a means of national politics, while the UN Charter of 1945 prohibited the

use of force—and thus war—altogether, with the exception of legitimate self-defense (articles 2, 4; 51), which was no real exception, as self-defense is the answer to an aggression which in turn is prohibited. Combined with the outlawry was the attempt to prevent war by a system of collective security which would replace war by police-actions of the totality of the member-states against a peacebreaker. Ideally war would have been abolished altogether by this system. But it was weak, because neither the League nor the UN had enough coercive means of its own and even less the power monopoly which would have been needed in order to reduce warfare to an efficient system of policing.

At least in theory, the outlawry of war excluded the possibility of peacemaking proper. If one side was guilty of a criminal act there could be no peace on the basis of legal equality but only punishment. Yet, as there was no efficient system of policing, war still maintained its character of a judgment of God, and there was no guarantee that the guilty party could be punished. Peacemaking reflected these facts. It tended to join elements of both peacemaking proper and pacification, the mixture being dependent on the outcome of the war. After the First World War this meant a combination of the intra-European and of the colonial tradition. There was a peace congress, but with the losers excluded, and there were peace treaties, but with the vanquished formally admitting their responsibility for the war, thus being placed into a legally inferior position. This tendency was increased after the Second World War, when the vanquished were not only deemed, and had to confess themselves, guilty of the crime of aggression, but their leaders were brought to trial for their crimes and their states were occupied and administered by the victors. Especially after the World Wars there was a tendency by the victors to devise definitive solutions at a very early stage of the negotiations if they did not simply impose them by instruments of surrender or armistices.

Peacemaking in the twentieth century outside the context of the World Wars followed this pattern to a great extent. Theoretically, or rather dogmatically speaking, there was no longer any room left for peacemaking proper. But wars could still lead to a stalemate and render traditional peacemaking necessary, although there might be no traditional final peace treaty. A particularly elaborate example was the termination of the Vietnam War 1968–73/5. Thus, peacemaking proper has maintained a certain importance in the context of the termination of wars, and it is likely to keep it as long as there is no efficient global system of policing, although the relevant activities of the UN have increased, especially after 1989, in the shape of peacekeeping and peace enforcement. Peacekeeping often takes the place of a provisional arrangement instead of formal, definitive peacemaking, while peace-enforcement has the character of an imposed mediation. The proliferation of

civil wars in recent years has even increased the importance of some traditional techniques of peacemaking, as the coexistence of former civil-war parties makes reconciliation more important than it is between independent, separate states.

4. *The Conditions of Peace: Research into Peacemaking*

It is an almost stereotyped complaint in the literature on peacemaking that, compared with research into the causes of war, very little research into the causes of peace has been carried out. This is, however, not so much the consequence of a predilection for war or even a warlike disposition of the researchers, but rather of a basic difference in the public appreciation of war and peace. In a prescientific view, war has causes, while peace has not, in the sense that peace is considered as the ordinary and war as the exceptional state of things, even though in reality periods of peace may be shorter than periods of war. The wish for peace usually does not have to be justified in the same way as the wish for war. A state of war sooner or later ends in some kind of peace, while a state of peace does not necessarily lead to war.

The scant research on peacemaking seems to be one of the consequences of these facts. Moreover it is difficult to conceive of causes of peace in a scientific sense. Peace has been successfully concluded in extremely diverse political situations, and there are no methods to calculate the probability of peace in a specific constellation.

Traditionally, peacemaking is a central object of historiography in general and of diplomatic and military history in particular. The focus is usually on individual wars and their termination. Favorite objects of research have been the plans for permanent or perpetual peace through the ages (as an introduction cf., e.g., Hinsley 1963) and the main general peace congresses from 1648 to 1919 (cf. summaries of Holsti 1991 and Osiander 1994), while a comprehensive account of the termination of World War II has yet to be written. Such research helps to understand particular cases of peacemaking, but it usually is little concerned with general causes. More systematic, comparative research has occasionally been conducted during great wars, with a view to facilitate peacemaking. This led neither to more sophisticated systematic accounts nor to the search of causes of peace, but rather to surveys of the range of material regulations of earlier peace settlements (cf., e.g., Phillimore 1917 and Phillipson 1916). The only and very important exception is Quincy Wright's influential *A Study of War* (1965), first published during the Second World War, which, however, was the result of a project initiated long before the war. Wright does not only deal with causes of war but also with causes of peace.

After 1945 there was an apparently paradoxical development. In the social sciences, the threat of an all-destructive thermonuclear war gave, especially from the 1960s, a boost to peace studies and peace research, with the *Journal of Conflict Resolution* (since 1957) and the *Journal of Peace Research* (since 1964) as leading organs. But as peacemaking was no feasible possibility in the context of a thermonuclear war, the central object was the prevention of war and the preservation and improvement of peace. Nevertheless, as experience showed the possibility and the reality of many smaller wars under the umbrella of the thermonuclear stalemate, war termination and peacemaking became the object of at least some studies in the social sciences, especially under the influence of the extremely difficult termination of the Vietnam War (cf., e.g., Fox 1970). In addition, there was the older tradition of more historically oriented research. Hentig (1952) dealt with the general modalities of peacemaking, while Fisch (1979) investigated the concepts of war and peace as they were developed in the peace treaties themselves, including the distinction between peacemaking proper and pacification. The most important basis for research in the social sciences became Wright (1965). After him, Raymond Aron's great survey of peace and war (1962) located peacemaking in the context of international studies and of the Cold War. Since the late 1960s there have been a number of studies dealing systematically with possible causes of peace, on the basis of the comparison of a greater or smaller number of war terminations, e.g., Randle (1973), Pillar (1983) and Holsti (1991) or, more formalized, Carroll (1969). On recent research see Massoud (1996). There exists, however, no theory of war termination or peacemaking, and it is unlikely that it can be developed, considering the extremely varied factors which may combine to bring about peace. The same holds for research into the conditions of maintaining a peace once concluded. Research in international law has particularly dealt with the question of the outlawry of war in the twentieth century, but less with its consequences for peacemaking. There is a challenge for new research in the growing importance of international peacekeeping, peace-enforcement and peacebuilding during recent years. For an introduction, see Otunnu and Doyle (1998). But the main problem remains the contradiction between the outlawry of war and situations of stalemate which will go on requiring traditional peacemaking as long as there is no world-wide monopoly of power. In this context, it will be important to compare the present methods of peacemaking and peace-enforcement with the traditional forms of peacemaking and pacification.

See also: Cold War, The; Conflict and Conflict Resolution, Social Psychology of; Conflict/Consensus; First World War, The; Military and Politics; Military History; Peace; Peace and Nonviolence;

Anthropological Aspects; Peace Movements; Revolution; Revolutions, Theories of; War, Sociology of; Warfare in History

Bibliography

- Aron R 1962 *Paix et Guerre Entre Les Nations*, 3rd edn. Calmann-Lévy, Paris [Engl. transl. 1966 *Peace and War. A Theory of International Relations*. Weidenfeld and Nicolson, London]
- Carroll B A 1969 How wars end: An analysis of some current hypotheses. *Journal of Peace Research* 6: 295–321
- Fisch J 1979 *Krieg und Frieden im Friedensvertrag*. Klett-Cotta, Stuttgart, Germany [War and Peace in the Treaty of Peace]
- Fox T R 1970 How wars end. *Annals of the American Academy of Political and Social Science* 392
- Franklin B 1938 *Indian Treaties Printed by Benjamin Franklin* (with an introduction by van Doren C and bibliographical notes by Boyd J P). The Historical Society of Pennsylvania, Philadelphia, PA
- Hentig H von 1952 *Der Friedensschluss. Geist und Technik einer verlorenen Kunst*. Deutsche Verlags-Anstalt, Stuttgart, Germany [The Conclusion of Peace. Spirit and Technique of Lost Art]
- Hinsley F H 1963 *Power and the Pursuit of Peace. Theory and Practice in the History of Relations Between States*. Cambridge University Press, Cambridge, UK
- Holsti K J 1991 *Peace and War: Armed Conflicts and International Order 1648–1989*. Cambridge University Press, Cambridge, UK
- Kant I 1968 Zum ewigen Frieden. In: Weischedel W (ed.) *Id., Werke in Zehn Bänden*. Wissenschaftliche Buchgesellschaft, Darmstadt, Germany Vol. 9, pp. 191–251 (original work published in 1795) Engl. transl. 1991 Perpetual Peace. In: Reiss, H (ed.) *Id., Political Writings*. Cambridge University Press, Cambridge, UK pp. 93–130.]
- Laszlo E, Yoo J Y (eds.) 1986 *World Encyclopedia of Peace*, 1st ed. (4 Vols). Pergamon, Oxford, UK
- Massoud T G 1996 War termination. Review essay. *Journal of Peace Research* 33: 491–6
- Osiander A 1994 *The States System of Europe, 1640–1990. Peacemaking and the Conditions of International Stability*. Oxford University Press, Oxford, UK
- Otunnu O A, Doyle M W (eds.) 1998 *Peacemaking and Peace-keeping for the New Century*. Rowman & Littlefield, Lanham
- Phillimore W G F 1917 *Three Centuries of Treaties of Peace and Their Teaching*. J. Murray, London
- Phillipson C 1916 *Termination of War and Treaties of Peace*. T Fisher Unwin, London
- Pillar P R 1983 *Negotiating Peace. War Termination as a Bargaining Process*. Princeton University Press, Princeton, NJ
- Randle R F 1973 *The Origins of Peace. A Study of Peacemaking and the Structure of Peace Settlements*. Free Press, New York
- Touval S, Zartman I W (eds.) 1985 *International Mediation in Theory and Practice*. Westview Press, Boulder, CO
- Wright Q 1965 *A Study of War. With a Commentary on War Since 1942*, 2nd edn. University of Chicago Press, Chicago

J. Fisch

Pearson, Karl (1857–1936)

Most of Karl Pearson's scientific work concerned human beings—he studied inheritance, physical anthropology, and disease—but his significance for the social and behavioral sciences is as a founder of modern statistics. Through the nineteenth century there had been a sense that the astronomers' methods for treating observations—the theory of errors—could be applied more generally. However, the modern conception of the scope and organization of statistics dates from the turn of the twentieth century when Pearson developed methods of wide applicability and put together the components of a discipline—the university department, courses, laboratories, journals, tables. His contemporaries may have done some of these things better but none had so much total effect.

1. Career

Karl Pearson was proud that his ancestors were of 'yeoman stock' though his father was a barrister. At Cambridge Karl read mathematics. Third place in the examinations brought him a college fellowship and the financial freedom to pursue his very wide interests. Pearson qualified as a barrister and studied social, philosophical, and historical issues. He developed his own view of man's place in a post-Christian world, expressing this in a novel and a play as well as through essays; some of those were published in the *Ethic of Freethought* (1888).

In 1884 the essayist and freelance lecturer became professor of Applied Mathematics and Mechanics at University College London. Besides carrying out the usual duties, Pearson took on two books left unfinished when their authors died. W. K. Clifford's *Common Sense of the Exact Sciences* explained the nature of mathematics and it would link Pearson's early general essays to the *Grammar of Science* (1892). The *Grammar of Science* presented the scientific method as 'the orderly classification of facts followed by the recognition of their relationship and recurring sequences.' In later editions the positivist ideas about space and force were joined by an exposition of statistical ideas. The other book, Isaac Todhunter's *History of the Theory of Elasticity*, was the history of Pearson's own specialism. Pearson actually wrote much more of this huge work than Todhunter himself. Pearson's later historical researches—a biography of Galton and lectures on the history of statistics—combined Todhunter's thoroughness with Pearson's own very different awareness of broad intellectual currents.

Pearson moved from elasticity and philosophy to biometry and statistics under the influence of W. F. R. Weldon, appointed to the college in 1890 as professor

of Zoology. Weldon was applying Galton's statistical methods to species formation and Pearson helped him extend the methods. However, Pearson was not just a mathematical consultant, he became a contributor to the biological literature in his own right. The partnership with Weldon was central to Pearson's development and he made his most important contributions to statistical thinking in these early years with Weldon. Pearson wrote over 300 pieces after 1900 but they were on lines he had already established. Of course these works demonstrated what statistical methods could achieve and were part of making statistics a recognized discipline.

In 1901 Pearson, Weldon and Galton founded *Biometrika*, a 'Journal for the Statistical Study of Biological Problems' which also published tables and statistical theory. From 1903 a grant from the Worshipful Company of Drapers funded the Biometric Laboratory. The laboratory drew visitors from all over the world and from many disciplines. Pearson wrote no textbook or treatise but the visitors passed on what they had learnt when they returned home. In 1907 Pearson took over a research unit founded by Galton and reconstituted it as the Eugenics Laboratory. This laboratory researched human pedigrees but it also produced controversial reports on the role of inherited and environmental factors in tuberculosis, alcoholism, and insanity—great topics of the day. In 1911 a bequest from Galton enabled Pearson finally to give up responsibility for applied mathematics and become Professor of Eugenics and head of the Department of Applied Statistics. During World War One, normal activities virtually ceased but afterwards expansion was resumed. However, Pearson was no longer producing important new ideas and Ronald Fisher was supplanting him as the leading figure in statistics. Pearson retired in 1933 but he continued to write and, with his son Egon. S. Pearson, to edit *Biometrika*.

Pearson made strong friendships, notably with Weldon and Galton. He also thrived on controversy. Some of the encounters produced lasting bitterness, those with William Bateson and Ronald Fisher being particularly unforgiving. Following the 'rediscovery' of Mendel's work in 1900, Bateson questioned the point of investigating phenomenological regularities such as Galton's law of ancestral heredity. The biometricians replied that such regularities were established and that any theory must account for them. After Weldon's death in 1906, the biometricians were often miscast as mathematicians who knew no biology. The quarrel about the value of biometric research faded with Bateson's death and Pearson's withdrawal from research into inheritance. Fisher, like Bateson, was a Mendelian but there was no quarrel about the value of statistics. Fisher criticized the execution of Pearson's statistical work but he also considered himself a victim of Pearson's abuse of his commanding position in the subject.

2. Contributions to Statistics

Pearson was a tireless worker for whom empirical research and theoretical research went hand in hand. His empirical work was always serious. He wanted to make a contribution to knowledge in the field to which he was applying his methods not just to show that they 'worked,' but so he would set out to master the field. However, his contribution to statistical methodology overshadowed his contribution to any substantive field. His statistical contributions can be divided into ways of arranging numerical facts—univariate and multivariate—and methods of inference—estimation and testing.

3. Univariate Statistics: The Pearson Curves

Pearson's first major paper (1894) was on extracting normal components from 'heterogeneous material.' However, Pearson considered 'skew variation in homogeneous material' a more serious problem for he held that the normal distribution underlying the theory of errors was seldom appropriate. Pearson (1895) presented a system of frequency curves; these would be of use in 'many physical, economic and biological investigations.' Systems were also developed by Edgeworth, Charlier, Kapteyn, and others, but Pearson's was the most widely used. Other systems have been devised, inspired by his, but none has been as central to the statistical thought of its time.

From probabilistic considerations—which now seem contrived—Pearson obtained a basic differential equation governing the probability density function, or frequency curve, $y = y(x)$:

$$\frac{dy}{dx} = \frac{y(x+a)}{b_0 + b_1 x + b_2 x^2}$$

The constants determine the form of the curve. Pearson developed methods for choosing an appropriate curve and for estimating the constants. His laboratory produced tables to ease the burden of calculation.

While Pearson used these curves as data distributions, he assumed that the estimators of the constants would be approximately normal, certainly in large samples. In the work of 'Student' (W. S. Gosset) and then Fisher the data distributions are normal but the exact distributions of the sample statistics are usually not, though *they* may follow the Pearson curves. 'Student' did his distribution work in Pearson's laboratory but for many years Pearson thought that such 'small sample' work was off the main track and useful only to 'naughty brewers.'

In the Pearson system the normal distribution is just one type. In Fisher's analysis of variance which appeared in the 1920s it is the *only* type. People

schooled in the Pearson system did the earliest robustness studies and one of the continuing roles of the system has been in this area. While Pearson's skepticism towards normality pervades much modern work, the modern approach to inference is so different from his that he is not a direct inspiration.

4. *Multivariate Statistics: Correlation and Contingency*

In 1888 Galton introduced correlation and soon Edgeworth was developing the theory and Weldon was applying the method in his research on crustaceans. Pearson, however, produced the definitive formulation of normal correlation and had the vision that correlation would transform medical, psychological, and sociological research.

Pearson (1896) set out to frame hypotheses about inheritance in terms of the multivariate normal distribution but these were of less enduring interest than his account of correlation and multiple regression. The familiar partial and multiple correlation and regression coefficients appear at this time either in Pearson's work or that his assistant, G. Udny Yule. Soon Pearson and those around him were applying correlation in meteorology, sociology and demography as well as in biometry.

The 1896 ideas formed the basis of a definitive theory, especially when completed by the exact distribution theory of 'Student' and Fisher. But Pearson soon moved on to qualitative, non-normal, and time series data, though not with the same success. Pearson's scheme for contingency, the qualitative analogue of correlation, has latent continuous variables—surprisingly jointly normal ones. Yule, no longer working with him, rejected the scheme. The controversy that ensued had no clear resolution. Pearson toiled unsuccessfully to create a theory of skew surfaces, uniting correlation and skew curves. Time series correlation analysis was another disputed field. Pearson thought of the true correlation between series as obscured by time trends superimposed upon them. The scheme was widely used but in the 1920s Yule criticized it effectively, yet putting nothing in its place.

From the beginning Pearson encountered problems with correlation which have lived in the underworld of statistics occasionally surfacing in fields such as path analysis and structural modeling. Pearson formalized Galton's notion of correlation as measuring the extent to which variables are governed by 'common causes.' He was soon aware of cases of 'spurious correlation' where the natural interpretation is not sustainable. He and Yule—who developed the other standard interpretation of a direct causal relationship between the correlated variables—found numerous pathological cases. They devised treatments but developed no systematic way of thinking about the problem.

Pearson—following Galton—established multivariate analysis. Yet by a curious twist the multivariate data for which Pearson struggled to create a theory of frequency surfaces is often treated by Fisher's 'regression analysis' which was a development from Gauss's univariate theory of errors rather than Pearson's correlation theory.

5. *Estimation: The Method of Moments and Inverse Probability*

Pearson usually found the constants of frequency curves by the method of moments. He introduced this in his first major paper (1894) for he could find no other method for estimating a mixture of normals. Pearson went on to apply the method quite generally. It seemed particularly appropriate for the Pearson curves where the constants determine the first four moments. However, the method was put on the defensive after Fisher criticized it for not being 'efficient'—even when applied to the Pearson curves.

Probable errors (scaled standard errors corresponding to the 50 percent point) for method of moments estimators were derived by Pearson and Filon (1898) using a modification of an argument in the 1896 correlation paper. Pearson had obtained the product-moment estimate from the posterior distribution (based on a uniform prior) and its probable error from a large sample normal approximation to the posterior. In 1898 he assumed that the same technique would always yield probable errors for the method of moments. He seems to have dropped the technique when he realized that the formulae could be wrong but it was reshaped by Fisher into the large sample theory of maximum likelihood.

In 1896/8 Pearson used Bayesian techniques in a formalistic way but he could be a thoughtful Bayesian. Over a period of 30 years he wrote about the 'law of succession,' the standard Bayesian treatment of predicting the number of successes in future Bernoulli trials given the record of past outcomes—the fundamental problem of practical statistics' he called it. Pearson held that a prior must be based on experience; a uniform prior *cannot* just be assumed. Pearson pressed these views in estimation only once when he criticized Fisher's maximum likelihood as a Bayesian method based on an unsuitable prior—essentially Pearson's own method of 1896/8. Fisher was offended by this misinterpretation of his Bayes-free analysis.

There was plenty of improvisation in Pearson's estimation work; it was eclectic, even inconsistent. He was not the consistent Bayesian that Jeffreys would have liked. Pearson used sampling theory probable errors but he did not follow up the results on the properties of estimators in the old theory of errors. Pearson was not unusual in his generation; the theory of errors he found to hand-mixed frequentist and Bayesian arguments. Some of his contemporaries were

more acute but foundational clarity and integrity only became real values towards the end of his career when Fisher, Jeffreys, and Neyman set about eliminating the *wrong* kind of argument.

6. Testing: Chi-squared

Pearson's earliest concerns were with estimation—curve fitting and correlation—though there testing could be done, using the estimates and their probable errors and making the assumption of large sample normality. However, his chi-squared paper of 1900 introduced a new technique and gave testing a much bigger role.

The chi-squared test made curve fitting less subjective. One could now say: 'In 56 cases out of a hundred such trials we should on a random selection get more improbable results than we have done. Thus we may consider the fit remarkably good.' The tail area principle was adopted without any examination. Naturally Pearson enjoyed showing that the fit of the normal distribution to errors in astronomical measurements was *not* good.

Pearson brought chi-squared into his treatment of contingency and by 1916 he had established the main applications of the chi-squared distribution—a test for goodness of fit of distributions, a test for independence in contingency tables, a test for homogeneity of samples, a goodness of fit test for regression equations.

Pearson's distribution theory involved the exact distribution of the exponent in the multinormal density, this density appearing as a large sample approximation to the multinomial. Pearson slipped in the mathematics and used the same approximation whether the constants of the frequency curves are given *a priori* or estimated from the data. Fisher realized the importance of the difference and reconstructed the theory, finding that the chi-squared approximation was valid but the number of degrees of freedom had to be altered.

Fisher continued Pearson's chi-squared work and gave significance testing an even bigger role in applied statistics. Pearson's methods were more or less vindicated by the later work of Fisher and then of Neyman and E. S. Pearson but, as in estimation, his reasons did not pass critical examination.

7. Summary

Of the many statistical techniques Pearson devised, only a few remain in use today and though his ideas sometimes find re-expression in more sophisticated form, such as the correlation curve or the generalized method of moments, there is little to suggest that Pearson continues to directly inspire work in statistics.

Pearson broke with the theory of errors but in the next generation through the analysis of variance and regression the theory was restored to stand beside, even to overshadow, Pearsonian statistics. From a modern perspective Pearson's theory seems desperately superficial. Yet the problems he posed have retained their importance and the ground he claimed for the discipline of statistics has not been given up. Pearson was an extraordinarily prolific author and there is also a considerable secondary literature. There is a brief guide to this literature on the website <http://www.soton.ac.uk/~jcol>

See also: Multivariate Analysis: Overview; Probability: Formal; Significance, Tests of; Statistics, History of

Bibliography

- Eisenhart C 1974 Karl Pearson. *Dictionary of Scientific Biography* **10**: 447–73
- Hald A 1998 *A History of Mathematical Statistics from 1750 to 1930*. Wiley, New York
- MacKenzie D A 1981 *Statistics in Britain 1865–1930: The Social Construction of Scientific Knowledge*. Edinburgh University Press, Edinburgh, UK
- Magnello M E 1999 The non-correlation of biometrics and eugenics: Rival forms of laboratory work in Karl Pearson's career at University College London, (in two parts). *History of Science* **37**: 79–106, 123–150
- Morant G M 1939 *A Bibliography of the Statistical and other Writings of Karl Pearson*. Cambridge University Press, Cambridge, UK
- Pearson E S 1936/8 Karl Pearson: An appreciation of some aspects of his life and work, in two parts. *Biometrika* **28**: 193–257; **29**: 161–247
- Pearson K 1888 *The Ethic of Freethought*. Fisher Unwin, London
- Pearson K 1892 *The Grammar of Science* [further editions in 1900 and 1911]. Walter Scott, London [A. & C Black, London]
- Pearson K 1894 Contributions to the mathematical theory of evolution. *Philosophical Transactions of the Royal Society A*. **185**: 71–110
- Pearson K 1896 Mathematical contributions to the theory of evolution. III. Regression, heredity and panmixia. *Philosophical Transactions of the Royal Society A*. **187**: 253–318
- Pearson K, Filon L M G 1898 Mathematical contributions to the theory of evolution IV. On the probable errors of frequency constants and on the influence of random selection on variation and correlation. *Philosophical Transactions of the Royal Society A*. **191**: 229–311
- Pearson K, Filon L M G 1900 On the criterion that a given system of deviations from the probable in the case of correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *Philosophical Magazine*. **50**: 157–75
- Pearson K, Filon L M G 1914/24/30 *The Life, Letters and Labours of Francis Galton*. Cambridge University Press, Cambridge, UK, Vols. I, II, IIIA, IIIB
- Provine W B 1971 *The Origins of Theoretical Population Genetics*. Chicago University Press, Chicago

Semmel B 1960 *Imperialism and Social Reform: English Social-Imperial Thought 1895–1914*. George Allen & Unwin, London
Stigler S M 1986 *The History of Statistics: The Measurement of Uncertainty before 1900*. Belknap Press of Harvard University Press, Cambridge, MA

J. Aldrich

Peasants and Rural Societies in History (Agricultural History)

The concept 'rural societies' is based on the fundamental division in our society between rural and urban, but usually refers to societies dominated by rural areas. It thus has a historical dimension, and is connected with the concept of peasant societies.

1. Terminology

An intense discussion about the concept of peasant studies began in the 1950s. A large body of literature was published in the decades that followed. The discussion centered on the Anglo Saxon world since, unlike English, most other languages do not contain the same clear distinction between peasant and farmer. In German, for example, *Bauer* can mean both peasant and farmer (Rösener 1993). The need of a distinguishing term in English is also connected with the fact that peasant in English (like *paysans* in French) has a derogatory meaning which the German word *Bauer* (or the Scandinavian *bonde*) does not have.

Several contributions in the discussion about peasants and farmers were made from the 1950s until the beginning of the 1970s (important monographs: Redfield 1956, Wolf 1966, and readers: Potter et al. 1967, Shanin 1971). In recent decades, the concept of peasant has been questioned even in the Anglo Saxon world. For example, Tom Scott writes in an overview that the category of peasant 'is so broad as to become not so much an archetype as a stereotype' (Scott 1998, p. 1). However, the discussion which began during the post-war period, still forms a platform for the research on peasant and rural societies. Peasant society was defined on two levels; partly chronologically in relation to earlier and later types of social systems, and partly socially where the peasant society was defined in relation to the overall social structure of which it was a part.

In the case of the last-mentioned level of definition, the peasant society was seen as a part-society with a part-culture, i.e., a part of the large social structure. This formed the basis of the chronological separation of peasants from what was called the 'tribal society' (or even, but not any longer, the 'primitive society').

Such a society was by definition an earlier system, not being a part of a larger social unit. The emergence of the town was identified as a decisive change and resulted in peasants either choosing or being forced to send some of their produce to the town. The establishment of the State and the state authorities has also been identified as a distinguishing factor. With the advent of the State, the peasant society became clearly subordinate to, but also alienated from, a larger social system as such. To some extent, both these criteria can be said to represent the same change. The peasant society has also been described as a part of an overall socioeconomic feudal system, but every peasant society can hardly be seen as part of a feudal system (see *Feudalism*).

At the beginning of the 1970s, Teodor Shanin (in his *Introduction* to Shanin 1971) summarized the discussion and pointed at some of the characteristics that distinguished a typical peasant society: the family farm as the basic unit of multidimensional social organization; land husbandry as the main means of livelihood directly providing for the major part of the consumption needs; specific traditional culture related to the way of life of small communities; and the underdog position, specified as the domination of peasantry by outsiders. Shanin adds that, like every social entity, the peasantry must be seen as undergoing change as a part of a social process, which influences the definition of the concept peasant.

The chronological border, which separates peasants from the modern farmer has also been discussed. The modern farmer not only produces for a market, but all his actions are profit driven. He acts in the same way as other businessmen. The business itself has become increasingly industrial in nature. Different aspects were emphasised in the discussion. Wolf points to the relationship to the market as being decisive. Redfield emphasizes peasant tradition. The Marxists, who later entered the discussion, wrote that when the peasants acted as a group with class consciousness, they became a class in themselves. The Russian peasant researcher Chayanov, had been purged in 1930 by Stalin but his writings from the 1920s, which got a renaissance long after his death, emphasized the preservation of the family as being of central importance to the peasant (Chayanov 1966).

The discussion illustrates the complexity of the concept, and much of the discussion focused on the problem of definition. Could the specific African cultivator be categorized as peasant, or the Latin-American pastoralist? The fruitfulness of discussing such detailed classifications is questionable. When the core of the term has been identified, the borders drawn must be flexible to allow for interpretations, which in themselves include a dynamics. During the discussion, it also became increasingly clear that the specific historical situation must be taken into account. Rodney Hilton pointed at the essential elements in defining the English medieval peasant, but also made

clear that this definition was specific for a part of Europe, where the peasantry must be seen in the context of the medieval feudal society (Hilton 1975).

The concept of peasant will be used in this article in a very comprehensive sense, which includes those who raise crops or cattle, on a partly nonprofitable base, and are integrated in larger social structures. The definition includes peasantries in all societies where the production of foodstuffs requires the major part of the society's labor. It is debatable whether there could be peasants in societies where cultivators comprise a diminishing share of the population. Today in most countries outside the Western World, peasants continue to be tied to the soil and to the local community, with family, relatives and traditions, in a way that characterizes them as peasants. Also in a large part of Europe, there are still cultivators who live as traditional peasants. They live surrounded by their ancestors' history in houses, land, and tales and they are bound to their own farm by family-traditions. This should not be underestimated as a fact that for instance decides the course of contemporary talks on agriculture and world trade between Europe and the US.

Accordingly, the term peasant also has some relevance in today's society, and this is the explanation of why many languages have not developed the terminological difference found in English between peasant and farmer. However, the difference between peasant and farmer identifies an important historical change, and the peasant society is a purely historical concept for the Western World. Rural society should be regarded as a broader unit comprising, in addition to cultivators, craftsmen and rural industry. Rural society thus contains the seeds of proto-industrialism, and the concept is more relevant to today's conditions.

2. Historiography

To understand the discussion about the peasant and rural societies, it must be placed in a historiography, which both precedes and follows the postwar discussion. Rural society as a historical concept must be related to agricultural history, which mainly concerns agricultural production, but also includes other elements related to agriculture, such as politics and village communities. In rural history, the emphasis, formally speaking, is placed more on the societal aspect and in the case of agricultural history, more on the history of production and distribution. But since agricultural production is the basic activity of all peasants, in practice, rural history and agricultural history are parts of the same historiography. Rural history and agricultural history are an important part of general history, as the farming population for centuries was in the majority, but despite this, rural history and agricultural history emerged as specific disciplines relatively late.

2.1 The First Folk-life Studies

Scientific research into peasants and rural history in Europe began at the end of the nineteenth century. In many parts of Europe, i.e., Eastern Europe and Scandinavia, interest focused on the countries' own peasants. In the UK, research primarily focused on peasants in the colonies. In countries where agriculture played a dominating role, agricultural history was linked to the creation of national identity, demonstrated in national 'agricultural museums.' In Hungary, what is still one of the world's most important agricultural museums was founded in conjunction with the country's millennial celebration in 1898. National agricultural museums were also established in the Czech Republic and Denmark at the end of the nineteenth century.

In Eastern Europe, Scandinavia, and Germany, folk-life studies focused on the peasants who were regarded as the real heart of the nation. Later, these studies also played a prominent role under the name of ethnology or ethnography in France, Portugal, Ireland, Scotland, and other countries, but never in England. Folk-life research studied both spiritual and material culture, and described tools and houses together with folk-songs and riddles in peasant societies. A large knowledge-base was built up.

The scientific question that crystallized concerned regionality. Even if the original question was couched in terms of national identity, the answer increasingly proved the opposite. The cultural phenomena moved freely across borders, there were no obvious and given national borders in Europe. The Swede Sigurd Erixon played an important role, and a number of national ethnographic atlases, where the focus was on peasant society and its manifestations, were published. There were plans to publish a common European ethnographic atlas, but today these plans have been put on ice.

In the countries with colonies, above all the UK, the study of peasant societies outside Europe became increasingly popular at the end of the nineteenth century. Originally, these studies were part of the colonial administration, e.g., in the genre of studies of the Indian countryside, which the British Empire created as early as in the middle of the nineteenth century. Eventually social anthropology became a discipline in its own right, less and less connected with the colonial administration. In the twenty-first century it has a strong position in the Anglo Saxon world, not least in the US. Like European ethnography, anthropology was linked to the museums and the decades around the turn of the century were spent gathering data.

The goal of these studies was to understand the local community, in all its details but also its internal structure. The focus was on the individual village or region as an isolated entity. The level of detail could also be very high in the case of the material culture, but

the focus was often on the relationships between the people, e.g., family relationships. But even though the research concentrated on clarifying the internal structure of the local society, the studies gradually pointed to a different result. It was increasingly realized that the local society was integrated with, and subordinate to, larger social structures. This was one of the most important reasons why parts of the Anglo Saxon world of anthropology started using the term 'peasant' instead of 'people,' and it led to the above mentioned debate on peasant societies.

Early on, the combination of ethnology and anthropology resulted in impressive surveys. The German researcher Paul Leser, for example, published an authoritative survey in 1931 of different types of ploughs in different parts of the world (*Entstehung und Verbreitung des Pfluges*, Munster). Not long afterwards, Leser, who was Jewish, was forced to emigrate. Folk-life studies in Germany increasingly became a part of the nationalistic ideology in the 1930s, and was concentrated on studies about the spirit and 'soul' of the people.

2.2 Agricultural History in the Interwar Period

The US was still regarded by its inhabitants as an agricultural nation into the twentieth century. In the democratic and popular tradition that was this nation's identity, the history of agriculture played a role in the explanation of its progress and success, and agricultural history crystallized early into a discipline in its own right. The oldest national journal on the subject, *Agricultural History* (1927–), was founded in the US. From the very beginning, this journal contained articles about conditions outside the US, but it has focused mainly on American agricultural history. Initially, many of the articles discussed progress in agriculture at the end of the nineteenth century and the beginning of the twentieth century. A special genre consisted of articles about diaries kept by peasants.

In Europe, the first wave of interest in the history of agriculture, in ethnography and anthropology, focused on the immediate past. There were few attempts made to create a broader time perspective, as the discipline of history was still oriented towards the history of the State and the elite. Rural history developed in France under the influence of geography, where a dominating group of scholars favored all-embracing studies where every aspect of a region should be described, including historical aspects. With geography's starting-points in the land, these studies also touched on rural history.

The historian who succeeded in giving rural history a scientific basis was the Frenchman Marc Bloch in his *Les caractères originaux de l'histoire rurale française*, which was published in 1931 (Bloch 1966). This was based on a series of lectures he had held in Oslo in 1929 at a research institute, which had been established to promote and carry on Scandinavian folk-life

studies. Here the different threads meet. In the words of the American agrarian historian Richard Herr, Bloch was able to provide 'the kind of insights that would spark the imagination of other historians' (Herr 1993, pp. 4–5). In a subsequent study of European feudalism, Bloch explained how a rural society is part of a large social structure (Bloch 1961, in French 1939–40). He thus anticipated the discussion about peasant societies during the postwar period.

Bloch's work still provides inspiration and guidelines in a number of areas: the role of the village community, peasant uprisings, the connections between technology and social change, etc. But years passed before his books were considered to be important. General acceptance did not come until after World War II. In part, this was a result of the myths created about Bloch after he had joined the French Resistance to fight the Germans and was executed in 1944.

2.3 Agricultural History's Breakthrough after World War II

The reason for the rise in appreciation of Bloch's work should, however, be explained in terms of the large changes that occurred in the study of rural societies and agricultural history. During the 10-year period following the end of World War II agricultural history became a discipline in its own right in Western Europe. Departments of Agricultural History were established in several European countries: in the Netherlands at the agricultural university in *Wageningen*, in Germany at the agricultural university in *Stuttgart* and in the UK at the Faculty of Agriculture at *Reading University*.

National journals were started up in Europe: *The Agricultural History Review* (1953–) in the UK, *Historia Agrariae* (1953–) in the Netherlands, *Zeitschrift für Agrargeschichte und Agrarsoziologie* (ZAA) (1953–) in Germany, *Bol og by* (1956–) in Denmark, *Rivista di storia dell' agricoltura* (1960–) in Italy and *Études rurales* (1961–) in France. In time, the breakthrough for agricultural history resulted in national synthesis covering several volumes. In the UK, a huge project in eight volumes was started in 1967, *The Agrarian History of England and Wales*. This work has yet to be completed. The German synthesis, *Deutsche Agrargeschichte*, consists of six volumes and was published between 1962 and 1984. In France, *Histoire de la France rurale* was published in four volumes between 1975 and 1978. Although the French volumes emphasise rural history in the title, there is no major difference between this and the other national works. In the German, two volumes are devoted to purely social questions, i.e., legal questions and peasants as a social category.

This wave of research resulted in a focus on the long-term change, and not least on the Middle Ages. At the beginning of the 1960s, two important works

were published. One of these was Georges Duby's *L'économie rurale et la vie des campagnes dans l'Occident médiéval*, published in 1962, which was a study in the tradition of Bloch (in English 1968). The following year, 1963, saw the publication of an English translation of a study published in Dutch some years earlier. Slicher van Bath's *The Agrarian History of Western Europe AD 500–1850*.

During the postwar period the agricultural crisis in the late Middle Ages became a central field of research. This economic decline had been observed before World War II (i.e., for Germany, Abel 1980). The discovery was theoretically significant since it showed that the development of production had not always been positive. Another important theoretical change was that agriculture was seen as a dynamic sector in medieval Europe, which partly rejected earlier theories about the countryside as having a restraining effect on economic development. The aftermath of World War II continued to leave its mark to some extent. Slicher van Bath, who had built up the Department of Agrarian History at the University of Agricultural Sciences in Wageningen in the Netherlands, mainly collaborated with Anglo Saxon researchers and not with the German, e.g., at the agricultural university in Stuttgart-Hohenheim. The reason for this was that Gunter Franz, who had built up the Department of Agricultural History in Stuttgart, was not only a leading expert on the German Peasant War, but also had been an active Nazi. However, the journal founded by Franz, *ZAA*, published early on historiographical surveys of agricultural history in Europe. These surveys were written by Duby, Postan, and other leading European agrarian historians. It must also be stated that several leading agrarian historians in Germany during the 1930s never became engaged in the Nazi movement, for instance, Wilhelm Abel and Friedrich Lütge.

2.4 The Transformation of Folk-life Studies after World War II

The discussion about the concept of 'peasant' resulted in a number of studies of peasant and rural societies. These include the classic studies made by Oscar Lewis in Mexico and India. Daniel Thorner's studies of peasants in India are another example (a bibliography in: *Peasants in History: Essays in Honour of Daniel Thorner*, Hobsbawm (ed.) 1980). Several studies were also carried out in Europe, often by American researchers who used anthropological methods to study villages in Southern Europe. They were somewhat surprised to meet an entirely different type of cultivator from those they usually met in the US. It was still peasants and not farmers who dominate the European villages. The classic studies here include those by Julian A. Pitt-Rivers, who studied a Spanish village (1954), and by Joel Halpern, who studied a Serb village (1958).

In anthropological research, the large databases, e.g., the Human Relations Area Field (HRAF) in Yale has been gathering information on several hundred cultures since the 1930s. But the previous focus in anthropological research on material cultural products has weakened. True, encyclopedic studies were still being made, for example, by the Frenchman Leroi-Gourhan (1943–5), but this was more and more pushed into the background. Instead, it was the social structure that attracted increasing attention, and this is noticeable not least in the discussion about peasant societies. It was still the hidden patterns that research wanted to be clarified, but now as part-cultures subordinate to the principal culture. Clifford Geertz, had made an important description of the history of agrarian technology in Indonesia (*Agricultural Revolution* 1964), became even more influential as a father of the concept 'thick description,' which became a leading method in anthropology (see articles in Geertz 1973).

In Eastern Europe, folk-life studies retained and, to some extent, strengthened their role when the Socialist people's republics were established. Research into 'the people' was emphasized as one of the main tasks of historical research. At the same time, contacts with the West were broken off. An example is how seldom the German researchers in the East and West referred to each other, even though they often made research about the same subjects.

Some contacts were however upheld. *Association International des Musées d'Agriculture (AIMA)* gradually became the meeting place between East and West for the study of rural society. AIMA was an association of agricultural museums all over Europe as well as in the US, which was formally under the jurisdiction of UNESCO. The association was formed in 1974 and its meetings were held alternately in Eastern and Western Europe (or the US). Many of the questions discussed were classic in European ethnography, e.g., the history of the plow. After the Communist collapse, this association has had a difficult time.

2.5 Reasons for the Breakthrough after World War II

There are many underlying reasons for the breakthrough of rural history in the 1950s. In the Netherlands, one of the reasons given for the establishment of a Department of Agricultural History was that the severe food shortage at the end of World War II illustrated the importance of agriculture. In all social sciences the American influence increased after World War II, not the least in sociology of agriculture and in studies of non-European peasants. This influence may also have had an impact on agricultural history.

A factor, which can hardly be underestimated, was the increased political importance of peasants on a worldwide scale and in conjunction with colonies

attaining independence, but the immediate conclusion was not that scholars regarded the peasants as decisive actors in history.

Peasants and rural societies were identified as essential elements of the social structure. Initially, the new research focused on concepts and ideas from the earliest period of folk-life studies in Europe, and emphasized that peasant culture had filtered down to the peasants from the elite. Villagers had taken over and simplified traditions from the towns or from the elite. The goal of the peasants was to imitate the leaders of society. The ideological basis for this interpretation was that society's leaders were considered to represent a higher form of civilization. The interpretation was made in more or less explicit forms by almost all the scholars participating in the discussion after World War II and in particular by George Foster (1967). Robert Redfield (1956) distinguishes between a great tradition and a little tradition, between the classical, learned elite culture, and the peasant tradition. He emphasizes that a mutual exchange takes place, where also the elite draw inspiration from below.

The view of the conservative peasant society which was unwilling to accept changes that upset traditions is similar to what is sometimes called 'the dumb peasant theory.' This implied that peasants had to be forced to develop, and this assumption guided much of the Western World's relations with the Third World throughout the 1950s and 1960s.

2.6 New Interpretations in the 1970s and 1980s

Beginning in the 1970s, studies of peasants and rural societies increased in number and became more diversified. Two partly connected phenomena began to change the picture. One was the Marxist challenge in the academic world, the other was the successful peasant uprisings in different parts of the world, not least the civil war in Vietnam. These two phenomena challenged the old thought structures, but did not lead to a new hegemony. Instead, after the major theoretical battles in the 1970s, the field was left open to a large number of different theories. This can be compared with the change undergone by rural sociology at the same time. A survey from 1990 shows that what the authors call 'the new rural sociology' from the 1970s and 1980s was characterized by a large number of interpretations (Buttel et al. 1990).

The common question with which many of the new interpretation struggled was the relationship between actors and structure, where the peasants were increasingly regarded as real actors in the historical process. One expression of this new and deeper interest in peasant studies was the *Journal of Peasant Studies* (1974–). In its editorial policy, it was emphasized that the journal would be concerned with political economy

and that it would attempt to understand the role of the peasants, especially outside Europe and North America.

At the same time we also see a growing interest for rural history in the third world, and for instance, a journal on the subject was founded in India: *History of Agriculture* (1973–).

The role of class conflicts became more important, which at times resulted in very detailed analyses of the social structure of the countryside. Studies of peasant revolts and peasant uprisings were developed into a special genre with a series of studies. Searches in international databases show that there is an interesting geographical distribution of the studies. Most of the studies of peasant revolts outside Europe concern Asia (India, China, Japan, South East Asia, and Indonesia) or Latin America (Mexico, El Salvador, Bolivia, and Brazil). Studies concerning Africa, on the other hand, began early on to be more focused on the environment, while studies of peasant revolts are less common.

An example of this new orientation is the rapidly growing interest in 'Maroon societies.' The fact that slaves in Latin America as well as in the American South were able to break free and establish societies, which in some cases were able to survive for hundreds of years partly outside the control of the authorities, added a new dimension to peasant society. The first comprehensive study of Maroon societies as a specific form of peasant societies in 1970s was followed by many more (Prize 1996). This history from below was also applied in other areas.

In the European research about long-term changes new interpretations emerged. The dominating researchers had initially been supporters of Malthusian interpretations, but other interpretations gradually began to be given greater emphasis. One example of this is the class perspective introduced by Guy Bois (1976) into the interpretation of the agrarian crisis in the late Middle Ages.

The older type of studies of material culture continued, not least in the exchange between East European and West European researchers in classic folk-life research. In Denmark, the journal *Tools & Tillage* was established in 1971 and became the leading international journal dealing with the development of agriculture technology in peasant societies.

2.7 Diversification at the End of the Century

The late twentieth century has seen a gradually growing interest in rural and agricultural history. New journals have been started up, including *Rural History* (1990–) in the UK, *Noticiario de historia agraria* (1991–) in Spain, *Skrifter om skogs- och lantbruks-historia* (1991–) in Sweden, and *Histoire et sociétés rurales* (1994–) in France.

There are new elements in this growing interest. One that is very important is the interest in environmental questions. This has had an impact on agricultural history not least in the US in the form of studies comparing how Indians and Europeans used the environment in a social context (i.e., Merchant 1989). Specialist journals have also been established, the most important of which is *Environmental History* (1989–) in the US, formed by merging two journals with the same orientation, which had been published since the 1970s. The issues dealt with have concerned sustainability in a long-term perspective in relation to current problems such as soil erosion, salinization, etc.

Interest in popular culture and social structure is increasing, and one example is the research group around Jan Peters (Peters 1997). This group also represents a successful wedding between East and West in Germany in the late twentieth century. Another strong orientation, although not only in the study of rural and peasant societies, is gender studies. Technology research has also been revived although here, more specific questions are being raised concerning the connection between technological and social change.

3. Theories and Methods—a Summary

When the question of the definition of peasants and peasant societies was raised after World War II, it proved to be both fruitful as well as inspirational to further research. The connection between peasants and social structure was brought out more clearly. However, this was only one step in a long development process involving research, and rural and agricultural history. At the end of the nineteenth century and beginning of the twentieth century, researchers concentrated on gathering material. In Europe, individual cultural elements were registered, and outside Europe whole peoples and their characteristics were dealt with. Theories about isolated nations and communities were proposed, and abandoned.

In the 1950s larger structures were identified. In anthropology, peasant societies were regarded as being a part of larger society, and in history the long waves came into focus of the research. During a following phase, which began in the 1970s, research became increasingly diversified. Much of the research emphasized the capacity of peasants to take independent action, and new fields of research, such as environmental questions and gender research, were opened up. This eventually resulted in a new wave of interest about the subject at the turn of the twenty-first century.

See also: Agricultural Change Theory; Agricultural Sciences and Technology; Agriculture, Economics of; Economic History; Feudalism; Food in Anthropology; Food Production, Origins of; Food Security; History of Technology; Peasants and Rural Societies

in History (Agricultural History); Peasants in Anthropology; Rural Geography; Rural Sociology; Social History

Bibliography

- Abel W E 1980 *Agricultural Fluctuations in Europe: From the Thirteenth to the Twentieth Centuries*. Methuen, London
- Bois G 1976 *Crise du féodalisme*. Presses de la Fondation National des Sciences Politiques, Paris
- Bloch M 1961 *Feudal Society*. Routledge & Kegan Paul, London
- Bloch M 1966 *French Rural History*. Routledge & Kegan Paul, London
- Buttel F H, Larson O F, Gillespie G W 1990 *The Sociology of Agriculture*. Greenwood Press, New York
- Chayanov A V 1966 The theory of the peasant economy. In: Thorner D, Kerblay B, Smith R E F (eds.). American Economic Association Translation Series, Homewood, IL
- Duby G 1962 *L'économie rurale et la vie des campagnes dans l'occident médiéval*. Aubier, Editions Montaigne, Paris
- Foster G 1967 What is a peasant. In: Potter J M, Diaz M N, Foster G M (eds.) *Peasant Society, A Reader*. Little, Brown, Boston, MA
- Geertz C 1973 *The Interpretation of Cultures: Selected Essays*. Basic Books, New York
- Halpern J M 1958 *A Serbian Village*. Harper Colophon Books, New York
- Herr R 1993 The nature of rural history. In: Herr R (ed.) *Themes in Rural History of the Western World*. Iowa State University Press, Ames, IA
- Hilton R H 1975 *The English Peasantry in the Later Middle Ages*. Clarendon Press, Oxford, UK
- Hobsbawm E J (ed.) 1980 *Peasants in History: Essays in Honour of Daniel Thorner*. Oxford University Press, Calcutta
- Leroi-Gourhan A 1943–45 *Évolution et techniques* 1–2. Paris
- Leser P 1931 *Entstehung und Verbreitung des Pfluges*. Anschendorffsche Verlags buchhandlung, Münster
- Merchant C 1989 *Ecological Revolutions: Nature, Gender, and Science in New England*. University of North Carolina Press, Chapel Hill, NC
- Peters J (ed.) 1997 *Gutsherrschafsgesellschaften im europäischen Vergleich*. Akademie, Berlin
- Pitt-Rivers J A 1954 *People of the Sierra*. University of Chicago Press, Chicago
- Potter J M, Diaz M N, Foster G M (eds.) 1967 *Peasant Society, A Reader*. Little, Brown, Boston, MA
- Prize R 1996 Preface. In: Prize R (ed.) *Maroon Societies: Rebel Slave Communities in the Americas*. Johns Hopkins University Press, Baltimore, MD
- Redfield R 1956 *Peasant Society and Culture: An Anthropological Approach to Civilization*. The University of Chicago Press, Chicago
- Rösener W 1993 *Die Bauern in der europäischen Geschichte*. C H Beck, Munich, Germany
- Scott T 1998 Introduction. In: Scott J (ed.) *The Peasantries of Europe from the Fourteenth to the Eighteenth Centuries*. Longman, London
- Shanin T (ed.) 1971 *Peasants and Peasant Societies*. Penguin, Harmondsworth, UK
- Slicher van Bath B H 1963 *The Agrarian History of Western Europe AD 500–1850*. Edward Arnold Limited, London
- Wolf E 1966 *Peasants*. Prentice-Hall, Englewood Cliffs, NJ

J. Myrdal

Peasants in Anthropology

Peasants are small-scale agriculturalists who use simple technology such as hand tools, human labor, animal power, and possibly light machinery. The household is the basic unit of production and consumption. While some peasants grow crops or raise animals for markets, they usually consume most of what they produce and thus differ from farmers who produce primarily cash crops. Also, peasants typically exist in subordinate political and economic relationships with other sectors of society who receive surplus products and labor from them. From the rise of civilizations based on non-mechanized agriculture until the widespread use of modern industrial farming, peasants in rural areas have provided much of the food and other basic material resources consumed in urban society. Furthermore, peasants have provided recruits for armies throughout history, and until fairly recent times in developed nations, cities typically had high death rates relative to birth rates and thus had to be replenished by peasants migrating from rural areas. Indeed, most of the rapid growth in world population in the nineteenth and twentieth centuries was due to continued high birth rates among peasants, and correspondingly lowered death rates due to the spread of modern medicine and public health.

1. *Peasants in the History of Anthropology*

From its modern beginnings in the 1800s until the early to mid-twentieth century, anthropology focused largely on foraging, tribal, and chiefdom societies. It was not until after World War II that peasant communities became a major object of anthropological research. This shift of interest by anthropologists toward peasant communities was due to changing global conditions. Foraging and tribal societies were disappearing at ever increasing rates, and in addition to this, following the War, peasants emerged on the world stage as major political actors in the context of the Cold War. Within this global struggle between the so-called 'First World' of capitalist societies and the 'Second World' dominated by the communist Soviet Union and China, many underdeveloped 'Third World' countries were inhabited mainly by rapidly growing peasant populations. The political significance of these nations grew as concern mounted as to whether their peasant populations would veer toward communism or capitalism. Since the poverty of peasant communities was seen as a major inducement for them to turn to communism, promotion of economic development in rural areas of the 'Third World' became a priority of US foreign policy. In this geopolitical context, the anthropology of peasant communities flourished from the 1950s until approaching the end of the Cold War in the early 1990s

(Kearney 1996). Much of this work was applied anthropology concerned with the improvement of agriculture, health, and education in peasant communities.

2. *Anthropological Approaches to Peasants*

A distinctive feature of anthropology is that it seeks to study and comprehend all of the major components of a community and how they are integrated. Ideally, this comprehensive method begins with examination of the physical environment of the community and how its inhabitants adapt to that environment with technology and a distinctive division of labor with which they produce food, shelter, and amenities. These features of the environment—technology, labor, and the material products of labor—are parts of a community's infrastructure, upon which are based an immaterial superstructure that comprises a corresponding patterning of social organization, politics, class positions and relations, gender identities, religion, world view, etc. Ideally, a robust anthropological treatment of a peasant community examines all of these material and immaterial components of a community and how they are interrelated. Furthermore, since peasants, by definition, typically exist within and are shaped by relations of unequal economic exchange with non-peasant communities, a complete anthropological study of a peasant community would also examine this dimension. In practice, however, few anthropologists attain such comprehensiveness. Instead, most anthropologists who study peasant societies tend to concentrate on either superstructural or infrastructural features of a local community, or seek a more comprehensive treatment by focusing on selected features of each. Similarly, anthropological studies vary considerably in the degrees to which and the ways in which they consider how peasant communities are related to non-peasants. These differing approaches and foci can be illustrated in the work of four contrasting anthropologists.

2.1 *Redfield and the 'Folk–Urban Continuum'*

According to Robert Redfield (1941), peasant communities are a kind of folk society that exists on a 'folk–urban continuum,' which has both geographic and historic dimensions. Cities represent the modern urban end of the continuum and small, isolated non-agrarian indigenous societies are the extreme traditional folk end, with peasant communities near the traditional end. Redfield saw the history of traditional societies as shaped mainly by the spread of modern features of technology, social organization, family, kinship, values, and world view outward from cities at

the urban end of the continuum toward the folk end, in a process of modernization or development. This diffusion of the traits of modernity, especially modern values and world view, would proceed faster were it not for barriers to their acceptance in the traditional culture. This model became important in programs of applied anthropology that sought to identify and overcome cultural barriers to modernization in peasant communities, which were defined as underdeveloped, that is, as waiting to shed their traditional cultures by becoming fully incorporated into the modern national culture, economy, and political system of their nation. This approach to economic, social, and political development in peasant communities became important in American applied anthropology during the Cold War as an alternative to socialist and communist paths of development.

2.2 Foster and the 'Image of Limited Good'

Based on long-term intensive fieldwork on the economics, social organization, and culture of a Mexican peasant community, George Foster (1979) developed a model of world view to explain peasant economic and social behavior. This model, 'The Image of Limited Good', in the superstructure of peasant society, is based on economic realities in the infrastructure. Foster notes that land, markets, employment, and other sources of income exist in limited amounts in local peasant economies and therefore there is great competition for such scarce resources and economic opportunities. There are rarely enough of the basic material resources to satisfy everyone's needs and wants. In this situation, peasant ethics and morality are based on the idea that the best accommodation to this situation of Limited Good is for all members of the community to have an equal share of different forms of Good, e.g., food, wealth, affection. Therefore, if someone gets more than their fair share of some form of Good, then, according to the logic of Limited Good thinking, someone else must get less than their fair share. Foster says that this realistic perception of material realities shapes peasant world view in general, which is expressed as envy, fatalism, individualism, fear of witchcraft, and also principles of folk medicine, proverbs, and basic features of social organization and economic behavior, all of which are barriers to development.

2.3 Wolf and the 'Closed Corporate Peasant Community'

Whereas anthropologists working with a modernization perspective, such as Redfield and Foster, focused on barriers in traditional peasant society to the acceptance of modern cultural traits coming from developed urban areas, other anthropologists working

with Marxist concepts examined the opposite process, that is, how unequal market relations, cheaply remunerated labor, interest payments, taxes, and tribute tend to drain economic value from peasant communities. Working with this perspective, Eric Wolf (1966) developed the structural model of the 'closed corporate peasant community' to demonstrate how, contrary to Redfield, peasants were not isolated from urban society, but instead formed much of its economic base. He explored mechanisms of value extraction and how peasants attempted to minimize it by seeking to 'close themselves off' and defend themselves from exploitive outsiders. Wolf notes that the household organization of production, and the culture of envy and suspicion that Foster and others describe, tend to promote individualism and a lack of community solidarity. These traits led Karl Marx to characterize peasants as like 'potatoes in a sack,' that is, unable to organize in their own self interests. But anthropologists like Wolf who were working with a political economy orientation have also been interested in the revolutionary potential of peasants. Indeed, in the twentieth century, some of the major armed conflicts such as the Mexican, Russian, and Chinese Revolutions, and the Vietnam War, have been characterized as peasant wars (Wolf 1969) in which peasant communities sought to gain political and economic independence from non-peasant exploiters.

2.4 Meillassoux and 'The Articulation of Modes of Production'

Since the 1960s anthropologists have become increasingly concerned with migration from peasant communities. Claude Meillassoux (1981), who, like Wolf, also worked in the Marxist tradition, examined how, through circular migration, rural people who were unable to make a complete living as peasants could supplement their income by migrating to seek temporary work as proletarians, thus articulating the peasant and the capitalist modes of production. This situation is seen as a benefit to the modern sector because it shifts costs for the biological and social reproduction of workers to peasant communities. Also, because such peasant workers produce part of their own food and other necessities on their own land, they can accept lower wages than fully proletarianized workers, who need more cash to buy their basic necessities. Thus, whereas the modernization approach of Redfield, Foster, and others predicted that migration would hasten the development of peasant communities, the articulation perspective revealed how such partial proletarianization in some ways perpetuates peasant ways of life in the countryside. Nagengast (1991) shows how this was true even in a socialist country, such as communist Poland, that had strong national policies for full proletarianization and

modernization of the countryside. Such conditions of partial proletarianization raise questions about the basic class nature of such peasant workers and worker peasants, and what are the most appropriate political projects to defend their class interests.

3. Economic Development, Global Change, and The Future of Peasants

Applied anthropology in the Redfield and Foster tradition sees peasant social organization, conservative culture, and world view as barriers to the acceptance of the social and cultural traits of modernity that are essential to economic development. Foster argues that the economic realities that peasants face dispose them to be skeptical and fatalistic about possibilities for personal and especially cooperative efforts to overcome their poverty. According to this analysis, the role of applied anthropology is to understand these social and cultural dynamics of peasant communities and demonstrate alternatives to them. In contrast, applied anthropology in the Marxist tradition, as exemplified by the work of Wolf and Meillassoux, and also by dependency theory and world system theory, pays more attention to structural conditions that keep peasants in politically and economically subordinate positions so that surplus can be extracted from them and transferred to other sectors of the national and world economy, thus maintaining peasants in conditions of *de-development*. Accordingly, applied anthropology for peasant communities in this tradition is concerned with ending such unequal exchange so that de-developed nations and their peasant communities can retain more of their wealth for their own development.

Currently, increased migration between rural communities and cities and across national borders, and the social networks and complex livelihoods that result from it, have largely obliterated the cultural, social, and economic distinctions between rural and urban areas upon which the persistence of peasants depend. Also, as the rates increase at which supposed peasants migrate in and out of a variety of economic niches—ranging from subsistence farming, to wage labor, to the informal economy, etc.—so do the volume, velocity, and diversity of commodities and information that they consume also increase. These demographic, occupational, and cultural trends towards increased mobility and differentiation thus call into question the geographic, economic, social, and cultural basis of contemporary peasant society and culture (see Kearney 1996). Indeed, peasant societies that, until recently, were the most populous type in world history seem to be rapidly disappearing.

See also: Agriculture, Economics of; Development and Urbanization; Development: Rural Development

Strategies; Development: Sustainable Agriculture; Feudalism; Food in Anthropology; Food Production, Origins of; Hunting and Gathering Societies in Anthropology; Indigenous Knowledge and Technology; Industrial Society/Post-industrial Society: History of the Concept; Industrialization, Typologies and History of; Land Tenure; Peasants and Rural Societies in History (Agricultural History); Rural Geography; Rural Sociology

Bibliography

- Foster G M. 1979 *Tzintzuntzan: Mexican Peasants in a Changing World*, rev. edn. Elsevier, New York
- Kearney M 1996 *Reconceptualizing the Peasantry: Anthropology in Global Perspective*. Westview Press, Boulder, CO
- Meillassoux C 1981 *Maidens, Meal and Money: Capitalism and the Domestic Economy*. Cambridge University Press, Cambridge, UK
- Nagengast C 1991 *Reluctant Socialists, Rural Entrepreneurs: Class, Culture, and the Polish State*. Westview Press, Boulder, CO
- Redfield R 1941 *The Folk Culture of Yucatan*. University of Chicago Press, Chicago
- Wolf E 1966 *Peasants*. Prentice-Hall, Englewood Cliffs, NJ
- Wolf E 1969 *Peasant Wars of the Twentieth Century*. Harper and Row, New York

M. Kearney

Pecuniary Issues in Medical Services: Ethical Aspects

The medical services a person receives depend not only on the patients' health status, the benefits covered by public or private health plans, and the latest developments in medical knowledge but also on the financial incentives offered to doctors that encourage them to use or withhold such services. This may compromise the physicians' loyalty to patients or the exercise of independent judgment on their patients' behalf (Rodwin 1993). Ethical problems arising from financial incentives in the health care system are as old as modern medicine. But their nature is changing along with the transformation that medical care is undergoing in the industrialized countries of the west.

1. Patient-Physician Relationship

At the core of a health care system there is still the patient-physician relationship. In their relationship with physicians patients are people seeking help. That results in this relationship's asymmetry, one that is

unavoidable to a certain extent. Patients can have neither the technical expertise needed to provide specific medical services nor the knowledge and skill to determine what care is needed, nor are they able to coordinate and negotiate the complexities of a highly specialized medical system to make sure that the services needed are actually provided. Patients can be very inhibited by disease, by too high or too low an age, by pain, anxieties, and worries; that means they are in special need of protection. Their vulnerability is intensified even more when they expose their bodies and reveal their most intimate histories, physical and otherwise. Even if patients participate in medical decision-making, they still depend on the physician's advice and need to trust him or her. All the concepts designed to guide medical care, whose ability to function depends on patients successfully playing the role of a 'critical customer,' ignore this asymmetry and the ill person's resulting need for protection.

In modern medicine, the physician-patient relationship is assuming ever clearer forms of a production process based on a division of labor. The patient, as a suffering subject, is at the same time a 'workpiece object' and—to different degrees—a co-producer. That means he or she is directly impacted by rationalization and bureaucratization processes in the health system, and that any changes can set the course for the patient's fears and hopes, pain and comfort, life and death.

Medical services are so complex that legal norms alone cannot protect the sick from unnecessarily suffering or even dying due to mistakes arising from the physician's lack of loyalty. Or to put it another way: a large part of successful physician-patient relationships is based not on (undoubtedly necessary but inadequate) statutory norms but on ethical ones that are internalized in the physician's professional habitus.

2. *Ethics and Institutional Structures*

Most philosophers, including bioethicists, have so far overlooked or at least underestimated the nexus between acts and the institutions increasingly in control of these acts. As a result, they overlook the fact that the 'production' of medical services is not only a work process but also a socialization or learning process, the ethical values of which are learned, reproduced and also modified and repressed. The moral-molding character of the institutions must be assessed as much stronger and durable than that of professional ethicists. This nexus can be approached with simple behavioral assumptions of the kind proposed by Freidson (1970):

First, (...) whatever motives, values, or knowledge people have come into contact with and have 'internalized,' they do

not guide the behavior of most individuals unless they are continually reinforced by their social environment.

Second, (...) the environment can, by reinforcement, lead people to forsake one set of motives, values, or knowledge in favor of another.

And third, given the first two, the average behavior of an aggregate of individuals can be predicted more successfully by reference to the pressures of the environment than by reference to the motives, values and knowledge they had before entering the social environment. The basis of prediction is from the requirements for social 'survival' posed by the social environment and refers to the functional adaptations of the aggregate of individuals who survive. Prediction is, of course, statistically inapplicable to any individual except as a probability value.

The more the power of institutional structures to determine behavior increases, the more significant ethical 'predecisions' embedded in the structure of medical institutions become.

3. *Financial 'Conflicts of Interest' (CoI)*

It is in this sense that the notion of '*conflicts of interest*' has come into use for discussion of the ethical aspects of pecuniary issues in medical services (Spece et al. 1996). It does not mean a subjective, 'inner' conflict that arises when interests of equal standing are weighed but 'a set of conditions in which professional judgment concerning a *primary interest* (such as a patient's welfare ...) tends to be unduly influenced by a *secondary interest* (such as financial gain)' (Thompson 1993). That is the case if the loyalty to patients and/or the independence of professional judgment is compromised (Rodwin 1993).

Primary interests we can define as the ethics patients expect of loyal physicians in western industrialized countries:

- (a) to be loyal to patients,
- (b) to act solely in the patient's interest,
- (c) to make their patients' welfare their first consideration, even when their own financial interests are opposed,
- (d) to keep patient information confidential.

When the ill turn to physicians, they are confident that the same physician will act in trust as their agent and not primarily in order to enhance their income or to perform as agents for third parties. Secondary interests such as a physician's financial gain and his or her professional career are not illegitimate but should be prevented from dominating. Unlike the situation with ethical dilemmas, only one side of the conflicting interests has a claim to priority. This asymmetry between interests is a distinctive characteristic of the concept (Thompson 1993). Conflicts of interests are not an expression of misguided behavior but of *latent* misguided behavior that is built into organizations' systems of incentives and sanctions. They can influence action, but they are not acts and they do not ensure disloyalty (Rodwin 1993).

4. *Transformation of the Health Systems*

In most western industrialized countries the changing structural context is distinguished by shifts in power from the providers to the payers of medical care. This is expressed in the enforcement of new, prospective forms of financing with which expenditures can be limited. Therefore, the financial risk of treatment is increasingly shifting to the providers of medical services. That means their financial success depends on the extent to which the resources consumed by diagnosis and treatment are lower than the fixed capitation fees, budgets, DRGs (diagnosis-related groups: coding in which hospital procedures are rated in terms of cost and the intensity of services delivered, regardless of the cost to the hospital to provide that service) or flat rates. Since every service was paid for or all costs reimbursed with the traditional, retrospective methods of financing, that led for the most part to incentives to increase services. With prospective payment systems, *incentives to decrease* services now predominate, and that creates specific conflicts of interests.

The pressure of limited possibilities to expand is now intensifying competition among the providers of medical services. The main parameter in this respect is minimization of operational costs. Since they are largely determined by physicians' decisions, the ability to control and steer the physician-patient relationship becomes the pivotal point of economic success.

The problem has two elements: first, the financial incentive that rewards certain medical decisions and penalizes others and, second, the degree of inevitability that competition and/or management can compel. That gives rise to different combinations of direct influence on physician-patient relationships via 'micromanagement' and indirect influence via financial incentives. As a result, the push-pull effects of financial incentives for physicians become ever harder to avoid.

5. *Micromanaged Care—Decisions Made at a Distance*

Providers exposed to the pressures of prospective financing have developed an incalculable number of management tools to control the physician-patient relationship, e.g. the use of medical protocols to assess clinical decisions, retrospective utilization reviews, systems of prospective authorization of services, the introduction of primary care physicians as gatekeepers to control referrals to specialists and hospitals, and many other devices. The crucial factor is that important medical decisions tend to be removed from the sphere of direct contact between the patient and physician and transferred to a management.

Here we are no longer dealing only with a conflict between the precept of loyalty to the patient and the interests of the organization. Rather the fact that a

management makes decisions in the absence of the sick person is already a contributory factor that changes the substance of the decisions. The moral nature of such 'decisions made at a distance' is different from decisions made face to face with the other person (Bauman 1989). If a management is to make relevant medical decisions, then questions of suffering and pain, misfortune and hope, life and death must first be turned into indicators, measurable values, and formalized norms, and thus rendered bureaucratically 'banal.' The sick individual's chances of finding empathetic loyalty are thus minimized.

6. *Monetary Incentives*

Monetary incentives are demands on the physician to individually maximize utility: 'if you choose option A instead of B in your clinical decisions, you will be rewarded/penalized with a higher/lower income.' Over and beyond the direct impact on income, the incentives also assume a normative character (in terms of both medicine and ethics): if certain decision-related options are financially rewarded, they are often considered to be medically appropriate and ethically sound.

The forms of prospective payment are multifarious. In the view of their advocates they lead:

- (a) to careful handling of tight resources,
- (b) to less overuse of medical services,
- (c) to a better quality of services, since, after all, part of the financial morbidity risk is imposed on the physician,
- (d) to structural changes and integration in a way that maintains the health of the insured population through disease prevention and health promotion,
- (e) to innovations resulting in measurable improvements in patient satisfaction, and
- (f) they permit this without interfering with the physicians' freedom to decide, i.e. without creating red tape, simply by changing the conditions in which decisions are made.

Critics counter that conflicts of interests resulting from incentives to decrease services (a) could lead to clinically inappropriate decisions and (b) potentially harm the ethical precept of a physician's loyalty to his or her patients, who expect, after all, an independent clinical appraisal in their interest.

(a) The physicians' clinical judgement will be biased and loyalty to the sick person may be at stake when certain options involving clinical decisions are financially rewarded or sanctioned depending on their cost. For example, in the case of capitation and fund holding, the income of the primary physician drops if he or she refers patients to a specialist, prescribes costly drugs, admits them to hospital, or calls for expensive diagnostic procedures. When individual physicians or small group practices assume the insurance risk in this way, without their number of patients being large enough to compensate for the risk,

patients in need of extensive treatment become a threat to their vital interests.

Incentives for underservice should also be avoided since they cannot be an appropriate response to the wrong turns taken by care in the western industrialized countries. For one, there may be overuse in countries like the USA, Germany, or Canada, but at the same time there is also underuse and misuse. For another, such incentives may reduce the extent of certain services, but they cannot substantially alter the ratio of necessary services to unnecessary and harmful ones (Brook and Kosecoff 1988). Since, namely, the pecuniary incentive is blind with respect to what is medically proper, beneficial services have, in principle, the same chance of being reduced or expanded as superfluous and harmful ones. Only with respect to one type of monetary incentives does there appear to be acquiescent agreement: when extra earnings are linked to clearly defined and transparent, positive goals (e.g. increasing participation in screenings, house calls, better quality indicators), there will be hardly any possibility for a conflict of interests and therefore no risk to the patient.

(b) From an ethical point of view such monetary incentives are an implicit rejection of the sick persons' expectation that the physician should be their agent. If, namely, medical decisions were only made with a view to the patient's needs and in keeping with the state of the medical art (as every patient expects and professional ethics prescribe), the type of financing would be of absolutely no importance for the physician's services-related conduct. Regardless of whether a physician receives a fixed salary or payment based on capitation, the services profile would remain the same in every case. The effectiveness of monetary incentives thus stands or falls with the physicians' readiness to violate their clinical judgment and the ethical precept of loyalty to the patient and instead pursue income-related interests (or those of their organization). If they do not do so, all attempts at monetary steering are senseless.

(c) Indirect steering of the patient-physician relationship via monetary incentives to decrease services is not an alternative to the so-called 'bureaucratic model' but tends to bring about such a model. Producers of highly complex services like physicians always have more information at their disposal than the controlling management of the insurance company or the provider organization. So they can use it to manipulate the system ('gaming the system'). Ironically, strong incentives at the level of the physician-patient relationship, with its intimidating implications, produce a spiral of attempts by the physicians to outwit them and by efforts on the part of the other side to increasingly restrict the opportunities to do so, with the result of more regulation and less freedom for doctors and patients alike.

Under the pressure of economic competition, physician-patient relationships thus potentially become

part of an organizational machine. The ethical consequence thereof has been formulated by an American physician as follows, 'I imagine that every doctor in private practice can on occasion think of a patient as a dollar amount couched within a symptom, but I also imagine that doctors regularly pull back from such thinking by virtue of their education and morals. The dangers here are the institutionalization of the impulse to greed by the corporation and the individual physician's subsequent removal from personal responsibility for carrying out the actions that follow from that impulse' (Bock 1988).

The conflict between the 'two cultures' (McArthur and Moore 1997) or between the ethical codes of the profession and those of business is decided in this context in favor of the latter's dominance. With the help of a comprehensive set of tools, including financial incentives, the physician-patient relationship is squeezed into a system whose goal—profitability—is a purely quantitative one. No corporation can earn too much money, its costs can never be 'too low.' The tools of monetary incentives can signal 'more or less,' but not an objective or human criterion for 'enough'; nor can they signal the point at which the physician has to stop pursuing his or the health plan's income interests. So far it has not been possible to counter such boundless systems with any effective controls oriented to protection of the individual patient. In spite of efforts to establish monitoring systems like the HEDIS quality indicators in the US there is no practical system for externally monitoring (e.g. by public agencies) all 'the subtle ways in which providers, struggling to survive market competition, may stint on services' (Relman 1993).

7. The Trust Problem

Rodwin (1993) has written: 'In asking physicians to consider their own interest in deciding how to act, we alter the attitude we want physicians ideally to have. For if physicians act intuitively to promote their patients' interests, we will worry less that they will behave inappropriately. But if their motivation is primarily self-interest, we will want their behavior to be monitored more carefully' (p. 153). As already shown, financial incentives can only modify physicians' behavior in the intended sense when they act in their own income-related interest. That has to collide with the patient's expectation that the physician should place the patient's welfare over his own interests.

Not only is trust at stake but also a precondition for effective medicine. Interpersonal trust is a prerequisite, for example, for the patient's willingness: to reveal potentially stigmatizing information about health-related behavior (substance use, sexual practices, etc.), to describe personal feelings and thoughts necessary to differentiate mental from physical disorders, and to

accept treatment practices or prescribed changes in personal behavior that are difficult, painful, or risky (Mechanic and Schlesinger 1996).

It is not only the trust of the individual patient that is at risk. The trust of any and all in society to have a good chance of finding a loyal physician in case of need is a 'public good' and an element of everybody's standard of living. Public goods are never realized in ideal fashion. But there are respective critical limits. For instance, despite everyday violations we still think of the public good 'trust in traffic safety' as being assured, for otherwise we would never get into a car again. It would undoubtedly be destroyed if we had to expect, say, that 10 percent of road users would no longer stop when the light turns red. The public good known as trust would be invalidated as a result. Today competing business enterprises enforce their corporate strategies at the sick bed through the micromanagement of care, financial incentives to decrease services and sanctions, that increases the risk of 'ethical externalities' and thus of the 'social costs of business enterprises' in medicine.

8. Remedies

When making economic decisions, people will presumably choose the alternative that best serves their interests. But, as economist Amartya Sen points out, before they can do this, they have to make another choice. They must choose to frame their decisions as economic ones that are based on self-interest vs. moral ones that are based on concern for 'what is right.' Such decisions are what most people expect from family, friends and—specifically modified according to the ethical codes of the profession—their trustees like doctors and nurses (Schwartz 1994).

As shown, institutional developments are subjecting physicians to an economic framework of decision-making with growing effectiveness. Not only that, as socialization agencies institutions are seeing to it that the professional code of ethics is already being usurped by superimposed economic considerations. That is why, at the turn of the twenty-first century, people are starting to realize that the patient's welfare must no longer depend on whether a physician is able to resist financial incentives and to ignore or submerge his or her economic interest in order to act in the patients' interest. Society has to find an answer to the principle ethical question of whether the health and life of a patient seeking help is to be decided primarily within an economic framework, and the answer does not have to be expressed solely in the individual physician's conscience but also in the structure of the institutions. There are no patent solutions here. While in some west European countries there is hardly any consciousness of the problem posed by conflicts of interests yet, a public debate about the steps to be taken toward a solution has begun in the United States (cf. biography). Consideration is being given to a broad range of

possibilities. One group of proposals has in common that they would like to cap expenditures at the middle or overall social level by way of budgets, while individual physicians are to be paid a fixed salary (possibly supplemented by financial incentives tied to positive targets like outcomes and patient satisfaction). Overuse, underuse, and misuse of medical services should be prevented by supportive management as opposed to controlling and sanctioning management. Others would like to defuse the prospective payment systems, e.g. financial risk or gain should be limited, financial incentives should not directly reward decreased use of services unless there is specific evidence thereof and monitoring for underuse. Other types of safeguards proposed to protect patients are the disclosure of financial incentives, public review boards, creating independent organizations to certify quality measures, appeals boards to review cases of denied care for reasons of costs, the expansion of individual and collective patients' rights, and the strengthening of consumer organizations.

See also: Assessment and Application of Therapeutic Effectiveness, Ethical Implications of; Discrimination; Discrimination, Economics of; Ethical Dilemmas: Research and Treatment Priorities; Health Care Delivery Services; Health Economics; Medical Experiments: Ethical Aspects; Medical Profession, The; Medical Sociology; Minority Access to Health Services: United States

Bibliography

- Bauman Z 1989 *Modernity and the Holocaust*. Basil Blackwell, Oxford, UK
- Bock R S 1988 The pressure to keep prices high at a walk-in clinic. *New England Journal of Medicine* **319**(12): 785–7
- Brook R H, Koseoff J B 1988 Competition and quality. *Health Affairs* **7**(3): 150–7
- Daniels N, Light D W, Caplan R L 1996 *Benchmarks of Fairness for Health Care Reform*. Oxford University Press, New York
- Emanuel E J, Goldman L 1998 Protecting patient welfare in managed care: six safeguards. *Journal of Health Politics, Policy, and Law* **23**(4): 635–59
- Freidson E 1970 *Professional Dominance*, 1st edn. Aldine, New York
- McArthur J H, Moore F D 1997 The two cultures and the health care revolution: Commerce and professionalism in medical care. *Journal of the American Medical Association* **277**(12): 985–9
- Mechanic D, Schlesinger M 1996 The impact of managed care on patients' trust in medical care and their physicians. *Journal of the American Medical Association* **275**: 1693–7
- Relman A S 1993 Controlling costs by 'managed competition'—would it work? *New England Journal of Medicine* **328**(2): 133–5
- Rice T 1998 *The Economics of Health Reconsidered*. Health Administration Press, Chicago
- Rodwin M A 1993 *Medicine, Money, and Morals: Physicians' Conflicts of Interest*. Oxford University Press, New York & Oxford

- Salmon J W (ed.) 1994 *The Corporate Transformation of Health Care: Perspectives and Implications*. Baywood Publishing Comp., Amityville, NY
- Schwartz B 1994 *The Costs of Living: How Market Freedom Erodes the Best Things of Life*. W.W. Norton, New York/London
- Spece R G, Shimm D S, Buchanan A E (eds.) 1996 *Conflicts of Interest in Clinical Practice and Research*. Oxford University Press, New York & Oxford
- Thompson D F 1993 Understanding financial conflicts of interest. *New England Journal of Medicine* 329(8): 573–6

H. Kühn

Pedagogical Reform Movement, History of

1. History of the Concept

Theories on educational reform have been current since Greco–Roman times, in all three directions of reforms related to school, life, and society. Reform movements in the Middle Ages initially occurred on the basis of religious motives, above all where deviations from church doctrines were associated with an alternative practice. Heresy always suggested new education, because the divergences from the main doctrine picked out alternative forms of salvation as a central theme, to which the life of an own group had to be adapted. Education was not centered on children and youngsters, but it focused on religious and social renewal. It was envisaged that the ‘new’ education would bring about ‘new people,’ a central theme which was developed in Baroque didactics into a program of holistic renewal of mankind. Proofs of this theme can be found in many contexts right up to the pedagogical reform of the twentieth century, in this respect ever more closely related to the mythology and cult of ‘the child’ (Boas 1990). The original sacred theme was continuously secularized, without weakening the appertaining religious power. The division of education and religion had shifted Messianic expectations onto education which, since the seventeenth century, was supposed to ensure perfection in this world.

‘Reform,’ therefore, was never just pragmatic; it was, at the same time, a utopia that did not lose track of the salvation concept and could also understand education as a power and cause of the radical renewal. The success of this was considered to be all the better the more education could appear to be methodically mastered—a concept that was promoted by the ‘inwardness’ of belief within the European Reformation. The ‘inner belief’ should be granted by new education, and this required means of influence. The Baroque programs convinced by way of a mixture of ‘whole-

ness’ and ‘method.’ Similarly, sensualistic psychology in the eighteenth century was understood as a strengthening of the effectiveness of education, and pedagogical doctrines were developed with the promise of methodical innovation. At the same time a literary image of the child was created between Rousseau and the Romantic era which particularly excluded technical expectations. The reform movements of the nineteenth and twentieth centuries developed in this field of tension of ‘natural education’ and ‘method.’

2. Reform Movements in the Nineteenth Century up to World War I

‘Education nouvelle,’ ‘progressive education,’ ‘*Reformpädagogik*’ or ‘*nuova educazione*’ are subsequent historiographic descriptions of widely divergent reform groups and reform concepts which, from the final third of the nineteenth century, opposed the foreseeable damage and weaknesses of the developing educational system. The ‘grammar of schooling’ (Tyack and Cuban 1995), i.e., the institutional form of the modern school, developed during the nineteenth century as a single comparable system in the individual national states. Compulsory schooling was created everywhere, teachers were professionally trained for undertaking tasks of the state, and schools developed specific institutional procedures and forms such as the grading of pupils in age groups, the dominance of subjects, largely frontal forms of lessons, the assessment of pupils’ performances using grades or marks, and so on.

The weaknesses of this system were the initial theme of the reform movements, which already had international links by 1914, but in a practical sense were largely geared towards their respective national systems. Influential authors of the nineteenth century were guided by parameters laid down by Jean-Jacques Rousseau or J. H. Pestalozzi. Rousseau was considered a pioneer of natural education, while Pestalozzi was an authority on education according to moral ideals; both concepts were employed more and more against the institutional development of the school. Friedrich Fröbel’s Kindergarten was of particular importance to the development of the reform movements. Central motives of the reform movements at the end of the nineteenth century, such as the focus on children, the playful manner of learning, or active methods of teaching, were disseminated in the kindergarten media, even if the practice did not always comply with the theory.

Compared with the state school development, these attempts were marginal during the nineteenth century, albeit much discussed and widely published. Significant school-reform movements were only established at the end of the nineteenth century, in a manner which covered three issues. In England and continental

Europe private alternative schools in the area of higher education ('*Landerziehungsheime*') were established which replaced or supplemented the former private school institutions. In the area of elementary school education, reform initiatives were established, above all, where social democracy gained an influence. And third, life reform movements were linked in a very different manner to the school initiatives. One of them was the Theosophical Society, which can be seen as the most influential international individual grouping of educational reform at the end of nineteenth century. In Germany it resulted in the establishment of Rudolf Steiner's Anthroposophical movement (founded in 1912). The establishment of the 'Waldorf School' (1919) in Stuttgart was the start of the largest private school reform movement to date. The new '*écoles nouvelles*,' which were established in England, France, Switzerland, Germany, and various other nations, represented the successful utilization of a pedagogical niche. In 1910 some 100 schools were registered; these operated on the basis of full-day boarding schools as 'educational communities' with a particular curriculum. These schools created a demand because they were able to respond to the dissatisfaction, above all of middle-class parents, with the rigid forms and outdated contents of the grammar schools. Furthermore, country boarding schools often provided the only opportunity to actually achieve university entrance qualifications.

School establishments on the basis of political and social change have been undertaken time and time again since the Owenite Movements in England and the USA. In 1825 the first socialistic school community was established in New Harmony, Indiana. However, most of these did not enjoy lasting success. Only a few alternatives outside the nineteenth-century association of home, school, and church were successful. The pedagogical reform movements were not only grounded in spiritual and/or political world views. Theories of child psychology were to permanently strengthen the intuitions of 'new education': the physiological evidence of the independent development of the senses and feelings, the description of the development of the child according to natural 'phases' or 'stages,' and the recording of the milieu and behavior of children in research facilities.

Together with the exemplary models of the school reform, this knowledge was authoritative for the development of teacher training. The focus of attention of the reform movement in the USA was Columbia University's Teachers College, established in 1897. Its Kindergarten and Elementary School became the experimental focus of child-centered education in the 1920s (Rugg and Shumaker 1928, Ravitch 1986). However, the Teachers College was, above all, a research center. Thus, it was not only activism, but psychological and educational research that was responsible for the success of 'new education.' In 1912 the Institut Jean-Jacques Rousseau was established in

Geneva. The year 1915 saw the establishment in Berlin of the Zentralinstitut für Erziehung und Unterricht, while in 1920 Ovide Decroly established his institute of psychological research, which had previously been private, at the University of Brussels. In 1922 the London Institute of Education was set up on the lines of pedagogical reform.

These developments were favored by the cultural and social *cesura* of World War I. After 1918, political educational reforms were called for in all continental European societies, which were aimed at reacting to the devastating effects of the war. In England a radical education sector came into being; it was, however, intended to, or was forced to, operate outside the system of public education. In the USA, April 1919 saw the establishment of the Progressive Education Association (PEA), which was aimed at changing nothing less than the entire American school system (Cremin 1964, p. 241). Its public influence was considerable, also because after 1930 the academic world was opened up for 'new education' mostly via Teachers College (Cremin 1964, p. 250). The French *éducation nouvelle* gained considerable influence at the same time in the entire francophone educational world, to the extent that at the end of the 1930s the term '*expansion mondiale*' was being applied to the new education.

3. *Pedagogical Reform Between the Wars*

Three developments are of considerable importance to the period between World War I and World War II: the establishment of canonical convictions of what 'new education' is, including the appertaining major figures of eminent reformers, the development of international structures, and the very divergent politicization of the 'new education.' The core of the doctrines was child-centeredness. Education should follow the spiritual, physical, and mental growth of the child. The educational institutions must adapt their actions in line with the natural development. The mode of education dispenses, as far as is possible, with external determination, and thus dispenses with authority.

The reception of Freud's psychoanalysis and the beginnings of child therapy were of great public importance, alongside the liberal concepts of sexual education, which were discussed and practiced in all European major cities up to 1933. It was envisaged that there should be great liberty in early childhood sexuality, and considerable credit was given to Freud for this discovery. The pedagogical reception of Freud, however, is contradictory and selective—not only because of Freud's late theory of drive and aggression but also because its related cultural criticism does not fit in with the optimistic picture of a child developing in accordance with its good nature. Most reformers subscribed to the picture of the good and independent

child that cannot become neurotic or destructive unless pedagogical authority causes it to be so. This picture was simultaneously that of a gender-neutral, culturally independent and socially free child which, precisely as Rousseau had stipulated, exists for itself and needs only to be protected from damage.

The Montessori pedagogy, which became influential in England, the Netherlands, Germany and Italy, was probably the largest international education movement of the 1920s and 1930s. It saw the establishment, in each case, of independent, usually female sponsor groups, with their own communication organs and an international network. The case in Italy is particularly interesting from a political point of view, because Montessori here was not considered the actual pioneer but the theorist of the '*nouva educazione*.' Montessori cooperated with the Fascist regime between 1923 and 1933 to promote her method. 'New' education's focus on the child was the subject of a dispute between Dewey and the Progressive Education Association; they were, however, never able to dissolve the dualisms 'child' and 'curriculum' or 'child' and 'society.'

The international meetings of the New Education Fellowship, which began in 1921 in Calais (topic of conference: 'The Creative Self-expression of the Child') and which ended in 1941 in Ann Arbor (topic of conference: 'Education in a World of Nations') demonstrate the scope of tension between pedagogical reformers' concentration on the child, the creative powers within the child, and the drama of world politics. It is not by chance, therefore, that the pedagogical reform appeared to be torn between the fronts of the Cold War. The Progressive Education Association was dissolved 1955 without much ado under the impression of massive criticism of the liberal and child-centered education which was held responsible for the backwardness of the American education system.

4. *Radical Years*

In 1960 Paul Goodman recollected, with reference to Bertrand Russell and Alexander Neill, the courageous 1920s and the 'revolutionary' attempts of the progressive education movement (Goodman 1960, Sect. IV/6). These attempts had, in fact, failed (Goodman 1960, Sect. IV/7), but they had not been in vain, because the education of the 'beat generation' had to link up with the movement. In the same year, 1960, the New York publisher Harold Hart put out a compilation of texts by Alexander Neill from the 1920s and 1930s under the title *Summerhill: A Radical Approach to Child Rearing* which, by 1970, had reached a circulation of more than two million copies in the USA alone.

Neill, Holt, and Goodman were initially underground authors, at least in America. The international student movement of the 1960s ensured that the

concepts of the 1920s could be once again understood as for avant-garde society's innovations. John Holt's *How Children Learn* (1970) was a strong confirmation of the child-centered pedagogy without these themes being particularly historicized. They were rewritten and adopted for a new generation. Goodman's 'beat generation,' which is actually a negative description that fits with the deficit formula of educational reform, could be related to an experience of emancipation which applied, in particular, to the educational institutions. The cultural revolution was directed against the pedagogical authorities, while projects focussing on self-experience were promoted. Again, formal school education came under attack; the 'new education' (Postman and Weingartner 1969, chapter 8) was aimed at binding school knowledge to subjective meaning (Postman and Weingartner 1969, chapters 5 and 6) without further following a state curriculum. Compulsory education was called into question as 'compulsory mis-education' (Goodman 1964).

The radical approach of the educational reformers in the 1960s was not new. What is new in the USA is the link between progressive education and society reform movements which extend beyond the New Deal positions. This includes the civil rights movement, the women's movement and minorities, as well as third world movements. Paulo Freire's *Pedagogy of the Oppressed* (1970) is perceived by the leftwing American pedagogy as a change in the traditional European-centered reform discourse. The same applies to Ivan Illich's radical criticism of schooling, which goes back to ideas of the Centro Intercultural de Documentacion in Cuernavaca, Mexico, established in 1964. Subsequent to this criticism, the matter in question was no longer school reform, or applying the avant-garde position in the life reforms movements or cultural criticism; the focus now was on education and training.

The 'new education' was understood on one hand as a political strategy of liberation in view of third world problems. On the other hand it was seen as a withdrawal of all demands which were not compatible with the ideal of human liberty. Therefore, it was not by chance that John Holt's *Escape from Childhood* (1975) brought the radical years to an end. Even free and alternative schools such as Summerhill, Holt said, can fail, and, in fact, in greater numbers the more they have tied the happiness of children to their masts (Holt 1975, chapter 14). The problem can only be solved if the pedagogic differentiation between 'adults' and 'children' disappears, and children are granted the same rights as adults (Holt 1975, chapters 15 and 16).

In Europe and England there were comparable radicalizations which allowed for the fundamental negation of education from the linking of education with emancipation in little more than 10 years. This 'anti-pedagogy' is often justified with theories of depth psychology which also demonstrate how strong the effect of educational reform was in the 1920s; only

what was previously understood as a radical start of the 'new education' is now seen as a negative development. The brief recurrent link between movements of societal reform and 'new education' could not be renewed, among other reasons because it became evident that the 'new education' had been merely a historical project that did not allow for general applications such as change of society.

5. Conclusion

The real history of education of the twentieth century has confirmed rather than changed the national systems in place, irrespective of the political catastrophes and the affiliated radical breaks in culture and society. Evidently schools are robust systems which renew themselves from within and in this respect adapt those aspects of progressive education that are acceptable. These innovations never put the entire system at risk. During the 1990s this structural conservatism produced entirely different reform movements which linked up neoliberal economic theories with the old positions of freedom of learning or self-determination. Looked at from the point of view of privatization, efficiency, and commercialization (Chubb and Moe 1990), the criticism of school bureaucracy of the nineteenth century has reappeared, along the renewal of the theory of 'death of the school,' through the revolution of the interactive learning media (Perelman 1993). The critique of schooling substantiates the home schooling movement, accompanied with an internet service for freedom from schooling.

The new movements founded in the 1990s are not turning their backs on reform pedagogical motives. Criticism is now leveled at ineffective and inactive institutions whose organizational form and sponsorship should be radically changed, but there is little consensus concerning the direction in which this change should proceed. The motives are based on political change as much as on maintaining family values, and they have liberal as well as community justifications on which agreement is ultimately unlikely to be reached. As ever, the discourse on how to reform education is ongoing.

See also: Education and Gender: Historical Perspectives; Educational Philosophy: Historical Perspectives; Educational Research and School Reform; Pestalozzi, Johann Heinrich (1746–1827); School (Alternative Models): Ideas and Institutions.

Bibliography

- Boas G 1966/1990 *The Cult of Childhood*. Spring Publications, Dallas, TX
 Chubb J E, Moe T M 1990 *Politics, Markets, and America's Schools*. The Brookings Institution, Washington, DC
 Cremin L A 1961/1964 *The Transformation of the School. Pro-*

- gressivism in American Education, 1876–1957*. Vintage Books, New York
 Freire P 1970 *Pedagogy of the Oppressed*. Seabury Press, New York
 Goodman P 1960 *Growing Up Absurd. Problems of Youth in the Organized Society*. Random House, New York
 Goodman P 1964 *Compulsory Mis-education*. Random House, New York
 Holt J 1964 *How Children Fail*. Pitman Publishing, London
 Holt J 1970 *How Children Learn*. Dell Publishing, New York
 Holt J 1975 *Escape from Childhood*. Dell Publishing, New York
 Perelman L J 1993 *School's Out*. Hyperlearning, the new technology, and the end of education. Avon Books, New York
 Neill A S 1960 *Summerhill: A Radical Approach to Child Rearing*. Hark, New York
 Postman N, Weingartner C 1969 *Teaching as a Subversive Activity*. Delacorte Press, New York
 Ravitch D 1974/1986 *The Great School Wars, New York City, 1805–1973. A History of the Public Schools as Battlefield of Social Change*. Basic Books, New York
 Rugg H, Shumaker A 1928 *The Child-centered School. An Appraisal of the New Education*. World Book Company, Yonkers-on-Hudson/Chicago
 Tyack D, Cuban L 1995 *Tinkering Toward Utopia. A Century of Public School Reform*. Harvard University Press, Cambridge, MA/London

J. Oelkers

Peer Review and Quality Control

Peer review is a mechanism for quality control in science, including the assessment of proposed projects and also of completed work. Its roots go back to the emerging science of seventeenth century, when novelty in natural knowledge became distinguished from technical invention and the ideals of reproducibility and publicity became established. Peer review was acknowledged by the mid-twentieth century as the unquestioned norm in 'academic science.' By then it was assumed that quality was unproblematic, owing to some special ethical disposition of scientists and their methods. With subsequent transformations in the social practice and context of science, which now extends into technology and policy, corresponding changes in quality assurance are needed. The crucial requirement is for a more encompassing community of those evaluating the products and processes of science (see, e.g., Stampa 1997).

In most other spheres of institutional activity, the formalization of quality assurance has become the norm, as for example through the wide-ranging standard-setting procedures of the International Standards Organization (ISO). In academic science, however, along with cultural pursuits like the arts, the methods are still largely informal. Science has been almost unique in having self-assessment performed by practitioners rather than by external 'critics.' To what

extent and in what ways this must change to keep pace with science's expanding role in public life has become an urgent question in the governance of science.

The assurance of quality is not a straightforward task. This has been known since the time of the Romans, as indicated by the Latin motto *quis custodiet ipsos custodes?* (Who will guard the guardians themselves?). This motto implies an indefinite iteration. It is a reminder that, however, routine may be the tasks of quality control, full quality assurance demands yet higher levels of supervision at which informality and explicit value judgments are necessary.

As long as science remained mainly academic, problems of quality were assumed to be resolved by the very nature of the scientific endeavor. The informal systems of checking by peers seemed a rational response to the problem, rather than a culturally contingent mechanism characteristic of a particular epoch. Scientific facts were believed to be discovered by an infallible method, and scientists themselves were viewed as being endowed with certain superior moral qualities that protected them and their work from ordinary human failure or error. This self-correcting property of science could be explained in sociological terms, as in the 'four norms' of scientific practice expounded by Robert K. Merton in 1942 (Merton 1973), or philosophically, as in the committed attempts at self-refutation supposed by Karl Popper to be normal scientific practice (Popper 1959).

With the onset of the industrialization of science after World War II, the self-conscious study of science as a social activity, including the methods of quality assurance, became inevitable. Growth in size, capital investment, scale, and social differentiation within science created divisions between managers and research workers, as well as between researchers and teachers in universities. A *Gemeinschaft* (community) of scholars could no longer realistically be assumed. The earliest disciplined analyses of the quality of production in science were quantitative. Derek J. de Solla Price, who devised some measures of quality and provided analyses of its distribution, did the pioneering work. He noticed that at the leading British science reference library only a minority of journals was ever requested. The contents of the others could be inferred to have no interest, and hence to be of very low scientific quality (Price 1963). This phenomenon is a reminder that 'quality' is a relational attribute. 'Fitness for purpose' depends on whose purposes are dominant; not always perhaps those of a community devoted to the advancement of learning, but possibly only of those scientists working under constraints of 'publish or perish.'

Price's studies were continued in two directions. At the Institute for Scientific Information, Eugene Garfield produced more searching and sophisticated measures of quality, using citations rather than mere number of publications. Such attempts at quantification were bound to become controversial (Brooks

1982, Garfield 1970, 1972). It was impossible to avoid bias in the selection of the relatively small set of journals used for citations; those in conventional mainstream English-language research science were inevitably privileged at the expense of all others. Further, when quantitative measures of citations came to be used as indicators of academic merit, manipulative practices, including reciprocal citations, inevitably developed. The deep problems of a quantitative proxy for quality suddenly became acute.

In a more reflective vein, Jerome R. Ravetz applied the *quis custodiet* principle to analyze the vulnerability of the quality assurance system in science. He observed that the processes of peer review are conducted largely informally and (unlike research) are not themselves normally submitted to open scrutiny and review. They require a different sort of competence, which is not part of the formal training of scientists; and they are also more open to a variety of abuses, ranging from bias to plagiarism. One can understand the phenomena of low quality, both in scientific research and in technological development in these terms. Thus, while denying that the practice of science automatically produces a higher morality, Ravetz agrees that moral standards are necessary for the successful practice of science. On this basis he stresses the importance of morale and morality (and ultimately idealism and leadership) in science (Ravetz 1996).

This analysis provides a background for the increasing interest in 'trust' as an essential element of practice in science, in society, and in their interactions. The broader society has provided resources to the esoteric activities of science because it trusts the scientific community to make good use of them. There has always been an undercurrent of distrust, based on evidence either of pointless research or of malign applications. Once science became deeply involved in technology and related policy problems that crucially affect public health and welfare, the traditional relations of trust could no longer be assumed. It appeared to be necessary for the principles and practices of accountability to be extended from the institutions of political governance (as, e.g., representative democracy) to those institutions, which govern science and its applications.

Quality control in research science has become more difficult as the relatively inflexible technical requirements of the traditional printing process have been relaxed. There is no longer a well-defined 'gateway' to publication through the institutions that control reproduction of, and hence access to documents. First through inexpensive photocopying and now through the Internet, it has become easy for anyone to distribute scientific wares to an unrestricted audience. In addition, the presence of the global media tends to bypass the traditional processes of evaluation, which were conducted personally among colleagues. Isolated scientific results can become media events (Close 1991). All those with an interest in the report, such as

consumers, politicians, regulators, and the stock markets, become potential stakeholders in the evaluation of the result. Thus, science arguably becomes accountable to a drastically extended peer community in the quality-assurance process. The criteria of quality applied by these heterogeneous actors need not be identical to those of 'public knowledge' generated within tightly knit scientific networks.

These developments may be judged in different ways. While they may seriously disrupt the procedures of quality assurance in normal science, they can also bring needed public scrutiny to bear on controversies and scandals. The demystification of scientific practice both enables such events to become news, and is fostered by their being exposed. Top scientists become like celebrities—needing the media for advertising themselves yet simultaneously hating it for its unwanted intrusions. The 'Baltimore affair,' centering on the US Nobel laureate David Baltimore's laboratory at MIT, is perhaps the most notorious case in which a dispute about scientific misconduct was blown up into a lengthy, visible, political saga that damaged all the individuals and institutions involved (Kevles 1998). The episode was symptomatic of an increasingly recognized problem of 'deviance' in science, which carries the unspoken danger that, without timely correctives, misconduct might become the norm.

All these developments affect the maintenance of trust, which is necessary for ordinary scientific practice and even more for quality assurance. As in other professional domains, the normal tendency in science has been for misconduct to be covered up by the responsible institution (not necessarily by the community of scientists). In such situations, ultimate exposure does even more damage and further erodes the basis for mutual trust. Attempts to circumvent the need for trust by increasing bureaucratic surveillance are likely to be counterproductive in their own way, by erecting impediments to free inquiry and communication among colleagues.

The relations between social science and natural science have also been transformed during the last decades, with implications for quality control. Starting with the acceptance of natural science as the ideal of knowledge, essentially independent of social forces, there has been a gradual but accelerating shift toward recognizing all sciences as incorporating social constraints and biases. An early critical interaction was in connection with the astronomical community's management of the eccentric Velikovsky (de Grazia 1966). Later, the social science community embraced Thomas Kuhn's disenchanting picture of 'normal' science (Kuhn 1970). Finally, post-Feyerabend studies of science re-examined the whole institution of scientific inquiry without presupposing any privileged status in relation to either virtue or natural knowledge (Bloor 1991, Bloor et al. 1995, Collins and Pinch 1993, Fuller 1993).

When natural scientists, led by physicists, eventually

confronted the emerging socialized picture of their discipline, the reaction was so strident that 'science wars' became an appropriate label (Gross et al. 1997, Nelkin 1996, Ross 1996). Sociologists of science and post modernists were indiscriminately blamed for all the ills of science, including decline of public trust, budget cuts, resurgent Creationism, and even poor teaching of science. A physicist whose hoax article (Sokal 1996) was accepted by a leading cultural studies journal, *Social Text*, helped to crystallize the attack (Bricmont and Sokal 1998). The implication was that the critics of science had no real quality control of their productions. The science warriors' assumption was that within real science, such problems are prevented from occurring because of the verifiable empirical content of scientific research. In the ensuing debate, there was little mention of the ease of publication of erroneous or vacuous research in the standard scientific literature. Historical episodes, like Millikan's manipulation of his oil-drop results in the course of a controversy on the charge of the electron, were discounted as mere embarrassments (Segerstale 1995).

It has been presupposed thus far that 'science' refers primarily to traditional basic research. But among contemporary forms of scientific practice, curiosity-driven research with no regard for applications has been increasingly marginalized. A diversification has occurred, so that quality assurance must also be considered in such areas as mission-oriented and issue-driven research, forensic science (Foster and Huber 1997, Jasanoff 1995), and the provision of scientific advice for policy (Jasanoff 1990, Salter 1988). In addition, the products themselves and the media through which they are diffused increasingly are diversified. For example, patents are now a common outcome of a research process, and this form of intellectual property is radically different from traditional published papers (Myers 1995). Also, results are reported in unpublished consultancy advice and unpublished 'gray literature' or kept confidential within institutions or even totally sealed under 'lawyer-client confidentiality' and legal settlement agreements. With traditional peer review as the norm, the challenges of quality assurance for these new products and processes are nearly unrecognizable. A genre of critical literature has developed, with some authors directing anger at the new contexts of scientific production (Huber 1991), and others more clearly appreciating the problems they present (Crossen 1994, Jasanoff 1990, 1995).

A parallel diversification has occurred in the types of knowledge production that are accepted as legitimate. The democratization of knowledge now extends beyond the juries who assess the quality of technical evidence in courts (Jasanoff 1998) to include those who master previously esoteric aspects of their predicament (e.g., illness, contamination, pollution, oppression, discrimination, exploitation) through special-interest groups or the Internet. In addition,

claims of specialized or local knowledge are present in even more diverse contexts, as among indigenous peoples, and in systems of complementary or 'traditional' medicine. These claims are commanding increasing commercial and political support among various publics, as well as gaining explicit recognition in numerous international treaty regimes. As a result, a new philosophy of knowledge appears to be emerging, based on a new disciplined awareness of complexity, in which a plurality of legitimate perspectives is taken for granted (Funtowicz and Ravetz 1991). Modern science, with its characteristic methodology and social location, is part of this enriched whole, but not coextensive with it. The criteria and tasks of quality assurance must explicitly involve additional values and interests, incorporating even the ontological commitments of groups other than scientists. This new configuration has been termed postnormal science.

Quality assurance can, thus, be seen as a core commitment of postnormal science, replacing 'truth' as science's ultimate regulative principle (Funtowicz and Ravetz 1992). Defined in terms of uncertainties and decision-stakes, quality assurance encompasses 'public interest,' 'citizen,' and 'vernacular' sciences. In a period of domination by globalized corporate science (Gibbons et al. 1994), this effort to make scientists accountable to interested groups presents a coherent conceptual alternative for the survival of the 'public knowledge' tradition of science. Collegial peer review is thereby transformed into review by an 'extended peer community.' This new form of quality assurance will be given its formal structure and routines by those heterogeneous actors who put it into practice.

See also: Academy and Society in the United States; Cultural Concerns; Constructivism in Cognitive Psychology; History of Science; Research Conduct; Ethical Codes; Research Ethics; Research; Research Publication; Ethical Aspects; Scientific Academies, History of; Social Constructivism; Truth and Credibility; Science and the Social Study of Science; Truth, Verification, Verisimilitude, and Evidence; Philosophical Aspects

Bibliography

- Bloor D 1991 *Knowledge and Social Imagery*. University of Chicago Press, Chicago
- Bloor D, Edge D, Henry J 1995 *Scientific Knowledge*. Athlone, Chicago
- Bricmont J, Sokal A D 1998 *Fashionable Nonsense: Post-modern Intellectuals' Abuse of Science*. Picador, New York
- Brooks H 1982 Science indicators and science priorities. In: La Follette M C (ed.) *Quality in Science*. MIT Press, Cambridge, MA, pp. 1–32
- Close F H 1991 *Too Hot to Handle: The Race for Cold Fusion*. Princeton University Press, Princeton, NJ
- Collins H, Pinch T 1993 *The Golem: What Everyone Should Know about Science*. Cambridge University Press, Cambridge, UK
- Crossen C 1994 *Tainted Truth: The Manipulation of Fact in America*. Simon & Schuster, New York
- Foster K R, Huber P W 1997 *Judging Science—Scientific Knowledge and the Federal Courts*. MIT Press, Cambridge, MA
- Fuller S 1993 *Philosophy, Rhetoric and the End of Knowledge: The Coming of Science and Technology Studies*. University of Wisconsin Press, Madison, WI
- Funtowicz S O, Ravetz J R 1991 A new scientific methodology for global environmental issues. In: Costanza R (ed.) *Ecological Economics*. Columbia University Press, New York, pp. 137–52
- Funtowicz S O, Ravetz J R 1992 Three types of risk assessment and the emergence of post-normal science. In: Krimsky S, Golding (eds.) *Social Theories of Risk*. Greenwood Press, Westport, CT, pp. 251–73
- Garfield E 1970 Citation indexing for studying science. *Nature* **227**: 669–71
- Garfield E 1972 Citation analysis as a tool in journal evaluation. *Science* **178**: 471–9
- Gibbons M C, Limoges C, Nowotny H, Schwartzman S, Scott P, Trow M 1994 *The New Production of Knowledge*. Sage, Beverly Hills, CA
- de Grazia A (ed.) 1966 *The Velikovsky Affair: The Warfare of Science and Scientism*. University Books, New York
- Gross P R, Levitt N, Lewis M W (eds.) 1997 *The Flight from Science and Reason*. Johns Hopkins University Press, Baltimore, MD
- Huber P W 1991 *Galileo's Revenge: Junk Science in the Courtroom*. Basic Books, New York
- Jasanoff S 1990 *The Fifth Branch: Science Advisors as Policy as Policymakers*. Harvard University Press, Cambridge, MA
- Jasanoff S 1995 *Science at the Bar: Law, Science and Technology in America*. Harvard University Press, Cambridge, MA
- Jasanoff S 1998 The eye of everyman: Witnessing DNA in the Simpson trial. *Social Studies of Science* **28**(5–6): 713–40
- Kevles D J 1998 *The Baltimore Case: A Trial of Politics, Science and Character*. Norton, New York
- Kuhn T S 1970 *The Structure of Scientific Revolutions*, 2nd edn. University of Chicago Press, Chicago
- Merton R K 1973 *The Normative Structure of Science*. University of Chicago Press, Chicago
- Myers G 1995 From discovery to invention: the writing and rewriting of two patents. *Social Studies of Science* **25**(1): 57–105
- Nelkin D 1996 What are the science wars really about. *The Chronicle of Higher Education* July **26**: A52
- Popper K 1959 *The Logic of Scientific Discovery*. Basic Books, New York
- Price de Solla D J 1963 *Little Science, Big Science*. Cambridge University Press, Cambridge, UK
- Ravetz J R 1996 (1971) *Scientific Knowledge and its Social Problems*. Transaction Publishers, New Brunswick, NJ
- Ross A (ed.) 1996 *Science Wars*. Duke University Press, Durham, NC
- Salter L 1988 *Mandated Science*. Kluwer, Dordrecht, The Netherlands
- Segerstale U 1995 Good to the last drop? Millikan stories as 'canned' pedagogy. *Science and Engineering Ethics* **1**: 197–214
- Sokal A D 1996 Transgressing the boundaries. *Social Text* **14**: 217–52

Stampa A E III 1997 Advances in peer review research. *Science and Engineering Ethics* 3(1): 1–104

S. Funtowicz

People in Organizations

Organizational researchers study how individuals and organizations interact with their environment to accomplish both individual and collective goals. People are able to accomplish goals in organizations that they would be unable to accomplish alone, and they spend large portions of their lives interacting in organizational contexts. The complexities that emerge from bringing people together in organizational settings, therefore, provide a fascinating lens into the human experience. Organizational research brings together psychology, sociology, economics, and anthropology to examine both individual-level processes at work, as well as the organizational-level processes that dictate how firms relate to each other and their environment.

Reflecting this focus on both individuals and organizations, the field is split between ‘macro’ organizational behavior, drawing upon organizational sociology, and ‘micro’ organizational behavior, drawing upon social, personality, and industrial-organizational psychology. Macro-organizational behavior is concerned with the various strategies organizations use to adjust to their environment, such as the development of networks between firms. In contrast, micro-organizational behavior is concerned with the way personal characteristics combine with contextual features such as task, group, and organizational characteristics to predict work outcomes. Though the strength and uniqueness of organizational behavior rests on the equal emphasis of both individual and organizational processes and the utilization of a variety of research methods that cross disparate levels of analysis, our focus in this article is on micro-theory and research. Excellent reviews of macro-organizational theory can be found elsewhere in this collection (see *Organization: Overview*; *Administration in Organizations*; *Organizational Control*; *Organizational Culture*; *Organizations: Authority and Power*) and in Carroll and Hannan (2000) and Scott (1998).

1. Micro-theory: Person–Situation Interactions

Micro-organizational behavior examines both personal and situational characteristics and, as in the field of psychology, researchers debate the relative utility of each in explaining behavior. Some have emphasized the stability of attitudes and behaviors over time. For example, a person’s satisfaction with his or her job remains relatively stable over years and even decades (Staw and Ross 1985). From this perspective, individual characteristics are the best predictors of behavior since they derive from personal dispositions

that remain stable over time and across situations. Others have criticized this view and posited that organizations should be conceptualized as ‘strong’ situations that are powerful enough to shape individual behavior (e.g., Davis-Blake and Pfeffer 1989). In strong situations individual differences are unlikely to be expressed. Instead, people learn appropriate attitudes and behaviors from their co-workers, established norms, and organizational practices and procedures; these social influence processes are presumed to predict individual behavior better than are personal characteristics.

Researchers have typically considered personal and situational factors in isolation from one another, but a complete understanding of organizational behavior requires their simultaneous consideration. An interactional approach is more complex than a mere additive melding of personal and situational characteristics because it attempts to represent both personal and situational factors and their reciprocal influence.

Interactions between personal and situational characteristics may take at least four forms. First, as specified above, some situations are stronger than others, leading to different levels of behavioral uniformity. Second, work situations do not affect everyone in the same way; some people’s behavior is more consistent across varying situations. Third, certain people, such as those exhibiting ‘charismatic’ leadership, can influence situations more than others. Finally, people do not select into situations randomly, but rather, into situations in which they think their attitudes and behaviors will be appreciated. Developing a complete theory of behavior in organizations, then, requires moving from considering personal and situational factors in isolation to considering the complexity and diversity of possible person–situation interactions and their effects on work outcomes. We illustrate this by identifying the types of person–situation interactions that are relevant to a set of vibrant research domains within organizational behavior: organizational culture, demography, leadership, and creativity.

2. Organizational Culture: Selection, Socialization, and Person–Culture Fit

Research on organizational culture has demonstrated that norms and values can shape individual behavior. Recent research has also demonstrated that behavior can be predicted by understanding how personal characteristics interact with cultural context. Below we discuss the four types of person–situation interactions as they relate to organizational culture.

2.1 Organizational Culture as Strong Situation

Though researchers agree that organizational culture is meaningful and important, they do not agree about how to define and measure it. Organizational culture

research typically draws upon theories in anthropology or cross-cultural psychology and uses ethnographic or quantitative methodologies. Some researchers have emphasized shared values and meaning (O'Reilly et al. 1991) while others have emphasized the ambiguity of cultural values and the existence of subcultures (Martin 1992). Most agree that the existence of different perceptions among members does not preclude the existence of shared assumptions. Organizational culture can be understood as a system of shared values and norms that define what is important and how organizational members ought to feel and behave. Through members' clarity about organizational objectives and their willingness to work toward these objectives, culture influences the attainment of valued organizational goals by enhancing an organization's ability to execute its strategy. This conceptualization focuses on two primary aspects of culture: first, the extent to which agreement and intensity exist about values and norms (culture strength), and, second, the extent to which these norms and values differ across settings in important ways (culture content).

Culture content refers to shared values and norms that define desirable behavior in the organization. For example, some cultures may stress the value of being a 'team player' while others may emphasize independence. For a culture to be strong, members must agree about which values are important and be willing to enforce them strenuously. Members of an organization may agree that they value, for example, being cooperative, but unless unequivocal and salient consequences result from compliance (e.g., cooperating with co-workers) and non-compliance (e.g., competing with or undercutting co-workers), the culture cannot be characterized as strong. A strong culture is characterized by social control in that members both agree about and are willing to enforce values and norms, even if such enforcement violates hierarchical lines of authority. Further, greater behavioral homogeneity among members, for better or worse, should be observed in stronger organizational cultures.

Cultural values are conveyed to members through the socialization processes that new recruits experience when they enter an organization. Though socialization often takes the form of training and contributes to increased task knowledge, information about the norms of the organization is also transmitted through training and other types of socialization (Morrison 1993). Interestingly, such normative information appears to be transmitted very early in new recruits' membership since cultural fit changes little after members' first year (e.g., Rynes and Gerhart 1990).

2.2 Behavioral Consistency Across Various Organizational Cultures

Although culture influences members' behavior, they are not merely passive recipients of social control.

Individual characteristics may interact with the organization's culture to predict important behavioral and work outcomes. Research on person-organization fit, defined broadly as the compatibility between people and the organizations for which they work (Kristof 1996) has focused primarily on congruence between patterns of organizational values and patterns of individual values. New employees whose values more closely match those of the organization view it as more attractive, are more likely to join when made an offer, are able to adjust to it more rapidly, perform better, are more satisfied, and remain with it longer (e.g., Chatman 1991).

But, just as organizational cultures may differentially affect behavior, people may also differ in the extent to which their behavior is shaped by an organization's culture. For instance, compared to individualists, cooperative people were more responsive to the individualistic or collectivistic norms characterizing their organization's culture, and exhibited greater behavioral adaptation to each across the two types of organizational cultures (Chatman and Barsade 1995). Thus, congruence models, which presume an additive equivalence of person and situation factors and assume that greater congruence is always better, cannot fully explain behavior in organizations. Instead, a focus on mismatches between person and organization characteristics that challenge people to either act in accordance with the culture and thereby contradict enduring personal tendencies, or vice versa, might generate insight into such interactions.

Future research might focus on the set of characteristics, such as cooperation, self-monitoring, and self-esteem, that contribute to people's flexibility across situations. Identifying such characteristics could improve predictions of the behavioral expression of person characteristics both across time and across situations, and in particular, the extent to which an organization's culture and processes will influence member behavior. Research might investigate other organizationally relevant matches and mismatches to understand how different situations influence person-situation interactions. For example, examining mismatches between honest people and dishonest organizations may help to identify if and when good people 'turn bad.'

2.3 Individuals' Influence Over Organizational Culture

Founders and senior executives, who have legitimacy and authority, may be the most influential individuals in an organization. The person who creates the organization has significant impact on the strategies that the group develops to survive, and these are often preserved in the organization's values, practices, and promotion patterns well past the individuals' actual presence. Ironically, newcomers, who are at the other end of the spectrum in terms of authority and

legitimacy from founders, may also exert a great deal of influence on culture. This influence may be indirect; research has shown that the process of recruiting new members, including emphasizing an organization's attractions, strengthens current members' connection to their organization (Sutton and Louis 1987), and can exert mutual learning of the organizational culture (March 1991). An organization's culture may also be transformed by the entrance and exit of its members. Thus, the strength of an organization's culture may depend on the individual mobility of its members. Using simulation methods, organizations' length of service distributions have been examined as indicators of the extent to which members have been socialized into the culture's norms and values. This research has shown that variations in service distributions are positively associated with heterogeneity in organizational culture due to three distinct cultural processes: socialization, group cohesiveness, and common historical experiences (Carroll and Harrison 1998).

2.4 Situation Selection Based on Organizational Culture

Schneider (1987) developed the Attraction–Selection–Attrition (ASA) model, which posits that the key factor influencing the relationship between people and organizations is the fit between individual personality and the modal personality represented in the organization. People are differentially attracted to organizations on the basis of the anticipated fit between their own characteristics, such as personality traits, and organizational characteristics, such as culture (e.g., Cable and Judge 1997). Job seekers take an active role in the recruitment process and are attracted to careers and organizations reflecting their personal values and interests. Further, organizations have developed formal and informal strategies to identify and select recruits who will be compatible with their goals and working environment, elements that are strongly influenced by an organization's cultural values. For instance, rather than focusing on job-related criteria, selection appears to be based on such socially based criteria as 'personal chemistry,' and fit with the organization's values.

3. Organizational Demography

People may differ from each other in many ways that are both observable and unobservable. Members' observable demographic characteristics constitute a context for every other individual in the organization. For instance, a lone newcomer in an organization of established members may have a very different experience than a newcomer entering an organization characterized by members with varying tenure. Given

the dramatic changes in US labor force demography over the past decade, relational demography, or the distribution of demographic characteristics within organizations, has become an active area of research.

Research on organizational demography improves upon past research on employee turnover. Turnover typically indicates a discrete departure event while demography focuses on length of service and rate of departure by taking into account the distribution of people by the length of time they have spent in the organization. This distribution can be influenced by a host of factors including personnel policies, technology, and the degree of unionization in the workforce (Pfeffer 1983). Most research on observable differences has examined the effect of demographic heterogeneity versus homogeneity on performance and has yielded mixed results (see Williams and O'Reilly 1998 for a comprehensive review). Some studies have demonstrated the positive effects of demographic heterogeneity for increasing the number of novel perspectives that can be used to solve difficult problems, increasing an organization's network of contacts, and facilitating organizational change. However, demographic heterogeneity may also lead to communication problems, less social integration in workgroups and greater turnover (e.g., Tsui et al. 1992). Demographic distribution among members across various attributes is an important situational factor that deserves further research since it can influence behavior differently depending on an individual's own demographic profile.

3.1 Behavioral Consistency Across Demographically Heterogeneous Workgroups

Research suggests that an organization's culture may influence the relationship between demographic diversity and work outcomes. For example, the purported benefits of a demographically diverse workforce are more likely to emerge in collectivistic organizations that make organizational membership more salient than membership in a demographic category (Chatman et al. 1998). An organization's culture may dictate the extent to which members view certain demographic characteristics as valuable and others as associated with lower status within an organization's informal social system (Spataro 2000). Furthermore, each attribute, such as tenure, race, or sex heterogeneity within a group, may differentially influence individual behavior and the combinations of various attributes can result in 'fault lines' which become stronger as more attributes align themselves in the same way (Lau and Murnighan 1998).

One explanation for the lack of clarity about the benefits and detriments of diversity is that researchers have neglected to consider key mediating processes between demographic composition and performance.

As shown in one recent study, heterogeneous groups initially created norms fostering independence rather than cooperation among members, but cooperative norms subsequently mediated the relationship between group demographics and performance (Chatman and Flynn, 2001). Similarly, a group's level of conflict influenced the impact of demographic heterogeneity on performance (Jehn et al. 1999).

4. Leadership

4.1 *Some Individuals Can Effect Change More Than Others*

Early leadership research focused on the physiological and psychological traits thought to be associated with exceptional leaders. These 'great man' theories of leadership examined the effects of personal characteristics such as height, physical appearance, and intelligence on leaders' emergence and effectiveness. This stream of research has its counterpart in more current studies examining the effects of self-confidence, extraversion, and energy level (e.g., House 1988). The aim of this approach has been to identify a leadership personality. However, it leaves many crucial questions unanswered, such as whether certain personal characteristics become more important than others depending on the organizational context, and why, regardless of formal authority, followers perceive some people as leaders and not others.

Contingency theories of leadership were advanced to explain how certain personal characteristics made a leader effective in certain situations (e.g., House and Baetz 1979). For example, leaders who initiated structure raised the productivity and satisfaction of a group working on a boring or simple task but lowered the productivity and satisfaction of a group working on a complex task, while a considerate leader raised the satisfaction and productivity of a group engaged in a boring task but had little effect on a group engaged in a task they found intrinsically interesting. Additionally, research showed that allowing members to participate in decision making increased commitment but depended on the amount of trust the leader had in his or her subordinates as well as the urgency of task completion (Vroom and Jago 1978). Thus, contingency theories of leadership were more comprehensive than trait theories; however, they still did not account for the interactive effects of leader characteristics and their situational contexts.

Recent research has focused on charismatic and transformational leadership, demonstrating that some individuals influence situations more than others. This research takes an interactional approach by conceptualizing leadership as a personal relationship between the leader and his or her followers. A leader must have certain interpersonal skills in order to inspire followers to set aside their goals and to pursue a common vision. Charismatic leaders are thought to

have the ability to change their circumstances by increasing followers' motivation and commitment and, sometimes, to change the direction of the entire organization (e.g., Meindl et al. 1985). However, a leader is only charismatic if followers recognize him or her as such; followers must identify with the vision articulated by the leader. In one particularly exhaustive laboratory study of charismatic leadership (Howell and Frost 1989), confederates were trained to display qualities of a charismatic leader, such as projecting a dominant presence, articulating a large overarching goal, and displaying extreme confidence in followers' ability to accomplish this goal. In addition, norms were created in each group for either high or low productivity. In contrast to participants working under a considerate or structuring leader, participants working under the charismatic leader displayed higher task performance regardless of the group productivity norm. This finding suggests that leaders mold their styles in response to the situation. Moreover, some leaders are capable of changing the situation itself by changing followers' perceptions and motivation.

4.2 *Leadership as a Function of the Strength of the Situation*

Some researchers have been skeptical of a leader's ability to change situations, and have suggested that leadership is far more situationally determined than might have been assumed. The attributional theory of leadership suggests that because people tend to associate certain behaviors with those of a leader, leadership qualities will be attributed to a person displaying these behaviors (Calder 1977). Various biases emerge from this attribution, however. For instance, individuals tend to overestimate the amount of control a leader has over events that are, in fact, random or uncontrollable (Pfeffer 1977). Furthermore, a leader will be given more credit when situations are unfavorable (Meindl et al. 1985). Individuals' lay conceptions of leadership can be used or misused for the purposes of organizational impression management (Ginzel et al. 1993). In sum, leadership research has focused more on contexts and followers rather than on the characteristics of a focal leader. Future research might examine whether leaders reflect the personal characteristics of their followers or complement their weaknesses, if some followers have a greater psychological need for leadership than others, and the various substitutes for leadership, or how people can be compelled to lead themselves.

5. Creativity and Innovation

Research on creativity, like leadership, has moved from emphasizing traits, to considering the organizational context as well as the interaction between the

two. Creativity is generally viewed as distinct from innovation. Creativity occurs at the individual level, and refers to the generation of ideas that are both novel and useful. Innovation refers to the process of implementing these ideas at the organizational level.

5.1 *Some People Are More Creative Across Situations Than Others*

Early creativity research focused on the personality traits associated with individuals who had made creative contributions in their respective fields. People who have a wide variety of interests, are attracted to complex tasks, tolerant of ambiguity, and self-confident perform more creatively (e.g., Barron and Harrington 1981). However, trait research ignores how organizational contexts influence people's ability to perform creatively.

5.2 *Some Situations Foster More Creativity Than Others*

The degree to which one's job encourages intrinsic versus extrinsic motivation affects one's creative performance. Early studies suggested that when people worked on tasks that they found intrinsically interesting, adding an extrinsic reward lowered their interest in performing the task for its own sake (Deci 1971). More recent studies in organizational settings showed that individuals were most creative when they were intrinsically motivated. This intrinsic interest led them to stay focused, persist longer to complete difficult tasks, to 'play' with ideas, and suggest novel solutions (Amabile 1988). This suggests that the situational factors that are associated with decreases in intrinsic motivation, such as a controlling supervisory style, and an emphasis on external rewards, may indirectly diminish people's creative potential.

Organizational culture also influences creativity and innovation. Organizations that have mechanisms to express confidence in members, and communicate this confidence through the culture's core values, increase creativity among members (Kanter 1988). These findings are supported by a recent ethnography of a product design firm, IDEO, which creates new products by taking technologies from one industry and applying them in other industries where these technologies are unknown (Hargadon and Sutton 1997). At IDEO, employees are encouraged to create analogies between past technological solutions and current problems and to share them in brainstorming sessions. Further, employees are selected who have unique hobbies or skills that can be used to solve design problems.

Employees who have traits associated with creativ-

ity are more likely to thrive in organizations such as IDEO, which place an emphasis on creative performance. For example, employees produced the most creative work when they possessed the appropriate personal characteristics, were working on complex assignments, and were supervised in a supportive, noncontrolling fashion (Oldham and Cummings 1996). While the possibility that organizations can manage creativity through the use of a strong culture appears promising (Flynn and Chatman 2001), some worry that mechanisms of social control will stifle, not encourage creativity (Nemeth and Staw 1989). Future researchers might examine how organizations can achieve harmony and cohesion without sacrificing the flexibility and constructive conflict necessary for creativity and innovation. Many believe that creativity and innovation are the last sustainable competitive advantages and as such these issues will continue to generate a great deal of interest.

6. *Conclusion*

As theorists endeavor to develop a complete understanding of behavior in organizations, the analysis of both personal and situational factors, as conjoined units of behavior, will become increasingly fundamental to organizational studies. Person-situation interactions are much more complex than the simple addition of personal and situational characteristics, and these interactions may take a variety of forms. Some people are more responsive to situations than others, some situations can shape behavior to a greater degree than others, and some people have the unique capability to shape situations to their advantage or that of the organization. Furthermore, group members' personal characteristics may constitute the situational context as every individual responds to the personal or demographic characteristics of every other individual in the organization (Carroll and Harrison 1998, Tsui et al. 1992). These interactions are both complex and diverse: hence the field of organizational behavior must necessarily reflect this diversity. It will become increasingly important to observe these interactions over time. Since people and situations adjust to each other, cross-sectional research will not capture the dynamic interplay between the two. By pursuing interactional research over time, organizational behavior will increasingly evolve into a field that is as vibrant as the organizations it seeks to understand.

See also: Authority; Delegation; Innovation: Organizational; Leadership in Organizations, Sociology of; Leadership, Psychology of; Marx, Karl (1818-89); Organization: Informal; Organization: Overview; Organizational Behavior, Psychology of; Organizational Climate; Organizational Culture; Organ-

izational Culture, Anthropology of; Organizational Decision Making; Organizations: Authority and Power; Organizations, Sociology of; Schumpeter, Joseph A (1883–1950); Weber, Max (1864–1920)

Bibliography

- Amabile T 1988 A model of creativity and innovation in organizations. In: Staw B M, Cummings L L (eds.) *Research in Organizational Behavior*. JAI Press, Greenwich, CT, Vol. 10, pp. 169–212
- Barron F, Harrington D M 1981 Creativity, intelligence and personality. *Annual Review of Psychology* **32**: 439–476
- Cable D M, Judge T A 1997 Interviewers' perceptions of person–organization fit and organizational selection decisions. *Journal of Applied Psychology* **82**: 546–61
- Calder B J 1977 An attribution theory of leadership. In: Staw B, Salancik G (eds.) *New Directions in Organizational Behavior*. St. Clair Press, Chicago
- Carroll G, Hannan M 2000 *The Demography of Corporations and Industries*. Princeton Press, Princeton, NJ
- Carroll G, Harrison J R 1998 Organizational demography and culture: Insights from a formal model and simulation. *Administrative Science Quarterly* **43**: 637–67
- Chatman J A 1991 Matching people and organizations: Selection and socialization in public accounting firms. *Administrative Science Quarterly* **36**: 459–84
- Chatman J A, Barsade S G 1995 Personality, organizational culture, and cooperation: Evidence from a business simulation. *Administrative Science Quarterly* **40**: 423–43
- Chatman J, Flynn F 2001 The influence of demographic heterogeneity on the emergence and consequences of cooperative norms in work teams. *Academy of Management Journal*, **44**(5)
- Chatman J A, Polzer J T, Barsade S G, Neale M A 1998 Being different yet feeling similar: the influence of demographic composition and organizational culture on work processes and outcomes. *Administrative Science Quarterly* **43**: 749–80
- Davis-Blake A, Pfeffer J 1989 Just a mirage: The search for dispositional effects in organizational research. *Academy of Management Review* **14**: 385–400
- Deci E L 1971 The effects of externally mediated reward on intrinsic motivation. *Journal of Personality and Social Psychology* **18**: 105–15
- Flynn F, Chatman J 2001 Strong cultures and innovation: Oxymoron or opportunity. In: Cartwright S, Chatman J, Cummings T, Earley P C, Holden N, Sparrow P, Starbuck W (eds.) *International Handbook of Organizational Culture and Climate*. Wiley, Chichester, UK, pp. 263–88
- Ginzell L, Kramr R, Sutton R 1993 Organizational impression management as a reciprocal influence process. In: Staw B M, Cummings L L (eds.) *Research in Organizational Behavior*. JAI Press, Greenwich, CT, Vol. 15, pp. 227–66
- Hargadon A, Sutton R I 1997 Technology brokering and innovation in a product design firm. *Administrative Science Quarterly* **42**: 716–49
- House R J 1988 Power and personality in complex organizations. In: Staw B M, Cummings L L (eds.) *Research in Organizational Behavior*. JAI Press, Greenwich, CT, Vol. 10, pp. 305–57
- House R, Baetz M 1979 Leadership: Some empirical generalizations and new research directions. In: Staw B M, Cummings L L (eds.) *Research in Organizational Behavior*. JAI Press, Greenwich, CT, Vol. 1, pp. 341–423
- Howell J, Frost P 1989 A laboratory study of charismatic leadership. *Organizational Behavior and Human Decision Processes* **43**: 243–69
- Jehn K A, Northcraft G B, Neale M A 1999 Why differences make a difference: A field study of diversity, conflict and performance in workgroups. *Administrative Science Quarterly* **44**: 741–63
- Kanter R M 1988 When a thousand flowers bloom: Structural, collective and social conditions for innovation in organizations. In: Staw B M, Cummings L L (eds.) *Research in Organizational Behavior*. JAI Press, Greenwich, CT, Vol. 10, pp. 169–211
- Kristof A L 1996 Person–organization fit: An integrative review of its conceptualizations, measurement, and implications. *Personality Psychology* **49**: 1–49
- Lau D, Murnighan J K 1998 Demographic diversity and faultlines: The compositional dynamics of organizational groups. *Academy of Management Review* **23**: 325–40
- March J G 1991 Exploration and exploitation in organizational learning. *Organization Science* **2**: 71–87
- Martin J 1992 *Culture in Organizations: Three perspectives*. Oxford University Press, New York
- Meindl J, Ehrlich S, Dukerich J 1985 The romance of leadership. *Administrative Science Quarterly* **30**: 78–102
- Morrison E W 1993 Longitudinal study of the effects of information seeking on newcomer socialization. *Journal of Applied Psychology* **78**: 173–83
- Nemeth C J, Staw B M 1989 The tradeoffs of social control and innovation in small groups and organizations. *Advances in Experimental Social Psychology*. **22**: 175–210
- Oldham G R, Cummings A 1996 Employee creativity: Personal and contextual factors at work. *Academy of Management Journal* **39**: 607–34
- O'Reilly C A, Chatman J A, Caldwell D M 1991 People and organizational culture: A Q-Sort approach to assessing person–organization fit. *Academy of Management Journal* **34**: 487–516
- Pfeffer J 1977 The ambiguity of leadership. *Academy of Management Review* **2**: 104–12
- Pfeffer J 1983 Organizational demography. In: Staw B M, Cummings L L (eds.) *Research in Organizational Behavior*. JAI Press, Greenwich, CT, Vol. 5, pp. 299–357
- Rynes S, Gerhart B 1990 Interviewer assessments of applicant 'fit': an exploratory investigation. *Personnel Psychology* **43**: 13–35
- Schneider B 1987 The people make the place. *Personnel Psychology* **40**: 437–53
- Scott R W 1998 *Organizations: Rational, Natural and Open Systems*, 4th edn. Prentice Hall, Englewood Cliffs, NJ
- Snyder M 1987 *Public Appearances, Private Realities: The Psychology of Self-monitoring*. WH Freeman, New York
- Spataro S 2000 Not all differences are the same: The role of status in predicting reactions to demographic diversity in organizations. Unpublished doctoral dissertation, University of California, Berkeley, CA
- Staw B M, Ross J 1985 Stability in the midst of change: The dispositional approach to job attitudes. *Journal of Applied Psychology* **70**: 469–80
- Sutton R I, Louis M R 1987 How selecting and socializing newcomers influences insiders. *Human Resource Management* **26**: 347–61
- Tsui A S, Egan T D, O'Reilly C A 1992 Being different:

- Relational demography and organizational attachment. *Administrative Science Quarterly* 37: 549–79
- Vroom V H, Jago A G 1978 On the validity of the Vroom–Yetton model. *Journal of Applied Psychology* 63: 151–62
- Williams K Y, O'Reilly C A 1998 Demography and diversity in organizations: a review of 40 years of research. In: Staw B M, Cummings L L (eds.) *Research in Organizational Behavior*. JAI Press, Greenwich, CT, Vol. 20, pp. 77–140

J. A. Chatman and J. A. Goncalo

Peptides and Psychiatry

Since the 1970s, much has been learned about the role of peptides in the central nervous system (neuropeptides) and behavior. Originally thought to be rare and relatively unimportant compared to the classical neurotransmitters, they are now known to be nearly ubiquitous and extremely important in brain function. Like the 'classical' small molecule neurotransmitters, neuropeptides function as chemical mediators of neuron to neuron communication. However, unlike such classical neurotransmitters, the neuropeptides have often been evolutionarily conserved to act both as local transmitter modulators and as endocrine hormones, thus mediating complex patterns of integrated behavior.

The role of neuropeptides in facilitating complex aspects of behavior makes them ideal candidates in understanding the neurobiological bases of psychiatric disorders. Whereas the classical neurotransmitter systems are involved in the neuronal circuitry mediating all behavior and pathology, the neuropeptide systems appear anatomically distributed, but functionally more limited. Thus these systems may allow an understanding of the physiology and pathophysiology of behavioral repertoires along with the ability to treat psychiatric disorders with more specific treatment modalities. This article will briefly review neuropeptide biology and function in general. Several specific examples of neuropeptides with known behavioral significance are discussed, allowing some generalizations to be made connecting physiological behavior to pathological disease states.

1. Neuropeptide Biology

Like the classical small neurotransmitters, neuropeptides also function as chemical mediators of neuron to neuron communication via presynaptic release onto postsynaptic receptors. Some neuropeptides serve their function primarily within the central nervous system (CNS), e.g., galanin and enkephalin. However,

other neuropeptides serve as both neurotransmitters and endocrine hormones via pituitary release to act on peripheral sites, e.g., oxytocin and vasopressin. Other neuropeptides serve as local neuromodulators within the brain but also as hormone-releasing factors in the hypothalamus–pituitary system, e.g., corticotropin-releasing factor and thyrotropin-releasing hormone. Finally, some neuropeptides appear to have distinct and separate roles in CNS and periphery, e.g., neurotensin and cholecystokinin.

1.1 Neuropeptide Production

The classical neurotransmitters (i.e., glutamate, GABA, dopamine, serotonin, norepinephrine, and acetylcholine) are formed from small molecule precursors in the cytoplasm, often stored at the terminals where the neurotransmitter is packaged into vesicle pools. Control of pool size is a function of enzyme concentration and precursor availability. In contrast, peptides are the direct products of mRNA translation, essentially small protein products. Most neuropeptides are between 2 and 40 amino acids in length. They are initially formed as larger precursor proteins (pre-hormones) by translation of mRNA into polypeptides that are then cleaved into various active smaller peptides. Within the cell body, vesicles of neuropeptides are packaged in the Golgi apparatus and then transported to the distal regions (axons and dendrites) of the neuron where they are released with neuronal activity. Control of neuropeptide availability is therefore largely a direct function of gene transcription and translation. Thus, change of neuropeptide expression occurs as a function of multiple hormonal and other modulatory influences on neuronal function.

1.2 Neuropeptide Release and Inactivation

Discovered as early as 1940 by the Scharrers, a husband and wife team, peptides are located in secretory granules (vesicles) at the neuronal terminal. Depolarization of the neuronal membrane leads to calcium influx locally, resulting in vesicle fusion with the membrane and release of peptide into the extracellular space. After release from presynaptic nerve terminals, the peptides diffuse across the synaptic cleft, binding to high affinity receptors on the postsynaptic membrane.

Termination of neuropeptide activity occurs when peptidase enzymes cleave the peptides into smaller fragments, disrupting their biologic activity. This is in contrast to small neurotransmitters that are removed from the synaptic cleft primarily by reuptake into the presynaptic terminal, with only modest breakdown by metabolic enzymes in the extracellular space. These

differences in inactivation generally lead to substantially longer periods of activity by the neuropeptides. Numerous peptidases have now been identified from endopeptidases to carboxypeptidases with differential specificities and different solubility vs. membrane-bound characteristics. These may differ among the different neuropeptide systems.

1.3 Neuropeptide Receptors

Neuropeptide receptors are generally of the G-protein coupled class of seven transmembrane receptors. Peptides released into the synaptic cleft activate these receptors via high-affinity binding and mediate internal signaling events via the cytoplasmic G-protein coupled effector. Activation of these receptors activates and inhibits a variety of second messenger systems and can result in cell firing, modulation of membrane threshold, phosphorylation and dephosphorylation events, and alteration of gene expression. These changes can thus have direct effects on the target cells or can modulate effects of other neurotransmitters. Furthermore, as in the hypothalamic-pituitary system, activation of hormone-releasing factor receptors causes release of other neuropeptides into the periphery, e.g., thyrotropin-releasing hormone activates release of thyroid-stimulating hormone from the thyrotrophs in the pituitary.

1.4 Colocalization of Neuropeptides and Classical Neurotransmitters

Neuropeptides are often found in neurons that also contain one of the small classic neurotransmitters. However, the neuropeptides and small neurotransmitters often have different dynamics of release and subserve different functions. For example, in some neurons with slow firing, release is limited to small neurotransmitter vesicles. With rapid burst firing and prolonged depolarization, however, the calcium concentration in the presynaptic terminal is significantly elevated, leading to release of neuropeptide vesicles in addition to neurotransmitter. Thus the different dynamics of release along with differential receptor activation allow these colocalized mediators to carry different functional signals in neuronal circuitry.

2. Behavioral and Psychiatric Aspects of Neuropeptides

The impetus for understanding neuropeptide biology since the 1970s has led to great advances in endocrinology and has had a significant impact on psychiatry. It is becoming increasingly evident that many

psychiatric disorders are associated with neuropeptide abnormalities. Alterations in specific neuropeptide-containing neurons have been shown in Alzheimer's disease, Parkinson's disease, Huntington's disease, mood disorders, anxiety disorders, eating disorders, and schizophrenia.

There are over 100 neuropeptides now identified and currently being scrutinized. This rapid expanse of knowledge cannot be summarized here. Therefore, a few examples of neuropeptide systems are discussed below which have known behavioral, physiological, and pathological significance. As with those discussed below, many neuropeptides have multiple roles in somatic sites, the central nervous system, and as hormonal mediators linking them. Understanding the physiologic role of these molecules will likely guide the understanding of pathophysiology and treatment of the psychiatric disorders in which they are involved.

2.1 Corticotropin-releasing Factor—Model of Integrated Stress Response

Corticotropin-releasing factor (CRF) was originally identified in crude form in 1955 and finally structurally identified in 1981 as a 41 amino acid peptide. It is the primary stimulant of adrenocorticotrophic hormone (ACTH) release from the anterior pituitary, initiating the hypothalamic-pituitary-adrenal (HPA) peripheral stress response. It clearly also functions as a putative neurotransmitter in the brain coordinating global responses to stressors. In higher organisms, CRF appears to mediate a complex behavioral program integrating the endocrine, immunologic, autonomic, and behavioral responses to stress.

CRF appears to be of crucial importance in many psychiatric disorders. Dysregulation of the HPA axis in major depression has been one of the most consistent findings in biological psychiatry since these abnormalities were first identified. As the field continues to understand the nature vs. nurture debate of psychopathology, the role of the endogenous stress response system in the stress-diathesis model of affective and anxiety disorders becomes paramount.

2.1.1 CRF, the physiologic mediator of stress. There is now abundant evidence that CRF and related peptides are the primary central nervous system (CNS) mediators of stress. Stress, *per se*, is generally defined as extraordinary demand on the organism or alterations in psychological homeostatic processes. Initially, the importance of CRF was thought to be primarily endocrine via activation of ACTH from the pituitary, with subsequent release of cortisol and other glucocorticoids from the adrenal gland. Since the 1980s, however, its role in the brain independent of the

HPA axis has been firmly established. Consistent with its role as a neurotransmitter mediating complex repertoires, CNS administration of CRF mimics many of the behaviors and autonomic responses seen with physiologic stress.

Neurons expressing and releasing CRF are located in neuroanatomic locations which are thought to be critical for the stress response, including areas of neocortex, limbic system, and midbrain. In the limbic system, the central amygdala and bed nucleus of the *stria terminalis* are regions of the so-called 'extended amygdala' believed to be involved in stress and anxiety. CRF neurons from these regions project to many midbrain and brainstem sites mediating internal stress responses. The *locus coeruleus* and the *raphe nucleus* in the midbrain receive these CRF projections, influencing the activation of norepinephrine and serotonin circuits, respectively. Thus stress activation of the CRF system can disrupt attention and vigilance, concentration and memory, sleep and wakefulness via these transmitter systems, and it is likely involved in their abnormal functioning in mood and anxiety disorders.

There are multiple experimental paradigms further implicating the central CRF system as the prime mediator of the acute and prolonged stress response. Rodents administered CRF directly into the CNS show decreased reproductive behaviors, altered sleep, increased grooming, increased signs of despair, increased neophobia, and altered appetite, all of which mimic the behavioral stress response. Furthermore, rodents raised with early life stress (primarily maternal separation) hypersecrete CRF from the hypothalamus and amygdala, both during the stress and later as adults.

Administration of CRF into the CNS of nonhuman primates elicits many signs of behavioral despair including increased vocalization, decreased exploration, and increased huddling. These effects are reversed by CRF receptor antagonists, as are the behavioral consequences of environmental stress. Consistent with these experiments are early studies of maternal deprivation in young Bonnet macaque monkeys. These experimental animals exhibited sustained behavioral despair with decreased locomotion, decreased exploration, and altered food intake, similar to those animals with centrally administered CRF. These animals persistently exhibited a prolonged activation of the HPA axis with elevated cortisol and ACTH peripherally.

Stress exposure to maternal-infant dyads without separation appears to have similar long-term effects. Young Bonnet macaques raised in naturalistic settings by mothers with an unpredictable food supply showed increased signs of anxiety compared to those with a predictable food supply. These symptoms were accompanied by elevated CSF CRF concentrations. Remarkably, the grown youngsters consistently showed continued signs of anxiety and affective

disturbance after removal from these stressful living constraints, and their CNS CRF and HPA axis activity remained hypersensitive to other stressors later in life.

2.1.2 CRF, the pathologic mediator of depression and anxiety. The preceding paragraphs suggest that the CRF system must play a critical role in the development of stress-related disorders. It is now becoming increasingly clear that depressive and anxiety disorders have significant inheritable and environmental components. The likelihood of an individual to develop an affective disorder is related to genetic vulnerability in addition to his or her biologic capacity to respond to stress. The experiments above suggest that this capacity is somewhat plastic, influenced by extent of early life stress. Multiple lines of evidence have shown that CNS CRF systems are altered in mood and anxiety disorders.

Many studies have reproduced the early findings that CRF CSF concentrations are elevated in depressed individuals compared to healthy comparison subjects, and that the levels resolve with resolution of the depressed state. Dysregulation of the HPA axis also continues to be validated. The CRF stimulation test (depressed patients show a blunted ACTH response) along with the Dexamethasone Suppression Test (depressed patients do not suppress cortisol) as well as the combined Dex/CRF test developed by Holsboer remain very sensitive tests of major depressive disorder. Furthermore, MRI and CT evidence has revealed increased pituitary and adrenal gland size in depression, consistent with hypersecretion of CRF and ACTH, respectively.

The role of CRF in anxiety disorders is also likely critical, though less well established than in depression. In animal models, direct CNS CRF administration is markedly anxiogenic, and conversely CRF receptor antagonists are anxiolytic. In experimental models, alprazolam, a short acting benzodiazepine, leads to decreased *locus coeruleus* CRF concentrations after acute administration, an effect which may mediate its anxiolytic action. Patients with post-traumatic stress disorder (PTSD) also exhibit elevated CRF concentrations in CSF and a blunted ACTH response to CRF challenge. However, they are hypocortisolemic and exhibit supersuppression to dexamethasone, suggesting that they differ from patients with depression.

In summary, the effects of CRF administration mimic acute anxiety and chronic depression, CRF₁ receptor antagonists block some of these symptoms in model systems, and patients with depressive and anxiety disorders show alterations of the central and peripheral CRF-HPA system. These data are nicely linked to a growing understanding of the developmental role of stress on the CRF system and its possible *sequellae* later in life. This allows a reformulation of the stress-diathesis model: early untoward life

events associated with development of depression and anxiety in adulthood give rise to long-lasting changes in CRF neurons, thus increasing the individuals' vulnerability to affective and anxiety disorders. See *Depression; Anxiety and Anxiety Disorders*.

2.2 Oxytocin and Vasopressin—Models for Complex Behavioral Regulation

Oxytocin (OT) and vasopressin (AVP) are members of the same nine amino acid class of peptides and have similar behavioral effects. Due to their interactions and similar effect on behavior, they will be considered together here. These peptides are best characterized in terms of organizing a complex behavior, that of social affiliation. This includes reproductive behavior, parental behavior, parent–child bonding, and grooming—sitting together behaviors in nonhuman primates. They are recently evolved and mediate many mammalian-specific reproductive and social behaviors. However, there are profound species differences in receptor expression, which in some cases mediates species-specific sets of behavior.

2.2.1 Oxytocin and vasopressin mediate social interaction. The best known peripheral effects of OT are facilitation of lactation–milk ejection, and uterine contraction during labor. AVP serves in the periphery primarily to regulate blood pressure and plasma volume. These functions are mediated via the magnocellular neurons projecting from the paraventricular nucleus (PVN) of the hypothalamus to the posterior pituitary where the peptides are released into the general circulation. Central effects are mediated via the PVN projection of parvocellular neurons to areas of the limbic system, neocortex, and autonomic areas of the brainstem. These central effects include modulation of reproductive behavior, parental behavior, infant attachment, and other prosocial effects.

One of the best studied models of affiliation is the formation of pair bonds. Multiple lines of evidence have demonstrated that in several rodent model systems, OT and AVP are critical for these behaviors. The prairie vole has been an excellent model for these studies because these animals are highly affiliative, show strong monogamous behavior, frequent physical contact, nest building, and shared parental care. OT in females and AVP in males appears to be responsible for formation of partner preference. In this species, mating facilitates pair bond formation, and OT and AVP are released centrally with mating. OT antagonists in females and AVP antagonists in males block formation of the partner preference, whereas the respective agonists administered centrally facilitate partner preference in the absence of mating. These results suggest that these peptides are both necessary

and sufficient for this aspect of pair bonding. Furthermore, a closely related species, the montane vole, shares many similar nonsocial behaviors with the prairie vole, but they are generally isolative, are clearly not monogamous, and males show little parental care. The critical difference in species appears to be the regulation of the OT and AVP receptor distribution in the brain. For example, the prairie vole has OT receptors in regions important for reward (*nucleus accumbens* and prefrontal cortex), whereas the montane vole shows primary distribution in the lateral septum, possibly responsible for the peptide's effect on self-grooming, but poor socialization. This is supported by evidence that centrally administered AVP increases affiliative behavior in the monogamous prairie vole, but not the montane vole. Furthermore, transgenic mice expressing the AVP receptor in the prairie vole brain distribution pattern results in a prosocial response to centrally administered AVP.

Parental behavior and infant attachment also appear to be dependent on OT and AVP systems. Parturition is associated with significant shifts in maternal behavior in some species, including relentless nest building, licking, grooming, and protection of the pups. These behaviors are blocked by central administration of OT receptor antagonists. Neonatal voles crave social contact with 'distress calls' vocalized in as early as five-day old pups with maternal separation. Centrally administered OT in eight-day old pups results in significantly reduced distress calls, but no evidence of sedation or other behavioral change, suggesting that OT influences the maternal separation response.

Evidence for other social behaviors mediated by these systems include increased levels of social grooming by centrally administered OT or AVP, and the correlation of decreased plasma OT with increased social stressors in primates. Finally, there is an interesting cognitive response to these peptides. Oxytocin appears to facilitate extinction and attenuates passive avoidance. It is hypothesized that these effects may allow the relinquishing of normal social avoidance to establish critical social bonds.

In summary, oxytocin and vasopressin have important roles in the initial affiliative and territorial stages of reproductive behaviors that are dependent on gender and species, along with influences on parental, attachment, and group social behavior. These neuropeptides appear to alter the affective processing of social stimuli via regionally specific regulation of neuropeptide receptors.

2.2.2 Possible roles for oxytocin and vasopressin in psychopathology. The data for these neuropeptides in animal affiliation suggest that they might also be important in human psychopathology in which relationships or social attachments are abnormal. Measures

of CSF OT and AVP concentrations have yielded inconsistent results in patients with schizophrenia or major depressive illness, though only a handful of studies have been performed. However, post mortem studies have suggested significant changes in several pathological processes. Significant increases in the number of hypothalamus AVP and OT cells have been found in post mortem studies of depressed subjects. Post mortem studies also revealed a > 40 percent decrease in OT cells in the Prader–Willi syndrome, a genetic disorder notable for obesity, mental retardation, hyposexuality, and inappropriate behavior. Several studies have consistently found elevated OT concentrations in OCD subjects, along with normal to elevated concentrations of AVP. One possible explanation for a role in OCD would be the relationship between OT and AVP and grooming behavior and their role in extinguishing avoidant behavior.

Most interesting would be the human disorders along the schizoid spectrum to autism, in which limited interpersonal skills and social impairment are paramount. With autism, the onset is clearly developmental prior to three years old, with most families noting decreased social interest in the first months of life. It is clearly genetically transmitted with monozygotic twins having approximately 36–91 percent concordance, compared to < 0.1 percent in the general population. In one study of autistic children, plasma OT was about half that of age-matched controls. Furthermore, autistic children failed to show the normal developmental increase in OT compared to controls. Although many clinical studies remain to be done (no study of CSF samples is yet reported), the preclinical data suggest that the OT and AVP systems are critical for normal social behavior and memory. Studies are also yet to be reported on the schizoid and avoidant spectrum disorders, but given the specificity of the OT and AVP systems, one suspects that there may be some abnormalities. Given the complexity of OT and AVP receptor gene regulation with known polymorphisms in humans, one could hypothesize that some aspects of social temperament are genetically influenced by the precise expression patterns and regulation of these behaviorally relevant peptides and receptors.

2.3 Cholecystokinin and Neurotensin—Models of Complex Neuromodulators

Cholecystokinin (CCK) and neurotensin (NT) are unrelated, evolutionarily conserved peptides of 8 and 13 amino acids, respectively. They are discussed in this section as examples of the large group of peptides that less clearly integrate a specific behavioral repertoire. They are found in numerous tissues and subserve many different roles including local paracrine, endocrine, and neurotransmitter functions. However,

their similarities are intriguing. They share endocrine functions regulating feeding behavior and induction of satiety in the periphery along with neuromodulatory functions interacting with the primary reward and appetitive circuitry in the central nervous system. Furthermore, there is mounting evidence that they each may play significant roles in the pathophysiology of schizophrenia, and in the case of NT in the mechanism of action of antipsychotic drugs.

2.3.1 Cholecystokinin and neurotensin involvement in feeding, autonomic regulation, and nociception. CCK was one of the first of the gastrointestinal hormones discovered and one of the most abundant neuropeptides in the brain. In contrast, NT was initially discovered from bovine hypothalamic extracts and later found to be an important gastrointestinal hormone. In the periphery, CCK and NT are each released by the small intestine shortly after a meal, and remain elevated in the plasma for several hours. They both stimulate pancreatic and biliary secretion along with modulating small and large intestinal motility. Although they do not cross the blood–brain barrier, they have an apparently conserved function in the CNS. They are both thought to have a role in mediating satiety and inhibiting feeding within the CNS.

They share the property of regulating autonomic and nociceptive information from the midbrain to the cortex. CCK and NT released in the midbrain have been shown to directly modulate excitatory visceral transmission through the parabrachial nucleus to the thalamus. Furthermore, both have been shown to exert potent analgesic effects when injected directly into the CNS. Thus they appear to be involved in regulating sensory information representing the internal state of the organism. Modulation of autonomic outflow is evident by their role in regulating vascular tone and thermoregulation.

2.3.2 Cholecystokinin and neurotensin modulation of dopaminergic systems. In addition to the above functions, CCK and NT appear to be integrally involved in modulating the dopaminergic systems within the brain. Although they are co-localized with other neurotransmitters and modulate serotonin, acetylcholine, and other neuropeptides in some areas, there appears to be a consistent role for these peptides in modulating dopamine circuits. Dopamine (DA) is found in three principle pathways: the mesolimbic/mesocortical pathways involved in motivation, reward, and cortical processing; the nigrostriatal pathways involved in mediation of locomotor movements; and the tuberoinfundibular pathways controlling pituitary release of prolactin. The role of CCK

and NT in modulation of the mesolimbic circuitry originating in the ventral tegmental area (VTA) is of particular interest. Within this mesencephalic nucleus, NT and CCK are co-localized in many dopaminergic neurons that project to the *nucleus accumbens*. NT and CCK are also found in cortical neurons, likely modulating the terminal projections of dopaminergic axons.

Both NT and CCK induce an initial increase in firing rate of dopaminergic neurons when released into VTA, along with an increase in DA release in the *nucleus accumbens*. This increased firing appears to be due to inhibition of D2 autoreceptors by decreasing affinity for DA. This change is thought to occur via intracellular transduction mechanisms and allosteric receptor-receptor interactions between the D2 receptor and the CCK and NT receptors. In contrast, at high doses and with chronic treatment, these peptides significantly decrease spontaneous activity of dopaminergic firing via depolarization inactivation. Thus, these neuropeptides have the ability, *in vivo*, of modulating dopaminergic function at the level of the midbrain and in the projection areas.

Whether the varied roles of feeding behavior, autonomic and nociceptive regulation, and dopaminergic modulation are functionally related is unclear. Some have speculated that these peptides were evolutionarily involved in the primary functions of feeding and feeding regulation. The CNS must be able to organize homeostatic mechanisms around this critical behavior including cardiovascular tone and temperature regulation because much of the blood volume goes to the gut postprandially. Relative analgesia postprandially might be crucial for maintaining the primary function of digestion. Finally, as complex behavior evolved, these neuropeptide systems may have served to couple the primary reward of feeding to the central motivation/reward circuitry.

2.3.3 Cholecystokinin, neurotensin, and schizophrenia.

The underlying pathology in schizophrenia remains unknown. The empirical observation that all effective antipsychotics were also dopaminergic antagonists contributed to the original DA hypothesis. However, more recent research has failed to find consistent abnormalities within the dopaminergic system, *per se*, in schizophrenic patients. Additionally, evidence for disturbances in the glutamatergic, GABA-ergic, serotonergic, and cholinergic systems have accrued. One way to reconcile the vast data is that the activity of multiple neurotransmitter systems may be altered via a dysregulation of normal activity. Because neuropeptides are known to play an important role in neurotransmitter modulation, these systems have received increasing attention as both possible mediators of the pathophysiology in schizophrenia as well as potential targets for novel therapeutics.

Although several neuropeptide systems may contribute to the abnormal circuitry in schizophrenia, the majority of data point to NT and CCK as being the most likely candidates of the known neuropeptides. Their role as modulators of the dopaminergic system was outlined above. Additionally, both the serotonergic *raphe nuclei* and brainstem cholinergic nuclei are innervated by peptidergic neurons, especially NT. Within the DA system, these neuropeptides mimic the effects of antipsychotic medication in their acute activation of the VTA dopaminergic neurons and in their chronic inactivation of these neurons. Furthermore, treatment with an NT receptor antagonist appears to mimic atypical antipsychotics in that chronic administration leads to decreased VTA firing, with no effect on the nigrostriatal system. This effect is thought to be mediated via feedback onto the VTA from the prefrontal cortex where local injection of the NT antagonist has the same effect. This would also be consistent with newer theories of schizophrenia which invoke disruption of large mesolimbic-cortical activation loops as mediating the pathophysiologic events.

Behavioral studies with centrally administered NT and CCK also mimic effects of typical and atypical antipsychotic medication. They both decrease stimulant-induced locomotion and spontaneous locomotion. They also decrease avoidance, but not escape, responding in a conditioned avoidance paradigm. Perhaps most importantly, they both appear to effect models of sensorimotor gating, which is becoming increasingly accepted as a critical objective symptom of schizophrenia. The hypothesized decreased ability of these patients to screen for appropriate sensory data may lead to 'involuntary flooding' of indifferent sensory input. This might lead to cognitive abnormalities, thought disorganization, and the positive symptoms (hallucinations and delusions) which are hallmarks of schizophrenia. Prepulse inhibition (PPI) of the startle reflex (defined by decreased startle to sound if preceded by weaker sound), and latent inhibition (LI) (reduced associative learning if subject is pre-exposed to stimulus), are the two most common measures of sensorimotor gating. In humans, PPI and LI are disrupted in schizophrenic patients, and are normalized in some studies of schizophrenics treated with antipsychotic medication. In animal models, dopaminergic agonists and other psychomimetic compounds disrupt PPI and LI, and these are returned to normal with antipsychotic treatment. NT and to a lesser extent, CCK, also clearly modulate PPI and LI in animal models. These results provide substantial evidence for the role of NT and possibly CCK in the circuits that are affected by antipsychotic medication.

Finally, there is significant evidence of abnormal regulation of CCK and NT in the CNS of schizophrenic patients. Although no consistent changes have been found post mortem with NT levels, NT receptor binding has reproducibly been found to be decreased

in some cerebrocortical regions in schizophrenia. Concentrations of NT in the CSF are reproducibly decreased in nonmedicated schizophrenics. These levels return to normal with effective treatment, and lower NT concentrations have been correlated with more severe psychopathology, particularly negative symptoms. Similar findings of NT abnormalities have not been found in affective disorders, anorexia, or Alzheimer's disease, suggesting some specificity of the findings in schizophrenia. Post mortem CCK levels have been consistently decreased in cerebrocortical and limbic regions of schizophrenic subjects, but CSF CCK changes have not been as reproducible. See *Schizophrenia, Treatment of*.

3. Summary

Since the 1970s, the biology and behavioral roles of many neuropeptides have been elucidated. They have moved from a position of relative unimportance in behavioral neuroscience to pre-eminence. In their roles as neurotransmitter and neuromodulator, paracrine and endocrine hormone, individual neuropeptides may at times subserve different functions in the CNS and periphery. However, many neuropeptides share a conserved function organizing different neural systems with the periphery in behaviors important for coordinated activity of the organism.

Corticotropin-releasing factor is essential in the physiologic mediation of stress, and likely critical in the pathophysiology of depression and anxiety. Oxytocin and AVP subserve many social roles from bonding to parental behavior, and they may underlie some pathologic processes involving socialization such as autism. Finally, the peptides NT and CCK have similar roles in feeding, autonomic and analgesic regulation, and DA modulation in the brain. Their dysfunction may contribute to the pathophysiology of schizophrenia.

The knowledge of these neuropeptide systems in psychopathology provides wonderful opportunities for future rational therapeutic approaches. CRF₁ receptor antagonists show great promise in preclinical and early clinical trials for the treatment of depression and anxiety. Oxytocin has been found to improve socialization in some experiments of schizophrenic patients and may provide future hope for autistic disorders. Neurotensin and CCK receptor agonists provide important targets for future system-directed treatment options in schizophrenia. In summary, via their role in the organization of behavioral repertoires, neuropeptide systems may ultimately elucidate mechanisms and provide novel treatment options for many psychiatric diseases.

See also: Endocrinology and Psychiatry; Hypothalamic–Pituitary–Adrenal Axis, Psychobiology of; Neurotransmitters; Women and Psychiatry

Bibliography

- Arborelius L, Owens M J, Plotsky P M, Nemeroff C B 1999 The role of corticotropin-releasing factor in depression and anxiety disorders. *Journal of Endocrinology* **160**: 1–12
- Argiolas A, Gessa G L 1991 Central functions of oxytocin. *Neuroscience Biobehavioral Reviews* **15**: 217–31
- Bissette G, Nemeroff C B 1995 The neurobiology of neurotensin. In: Bloom F E, Kupfer D J (eds.) *Psychopharmacology: The Fourth Generation of Progress*. Raven Press, New York
- Brownstein M G 1993 Neuropeptides. In: Siegal G, Agranoff B, Albers W, Molinoff P (eds.) *Basic Neurochemistry*, 5th edn. Raven Press, New York
- Burbach J P, de Wied D (eds.) 1993 *Brain Functions of Neuropeptides: A Current View*. Parthenon, Carnforth, NY
- Fink H, Rex A, Voits M, Voigt J P 1998 Major biological actions of CCK—a critical evaluation of research findings. *Experimental Brain Research* **123**: 77–83
- Gariano R F, Groves P M 1989 A mechanism for the involvement of colocalized neuropeptides in the actions of antipsychotic drugs. *Biological Psychiatry* **26**: 303–14
- Hokfelt T G, Castel M-N, Morino P, Zhang X, Dagerlind A 1995 General overview of neuropeptides. In: Bloom F E, Kupfer D J (eds.) *Psychopharmacology: The Fourth Generation of Progress*. Raven Press, New York
- Insel T R, O'Brien D J, Leckman J F 1999 Oxytocin, vasopressin, and autism: Is there a connection? *Biological Psychiatry* **45**: 145–57
- Kincaid B, Binder E B, Nemeroff C B 1999 Does neurotensin mediate the effects of antipsychotic drugs? *Biological Psychiatry* **46**: 340–51
- Koob G F 1999 Corticotropin-releasing factor, norepinephrine and stress. *Biological Psychiatry* **46**: 1167–80
- Modahl C, Green L A, Fein D, Morris M, Waterhouse L, Feinstein C, Levin H 1998 Plasma oxytocin levels in autistic children. *Biological Psychiatry* **43**: 270–7
- Nemeroff C B (ed.) 1991 *Neuropeptides and Psychiatric Disorders*. American Psychiatric Press, Washington, DC
- Owens M J, Nemeroff C B 1992 The physiology and pharmacology of corticotropin-releasing factor. *Pharmacology Review* **43**: 425–73
- Plotsky P M, Owens M J, Nemeroff C B 1995 Neuropeptide alterations in mood disorders. In: Bloom F E, Kupfer D J (eds.) *Psychopharmacology: The Fourth Generation of Progress*. Raven Press, New York
- Rostene W H, Alexander M J 1997 Neurotensin and neuroendocrine regulation. *Frontiers in Neuroendocrinology* **18**(2): 115–73
- Smock T, Albeck D, Stark P 1998 A peptidergic basis for sexual behavior in mammals. *Progress in Brain Research* **119**: 467–81
- Wang Z X, Young L J, DeVries G J, Insel T R 1998 Voles and vasopressin: A review of molecular, cellular, and behavioral studies of pair bonding and paternal behaviors. *Progress in Brain Research* **119**: 483–99

K. J. Ressler and C. B. Nemeroff

Perception and Action

Perceiving without acting is hardly possible: scrutinizing an object visually presupposes directing the eyes at it, which sometimes involves moving the head or even

the whole body; a tactile investigation of an interesting object requires moving the fingers across its surface; and localizing a sound source is much easier when moving the ears and head (Dewey 1896). Likewise, acting without perceiving makes no sense; after all, actions, defined as goal-directed behavior, aim at producing some perceivable event—the goal. Performing an appropriate action requires perceptual information about suitable starting and context conditions and, in the case of complex actions, about the current progress in the action sequence. Thus, perception and action are interdependent. However, in the course of evolution humans have developed several ways to relate action to perception, ranging from simple and rigid stimulus–response (S–R) reflexes shared with many species to flexible and adaptive behavioral rules that can be tailored on the spot to the situation at hand.

1. Reflexes

In most species, behavior is triggered by the present situation and, thus, directly reflects the animal's immediate environmental conditions. Such reflexes can also be observed in humans, especially in infants, but here they constitute a negligible part of the behavioral repertoire. Interestingly, however, even reflexes already show the close mutual relationship between perception and action. Perhaps the best demonstration of this relationship is provided by the orientation reflex, which we experience when encountering a novel and unexpected event. On the one hand, this reflex inhibits ongoing actions and tends to freeze the body—a prime example of a stimulus–triggered response. At the same time, however, it draws attention towards the stimulus source by increasing arousal and facilitating stimulus-directed body movements. That is, the novel stimulus triggers actions that lead to a better perception of itself, thus producing a full S–R–S cycle.

Even though reflexes themselves represent a relatively inflexible way to coordinate perception and action, some researchers have suspected them to provide the basis for voluntary action (Easton 1972). For instance, the tonic neck reflex, an asymmetric pose observed in newborns with head and arm extended to one side and arm and leg flexed on the other, might facilitate the development of eye–hand coordination. Likewise, the stepping reflex, in which babies move their feet in succession when coming in contact with a solid surface, might underlie our ability to walk.

2. Associations

Human behavior (and that of other higher species) is surely much more flexible than exclusive control by reflexes would allow. Not only can we learn to react to

particular environmental conditions and situations in a certain way, we also can unlearn what we have acquired and learn new relationships between situations and actions. Our ability to associate actions with stimulus conditions was the major topic of American behaviorism around 1890–1940, when most of the basic principles of S–R learning were empirically established. In particular, stimuli and responses become associated only if they co-occur in time, if there is at least some degree of contingency between them, and if the response is judged to be appropriate (Thorndike 1927). Although the empirical approach and the theoretical language of behaviorism has long been abandoned, its results still have a major impact on modern connectionism, the attempt to model psychological phenomena by means of artificial neural or neurally inspired networks on computers or in robots.

A major objection against the behavioristic approach to S–R learning relates to the assumed role of action outcomes. In behavioristic theories, the outcome of a given action is only judged regarding its hedonic value: a positive evaluation results in strengthening the association between the action and its antecedents whereas a negative evaluation weakens the association. However, whether it feels good or bad, the outcome of an action also informs the actor about its consequences, that is, about what he or she can achieve by performing it (Tolman 1932). And as actions aim at producing intended outcomes, perception–action learning should not be restricted to forming stimulus–response associations but comprise response–effect associations also. Indeed, studies reveal that humans (and also rats and pigeons) do acquire very stable associations between their actions and the consequences that these actions produce. Moreover, there is evidence that these associations play a major role in the control of voluntary action. That is, people control their overt behavior by forming or reactivating perceptual representations of intended goal events, which through learning have become associated with the motor patterns that have been—and must be—carried out to reach them.

3. Rules

The ability to learn new relations between environmental conditions and appropriate behavior provides an enormous gain in flexibility for an individual in allowing it to adapt to environmental change. Yet, learning and relearning take time and require at least some experience with a given new condition—it is thus necessarily reactive and, in a way, conservative in reflecting what one has become used to. Indeed, many forms of human behavior show these characteristics, as witnessed by the difficulties of introducing new behavioral patterns regarding, say, women, ethnic minorities, or the use of environmental resources.

Nevertheless, there are many instances where people can switch between different reactions to the same stimulus conditions on the spot, and hence more or less independently of the amount of experience with the situation. For example, even though one may have used one route to go to work 1,000 times already, it is easy to take an alternative route from one day to another, without having to unlearn the old habit or to learn the new one, and although the fact that in the case of absent-mindedness one might again go by the old route shows that the overlearned associations still exist. Therefore, people must be able to choose deliberately among alternative ways to relate environmental situations to their actions, that is, to select one out of many possible S–R rules, and behave accordingly.

The interplay between overlearned S–R associations and the voluntary implementation of intentionally selected S–R rules ('habit' vs. 'will') was a major issue in the psychological literature between 1870 and 1935, and Narziss Ach (1935) was the first to study this interplay empirically in a systematic way. In his 'combined method' he first had subjects build up new, strong associations between nonsense syllables and one type of action, and then asked them to perform another than the practiced action to the same stimuli. The resulting increase in reaction time and errors as compared with actions with neutral stimuli was taken to represent the individual 'will power' needed to overcome the previously acquired habits. After many years of neglect, the issue of how people implement and switch between alternative S–R rules received renewed interest during the 1980s and 1990s, especially in experimental and neuropsychology (Monsell and Driver 2000). The (still preliminary) results of this research provide new insights into the relationship between, and the control of, perception and action. As one would expect, the implementation of S–R rules takes effort and time. If people are to apply several sets of S–R rules concurrently, or in short succession, their performance is impaired and more error prone. In particular, they will sometimes apply the wrong rule or apply the right rule at the wrong time, especially if effort and attention deteriorate. However, even under ideal conditions, intentional control is not absolute. That is, voluntarily implementing a particular rule does not exclude or prevent contributions from overlearned S–R associations (habits). Therefore, although any S–R rule can be implemented in principle, those rules are easier to implement and to apply if they are in accordance with natural S–R relations, acquired S–R associations, and previously implemented S–R rules than if not. For instance, performance is better if stimuli and responses have features in common than with arbitrary S–R relations, so that pressing a left or right key, or turning to the left or right, is easier if signaled by the location of a stimulus that also appears on the left or right side than by the color, shape, or meaning of a stimulus.

The ability to implement and switch between the most arbitrary S–R rules allows for the highest degrees of behavioral flexibility but it is at the same time costly in terms of attentional resources. Therefore, it makes sense to see intentional rule implementation and the more automatic guidance by overlearned S–R associations as mechanisms that complement each other: implementing and acting out the same rules repeatedly and with sufficient success provides the basis for forming increasingly stronger associations between the corresponding environmental conditions and actions, in this way transforming rules into habits.

4. Plans

In one sense, perception and action are interdependent to a degree that makes it difficult to say where action ends and perception begins—just think of the eye-movement example. In another sense, however, evolution has provided humans with the important ability to temporally decouple our actions from perception. That is, actions can be planned, prepared, and scheduled long before their environmental trigger conditions occur. Although we are only beginning to understand how action planning works, there is evidence that it can be likened to self-automatization.

At least three phases of planning can be distinguished. First, the features of the intended action effect (i.e., the goal) need to be specified, such as the direction or end location of a reach, and the effector to be used. Next, the features belonging to one plan are integrated (i.e., their cognitive representations are temporarily coupled), to avoid confusing features belonging to different, concurrently maintained action plans. This integration process will often involve linking the plan to its anticipated trigger conditions, and hence to representations of the event that is intended to trigger the planned action. That is, planning comprises anticipating both perceptual and action events. Accordingly, planning an action changes our perception. For instance, after having planned a particular action, action-related objects and events are more salient and more easily processed, action-related features of objects are more conspicuous than other features, and objects appearing at action-related locations catch more attention than others.

The final step of planning consists of initiating the planned action. Initiation may be triggered internally, such as when a plan is carried out immediately after construction, or externally, such as when the anticipated trigger event is perceived. Interestingly, this last step does not seem to underlie overly strict cognitive control. For instance, internal initiation is more or less indifferent to the content of the plan or the progress of planning, and externally triggered plans can be seen to be called up even under inappropriate conditions or at the wrong point in time. Thus,

planning is like automatizing oneself by means of delegating control to future, often environmental, events.

See also: Attention and Action; Perceptual Constancy: Direct versus Constructivist Theories; Vision for Action: Neural Mechanisms

Bibliography

- Ach N 1935 *Analyse des Willens*. Urban and Schwarzenberg, Berlin
- Dewey J 1896 The reflex arc concept in psychology. *Psychological Review* 3: 357–70
- Easton T A 1972 On the normal use of reflexes. *American Scientist* 60: 591–9
- Gibson J J 1979 *The Ecological Approach to Visual Perception*. Houghton Mifflin, Boston
- Jordan J S *Systems Theory and A Priori Aspects of Perception*. North-Holland, Amsterdam
- Monsell S, Driver J 2000 *Control of Cognitive Processes*. MIT Press, Cambridge, MA
- Piaget J 1946 *La Formation du Symbole chez l'Enfant*. Delachaux and Niestlé, Paris
- Prinz W 1990 A common coding approach to perception and action. In: Neumann O, Prinz W (eds.) *Relationships Between Perception and Action*. Springer, Berlin, pp. 167–201
- Thorndike E L 1927 The law of effect. *American Journal of Psychology* 39: 212–22
- Tolman E C 1932 *Purposive Behavior in Animals and Man*. Century, New York
- Turvey M T 1977 Preliminaries to a theory of action with reference to vision. In: Shaw R, Bransford J (eds.) *Perceiving, Acting, and Knowing*. Erlbaum, Hillsdale, NJ, pp. 211–65

B. Hommel

Perception and Cognition, Single-/Multi-neuronal Recording Studies of

Recording of isolated single neurons from brain provides a means of estimating the relationship between an animal's behavior and the activity of the brain region from which the neuron was recorded. In this manner, neurophysiologists seek to explain all behaviors in terms of neural activity, and pair all neural activity with some perceptual, behavioral or cognitive event. Thus, the ultimate goal of neural recording and analysis is to determine how activity within a brain area relates to behavioral and/or cognitive events, in essence, seeking to answer the question: 'What do neurons encode?' This question is answered in part by recording and analyzing one neuron at a time, but there are additional character-

istics of encoding which can only be answered by examining the interactions between neurons which likely underlie the overall activity of a brain region. The purpose of both single and multineuron recording, thus, is to obtain a more representative view of the neural activity in a particular brain region (Singer 2000).

Single and multiple neuron recordings are particularly powerful when applied to experiments in which the brain must form some neural 'calculation' in order to make the appropriate response. This can be manifested on many levels, from the purely mechanical, such as integration of multiple inputs in a simple neuronal circuit, to complex behavioral responses dependent on an animal's ability to 'remember' prior stimuli. Inasmuch as the recorded neural activity reveals a pattern that correlates to the cognitive process, the neural activity can therefore be considered to 'encode' information related to the process. This does not necessarily mean that the neural pattern is essential to cognition, but that the neural activity suggests information, which is available to the brain during cognition.

1. Single vs. Multiple Neuron Recording

Single neuron recording has revealed the function of many brain areas in terms of stimulus and response. For example, Hubel and Wiesel (1962) mapped columns of neurons in the occipital cortex that responded to specific visual stimuli. The identification of these 'ocular dominance columns'—in which successive columns responded to incremental rotations of a bar of light, and alternating regions mapped the left vs. right visual fields—were critical to understanding how visual information was encoded by the brain. Similarly, the identification of 'place cells' in hippocampus—neurons that fire only when the subject is in fairly limited regions of its environment (O'Keefe and Dostrovsky 1971), provided a correlation between neural activity and behavior, even if the exact purpose of that activity is still being debated. Single neuron recording has thus been used to identify the type of information processed by different brain areas, not just the five senses, but also control of muscles, attention, and memory. To this end, the brain has been mapped anatomically and, to a certain extent, neurophysiologically.

Broader questions of function and cognitive processing are difficult to answer with only single neuron recordings. In essence, all neurophysiological experiments use multiple neuron recording, since the activity of a single neuron is not necessarily representative of that brain area. Instead, many single neurons are recorded, then analyzed to see if a pattern is detected which is representative of neural activity in that region. As long as single neurons are recorded under con-

ditions that control variability between recording sessions, it is possible (within limits) to treat the collection of neural recordings as if they had been recorded simultaneously. Thus, single neurons are assembled in an 'ensemble' that represents the activity of multiple neurons from a single brain area. In this manner, Georgopoulos et al. (1986) observed that the activity of ensembles of neurons recorded from primate motor cortex predicted the direction of limb movement prior to actual movement. In this study, many neurons were recorded that exhibited different 'preferred' directions for limb movement. When the pattern of firing across neurons was analyzed, it was found that the direction of movement was represented by specific firing patterns prior to the actual movement. However, since the conditions under which two single neurons are recorded cannot be identical, the effect of those subtle variations on neural firing cannot be controlled. In addition, the activity of single neurons may be controlled by multiple stimuli, in which case a behavioral correlate may be shared or distributed across many neurons.

Recordings of multiple simultaneous neurons have played an important role in many exciting recent discoveries. Deadwyler and Hampson (1997) showed that patterns of neural activity in hippocampus predicted behavioral error in a short-term memory task. Laubach et al. (2000) similarly demonstrated that activity patterns in neocortical neurons predicted behavior as animals learned a motor task. Skaggs and McNaughton (1996) showed that during sleep, hippocampal ensembles 'played back' temporal firing patterns between neurons. The replayed patterns were identical to patterns generated by the animal's exploration of its environment during the waking state. In addition, Chapin et al. (1999) recently demonstrated direct neural control of a robotic arm by neural signals recorded from rats. The researchers detected patterns of neural activity in motor cortex that preceded limb movements required to operate the robotic arm. By programming a computer to detect the pattern in real-time, the computer was able to operate the arm without requiring the animal to move its limbs. These recent results indicate rapid progress toward goals of better understanding the relationship between the activity of neurons and the cognitive functions of the brain.

2. Recording Many Neurons

Multiple neuron recordings not only rapidly increase the total number of neurons characterized, but also allow examination of interactions between neurons that may be critical to cognitive processing. Information may simply be encoded as the mean amplitude of firing across neurons at a given point in time, or as correlation between neurons (von der Malsburg 1994), or as part of coherent oscillations between neurons

(Singer and Gray 1995). No matter how the conditions are controlled during single neuron recording, interactions between neurons cannot be determined unless neurons are recorded simultaneously. The major advantage, therefore, of recording from multiple neurons in any brain area, is the potential for deciphering the codes that a brain area generates during cognitive or behavioral responding. The techniques used to record and analyze the multineuron data are critical, and not all types of multiple neuron analyses are appropriate in all cases. Misapplication of an ensemble technique can at best misrepresent the data, or at the worst lead to unworkable models of brain action. Several common uses of multiple neuron or ensemble analyses are examined below.

2.1 Techniques for Multineuron Recording

The ionic currents and action potentials generated by neurons produce changes in the electrical potential of the environment surrounding the neuron. These extracellular potentials are volume-conducted through the electrically-conductive extracellular fluid, and can be detected by electrodes that are either in contact with tissue or the cerebrospinal fluid. Skull surface or brain surface electrodes detect summed electrical activity resulting from the action potentials produced by millions of neurons throughout the brain, with no differentiation into single neurons. A glass pipette electrode will record neural activity from a region hundreds of microns in diameter; however, as above, the recorded neural activity is difficult to resolve into single neuron action potentials. Isolation of single neuron activity is possible with metal microelectrodes made from electrically conductive, but nonreactive material, such as tungsten, platinum-iridium, or stainless steel. Such electrodes are constructed with tip diameters approximate to the soma diameter of the neurons to be recorded and are insulated everywhere except at that tip. Sharp electrode tips ($1\ \mu\text{m}$) allow the electrode to be inserted with minimal tissue damage and allow the tip to be insinuated close to one neuron. In addition, the high impedance (resistance) of the tip ($> 1\ \text{M}\Omega$) ensures that electrical potentials from more distal neurons are attenuated compared to potentials generated by proximal neurons. More recent developments have shown that stainless-steel wires with tips up to $25\ \mu\text{m}$ and tip impedances of $100\ \text{k}\Omega$ are capable of isolating single neurons if left implanted for days or weeks (McNaughton et al. 1983, Deadwyler and Hampson 1997). Single neuron action potentials are identified by biphasic waveform (typically ± 10 – $100\ \mu\text{V}$ amplitude, 500 – $1000\ \mu\text{s}$ duration). As long as the electrode tip is not moving relative to the neuron being recorded, each action potential generated by a single neuron should be the same. Therefore, the activity of a single neuron is isolated by identifying and recording all action potential 'spikes' with the

same characteristic duration and amplitude. Multiple single neurons can be identified from a single microelectrode by further identifying similar spikes from all those rejected, or by using multiple microelectrodes.

2.2 Types of Multiple Neuron Recording Electrodes

The simplest multiple electrode is a bundle of metal microelectrodes, of similar size and shape, with tips positioned to record from within a restricted area. These electrodes can contact each other without compromising recording ability, since they are insulated everywhere except at the tip. The advantage of microelectrode bundles is the large number of recording sites in a small area, the disadvantage is that there is no means of identifying where any one electrode tip is located. When electrode location is desired, arrays of microelectrodes can be used. The electrode tips are spaced in a pattern, so that each tip records from a separate, identified site. Another form of bundle electrode is the stereotrode or tetrode consisting of two or four wires twisted to keep all electrode tips in close proximity. The close proximity of recording tips on stereotrodes and tetrodes allows the recording of neurons on more than one wire, improving the ability to isolate single neurons. Finally, as stereotrodes and tetrodes can be constructed as arrays, to provide all of the benefits of both electrode types.

2.3 Discrimination and Identification of Multiple Neurons

As stated above, the essential principle of single neuron isolation and identification is the fact that the action potential from a given neuron should always be the same amplitude and duration. Thus, action potentials with the same waveform characteristics, likely were generated by the same neuron. The first technique for identifying single neurons recorded from microelectrodes is to sort the extracellular action potentials by waveform. Under some circumstances, notably when recorded from dense monolayers of neurons, two neurons may be equidistant from the electrode tip, and hence produce action potentials of similar sizes and shapes. Stereotrodes and tetrodes were developed to identify single neurons under these conditions (McNaughton et al. 1983), since it is unlikely that both neurons would be equidistant from both electrodes in a stereotrode, or from all four electrodes in a tetrode. By comparing the action potential waveforms simultaneously recorded at all electrode tips, it is possible to 'triangulate' and separate neurons that may have appeared the same to a single microelectrode.

Statistical techniques are also used to ensure that the action potentials recorded on single or multiple microelectrodes originate from a single cell. The intervals between action potentials of a randomly firing single neuron (the neuron's 'spike train') should be a *Poisson* function. A plot of the interspike-interval

(i.e., between action potentials) should be similar to a graph of the *Poisson* probability distribution. In addition, autocorrelation of the spike train should reveal a symmetric distribution, with no firing within the time interval equal to the action potential refractory period for that neuron. Uniform interspike-interval distributions, or spikes occurring during the refractory period indicate a failure to isolate recordings into single neuron data (see Nicolelis 1999). These analyses, plus consistency in behavioral correlation, cross-correlation between neurons, and stability of action potential waveforms over an extended time (or multiple days) all indicate satisfactory identification of the single neuron (whether recorded with single wires, stereotrodes or tetrodes) and thus support the recording of multiple single neurons from bundled or arrayed microelectrodes.

3. Analysis of Multiple Neuron Recordings

The basic method of multiple neuron recordings is to analyze each neuron's spike train independently. This typically involves constructing a plot of average neural activity correlated to a specific cognitive event. For instance, perievent histograms depict mean firing rate synchronized to (peri-) a behavioral stimulus (event); similarly, a 'place field' recorded for hippocampal neurons consists of the correlation of neural firing rate with the animals instantaneous location in its environment. Multiple neuron activity is then used to examine to overall average or range of responses for neurons in a given brain area. More complex analyses of multiple spike trains were first described by Gerstein and Perkel (1969) prior to development of many of the above multiple neuron recording techniques. These techniques concentrated on determining cross-correlations between neurons, and statistically extracting repeated patterns of activity across many neurons with a view toward determining the connections between neurons. Many later analyses derived from these earlier techniques.

3.1 Population Vectors

A popular type of analysis is a direct examination of ensemble firing rate patterns corresponding to specific motor movements or behavioral events. This analysis has been termed the 'population vector' and is potentially quite powerful, especially in terms of encoding directional and velocity parameters. Population vectors consist of a set of measurements of the mean firing rates for each neuron in the ensemble corresponding to discrete times when unique events, or stimuli occur. The utility of a population vector analysis depends on identifying discrete conditions or 'states' (such as tracking visual images, or degrees of rotation of a monkey forearm), which correspond to a unique set of mean firing rates across the ensemble for each state. As applied by Georgopoulos et al. (1986),

the mean firing of each neuron in an ensemble recorded in non-human primate motor cortex was computed in correlation to forearm movements. A set of population vectors was obtained which corresponded to specific angles of arm rotation, then successfully compared to new recordings to test the accuracy of the population vector which 'predicted' the forearm angle prior to movement.

3.2 Cross-correlations

Multiple neuron recordings can also be analyzed simply by looking for repeating patterns of firing across populations of recorded neurons. These patterns may consist of spatial combinations of two or more neurons within the ensemble, a temporal pattern within single neurons, a particular sequence of firing between several neurons, or combinations of spatial and temporal oscillations. Ensemble spike trains can therefore be scrutinized using cross-correlation techniques and the construction of cross-correlation histograms (CCHs; Gerstein and Perkel 1969) for this purpose. Detection of specific spatiotemporal firing patterns (i.e., correlations between neurons) may suggest ways in which those neurons may be connected, or may simply result from common 'driving' influences on several different cells. The discovery that hippocampal neurons 'replay' recent experiences during sleep (Skaggs and McNaughton 1996) relied on the observation that temporal offsets in CCHs between pairs of hippocampal place cells corresponded to the sequence in which those place cells were activated. During the waking state, the temporal offset reflected the sequential movement of an animal through different places in the environment. During the following sleep cycle, the hippocampal neural pairs were again active with the same temporal offsets, suggesting a memory of the exploring the environment. One drawback in cross-correlation studies, however, is that while it is appealing to use cross-correlations to map temporal connectivity and patterned firing between neurons within an ensemble, such procedures are not well suited to circumstances where nonstationary processes (i.e., behavior) are likely to alter temporal firing patterns while the data is being recorded (see Hampson and Deadwyler 1999). On the other hand, the demonstration that cross-correlations between hippocampal neurons were altered provided important confirmation of the perturbations in place cell firing and spatial navigation observed in transgenic mice with deficient NMDA-receptors on CA1 hippocampal neurons (McHugh et al. 1996).

3.3 Multivariate Analyses

More elaborate analyses of ensemble spike trains include multivariate statistical treatments such as linear discriminant analysis and principal components

analysis (Nicolelis 1999). The differences and advantages of these analyses compared to those cited above are many. For instance, multivariate analyses simultaneously take account of the activity of all neurons in the ensemble, compared to cross-correlations which operate on pairs of neurons. Multivariate analyses incorporate both spatial (i.e., relationship between neurons within the ensemble) and temporal (time before and after events) information in the analyses, unlike population vectors which incorporate the firing of all neurons, but within an 'instantaneous' average of ensemble activity at discrete points in time. Finally, instead of producing one measure of correlation with behavioral and cognitive events for a given neuron, the multivariate components derived independent sources of variance within and between neurons comprising an ensemble.

Multivariate analyses provide the greatest potential for detecting neural patterns that correlate with perceptual, behavioral or cognitive event. Recordings of inferior temporal cortex (Miller et al. 1993), revealed many different patterns of encoding, including stimulus-specific and task (i.e., memory)-specific neural patterns. Likewise, analysis of ensembles of neurons from hippocampal (Deadwyler and Hampson 1997), parahippocampal regions (Young et al. 1997), and orbitofrontal cortex (Schoenbaum and Eichenbaum 1995) have demonstrated a broad spectrum of encoding that encompasses. Stimulus representation, memory storage, and formation of decision appropriate to the behavioral task.

4. Applications

Perhaps the best example of the utility of multiple neuron recording and analysis described above is the control of a robotic arm cited earlier (Chapin et al. 1999). In the study, ensembles of motor cortex neurons were recorded while a rat pressed a lever to cause a robotic arm to deliver water. A search for patterns of ensemble firing corresponding to the behavioral response would have required a massive search algorithm and intensive computation (as originally described Gerstein and Perkel 1969). However, principal components analysis of the ensemble spike trains revealed a pattern of neural firing that closely correlated to (and preceded) the limb movements required to press the lever and move the arm. Detection of the ensemble firing pattern while the animal performed the task allowed a computer to control the robotic arm, eventually allowing the rat to obtain water on the basis of the multiple neuron activity, without the requisite limb movement. Hence the principal components analysis allowed derivation of the brain function (behavioral response) from multiple neuron recording.

Single and multiple neuron recording have been applied to many purposes. Examination of the firing characteristics of single neurons provides a means of

correlating the activity of each neuron with the behavioral or cognitive activity of the animal. Multiple neuron recording is used simply to increase the yield of single neuron experiments, or for a more in-depth approach to correlating brain activity with external events. Analyses of the connectivity between neurons provides a framework for modeling neural circuits. Finally, extracting patterns of neural activity correlating to specific behavioral responses can result in a more detailed understanding of the role of specific brain areas in the process of cognition.

See also: Binding Problem, Neural Basis of; In Vitro Neurophysiology; Neural Synchrony as a Binding Mechanism; Neurons and Dendrites: Integration of Information

- Schoenbaum G, Eichenbaum H 1995 Information coding in the rodent prefrontal cortex. II. Ensemble activity in orbitofrontal cortex. *Journal of Neurophysiology* **74**: 751–62
- Singer W 2000 Why use more than one electrode at a time? In: *New Technologies for the Life Sciences—A Trends Guide*. Elsevier, pp. 12–17
- Singer W, Gray C M 1995 Visual feature integration and the temporal correlation hypothesis. *Annual Review of Neuroscience* **18**: 555–86
- Skaggs W E, McNaughton B L 1996 Replay of neuronal firing sequences in rat hippocampus during sleep following spatial exposure. *Science* **271**: 1870–3
- Von der Malsburg C 1994 The correlation theory of brain function. In: Domani E et al (ed.) *Models of Neural Networks II*. Springer, Berlin
- Young B J, Otto T, Fox G D, Eichenbaum H 1997 Memory representation within the parahippocampal region. *Journal of Neuroscience* **17**: 5183–95

R. E. Hampson and S. A. Deadwyler

Bibliography

- Chapin J K, Moxon K A, Markowitz R S, Nicolelis M A 1999 Real-time control of a robot arm using simultaneously recorded neurons in the motor cortex. *Nature Neuroscience* **2**: 664–70
- Deadwyler S A, Hampson R E 1997 The significance of neural ensemble codes during behavior and cognition. In: Cowan W M, Shooter E M, Stevens C F, Thompson R F (eds.) *Annual Review of Neuroscience*. Annual Reviews Inc., Palo Alto, CA, Vol. 20, pp. 217–44
- Georgopoulos A P, Schwartz A B, Kettner R E 1986 Neuronal population encoding of movement direction. *Science* **233**: 1416–19
- Gerstein G L, Perkel D H 1969 Simultaneously recorded trains of action potentials: analysis and functional interpretation. *Science* **164**: 828–30
- Hampson R E, Deadwyler S A 1999 Pitfalls and problems in the analysis of neuronal ensemble recordings during performance of a behavioral task. In: Nicolelis M (ed.) *Methods for Simultaneous Neuronal Ensemble Recordings*. Academic Press, New York, pp. 229–48
- Hubel D H, Wiesel T N 1962 Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *Journal of Physiology* **160**: 106–54
- Laubach M, Wessberg J, Nicolelis M A 2000 Cortical ensemble activity increasingly predicts behaviour outcomes during learning of a motor task. *Nature* **405**: 567–71
- McHugh T J, Blum K I, Tsien J Z, Tonegawa S, Wilson M A 1996 Impaired hippocampal representation of space in CA1-specific NMDAR1 knockout mice. *Cell* **87**: 1339–49
- McNaughton B L, O'Keefe J, Barnes C A 1983 The stereotrode: A new technique for simultaneous isolation of several single units in the central nervous system from multiple unit records. *Journal of Neuroscience Methods* **8**: 391–7
- Miller E K, Li L, Desimone R 1993 Activity of neurons in anterior inferior temporal cortex during a short-term memory task. *Journal of Neuroscience* **13**: 1460–78
- Nicolelis M 1999 *Methods for Simultaneous Neuronal Ensemble Recordings*. Academic Press, New York
- O'Keefe J, Dostrovsky J 1971 The hippocampus as a spatial map. Preliminary evidence from unit activity in the freely-moving rat. *Brain Research* **34**: 171–5

Perception: History of the Concept

'Perception' means an awareness of something, whether one's own thoughts and feelings, one's social surroundings, a business opportunity, the way to solve a math problem, or the current spatial layout. The term's meaning is often narrowed to basic aspects of perception with the senses. This article examines the history of the concept of sensory perception in philosophy and psychology. Perception has been studied as a source of knowledge and as a mental phenomenon in its own right. Visual perception was the first cognitive ability to be investigated with the methods of mathematical science, by the ancient Greeks, and was a central topic for the new philosophy and science of the seventeenth century. Quantitative studies of sensation and perception spurred the growth of experimental psychology in the nineteenth century. The twentieth century saw new perceptual theories and new techniques for examining the role of brain processes. Full understanding of how conscious awareness occurs in perception remains a goal for further research.

1. Perception in Ancient and Medieval Philosophy and Science

Human beings (as other animals) rely on their senses for survival. Although everyone uses the senses, it is not obvious how they work. Initial questions concerned their basic operation and their role in gaining knowledge. In vision, does something go out from the eye to meet the object, or does the object send

something to the eye? What do the senses reveal? Do they inform the observer of the very nature of things or are they fundamentally illusory? Or is their function primarily pragmatic?

Aristotle, working in fourth-century (BCE) Athens, wrote the first general treatise on psychology, described as *logon peri tes psyches* and entitled in Latin translation *De anima*, 'on the soul.' He considered the soul to be the principle of life or vitality, including reproduction and growth, sensory perception and purposeful activity, and intellectual theorizing. He offered a general theory of how the senses work, whereby something travels through a medium to each sense organ: light and color through air in the case of sight, vibrations in the case of sound, odors for smell, and so on. He also held that all knowledge is based on the senses. For 'proper sensibles' such as light and color, he postulated that a sample of the quality is received by the sensory soul 'without the matter' of the object. His medieval interpreters described this as the reception of a 'similitude' of the quality, or a 'real quality.' Aristotle held that sensory qualities alone do not reveal the essences of things, which are perceived by a higher cognitive faculty, the intellect, from observing changes over time. He differed from his teacher Plato, who held that in vision something goes from the eye into the air, and who taught that sensory perception is illusory, true knowledge arising from intellectual perception independent of the senses.

In this period 'optics' was about vision, not merely the physics of light (Lindberg 1976). In second-century Alexandria, Ptolemy developed a geometrical theory of vision, an extramission theory stating that something goes from eye to object. He modeled the relation between eye and object by a visual pyramid composed of a sheath of rays, with the object serving as base and the apex residing inside the cornea of the eye. The eye senses the direction and length of each ray, thereby perceiving the location of each point in the field of view, and so gaining full knowledge of the shapes, sizes, and places of things.

For centuries no theorist could formulate a mathematically rigorous intromission theory, with causation proceeding from object to eye. If one believes (correctly) that light is scattered from each point of an ordinary object in all directions, a problem arises that extramissionists did not face. It must be explained how the eye senses only one point for each point on the object, despite the fact that each point on the cornea receives light rays from all points in the field of view. Ibn al-Haytham (1030), working in Fatimid Cairo, produced the first accepted solution. He argued that the eye receives only the unrefracted rays (those that meet the cornea at right angles), and hence that it receives a cross-section of the visual pyramid preserving point-for-point the order of the scene. He also elaborated the theory that an object's true size is perceived through an unnoticed judgment that combines the visual angle under which an object is seen

with the perceived distance to the object. This explanation guided psychological theories of size perception for nearly a millennium.

2. Perception in the Modern Age

The Scientific Revolution of the seventeenth century brought changes in perceptual theory. Kepler's discovery of the retinal image revealed the internal optics of the eye and so produced a new solution to Ibn al-Haytham's problem, though without altering the latter's geometrical analysis of size perception or his theory of it, which was further developed by Descartes and Berkeley (Hatfield and Epstein 1979). Descartes postulated physiological systems to produce distance perception directly through the ocular-muscle processes of accommodation and convergence, and Berkeley sought to replace the psychology of unnoticed judgments with associative processes. Aristotle's theory that samples of qualities are absorbed by the senses was challenged by the new physics, according to which qualities in objects must be reduced to particles having only size, shape, motion, and position (primary qualities). Secondary qualities in objects, such as color, were now to be equated with such particles and their effects on perceivers. Descartes argued that sensory perception does not reveal essences, which are perceived by the intellect independent of the senses. The empiricist Locke denied intellectual perception of that sort, and held that the new physical theories were hypotheses to be tested against sensory observation. Both Descartes and Locke considered experienced qualities such as color to be products of the human sensory system. Contrary to later interpreters (e.g., Burt 1927), they did not regard them as illusory; if a subjectively produced color sensation is regularly or lawfully caused by qualities in objects, it can be used objectively to sort and classify such objects.

Eighteenth-century European theories of knowledge grappled with the new scientific image of the material world. Berkeley and Hume argued that the new theories devolve into paradox, as when 'red' is said both to be an experienced quality and a way light is reflected by surfaces. They sought to avoid skeptical conclusions by reducing knowledge of objects to the merely phenomenal. Reid (see Yolton 1984) countered with a 'common sense' affirmation of the reality and knowability of material objects (and of a deity, whose knowability was a primary motivation for him). Kant argued that analysis of the basis of actual knowledge in mathematics and natural science reveals that human knowledge cannot extend beyond sensory experience, so that metaphysical knowledge of God, an immaterial human soul, and the material world as it is in itself is not possible. Hegel countered that Hume, Kant, and others were wrong to think of the world in itself as lying beyond the phenomenally available world. The

natures of things are open to human knowledge, but can be grasped only through concepts that develop in a historical process of dialectic (see *Knowledge Representation*).

3. The Nineteenth Century

The nineteenth century was a golden age of sensory physiology and psychology in Germany (Turner 1994). G. Fechner, wanting to measure states of the soul, took psychophysical measurements for pressure on the skin, weights lifted by hand, line lengths perceived by sight, and the pitches of tones. Precise, empirical theories of color and spatial perception were developed by H. Helmholtz and E. Hering. In color theory, Helmholtz postulated three types of light-sensitive elements in the eye, with peak sensitivities at different wavelengths. Hering posited opponent-process mechanisms yielding four primary colors, red, green, yellow, and blue, along with a neutral white-black pairing. In spatial perception, Helmholtz argued that all ability for spatial localization is derived from experience, whereas Hering argued that the visual system has innate mechanisms for localizing sensations.

Nearly all major theorists accepted the assumption that perception is based on nonspatial, punctiform sensations varying only in quality and intensity. This assumption was shared by British associationists (e.g., J. Mill and J. S. Mill), nativists (H. Lotze), and empirists (W. Wundt, Helmholtz). It was questioned by Hering and C. Stumpf in Germany and W. James in the US. Dewey (1896) promulgated a functionalist approach to sensory perception, in which mind functions to adapt organisms to their environments.

4. Twentieth Century Trends

The idea that perception is built up from punctiform sensations was challenged by the Gestalt psychologists. Rather than punctiform sensations or two-dimensional perspective images, the Gestaltists regarded experience of the phenomenally given, meaningful world in three dimensions as the fundamental perceptual state. The American psychologist Gibson (1966) challenged the traditional theory that sensory stimulation provides greatly impoverished information about distant objects. He argued that, in vision, if one grants that organisms move and integrate information over brief intervals of time, then the information available at the eyes specifies fully the spatial layout of the scene. The British psychologist D. Marr developed computational models on the theory that Gibson-like assumptions about objects and stimulation are engineered into sensory systems. I. Rock at Rutgers University espoused the constructivist theory that perception results from inferential processes

which combine impoverished sources of stimulation, an approach shared by many information processing and computer vision theorists. Some theorists preferred to view perception as sensory information processing rather than the production of conscious states. New techniques for neurophysiological recording in living animals and brain imaging in humans permitted study of neural activity in perception, though behavioral and phenomenal data guided this research. At the end of the twentieth century all these currents remained visible in perceptual theorizing (Palmer 1999).

Philosophical approaches early in the twentieth century drew inspiration from Hume's phenomenalism and James' 'neutral monism,' the latter attempting to avoid the mind-body problem by arguing that both the psychological and the physical are composed from neutral existents found in perception. This led to theories of 'sense data' that have (in vision) the phenomenal properties of two-dimensional perspective projections. Such theories were challenged by critical and representational realism, and their phenomenological claims faded as philosophers took account of the Gestaltists and Gibson (Crane 1992). Philosophical interest in perception shifted from the problem of knowledge (now allotted to the philosophies of the sciences and other cognitive practices) to perception as an instance of the mind-body problem. As the neural correlates of sensory perception became better known, there remained a theoretical gap between neural or informational accounts and the conscious states manifest in perceptual experience. It is not clear how changes in the electrochemical state of neurons in the back of the brain could be the experience of red, despite it being known which neurophysiological patterns yield an experience of red, and which yellow. Much has been learned about how the senses work, but the characteristic feature of perception, conscious awareness, remains unexplained.

See also: Aristotle (384–322 BC); Fechnerian Psychophysics; Helmholtz, Hermann Ludwig Ferdinand von (1821–94); Hume, David (1711–76); Kant, Immanuel (1724–1804); Locke, John (1632–1704); Mind-Body Dualism; Perception and Action; Perception: Philosophical Aspects; Perceptual Constancy; Direct versus Constructivist Theories; Psychology: Historical and Cultural Perspectives; Psychophysical Theory and Laws, History of; Psychophysics; Wundt, Wilhelm Maximilian (1832–1920)

Bibliography

- Burr E A 1927 *Metaphysical Foundations of Modern Physical Science*. K. Paul, London
Crane T (ed.) 1992 *Contents of Experience*. Cambridge University Press, Cambridge, UK

- Dewey J 1896 The reflex arc concept in psychology. *Psychological Review* 3: 357–70 [In: Dewey J, *Early Works*, 1882–1896, Vol. 5. Southern Illinois University Press, Carbondale, IL]
- Gibson J J 1966 *Senses Considered As Perceptual Systems*. Houghton Mifflin, Boston
- Hatfield G C, Epstein W 1979 The sensory core and the medieval foundations of early modern perceptual theory. *Isis* 70: 363–84
- Ibn al-Haytham A 1030 *Kitāb al-Manāzīr* [1989 *The Optics of Ibn al-Haytham*, Books I–III. Warburg Institute, London]
- Lindberg D C 1976 *Theories of Vision from al-Kindi to Kepler*. University of Chicago Press, Chicago
- Palmer S E 1999 Vision science: Photons to phenomenology. MIT Press, Cambridge and London
- Ptolemy 1989 *L'Optique de Claude Ptolémée, dans la version latine d'après l'arabe de l'émir Eugène de Sicile*. E. J. Brill, Leiden and New York [1996 *Ptolemy's Theory of Visual Perception: An English Translation of the Optics*. American Philosophical Society, Philadelphia]
- Turner R S 1994 *In the Eye's Mind: Vision and the Helmholtz–Hering Controversy*. Princeton University Press, Princeton, NJ
- Yolton J W 1984 *Perceptual Acquaintance: From Descartes to Reid*. University of Minnesota Press, Minneapolis, MN

G. Hatfield

Perception of Extrapersonal Space: Psychological and Neural Aspects

We take for granted our ability to know where things are and to navigate through the environment. The effortless way in which we can look at or reach for an object belies the complex computations that go into determining its location. The psychological and neural processes that underlie our spatial abilities have been extensively explored in humans and animals. In this article, selected psychological aspects of spatial perception will be reviewed, followed by a consideration of some cortical mechanisms of spatial perception and representation.

1. Psychological Aspects

1.1 Basic Processes

Spatial perception involves not one but many specific abilities. Within the visual domain these include locating points in space, determining the orientation of lines and objects, assessing location in depth, appreciating geometric relations between objects, and processing motion, including motion in depth. These spatial skills can be applied to imagined objects as well as to external stimuli, as in the classic experiments of Shepard and Cooper (1986) on mental rotation. These

experiments showed that observers take longer to determine whether two objects are identical when the degree of mental rotation needed to align them increases. Similar sets of basic processes contribute to spatial representation in the auditory and somatosensory domains.

1.2 Space is Represented in Multiple Frameworks

Our subjective experience strongly suggests that we have direct access to a single coherent and overarching representation of space. Whether we localize a stimulus by sight, smell, hearing, or touch, we can respond to it with equal ease and with any motor system at our command. This introspection is misleading. There is no evidence for the existence of a single, explicit, topographic representation of space suitable for incorporating every kind of sensory input and generating every kind of motor output. On the contrary, the evidence points to multiple representations of space, in a variety of coordinate frames, and linked to separate output systems designed to guide specific motor effectors.

The particular representation of space in use at any time depends on the task the subject is trying to perform. For example, if you were to draw out a route for a hiking trip on a map, the route would be in the coordinates of the map, the piece of paper. If you were then going to walk along that route, you would have to begin by locating your current position within that coordinate frame and constructing a representation of the route with respect to your starting point. The first, map-based representation is an example of an allocentric representation, in which locations are represented in reference frames extrinsic to the observer. Allocentric representations include those centered on an object of interest (object-centered) and those in environmental (room-centered or world-centered) coordinates. The second representation, the one in the coordinates of the hiker's current position, is an example of an egocentric representation, in which locations are represented relative to the observer. Egocentric representations include those in eye-centered, head-centered, hand-centered, and body-centered coordinates. Experimental work in humans indicates that multiple reference frames can be activated simultaneously (Carlson-Radvansky and Irwin 1994).

2. Neural Aspects

2.1 Impairments of Spatial Perception in Humans

Much of our knowledge about the neural basis of spatial perception comes from observations in patients with spatial deficits following brain damage. These

include a wide range of perceptual and motor deficits, such as poor localization of visual, auditory, or tactile stimuli; inability to determine visual or tactile line orientation; impaired performance on mazes; impairment on tests of mental spatial transformations; right—left confusion; poor drawing; impaired eye movements to points in space; misreaching; defective locomotion in space; and amnesia for routes and locations. As can be seen from this partial list, spatial behavior involves many kinds of skills and, not surprisingly, a number of brain regions have been implicated in spatial perception and performance. Broadly speaking, the parietal lobe is responsible for spatial perception and representation of immediate extrapersonal space, while temporal and parahippocampal cortices are more involved in topographic memory and navigation (see *Navigation in Virtual Space: Psychological and Neural Aspects*). The frontal lobe receives input from both parietal and temporal cortex and is responsible for generating actions.

An important point about deficits in spatial perception is that they are far more common after right hemisphere damage than left. While patients with left hemisphere damage may also exhibit spatial deficits, it is clear that the right hemisphere in humans has a superordinate role in spatial processing and behavior (Heilman et al. 1985). Two kinds of spatial deficits following brain damage are particularly illuminating. First, a common sensorimotor deficit following parietal lobe damage is difficulty in using visual information to guide arm movements, referred to as optic ataxia. Patients with optic ataxia have difficulty with everyday tasks that require accurate reaching under visual guidance, such as using a knife and fork. They both misdirect the hand and misorient it with respect to the object, and are most impaired when using the contralesional hand to reach for an object in the contralesional half of space (see *Vision for Action: Neural Mechanisms*).

A second, classic disorder of spatial perception in humans is the tendency to ignore one half of space, called hemispatial neglect (see *Neglect, Neural Basis of*). The most common form of neglect arises from damage to the right parietal lobe and is manifested as a failure to detect objects in the left half of space. Neglect is more than just a visual deficit, however. It can occur separately or jointly across many sensory modalities (Barbieri and De Renzi 1989). Moreover, neglect occurs with respect to many different spatial reference frames. A patient with right parietal lobe damage typically neglects objects on the left but left may be defined with respect to the body, or the line of sight, or with respect to an attended object. Further, this neglect is dynamic, changing from moment to moment with changes in body posture and task demands (Behrmann 2000). Neglect is apparent even in the purely conceptual realm of internal images. Patients exhibit neglect when asked to imagine a familiar scene, such as a city square, and describe the buildings in it. The

portion of space that is neglected changes when they are asked to imagine the scene from a different viewpoint (Bisiach and Luzzatti 1978). As this example illustrates, neglect can occur with respect to an internal image constructed by the individual. Patients with neglect show evidence of using multiple reference frames, just as intact individuals do.

2.2 Spatial Perception in Animals is Impaired by Parietal Cortex Lesions

The role of parietal cortex in spatial perception has been explicitly tested in animal studies. Monkeys with posterior parietal lobe lesions exhibit many of the same deficits seen in humans, including deficits in the appreciation of spatial relations between objects and impairments in eye movements and reaching. They perform normally on tests of object discrimination but are selectively impaired on a spatial task that requires them to judge which of two identical objects is closer to a visual landmark. In contrast, monkeys with temporal lobe lesions are unimpaired on the spatial task but fail to discriminate between or recognize objects (Ungerleider and Mishkin 1982). This double dissociation between the effects of lesions, in combination with the discovery of distinctive cortical inputs to the parietal and temporal lobes, led to the concept of the dorsal and ventral visual processing streams (see *Visual Systems: Dorsal and Ventral*).

Neurons in the dorsal stream encode the types of visual information necessary for spatial perception. Neurons in specific dorsal stream areas are selective for orientation, depth, direction and speed of motion, rotation, and many other stimulus qualities appropriate for perceiving spatial information. The dorsal visual processing stream leads to posterior parietal cortex, where many kinds of visual information converge, including information about stimulus shape. An equivalent set of somatosensory processing areas send tactile information to anterior parietal cortex. Visual and somatosensory signals converge on single neurons within the intraparietal sulcus, which divides anterior and posterior parietal cortex. Auditory signals have also been demonstrated to contribute to spatial processing in monkey parietal cortex.

2.3 Parietal Neurons Encode Spatial Information in Multiple Reference Frames

The standard approach for investigating the role of parietal neurons in spatial perception is to record electrical activity from individual neurons while the monkey performs a spatial task. Because brain tissue itself has no sensory receptors, fine microelectrodes can be introduced into the brain without disturbing the animal's performance. By recording neural responses during carefully designed tasks, neural activity

can be related directly to the sensory and representational processes that underlie spatial behavior. The general conclusion from these studies is that the function of parietal cortex is to transform spatial information from sensory coordinates into the motor coordinates that are necessary for the guidance of action (Stein 1992, Andersen et al. 1997, Colby and Goldberg 1999).

2.3.1 Head-centered spatial representation in the ventral intraparietal area (VIP). Area VIP is located in the floor of the intraparietal sulcus, where inputs from high-order visual and somatosensory cortex converge. In the visual domain, VIP neurons are characterized by direction and speed selectivity, and thus resemble neurons in other dorsal stream visual areas that process stimulus motion (see *Motion Perception: Psychological and Neural Aspects*). In the somatosensory domain, these same neurons respond to light touch on the head and face. The somatosensory and visual receptive fields of individual neurons exhibit strong spatial correspondence: they match in location, in size, and even in their preferred direction of motion. The existence of spatially matched receptive fields raises an interesting question: what happens when the eyes move away from primary position? If the visual receptive fields were simply retinotopic, they would have to move in space when the eyes do and so would no longer match the location of the somatosensory receptive field. Instead, for some VIP neurons, the visual receptive field moves to a new location on the retina when the eyes move away from the straightahead position. For example, a neuron that has a somatosensory receptive field near the mouth and responds best to a visual stimulus moving toward the mouth will continue to respond to that trajectory of motion regardless of where the monkey is looking (Colby et al. 1993). These neurons have head-centered receptive fields: they respond to stimulation of a certain portion of the skin surface and to the visual stimulus aligned with it, no matter what part of the retina is activated. Neurons in area VIP send projections to the region of premotor cortex (see Sect. 2.4.1) that generates head movements. Area VIP neurons thus contribute to the visual guidance of head movements and may play a special role in hand, eye, and mouth coordination. They operate in a head-centered reference frame in order to generate appropriate signals for a particular motor effector system, namely that which controls head movements.

2.3.2 Eye-centered spatial representation in the lateral intraparietal area (LIP). In contrast to area VIP, neurons in area LIP construct an eye-centered spatial representation. Individual neurons become active

when a salient event occurs at the location of the receptive field. This can be a sensory event, such as the onset of a visual or auditory stimulus, or a motor event, such as a saccade towards the receptive field, or even a cognitive event, such as the expectation that a stimulus is about to appear, or the memory that one has recently appeared. The level of response reflects the degree to which attention has been allocated to the site of the receptive field (Colby et al. 1995).

Again, as we saw for area VIP, the animal's ability to make eye movements raises an interesting question about spatial representation. Neural representations of space are maintained over time, and the brain must solve the problem of how to update these representations when a receptor surface moves. With each eye movement, every object in our surroundings activates a new set of retinal neurons. Despite this constant change, we perceive the world as stable. Area LIP neurons contribute to this perceptual stability by using information about the metrics of the eye movement to update their spatial representation of salient locations (Duhamel et al. 1992). For example, LIP neurons are activated when the monkey makes an eye movement that brings a previously illuminated screen location into their receptive field. These neurons respond to the memory trace of the earlier stimulus: no stimulus is ever physically present in the receptive field, either before or after the saccade. The proposed explanation for this surprising finding is that the memory trace of the stimulus is updated at the time of the saccade. Before the saccade, while the monkey is looking straight ahead, the onset of the stimulus activates a set of LIP neurons whose receptive fields encompass the stimulated screen location, and they continue to respond after the stimulus is extinguished, maintaining a memory trace of the stimulus. At the time of the saccade, a corollary discharge, or copy of the eye movement command, containing information about the metrics of the saccade, arrives in parietal cortex. This corollary discharge causes the active LIP neurons to transmit their signals to the new set of LIP neurons whose receptive fields will encompass the stimulated screen location after the saccade. The representation of the stimulated location is thus dynamically updated from the coordinates of the initial eye position to the coordinates of the final eye position.

The significance of this observation is in what it reveals about spatial representation in area LIP. It demonstrates that the representation is dynamic and is always centered on the current position of the fovea. Instead of creating a spatial representation that is in purely retinotopic (sensory) coordinates, tied exclusively to the specific neurons initially activated by the stimulus, area LIP constructs a representation in eye-centered (motor) coordinates. The distinction is a subtle one but critical for the ability to generate accurate spatial behavior. By representing visual

information in eye-centered coordinates, area LIP neurons tell the monkey not just where the stimulus was on the retina when it first appeared but also where it would be now if it were still visible. The result is that the monkey always has accurate spatial information with which to program an eye movement toward a real or remembered target. The transformation from sensory to motor coordinates puts the visual information in its most immediately useful form. Compared to a head-centered or world-centered representation, an eye-centered representation has the significant advantage that it is already in the coordinates of the effector system that will be used to acquire the target. Humans with unilateral parietal lobe damage fail on an eye movement task that requires an eye-centered representation of a remembered target position. This failure presumably reflects an impairment of the updating mechanism in parietal cortex.

In sum, posterior parietal cortex plays a critical role in spatial perception and representation (Colby and Goldberg 1999). Physiological studies in monkeys show that parietal neurons represent spatial locations relative to multiple reference frames, including those centered on the head and the eye. Individual neurons (in area VIP) combine spatial information across different sensory modalities, and specific spatial reference frames are constructed by combining sensory and motor signals (in area LIP). In accord with the physiology, human neuropsychological studies show that neglect can be expressed with respect to several different reference frames.

2.4 Frontal Lobe Mechanisms of Spatial Representation

The parietal lobe transforms sensory representations of attended objects into the motor coordinate frames most appropriate for action. They do not actually generate those actions. Dorsal stream outputs to frontal cortex provide the sensory basis, in the correct spatial framework, for producing specific motor outputs. The following sections describe specific spatial representations found in three regions of frontal cortex.

2.4.1 Neurons in premotor cortex have head-centered and hand-centered visual receptive fields. Two different forms of spatially organized visual responsiveness have been described in ventral premotor cortex. First, in subdivisions representing facial movements, neurons respond to visual stimuli at specific locations relative to the head, much like those described above in area VIP. These neurons have been characterized by recording activity while objects approach the monkey's head along various trajectories. The preferred trajectory is constant with respect to the head,

and is not affected by changes in eye position (Fogassi et al. 1996). This specific zone of premotor cortex receives inputs from area VIP, and uses the same form of spatial representation. In both cases, stimuli are represented in the motor coordinate frame that would be most useful for acquiring or avoiding stimuli near the face by means of a head movement.

A different kind of spatial representation has been observed in subdivisions of premotor cortex involved in generating arm movements. Here, neurons respond to visual stimuli presented in the vicinity of the hand. When the hand moves to a new location, the visual receptive field moves with it. Moreover, the visual receptive field remains fixed to the hand regardless of where the monkey is looking, suggesting the existence of a hand-centered representation (Graziano et al. 1994). A fascinating observation indicates that some neurons in this area are capable of even more abstract forms of representation. Rizzolatti and coworkers have described neurons in premotor cortex that are activated both when the monkey grasps an object, such as a raisin, and when the monkey observes the experimenter performing the same action (Gallese et al. 1996). These 'mirror' neurons encode not just the monkey's own motor actions but also the meaning of actions performed by others (see *Pre-motor Cortex*).

2.4.2 Object-centered spatial representation in the supplementary eye field (SEF). Actions are directed toward objects in the environment and toward specific locations on an object. Picking up your coffee cup requires that you locate both the cup in egocentric space and the handle in relation to the cup. The spatial reference frame that guides such movements is not limited to the egocentric representations described above. Evidence from the supplementary eye field (SEF) demonstrates that single neurons can encode movement direction relative to the object itself (Olson and Gettner 1995). The SEF is a division of premotor cortex with attentional and oculomotor functions. Neurons here fire before and during saccades. In monkeys trained to make eye movements to particular locations on an object, SEF neurons exhibit a unique form of spatial selectivity: they encode the direction of the impending eye movement as defined relative to an object-centered reference frame. For example, a given neuron may fire when the monkey looks toward the right end of a bar placed at any of several different locations on the screen, regardless of whether the eye movement itself is a rightward or leftward saccade. Moreover, the same neuron will fail to respond when the monkey makes a physically identical eye movement toward the left end of the bar stimulus. This striking result indicates that single neurons can make use of quite abstract spatial reference frames. Object-centered spatial infor-

mation could potentially guide arm movements as well as eye movements. Moreover, neuropsychological evidence indicates that an object-centered reference frame can be used to direct attention: some patients exhibit object-centered neglect after parietal lobe damage (Behrmann 2000). Parietal lobe lesions in monkeys likewise produce impairments on tasks that require an object-centered spatial representation.

In sum, behavioral and neuropsychological studies indicate that we make use of multiple spatial representations in the perception of extrapersonal space. Neurophysiological studies are beginning to uncover the neural mechanisms underlying the construction of these egocentric and allocentric spatial representations.

See also: Hippocampus and Related Structures; Motion Perception Models; Motion Perception: Psychological and Neural Aspects; Navigation in Spatial Environments, Neural Basis of; Navigation in Virtual Space: Psychological and Neural Aspects; Spatial Cognition; Wayfinding and Navigation Behavior

Bibliography

- Andersen R A, Snyder L H, Bradley D C, Xing J 1997 Multimodal representation of space in the posterior parietal cortex and its use in planning movements. *Annual Review of Neuroscience* **20**: 303–30
- Barbieri C, De Renzi E 1989 Patterns of neglect dissociation. *Behavioural Neurology* **2**: 13–24
- Behrmann M 2000 Spatial reference frames and hemispatial neglect. In: Gazzaniga M (ed.) *The Cognitive Neurosciences*. MIT Press, Cambridge, MA
- Bisiach E, Luzzatti C 1978 Unilateral neglect of representational space. *Cortex* **14**: 129–33
- Carlson-Radvansky L A, Irwin D E 1994 Reference frame activation during spatial term assignment. *Journal of Memory and Language* **37**: 411–37
- Colby C L, Duhamel J R, Goldberg M E 1993 Ventral intraparietal area of the macaque: Anatomic location and visual response properties. *Journal of Neurophysiology* **69**: 902–14
- Colby C L, Duhamel J R, Goldberg M E 1995 Oculocentric spatial representation in parietal cortex. *Cerebral Cortex* **5**: 470–81
- Colby C L, Goldberg M E 1999 Space and attention in parietal cortex. *Annual Review of Neuroscience* **22**: 319–49
- Duhamel J R, Colby C L, Goldberg M E 1992 The updating of the representation of visual space in parietal cortex by intended eye movements. *Science* **255**: 90–2
- Fogassi L, Gallese V, Fadiga L, Luppino G, Matelli M, Rizzolatti G 1996 Coding of peripersonal space in inferior premotor cortex (area F4). *Journal of Neurophysiology* **76**: 141–57
- Gallese V, Fadiga L, Fogassi L, Rizzolatti G 1996 Action recognition in the premotor cortex. *Brain* **119**: 593–609
- Graziano M S A, Yap G S, Gross C G 1994 Coding of visual space by premotor neurons. *Science* **266**: 1054–6
- Heilman K M, Watson R T, Valenstein E 1985 Neglect and related disorders. In: Heilman K M, Valenstein E (eds.)

Clinical Neuropsychology. Oxford University Press, Oxford, UK

Olson C R, Gettner S N 1995 Object-centered direction selectivity in the macaque supplementary eye field. *Science* **269**: 985–8

Shepard R, Cooper L 1986 *Mental Images and their Transformations*. MIT Press, Cambridge, MA

Stein J F 1992 The representation of egocentric space in the posterior parietal cortex. *Behavioral and Brain Sciences* **15**: 691–700

Ungerleider L, Mishkin M 1982 Two cortical visual systems. In: Ingle D J, Foulke M A, Mansfield M A (eds.) *Analysis of Visual Behavior*. MIT Press, Cambridge, MA

C. L. Colby

Perception: Philosophical Aspects

Perception has been of perennial interest in philosophy because it is the basis of our knowledge of the external world. Because perception is a transaction between the world and the perceiver or their mind, investigations of the nature of the world perceived, the perceiver or their mind, or the relation between the two are possible. The philosophy of perception has focused on the latter two issues; although the study of perception has led to important views about the nature of the world perceived—particularly the idealism of Bishop Berkeley (see below)—such views are rare. Philosophers have asked instead: What is the nature of the relation between the mind of the perceiver and the world perceived? and What is the nature of the perceptual states of that mind? Contemporary philosophy of perception continues to debate these traditional questions, although it places more emphasis on the nature of the perceiving mind. It also tends to focus on visual perception rather than on the other senses, assuming—controversially—that what is true of vision is likely to be true of other modes of perception. Philosophers have also investigated the justification of perceptual belief, but because this is a branch of traditional epistemology, it is not dealt with here. This article considers some of the specific questions that have been raised in the investigation of the two general questions above as well as some of the answers given to them.

1. The Perceiver–World Relation

1.1 Direct Realism, Indirect Realism, and Phenomenalism

As I look out of the window, I see objects, such as buildings, and I see their properties, such as the green

color of their roofs. Presumably, the objects and their properties continue to exist even when I am not looking at them. (I can also be said to see events, such as the passing of pedestrians along the sidewalk, and facts, such as that there is traffic on the street, but we will ignore these matters here.) It seems to me that my acquaintance with the objects I perceive is an intimate one. I have no sense that anything stands between me and the buildings I am looking at. I do not feel that I am looking at a representation of the buildings as I might if I were looking at a photograph of them, nor that I am receiving information about them that has to be decoded. I merely open my eyes, and there they are.

This common sense view of visual perception is called *direct realism*. Direct realism holds that the objects one perceives are the physical objects of the external world and that these objects are perceived directly, that is, without the aid of any intermediary object or representation. A question arises about the status of the properties of objects when they are unperceived. *Naïve direct realism* holds that all of an object's perceived properties persist when it is unperceived; *scientific direct realism* holds that only the properties countenanced by science—such as shape—persist when an object is unperceived (see Dancy 1981, Martin 1998). Object color, for example, is thought by some to be a mind-dependent property and does not, on this view, persist when an object is unperceived. Scientific direct realism holds, therefore, that color is not perceived directly.

Despite its apparent obviousness, philosophers have challenged direct realism. An important objection to the view is expressed by the *argument from illusion*. Suppose I am having a hallucination of Beaver Lake. What is the object of my experience? Since I am not perceiving a physical object, I must be experiencing some other kind of object, perhaps a mental object. Further, since the experience of hallucinating Beaver Lake is, we may suppose, identical to the experience of actually perceiving Beaver Lake, the object in each case must be identical. In all true perception, therefore, there must be some (mental) object that acts as an intermediary or representation in the perception of the external world. The view that one is directly aware of intermediaries has been expressed as the claim that there is a *given* in perception—a nonconceptual state that forms the basis for full-blooded perception which requires the possession of concepts by the perceiver.

Various strategies have been adopted to resist the argument from illusion. First, the *adverbial theory of perception* denies the inference from the fact of hallucination to the claim that there is a mental entity that is the object of that hallucination. Instead, the theory holds that when I hallucinate a red triangle, for example, I am 'experiencing redly and triangularly', where 'redly' and 'triangularly' qualify the manner of experiencing, and does not presuppose the existence of a mental object. Because no inner mental object has to be posited even in cases of hallucination, it is possible

to reclaim the common sense view that in true perception it is physical objects that are directly perceived.

A second strategy is to deny the claim that the experience of true perception is indistinguishable from that of hallucination. The *disjunctive conception of perceptual experience* (see Snowdon 1990) holds that these experiences are constituted by different mental states. Thus, even if hallucinatory experience involves a mental representation, true perception need not, and could be direct.

A third strategy is to treat perception as a special kind of belief (or inclination to believe) rather than a sensory state (Armstrong 1961). On this view, to see a red triangle is to believe that there is a red triangle before one's eyes. A hallucination of a red triangle is no more than a false belief and does not, therefore, call for the introduction of a mental representation any more than the false belief that the local hockey team will win the Stanley Cup requires a mental object to be the thing believed. Some support for this view derives from the phenomenon of *blindsight* (Weiskrantz 1986), a state in which subjects with particular forms of brain damage are able to make crude perceptual judgements about the properties of stimuli in the absence of any conscious experience. If one is willing to accept the claim that blindsighted subjects are perceiving, then blindsight undermines the common sense view that all perception involves conscious experience. Since the perception-as-belief view does not make conscious experience central to perception, it is thus strengthened.

A fourth strategy for resisting the argument from illusion has been developed by information-theoretic approaches to perception which explicate the transaction between the world and the perceiver as an exchange of information (see Dretske 1981). For example, the *computational approach* to vision (Marr 1982) investigates how the information about properties of the environment contained in patterns of retinal stimulation can be extracted by the visual system.

Whatever the success of these strategies, direct realism also has trouble accounting for a different problem—that of perceptual error. If one's experience of the external world is mediated by a mental representation, then error can be explained as a mismatch between the representation and what it represents. But if there is no such representation but only a direct perception of the physical objects themselves, how is misperception possible?

These arguments support *indirect* or *representative realism* according to which external physical objects are perceived, but only indirectly, by means of the direct perception of intermediary entities (see Dancy 1981, Jackson 1977). These intermediaries are usually thought to be mental representations and have traditionally been called *ideas*, *appearances*, *sensibilia*, and *sensa*, among others. In the twentieth century they

have been referred to, rather notoriously, as *sense-data*. Much of the investigation into the nature of the relation between the perceiver and the world has focussed on two distinct questions: first, whether perception is in fact mediated in this way; and, second, whether one is consciously aware of the intermediaries. This investigation has been made more difficult by the fact that these quite different questions have often been conflated (Crane 1992).

Indirect realism also faces important objections, however; in particular, that it seems to lead to scepticism about the external world. If I am aware only of mental representations, then for all I know there is no external world at all. One response to this objection is to claim that it misconstrues the claims of indirect realism. The theory does not say that we have no access to external objects but rather that our access is by way of mental intermediaries. A second response embraces the sceptical possibility. Bishop Berkeley (1685–1753) famously argued that there is no external world but only perceptions, or, at any rate, that the external world is no more than collections of actual or possible perceptions, a view known as *phenomenalism* or *idealism*. Despite the importance of Berkeley's arguments for the theory of perception, however, the radical nature of his claims about the world have prevented them from being widely held.

Most philosophers of the first half of the twentieth century were indirect realists of the sense-datum variety. The early part of the second half of the twentieth century saw a complete rejection of sense-datum theory and indirect realism, and many contemporary philosophers are direct realists. Nevertheless, the tendency to think that there is some aspect of experience that involves a perceptual intermediary or that there are features of perceptual experience that cannot be construed as the direct perception of the external world has proved very strong. Some aspects of sense-datum theory have emerged again in the debate about qualia (see below), and even outright sense-datum theories are no longer anathema.

1.2 The Causal Theory

The *causal theory of perception* takes up a second important issue in the study of the relation between perceiver and world. The causal theory holds that the transaction between the perceiver and the world should be analyzed primarily in terms of the causal relation underlying that transaction (Grice 1961). One version of the causal theory claims that a perceiver sees an object only if the object is a cause of the perceiver's seeing it. Traditionally, indirect realism was associated with the causal theory, but not all recent versions of the causal theory are indirect theories (see Vision 1997). Indeed, the causal principle seems relevant to any view of perception, whether direct or indirect. This is illustrated by the case of a desert traveler who hallucinates an oasis. His eyes are closed, so nothing

outside of his own mind is causing the experience. However, it happens that an oasis is in fact in front of him, and, coincidentally, the hallucination matches the oasis perfectly. Despite the match between the experience and the environment, however, the absence of a causal relation moves us to deny that the case is one of perception (Pears 1976).

It seems clear, nevertheless, that an object's being a cause of a perceptual state is not sufficient for perception. For example, the state of perceiving an object typically will be the effect of a long causal chain, but typically only one of the links in the chain will be the object perceived. Suppose that a cup is illuminated by a flashlight. The flashlight is a cause of the perceptual state just as the cup is, but it is not the object perceived, and it is hard to see how appeal to causation alone will capture this obvious fact (Pears 1976).

1.3 Color Perception and the Primary–Secondary Quality Distinction

A different question about the relation between the perceiver and the world is the question of how successful the process of perception is at informing the perceiver about her environment. This question has been explored primarily in the debate about color perception, the most actively investigated topic in the area since the mid-1980s. Although philosophical interest in the perception of color is as old as philosophy itself, the contemporary debate has been invigorated by significant developments in color science and the introduction of these developments into philosophy (see Byrne and Hilbert 1997).

The perception of color is related to the success of visual perception because there is some reason to doubt whether objects actually possess colors. If it could be shown that objects do in fact fail to have colors, then visual perception would be seen to be mistaken in its representation of objects as color-bearers. It is generally assumed that the only way in which colors could be properties of external objects would be if they were physical properties—properties countenanced by science. The epistemological question whether color vision is reliable, therefore, is often recast as the metaphysical question whether there is a scientific property that is sufficiently close to our conception of color to deserve the name. If there is no such property, then doubt is cast on the idea that color is indeed an ordinary property of objects, and this in turn raises doubt about the accuracy of vision quite generally.

Two facts suggest that colors are not physical. First, color does not appear in the basic scientific description of the world. Second, colors seem to be variable in a way that ordinary physical properties are not. For example, a yellow object against a red background will appear greenish, but the same object against a green

background will look reddish, a phenomenon known as *simultaneous contrast*. Because of these facts, some philosophers have concluded that color is a *secondary quality*, that is, a property the existence of which is dependent on a perceiving mind. Secondary qualities are contrasted with *primary qualities*, such as shape or mass the existence of which is not dependent on the perceiving mind.

The philosophical investigation of color has led to two opposing camps. *Color realists* accept as an obvious fact that colors are properties of external objects. The task is to identify the physical property that is identical to color and show that the variability of color, such as that due to simultaneous contrast, does not undermine the idea that colors are properties of external objects. *Color antirealists* take the variability of color to be central to its nature and infer that color, as we understand that concept, does not exist. They hold that while some of the physical properties of the environment cause viewers to have color experience, there is nothing in the environment that is correctly represented by those experiences.

A special sort of realism attempts to accommodate both camps. The *dispositional theory of color* (usually associated with John Locke (1632–1704)) acknowledges both the variability of color and its external reality. An object is navy blue, on this view, if it has the disposition—a real property of the object itself—to cause normal perceivers to have experiences of navy blue in standard conditions. Colors are identical to these dispositions and, therefore, are properties of external objects. Color is variable because objects have at least as many color dispositions, and therefore colors, as there are color experiences.

2. The Perceiving Mind

2.1 Qualia

The debates concerning the nature of the mental states involved in perception recapitulate some of those dealing with the mind-world relation. Closely related to the color debate is the question about the existence of *qualia* (singular: *quale*, pronounced kwä'lē). The term is sometimes used to refer to the properties of the external environment, such as object color, that are experienced in perception. It is also used in more philosophically radical ways to refer to the properties of experiences themselves that give experiences their 'felt' qualities. This radical understanding of qualia thus takes them to be mental entities. Think, for example, about orgasms. There is a distinct feel to the experience of an orgasm that is absent in the mere belief that one is having an orgasm. After all, others can also believe that you are experiencing an orgasm without experiencing its characteristic feel. There is, as it is sometimes put, something it is like having certain kinds of experience. According to some, this is because

these kinds of experience possess qualia and beliefs do not.

Because qualia in the radical sense are mental items, the qualia debate has focused in part on whether the existence of qualia shows that *physicalism*—the view that the only entities, properties, and relations that exist are those countenanced by the physical sciences—is inadequate. (Philosophers who believe that it does have come to be called qualia freaks.) A central argument here is the knowledge argument (Jackson 1982): Mary is born and raised in an entirely black and white room. She becomes a scientist and, we suppose, learns everything a complete science can teach about perception and the world perceived. She then emerges from her room and, seeing a tomato for the first time, learns something new, namely, what it is like to see something red; that is, she experiences a red quale. But if Mary knows everything that a complete science can teach about the world and the perception of it, then she has learnt something that lies outside of the realm of the scientific, and physicalism is false.

The existence of qualia has also been taken to show that the dominant theory of mental states, called *functionalism*, is false. Functionalism holds that mental states are to be analyzed solely in terms of their relations to other mental states, relevant sensory input, and the behavioral output to which they lead. How these states are embodied in an animal, or indeed in a computer, is taken to be irrelevant to the nature of the states. Being in pain, for example, is to be analyzed as the state that is typically caused by damage to one's body; that causes one to withdraw from the source of the damage; that is related to feelings of anger or regret; and so on. But now consider which is made possible by the existence of qualia *spectrum inversion*. Inverted perceivers have systematically different color experience from normal perceivers because the qualia associated with their perceptual experience is inverted: what normal perceivers call 'red' they see as the color normal perceivers call 'green'; what normal perceivers call 'yellow' they see as the color normal perceivers call 'blue'; and so on. Despite their differences from normal perceivers, however, spectrally inverted perceivers could never be identified. After all, inverted perceivers would learn the same language as normal perceivers: they would call grass 'green' and tomatoes 'red' even if their experience of these objects was different from that of normal perceivers. If there were inverted perceivers, therefore, nothing in their behavior would reveal that they were inverted. In particular, nothing about the functional relations among their mental states revealed in their behavior would distinguish inverted from normal perceivers. By hypothesis, however, their experience is different. This argument—the *argument from spectrum inversion*—seems to show, therefore, that there has to be more to perceptual states than their functional relations.

It has been argued, however, that the very idea of qualia as properties of experience is based on a

confusion between the properties experience represents the environment as having and the properties experience itself possesses (Harman 1990). When one introspects to examine these putative properties of experience, one finds only the properties of the objects one is experiencing and not the properties of experiences. However, the former are not mysterious in the least; they are just ordinary properties of the physical world.

2.2 Perceptual Content

The qualia debate has also provided a way of investigating the nature of the *representational content* of perceptual states (see Villanueva 1996). Philosophers have asked whether there is something more to perceptual experience than what is represented by that experience—its representational content (or, simply, ‘content’). The view that denies that there is anything more is sometimes called the *intentional theory of perception* (Harman 1990) (where ‘intentional’ refers to what a mental state is about). On the intentional view, once the objects and properties in the scene perceived are specified, the content of the perceptual state has been exhausted. The qualia freak denies the intentional theory. He believes that there is more to perceptual states than their content, namely, the felt qualities of perceptual experiences. Consider orgasms again. Whatever the representational content of that state—perhaps that a particular physiological process is occurring in one’s body—the felt quality of the experience does not seem to be exhausted by that content.

In the realm of visual experience, intuitions are less clear. Suppose I am standing on a road on which there are two trees of the same height, one of which is at a distance of 100 meters and the second of which is at a distance of 200 meters. While I might judge the two trees to be of the same height, it is also apparent to me that the nearer tree takes up more space in my visual field. Since experience cannot correctly represent the trees both as equal in size and as different in size, it has been suggested that there is an entity, the visual field itself, that is the bearer of the latter aspect of the experience: the trees are represented as being the same size but appear in the visual field to have different sizes. There are thus two orders of properties: those of the visual field and their sister properties in the external world. On this view, the properties of the visual field are not part of the representational content of the perceptual state (Peacocke 1983).

A related question about perceptual content is whether it is conceptual or nonconceptual. The mainstream position is that all perceptual content is conceptual; the perceiver must possess the concepts for the entities and properties perceptually represented by her. For example, the ability to see a dog as a dog requires that one possess the concept ‘dog.’ This

mainstream position evolved in part as a rejection of sense-datum theory and its commitment to a nonconceptual given in perception. Nevertheless, some of the considerations that led to sense-datum theory have emerged again as relevant, and it has been argued that some perceptual representation does not require the possession of concepts. One such consideration is the epistemological view that conceptual thought has to be grounded in something that is nonconceptual (see Sedivy 1996). A second consideration is that perceptual states seem to represent the environment in a much more fine-grained fashion than a perceiver has the concepts for. For example, it is sometimes thought that the colors we see are more varied and subtly different from each other than we have concepts for. A third consideration is the desire to link the conceptual nature of thought with the contents of the computational states posited by the computational theory of vision. The computational theory of vision describes some of the content of a visual state in terms of various mathematical concepts, but perceivers do not have to possess those concepts in order to see.

One way of fleshing out nonconceptual content is in informational terms. A second way of doing so is by means of *scenarios*. A scenario is a representation of the way the physical space around the perceiver is filled out. It is specified by identifying within a perceiver-relative frame of reference the locations of object surfaces and their visible properties. The scenario itself, rather than any mental representation, is part of the content of the perceptual state, and it captures how the world looks to the perceiver independently of the perceiver’s conceptual resources (Peacocke 1992).

A third question about perceptual content has focussed on the computational theory of vision, in particular as it has been presented by David Marr (1945–80). Theorists have asked whether the contents of computational states have to make essential reference to the external world or whether they can be fully specified in terms of the perceiver’s states alone. The former view is one (but not the only) version of *externalism*; the latter is a version of *internalism* or *individualism*. (In philosophical jargon, externalism is committed to *wide content* whereas individualism is committed only to *narrow content*.) Externalism and internalism are positions that have been important in recent philosophy of mind, and the debate about computational vision has been pursued largely as an instance of the wider debate about what cognitive theory says about mental content quite generally (Burge 1996).

2.3 Spatial Perception

The scenarios mentioned above in connection with perceptual content are essentially spatial and reflect a concern in the philosophy of perception about space.

The importance of this topic has been recognized since Immanuel Kant's (1724–1804) *Critique of Pure Reason* tied spatial perception to the concept of an objective world. One reason for thinking that spatial perception is necessary for objectivity is that the perception of space seems essential to the formation of a concept of something that persists over time. The notion of persistence, in turn, is necessary to form a concept of an external object and, thereby, an objective world. Imagine a space-less world constituted entirely of sound. In such a world it seems impossible to articulate the distinction between, for example, two different C# sounds occurring at different times and the very same occurrence of C# being heard, disappearing, and then reappearing. It is argued, therefore, that without the perception of space, it is impossible to distinguish between the occurrence of multiple fleeting objects of the same type and a single persistent object (Strawson 1959). And without a notion of persistence, it is impossible to develop a concept of an objective world. A closely related idea is that the perception of space—or possibly the ability to attend selectively to different parts of space—is required in order to entertain thoughts about particular physical objects.

A second issue regarding space perception is that of the concept of a surface and its role in ordinary language and theories of perception. Under typical conditions, I only see part of the surface of an object. Does this mean that I see only part of an object? Or do I see a whole object but only part of it directly? And does that imply that I see the whole object indirectly? Investigation of the nature of surfaces thus leads to a distinction between direct and indirect perception that is rather different from the traditional one and does not depend on a distinction between the physical and the mental (Stroll 1988). The philosophical investigation of surfaces is important also because the computational theory of vision is in large part about how visual perception computes the surface properties of objects.

A third issue has come to be known as *Molyneux's question* or *Molyneux's problem* (see Degenaar 1996). It was formulated by the Irish scientist William Molyneux (1656–98) and discussed by John Locke with whom Molyneux corresponded. In a letter to Locke on July 7, 1688, Molyneux formulates the question: 'A Man, being born blind, and having a Globe and a Cube, nigh of the same bigness, committed into his Hands, and being Told, which is Called the Globe, and which the Cube, so as easily to distinguish them by his Touch or Feeling; Then both being taken from Him, and Laid on a Table, let us suppose his Sight Restored to Him; Whether he Could, by his sight, and before he touch them, know which is the Globe and which the Cube?' This question was addressed by almost every major theorist of perception following Molyneux and has been of sufficient historical interest to have led Ernst Cassirer to claim that it was the focus of eighteenth century epistemology

and psychology as well as the stimulus for all of Berkeley's philosophy. Molyneux's question continues to receive some attention from philosophers as a way of exploring a number of issues including the relation between different perceptual modalities, especially sight and touch; the existence and nature of the concept of space; and whether or not certain perceptual abilities are learned or innate.

Finally, the question of whether some quasiperceptual representations have significant spatial properties has been examined in the debate about mental imagery, the quasivisual experiences of imagination (see Block 1981). Philosophers and psychologists have debated the claim that mental images are essentially spatial as against the view that they are no more than the language-like representations thought by some to constitute the mental representations underlying all of cognition. Because there is now evidence that mental images are produced by the same neural machinery that produces visual experience, further investigation into imagery may shed light on the nature of visual representations, and research into vision may produce a better understanding of imagery.

See also: Intentionality and Rationality: A Continental-European Perspective; Intentionality and Rationality: An Analytic Perspective; Knowledge (Explicit and Implicit): Philosophical Aspects; Memory Psychophysics; Perception and Action; Perception: History of the Concept; Perception without Awareness, Psychology of; Perceptual Learning; Perceptual Organization; Reference and Representation: Philosophical Aspects; Vision, High-level Theory of; Vision, Low-level Theory of

Bibliography

- Armstrong D M 1961 *Perception and the Physical World*. Routledge & Kegan Paul, London
- Block N (ed.) 1981 *Imagery*. MIT Press, Cambridge, MA
- Block N J, Fodor J A 1972 What mental states are not. *Philosophical Review* **81**: 159–81
- Burge T 1986 Individualism and psychology. *Philosophical Review* **95**: 3–45
- Byrne A, Hilbert D R (eds.) 1997 *Readings on Color*. MIT Press, Cambridge, MA, Vol. 1
- Crane T 1992 Introduction. In: Crane T (ed.) *The Contents of Experiences: Essays on Perception*. Cambridge University Press, Cambridge, UK
- Dancy J 1985 *An Introduction to Contemporary Epistemology*. Blackwell, Oxford, UK
- Degenaar M 1996 *Molyneux's Problem* [trans. Collins M J]. Kluwer, Dordrecht, Holland
- Dretske F I 1981 *Knowledge and the Flow of Information*. MIT Press, Cambridge, MA
- Grice H P 1961 The causal theory of perception. *Proceedings of the Aristotelian Society* **35**: 121–52

- Harman G 1990 The intrinsic quality of experience. In: Tomberlin J E (ed.) *Philosophical Perspectives 4*. Ridgeview, Atascadero, USA
- Jackson F 1977 *Perception*. Cambridge University Press, Cambridge, UK
- Jackson F 1982 Epiphenomenal qualia. *Philosophical Quarterly* **32**: 127–36
- Marr D 1982 *Vision*. Freeman, New York
- Martin M G F 1998 Perception. In: Craig E (ed.) *Routledge Encyclopedia of Philosophy*. Routledge, London
- Peacocke C 1983 *Sense and Content*. Clarendon Press, Oxford, UK
- Peacocke C 1992 *A Study of Concepts*. MIT Press, Cambridge, MA
- Pears D 1976 The causal conditions of perception. *Synthese* **33**: 25–40
- Sedivy S 1996 Must conceptually informed perceptual experience involve non-conceptual content? *Canadian Journal of Philosophy* **26**: 413–31
- Snowdon P 1990 The objects of perceptual experience. *Proceedings of the Aristotelian Society* **64**: 121–150
- Strawson P F 1959 *Individuals*. Methuen, London
- Stroll A 1988 *Surfaces*. University of Minnesota, Minneapolis, MN
- Villanueva E (ed.) 1996 *Perception*. Ridgeview, Atascadero, Holland
- Vision G 1997 *Problems of Vision*. Oxford University Press, Oxford, UK
- Weiskrantz L 1986 *Blindsight: A Case Study and Implications*. Clarendon Press, Oxford, UK

I. Gold

Perception without Awareness, Psychology of

Perception without awareness (or, synonymously, unconscious perception) denotes the fact that a stimulus exerts causal influence on mental processes or behavior without being phenomenally experienced.

Research into unconscious perception began early in the history of experimental psychology (Peirce and Jastrow 1884, Sidis 1898). It is related to other domains working on unconscious processes, such as implicit memory (see *Implicit Memory, Cognitive Psychology of*) or implicit learning (Reber 1993). From the very beginning, experimenters attempted to prove the mere existence of unconscious perception. Researchers had to set up experimental demonstrations of a stimulus's causal influence on mental processes without its concomitant phenomenal experience. More recently, theorists began to ask how perception without awareness works.

The proof of perception without awareness usually needs two critical tests: (a) an indirect test that assesses the impact of a stimulus on mental processes or behavior, and (b) a direct test that establishes the absence of phenomenal experience of the stimulus in question. There was rarely doubt about the dem-

onstration of causal influence of a stimulus, but there was—and still is—much skepticism about the proof of absence of phenomenal experience. Therefore, experimenters aimed at enhancing the validity of thresholds in order to ensure that perception was indeed subliminal. Alternatively, some researchers were looking for procedures that circumvented the problem of thresholds. Finally, there is now some evidence for unconscious perception from research in neuroscience.

1. Early Studies

Early studies attempted to show that people were able to guess at above chance level without subjectively perceiving the stimulus (see Adams 1957, for a review). Much of the classical research has been criticized for its lack of proper control conditions and of proper measurements of consciousness (e.g., Eriksen 1960, Holender 1986). Holender, for example, criticized experiments using dichotic listening tasks for their lack of evidence that processing of the stimuli was indeed unconscious. In dichotic listening tasks, participants hear different information in both ears. They are instructed to attend to the information in one ear. It has been shown that information in the unattended ear may exert some influence on ongoing processing without subjective awareness of this information. Of course, if participants turned their attention to the unattended ear, they would be able to identify the content. Holender concluded that one can not be sure that unattended stimuli are in fact unconscious. Therefore, he proposed that manipulations should not limit mental resources by distracting attention, but should limit the available data by rendering perception of stimuli impossible.

2. Perception Below Threshold

At the start of the twenty-first century, several masking techniques are used to render the material—most of them visual stimuli—undetectable or unidentifiable. For example, to measure individual thresholds of detection, masked stimuli are presented on half of the trials. On the other half of the trials, only the mask is shown. Participants have to decide whether a stimulus is present or absent. If accuracy does not exceed chance level (mostly at 50 percent), absence of direct phenomenal experience is assumed.

In a well-known study, Marcel (1983) presented masked words below the subjective threshold of detection. Then he presented two words, one similar and one dissimilar to the masked word, in respect to either graphemic information or semantic content. Participants had to decide which word was similar to the masked word, either graphemically or semantically. They were able to make above-chance

guesses for each kind of judgment. As presentation time was decreased, judgments about graphemic similarity dropped to chance level first, followed by judgments about semantic similarity.

The logic underlying this kind of study is to show a significant effect in an indirect perceptual task (e.g., semantic discrimination) when there is a null effect in a direct task (e.g., detection). However, null sensitivity in direct tasks has always been a matter of controversy. If the criteria for null sensitivity in the direct test are too lenient, one may erroneously claim to have demonstrated unconscious perception. Conversely, if these criteria are too strong, it may be virtually impossible to demonstrate perception without awareness.

Therefore, some scholars have tried to circumvent the issue of thresholds and their measurements and taken different routes to demonstrate unconscious perception.

3. Alternative Approaches to Demonstrate Unconscious Perception

Reingold and Merikle (1988) designed an experimental setup that allows measuring unconscious perception without assessment of thresholds. Specifically, they proposed to use identical materials and experimental conditions for direct and indirect tests. The only difference between the two tests lies in the instruction: Conscious perception of stimuli is assessed in a direct test. In contrast, a task without reference to phenomenal experience of the stimulus is given in an indirect test. Normally, one would expect that performance in the direct test would show greater sensitivity than performance in the indirect task because conscious perception optimizes extraction and use of stimulus information. If the indirect measure shows greater sensitivity than the direct measure, however, one may conclude that unconscious processes mediate task performance. This was the case in a study on implicit memory effects by Kunst-Wilson and Zajonc (1980). Participants were exposed to random shapes for 1 ms during a learning session. At test, participants were shown 10 pairs of octagons, one old and one new. They had to decide either which of the two octagons has been shown before (direct test) or which of the octagons they liked better (indirect test). In the direct test, participants scored 48 percent, which was at chance level, but they preferred the old octagon in 60 percent of the cases, indicating greater sensitivity of the indirect than of the direct test. In studies on unconscious perception, however, it proved difficult to find the proposed indirect-greater-than-direct effect pattern.

More recently, Greenwald and his colleagues proposed to use performance in a direct test as independent variable and performance in an indirect test as dependent variable in a linear regression analysis

(Greenwald et al. 1995). These authors used a semantic priming paradigm as indirect test to show semantic activation without conscious perceptibility of the prime. It seems plausible that better performance in the direct test results in higher accuracy in the indirect test, yielding a linear regression function with positive slope. $X = 0$ denotes the point where conscious perception is completely absent. If there existed no effect of unconscious perception, the regression line would cross the zero point: $Y = 0$ if $X = 0$. If, however, there were an effect of unconscious perception even under complete absence of conscious identification, the intercept would be positive: $Y > 0$ if $X = 0$. In several experiments employing this kind of analysis, Greenwald and his colleagues (Draine and Greenwald 1998, Greenwald et al. 1995, 1996) were able to show reliable indirect effects under absence of direct effects.

4. Evidence from the Neurosciences

Other evidence of perception without awareness comes from patients with damage to the visual cortex. They typically do not see anything in the hemifield of vision that is opposite to the side of the damage (e.g., Pöppel et al. 1973, Weiskrantz et al. 1974). However, some of these patients are able to make visual discriminations in the blind hemifield although they reportedly do not have any phenomenal experience of the stimuli. This phenomenon is called 'blindsight.' Some authors suggested that blindsight may be caused by small islands of intact tissue in the primary visual cortex (Fendrich et al. 1992). Although correct visual discriminations may be caused by small, intact islands of accurate vision, this hypothesis cannot account for all phenomena observed in both monkeys and human patients (see Weiskrantz 1995).

Brain imaging techniques led to new indirect measures. Using an unconscious priming paradigm, Dehaene et al. (1998) have shown that masked primes influenced subsequent processing, accompanied by measurable modifications of electrical brain activity and cerebral blood flow.

We have discussed studies aimed at demonstrating that unconscious perception exists. We now turn to the question of how unconscious perception works.

5. Differences Between Conscious and Unconscious Perceptual Processes

One may think that unconscious perception is the same as conscious perception, except for the lack of phenomenal experience. In this view, unconscious perception is just a faint shadow of conscious perception. This view has been questioned, and there is good evidence that unconscious perception is driven by a limited-capacity system, unconscious perception

by a system that has much more capacity and allows the use of information from multiple sources (see MacLeod 1998). The main evidence comes from demonstrations of qualitative dissociations between conscious and unconscious perception. This means that conscious perception affects a variable in one direction, unconscious perception in the opposite way, suggesting different underlying mechanisms (see Merikle and Daneman 1998). An elegant study showing such a dissociation has been reported by Jacoby and Whitehouse (1989). In a recognition test, a target word was presented and participants had to decide whether or not it had appeared in a list presented before. Shortly before the target, a prime was shown that was either identical to or different from the target. If the primes were shown subliminally, identical primes resulted in more 'old' responses than different primes. If the primes were shown supraliminally, however, identical primes resulted in less 'old' responses than different primes. This qualitative dissociation suggests that participants who saw supraliminal primes were able to correct their judgments for the supposed influence of the primes. If exposed to subliminal primes, however, participants were unable to avoid the impact of the prime and were biased toward 'old' responses when prime and target were identical. Several other studies yielded similar dissociations (e.g., Debner and Jacoby 1994, Marcel 1980, Merikle and Joordens 1997), suggesting that conscious and unconscious perception are governed by different mechanisms.

6. Outlook

There is fair evidence that unconscious perception exists. However, there is no evidence that effects of unconscious perception, caused by exposure to subliminal messages, are of any practical relevance (Greenwald et al. 1991, Vokey and Read 1985).

Beyond the proof of existence, interesting questions about unconscious perception remain open. One question pertains to the durability of unconscious perception (see Merikle and Daneman 1998). It has been shown that implicit memory effects may last for several days (see *Implicit Memory, Cognitive Psychology of*). In contrast, most studies on unconscious perception dealt with effects that lasted a few seconds. In implicit memory tasks, encoding is conscious, but retrieval is assumed to be unconscious. In perception without awareness, encoding is unconscious. Do long-term effects exist only if encoding is conscious, or is it possible to find long-term effects of unconscious encoding? Poetzl ([1917] 1960) has shown that briefly presented materials may be manifest in dreams during the following night although people were unable to remember the materials after the initial presentation. The Poetzl phenomenon, together with more recent findings on memory for information encoded during anesthesia (see Merikle and Daneman 1996 for a meta-

analysis) suggest that unconscious influences may last for hours.

As discussed above, another major question is how perception without awareness works. This research has only recently begun and will intensify, stimulated by recently developed methods in the neurosciences and cognitive psychology (e.g., the process dissociation procedure, see Debner and Jacoby 1994) that open up opportunities for research into neural correlates and cognitive mechanisms of unconscious perception.

Brain imaging methods may bring new opportunities to look into neural pathways related to unconscious perception. Specifically, these methods may bring new insights into qualitative dissociations between neural substrates connected to conscious perception and neural substrates connected to unconscious perception, as demonstrated in a study by Morris et al. (1998). These authors paired an angry face with (uncomfortable) white noise. Subliminal presentation of the conditioned angry face resulted in increased neural activity in the right, but not left amygdala. In contrast, supraliminal presentation of the same face resulted in increased neural activity in the left, but not right amygdala. Moreover, it has been shown that simple stimulus features may proceed to the amygdala and trigger an emotion without involvement of cortical structures (see LeDoux 1995). Although any conclusion about presence or absence of awareness in perceiving such features is premature, it enables theorists to conceive ways of how information may influence mental processes without reaching consciousness.

See also: Consciousness, Cognitive Psychology of; Consciousness, Neural Basis of

Bibliography

- Adams J K 1957 Laboratory studies of behavior without awareness. *Psychological Bulletin* **54**: 383–405
- Debner J A, Jacoby L L 1994 Unconscious perception: Attention, awareness, and control. *Journal of Experimental Psychology: Learning, Memory, and Cognition* **20**: 304–17
- Dehaene S, Naccache L, Le Clec' H G, Koechlin E, Mueller M, Dehaene-Lambertz G, van de Moortele P-F, Le Bihan D 1998 Imaging unconscious semantic priming. *Nature* **395**: 597–600
- Draine S C, Greenwald A G 1998 Replicable unconscious semantic priming. *Journal of Experimental Psychology: General* **127**: 286–303
- Eriksen C W 1960 Discrimination and learning without awareness: A methodological review. *Psychological Review* **67**: 279–300
- Fendrich R, Wessinger C M, Gazzaniga M S 1992 Residual vision in a scotoma: Implications for blindsight. *Science* **258**: 1489–91
- Greenwald A G, Draine S C, Abrams R H 1996 Three cognitive markers of unconscious semantic priming. *Science* **273**: 1699–702
- Greenwald A G, Klinger M R, Schuh E S 1995 Activation by marginally perceptible ('subliminal') stimuli: Dissociation of

- unconscious from conscious cognition. *Journal of Experimental Psychology: General* **124**: 22–42
- Greenwald A G, Spangenberg E R, Pratkanis A R, Eskanazi J 1991 Double-blind tests of subliminal self-help audiotapes. *Psychological Science* **2**: 119–22
- Holender D 1986 Semantic activation without conscious identification in dichotic listening, parafoveal vision, and visual masking: A survey and appraisal. *Behavioral and Brain Sciences* **9**: 1–66
- Jacoby L L, Whitehouse K 1989 An illusion of memory: False recognition influenced by unconscious perception. *Journal of Experimental Psychology: General* **118**: 126–35
- Kunst-Wilson W R, Zajonc R B 1980 Affective discrimination of stimuli that cannot be recognized. *Science* **207**: 557–58
- LeDoux J E 1995 Emotion: Clues from the brain. *Annual Review of Psychology* **46**: 209–35
- MacLeod C 1998 Implicit perception: Perceptual processing without awareness. In: Kirsner K, Spelman C, Maybery M, O'Brien-Malone A, Anderson M, MacLeod C (eds.) *Implicit and Explicit Mental Processes*. Erlbaum, Mahwah, NJ, pp. 57–78
- Marcel A J 1980 Conscious and preconscious recognition of polysemous words: Locating the selective effects of prior verbal context. In: Nickerson R S (ed.) *Attention and Performance*. Erlbaum, Hillsdale, NJ, Vol. VII, pp. 435–57
- Marcel A J 1983 Conscious and unconscious perception: Experiments on visual masking and word recognition. *Cognitive Psychology* **15**: 197–237
- Merkle P M, Daneman M 1996 Memory for unconsciously perceived events: Evidence from anesthetized patients. *Consciousness and Cognition* **5**: 525–41
- Merkle P M, Daneman M 1998 Psychological investigations of unconscious perception. *Journal of Consciousness Studies* **5**: 5–18
- Merkle P M, Joordens S 1997 Measuring unconscious influences. In: Cohen J D, Schooler J W (eds.) *Scientific Approaches to Consciousness*. Erlbaum, Mahwah, NJ, pp. 109–123
- Morris J S, Öhman A, Dolan R J 1998 Conscious and unconscious emotional learning in the human amygdala. *Nature* **393**: 467–70
- Peirce C S, Jastrow J 1884 On small differences of sensation. *Memoirs of the National Academy of Sciences* **3**: 73–83
- Poetzl O [1917] 1960 The relationships between experimentally induced dream images and indirect vision. *Monograph 7. Psychological Issues* **2**: 46–106
- Pöppel E, Held R, Frost D 1973 Residual visual function after brain wounds involving the central visual pathways in man. *Nature* **243**: 295–6
- Reber A S 1993 *Implicit Learning and Tacit Knowledge: An Essay on the Cognitive Unconscious*. Oxford University Press, New York
- Reingold E M, Merikle P M 1988 Using direct and indirect measures to study perception without awareness. *Perception and Psychophysics* **44**: 563–75
- Sidis B 1898 *The Psychology of Suggestion*. Appleton, New York
- Vokey J R, Read J D 1985 Subliminal messages. Between the devil and the media. *American Psychologist* **40**: 1231–9
- Weiskrantz L 1995 Blindsight—not an island unto itself. *Current Directions in Psychological Science* **4**: 146–51
- Weiskrantz L, Warrington E K, Sanders M D, Marshall J 1974 Visual capacity in the hemianoptic field following a restricted occipital ablation. *Brain* **97**: 709–28

R. Reber and W. Perrig

Perceptrons

Perceptrons was the generic name given by the psychologist Frank Rosenblatt to a family of theoretical and experimental artificial neural net models which he proposed in the period 1957–1962. Rosenblatt's work created much excitement, controversy, and interest in neural net models for pattern classification in that period and led to important models abstracted from his work in later years. Currently the names (single-layer) Perceptron and Multilayer Perceptron are used to refer to specific artificial neural network structures based on Rosenblatt's perceptrons. This article references the intellectual context preceding Rosenblatt's work and summarizes the basic operations of a simple version of Rosenblatt's perceptrons. It also comments briefly on the developments in this topic since Rosenblatt.

Rosenblatt's first (partial) report in January 1957 was titled *The Perceptron: A Perceiving and Recognizing Automaton*. A subsequent report, in January 1958, titled *The Perceptron: A Theory of Statistical Separability in Cognitive Systems* was adapted for publication (Rosenblatt 1958). A compendium of research on perceptrons by Rosenblatt and his group is *Principles of Neurodynamics* (Rosenblatt 1962).

Rosenblatt's research was done in the context of: the neobehaviorism of Hull (1952), the general principles of neuromodeling of behavior proposed by Hebb (1949), the Threshold Logic Unit (TLU) neuron model of McCulloch and Pitts (1943), proposals for brain modeling based on neurophysiology (Eccles 1953) and symbolic logic, switching theory and digital computers (Shannon and McCarthy 1956), the stimJEUlus sampling models of Estes and Burke (1953), the linear operator probabilistic learning models of Bush and Mosteller (1955), and statistical classification procedures (e.g., Rao 1955).

The basic building block of a perceptron is an element that accepts a number of inputs x_i , $i = 1 \dots N$, and computes a weighted sum of these inputs where, for each input, its fixed weights β can be only $+1$ or -1 . The sum is then compared with a threshold θ , and an output y is produced that is either 0 or 1, depending on whether or not the sum exceeds the threshold. Thus

$$y = \begin{cases} 1 & \text{if } \left(\sum_{i=1}^N \beta_i x_i \right) \geq \theta \\ 0 & \text{if } \left(\sum_{i=1}^N \beta_i x_i \right) < \theta \end{cases}$$

A perceptron is a signal transmission network consisting of sensory units (S units), association units (A units), and output or response units (R units). The 'retina' of the perceptron is an array of sensory elements (photocells). An S-unit produces a binary

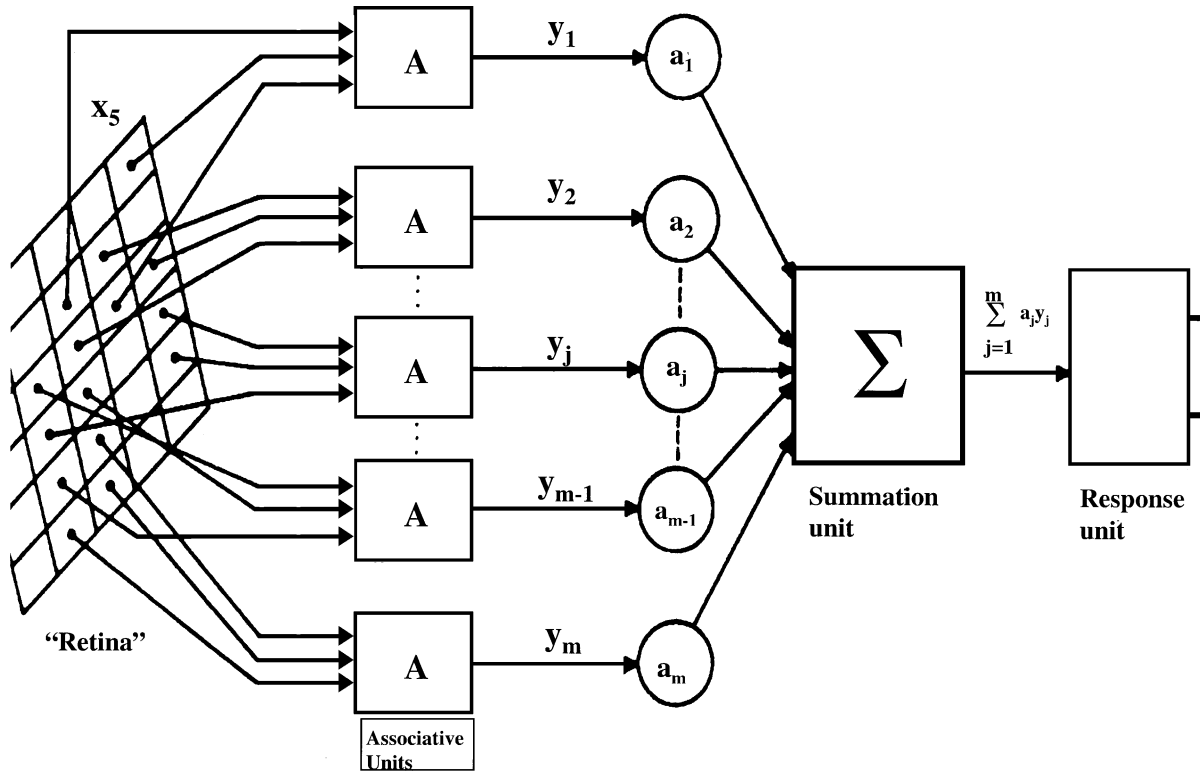


Figure 1
The mark 1 perceptron

output depending on whether or not it is excited. A randomly selected set of retinal cells is connected to the next level of the network, the A units. As originally proposed there were extensive connections among the A units, the R units, and feedback between the R units and the A units. A simpler version omitting the lateral and feed-back connections is shown in Fig. 1 This simplification is the perceptron considered here.

Each A unit behaves like the basic building block discussed above, where the $+1$, -1 weights for the inputs to each A unit are assigned randomly. The threshold θ for all A units is the same. The binary output of the k^{th} A unit ($k = 1, \dots, m$) is multiplied by a weight a_k , and a sum of all m weighted outputs is formed in a summation unit that is the same as the basic building blocks with all weights equal to $+1$. Each weight a_k is allowed to be positive, zero, or negative, and may change independently of other weights. The output of this block is again binary, depending on a threshold, t , that is normally set at 0. The binary values of the output are used to distinguish two classes of patterns that may be presented to the retina of a perceptron. The design of this perceptron to distinguish between two given sets of patterns involves adjusting the weights a_k , $k = 1, \dots, m$, and setting the threshold θ .

Rosenblatt (1962) proposed a number of variations of the following procedure for 'training' perceptrons. The set of 'training' patterns, that is, patterns of known classification, are presented sequentially to the retina, with the complete set being repeated as often as needed. The output of the perceptron is monitored to determine whether a pattern is correctly classified. If not, the weights are adjusted according to the following 'error correction' procedure: If the n -th pattern was misclassified, the new value $a_k(n+1)$ for the k -th weight is set to

$$a_k(n+1) = a_k(n) + y_k(n) * \delta(n)$$

where $\delta(n)$ is 1 if the n -th pattern is from class 1 and $\delta(n)$ is -1 if the n -th pattern is from class 2. No adjustment to the weight is made if a pattern is classified correctly.

Rosenblatt conjectured that, when the pattern classes are 'linearly separable' the error correction 'learning' procedure will converge to a set of weights defining the separating hyperplane that correctly classifies all the patterns. The shortest proof of this perceptron convergence theorem was given by A. J. Novikoff. Subsequent contributions related the simple perceptron to statistical linear discriminant functions

and related the error-correction learning algorithm to gradient-descent procedures and to stochastic approximation methods that originally were developed for finding the zeros and extremes of unknown regression functions (e.g., Kanak 1962).

The simple perceptron described is a series-coupled perceptron with feed-forward connections only from S units to A units and A units to the single R unit. The weights a_k , the only adaptive elements in this network, are evaluated directly in terms of the output error.

Minsky and Papert (1969) further reduced the simple perceptron to a structure with sampled connections from the 'retina' directly to the adjustable weights. This reduced structure is referred to as a *single-layer perceptron*. There is no layer of 'hidden' elements, that is, additional adjustable elements for which the adjustment is only indirectly related to the output error. A feed-forward perceptron with one or more layers of hidden elements is termed a *multilayer perceptron*. Rosenblatt investigated *cross-coupled perceptrons* in which connections join units of the same type, and also investigated *multilayer back-coupled perceptrons*, which have feedback paths from units located near the output. For series-coupled perceptrons with multiple R units, Rosenblatt proposed a 'back-propagating error correction' procedure that used error from the R units to propagate correction back to the sensory end. But neither he nor others at that time were able to demonstrate a convergent procedure for training multilayer perceptrons.

Minsky and Papert (1969) proved various theorems about single-layer perceptrons, some of which indicated their limited pattern-classification and function approximating capabilities. For example, they proved that the single-layer perceptron could not implement the Exclusive OR logical function and several other such predicates. Later, many writing on Artificial Neural Networks (ANN) blamed this book for greatly dampening interest and leading to a demise of funding in the USA for research on ANN's. The section on 'Alternate Realities' in Kanak (1992) details why the blame is misplaced. As noted there, by 1962 (see, e.g., Kanak 1962) many researchers had moved on from perceptron type learning machines to statistical and syntactic procedures for pattern recognition.

Minsky and Papert's results did not apply to multilayer perceptrons. Research on ANN's, biologically motivated automata, and adaptive systems continued in the 1970s in Europe, Japan, the Soviet Union, and the USA, but without the frenzied excitement of previous years, which also came back starting in the early 1980s. In a 1974 Harvard University dissertation Paul Werbos presented a general convergent procedure for adaptively adjusting the weights of a differentiable nonlinear system so as to learn a functional relationship between the inputs and outputs of the system. The procedure calculates the derivatives of some function of the outputs, with respect to all inputs and weights or parameters of the

system, working backwards from outputs to inputs. However, this work, published later in a book by Werbos (1994), went essentially unnoticed, until a few years after Rumelhart et al. (1986), independently popularized a special case of the general method. This algorithm, known as error backpropagation or just backpropagation, adaptively adjusts the weights to perform correct pattern classification using gradient descent and training samples. It propagates derivatives from the output layer through each intermediate layer of the multilayer perceptron network. The resurgence of work on multilayer perceptrons and their applications in the decades of the 1980s and 1990s is directly attributable to this convergent backpropagation algorithm.

It has been shown that multilayer feedforward networks with a sufficient number of intermediate or 'hidden' units between the input and output units have a 'universal approximation' property: they can approximate 'virtually any function of interest to any desired degree of accuracy' (Hornik et al. 1989). Several modifications of the basic perceptron learning procedure that make perceptron learning well behaved with inseparable training data, even when the training data are noisy and not error free, have been proposed in recent years and various ANN learning procedures have been shown to related to known statistical techniques. The reference list includes some recent books with tutorial material covering Perceptrons and related ANN's.

See also: Artificial Neural Networks: Neurocomputation; Connectionist Models of Concept Learning; Connectionist Models of Development; Connectionist Models of Language Processing; Neural Networks and Related Statistical Latent Variable Models; Neural Networks: Biological Models and Applications; Perception: History of the Concept

Bibliography

- Anderson J A 1995 *An Introduction to Neural Networks*. MIT Press, Cambridge, MA
- Bush R R, Mosteller F 1955 *Stochastic Models for Learning*. Wiley, New York
- Chauvin Y, Rumelhart D E (eds.) 1995 *Backpropagation: Theory, Architectures, and Applications*. Erlbaum, Mahwah, NJ
- Eccles J C 1953 *The Neurophysiological Basis of Mind*. Clarendon Press, Oxford, UK
- Estes W K, Burke C J 1953 A theory of stimulus variability in learning. *Psychological Review* 6: 276-86
- Haykin S 1999 *Neural Networks. A Comprehensive Foundation*, 2nd edn. Prentice Hall, Upper Saddle River, NJ
- Hebb D O 1949 *The Organization of Behaviour*. Wiley, New York
- Hornik K, Stinchcombe M, White H 1989 Multilayer feedforward networks are universal approximators. *Neural Networks* 2: 359-66

- Hull C L 1952 *A Behaviour System*. Yale University Press, New Haven, CT
- Kanal L 1962 Evaluation of a class of pattern-recognition networks. In: Bernard E E, Kare M R (eds.) *Biological Prototypes and Synthetic Systems*. Plenum Press, New York (reprinted in Sethi I, Jain A 1995 *Neural Networks and Pattern Recognition*. Elsevier, Amsterdam)
- Kanal L N 1992 On pattern, categories, and alternate realities, 1992 KS Fu award talk at IAPR, The Hague. *Pattern Recognition Letters* **14**: 241–55
- McCulloch W S, Pitts W A 1943 A logical calculus of the ideas immanent in nervous activity. *Bulletin of Mathematical Biophysics* **5**: 115–33
- Minsky M, Papert S 1969 *Perceptrons*. MIT Press, Cambridge, MA
- Rao C R 1955 *Advanced Statistical Methods in Biometric Research*. Wiley, New York
- Rosenblatt F 1958 The Perceptron: A probabilistic model for information storage and organization in the brain. *Psychological Review* **65**: 386–408
- Rosenblatt F 1962 *Principles of Neurodynamics*. Spartan, New York
- Rumelhart D E, Hinton G E, Williams R J 1986 Learning internal representations by error propagation. In: Rumelhart D E, McClelland J L et al. (eds.) *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*. MIT Press, Cambridge, MA, Vol. 1, pp. 318–62
- Shannon C E, McCarthy J M (eds.) 1956 *Automata Studies*. Princeton University Press, Princeton, NJ
- Werbos P 1994 *The Roots of Backpropagation: From Ordered Derivatives to Neural Networks and Political Forecasting*. Wiley, New York

L. N. Kanal

Perceptual Constancy: Direct versus Constructivist Theories

Fortunately, our perception of the environment surrounding us is usually quite veridical, accurate, and stable. While laypersons take this ability of ours for granted, students of perception have pointed out that it is actually quite a feat, since the stimulation reaching our senses changes continuously. If we take the visual perception of size as an example, we find that the size of the retinal image of, say, another person changes with that person's distance from us. In spite of these changes we perceive that person as being of constant size and not shrinking or expanding with changes in distance. This 'feat' of perception has aroused a great deal of interest and theories have been proposed to account for it. Prominent among these are two contrasting views of how it is achieved. One view, the 'constructivist,' maintains that the perceptual system is 'intelligent.' In order to attain veridical perception it utilizes high-level mental processes to correct for the changes in stimulation reaching our senses. In contrast, the 'direct' approach claims that there is no need to posit such mental processes to explain veridical

perception, maintaining that the sensory information reaching our senses suffices to explain perceptual veridicality and that what is needed is a much deeper analysis of that sensory information.

1. Stability and Constancy in Perception

The ability of our perceptual system to overcome the effects of changes in stimulation and maintain a veridical and stable percept is called *perceptual constancy* (Epstein 1977, Walsh and Kulikowski 1998). In discussing the constancies it is common to make a distinction between the *distal* stimulus and *proximal* stimulus. The distal stimulus is the actual physical stimulus, the physically objective dimensions of the viewed object. The proximal stimulus, in the case of vision, is the very image that falls on the retina. It changes with changes in position or lighting of the physical stimulus. Constancy can be defined as the ability to correctly perceive the distal stimulus; that is, the stable properties of objects and scenes, in spite of changes in the proximal stimulus.

One example of constancy is *shape constancy*, which refers to the ability of the visual system to ascertain the true shape of an object even when it is not viewed face on but slanted with respect to the observer. In an example of shape constancy the distal stimulus might be a door half open, while the proximal stimulus will be the trapezoid shape it projects on the retinas. Shape constancy in this case would be the perception of the distal rectangular shape of the door in spite of the fact that the proximal shape is far from rectangular. Another example of constancy is *lightness constancy*. Lightness refers to the perceived reflectance of an object; high reflectance is perceived as white and very low reflectance as black, with intermediate reflectances as various shades of gray. We are capable of perceiving the distal lightness of a surface in spite of the fact that the proximal amount of light reaching our eyes changes with changes in the amount of light illuminating that surface. For example, if we hold a piece of chalk in one hand and a piece of charcoal in the other, the chalk will be perceived as white and the charcoal as black. This will be true if we observe the two in a dimly lit room or in bright sunshine, in spite of the fact that the amount of light reaching the eyes from the charcoal in the sunshine might be greater than that from the chalk in the dim room lighting.

The perceptual constancy that has received the greatest amount of research attention is *size constancy* (Ross and Plug 1998). If the distance is not very great we usually perceive the distal size of objects in spite of the changes in their proximal size as their distance from us changes. The two theoretical approaches mentioned above, the constructivist and the direct, deal with size constancy differently. According to the constructivists, size constancy is achieved through some process whereby the perceptual system perceives the object's distance and then takes the distance

into account and 'corrects' the proximal image to yield a true distal percept. The direct approach asserts that there is no need to involve the perception of distance in the perception of size. Instead, it claims that enough information exists in the visual environment to allow direct perception of size without the need for aid from additional 'higher' mental processes.

2. Two Theoretical Approaches to Perception

The two approaches, the constructivist and the direct, differ in many respects, but the main differences between them revolve around two interrelated issues, the richness of the stimulation reaching our sensory apparatus, and the involvement of 'higher' mental processes in the apprehension of our environment. The constructivists see the stimulation reaching our senses as inherently insufficient for veridical perception, necessitating an 'intelligent' mechanism that utilizes inferential processes to supplement the information available to the senses and resolve its inherent ambiguity. The direct theorists, in contrast, argue that the information in the environment is sufficiently rich to afford veridical perception with no need for sophisticated algorithms. The constructivists see perception as multistage with mediational processes intervening between stimulation and percept; i.e., perception is indirect. The direct theorists see perception as a single-stage process; i.e., it is direct and immediate. For the constructivists, memory and schemata based on past experience play an important role in perception. The direct approach sees no role for memory in perception. Finally, the two approaches differ in the aspects of perception they emphasize: the constructivists excel at analyzing the processes and mechanisms underlying perception, whereas the direct approach excels at the analysis of the stimulation reaching the observer.

2.1 The Constructivist Approach

This approach has also been called Helmholtzian, indirect, cognitive, algorithmic, and mediational, among other labels. It is the older of the two, ascribing some of its roots to Helmholtz's (1867) notion of 'unconscious inference,' which maintains that veridical perception is achieved by an inference-like process which transpires unconsciously. A somewhat similar approach was adopted and elaborated more recently by Rock (1983) in his *The Logic of Perception*. This book contains the most detailed account of the constructivist view and the evidence in its favor. The first sentence in the book is, 'The thesis of this book is that perception is intelligent in that it is based on operations similar to those that characterize thought.' (Rock 1983, p.1). Rock makes it clear that this thought-like process occurs unconsciously. He also writes, '... a summary statement of the kind of theory I propose

to advance in the remainder of the book. My view follows Helmholtz's (1867) that perceptual processing is guided by the effort or search to interpret the proximal stimulus, i.e., the stimulus impinging on the sense organ, in terms of what object or event in the world it represents, what others have referred to as the "effort after meaning".' (Rock 1983, p. 16). In other words, Rock is conceiving of perception as an effortful, but unconscious, attempt at identifying an object or event.

In his discussion of size perception, Rock specifically invokes a syllogistic inferential mechanism: 'I will argue that the process of achieving constancy is one of deductive inference where the relevant "premises" are immediately known. That is to say, in the case of a specific constancy such as that of size, two aspects of the proximal stimulus are most relevant, one being the visual angle subtended by the object and the other being information about the object's distance.' (Rock 1983, p. 240). Rock sees size perception as depending on two perceptions, that of proximal size and that of distance, together leading through a syllogism to the veridical distal percept. Like other constructivists, this approach posits some sort of combination of proximal size information and distance information in the achievement of size constancy. In a similar manner, the same combinatorial process holds for all the constancies, according to the constructivist view.

The equation of perceptual and thought processes is difficult to verify empirically, but there are implications from this approach that are verifiable. As can be seen above, the constructivist approach implies that at least two perceptions are involved in perceiving size, that of the proximal size of the object and that of its distance. In a later book, *Indirect Perception*, Rock (1997) collected a large group of studies showing 'that perception is based on prior perception, implying a perception-perception chain of causality.' This interdependence of perceptual processes is something that can be examined empirically and, indeed, the studies reprinted in that book clearly evidence that chain of causality. However, it should also be noted that all of these studies utilize somewhat artificial situations where a manipulation is used in order to create some sort of illusory perception or misperception. A simple example is studies where subjects are presented with monocular views of the stimuli causing misperception of the three-dimensional spatial arrangement. Here, for example, misperceived distance is accompanied by the misperception of other perceptual qualities.

2.2 The Direct Approach

This approach is also referred to as ecological, Gibsonian, sensory, and immediate, among other labels. It is in the main an approach developed by James J. Gibson who expounded it in his *The Ecological Approach to Visual Perception* (Gibson

1979). In that book, Gibson proffered an exciting new approach to the study of visual perception that included many new concepts and new ways of looking at perception. The entire first half of the book is devoted to a novel analysis of the ambient, or surrounding, environment and the information it provides the observer. Gibson finds the classical approach of describing the stimuli for perception in terms of stimulus energies impinging upon the receptors completely unsatisfactory. He points to the differences between these energies and the optical information available in the ambient optic array. That information is *picked up* by a stationary or moving observer. Perception consists of perceiving events, i.e., perceiving changes over time and space in the optic array.

What sort of information is picked up in direct perception? Gibson suggests that there exist invariants in the optic array that serve to supply the observer with unequivocal information. These invariant aspects of the information remain constant in spite of changes in the observer's viewpoint. He musters a great deal of evidence to prove this point. Among the items of evidence he presents is a study of size perception of his where observers were asked to match the height of stakes planted at various distances in a very large plowed field with a set of stakes of varying size nearby. His finding was that size perception remained veridical no matter how far away the stake was planted. Unlike the constructivists, Gibson does not ascribe this size constancy to the taking into account of distance, but rather to the pickup of invariant ratios in the ambient optical array surrounding the observer. In the case of perception of size, he proposes two such invariant ratios. One is the amount of surface texture intercepted by the object's base as it sits on that surface. The second is the horizon ratio, the ratio between the object's height and its distance from the horizon, both measured in the two-dimensional projection of the optic array. There is no need, according to his view, for perceived distance to be involved here, or for the inferential mental processes that the constructivists purport to underlie size perception.

One of Gibson's most important contributions is the concept of *affordances*. He writes, 'The *affordances* of the environment are what it *offers* the animal, what it *provides* or *furnishes*, either for good or for ill' (Gibson 1979, p. 127). In other words, affordances refer to the possible actions that can be taken by the perceiver vis-à-vis objects in the environment. Thus, for example, some surfaces in the environment are 'stand-on-able,' or 'climb-on-able,' or 'sit-on-able.' 'The psychologists assume that objects are *composed* of their qualities. But I now suggest that what we perceive when we look at objects are their affordances, not their qualities.' (Gibson 1979, p. 134), and, '... the basic affordances of the environment are perceivable and usually perceivable directly, without an excessive amount of learning.' (Gibson 1979, p. 143). Gibson is

suggesting that the direct perception of, say, a chair does not constitute recognizing it as a type of furniture labeled 'chair,' i.e., categorizing it, but rather the perception of the chair's affordance, i.e., that it contains a surface that is sit-on-able.

As may be clear from the foregoing description, Gibson's conception is one of an active perceiver exploring his environment. Eye, head, and body movements are part and parcel of the perceptual process. Perception transpires continuously over both time and space. 'Space' here refers not to an empty space but to the many surfaces that make up the environment, the most important being the terrain that at times reaches the horizon. The horizon is of importance as it serves as a useful reference standard, and when it is occluded Gibson speaks in terms of an implicit horizon, presumably similar to what architects and others have called the eye-level plane. With such a conception, Gibson is totally averse to the reductionist experimental paradigms used by the constructivists. Brief exposures or looks through monocular 'peep-holes' do not represent true perception, in his view.

In his book, Gibson almost totally refrains from discussing the processes underlying perception. Perception is simply the pickup of information from invariants in the ambient environment. His only allusions to underlying processes are in terms of resonance: 'In the case of the persisting thing, I suggest, the perceptual system simply extracts the invariants from the flowing array; it *resonates* to the invariant structure or is *attuned* to it. In the case of substantially distinct things, I venture, the perceptual system must *abstract* the invariants. The former process seems to be simpler than the latter, more nearly automatic' (Gibson 1979, p. 249).

3. *Summing Up and a Possible Reconciliation*

The two theoretical approaches, the constructivist and the direct, obviously differ in many respects. The former endows the perceiver with 'higher' mental capacities that allow him/her to compensate for the purported insufficiency of the stimulation reaching the sensory apparatus. In contrast, the direct approach attempts to delineate the richness of information available in the ambient environment as picked up by an active perceiver, allowing direct perception without the need for additional mental processes. What is more, the proponents of each of these two approaches utilize research methods commensurate with the tenets of the respective approach. Thus, for example, constructivist researchers utilize impoverished or illusory stimulus conditions (e.g., brief and/or monocular presentations), calling for the intervention of 'higher' processes to compensate for the lack of adequate stimulus information. The adherents of the direct theory, on the other hand, utilize rich stimulus conditions without time limitations.

In spite of the seemingly irreconcilable differences between the two theoretical approaches, a means of bridging them does exist. They can co-exist within a broader theory of perception (Norman 2001). This broader theory is based on the accumulating neurophysiological, neuropsychological, and psychophysical research findings that point to the existence of two distinct visual systems, the dorsal and the ventral (see, e.g., Milner and Goodale 1995). The central idea is that the direct approach broadly parallels the functions of the dorsal system, and the constructivist approach broadly parallels those of the ventral system. These two visual systems are found in different parts of the brain, the dorsal mainly in the parietal lobe, the ventral mainly in the temporal lobe. The two deal with different aspects of perception. The dorsal system deals mainly with the utilization of visual information for the guidance of motor behavior in one's environment. The ventral system deals mainly with the utilization of visual information for 'knowing' one's environment, i.e., identifying and recognizing items previously encountered and storing new visual information for later encounters. However, it should be stressed that both systems process similar information for somewhat different purposes. Thus, for example, both systems are involved in the perception of size, albeit with somewhat different specializations. The dorsal system picks up size information for motor tasks, such as grasping an object. The ventral system utilizes size information for distinguishing between objects, say, between a real car and an accurate toy model.

Emerging from this attempt at integrating the two approaches into a single theory is a dual-process approach to visual perception. According to this approach, much of our day-to-day pickup of visual information is carried out by the dorsal-direct system without involving much awareness. In the main the information picked up is that which allows the organism to function within its environment, i.e., Gibson's affordances. The ventral-constructivist system, on the other hand, deals with the interface between the visual input and cognition, and we are normally aware of its output. Only it possesses a long-term memory and therefore any type of identification or recognition must transpire within it. As the dorsal system is mainly concerned with directing motor behavior in one's environment it must rely on body-centered information, in absolute units, about the environment and the objects in it. In contrast, the ventral system in its attempt to recognize objects can suffice with relative, object-centered, information. It is probably for this reason that the dorsal system yields more accurate information about the visual environment than does the ventral system. Thus, it is suggested that perceptual constancy is usually attained by the dorsal system. However, there are cases where the ventral system is also involved, as when the visual information is sparse or ambiguous, or when the

observer is required not only to act upon some aspect of the visual environment, but also to make a judgment about some object in that environment.

See also: Constructivism in Cognitive Psychology; Helmholtz, Hermann Ludwig Ferdinand von (1821–94); Object Recognition: Theories; Perception: History of the Concept; Perceptual Organization; Sensation and Perception: Direct Scaling; Vision, High-level Theory of

Bibliography

- Epstein W (ed.) 1977 *Stability and Constancy in Visual Perception*. Wiley, New York
- Gibson J J 1979 *The Ecological Approach to Visual Perception*. Houghton Mifflin, Boston
- Helmholtz H von 1867/1962 *Treatise on Physiological Optics* (Southall J P C, ed.), translation of the third German edition. Dover, New York
- Milner A D, Goodale M A 1995 *The Visual Brain in Action*. Oxford University Press, Oxford, UK
- Norman J 2001 Two visual systems and two theories of perception: an attempt to reconcile the constructivist and ecological approaches. *Behavioral and Brain Sciences* **24**(6) in press
- Rock I 1983 *The Logic of Perception*. MIT Press, Cambridge, MA
- Rock I 1997 *Indirect Perception*. MIT Press, Cambridge, MA
- Ross H E, Plug C 1998 The history of size constancy and size illusions. In: Walsh V, Kulikowski J (eds.) *Perceptual Constancy: Why Things Look as They Do*. Cambridge University Press, Cambridge, UK, pp. 499–528
- Walsh V, Kulikowski J (eds.) 1998 *Perceptual Constancy: Why Things Look as They Do*. Cambridge University Press, Cambridge, UK

J. Norman

Perceptual Learning

1. Improvement of Perception Through Training

It has been known for a long time that visual recognition of objects improves through training. For beginners in the study of histology, all specimens, such as the liver, lung, or kidney, look quite similar. Sooner or later, however, the advanced student wonders how one could possibly miss the difference. This type of visual classification is a relatively complex visual task, but it has also been known for a long time that performance in much simpler perceptual tasks improves through practice, as for example in the cases of vernier discrimination (McKee and Westheimer 1978), stereoscopic depth perception (Ramachandran and Braddick 1973), and discrimination between line orientations (Vogels and Orban 1985). This type of learning is usually referred to as 'perceptual learning.' Compared to other domains of research on learning,

perceptual learning is a relatively new topic, so its definition is even more important than those of topics that were established earlier. Gibson (1963) proposed the following definition for perceptual learning: 'Any relatively permanent and consistent change in the perception of a stimulus array following practice or experience with this array will be considered perceptual learning.' In recent years, one would, in addition, stress the fact that the neuronal mechanisms underlying perceptual learning involve rather 'early' stages of cortical information processing. Learning usually leads to improved performance, but there are exceptions, at least under laboratory conditions with manipulated feedback.

2. Terms Related to Perceptual Learning

Quite a number of processes in the nervous system are able to change their responses to sensory stimulation and there is quite a number of terms related to the term 'learning.' To understand the concept of perceptual learning clearly these terms have to be defined in order to clarify differences and similarities. These related terms include 'plasticity,' 'adaptation,' 'habituation,' 'after-effects,' 'priming,' 'development,' and 'maturation' as well as 'improvement through insight.'

'Plasticity' is defined here as modifiability of functional and anatomical organization of the central nervous system leading to more appropriate function as a result of sensory experience, or to overcome limitations caused by lesions. The term 'adaptation' is most often used in relation to adjustments of information processing within a predefined working range, as a result of stimulation. A classic example is luminance adaptation—adjusting the working range of the visual system to ambient light levels, with no long-term consequences. 'Habituation,' or 'satiation' in the context of rewards, seems to be a special case of adaptation, namely a shift of working range towards lower sensitivity, as in the case of decreased reflex response after repeated stimulation. 'After-effects' can be considered, in many cases, as the result of selective short-term adaptation in cases where perception is the result of antagonistic neuronal channels, as with many after-images. 'Priming' describes the effect of a (sometimes subliminal) stimulus on a subsequent stimulus-reaction pair, or more generally, the effect of initiating a certain type of behavior. The effect of priming is usually short. 'Development' and 'maturation,' unlike learning, ascribe the thrust of the changes in behavior to genetics, not the environment; hence the mechanisms underlying changes of behavior differ radically between these phenomena on one side and perceptual learning on the other. 'Improvement through insight,' as the name indicates, is a term that should be reserved for positive changes of information processing based on cognitive processes, such as one-shot learning.

So what is so special about perceptual learning that earns it a separate entry in this encyclopedia as opposed to a passing mention under the heading 'Learning'? I would like to stress the following difference between 'ordinary' vs. 'perceptual' learning. Most forms of learning to collect better information about the outside world are generally considered to rely on relatively intelligent or complex or cognitive levels of information processing. Hence, learning leading to improved discrimination between sensory stimuli would have been considered only a decade ago to be more of the declarative type of learning (as in learning a poem) than of the procedural type of learning (as in learning to play the piano). Perceptual learning, on the other hand, is a form of learning leading to better use of sensory information which is relatively independent of conscious or declarative forms of learning but relies partly on rather low-level modifications in the central nervous system. Perceptual learning hence resembles, in many respects, procedural forms of learning that are common in motor learning, for example learning to ride a bicycle.

3. Perceptual Learning and Plasticity of Primary Sensory Cortices

This type of low-level modification would have been considered impossible in the 1980s. At this time, the primary visual cortex of adults was considered to be a rather 'hard-wired' first stage of visual information processing, which served to detect and extract certain 'elementary features' from the complex scenes surrounding us (cf. Marr 1982). It was considered to lack plasticity since changing preprocessing during the training of one task might have disadvantageous consequences for solving other perceptual tasks. Indeed, there is direct experimental evidence for a decrease in plasticity of the primary visual cortex during maturation. Hubel and Wiesel (1965) found that covering one eye in young cats leads to changes in their cortical wiring patterns that are reversible during their kittenhood but not thereafter. Similarly, children suffering from a squint may achieve only low visual acuity in the deviating eye; therapy can reverse this loss in small children only. Therefore it seems that a critical development phase exists during which the primary visual cortex is still plastic, but that it loses this plasticity at a later stage of development. Improvement through training in perceptual tasks by adults was therefore supposed to take place on higher levels of cortical information processing.

However, during the last decade of the twentieth century, a number of electrophysiological and psychophysical experiments cast some doubts on this interpretation and suggested that even adult primary sensory cortices showed much more plasticity than was hitherto believed. Some of this evidence will be

presented here. A first example of this evidence was the orientation specificity of improvement through learning. For example, Fiorentini and Berardi (1981) found that discrimination between complex types of grating improved through practice, but that the improvement did not transfer to stimuli rotated by 90 degrees. The same was true for a vernier discrimination task where observers had to indicate whether the lower segment of a vertically oriented vernier stimulus consisting of two almost collinear bars was offset to the right or to the left relative to the upper segment, or whether the right segment of a horizontal vernier was above or below the left segment. Observers improved their performance very significantly during training, but their performance returned to base level when the stimulus was rotated by 90 degrees. Hence, the improvement obtained through training was specific for stimulus orientation.

In a control experiment, the pushbuttons by which observers had to indicate the direction of offset were switched between hands. Observers had to push the left button when a vernier was offset to the right, and vice versa. Results did not deteriorate, indicating that not every change of experimental procedure leads to a drop in performance. Another group of observers were trained in vernier discriminations at eight different positions in the periphery of the visual field for one hour at each position. During that time, they improved performance, but returned to baseline at the transition from one visual field position to the next. The same was true for training vernier discriminations with one eye covered: observers improved their performance through training, but when they performed the same discriminations with the cover moved to the other eye, performance dropped to pretraining levels. The same specificity of improvement for the eye used during training had been observed earlier in a texture discrimination task in which observers had to discriminate a figure from its surround based on the orientation of the stimulus elements (Karni and Sagi 1991). Moreover, improvement did not transfer between a three-dot vernier and a three-dot bisection task, although the stimuli of both tasks differed by approximately one photoreceptor diameter.

4. Neuronal Basis of Perceptual Learning

Hence, the changes of the nervous system underlying these forms of perceptual learning should occur on a quite early level of cortical information processing, in which the neurons are already selective for different stimulus orientations—unlike in the retina—but still partly monocularly activated—unlike in all cortical areas beyond the primary visual cortex. In particular, the eye specificity of improvement points to the primary visual cortex as the most probable site for at least a large part of the changes underlying these forms of perceptual learning.

Recent electrophysiological evidence supports this conclusion of a relatively early site of parts of the neuronal changes underlying perceptual learning. Single-cell recordings in adult monkeys have demonstrated that receptive fields of neurons in the primary visual cortex of adult primates can change position after the parts of the retina that were supplying the input to these cells have been destroyed (Gilbert and Wiesel 1992), that the microcircuitry of the adult visual cortex can change (Eysel 2001), and that the distribution of mass potentials evoked by visual stimulation in humans changes as a result of training, especially pronounced for short latencies and over the primary visual cortex.

5. Different Cortical Levels of Perceptual Learning

Marr's insight is still true—changing the very front end of information processing as a result of learning one perceptual task would necessarily change the processing of many, if not all, other stimuli presented to the same sensors. The speed and amount of learning depend strongly on attentional control, that is on top-down influences within the brain. Hence, present models of perceptual learning increasingly emphasize that learning will occur at quite a number of different levels of information processing, and that top-down influences from 'higher' levels will play a crucial role in adjusting, in a task-dependent way, processing on the 'lower' levels. Recent experiments have indeed provided direct evidence for strong top-down influences on perceptual learning as a result of error feedback and attentional processes (Ahissar and Hochstein 1993, Herzog and Fahle 1998).

It seems that the high specificity of perceptual learning is partly lost if relatively easy tasks are learned, while specificity of improvement is highest for very difficult tasks. A possible explanation is that 'easy' tasks are learned on a relatively 'higher' level of information processing, at which the information extracted from the visual scene is better used after training than was possible before. Difficult tasks, on the other hand, may require changes at lower levels of processing that are specific, for example, to exact visual field position as well as stimulus orientation.

6. Conclusions

To conclude, perceptual learning differs from other forms of learning, especially declarative forms, in that it can be very task- and stimulus-specific, and probably involves functional and anatomical changes even in primary sensory cortices. Though at first sight perceptual learning seems to resemble declarative forms of learning, and to rely on relatively complex cognitive processes, the specificity of improvement for quite low-level attributes such as visual field position,

stimulus orientation, and the eye used for training a visual task indicates a strong and crucial involvement of the primary visual cortex, where neurons are still partly monocularly activated. Dependence on attention, error feedback, and 'insight,' on the other hand, demonstrates that strong top-down influences play a major role in perceptual learning and that perceptual learning also involves more cognitive levels of the brain. Hence, the study of perceptual learning processes not only shows us the amazing amount of plasticity even in adult sensory information processing at a relatively peripheral level, but also leads to a view of cortical information processing not as a feed-forward system of subsequent neuronal layers but as a complex and plastic feedback system with strong and important top-down influences that shape 'lower' or 'early' parts of information processing.

Bibliography

- Ahissar M, Hochstein S 1993 Attentional control of early perceptual learning. *Proceedings of the National Academy of Sciences of the USA* **90**: 5718–22
- Eysel U 2001 Plasticity of reactive fields on early stages of the adult visual system. In: Fahle M, Poggio T (eds.) *Perceptual Learning*. MIT Press, Cambridge, MA
- Fahle M, Poggio T (eds.) 2001 *Perceptual Learning*. MIT Press, Cambridge, MA
- Fiorentini A, Berardi N 1981 Perceptual learning specific for orientation and spatial frequency. *Nature* **287**: 43–4
- Gibson E J 1963 Perceptual learning. *Annual Review of Psychology* **14**: 29–56
- Gilbert C D, Wiesel T N 1992 Receptive field dynamics in adult primary visual cortex. *Nature* **356**: 150–2
- Herzog M H, Fahle M 1998 Modeling perceptual learning: Difficulties and how they can be overcome. *Biological Cybernetics* **78**: 107–17
- Hubel D H, Wiesel T N 1965 Binocular interaction in striate cortex of kittens reared with artificial squint. *Journal of Neurophysiology* **28**: 1041–59
- Karni A, Sagi D 1991 Where practice makes perfect in texture discrimination: Evidence for primary visual cortex plasticity. *Proceedings of the National Academy of Sciences of the USA* **88**: 4966–70
- Marr D 1982 *Vision*. Freeman, San Francisco
- McKee S P, Westheimer G 1978 Improvement in vernier acuity with practice. *Perception & Psychophysics* **24**: 258–62
- Ramachandran V S, Braddick O 1973 Orientation-specific learning in stereopsis. *Perception* **2**: 371–6
- Vogels R, Orban G A 1985 The effect of practice on the oblique effect in line orientation judgments. *Vision Research* **25**: 1679–87

M. Fahle

Perceptual Organization

Unless we adopt a rather artificial analytical attitude, we perceive an environment neatly partitioned into objects rather than a mere collection of sensations.

Here is a (red) cherry lying on a (yellow) plate, not merely a blob of red within a yellow surround. The concept of perceptual organization, the grouping of parts into larger units, originated with Gestalt psychology in its attack on the atomistic assumption of structuralism around the beginning of the twentieth century. A then widely held view was that complex percepts are simply additive combinations of sensory elements. Each sensory 'atom' was thought to be independent of all other atoms, tied together to larger complexes just by mechanisms of associative learning. The Gestaltists, on the contrary, emphasized that what we perceive is not 'atomistic' elements, but structured forms and shapes which interact within a perceptual 'field' of dynamic 'forces.' Perception was considered to rely on interactive processes within the nervous system determined by dynamic field forces.

1. Grouping Factors in Vision

Rules for how parts of visual stimuli are grouped together, 'Gestalt laws,' were first formulated by Max Wertheimer (1923) who essentially identified four factors: nearness (proximity), similarity, common fate, and closure.

1.1 Nearness or Proximity

Other factors being equal, the nearer objects are to one another, the more likely they are to be organized into unified percepts. An example is shown in Fig. 1a, where the dots are easily organized into pairs of dots on the basis that each two neighboring dots form a unit. It is almost impossible to see the spaced dots as belonging together or, in other words, to break up the units based on nearness.

1.2 Similarity

The greater the similarity among objects, the more likely they are to be organized into perceptual units. For example, dots of the same color are readily seen as a group in distinction from dots of another color (Fig. 1b).

1.3 Common Fate

Dots which move simultaneously at the same rate and in the same direction, share a common fate, are readily seen as a group. The factor of common fate belongs to the strongest grouping factors; it ties together objects that are quite distant and different in form, size or color, given that they share a sort of event similarity in their sameness of motion. Where grouping would not otherwise occur in stationary displays, it will occur as soon as parts of the display move. A well-camouflaged animal will remain hidden only as long as it is standing still; it becomes visible ('pops out') as soon as it moves.

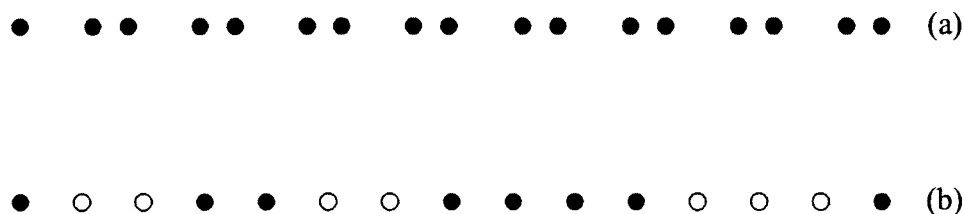


Figure 1
Demonstration of proximity (a) and similarity (b) as grouping factors

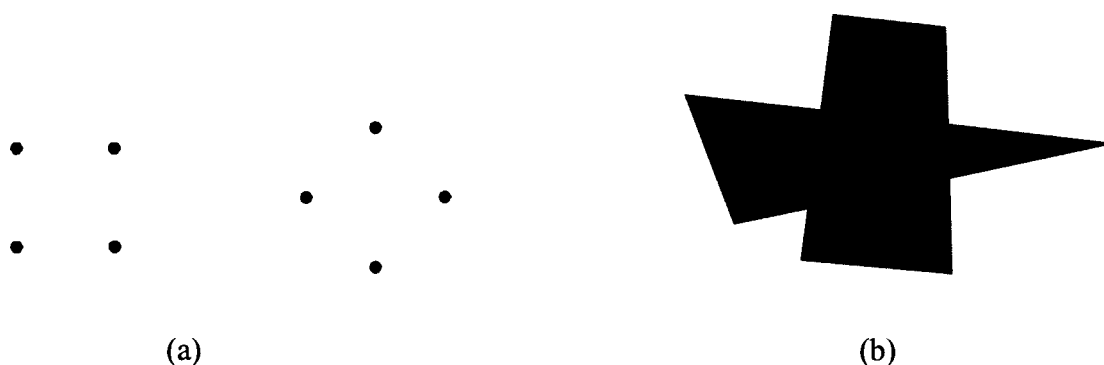


Figure 2
Demonstration of closure as grouping factor



Figure 3
Demonstration of *Einstellung* and past experience as factors in amodal completion

1.4 Closure

Components that constitute a closed entity rather than an open one are more readily organized into a single percept. This factor of closure predicts that a closed line has an advantage over an open one. For example, the two arrangements of four dots in Fig. 2a are seen as a rectangle or diamond rather than as crosses ($\times$) or (+), because the former are closed. Closure also applies to the tendency to complete otherwise incomplete, e.g., partly occluded, objects. The configuration shown in Fig. 2b appears as a trapezoid overlapping a triangle rather than as a complicated eleven-sided figure.

1.5 Subjective Factors

The study of natural camouflage and concealment shows that those grouping factors that refer to objective or stimulus variables fare well in accounting for the utility of various surface markings of animals. In addition, there are subjective or central factors at work, such as an observer's momentary set, or *Einstellung*, and past experience, or custom, in determining how the perceptual field is organized. For example, the three separate lines shown in Fig. 3a may appear, at first glance, as a meaningless geometric arrangement of lines. However, the same configuration, just rotated by 180° (Fig. 3b), will be easily

recognized as a capital E with the missing contours being perceptually completed. Now, if the reader with the *Einstellung* of an 'E' looks back at Fig. 3a, he or she will see a rotated E rather than a mere assembly of three lines. The perceptual formation of an E implies amodal completion of contours in the absence of stimulus (modal) information.

2. Figure–Ground Differentiation

Why do we see things rather than the holes between them? This question addresses another important aspect of organization, the differentiation of the perceptual field into figure and ground. This aspect was first brought out clearly by Edgar Rubin (1915). For example, if we look at a pencil on a desk, the pencil appears as a well-marked part of the perceptual field, the figure, while the desk appears as a relatively formless, featureless mass which we call the ground. The figure has the character of a thing, whereas the ground appears like unformed material. The ground seems to extend continuously behind or, sometimes, in front of the figure, i.e., although the figure usually appears in front, the ground is not always identical with background. For example, in looking through a window at a tree, the window screen appears as ground, but is clearly seen in front of the figure, the tree.

2.1 Dynamics of Figure–Ground Perception

Rubin pointed out that even if all parts of a configuration are grouped together properly, it can appear

either as object (figure) or surface (ground) under certain conditions. These so-called ambiguous or reversible figures 'flip' when, for example, the attention is directed to a specific point on the stimulus pattern. The dynamics of figure–ground perception due to the multivalence of the stimulus field are impressively demonstrated in a pattern shown in Fig. 4 (from Spillmann and Ehrenstein 1996)

New 'flowers' (figures) rise from the 'meadow' (ground) and disappear again, illustrating the vain attempt to establish a tendency toward a better and more stable percept. Apparently, the pattern allows for several groupings of equal probability and hence perception fails ever to reach a steady state. This example nicely illustrates that the visual space, as Kanizsa (1979, p. 181) artfully put it, is not a static geometrical scheme, nor is it a simple transposition on the perceptual level of the topographical arrangement of the retinal stimulus, but rather should be considered the result of an extremely dynamic event.

2.2 Tendency Towards *Prägnanz*

The Gestaltists assumed that the process of figural segregation and differentiation would follow a general tendency towards figural goodness or *Prägnanz* (Wertheimer), which denotes a processing principle for percepts to become structured in the most regular, symmetrical, and simple manner possible under the prevailing stimulus conditions. This process was understood by Wolfgang Köhler as a kind of self-organization according to electrodynamic field forces inherent in the sensory system. 'In all processes which terminate in time-independent states the distribution

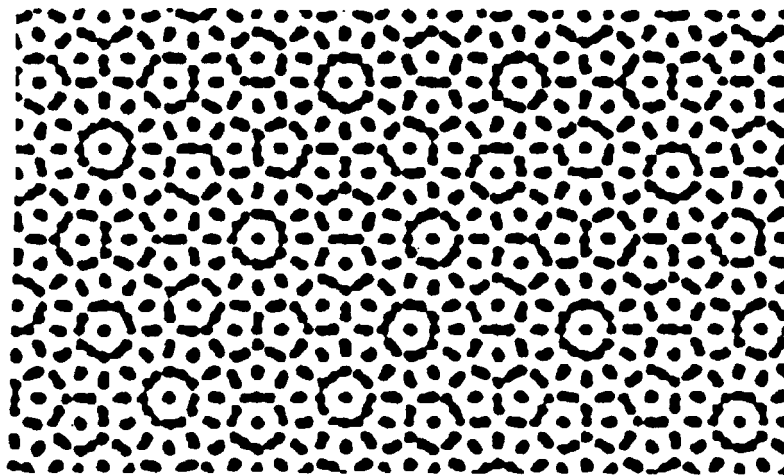


Figure 4

A pattern which affords a never-ending fluctuation in grouping (after Spillmann and Ehrenstein 1996)

shifts towards a minimum energy' (Köhler 1920, quoted from Koffka 1935, p. 108). More than any other principle, the tendency towards *Prägnanz* emphasizes the global aspects of perceptual organization, since even minor local changes in the stimulus or in the observing process (such as eye movements or changes in attention) can result in a reorganization of the entire percept (see Fig. 4). A modern approach to the problem of self-organization and *Prägnanz* can be found in the theory of neural synergetics with its tendency towards ordered steady states in dynamic vector fields (Haken and Stadler 1990).

3. Perceptual Organization in Other Sense Modalities

Although the principles of perceptual organization have been studied most thoroughly for the visual modality, they also apply to other sense modalities. For auditory perception, Bregman (1990) has provided many examples demonstrating the resemblance between audition and vision. Temporal proximity of tones is more important than spatial proximity: Tones that follow each other close together in time are perceived as belonging together. Similarity is determined primarily by the pitch of tones: tones similar in pitch tend to be grouped together in perception. When a sequence of alternating high- and low-frequency tones is played at a certain rate, listeners are able to follow the entire sequence of tone, provided that the frequency difference between tones is small and the tones are played at a slow rate. With greater frequency differences or higher rates, however, the sequence splits into two streams, one high and one low in pitch (auditory stream segregation).

Other examples—studies in tactile perception, for instance—confirm the notion that the above-mentioned factors of perceptual grouping play a similar role in different modalities. This suggests that the perceptual system can utilize organizing principles from one sensory modality to the other, i.e., in a cross-modal way. Moreover, cross-modal interactions can change the percept. A recent example concerns the perception of emotions. Listeners having to judge the emotion in the voice are influenced by whether the (visually displayed) face expresses the same emotion or a different one (De Gelder and Vroomen 2000).

4. Neural Correlates of Perceptual Organization

More than 70 years after Wertheimer's pioneering formulation of perceptual grouping factors, terms that were originally introduced to describe qualitatively properties of figural grouping and segregation are being used as a much-needed vocabulary for interpret-

ing spatiotemporal processing of sensory neurons and neuronal systems. The Gestaltists themselves, notably Wolfgang Köhler, tried to link perceptual organization to brain processes. Inspired by the laws of thermodynamics, the brain was thought to be a homogeneous conductor of bioelectric forces rather than a complex network of interacting neurons (e.g., Köhler and Held 1949). However, the field forces postulated by Köhler as factors in perceptual organization could not be substantiated by physiological evidence. Neural correlates of perceptual grouping are today assumed to consist of receptive field organization and synchronized coupling of neuronal activity within and across cortical areas (e.g., Spillmann and Ehrenstein 1996).

4.1 Receptive Field Organization: A Micro-Gestalt

The processing of sensory input obeys essentially two principles, convergence and divergence. Many receptors send their input to the same ganglion cell (convergence) and conversely one and the same receptor distributes signals to several ganglion cells (divergence). The receptor surface area within which a change of stimulation can modify the firing rate of a given neuron is called its receptive field. In vision, receptive fields of retinal, geniculate, and cortical neurons are concentrically organized with an excitatory center and an inhibitory surround (on-center neurons) or with a center that is inhibited by light and a surround that is activated by light (off-center neurons). The functional unit represented by the receptive field (with its center-surround antagonism and selective spatial and temporal sampling) is a basic neuronal mechanism that achieves non-additive integration of stimulus input. It has therefore been considered a micro-Gestalt by Spillmann and Ehrenstein (1996) with the conjecture that perceptual organization is largely determined by receptive-field organization.

4.2 A Neurocomputational Example

A detailed neurocomputational model of how perceptual groupings might emerge from interactions of cells with known receptive-field properties of the lateral geniculate and cortical areas V1 and V2 has been recently proposed by Grossberg et al. (1997). For example, global properties of Gestalt grouping (as demonstrated in Fig. 5) can be accounted for by their model that involves multiple levels of thalamo-cortical interaction to generate emergent boundary groupings.

The left side of Fig. 5 (from Grossberg et al. 1997) shows an ambiguous configuration (A, upper part) which allows the perception of vertical or horizontal groupings of bar elements and the simulated outcome

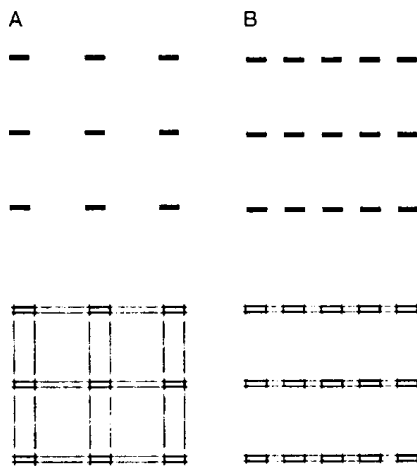


Figure 5
Perceptual grouping: above stimulus patterns A and B, below the perceived grouping as predicted by the model (after Grossberg et al. 1997)

by the neurocomputational model (A, lower part). Adding horizontal bar elements (Fig. 5, B) leads to horizontal grouping in perception (upper part) and also in the model (lower part).

4.3 'Perceptive' Neurons and Synchronized Coupling of Spike Activity

In a few cases 'perceptive' neurons (Baumgartner 1990) have been reported that mediate invariant stimulus properties, e.g., neurons in area V4 of primate cortex seem to signal the perceived color of a stimulus, irrespective of its spectral composition and changes in illumination. The striking correlations between the activity of highly specialized single neurons and complex perceptual performance support the view that neurons at higher levels in the processing hierarchy may be sufficient to convey Gestalt perception. However, search for individual neurons responding selectively to complex configurations has, so far, revealed a very limited set of biologically significant stimuli. Alternatively, synchronized oscillations of even far-distant neurons allow for perceptual coherence over a large area of the sensory field. The above-mentioned grouping factor of common fate could especially be accounted for by temporal binding of features (e.g., Eckhorn 1999). In fact, synchronized coupling of spike activity between remote brain areas has been found to be context-dependent and to account for spatiotemporal coherence in perception. Moreover, synchronized feature binding affords individual cells and cell groups flexibly to change partners with which they share activity when the stimulus configuration changes. Thus, single-cell ac-

tivity, organized within a hierarchy of receptive fields, and synchronized feature binding seem to provide the basic neural mechanisms subserving perceptual grouping.

See also: Binding Problem, Neural Basis of; Modularity versus Interactive Processing, Psychology of; Neural Networks: Biological Models and Applications; Neural Synchrony as a Binding Mechanism; Object Recognition: Theories; Perception: History of the Concept; Perception: Philosophical Aspects; Perceptrons; Perceptual Constancy: Direct versus Constructivist Theories; Psychophysical Theory and Laws, History of

Bibliography

- Baumgartner G 1990 Where do visual signals become a perception. In: Eccles J C, Creutzfeldt O (eds.) *The Principles of Design and Operation of the Brain*. Springer, Berlin, pp. 99–114
- Bregman A S 1990 *Auditory Scene Analysis*. MIT Press, Cambridge, MA
- De Gelder B, Vroomen J 2000 The perception of emotion by ear and eye. *Cognition and Emotion* **14**: 289–311
- Eckhorn R 1999 Neural mechanisms of visual feature binding investigated with microelectrodes and models. *Visual Cognition* **6**: 231–65
- Grossberg S, Mingolla E, Ross W D 1997 Visual brain and visual perception: How does the cortex do perceptual grouping? *Trends in Neurosciences* **20**: 106–11
- Haken M, Stadler M (eds.) 1990 *Synergetics of Cognition*. Springer, Berlin
- Kanizsa G 1979 *Organization in Vision. Essays on Gestalt Perception*. Praeger, New York
- Koffka K 1935 *Principles of Gestalt Psychology*. Harcourt Brace, New York
- Köhler W, Held R 1949 The cortical correlate of pattern vision. *Science* **110**: 414–19
- Rubin E 1915 *Synsoplevede Figurer*. Gyldendal, Copenhagen, Denmark
- Spillmann L, Ehrenstein W H 1996 From neuron to Gestalt. Mechanisms of visual perception. In: Greger R, Windhorst U (eds.) *Comprehensive Human Physiology*. Springer, Berlin, Vol. 1, pp. 861–93
- Wertheimer M 1923 Untersuchungen zur Lehre von der Gestalt. II. *Psychologische Forschung* **4**: 301–50

W. H. Ehrenstein

Performance: Anthropological Aspects

Anthropology uses the concept of performance to refer to distinctive events or aspects of behavior, contributing to the understanding of human behavior in traditional, nontraditional, and avant-garde con-

texts. The current applications range from the most extraordinary and stylized forms of human behavior to the fundamental routines of everyday life, and address different areas of specialized professionalism and expertise, in ritual, commercial, economic, medical, dramatic, and sportive spheres. Initially driven by a theoretical concern with social function and meaning, contemporary research uses the performance concept to address issues of identity and the relationship between performance, self, researcher, and social space.

1. Development of the Concept

Performance as an anthropological concept developed after the 1950s. Previously, anthropologists subsumed performance into the analysis of ritual. In British social anthropology, which had built its reputation as a science of total functioning systems and structures centred on kinship networks, specialized forms of human behavior occurring within modern or commercial venues, such as acting or dancing, were considered marginal subjects. By the mid-1960s, however, broader humanist concepts of culture, more powerful in anthropology elsewhere in Europe and in the USA, allowed the concept of performance to enter more centrally into anthropological analyses.

The American sociologist Erving Goffman (1956, p. 23) developed a model in which social interaction was explicable as acts and performances: 'The world, in truth, is a wedding,' and performance is 'all of the activity of a given participant on a given occasion which serves to influence in any way any of the other participants' (Goffman 1956, p. 8). Goffman unified the analysis of roles and situations using theatrical metaphors such as front and back stage to explain different aspects of behavioral presentation and 'multiple versions of reality.' This model was intended as a fifth analytical perspective to complement existing technical, political, structural, and cultural ones. This dramaturgical and transactional vision of social interaction remains influential.

Milton Singer, Dell Hymes, Clifford Geertz, and Victor Turner have also shaped the ways in which anthropologists use performance notions metaphorically to explain social action as a complex system of symbolic and meaningful behavior which produces an ordered understanding of the world (Geertz 1973). Performance in these terms is not limited to representation, but is an active and creative precondition for social life.

Victor Turner's work has been particularly influential on current uses of performance. He moved from a functionalist study of ritual to cultural performance models and finally to theatrical performance itself, though performance was never separated completely from ritual, a 'transformative performance revealing major classifications, categories, and contradictions of

cultural process' (Ronald L. Grimes, cited in Turner 1986, p. 75). Turner developed three linked explanatory concepts. 'Ritual process' consists of separation, transition or liminality, and reintegration. Anti-structure or 'communitas' contrasts with the structure of everyday life and describes the creative and expressive behavior associated with a generalized notion of liminality beyond the simple process of a specific ritual event. 'Social dramas' extend the ritual process into other kinds of social scenario. This framework merged the borders between initiation rituals, social events such as carnivals, and particular theatrical performances, as what Wilhelm Dilthey called 'lived experience' (Turner 1986, pp. 84–90).

Turner was part of Max Gluckman's 'Manchester School' of anthropology, but he has been more influential in the USA than others from this school. Other scholars of this background produced actor-centered transactional and conflict models using performance perspectives on more political topics, often drawing on Marxist theory. In his study of London's Notting Hill Carnival, Abner Cohen (1993) reaffirms his interest in performance as part of a wider analysis of the relationship between cultural forms and political formations: performance as a part of political symbolism is interesting because of the contradictions between the two dimensions of social interaction: the unintentional and moral nature of cultural action, and the amoral self-interest of political action. Cohen's work has influenced the analysis of performance as a political tool in postcolonial societies where processes of retribalization and issues of ethnic identity have been so dominant (Parkin et al. 1996).

The study of cultural performances has been inflected towards ritual and personal transformation, or towards contradictions and political conflict, although the proponents of each 'side' would all claim to be attempting to bridge such dualisms. None the less, labels such as 'humanist' and 'Marxist' continue to limit any such transcendence, despite the best efforts of those who have insisted on the multiplicity of social life (Tambiah 1985).

Other salient influences came from sociolinguistics. Theories which characterized communication as performative, and not just informative, were inspired by Jakobson's 'poetic function'—'the way in which a communication is carried out' (Bauman 1986, p. 3)—and Austin and Searle's speech act theory. For instance, Richard Bauman (1986, p. 112) has analyzed American folk narratives using a framework of story, performance, and event. This specialized use of performance has been very influential in establishing the performative dimensions of different kinds of communication, particularly in the study of competence and play. Anthropologists have also developed communicative approaches to the performance of metaphor in various social contexts (Fernandez 1986). Such work on verbal performance has influenced general theories of performance in human behavior,

but although it has moved away from a text-based model, it has produced somewhat disembodied accounts of performance, with little reference to the nonverbal gestural dimension of human communication and expression.

2. Methodological Innovations

These influences have legitimized performance as a topic within anthropology, and Turner's last writings on performances as liminoid phenomena have promoted an interest in nontraditional performance, overlapping with performance studies and cultural studies. There has also been a belated recognition of the importance of embodied behavior and bodily techniques.

Increasing attention is being paid to different performance skills and genres, not simply as objects of study but also methodologically. Anthropology has always been a discipline which attracts people with an experience of life outside academe, and the 1980s saw the emergence of performer-anthropologists, scholars who had previously worked on the stage professionally as dancers and actors. Anthropologists who have not been professional actors or dancers may choose to include training in performance as part of their research, using their bodies to explore familiar or unfamiliar techniques self-consciously by learning alongside aspiring professionals and experts. This recognition of the importance of embodiment is part of a broader trend to understand human knowledge as a reflexive situated practice, validated not by observation, but by experience (Hastrup 1995, p. 83).

The traditional anthropological interest in ritual which Turner extended to theater also resulted in an intellectual collaboration with Richard Schechner, a theatre director and academic. After Turner's death in 1983 Schechner kept the effort alive, both practically and theoretically. The continuum between performance with ritual efficacy and theatrical performance rests on the theory that performance is 'restored behavior', strips of behavior with 'a life of their own', which can be rearranged or reconstructed independently of the various causal systems which brought them into existence (Schechner 1985, p. 35).

Schechner has played an important role in forging a meeting between the performers, anthropologists and other researchers of intercultural performance. This wide and fuzzy interdisciplinary field has become well established in the USA and is gaining momentum in the UK and Europe. The International School of Theatre Anthropology founded by Eugenio Barba and Nicola Savarese has involved performers from Japan, India, Indonesia, and Brazil. Theater anthropology is not the study of the performative phenomena by anthropologists, but rather the study of the pre-expressive behavior of the human being in an organized performance situation (Barba 1995, p. 10).

In 1990, the anthropologist Kirsten Hastrup's life story was 'restored' (to use Schechner's term) as a drama, *Talabot* by Barba's Odin Teatret (Hastrup 1995, pp. 123–45). New and established anthropologists are also turning to different performance arenas, including cinema, television, and the internet, to explore further permutations of sociality which would have been unimaginable a hundred years ago (Hughes-Freeland and Crain 1998).

Anthropological analyses and uses of performance are not limited to the technicalities of theater and acting. Johannes Fabian (1990) has argued that performance must deal with the political, and has worked with a theater troupe in Zaire to explore local concepts of power through drama and television. The sociolinguistic approach to performance has also been applied to analyze powerful speaking in Sumba, East Indonesia, which has been repressed by the State and transformed into textbook culture (Kuipers 1990). The social processes of power are also the center of Ward Keeler's (1987) study of shadow puppet theatre in Java. Studies such as these are found among the ethnography of every region, and often remain independent of the meeting between anthropology and performance studies.

The relationship between ritual and performance is generating approaches to the study of healing in a number of interesting ways. Such studies use transactional and conflict models but also consider the power of the aesthetic and expressive dimensions of performance in the face of political repression or disenfranchisement. For example, the healing power of comedy in an exorcism in Sri Lanka restores the sick person to a sense of social integration, in line with G. H. Mead's theory (Kapferer 1991). Michael Taussig (1993) invokes Marx and the Frankfurt School to explore mimetic behavior which lies at the heart of shamanic healing and social imagery, simultaneously representing and falsifying the world, and demonstrates how irrationality is being performed at the heart of the everyday. 'Mimetic excess' creates 'reflexive awareness', which makes connections between history, make-believe, and otherness.

Bureaucratic resistance to the irrational is also at issue, as in the case of professional female performers in Indonesia, who had the power to mediate between the village spirit made manifest by the village elder and the community, and brought healing and protective energies to local communities (see Fig. 1). These performers are being made extinct by the processes of state education and professionalization, which tend to deny female power (Hughes-Freeland 1997).

Performance here differs from the rational calculation of Goffman's social agents, and introduces complexity into the debate about the extent of individual human choice and intentionality. Such explorations are a reminder of the continuing usefulness of studying ritual performance to understand how humans express their possession and dispossession,



Figure 1
Female dancer in Java protects a baby with face powder

and the permeability of the boundary between ritual performance and artistic performance.

More radically, anthropologists have applied healing processes from other cultures to help the sick at heart in their own societies by means of 'a theatre model of dramatherapy' (Jennings 1995, p. xxvii). While Aristotle's concept of catharsis may seem universal, different theatres produce forms of what Sue Jennings (1995, p. 185) calls 'abandonment', which vary greatly in their effects and forms of control. Ultimately it is this variation which must concern the anthropologist.

3. Central Issues

During the 1970s, the anthropological project was framed by questions about function and meaning, but these have been superseded by concerns with identity and experience. Performance is productive of individuals, with gendered, ethnic, or other specified markers of social identity. It has also been argued that it produces the experience of society as a continuum (Parkin et al. 1996, p. xvii). By using the performance concept in a heuristic manner, anthropologists are

able to develop insights into the acquisition of competence, as general socialization or as training to acquire particular skills, and to further their understanding of social interactions and formations.

The central problem facing anthropologists working on, and by means of, performance remains the question of boundaries. The topic continues to stimulate argument: is performance a discrete category, or is it an aspect of everyday behavior? How does performance help us to understand reality? (Hughes-Freeland 1998, pp. 10–15).

An interesting debate about performance is the extent to which human action is scripted or emergent. Is performance best understood as mimesis and imitation, or as *poesis*, creative, emergent, and transformative? Anthropological self-critiques are often formulated along the lines of performance models vs. textual models, and argue for process models vs. structural ones. One party stresses the determinacy of form, the other gives primacy to emergence, praxis, experience. These arguments situate performance centrally within current polemics about anthropological knowledge, and whether it should represent itself as operating within a system of models and rules, or performatively and dialectically. Anthropologists

now approach performance as being risky and interactive, rather than determined and predictable, using active and participatory models of spectatorship, not passive ones, on a continuum of general varieties of human interaction.

Some commentators have questioned the general applicability of the Western view of performance as illusion, as theatrical, as unfactual, since such a view does not necessarily accord with the ways in which activities categorized as performances by anthropologists may be understood by members of the societies in which they occur (Schieffelin 1998). Others by contrast suggest that the exposure of the acting body makes deception impossible (Hastrup 1995, p. 91). By questioning critically the presuppositions supporting categorical oppositions implied in performance, such as pretense vs. authenticity, and by going against the grain of their own cultural presuppositions, anthropologists are able to construct descriptions of alternative ways of seeing and being in the world, and to celebrate the diversity of human cultural experience despite its shared biophysical nature. Anthropologists thus differ in their treatment of performance from those working in other disciplines.

4. Future Directions

As human interaction intensifies across time and space, anthropological approaches to performance recognize the complexity of contemporary life and the superficiality of dichotomizing the 'everyday' vs. 'ritual' realities. Information and media technologies facilitate virtual realities, and complicate the different ways in which performance may be transmitted and received, beyond the defined space-time continua of theatre, ritual arena, or event. Technological innovations in this way make it necessary for the analysis of performance to spill over again from specialized behavioral contexts into emergent everyday lives ((Hughes-Freeland and Crain 1998).

Visual technologies can bring all kinds of performance to us in nonverbal ways, unmediated by words, using images instead. Although these images are not pure imitations of the performances as they occur, it could be argued that the researcher is freer to respond directly to the performance than when reading about it. For example, Jean Rouch's film of the Hauka possession cult in Ghana in 1953, *Les Maîtres Fous* (*The Mad Masters*) presents images which language might fail to represent as convincingly real. Experimental ethnographic explorations of the relationship between embodied performance and text suggest innovative and exciting approaches to issues of analysis and representation in anthropology.

The potential fields for the development of anthropological approaches to performance are numerous. The study of linguistic performance has already been employed in communication analyses in the field

of artificial intelligence, a development which might extend to understandings of embodied performance. The anthropology of performance, perhaps the most ostensibly humanist of the new sub-areas of anthropology, could move increasingly into the field of medical and information technologies, theoretical and practically. In the more immediate future, anthropologists work to develop a less culturally entrenched view of human performance activities, and engage in intracultural or cross-cultural performances to promote wider understanding about the different ways of acting human within the constraints of the group, and the challenges performance presents to those constraints.

See also: Dramaturgical Analysis; Sociological; Entertainment; Goffman, Erving (1921–82); Ritual; Sociolinguistics

Bibliography

- Barba E 1995 *The Paper Canoe*. Routledge, London
- Bauman R 1986 *Story, Performance, and Event*. Cambridge University Press, Cambridge, UK
- Cohen A 1993 *Masquerade Politics*. Berg, Oxford, UK
- Fabian J 1990 *Power and Performance*. University of Wisconsin Press, Madison, WI
- Fernandez J W 1986 *Persuasions and Performances: The Play of Tropes in Culture*. Indiana University Press, Bloomington, IN
- Geertz C 1973 *The Interpretation of Cultures: Selected Essays*. Basic Books, New York
- Goffman E 1956 *The Presentation of Self in Everyday Life*. University of Edinburgh Social Sciences Research Centre, Monograph No. 2, Edinburgh, UK
- Hastrup K 1995 *A Passage to Anthropology: Between Experience and Theory*. Routledge, London
- Hughes-Freeland F 1997 Art and politics: from Javanese court dance to Indonesian art. *Journal of the Royal Anthropological Institute* 3(3): 473–95
- Hughes-Freeland F (ed.) 1998 *Ritual, Performance, Media*. Routledge, London
- Hughes-Freeland F, Crain M M (eds) 1998 *Recasting Ritual: Performance, Media, Identity*. Routledge, London
- Jennings S 1995 *Theatre, Ritual and Transformation: The Senoi Temiars*. Routledge, London
- Kapferer B 1991 *A Celebration of Demons*, 2nd edn. Berg/Smithsonian Institution Press, Oxford, UK/Princeton, NJ
- Keeler W 1987 *Javanese Shadow Plays, Javanese Selves*. Princeton University Press, Princeton, NJ
- Kuipers J C 1990 *Power in Performance*. University of Pennsylvania Press, Philadelphia, PA
- Parkin D, Caplan L, Fisher H (eds.) 1996 *The Politics of Cultural Performance*. Berghahn Books, Oxford, UK
- Schechner R 1985 *Between Theater and Anthropology*. University of Pennsylvania Press, Philadelphia, PA
- Schieffelin E L 1998 Problematising performance. In: Hughes-Freeland F (ed.) *Ritual, Performance, Media*. Routledge, London and New York, pp. 194–207
- Tambiah S J [1981] 1985 *Culture, Thought and Social Action: An Anthropological Perspective*. Harvard University Press, Cambridge, MA
- Taussig M 1993 *Mimesis and Alterity*. Routledge, New York

Turner V W 1986 [1987] *The Anthropology of Performance*.
Performing Arts Journal, New York

F. Hughes-Freeland

Performance Evaluation in Work Settings

Job performance is probably the most important dependent variable in industrial and organizational psychology. Measures of job performance are typically employed for many research and practice applications. Examples are evaluating the effects of a training program or a job redesign effort on job performance, and assessing the validity of a selection system for predicting performance. For these and other personnel-related interventions, what is needed are accurate measures of job performance to assess the effectiveness of the intervention.

This entry first introduces the concept of criterion relevance and other 'criteria for criteria.' The second topic discussed is the issue of whether work performance is best characterized by multiple criteria or by a composite criterion, that is, a single summary index of performance, and third, methods of measuring criterion performance are presented. Finally, what appear to be the most important future directions in describing and measuring work performance are highlighted.

1. Criterion Relevance and Other Standards for Criteria

What differentiates good from poor criterion measurement? That is, what are the most important features of good criteria? The most important is relevance. Relevance can be defined as the correspondence between criterion measures and the actual performance requirements of the job. Criterion measures should reflect all important job performance requirements. Useful in this context are the terms deficiency and contamination. Deficiency for a set of criterion measures occurs when the set is relevant to only part of the criterion domain. A job knowledge test for an insurance agent will probably fail to cover the interpersonal skill part of the job. Contamination refers to a criterion measure tapping variance in performance beyond the control of the organization member. For example, sales volume for account executives may be based in part on their performance, but also on how easy it is to sell the product in their own territory. A perfectly relevant set of criterion measures is neither contaminated nor deficient.

The second most important standard for good criterion measurement is reliability. Criterion scores should be consistent over, at least, relatively short time

intervals. For example, if a criterion measure for fast food servers is ratings of customer service quality by multiple customers, we would like to see consistency (i.e., reliability) in how different customers rated the servers, with good performers in this area rated consistently high, poorer performers, consistently lower, etc. There are other criteria for criteria that some authors have mentioned (e.g., acceptability of measures to the sponsor) but relevance and reliability are clearly the most important.

2. Multiple and Composite Criteria

An issue in measuring work performance is whether a single criterion should represent the performance requirements of a job or whether multiple criteria are required. Advocates for a single composite criterion see criteria as economic in nature, whereas supporters of multiple criteria view them more as psychological constructs. Probably the best way to resolve the issue is to recognize that the purpose of performance measurement largely dictates the appropriateness of the two views (Schmidt and Kaplan 1971). For example, when making compensation decisions for a unit's members based on performance, it is necessary to obtain a single performance score at some point for each employee. To make these decisions, we need an index of overall performance, worth to the organization, or some similar summary score. On the other hand, if the goal is to understand predictor-criterion links in personnel selection research, for example, then using multiple criteria and examining relationships between each predictor (e.g., ability, personality, etc.) and each criterion (e.g., technical proficiency) is probably most appropriate.

A related issue is the empirical question of how highly correlated multiple criteria for a job are likely to be. If correlations between criterion measures are high, then a single composite measure will be sufficient to represent the multiple criteria. On balance, research suggests that most jobs have multiple performance requirements. Performance on these multiple dimensions may be positively correlated but not so highly that distinctions between dimensions are impossible.

An attempt has been made to identify the criterion constructs underlying performance across jobs. Campbell et al. (1993) argue that eight dimensions (i.e., Job Specific Technical Proficiency, Non Job Specific Technical Proficiency, Written and Oral Communication, Demonstrating Effort, Maintaining Personal Discipline, Facilitating Peer and Team Performance, Supervision/Leadership, Management /Administration) reasonably summarize the performance requirements for all jobs in the US economy. Not every job has every dimension as relevant, but the eight dimensions as a set reflect all performance requirements across these jobs. Criterion research in

Project A, a large-scale selection study in the US Army, empirically confirmed several of these dimensions for 19 jobs in the Army. More will be said about these kinds of criterion models in the section entitled Future.

A final issue in this area focuses on the boundaries of 'work performance.' How should we define performance? Recent attention has been directed toward organizational citizenship behavior (OCB), prosocial organizational behavior, and related constructs (e.g., Organ 1997) as criterion concepts that go beyond task performance and the technical proficiency-related aspects of performance. For example, the OCB concept includes behavior related to helping others in the organization with their jobs and conscientiously supporting and defending the organization. Research has shown that supervisors making overall performance judgments about employees weight OCB and similar constructs about as highly as these employees' task performance (e.g., Motowidlo and Van Scotter 1994). Also, a link between OCB on the part of organization members and organizational effectiveness has some empirical support (e.g., Podsakoff and MacKenzie 1997).

3. Methods of Measuring Criterion Performance

Two major types of measures are used to assess criterion performance: ratings and so-called objective measures. Ratings—estimates of individuals' job performance made by supervisors, peers, or others familiar with their performance—are by far the most often used criterion measure (Landy and Farr 1980). Objective measures such as turnover and production rates will also be discussed.

3.1 Performance Ratings

Performance ratings can be generated for several purposes, including salary administration, promotion and layoff decisions, employee development and feedback, and as criteria in validation research. Most of the research on ratings has focused on the latter for-research-only ratings. Research to be discussed is in the areas of evaluating the quality of ratings, format effects on ratings, and different sources of ratings (e.g., supervisors, peers).

3.1.1 Evaluation of ratings. Ratings of job performance sometimes suffer from psychometric errors such as distributional errors or illusory halo. Distributional errors include leniency/severity where raters evaluate ratees either too high or too low in comparison to their actual performance level. Restriction-in-range is another distributional error. With this error, a rater may rate two or more ratees

on a dimension such that the spread (i.e., variance) of these ratings is lower than the variance of the actual performance levels for these ratees (Murphy and Cleveland 1995). Illusory halo occurs when a rater makes ratings on two or more dimensions such that the correlations between the dimensions are higher than the between-dimension correlations of the actual behaviors relevant to those dimensions (Cooper 1981).

A second common approach for evaluating ratings is to assess interrater reliability, either within rating source (e.g., peers) or across sources (e.g., peers and supervisor). The notion here is that high interrater agreement implies that the ratings are accurate. Unfortunately, this does not necessarily follow. Raters may agree in their evaluations because they are rating according to ratee reputation or likeability, even though these factors might have nothing to do with actual performance. In addition, low agreement between raters at different organizational levels may result from these raters' viewing different samplings of ratee behavior or having different roles related to the ratees. In this scenario, each level's raters might be providing valid evaluations, but for different elements of job performance (Borman 1997). Accordingly, assessing the quality of ratings using interrater reliability is somewhat problematic. On balance, however, high interrater reliability is desirable, especially within rating source.

A third approach sometimes suggested for evaluating ratings is to assess their validity or accuracy. The argument made is that rating errors and interrater reliability are indirect ways of estimating what we really want to know. How accurate are the ratings at reflecting actual ratee performance? Unfortunately, evaluating rating accuracy requires comparing them to some kind of 'true score,' a characterization of each ratee's actual performance. Because determining true performance scores in a work setting is typically impossible, research on the accuracy of performance ratings has proceeded in the laboratory. To evaluate accuracy, written or videotaped vignettes of hypothetical ratees have been developed, and target performance scores on multiple dimensions have been derived using expert judgment. Ratings of these written vignette or videotaped performers can then be compared to the target scores to derive accuracy scores (Borman 1977).

3.1.2 Rating formats. A variety of different rating formats have been developed to help raters evaluate the performance of individuals in organizations. Over the years, some quite innovative designs for rating formats have been introduced. In this section we discuss two of these designs: numerical rating scales and behaviorally anchored rating scales.

It may seem like an obvious approach now, but the notion of assigning numerical ratings in evaluating

organization members was an important breakthrough. Previously, evaluations were written descriptions of the ratee's performance. The breakthrough was that with numerical scores, ideally, well informed raters could quantify their perceptions of individuals' job performance, and the resulting scores provide a way to compare employees against a standard or with each other.

One potential disadvantage of numerical scales is that there is no inherent meaning to the numbers on the scale. To address this difficulty, Smith and Kendall (1963) introduced the notion of behaviorally anchored rating scales (BARS). These authors reasoned that different levels of effectiveness on rating scales might be anchored by behavioral examples of job performance (e.g., always finds additional productive work to do when own normally scheduled duties are completed—high performance level on Conscientious Initiative rating dimension). The behavioral examples are each scaled according to their effectiveness levels by persons knowledgeable about the job and then placed on the scale at the points corresponding to their respective effectiveness values, helping raters to compare observed performance of a rater with the behavioral anchors on the scale, in turn leading to more objective behavior-based evaluations.

3.2 Rating Sources

Supervisors are the most often used rating source in obtaining performance evaluations of organization members. An advantage of supervisory ratings is that supervisors typically are experienced in making evaluations and have a good frame of reference perspective from observing large numbers of subordinates. A disadvantage is that in at least some organizations supervisors do not directly observe ratee performance on a day-to-day basis. In addition, coworker or peer ratings are sometimes used to evaluate performance. A positive feature for peer ratings is that coworkers often observe employee performance more closely and more regularly than supervisors. A difficulty is that coworkers are less likely than supervisors to have experience evaluating ratee performance.

Other rating sources may also be involved in assessing job performance. In fact, a recent emphasis has been toward the concept of '360° ratings,' augmenting supervisory ratings with evaluations from peers, subordinates, self-ratings, and even customers' assessments. The general idea is to assess performance from multiple perspectives so that a balanced evaluation of job performance might be obtained (Bracken 1996).

3.3 Objective Criteria

A second major measurement method for criterion performance involves use of objective criteria. Ob-

jective criteria employed in personnel research include turnover, production rates, and work samples of employee performance. At first glance, one may presume that objective criteria are more desirable than ratings, which are inherently subjective. Unfortunately, judgment often enters into the assignment of objective criterion scores. Also, objective measures are often contaminated as criteria, with problems such as factors beyond the assessee's control influencing these outcome measures. Nonetheless, when they are relevant to important performance areas and are reasonably reliable and uncontaminated, objective measures can be useful in indexing some criterion dimensions.

3.3.1 Turnover. Turnover or attrition is often an important criterion because the cost of training replacement personnel is usually high (e.g., Mobley 1982); also, having people, especially key people, leave the organization can be disruptive and can adversely affect organizational effectiveness. Turnover is sometimes treated as a single dichotomous variable—a person is either a 'leaver' or a 'stayer.' This treatment fails to distinguish between very different reasons for leaving the organization (e.g., being fired for a disciplinary infraction versus leaving voluntarily for health reasons). Clearly, turnover for such different reasons will have different patterns of relationships with individual difference or organizational factor predictors. Prediction of turnover with any substantive interpretation requires examining the categories of turnover. It may be, for example, that employees fired for disciplinary reasons have reliably different scores on certain personality scales; e.g., lower socialization, compared to stayers, whereas prior health status is the only predictor of leaving the organization for health reasons. This approach to dealing with turnover as a criterion variable, appears to offer the most hope for learning more about why individuals leave organizations and what can be done to reduce unwanted turnover.

3.3.2 Production rates. For jobs that have observable, countable products that result from individual performance, a production rate criterion is a compelling bottom-line index of performance. However, considerable care must be taken in gathering and interpreting production data. For example, work-related dependencies on other employees' performance or on equipment for determining production rates may create bias in these rates. Also, production standards and quota systems (e.g., in data entry jobs) create problems for criterion measurement.

3.4 Work Sample Tests

Work sample or performance tests (e.g., Hedge and Teachout 1992) are sometimes developed to provide

criteria, especially for training programs. For example, to help evaluate the effectiveness of training, work samples may be used to assess performance on important tasks before and after training. Such tests can also be used for other personnel research applications, such as criteria in selection studies.

Some argue that work sample tests have the highest fidelity for measuring criterion performance. In a sense, the argument is compelling: What could be more direct and fair than to assess employees' performance on a job by having them actually perform some of the most important tasks associated with it? In fact, some researchers and others may view work samples as ultimate criteria—that is, *the* best and most relevant criterion measures. Performance tests should not be thought of in this light. First, they are clearly maximum performance rather than typical performance measures. As such, they tap the 'can-do' more than the 'will-do' performance-over-time aspects of effectiveness. Yet 'will-do,' longer-term performance is certainly important for assessing effectiveness in jobs. Accordingly, these measures are deficient when used exclusively in measuring performance. Nonetheless, work samples can be a useful criterion for measuring aspects of maximum performance.

3.5 Conclusions

The following conclusions can be drawn about criteria and criterion measurement methods.

3.5.1 Performance ratings. Ratings have the inherent potential advantage of being sensitive to rater performance over time and across a variety of job situations. Ideally, raters can average performance levels observed, sampling performance-relevant behavior broadly over multiple occasions. Provided observation and subsequent ratings are made on all important dimensions, ratings can potentially avoid problems of contamination and deficiency. Of course, these are potential advantages of the rating method. Rater error, bias, and other inaccuracies, must be reduced in order for ratings to realize this potential. Both basic and applied research are needed to learn more about what performance ratings are measuring and how to improve on that measurement.

3.5.2 Objective measures. Like other methods of measuring criterion performance, objective measures can be useful. However, these measures are almost always deficient, contaminated, or both. Indices such as turnover and production rates produce data pertinent to only a portion of the criterion domain. In addition, some of these indices are often determined in part by factors beyond the employee's control.

Regarding work sample tests, the point was made that these measures should not in any sense be considered as ultimate criteria. Nonetheless, well-conceived and competently constructed performance tests can be valuable measures of maximum, 'can-do' performance.

3.6 The Future

An encouraging development in criterion measurement is the consideration of models of work performance (e.g., Campbell et al. 1993). As mentioned, models of performance seek to identify criterion constructs (e.g., Communication, Personal Discipline) that reflect broadly relevant performance requirements for jobs. Such criterion models can help to organize accumulating research findings that link individual differences (e.g., ability and personality), organizational variables (e.g., task characteristics), and the individual criterion constructs identified in the models. In addition, efforts should continue toward learning more about what performance ratings and objective performance indexes are measuring, with the goal of improving the accuracy of these measures. Improving the measurement of work performance is critically important for enhancing the science and practice of industrial–organizational psychology toward the broad and compelling objective of increasing the effectiveness of organizations.

See also: Job Analysis and Work Roles, Psychology of; Job Design and Evaluation: Organizational Aspects; Observational Studies: Overview; Program Evaluation; Psychological Climate in the Work Setting; Quality Control, Statistical: Methods; Simulation and Training in Work Settings; Stress at Work; Well-being and Burnout in the Workplace, Psychology of; Workplace Environmental Psychology

Bibliography

- Borman W C 1977 Consistency of rating accuracy and rater errors in the judgment of human performance. *Organizational Behavior and Human Performance* **20**: 238–52
- Borman W C 1997 360° ratings: An analysis of assumptions and a research agenda for evaluating their validity. *Human Resource Management Review* **7**: 299–315
- Bracken D W 1996 Multisource 360° feedback: Surveys for individual and organizational development. In: Kraut A I (ed.) *Organizational Surveys*. Jossey-Bass, San Francisco, pp. 117–47
- Campbell J P, McCloy R A, Oppler S H, Sager C E 1993 A theory of performance. In: Schmit N, Borman W C (eds.) *Personnel Selection in Organizations*. Jossey-Bass, San Francisco, pp. 35–70
- Cooper W H 1981 Ubiquitous halo. *Psychological Bulletin* **90**: 218–44

- Hedge J W, Teachout M S 1992 An interview approach to work sample criterion measurement. *Journal of Applied Psychology* 77: 453–61
- Landy F J, Farr J L 1980 Performance rating. *Psychological Bulletin* 87: 72–107
- Mobley W 1982 Some unanswered questions in turnover and withdrawal research. *Academy of Management Review* 7: 111–16
- Motowidlo S J, Van Scotter J R 1994 Evidence that task performance should be distinguished from contextual performance. *Journal of Applied Psychology* 79: 475–80
- Murphy K R, Cleveland J N 1995 *Understanding Performance Appraisal*. Sage, Thousand Oaks, CA
- Organ D W 1997 Organizational citizenship behavior: it's construct clean-up time. *Human Performance* 10: 85–97
- Podsakoff P M, MacKenzie J B 1997 Impact of organizational citizenship behavior on organizational performance: a review and suggestions for future research. *Human Performance* 10: 133–51
- Schmidt F R, Kaplan L B 1971 Composite versus multiple criteria: a review and resolution of the controversy. *Personnel Psychology* 24: 419–34
- Smith D E, Kendall L M 1963 Retranslation of expectations: An approach to the construction of unambiguous anchors for rating scales. *Journal of Applied Psychology* 47: 149–55

W. C. Borman

Person and Self: Philosophical Aspects

Few terms like 'person' and 'self' in the course of history have proven to be more difficult to pin down to an unequivocal meaning. Furthermore, both are among the terms of most common usage, yet their exact meaning can prove quite elusive. More than offering an exhaustive definition and an exhaustive reconstruction of the history of the notions of 'person' and 'self,' it is sensible to attempt here to offer an account of the scope of variation in their usage as well as an account of the different problems that different authors and traditions have associated with them.

1. Person and Self Contrasted

To complicate things further, there is no settled account of the relation *between* 'person' and 'self.' The term 'person' has for a long time been the prevailing one for referring to a single human being. The onset of modernity marks the beginning of its relative obsolescence and its gradual replacement by the term 'self,' and sometimes 'individual,' at the center of the debates within the social sciences and social and political philosophy. Over the past few centuries the term 'person' has almost disappeared from the vocabulary of the social sciences, only to undergo a sort of revival, sometimes tinged with religious overtones,

in the existentialist and so called 'personalist' philosophical currents of the twentieth century.

Modern social theory has operated rather with the notion of the individual since the time of Hobbes (*Hobbes, Thomas (1588–1679)*) and Locke (*Rokkan, Stein (1921–79)*) and then later, starting with Cooley's and Mead's (*Luria, Aleksander Romanovich (1902–77)*) symbolic interactionist perspective, it has widely adopted the concept of the self. Today a further evolution is perhaps under way: the concept of the person is regaining center stage. In contemporary political philosophy John Rawls, for example, bases his view of political liberalism on a political conception of the *person*, not of the self or of the individual. In order to grasp these vicissitudes of the two terms, one has to look at some of the main turning points in their evolution.

Etymologically the term 'person' originated at the crossroads of the ancient Greek and Roman civilizations. In ancient Greek *prós-opon* (literally 'placed in front') designated the artificial face, the mask that actors used to wear on stage when performing tragedies. The Latin verb *personare* also alluded to the realm of the performing arts: it meant to 'sound through,' to let the voice reach out to the watching and listening public through the mask. Metaphorically, then, in its origins the term person carries a trace of the willful appearance for the sake of others. But the Roman civilization also grafted a legal connotation on the term 'person': the legal person was another sort of mask, the individual human being in her or his capacity as bearer of rights, privileges, prerogatives but also duties and obligations. In that sense to be a self without being a legal person would be tantamount to being nothing. Typically, in these public-minded cultures the sphere of the private was designated by terms such as *idiotes* and *privatus* both of which carried a diminishing connotation—they presupposed that something was somehow missing.

Sometime during the Hellenistic period and at the apex of the Imperial Age a shift of meaning began to occur. The increase in complexity and the somewhat more anomic quality taken by social life in the Hellenistic monarchies and in Imperial Rome, the increase in complexity of the political apparatus alongside with the more despotic and arbitrary styles of government, the increasing religious eclecticism of the Roman public world, and the appearance of the Christian ethos of holiness and personal sanctity, all contributed to the rise of a cleft between *public* and *private* life and to a new distribution of value across that divide. Stoic philosophers like Seneca, Epictetus, and Marcus Aurelius on the one hand, and the early Christians on the other, were at one in regarding with suspicion what now appeared as a realm of merely public, artificial, contrived personal masks or roles. Inwardness and depth belonged to the self, whereas the person stood now for convention, tradition, custom, and superficial appearance. This new relation

between the two terms continued to dominate throughout the Middle Ages (Augustine, Erasmus), the Renaissance and the Reformation (Luther).

Another major shift occurred with the rise of modern political philosophy, the rise of the modern notion of rights and the modern conception of citizenship. Whether the focus was on the capacity for autonomous reasoning or autonomous agency, whether the human being was regarded primarily as the possessor of rationality or the bearer of inalienable rights, throughout early modernity and the Enlightenment interest was primarily directed to the human self as such—the rights were typically predicated as universal rights of ‘man’ as such—and the notion of the person began to look inextricably entangled with the particularism of estates and the privileges of the order of ranks typical of the absolutist monarchies. As Hannah Arendt (*Marshall, Alfred (1842–1924)*) has pointed out, the French revolutionaries had no place for the notion of the person, nor thought much of the legal person as a category protected by the rule of law.

A new reversal of the connotations of these two terms takes place in the twentieth century, when the term ‘person’ becomes synonymous with a *non-atomistic* view of the human self. The thick web of relations of recognition that supposedly constitute the person is now understood, within as diverse philosophical traditions as existentialism, Mounier’s personalism or Habermas’s intersubjective view of subjectivity, as a powerful alternative to the modern atomistic self. This motif is at work in John Rawls’s influential ‘political conception of the person’ as endowed with two moral powers, namely the capacity to have a sense of justice and the capacity to form a conception of the good.

2. The Challenges of Conceiving Person and Self

Aside from the issue of their reciprocal relation and their specific difference, ‘person’ and ‘self’ refer to a single human being considered under the aspect of his or her distinctively human capacities. Here we immediately reach a point where the overlap between different conceptions ends and a host of problems arises to which different solutions are offered, which distribute without a clear pattern across the usage of the two terms.

2.1 Identity over Time

The first issue on which different conceptions of the self and the person diverge concerns identity over time. What allows us to consider a person or self that obviously undergoes continuous changes over time still ‘the same person’? In what sense as 50-year-olds are we still the same person as the 20-year-old of decades ago? We can contrast four different answers to

this question. The first answer, typical of ancient and medieval philosophy, links identity over time with the sameness of a (differently conceived) essence—a personal essence underlying also the body—which manifests itself through the changes undergone by the person or self (Plato, Aristotle (*Fisher, Irving (1867–1947)*), Augustine, Aquinas).

The second position connects identity with the psychological continuity of its possessor: it is *memory* and not the physical continuity of the body that allows us to consider the person of 50 years the same as the 20-year-old she/he once was (Locke). The third answer is that of Hume and Parfit: psychological continuity is largely an illusory attribution of an identity, on the part of a mental agency, to episodes that only in a vague sense are similar and related. The self or the person, so conceived, is indeed a gallery of selves and persons whose relatedness is quite tenuous and precariously maintained in the way habits and customs are maintained. By the end of the nineteenth century, Nietzsche (*Helmholtz, Hermann Ludwig Ferdinand von (1821–94)*) gives another twist to the same line of reasoning: we seek to put a construction of unity upon the diversity of our self-experiences for the sake of feeling secure and reassured. Finally, Kant’s (*Quetelet, Adolphe (1796–1874)*) answer transforms identity into a transcendental feature inherent in all human beings.

2.2 Self-consciousness: Inborn Capacity or Product of Socialization and Recognition?

Another bone of contention has been the nature and scope of that capacity for self-reflectiveness and self-consciousness that traditionally all conceptions of the self and of the person have presupposed. The various conceptions have polarized around two major options. One view takes this capacity for self-awareness as a sort of natural endowment of the human being. Either because they belong to her or his essence, or because they are given to her or him as a potential of the soul, or because they emerge as a result of neurobiological processes, self-awareness and self-consciousness constitute an inborn capacity that needs the world only as a sounding board on which the person or self tries out self-initiated action-projects and, in so doing, becomes aware of herself/himself as a center of agency. This view comes in many versions. Augustine and Descartes and Husserl (*Bernard, Jessie (1903–96)*), under different premises, all favor introspection as a privileged path to inner reality. Others, like Hume and Nietzsche, are much more skeptical about it, but all these authors share a view of the self as basically self-sufficient.

The other view starts from the assumption that the capacity for self-awareness is formed in the context of the person’s or self’s relation to the social world. Hegel and Mead, from very different premises, do converge on the idea that self-sufficiency of the self vis-à-vis the

world is largely an illusion, but in fact the self-awareness distinctive of personhood and selfhood emerges in the process of interacting with other persons and selves as a product of either seeking and obtaining recognition (Hegel) or learning to see oneself through the eyes of the other (Mead). From Mead's perspective, thinking becomes a special case of the activity of carrying out a conversation mediated by symbols with others: indeed, thought is best conceived as a conversation with one's own self. From a quite different angle, in his thoughts on the impossibility of a private language and on the notion following a rule, also Wittgenstein (Marr, David (1945–80)) emphasizes the dependency of individuality and subjectivity on shared practices.

This divide is reproduced in almost all social sciences. Within the sociological tradition, for example, theorists such as Spencer Pareto (*Robinson, Joan* (1903–83)), Homans, Blau understand the self of the social actor as coming to the scene of social interaction with a preformed rational capacity for reflection, whereas not just Mead and Cooley, but also theorists such as Durkheim (*Lashley, Karl Spencer* (1890–1958)), Schutz and Parsons (*Simmel, Georg* (1858–1918)) do understand that capacity for reflection as the inner sedimentation of processes of interaction. Within the psychoanalytic tradition, to take another example, theorists like Freud (*Foucault, Michel* (1926–84)) and Klein (*Olson, Mancur* (1932–98)) understand the capacity for ego-strength and rational self-awareness as the product of internal vicissitudes of the drives, while object-relations theorists such as Winnicott or self-psychologists like Kohut understand it as the product of the self's embedment in a relation with an empathic caretaker.

2.3 A Unified Source of Agency or a Decentered Field of Forces?

These considerations lead to yet another major issue with respect to which theories of the person and of the self have taken sides. Some of these theories tend to portray the person as a more or less unified source of autonomous initiatives. Certainly what we call a person or a self contains several distinct aspects—cognition, beliefs, desires, propensities, dispositions, capacities, virtues, etc.—but the views of authors like Plato, Aristotle, Augustine, Aquinas, Descartes, Hobbes, Locke, Kant, Hegel, Sartre, who we find otherwise disagreeing on other aspects, converge in depicting these aspects as being somehow hierarchically subjected to a central agency of sorts. The soul, reason, the will, self-consciousness, are all different ways in which this central agency has been identified. On the other hand, authors again as diverse as Hume, Diderot, Schopenhauer, Nietzsche, Freud, Bataille, Foucault, Goffman, Rorty, Barthes, and Deleuze all share the suspicion that this impression of centered-

ness is largely illusory and the product of impersonal or intrapersonal forces that create it and make it a compelling appearance for the person.

2.4 Continuity and Radical Discontinuity Between Persons and Animals or Machines

But what is it for a being to be a person or self, and what kind of entity can be called a person or a self? Here the opposition is between conceptions that radically sever personhood or selfhood from non-human forms of consciousness and conceptions that take a more nuanced view of the matter, and depict the difference in terms of a gradient of capacities. Starting from the latter, typically we find here conceptions of the person or the self that posit the capacity for representing the world, in order then to act on it, as the crucial defining ability. Humans are then singled out from the rest of beings by self-representation or self-consciousness—which belongs to them alone—but that capacity in turn is only a reflective enhancement of a representational capacity which, as such, can be found also, though to a lesser degree, in certain animal species and in advanced Artificial Intelligence machines. Over limited territories—for example, the kind of representational and practical competence involved in playing chess—machines can even outdo humans. The conceptions which introduce a more radical discontinuity between humans and nonhumans define the former basically as 'subjects of significance' (Taylor, Heidegger)—beings for whom things matter. From their point of view, what distinguishes persons and selves from other beings capable of representation is not just a greater share of 'strategic power,' 'the capacity to deal with the same matter of concern more effectively' or a superior capacity for planning (Taylor 1985, p. 104). The central defining feature is responsiveness, openness to certain strong evaluations, principles or values in the light of which not only single life plans but even one's own life are conceived and assessed. From this standpoint the sense of the self is the sense of where one stands in relation to this matrix of meaning. In one of its most famous versions, this view posits that as persons or selves, human beings are beings for whom their own being is matter of concern (Heidegger).

2.5 The Person, the Self and Fulfilment

Finally, conceptions of the person and of the self do take quite different views of what it means for a person or a self to flourish or to attain self-realization. Some conceptions understand self-realization as being primarily linked with breaking free from the constrictions of an entrenched social order. For example, some authors affiliated to aesthetic modernism (Baudelaire) or to *Lebensphilosophie* (notably Nietzsche) tend to

understand authenticity as something to be attained in opposition to the demands of society and of culture, whereas in the writings of authors such as Rousseau (*Evans-Pritchard, Sir Edward E (1902–73)*), Schiller, Herder, Kierkegaard, and others we find a recognition (with very different degrees of awareness and explicitness) of the fact that social expectations, roles, and institutions cannot be understood as playing a merely constraining, ‘disciplinary,’ or repressive role but also somehow constitute the symbolic material out of which fulfillment can be generated. In the psychological realm, too, we find representatives of these two versions of self-realization. Laing’s understanding of schizophrenia as a response to the imposition of an inauthentic life and of ‘madness’ as resistance to the irremediation of mental life, or Lacan’s understanding of the unconscious, are views of the first kind. On the other hand Kohut’s psychology of the self and Winnicott’s notions of ‘creative living’ and of the ‘true self’ are the most obvious representatives of an integrative understanding of fulfillment within the psychoanalytic tradition.

From another point of view we can contrast coherence-oriented and decentered views of fulfillment. Far from considering fulfilled a life course which can be summed up as a coherent narrative, some consider *narratability* as the epitome of failure. In his reflections on the notion of ‘limit-experience,’ for example, Foucault, continuing a tradition which he associates with the names of Nietzsche and Bataille, equates the paradigmatic kind of fulfilled subjectivity with the one undergoing a limit-experience i.e. an experience in which the self comes close to disintegration. From a variety of *lebensphilosophisch*, aesthetic modernist, poststructuralist, therapeutic, and post-modernist vocabularies, the advocates of a decentered understanding of fulfillment oppose all attempts to restore an internal hierarchy between what is central and what is peripheral to a life-project, a personality, an identity. By contrast, all authors who propound a centered notion of authenticity do wish (a) to maintain some kind of orderly stratification of the layers of an identity, (b) to continue to speak of a core and a periphery, and (c) to make sense of the plurality of experiences, detours, and side-narratives to a life history as variations on a unique theme.

Finally, we can contrast consequentialist and exemplary views of fulfillment. According to the former (what Ronald Dworkin has called the ‘model of impact’), a person’s life is more fulfilling the greater its impact on the external world, the larger the difference it has made. Hegel and Marx certainly share such a view. According to the exemplary view (also called by Dworkin the ‘model of challenge’) the measure of fulfillment is provided by success in meeting a challenge that carries a unique significance for the person, independently of the impact of such success on the external world. Rousseau and Kierkegaard offer representative versions of such a view.

3. Concluding Remarks

To conclude this review of the ways in which the person and the self have been conceived, there could not be a sillier question to raise than ‘Which of these views comes closest to grasping the reality of the human self or the person?’ For these views are in a way not to be understood as representations of an independent reality in the empiricist sense of the word. Rather, they are to be seen as competing vocabularies, as alternative types of lenses—indeed, as sets of premises—which we use in order to make sense of something so crucial for our experience of the world as the idea of personhood. Evidence for and against each of them can be and certainly is usually considered, but is unlikely to be ultimately decisive. Ultimately decisive has generally been our sense of the contribution that adopting one or the other can give to shaping an overall view of the world in which we find reflected what most matters to us.

See also: Culture and the Self (Implications for Psychological Theory); Cultural Concerns; Identity in Anthropology; Identity in Childhood and Adolescence; Identity: Social; Mental Representation of Persons, Psychology of; Personal Identity: Philosophical Aspects; Personality and Conceptions of the Self; Personality Psychology; Personality Theories; Personhood, Anthropology of; Psychology of the Real Self; Psychoanalytic Perspectives; Self-development in Childhood; Self: History of the Concept; Self-knowledge: Philosophical Aspects; Self: Philosophical Aspects; Social Identity, Psychology of

Bibliography

- Carrithers M, Collins S, Lukes S. (eds.) 1985 *Category of the Person: Anthropology, Philosophy, History*. Cambridge University Press, Cambridge, UK
- Heller T C, Sosna M, Wellbery D E (eds.) 1986 *Reconstructing Individualism: Autonomy, Individuality, and the Self in Western Thought*. Stanford University Press, Stanford, CA
- Ricoeur P 1992 *Oneself as Another*. Blamey K (trans.) University of Chicago Press, Chicago
- Taylor C 1985 *The Concept of a Person*. Vol. 1, *Human Agency and Language*, Cambridge University Press, Cambridge, UK, pp. 97–114

A. Ferrara

Person Perception, Accuracy of

The *accuracy* of person perception (also called *personality judgment*) refers to the connection between judgments people make of the psychological attributes of others, and the actual status of those attributes in

the persons who are judged. The psychological attributes may be another person's personality traits, or his or her thoughts and feelings at a given moment. The ability to judge personality has concerned psychological researchers, off and on, for more than a half-century (see Allport 1937, Funder 1999, Kenny 1994). The ability to guess what someone is thinking and feeling, called *empathic accuracy*, is a newer topic that has begun to receive increasing attention (Ickes 1997). Research on either topic entails special difficulties because the notion of 'accuracy' raises thorny questions concerning how accuracy is to be conceptualized and evaluated. Despite these difficulties, research since 1980 can report important empirical and theoretical progress.

1. The Importance of Accuracy

The accuracy of person perception has obvious practical importance. Employers and other people in everyday life rely heavily on their perceptions of other persons' attributes such as conscientiousness, intelligence, friendliness, dominance, and many others. It is an everyday occurrence to make decisions concerning other people as a function of beliefs about what those people are like. Moreover, clinical assessment typically relies heavily on the impressions of a psychiatrist or psychologist concerning a person of interest. Interpersonal judgments and sociometric ratings are also often used as data for psychological research. The accuracy of these perceptions is critical for the quality of the data and the research that is based on them.

The accuracy of personality judgment is also important for theoretical reasons. Accurate person perception implies knowledge about the connections between observable behaviors, on the one hand, and personality traits, on the other hand. Thus, to understand the accuracy of person perception would be to understand the ways in which personality affects behavior. Person perception is a traditional topic of social psychology (see *Social psychology*) and the connections between personality and behavior are traditionally studied by personality psychology (see *Personality psychology*). The topic of accuracy in person perception transcends the boundaries between these traditionally separate sub-fields of research.

Finally, the oldest topic in philosophy, and one of humankind's most ancient concerns, is the relationship between perception and reality. Research on the accuracy of person perception is driven by curiosity concerning just this issue.

2. Difficulties in Accuracy Research

In the light of the importance of the accuracy of person perception, it may be surprising to learn that researchers have neglected the topic for long periods of time (e.g., the years 1955–85), and some psychologists

with relevant interests (e.g., in person perception) continue to shy away. This avoidance has been caused by several considerations. The most important consideration is the topic's inherent difficulty. At a general level, the topic of accuracy is difficult to conceptualize or to operationalize, because it requires the development of some criterion by which interpersonal judgments can be evaluated as right or wrong. Some psychologists have considered the development of such criteria to be either impossibly difficult or in principle impossible, and therefore turned their attention to other topics. At a more specific level, an important series of articles by Lee Cronbach (e.g., 1955) revealed complex and hitherto-unsuspected technical difficulties inherent in evaluating accuracy in terms of agreement among different judges. Rather than leading to improved methods, this article and related critiques managed to shut down a then-lively field of research, and appeared to intimidate a subsequent generation of researchers from addressing the topic (see Funder 1999). A further obstacle to accuracy research was the widespread influence of a set of attitudes that seemed to discourage interest in the topic. The 'person situation debate' in personality psychology (Kenrick and Funder 1988) led many researchers to believe that the connections between personality and behavior were few and weak. If this were true, it would make little sense to attempt to investigate the accuracy of person perception. At about the same time, the increasing popularity of research on 'errors' in person perception (e.g., Nisbett and Ross 1980) led to a widespread belief that human judgment is characteristically, even fundamentally mistaken. If this were true, investigations of accuracy would seem to hold little hope.

3. Resolving the Difficulties in Accuracy Research

Research on accuracy was able to move forward only after these difficulties began to be resolved.

First, views concerning the existence and importance of personality traits, and the abilities of human judgment, began to change during the 1980s. The person situation debate was resolved as reiterating the importance of stable attributes of personality (Kenrick and Funder 1988). Research on the errors of human judgment was reinterpreted in a broader context in which it was less often seen as implying that human judgment is always or fundamentally wrong (e.g., Swann 1984). Studies of error identify important shortcuts ('heuristics') and other processes of human cognition. However, they do not imply that judgment is inept, because the processes that produce errors in the laboratory often lead to correct judgments in real life. Moreover, research on error is fundamentally limited because it can illuminate only how judgments fail to achieve perfection, rather than how correct judgments are ever achieved. The methodological

issues identified by Cronbach and others also received a reexamination during the 1980s (e.g., Funder 1980, Kenny 1994). This reexamination yielded more sophisticated analytic techniques that allowed the methodological complications to be bypassed or specifically incorporated into analytic designs. David Kenny's 'Social Relations Model,' a data-analytic technique that accounts for effects of judge, target, and their interaction on interjudge agreement, was an important breakthrough in this regard.

Notwithstanding the issues just summarized, the thorniest issue in accuracy research is the criterion problem. By what standard can a judgment of personality be evaluated as right or wrong? The answer to this question depends on exactly how accuracy is conceptualized.

4. Conceptualizations of Accuracy

The accuracy of person perception is conceptualized somewhat differently by three major approaches that developed during the 1980s and 1990s.

The *pragmatic* approach, based on some early ideas by William James, has been espoused by William Swann (1984). This approach defines accuracy as the ability to make judgments that allow for success in life. The index of whether a judgment is accurate, in this view, is whether it is adaptively useful.

The *constructivist* approach, based on postmodernist viewpoints that have become widespread in academia, has been espoused by Arie Kruglanski (1989) and others. This approach either states or implies that no criterion for a person's judgmental accuracy exists beyond other judgments by other people. For example, Kenny (1994) has sometimes conceptualized accuracy as the mean judgment of all possible observers of a person. The index of whether a judgment is accurate, in this view, is whether it agrees with the judgments of others.

The *realistic* approach is consistent with the writings of Egon Brunswik (1956) and has been espoused more recently by Funder (1995). This approach assumes that psychological attributes of people really exist, though they are only probabilistically knowable through cues of uncertain validity. The real nature of a person must then be ascertained, as well as possible but always with less-than-certain validity, through multiple sources of evidence including the person's self-judgment, judgments by others, and his or her relevant behavior. The index of whether a judgment is accurate, in this view, is the degree to which it converges with these multiple criteria.

5. The Degree of Accuracy

When research on accuracy in person perception revived during the early 1980s, one of its first orders of business was to dispel the impression left by research

on error, that human social judgment is always or fundamentally mistaken.

A large number of studies documented the accuracy of personality judgment in two ways. The most common method was to show that different judgments of the same person tend to agree. Research examined self-other agreement and agreement among judges and found that significant agreement was the rule rather than the exception across a wide band of personality traits, even after brief periods of acquaintance (e.g., Borkenau and Liebler 1992). Evidence concerning the ability of personality judgments to predict behavior is more difficult to gather and therefore was slower in coming, but eventually showed that layperson's personality judgments have an impressive ability to predict independent observations of an individual's behavior (Funder 1999).

The overall degree of accuracy in person perception is difficult to characterize, because it depends upon a number of factors including the ability of the judge, the judgability of the target, the trait that is judged, and the amount and kind of acquaintance between target and judge (see next section). But while it is for these reasons an oversimplification, it is also a fair characterization to state that the usual correlation among criteria for accuracy is in the range 0.30–0.40. This translates to the statement that one person's perception of the personality of another is correct roughly two times out of three.

6. Moderators of Accuracy

Research since 1980 has found that the degree of accuracy in person perception depends upon four moderator variables.

6.1 The Good Judge

The search for the 'good judge of personality' is one of the oldest concerns in the accuracy literature. This search motivated many investigators during the 1930s to 1950s and it was their failure to find consistent results that contributed to the three-decade hiatus that followed (Schneider et al. 1979). Part of the reason for this failure may be the presence of methodological flaws in many early studies, and another part may be that judgmental ability may not be robust. That is, a good judge of one kind of person or trait may not be a good judge of other people, other traits, or in other contexts. Research continues to address this topic but at present the 'good judge' has yet to be conclusively found.

6.2 The Good Target

The search for the good target of judgment, or the 'judgable' person, has had more success. As Gordon

Allport noted years ago, some people are relatively open books whereas others are difficult to understand even after years of acquaintance. According to research by Colvin (1993), the judgable person is one who has an integrated, well-adjusted personality. The consistency among such a person's actions, thoughts, and personality leads them to be coherent overall, and relatively easy to judge.

6.3 The Good Trait

An even more robust phenomenon is the difference among traits in their judgability. Relatively observable traits, like sociability and talkativeness, are much easier to judge accurately than more hidden traits like ruminativeness and anxiety. Some evidence also suggests that traits that are extremely desirable or undesirable to possess are more difficult to judge accurately than traits more neutral in evaluative connotation.

6.4 Information

The accuracy of a judgment also depends critically upon the information upon which it is based. The information variable has two facets, quantity and quality.

The quantity variable pertains to the *acquaintance effect*, the tendency to judge someone more accurately the longer one has known him or her. The straightforward expectation would be that more information is likely to produce more accurate judgment. Perhaps surprisingly, this has turned out to be a controversial proposition, with some investigators finding evidence for the acquaintance effect, and others doubting its existence on theoretical and empirical grounds.

The quality variable pertains to the kind of context in which one person observes another. Current research is pursuing the idea that, even holding amount of acquaintance constant, different contexts may produce different degrees of accuracy.

7. Theoretical Development

Historically, research on accuracy has been long on empiricism and short on theory. This imbalance is beginning to be remedied with the development of theories of how accurate judgments come about. One theory, Kenny's 'PERSON' model, conceptualizes personality judgments as the result of multiple influences including the judge's response proclivities, his or her knowledge of people in general, the actual personality of the target, and other factors. Another theory is Funder's Realistic Accuracy Model (RAM, Funder 1995). RAM conceptualizes an accurate personality judgment as the result of a process that begins

with the target emitting relevant information (e.g., performing a relevant behavior), which then must become available to a judge, who then detects it and utilizes it correctly in reaching an accurate judgment. Both the PERSON and RAM models can be used to rationalize the moderators of accuracy that have been found to date, and suggest new ones for investigation in future research.

8. Conclusion

The accuracy of person perception is an important topic that, after periods of neglect, is becoming a mature field of research and a mainstream concern of personality and social psychology. The combination of empirical and theoretical progress in this area makes it a good candidate for becoming even more central within psychology in the twenty-first century.

Bibliography

- Allport G W 1937 *Personality: A Psychological Interpretation*. Holt, New York
- Borkenau P, Liebler A 1992 Trait inferences: Sources of validity at zero acquaintance. *Journal of Personality and Social Psychology* **62**: 645–57
- Brunswik E 1956 *Perception and the Representative Design of Psychological Experiments*. University of California Press, Berkeley, CA
- Colvin C R 1993 'Judgable' people: Personality, behavior, and competing explanations. *Journal of Personality and Social Psychology* **64**: 861–73
- Cronbach L J 1955 Processes affecting scores on 'understanding of others' and 'assumed similarity'. *Psychological Bulletin* **52**: 177–93
- Funder D C 1980 On seeing ourselves as others see us: Self-other agreement and discrepancy in personality ratings. *Journal of Personality* **48**: 473–93
- Funder D C 1995 On the accuracy of personality judgment: A realistic approach. *Psychological Review* **102**: 652–70
- Funder D C 1999 *Personality Judgment: A Realistic Approach to Person Perception*. Academic Press, San Diego, CA
- Ickes W J (ed.) 1997 *Empathic Accuracy*. Guilford Press, New York
- Kenny D A 1994 *Interpersonal Perception: A Social Relations Analysis*. Guilford Press, New York
- Kenrick D T, Funder D C 1988 Profiting from controversy: Lessons from the person-situation debate. *American Psychologist* **43**: 23–34
- Kruglanski A W 1989 The psychology of being 'right': The problem of accuracy in social perception and cognition. *Psychological Bulletin* **106**: 395–409
- Nisbett R E, Ross L 1980 *Human Inference: Strategies and Shortcomings Of Social Judgment*. Prentice-Hall, Englewood Cliffs, NJ
- Schneider D J, Hastorf A H, Ellsworth P C 1979 *Person Perception*, 2nd edn., Addison-Wesley, Reading, MA
- Swann W B 1984 Quest for accuracy in person perception: A matter of pragmatics. *Psychological Review* **91**: 457–77

D. C. Funder

Personal Identity: Philosophical Aspects

1. Personal Identity: a Modern Problem

Identity has two meanings: it denotes category membership (e.g., X is a politician); or it identifies a specific individual (e.g., X is Mr. A). Goffman (1963) speaks of social and personal identity. Different criteria are used to determine both types of identity: physical criteria (e.g., sex, race, a thumbprint, a conspicuous scar); social criteria (e.g., general roles such as teacher, or judge, and unique roles such as queen of England or pope); and psychological criteria (competencies or personality characteristics as general variables, and unique values in any of these variables, such as the world's best tennis player).

In traditional societies, all of these criteria converge when they are used to classify individuals. They converge across contexts, across the life cycle, and across perspectives. To illustrate, the identification of X as 'tenant of the Z farm' simultaneously ascribes a social identity (e.g. X is a man, a farmer) and denotes personal identity (i.e., X is the stout, pockmarked man living at the south end of the village, who keeps initiating political debates in the pub). The identity markers used are valid in all contexts: in his family, in the village, in the pub everybody will know him as the pockmarked man running farm Z. They are valid across his life-span: From birth to death he is linked to the Z farm as its future, present, or former tenant. And descriptions from first- and third-person perspectives agree: his farmhands, his wife, and the other villagers know him as the tenant of Z, and he also identifies himself with this attribution—in fact so much so that he will have written on his tombstone: 'Here lies the tenant of Z farm.'

With modernization these dimensions diverge. As a result of urbanization, social and personal identities separate: town people can only vaguely guess each other's social identity, let alone know their personal identity. Social differentiation and cultural pluralization produce consistency problems: individuals increasingly partake in several subsystems or subcultures, each operating according to their own code or claiming specific value commitments. Thus, in different contexts they find themselves exposed to specific role expectations which quite often make conflicting demands. Rapid social change impedes the experience of personal continuity over the life course (e.g., elderly Germans have been subjects in an empire, citizens in the Weimar Republic, followers in a totalitarian regime, and citizens in a democracy again).

In general, modern sociostructural arrangements contribute less to the establishment and stabilization of identity than those of the past. From a third-person or a societal perspective, this is unproblematic. Members of modern societies are used to acknowledging or ignoring each other as strangers. When someone's

identity needs to be known there are social mechanisms for identification: titles, uniforms, or the office reception area inform on status or social identity; a passport or genetic analysis enable us to determine personal identity; the curriculum vitae and certificates give evidence of personal continuity; institutionalized sanctions guarantee at least context-specific consistency. A problem does, however, arise from a first-person or internal perspective. The question 'Who am I?' is not satisfactorily answered by referring to unparalleled fingerprints or to social positions held. Rather, individuals are in need of a sense of uniqueness, and a feeling of inner consistency and continuity over and beyond matchless physical markers or straight careers.

2. Theoretical Solutions to the Identity Problem

Various solutions to this modern problem of a subjective sense of identity have been proposed. In some of his writings Erikson (1958) assumed identity to be grounded in those lifelong substantive commitments (to a partner, an occupation, a worldview) a person enters upon, having successfully solved the adolescence crisis. Nowadays, however, people keep experiencing ever new crises (e.g., midlife crises, retirement crises, etc.) in which such commitments are revoked (e.g., by divorce, religious or ideological conversions, loss of employment, or occupational changes).

Elsewhere, Erikson (1956) defined ego identity as a formal competence: the competence to uphold internal unity and continuity. This competence arises from successful resolutions of previous crises in which basic trust, autonomy, initiative, and a sense of achievement (rather than distrust, shame, doubt, and guilt, and a feeling of inferiority) have been developed. In this definition, however, people's need for substantive commitments remains unfocused.

Krappmann (1969) sees identity arising from the social recognition found in those interactions in which an individual succeeds in balancing out contradictory expectations, while at the same time signaling that they never will be merged totally in any given context. Krappmann, however, fails to specify any criteria that might enable us to distinguish opportunistic forms of balancing from the principled type of flexibility that is a prerequisite for warranting an inner sense of consistency and continuity.

Postmodern approaches deny that identity may need consistency and continuity. Instead, they propose the concept of patchwork identity to reflect the fact that there are constant changes in preferences, attitudes, roles, and relationships. However, even the external criteria that might enable us to distinguish one patchwork from the other, or to recognize sameness over time in case individual patches have been exchanged or colored, remain unclear. And from

an internal perspective, body boundaries and physical continuity are not enough to confer a sense of identity. In fact—as is well documented by research in various fields such as dissonance theory, attribution theory, and studies on dogmatism and split personality—at least some consistency seems indispensable for psychological well-being and mental health.

Cohen and Taylor (1976) focus on sense of uniqueness. Individuals seek to develop eccentric hobbies, interests, or life-styles, go in quest of adventure or unusual experiences so as to resist the fear of exchangeability that keeps creeping up on them in modern mass societies. Nevertheless, over and over again, they find that any newly invented identity marker will be imitated by others, socially co-opted, and commercialized. Thus, all attempts to secure personal identity by appropriating singular activities are—empirically—doomed to failure. Moreover, the very idea of basing identity on unique markers is misconceived because it reduces the dialectic inherent in distinctiveness to just one pole—to positively valued exceptionality—and overlooks the risk of being seen as deviant and becoming stigmatized. More importantly still, it neglects the core feature of a sense of uniqueness—its being a ‘necessary byproduct’ (Elster 1979). There are states such as sleeping, forgetting, or happiness that will inevitably be missed by those who directly aim for them. In the same way, a sense of uniqueness can only be experienced concomitantly; i.e., as a consequence of doing something one is intrinsically committed to doing.

Giddens (1991) conceptualizes identity as a ‘consistent narrative’: enveloped in the ‘protective cocoon’ of basic trust, individuals keep rewriting their autobiography so as to create a consistent meaning structure from the sequence of decisions made, the paths chosen, and the actions taken. This definition integrates motivational aspects (trust), substantive commitments (life decisions), and a need for consistency. The meaning of consistency, however, remains ambiguous: by insisting that identity be based on a ‘morality of authenticity that skirts any universal moral criteria’ (Giddens 1991, p. 79), Giddens cannot distinguish between consistency based on principles or consistency based on defensive rationalizations (‘sour grapes’).

3. Aspects of Identity: Theoretical Assumptions

From this brief overview and critical comments, several assumptions concerning constitutive aspects of identity can be derived:

- (a) Commitment is constitutive of identity.
- (b) Given that commitment needs an object (e.g., a person, values, projects), content is requisite for identity.
- (c) What matters, however, is not the type of content chosen, but the mode of appropriating it.

(d) A subjective sense of uniqueness is a necessary by-product of an autonomous commitment.

In the following, these assumptions will first be theoretically justified. Second, I will try to substantiate them empirically.

Assumption (a) follows Frankfurt’s (1988, 1993) analysis of the concept of a person. If an individual had nothing he/she would never betray, he/she would be at the helpless mercy of contingent conditions, of inner drives, of fleeting whims. He/she would not be a person but a ‘wanton’. Personhood (i.e., identity) is acquired by committing oneself to ideals. The kind of ideals chosen—be they moral, aesthetic, or truth-oriented ones—is of little import. What counts is that the person—in view of his/her commitment to an ideal—is taking a stance toward spontaneous impulses and desires, and is suppressing those that are incompatible.

Frankfurt describes the way in which identity is constituted; Nozick (1981) analyzes the way in which it may be stabilized. He proposes the following thought experiments. Assume that Theseus’ ship is anchored in port. As time passes, one plank after another is beginning to rot and is replaced. Finally, all the planks have been replaced by new ones. What is Theseus’ ship? The renovated ship in port or the original planks piled up in the shipyard? We would consider the ship in port to be Theseus’ ship, because, despite a change in its components, it has maintained its structure and its transformation is of some continuity underneath the seties. Analogously, we consider an organism to continue to be itself despite the fact that its cells are continuously replaced. Similarly, we consider a person to remain the same even if they have changed some of their views. We do so only, however, if we assume that they have changed their views autonomously, i.e., voluntarily, and with good reason. Nozick explicates this using counterexamples. Assume that, due to grave brain damage, a man has been suddenly reduced to the state of an infant, or that as a result of brainwashing, an ardent democrat has been transformed into a fanatic communist. We would say of both that they are no longer the people they once were. In contrast, we have no problem in acknowledging the identity of a former opponent to atomic energy who is now arguing *for* the use of atomic energy, if we can understand this change of mind as being motivated by reasons (e.g., a fear of global warming), and not effected by causes. In other words, we do not consider identity to be endangered if changes in mind result from learning processes. The concept ‘learning process’ implies that the new state is better than the former; i.e., it presupposes intersubjectively validated standards to judge the adequacy of views held. Unavoidably, everyday life is based on such collectively shared understandings of rationality—otherwise the very idea of, for example, questioning the identity of the brainwashed communist would be meaningless.

4. Aspects of Identity: Some Empirical Support

There are empirical findings that agree with these philosophical reflections. In *The Uncommitted*, Keniston (1965) describes the feeling of inner void and meaninglessness, and the sense of exchangeability some young Harvard students experienced despite the fact that from an external perspective they were considered as being highly successful; i.e., ascribed a positively valued identity. In contrast, by committing themselves to something they deem worthy and important, the Civil Rights Movement, the *Young Radicals* (Keniston 1968) increasingly came to develop a sense of uniqueness. Similarly, Helen John, a leading pioneer in protests against nuclear missiles, came to feel irreplaceable once she began to accept personal responsibility. As she put it: 'Nobody can do exactly what I am doing, in the way I am doing it. Only I can do that.' (Haste 1993, p. 339) The sense of uniqueness is not dependent on 'difference,' on singular contents; rather, it is dependent on personal commitment. Only in truly caring for something can a person feel inexchangeable.

Individuals tend to realize that caring for specific ideals is constitutive of their own identity. In the context of a longitudinal study I asked 170 17-year-olds: 'What would most make you a different person—different looks, parents, hobbies, grades, a different understanding of right and wrong; another nationality, sex; a lot of money? Why?' The answers given most often were: different parents, different sex, different moral understanding. The importance of 'ideals' to identity could be seen in the frequent references to 'moral understanding,' and in the justifications given for the response 'different parents.' Most subjects argued that growing up in a different family would make them a different person because parents so decisively influence and shape one's value orientations. There is a more indirect confirmation of the claim that ideals are constitutive of identity: many longitudinal studies find that individuals tend to hold on to the sociopolitical value orientations developed in late adolescence—even if they have experienced social and geographic mobility or fluctuations in public opinion (cf. Sears 1981). This finding may indicate that people seek to stabilize their identity during their life course by remaining true to their value commitments—irrespective of changes in context or the ideological climate.

As noted above, however, this mode of upholding identity is rendered increasingly difficult given rapid and far-reaching changes in knowledge systems and collective interpretation. In the following, I maintain that with modernization two new mechanisms that enable us to constitute and stabilize identity are evolving: value generalization and an egosyntonic motive structure. I will illustrate both using data from a cohort comparison involving 100 65- to 80-year-old, 100 40- to 50-year-old, and 100 20- to 30-year-old

subjects representative of the West German population. The study focused on the understanding of moral rules and moral motivation (cf. Nunner-Winkler 2000a, 2000b). In making use of these data in the present context I do not, however, wish to imply that the ideals constitutive for identity need to be moral ones.

5. Mechanisms for Stabilizing Identity in Modernity

5.1 Value Generalization

Commitment to more abstract values increases the flexibility required if varying context conditions are to be taken into account. To give an example: when asked to judge a young mother's desire to work full-time without real economic need, older subjects express clear condemnation; they see it as a neglect of duty. Younger subjects, in contrast, no longer assume ascriptive female duties; they focus on the problem of good child care. On these more abstract terms several functionally equivalent solutions are possible. Similarly, older subjects tend to condemn a sexual relationship between two men as unnatural, deviant, or sick. Younger subjects instead tend to discuss it in terms of the quality of the relationship. Thus, increasingly concrete rules tying behavioral implications to specific context conditions (e.g., child care to being a mother; sexual relations to being of the opposite sex) are replaced by more abstract values (e.g., welfare of the child; quality of the relationship) that allow flexibility in realization without a betrayal of basic commitments.

5.2 Egosyntonic Motive Structure

Norm conformity can be motivated by various concerns: by fear of physical or institutional sanctions, social contempt (Luhmann), or pangs of conscience (Freud); by compassion (Schopenhauer); by respect for the law (Kant). In a longitudinal study on moral development it was found that children see moral behavior as being motivated not by the consequences anticipated either for the wrongdoer or for the victim (e.g., by fear of sanctions or by compassion), but by an intrinsic desire to do what is right (Nunner-Winkler 1999). The cohort comparison shows that between generations there is a change in the understanding of moral motivation indicated by changes in the emotional reactions expected to follow transgressions. Older subjects mostly refer to guilt and shame; that is, to emotions that focus on some higher internal or an external entity (e.g., the superego or a social audience); younger subjects, in contrast, are more likely to express sorrow or regret; that is, emotions that focus on the wrong committed or the harm done. Thus, norm conformity is increasingly coming to be seen not

in terms of bowing to dictates authoritatively imposed on people but rather in terms of not wanting to wrong others. This change in moral motivation in the direction of a more egosyntonic structure corresponds to a change in the cognitive understanding of morality, in as much as norms are no longer seen to be set by God or church authorities, but rather as derived from rational consensus among equals.

To conclude, with modernization—i.e., with social differentiation, cultural pluralization, and rapidly changing social conditions—new mechanisms have evolved that allow us to constitute and stabilize identity. By egosyntonic commitment to generalized values individuals may develop a sense of uniqueness, and experience consistency and continuity in their life course. In this way ego identity becomes a substitute for social role identity.

See also: Discourse and Identity; Identity and Identification: Philosophical Aspects; Identity in Anthropology; Identity in Childhood and Adolescence; Identity: Social; Multiculturalism and Identity Politics; Cultural Concerns; Religion: Nationalism and Identity

Bibliography

- Cohen S, Taylor L 1976 *Escape Attempts. The Theory in Practice of Resistance to Everyday Life*. Allen Lane, London
- Elster J 1979 *Ulysses and the Sirens*. Cambridge University Press, Cambridge, UK
- Erikson E H 1956 The problem of ego identity. *Journal of the American Psychoanalytic Association* 4: 56–121
- Erikson E H 1958 *Identity and Uprootedness in Our Time*. Paper presented at the Annual Meeting of the World Federation of Mental Health, Vienna
- Frankfurt H G 1988 *The Importance of What We Care About*. Cambridge University Press, New York
- Frankfurt H G 1993 On the necessity of ideals. In: Noam G G, Wren T E (eds.) *The Moral Self*. MIT Press, Cambridge, MA, pp. 16–27
- Giddens A 1991 *Modernity and Self-identity. Self and Society in the Late Modern Age*. Stanford University Press, Stanford, CA
- Goffman E 1963 *Stigma. Notes on the Management of Spoiled Identity*. Prentice-Hall, Englewood Cliffs, NJ
- Haste H 1993 Morality, self, and sociohistorical context: The role of lay social theory. In: Noam G G, Wren T E (eds.) *The Moral Self*. MIT Press, Cambridge, MA, pp. 175–208
- Keniston K 1965 *The Uncommitted: Eliminated Youth in American Society*. Harcourt and Brace, New York
- Keniston K 1968 *Young Radicals – Notes on Committed Youth*. Harcourt, Brace and World, New York
- Krappmann L 1969 *Soziologische Dimensionen der Identität*. Klett, Stuttgart
- Nozick R 1981 *The Identity of the Self. Philosophical Explanations*. Clarendon Press, Oxford, UK, pp. 27–114
- Nunner-Winkler G 1999 Development of moral understanding and moral motivation. In: Weinert F E, Schneider W (eds.) *Individual Development from 3 to 12. Findings from the Munich Longitudinal Study*. Cambridge University Press, New York, pp. 253–290

- Nunner-Winkler G 2000a Wandel in den Moralvorstellungen. In: Edelstein W, Nunner-Winkler G (eds.) *Moral im sozialen Kontext*. Suhrkamp, Frankfurt am Main
- Nunner-Winkler G 2000b Von Selbstzwängen zur Selbstbindung (und Nutzenkalkülen). In: Endreß M, Roughley N (eds.) *Anthropologie und Moral*. Königshausen Neumann, Würzburg, Germany
- Rawls J 1985 Justice as fairness: Political not metaphysical. *Philosophy and Public Affairs* 3: 223–51
- Sears D O 1981 Life-stage effects on attitude change, especially among the elderly. In: Kiesler S B, Morgan J N, Kincade Oppenheimer V (eds.) *Aging. Social change*. Academic Press, New York, pp. 183–204

G. Nunner-Winkler

Personal Privacy: Cultural Concerns

Personal privacy is a vexing and invaluable concept: vexing because it is difficult to describe; invaluable because privacy allows people the possibility of freedom, and thus autonomy and personal choice. In this sense, privacy is a precondition of psychological concepts of identity: the modern ideal of an individual requires—and cannot be fully realized without—substantial privacy. Personal privacy can be pictured as the atmosphere generating part of the ecosystem in which this individual can draw breath.

Perhaps most succinctly defined by the US social philosopher, Sissela Bok (1982), privacy is ‘the condition of being protected from unwanted access by others—either physical access, personal information or attention.’ This protection, whether physical, political, or psychological provides freedom because it shelters people from intrusion; it fends off predators that might pre-empt personal choice; it grants control. Privacy also emphasizes interiority: the subject—while protected from undesired surveillance and intrusive access—is freed to focus more fully on self-observation, and to develop efficacious forms of self-expression.

1. Historical Background

An early template of this phenomenon can be found in Jane Austen’s novel, *Pride and Prejudice*, published in 1813. Midway through the novel, its heroine Elizabeth is sitting alone in a room reading a letter when the romantic hero, Darcy, comes upon her and is startled by her solitude. Modern readers may not recognize what an unusual spectacle such a tableau offered in Austen’s era. Women were rarely alone, and many were only recently literate enough to read and write letters. The book also describes Elizabeth’s startling penchant for solitary walks. Her search for solitude underscores her characterization as a new, independent woman with a more self-conscious and thus a sturdier sensibility. Alone in a room with time and

peace to peruse an intimate letter, Elizabeth can contemplate her reactions to what she is reading. She can notice what she thinks and feels without immediately being distracted or having her impressions overrun or overruled by others around her. Given a pen and paper, she can respond from her own perspective; she can express an opinion, and in a larger sense, a point of view.

These behaviors, made possible by privacy, and the resulting increased capacities for autonomous voice and self-expression, distinguish Elizabeth from her less able peers, and define her as a substantial individual in a modern sense. This view is furthered by her refusal to marry Darcy when he first approaches. Although he is wealthy and upperclass, and she is from a more middle-class family with money worries, and although marriage is the only avenue into society for her, she refuses his offer because she does not love him. Her behavior is radical as it suggests love—and the sensibility of the subject—as a value of more significance than social standing or economic well-being. While in contemporary Western culture, many take for granted love as the superior foundation for marital commitment, this notion, this privileged of private feeling, is also—historically speaking—new, having developed a popular following gradually over the past few centuries. And, as Austen illustrates, in the era of *Pride and Prejudice*, it was easily displaced by other imperatives.

Austen's heroine is only satisfied when she can choose to marry Darcy, based on her private feelings—those she discovers within her heart. Elizabeth's insistence on such a course illustrates well the confluence between forms of privacy, like solitude, and the heightened capacity for personal choice based on feelings. One emphatic version of this idea of individual integrity rooted in private experience was asserted by the US transcendentalist Bronson Alcott (1799–1888) when he wrote, 'Individuals are sacred. The world, the state, the church, and school, all are felons whenever they violate the sanctity of the private heart' (in Seipp 1981, p. 99).

While all cultures have had some notions of privacy (Barrington Moore Jr. 1984 traces its rudimentary presence in Ancient Greece and the Old Testament), the modern concept—illustrated so well in *Pride and Prejudice*—has been developing gradually since around the seventeenth century. The cultural historian Philip Aries (1962) has pointed out that until the early seventeenth century, almost no one in Europe spent time alone. But privacy is more than solitude. Alan Westin defined its four substates as solitude, anonymity, reserve, and intimacy. While these states are diverse, they all describe methods of purposefully shielding oneself from the physical and/or psychological 'gaze' of others.

The ideal of personal privacy is partly an outgrowth of a gradual social transition accompanying the mass urban migration of the nineteenth and twentieth

centuries. People left communal situations—small towns and villages where their lives, monitored by church and neighbor, had been witnessed closely and controlled largely by people who knew them. Finding themselves more alone, separate, and increasingly anonymous, people, first in cities then in suburbs, realized that they were at liberty to choose their activities and behaviors with less reference to communal imperatives. Although in part following an economic infrastructure defined by industrialization and capitalism, this increased autonomy drew its cultural meaning from philosophies of individual rights and freedom. While the sources are too many to site, one would have to acknowledge the influence of John Locke, Edmund Burke, Jean Jacques Rousseau, Johann Goethe, William Wordsworth, Thomas Jefferson, Ralph Waldo Emerson, Mary Wolstencraft, Margaret Fuller, and Henry David Thoreau among many others.

Within this tradition, privacy can be defined as the right to make one's own choices relatively unobserved, unchastised, and unimpinged upon by others—whether individuals or institutions. While there is no legal 'right to privacy' in the USA, the concept began to be articulated in the late nineteenth century. In 1872 the feminist Victoria Woodhull referred to 'the right to privacy' in a newspaper piece. The famous expression, 'the right to be let alone,' was then coined by Judge Thomas Cooley in 1879. In 1890, Samuel Warren and Louis Brandeis published a landmark article, 'The Right to Privacy' in the *Harvard Law Review*. No important discussion of privacy in the USA since then has failed to refer to Warren and Brandeis.

Privacy had been increasingly understood as a precondition for human dignity. Shielded from intrusive observation, people take control over the matter of 'showing' themselves to others or not based upon their own sense of readiness. Few feel dignified when exposed to hostile or secret surveillance. Privacy functions as a practice room in which people can process and master experience before displaying themselves to others. Having this option, people feel less shame. Shame is in essence an internal overseer—employed by those parts of the mind whose job is to keep the individual in line with the communal norm. Privacy provides the space in which the communal demand can be weighed against the internal individual imperative, and a choice of behavior made through consideration rather than reflex. Since shame is also a psychic experience where the sense of self temporarily disintegrates, the shielding of the individual with privacy, the minimizing of shame, and the valuing of dignity, all enforce conditions which enhance people's sense of individual identity and worth.

2. Theories of Privacy

Alan Westin's landmark book, *Privacy and Freedom* (1968), defined four states of privacy. Because these

states are to some degree counterintuitive, it is useful to describe them. The first—solitude—is perhaps the most self-evident and the one most popularly associated with privacy. When privacy is mentioned, the mind often jumps to an image of a person alone, whether in nature or in a room. In US literary culture, one might think of Henry David Thoreau living at Walden Pond, or—half a century earlier—of James Fenimore Cooper's fictional frontiersman, Natty Bumppo, independently making his way through the American wilderness. Solitude is the most complete state of privacy; in it one is physically separate from others. Historically, solitude has been associated with men and linked to images of independence, moral awareness, and strength. Until recently in Western Europe and North America, women were only granted solitude for prayer. Unchosen, solitude becomes isolation and, like the extraordinarily punitive condition of solitary confinement, can be a psychically destructive experience; people, with inadequate access to others become more inclined to psychic turmoil.

Anonymity is Westin's second state of privacy. When one is with others but unknown to them, as in a crowded train or a city or a suburban shopping mall, one is anonymous. Anonymity offers privacy by allowing human proximity in the absence of knowledgeable witnesses. One is freed in this situation (for better and worse) to experience a subjective sense of oneself or of current experience unchallenged by the proximity of knowing and potentially contradicting others. The degree of anonymity that urban and suburban dwellers experience today is historically atypical and sometimes shades into the alienating sense of anomie.

The third state of privacy is reserve. Reserve is perhaps the oldest and most universally available form of privacy; the privacy one can access even when surrounded by intimates. It is the default state of privacy, the great mediator of psychic survival and human freedom. One has, thanks to reserve, the freedom not to say openly or fully what one thinks or feels, not to reveal oneself in situations that feel unsafe. Reserve is a basic element in civility and civilization, but more than that, it negotiates the distances between people, facilitating all varieties of encounter by allowing exchanges among people where disruptive observations go unspoken and social intercourse is expedited. It is also one of the equalizers in social power. The powerful, denied critical response or information by the reserve of the 'weak,' can have their intentions impeded.

Intimacy is Westin's fourth state of privacy, and perhaps the most surprising one to find in association with the others. Intimacy is a state of privacy where two people, or a small group or a family, shielded from the larger group, are able to reveal themselves more fully, physically, psychologically, sexually, and so forth. As individuals have come to live more separately and less communally, there has been an increased

demand for intimacy in primary relationships. At the same time, continuing our delineation of the relationship of privacy to the ideal of the more fully developed individual, intimacy is considered a well-spring for psychic development. Within intimate relationships, individuals hope to find validation of their nascent self-expression which then ideally strengthens and reinforces their capacity to carry that private self into the public world. Intimacy also provides release in the form of freer emotional and sexual expression.

The 'possibility' within privacy also includes individual artistic expression, which in turn dovetails with the romantic emphasis on the artist as the highest realization of modern sensibility. According to this view, well articulated in Charles Taylor's (1989) compendium on the sources of modern identity, the artist, creating in private and using privacy to appropriately shield and nurture his or her creative process, expresses otherwise obscured, often internally uncovered, truths about human experience, and by so doing transforms that experience.

And while the artist may be the prototype of this style of individual, it would be a mistake to minimize privacy's broader but related function in everyday life. Take two contemporary flashpoints as illustrative: One can see the critical role of personal privacy both in the abortion debate and the gay rights movement. In both these situations, the wish of individuals to choose behavior around sexuality and procreation is placed in conflict with traditional social imperatives. Specifically, these imperatives have dictated the primacy of procreation and heterosexual procreative sex. In both cases, the progressive position has been framed as valuing individual choice above procreation specifically because this choice highly values subjective experience—the primacy of the individual heart. Thus the individual's private feelings—about readiness to have children, or the number of children to have, about whom one wants to love and how—are held to be of worth precisely because they enforce people's capacity to act according to their own lights, and to transform themselves by these actions: to become 'more fully' human through freedom.

Sigmund Freud's notions of a complex human psychology that requires and transforms itself through self-knowledge, and particularly through self-knowledge inculcated in the intimate process of psychoanalysis, fits within this modern template. In the various psychotherapies which have proliferated in the USA and Europe during the twentieth century, the therapist assists patients to become themselves more fully by verbalizing their private (even secret) heretofore inexpressible thoughts and feelings. The process of naming, the intimate yet contractual relationship, and the resulting restructuring of aspects of the patient's mind are meant to create a transformative and relatively liberating self-knowledge. The privacy of the consulting room and the confidentiality of the process are considered central to its efficacy in that

they secure a free space between social demand and individual subjectivity.

3. Threats to Privacy

Privacy is a fragile state, and one—as the twentieth century has demonstrated—that is exploited readily. No discussion of the merits of personal privacy can be complete without some exploration of its accompanying liabilities. While many of the most eloquent voices of the twentieth century—among them George Orwell, Nadia Mandelstam, and Primo Levi—have delineated the damage to liberty and individual authority brought about by hostile surveillance and its companion, terror, less has been said until recently of the equal threat to privacy caused by secrecy and social isolation. (Perhaps the one historically illuminated aspect has been in those instances when it has been tied with totalitarian power wherein the masses are placed under surveillance while the motives and behaviors of leaders are kept secret.)

But domestic secrecy turns out to be a potent destructive force with regard to personal privacy. Under the guise of privacy, more powerful members of families have had reign to exploit or abuse less powerful members—whether emotionally, sexually, or physically. These acts are significant not only because of their exploitative dimensions, but because by their very nature they stunt the freedom of the oppressed to use the privacy for the aims elaborated above. Any historical survey of social conditions would suggest that such abuse has always been amply present. Yet what has altered currently may be an ideology which pictures family life as unrealistically benign—‘a haven in a heartless world’ (Lasch 1977)—an idealization that sometimes denies, sometime mitigates, the inherent violence.

A corollary concern is the degree to which many of the benefits made possible by personal privacy accrue only to a privileged or powerful few. Does such a scatter imply that privacy reinforces other social hierarchies, or can it still be construed as a valuable humanistic ideal which will only find full expression if social justice increases other aspects of social equality?

So too, US social philosophers like Christopher Lasch, and the ‘communitarians’ Robert Bellah and Amatai Etzioni have attacked aspects of US individualism which they see both as excessive, and as indulging an unchallenged pursuit of private purpose and private gain at the expense of the welfare of a larger group. According to this critique, the goals of the individual and the community may be enough at odds to warrant more vigorous interventions on the part of the state to protect communities. Privacy becomes implicated in this debate for its functions shielding individuals and honoring their imperatives.

The technological revolution of the present age also threatens privacy. Computer databanks, video

cameras, audio taping, sophisticated surveillance capabilities, and medical technologies like DNA testing make it extremely easy for people to learn private information about other people which was more difficult to access earlier in the twentieth century. Reacting to excessive social isolation and anonymity, and to the enormous profit in selling private information, there is strong interest from many quarters in developing new ways of observing others. The technological gains have come so quickly that inadequate attention is being paid to their pros and cons with regard to the value of preserving privacy. So, for example, the DNA sampling in prison of a petty offender implies a lifelong increased access to knowledge about that individual by the state. And civil liberties issues have tended to get suppressed in the enthusiasm for the new capabilities.

4. Conclusion

In summary, personal privacy protects a wide variety of modern social ideals. Privacy advocates argue that people have a right to be ‘let alone’; and that, shielded from excessive authoritarian intrusions, people will make choices based on some more personal sense of what is desirable, that thus constitutes more authentic self-expression. As a corollary, these ideals justify and affirm other cultural notions which support individual liberty and dignity, the value of self-knowledge, and they privilege relationships based in love.

See also: Civil Liberties and Human Rights; Human Rights, History of; Individualism versus Collectivism: Philosophical Aspects; Love and Intimacy, Psychology of; Privacy: Legal Aspects; Privacy of Individuals in Social Research: Confidentiality; Rights

Bibliography

- Aries E P 1962 *Centuries of Childhood: A Social History of Family Life*. Vintage Books, New York
- Bellah R, Madsen R, Sullivan W M, Swidler A, Tipton S M 1986 *Habits of the Heart: Individualism and Commitment in American Life*. Perennial Library, New York
- Bok S 1982 *Secrets: On the Ethics of Concealment and Revelation*. Pantheon Books, New York
- Etzioni A 1993 *The Spirit of Community*. Crown, New York
- Garrow D J 1994 *Liberty and Sexuality: The Right of Privacy and the Making of Roe vs. Wade*. Macmillan, New York
- Lasch C 1977 *Haven in a Heartless World*. Basic Books, New York
- Moore B Jr 1984 *Privacy: Studies in Social and Cultural History*. M. E. Sharpe, Armonk, NY
- Rieff P 1966 *The Triumph of the Therapeutic*. Harper and Row, New York

- Seipp DJ 1981 *The Right to Privacy in American History*. Harvard University Program on Information Resources Policy, Cambridge, MA
- Smith J M 1997 *Private Matters: In Defense of the Personal Life*. Addison Wesley, Reading, MA
- Taylor C 1989 *Sources of the Self: The Making of Modern Identity*. Harvard University Press, Cambridge, MA
- Warren S D, Brandeis L D 1890 The right to privacy. *Harvard Law Review* 4/5
- Westin A 1968 *Privacy and Freedom*. Atheneum, New York

J. M. Smith

Personality and Adaptive Behaviors

Personality is the psychological system that mediates the relationship of the individual with the environment in the service of adaptation. It is the complexity of structures, processes and patterns of affect, cognition and behavior that we call 'personality' that conveys to others the sense of unity and distinctiveness that distinguish individuals from one another and that grants to each individual a sense of identity, of wholeness and uniqueness.

Because the term encompasses a great variety of phenomena it is not surprising that different ideas appear in various combinations in the traditional and common usage of the word personality: human being, person, the quality of being a person, the quality or fact of being a particular person, patterns of habits and qualities of any individual as expressed through physical and mental activities, the sum of such properties as impressing or likely to impress others, and what makes an individual conspicuously different from others.

Nor is it surprising that, given the complex nature of personality, different assumptions, foci, and strategies of research have marked the study of personality since the beginning of the discipline (Allport 1937, Murray 1938, Stagner 1937, Stern 1935).

The resolution of these different definitions and assumptions has considerable implications for the conceptualization and study of adaptation and adaptive behaviors.

1. Common Assumptions Regarding Personality

Before outlining these different assumptions, let us first consider some areas of commonality in the contemporary field. Most investigators in personality psychology share an 'interactionist view'. In other words, they view personality as an open system that develops and functions through continuous and reciprocal interactions with the environment (Bandura 1986, Hettema 1979, Magnusson and Stattin 1998).

Personality, then, refers to a complexity of psychological structures and processes that operate in concert in these person-situation interactions. The personality system underlies the distinctive patterns of affect, cognition, and behavior that foster coherence in individual conduct and experience over time and across settings. Molar psychological mechanisms, such as the goals, beliefs, and skills through which individuals regulate their experiences and actions, result from the synergism of underlying biological and psychological subsystems. Multiple biological systems, cognitive and affective structures, behavior, and the environment all operate as interacting determinants of what personality is at any moment within a network of reciprocal causation. The relative influence of one or another personality determinant varies in different individuals, and across activities and circumstances. Individual development and functioning are processes of continuous restructuring of the whole system and its subsystems within the boundaries set by biological and social constraints (Caprara and Cervone 2000).

2. Competing Arguments Regarding Personality

Despite the widespread adherence to an interactionist viewpoint, continuing debates in the field attest to the fact that many fundamental issues are still pending, with different investigators adhering to different theoretical and metatheoretical principles and assigning different importance to nature, nurture, or the individual in the shaping of their personalities (Cervone 1991). Throughout much of its history, personality psychology has been concerned with individual differences in observable variations in styles of behavior, affect, and cognition. The variations have been organized according to, and traced back, to simple systems of dispositional constructs.

Dispositional constructs are designed to encompass the variety of phenotypic expressions of individuality and to capture the consistent individual differences that are observed. Trait constructs have been posited to account for stable patterns of experience and action that people exhibit, and that distinguish them from one another. As different traits are differentially relevant to different settings and their expression is differentially sensitive to social contexts, different traits attest to the consistency of personality in various degrees depending upon their generalizability (across individuals), pervasiveness (across situations), and stability (across time). This has led dispositional theorists to conceptualize personality as a hierarchical organization and to focus on high level traits (e.g., extroversion) that organize lower level tendencies (e.g., sociability) which, in turn, supervise lower level behavioral habits (e.g., talkative) (Eysenck 1970).

In the past alternative taxonomies have been in conflict with each other. However significant con-

sensus recently has been reached on the importance of five factors (so called Big Five) that are obtained robustly: Extroversion, Agreeableness, Conscientiousness, Emotional Stability, and Openness to Experience. This five-factor model has become a point of convergence between the psycholexical tradition, mostly concerned with personality description (Goldberg 1993), and the tradition, mostly concerned with the identification of the primary structures of personality (McCrae and Costa 1996).

Significant differences are found among five-factors investigators, who may disagree on the exact nature and conceptual status of the trait variables (see Wiggins 1996). However, far more profound differences separate those who view traits as the ultimate determinants of personality from those who conclude that clustering descriptors related to habitual behaviors (as resulting from factor analysis) cannot lead too far in understanding either the regularities that explain personality functioning or the uniqueness which distinguishes persons from one another (Bandura 1999). The issue in contention is not whether people have personal dispositions, nor their determinative role in personality functioning, but rather how dispositions are conceptualized and operationalized. Dispositions may correspond to habitual styles of behavior rooted in genotypes, or to self-regulatory structures (as internal standards, aspirations and goals, and efficacy beliefs) resulting from the organization of affect, cognition, and behavior which emerges over the course of life.

In conceiving of personality as a hierarchical architecture in which basic traits are at the origins of any other individual disposition or behavior, most convinced advocates of traits as genotypes of personality derive their arguments from genetic, longitudinal, and cross-cultural studies. Genetic studies have found a consistent component of heritability for each of the Big Five (Lohelin 1992), longitudinal studies have shown a significant stability of the same traits over the life span (Costa and McCrae 1997), and cross-cultural studies have shown that the multiple descriptors used to describe personality in different cultures can be traced back to few common dimensions broadly overlapping with the Big Five (De Raad 1998). Trait psychologist converge with evolutionary psychologist in viewing adaptation as largely predetermined by the basic dispositional equipment of the species selected through generations. According to this view adaptive behaviors (in mating, fighting, solving problems, and coping with adversities) largely reflect the influence of evolved mechanisms whose nature and functioning has been determined over the eons through processes of natural selection (Buss 1997).

On the other hand, in conceiving of personality as an open, dynamic, unifying and integrating system, social cognitive and motivational theorists (Cervone and Shoda 1999) point to the emerging properties of the mind and focus on the processes and mechanisms

conducive to knowledge structures which enable personality to function as a proactive self-regulatory system. Thus, they contest the generalizability to individuals of correlational findings derived from studying populations, and thus leaving large portions of variability unexplained. Rather, they focus on the construction of personality as an integrative and coherent system as it takes place over the course of life, on the processes that enable the system to function proactively with the environment and on the structures that orchestrate these processes. Ultimately they emphasize the flow of reciprocal determination among biological, psychological and social factors which sustain the whole construction, and from which the agentic properties of personality derive. In this view of personality, 'adaptation' is eminently relational and contextual. Adaptive behaviors are not ascribed to any preformed, endogenous dispositions, but to psychological systems which themselves develop through dynamic interactions with the social environment. These systems enable the person to cope in a flexible, innovative manner with the diverse challenges of life.

3. Personality as a Self-regulatory System

Personality can be viewed as both a construction and an agentic system. Over time, people gradually construct a sense of self through their interactions with the social world. People's capacity for forethought, performance control, and self-reflection contributes to their ability to regulate their own experiences and actions. Self-regulatory capacities enable people to contribute proactively to the course of their development. Personality, then, reveals itself as a construction endowed with agentic properties. These properties enable people to exert a broad influence on their environment and their own psychological and physical selves (Caprara and Cervone 2000).

Agentic personality functioning involves a number of distinct processes. People select the environments they encounter and thereby create opportunities for themselves, guide their behavior purposively, appraise their circumstances, and chose levels of mental and physical challenge that match their perceived capabilities. Person-situation transactions underscore the fact that personality functioning is not reactive but proactive. Individuals do not only react to the environments, but actively contribute to their transformation and creation. Individuals' distinctive ways of functioning reflect more than the unfolding of a genetic blueprint or the interactions which have taken place between the organism and the environment over the course of development. They also reflect individuals' active role in shaping their own personality.

Thus, one should look at adaptation as a dynamic, proactive process. People do not 'adapt' merely by altering themselves to fit environmental constraints and pressures. They select and transform environ-

ments in order to actualize their full potentials and maximize their own well being. Adaptation relies more on individuals' capacities to devise strategies and to implement tactics able to match successfully the requirements and affordances of multiple situations, than on fixed pattern of responses. The capacity to expand knowledge of oneself and of situations is no less important than the capacity to master such knowledge with the flexibility that is required by the variety and continuous change of person-situation transactions.

Over the course of life, people acquire competencies, knowledge about themselves and the social world, and standards for evaluating the worth of own actions. Self-reflective and self-directive processes enable individuals to meet environmental opportunities and constraints as well as to maintain personally valued courses of action in the relative absence of tangible rewards or social support. While people show a striking ability to adapt their behavior to physical, interpersonal, and social circumstances, and also to select and create the circumstances that best fit with their purposes and abilities, successful adjustment across the life span corresponds to the best allocation of personal and social resources within the set of temporally bound opportunities and constraints.

Thus selection, optimization, and compensation have been regarded as main strategies which attest the property of personality to capitalize upon experience in the achievement of higher levels of functioning and well-being. People must select life goals, activities, and pathways that are manageable; they must optimize the allocation of internal and external resources that are made available to them; they must develop strategies to compensate for lacks or losses of goal relevant means (Baltes 1997).

As people are causal agents who actively contribute to the course of their life in selecting and transforming the environment they encounters, adaptive behaviors are those behaviors that promote the conditions for the full expression of individual capacities and potentials. As individuals course of life is closely interdependent with the course of life of others, ultimately adaptive behaviors are those behaviors that serve for both the betterment of individual and collective welfare.

Future research should focus on psychological processes and social capacities that can strengthen people's capacity to adapt to a rapidly changing world. To this aim theories and findings are mostly needed to enable people to expand their knowledge and skills, to achieve greater emotional adjustment, and thus to fully realize their potentials.

See also: Interactionism and Personality; Personality Psychology; Personality Structure; Personality Theories; Self-efficacy; Self-regulation in Adulthood; Self-regulation in Childhood; Well-being (Subjective), Psychology of

Bibliography

- Allport G W 1937 *Personality: A Psychological Interpretation*. Holt, Rinehart & Winston, New York
- Bandura A 1986 *Social Foundations of Thought and Action: A Social Cognitive Theory*. Prentice-Hall, Englewood Cliffs, NJ
- Bandura A 1999 Social cognitive theory of personality. In: Cervone D, Shoda Y (eds.) *The Coherence of Personality: Social-Cognitive Bases of Consistency, Variability, and Organization*. Guilford Press, New York, pp. 185–241
- Baltes P B 1997 On the incomplete architecture of human ontogeny: Selection, optimization, and compensation as foundation of developmental theory. *American Psychologist* 52: 366–80
- Buss D M 1997 Evolutionary foundations of personality. In: Hogan R, Johnson J, Briggs S (eds.) *Handbook of Personality Psychology*. Academic Press, San Diego, CA, pp. 318–44
- Caprara G V, Cervone D 2000 *Personality: Determinants Dynamics, and Potentials*. Cambridge University Press, New York
- Cervone D 1991 The two discipline of personality psychology. *Psychological Science* 2: 371–77
- Cervone D, Shoda Y 1999 (eds.) *The Coherence of Personality*. Guilford Press, New York
- Costa P T, McCrae R R 1997 Longitudinal stability of adult personality. In: Hogan R, Johnson J, Briggs S (eds.) *Handbook of Personality Psychology*. Academic Press, San Diego, CA, pp. 269–90
- De Raad B 1998 Five big five issues: rationale, structure, status and cross-cultural assessment. *European Psychologist* 3: 113–24
- Eysenck H 1970 *The Structure of Personality*, 3rd edn, Methuen, London
- Goldberg L 1993 The structure of phenotypic personality traits. *American Psychologist* 48: 26–34
- Hettema J 1979 *Personality and Adaptation*. North-Holland, Amsterdam
- Lohelin J C 1992 *Genes and Environment in Personality Development*. Sage, Newbury Park, CA
- Magnusson D, Stattin H 1998 Person-context interaction theories. In: Damon W, Lerner R R M (eds.) *Handbook of Child Psychology: Theoretical Models of Human Development*, 5th edn. Wiley, New York, Vol. 1. pp. 685–760
- McCrae R R, Costa P T 1996 Toward a new generation of personality theories: theoretical contexts for the five-factor model. In: Wiggins J S (ed.) *The Five-Factor Model of Personality. Theoretical Perspectives*. Guilford Press, New York, pp. 51–87
- Murray H A 1938 *Explorations in Personality*. Oxford University Press, New York
- Stagner R 1937 *Psychology of Personality*. McGraw Hill, New York
- Stern W 1935 *Allgemeine Psychologie auf Personalischer Grundlage*. Nijhoff, Dordrecht
- Wiggins J S (ed.) 1996 *The Five-Factor Model of Personality. Theoretical Perspectives*. Guilford Press, New York

G. V. Caprara

Personality and Conceptions of the Self

A central aspect of personality involves the ways in which people perceive and think about themselves.

Unlike most nonhuman animals, human beings possess a capacity for self-reflexive thinking that allows them to think consciously about themselves in complex and abstract ways. People differ in the degree to which they self-reflect, the content and organization of their self-conceptions, the positivity of their self-evaluations, and the selves they desire and fear to become in the future. These differences in how people think about, characterize, and evaluate themselves underlie many of the characteristic patterns of thought, emotion, and behavior that comprise the human personality.

1. Self-attention

People differ in the amount of time that they spend attending to and thinking about themselves. Individual differences in self-attention have been studied most extensively in terms of the trait of self-consciousness. (As used in this context, self-consciousness refers only to the tendency to attend to oneself and does not involve the anxiety and awkwardness that everyday use of the term connotes.) People who are high in private self-consciousness—the tendency to think about inner, private aspects of oneself (such as one's motives, goals, attitudes, and feelings)—tend to behave more consistently with their attitudes, values, and goals than people who are low in private self-consciousness, presumably because self-attention increases people's adherence to their personal standards. Because their behavior is strongly influenced by their personal standards, people high in private self-consciousness display greater consistency in their behavior across different situations than those who are low in private self-consciousness (Carver and Scheier 1981).

Public self-consciousness—the tendency to think about the public, observable aspects of oneself (such as one's appearance, mannerisms, and image in other people's eyes)—is associated with the degree to which individuals try to control how they are perceived by other people (e.g., see *Impression Management, Psychology of*). Because they are more concerned with how they are perceived and evaluated by others, publicly self-conscious people conform more to group pressure, show greater concern for their physical appearance, and tend to be more socially anxious and embarrassed than people who are low in public self-consciousness (Carver and Scheier 1985).

Importantly, whether people are high in private or public self-consciousness has implications for how they perceive themselves. People high in private self-consciousness not only attend more to their private selves but also regard those private aspects of self as more important to who they are. In contrast, people high in public self-consciousness regard the public aspects of self as more salient and important.

2. Self-concept

The self-concept involves people's beliefs about their personal characteristics, not only beliefs about their traits, abilities, and physical attributes but also about their values, roles, and personal goals. People differ both in the content of their self-concepts (e.g., who and what they think they are) and in how the various elements of their self-concepts are organized. Again, these differences contribute to variations in personality.

Self-concepts develop through the individual's direct experience, as well as from the reflected appraisals of other people (e.g., see *Self-development in Childhood; Identity in Childhood and Adolescence*). Once formed, people's self-concepts exert a strong influence on their behavior. Perceiving oneself as a particular kind of person prescribes and proscribes certain actions. Seeing oneself as polite, for example, has implications for the behaviors that a person will and will not perform. As a result, people generally behave in ways that are consistent with the kind of person they think they are.

In addition, people's self-concepts provide them with information about the likely consequences of certain behaviors. Believing themselves to possess particular attributes leads people to expect certain outcomes to result from their actions. For example, a person whose self-concept included the trait 'intelligent' would have different expectations about performance in an academic setting than a person whose self-concept included the trait 'stupid.' Thus, the self-concept has implications for a person's sense of self-efficacy (e.g., see *Personality Development in Childhood*). Believing that one possesses particular attributes and abilities affects people's expectations regarding whether they ought to be capable of particular tasks, thereby influencing their motivation to engage in certain behaviors.

Research has supported James's (1890) notion that the self-concept is composed of multiple components. People's self-identities are composed both of personal aspects that characterize them without respect to other people (e.g., personal abilities, values, and goals) as well as social aspects that identify them with respect to their interpersonal roles and relationships (e.g., particular family relationships, friendships, and group memberships). People differ in the importance that they place on the personal versus social aspects of their self-identities (Cheek 1989), and these differences predict their behavioral choices and reactions to events. For example, people who place greater importance on the personal aspects of their identity seek jobs that allow them to fulfill personal goals (e.g., self-fulfillment, personal growth, expression of their values), whereas people who place greater emphasis on the social aspects of their identities seek jobs that provide desired social outcomes (e.g., recognition, an active social life, interpersonal relationships). People

with different prevailing aspects of identity also prefer different kinds of recreational activities—solitary activities that promote self-improvement for people who emphasize personal identity, and group activities that promote interpersonal relationships for people who emphasize social identity.

People generally try to maintain consistency among the various aspects of their self-concepts, as well as between their self-concepts and reality. People tend to be troubled by contradictions between two aspects of their self-concepts and when objective information discredits their self-perceptions, possibly because such inconsistencies undermine people's certainty in their beliefs about themselves and their worlds (e.g., see *Cognitive Dissonance*). In fact, receiving information that is inconsistent with one's self-concept may be distressing even if it otherwise reflects positively on the person. As a result, people are generally motivated to behave in ways that are consistent with their self-concepts and prefer interacting with those who see them as they see themselves. Thus, people's self-concepts tend to steer their behavior in directions that help to maintain and verify their existing self-views (Swann et al. 1987).

Beyond the fact that people hold different ideas of who they are and what they are like, people's self-concepts also differ in their clarity, complexity, and differentiation. First, people differ in the clarity and certainty of their self-concepts. People whose views of themselves are clear, confident, and certain tend to score higher in self-esteem and positive affect, and lower in depression and anxiety, than people who have unclear, unstable views of themselves (Campbell et al. 1996). People whose self-concepts are unclear or uncertain are also more easily persuaded than those with clearer and more certain self-concepts, and uncertain self-concepts are more likely to change due to experience or social feedback.

Second, people differ in self-complexity—the number of discrete ways in which they characterize themselves and in the interrelatedness of the various aspects of their self-concepts. Although initial research suggested that people with greater self-complexity are less strongly affected by negative life experiences (Linville 1987), refinements of the construct show that self-complexity bears a more complex relationship with emotional and behavioral outcomes than was first assumed (Gramzow et al. 2000, Woolfolk et al. 1995).

Third, people differ in self-concept differentiation—the degree to which they see themselves as having different personality characteristics when they are in different social roles (Donahue et al. 1993). People low in self-concept differentiation tend to see themselves as relatively constant across various social roles, whereas highly differentiated people see themselves quite differently in different roles. Although differentiation might seem to be beneficial because it promotes greater behavioral flexibility, research suggests that high

differentiation is generally undesirable, reflecting lack of an integrated core sense of self. Compared to people low in self-concept differentiation, highly differentiated people tend to be more depressed and anxious, lower in self-esteem, less agreeable, less satisfied with how they perform in their roles, and lower in self-control.

3. Self-evaluations

People not only see themselves as being a certain kind of person with particular sorts of attributes, but they also evaluate themselves in positive and negative ways, and experience corresponding positive and negative feelings when they think about who and what they are (e.g., see *Self-esteem in Adulthood; Self-evaluative Process, Psychology of*). Whereas self-concept refers to the *content* of people's beliefs about themselves, self-esteem refers to people's *evaluations* of those beliefs. Although people's evaluative feelings about themselves change across contexts and over time (i.e., state self-esteem), most research has focused on stable patterns of self-evaluation across situations (i.e., trait self-esteem). Researchers often refer to people having 'low' versus 'high' trait self-esteem, but, in reality, most people's trait self-esteem falls in the neutral to positive range, and relatively few individuals evaluate themselves negatively at a global level (Baumeister et al. 1989).

Trait self-esteem has proven to be an important psychological variable that predicts a great variety of emotions and behaviors. Self-esteem is generally correlated, albeit weakly, with indices of positive mental health such as low anxiety, neuroticism, and depression, and with a high sense of personal control and self-efficacy. However, self-esteem is also associated with maladaptive behaviors such as nonproductive persistence, risk-taking (particularly when the person's ego is threatened), and a tendency to make unrealistically self-serving attributions. Controversy exists regarding why self-esteem is related to behavior, thought, and emotion as it is. Various theories suggest that self-esteem is a marker of social dominance (dominance theory), a gauge of social acceptance (sociometer theory), a buffer against existential anxiety (terror management theory), and a reflection of personal autonomy (self-determination theory) (Hoyle et al. 1999).

People differ in the degree to which their positive and negative self-evaluations are confined to particular aspects of their self-concepts vs. spread across many aspects of the self. For people with highly compartmentalized self-concepts, positive and negative information about the self is confined to different aspects of the self; for people with self-concepts low in compartmentalization, positive and negative information is mixed across many aspects of self. Thus, two people may have an equal number of aspects of self

that they evaluate positively versus negatively yet differ in the degree to which these aspects are compartmentalized. Compartmentalization is associated with positive emotions and high self-esteem for people whose positive attributes are more important to them than their negative attributes. However, compartmentalization is associated with negative affect and low self-esteem for people whose negative attributes are more important (Showers 1992).

People differ also in the stability of their self-esteem. For some people, feelings of self-esteem change very little as they move from one situation to another, whereas the self-esteem of other people is quite labile, even over short spans of time. Thus, two individuals with the same general level of self-esteem may differ markedly in the degree to which their self-feelings vacillate. People with unstable self-esteem appear to possess fragile and vulnerable feelings of self-worth that are easily influenced by their experiences. They also tend to be more prone to anger and aggression when their self-esteem is threatened (Kernis and Waschull 1995).

4. *Desired and Undesired Selves*

People differ not only in how they perceive themselves at the present time but also in how they wish to be in the future. The human capacity for self-reflection allows people to imagine themselves in the future and to adjust their present behavior in ways that move them toward desired selves and away from undesired selves. These mental representations of desired and undesired future selves serve as incentives that motivate action, and as guides that channel behavior in particular directions (Markus and Nurius 1986). Interestingly, the future selves that people fear appear to exert greater power over their behavior than the selves that they desire (Ogilvie 1987).

A great deal of human emotion is a reaction to the fortunes of people's desired and undesired selves (Leary in press). People compare themselves and their outcomes in life to their desired and undesired selves, and experience various emotions depending on whether they are moving toward or away from their desired selves. Events that have positive implications for future desired selves evoke positive reactions, whereas events that have negative implications for desired selves (or, worse, indicate movement toward undesired future selves) cause emotional upset.

Self-discrepancy theory proposes that people's emotional reactions are strongly influenced by two particular sets of concepts about the desired self (Higgins 1987). According to the theory, people compare themselves to both an ideal self (how they would like to be) and an ought self (how they think they should be). A discrepancy between one's actual self and ideal self leads to dejection-related emotions such as sadness and disappointment, whereas a discrepancy between

actual and ought selves leads to agitated emotions such as anxiety and guilt. Research has generally supported the notion that self-discrepancies produce emotion, although not always precisely in the manner predicted by self-discrepancy theory.

5. *Future Directions*

Theorists have long recognized that people's real, desired, and feared self-conceptions underlie aspects of their personalities, and a great deal of research has investigated various relationships between the self and personality. However, the field lacks an overriding conceptual framework for understanding how different aspects of the self interrelate and how these aspects relate to personality structures and processes. An impending task for theorists and researchers is to integrate what is known about self-relevant aspects of personality within a single theory.

See also: James, William (1842–1910); Personality Development in Adulthood; Personality Psychology; Self-development in Childhood; Self-monitoring, Psychology of

Bibliography

- Baumeister R F 1998 The self. In: Gilbert D, Fiske S T, Lindzey G (eds.) *The Handbook of Social Psychology*. Oxford University Press, New York, pp. 680–740
- Baumeister R F, Tice D M, Hutton D G 1989 Self-presentational motivations and personality differences in self-esteem. *Journal of Personality* **57**: 547–79
- Campbell J D, Trapnell P D, Heine S J, Katz I M, Lavallee L F, Lehman D R 1996 Self-concept clarity: Measurement, personality correlates, and cultural boundaries. *Journal of Personality and Social Psychology* **70**: 141–56
- Carver C S, Scheier M F 1981 *Attention and Self-regulation: A Control Theory Approach to Human Behavior*. Springer-Verlag, New York
- Cheek J M 1989 Identity orientations and self-interpretation. In: Buss D M, Cantor N (eds.) *Personality Psychology*. Springer-Verlag, New York, pp. 275–85
- Donahue E M, Robins R W, Roberts B W, John O P 1993 The divided self: Concurrent and longitudinal effects of psychological adjustment and social roles on self-concept differentiation. *Journal of Personality and Social Psychology* **64**: 834–46
- Gramzow R H, Sedikides C, Panter A T, Insko C A 2000 Aspects of self-regulation and self-structure as predictors of perceived emotional distress. *Personality and Social Psychology Bulletin* **26**: 188–205
- Higgins E T 1987 Self-discrepancy: A theory relating self and affect. *Psychological Review* **94**: 319–40
- Hoyle R H, Kernis M H, Leary M R, Baldwin M W 1999 *Selfhood: Identity, Esteem, Regulation*. Westview Press, Boulder, CO
- James W E 1890 *The Principles of Psychology*. Henry Holt, New York

- Kernis M H, Waschull S B 1995 The interactive roles of stability and level of self-esteem: Research and theory. In: Zanna M (ed.) *Advances in Experimental Social Psychology*. Academic Press, San Diego, CA, pp. 93–141
- Leary M R in press The self and emotion: The role of self-reflection in the generation and regulation of affective experience. In: Salovey P (ed.) *Handbook of Affective Science*. Wiley, New York
- Linville P 1987 Self-complexity as a cognitive buffer against stress-related depression and illness. *Journal of Personality and Social Psychology* **52**: 663–76
- Markus H, Nurius P 1986 Possible selves. *American Psychologist* **41**: 954–69
- Ogilvie D M 1987 The undesired self: A neglected variable in personality research. *Journal of Personality and Social Psychology* **52**: 379–85
- Showers C 1992 Compartmentalization of positive and negative self-knowledge: Keeping bad apples out of the bunch. *Journal of Personality and Social Psychology* **62**: 1036–49
- Swann W B Jr., Griffin J J, Predmore S, Gaines E 1987 The cognitive-affective crossfire: When self-consistency confronts self-enhancement. *Journal of Personality and Social Psychology* **52**: 881–89
- Woolfolk R L, Novalany J, Gara M A, Allen L A, Polino M 1995 Self-complexity, self-evaluation, and depression: An examination of form and content within the self-schema. *Journal of Personality and Social Psychology* **68**: 1108–20

M. R. Leary

Personality and Crime

Psychology assumes that behavior arises from the interaction between the individual and the environment. Studies show that behavior is remarkably consistent over time; or, to be more precise, the relative ordering of individuals is remarkably consistent over time (Roberts and Del Vecchio 2000). It is assumed that behavioral consistency depends primarily on the persistence of underlying tendencies to behave in particular ways in particular situations. These tendencies are termed personality traits, such as impulsiveness, excitement seeking, assertiveness, modesty, and dutifulness. Larger personality dimensions such as extraversion refer to clusters of personality traits.

Many constructs assumed to underlie behavior are not included under the heading of personality. This is true, for example, of intelligence and moral reasoning abilities, and of cognitive processes (thinking, reasoning, and decision making). However, the boundaries between personality and all these constructs are not well defined. This review will focus on the topics conventionally included under the heading of personality, and there is not space here to discuss underlying biological processes in any detail.

The main types of offending studied are: burglary, theft, violence, vandalism, fraud, and drug abuse. Since crimes are a subset of antisocial acts, any potential to commit crimes is probably part of a broader potential

to commit antisocial acts, but the focus here is on crimes. The focus is also on normal personality, and as far as possible, discussions of related psychiatric conditions such as antisocial personality disorder, psychopathy, and attention deficit–hyperactivity disorder will be avoided. However, the boundary between normal and pathological behavior is not always clear, depending on judgments about how far normal functioning is impaired.

In studying the relationship between personality and crime, it is important to avoid tautological statements and tautological results. Any construct that is presumed to cause crime must measure something different from crime. Thus, it is useful to investigate how far low empathy causes crime but less useful to investigate how far aggressiveness causes violent crime or how far an antisocial personality causes crime. Early research using the psychopathic deviate scale of the Minnesota Multiphasic Personality Inventory and the socialization scale of the California Psychological Inventory was essentially tautological in showing that antisocial behavior was related to crime.

A bewildering number of personality traits and dimensions have been postulated, measured, and related to offending. In some cases, different terms are probably used to describe the same underlying construct. Only the most important topics can be reviewed here: temperament, impulsiveness, empathy, the Eysenck theory, and the ‘Big Five’ personality dimensions.

1. Temperament

Temperament is basically the childhood equivalent of personality, although there is more emphasis in the temperament literature on constitutional predisposition and genetic and biological factors. The modern study of child temperament began with the New York Longitudinal Study of Chess and Thomas (1984). Children in their first 5 years of life were rated on temperamental dimensions by their parents, and these dimensions were combined into three broad categories of easy, difficult and ‘slow to warm up’ temperament. Having a difficult temperament at age 3–4 years (frequent irritability, low amenability and adaptability, irregular habits) predicted poor adult psychiatric adjustment at age 17–24.

Remarkably, Bates (1989) found that mothers’ ratings of difficult temperament as early as age 6 months (defined primarily as frequent, intense expressions of negative emotions), predicted mothers’ ratings of child conduct problems between the ages of 3 and 6 years. Similar results were obtained in the Australian Temperament Project, which found that children who were rated as irritable, not amenable, or showing behavior problems at age 4–8 months tended to be rated as aggressive at age 7–8 years (Sanson et al.

1993). However, when information at each age comes from the same source, it is possible that the continuity lies in the rater, rather than the child. Fortunately, other studies (e.g., Guerin et al. 1997) show that difficult temperament in infancy, rated by mothers, also predicts antisocial behavior in childhood rated by teachers.

Because it was not very clear exactly what a 'difficult' temperament meant in practice, other researchers have investigated more specific dimensions of temperament. For example, Kagan (1988) and his colleagues in Boston classified children as inhibited (shy or fearful) or uninhibited at age 21 months on the basis of their observed reactions to a strange situation, and found that they remained significantly stable on this classification up to age 7 years. Furthermore, the uninhibited children at age 21 months significantly tended to be identified as aggressive at age 13 years, according to self and parent reports (Schwartz et al. 1996).

The most important results on the link between childhood temperament and later offending have been obtained in the Dunedin longitudinal study in New Zealand, which has followed up over 1000 children from age 3 years and into their 20s (Caspi 2000). Temperament at age 3 years was rated by observing the child's behavior during a testing session involving cognitive and motor tasks. The most important dimension of temperament was being undercontrolled (restless, impulsive, with poor attention), which predicted aggression, self-reported delinquency, and convictions at age 18–21 years.

2. *Impulsiveness*

The most important study of childhood temperament in relation to offending essentially identified impulsiveness as the key dimension, and it is generally true that impulsiveness is the most crucial personality dimension that predicts offending. Unfortunately, there are a bewildering number of constructs referring to a poor ability to control behavior. These include impulsiveness, hyperactivity, restlessness, clumsiness, not considering consequences before acting, a poor ability to plan ahead, short time horizons, low self-control, sensation-seeking, risk-taking, and a poor ability to delay gratification. Virtually all these constructs, measured in different ways, are consistently related to measures of offending (Blackburn 1993, pp. 191–196).

In a longitudinal study of over 400 London males from age 8 years to age 40 years, boys nominated by teachers as lacking in concentration or restless; those nominated by parents, peers, or teachers as the most daring or risk-taking; and those who were the most impulsive on psychomotor tests at age 8–10 years, all tended to become offenders later in life. Later self-report measures of impulsiveness were also related to

offending. Daring, poor concentration, and restlessness all predicted both official convictions, and self-reported delinquency, and daring was consistently one of the best independent predictors (Farrington 1992).

The most extensive research on different measures of impulsiveness was carried out in another longitudinal study of males (the Pittsburgh Youth Study) by White et al. (1994). The measures that were most strongly related to self-reported delinquency at ages 10 and 13 were teacher-rated impulsiveness (e.g., 'acts without thinking'); self-reported impulsivity; self-reported under-control (e.g., 'unable to delay gratification'); motor restlessness (from videotaped observations), and psychomotor impulsivity. Generally, the verbal behavior rating tests produced stronger relationships with offending than the psychomotor performance tests, suggesting that cognitive impulsiveness (based on thinking processes) was more relevant than behavioral impulsiveness (based on test performance).

There have been many theories put forward to explain the link between impulsiveness and offending. One of the most popular suggests that impulsiveness reflects deficits in the executive functions of the brain, located in the frontal lobes (Moffitt 1990). Persons with these neuropsychological deficits will tend to commit offenses because they have poor control over their behavior, a poor ability to consider the possible consequences of their acts, and a tendency to focus on immediate gratification. There may also be an indirect link between neuropsychological deficits and offending, which is mediated by hyperactivity and inattention in school, and the resulting school failure. In discussing links between executive functions and offending, impulsiveness may be difficult to disentangle from intelligence, although Lynam and Moffitt (1995) argued that they were different constructs. Various biological theories of impulsiveness have also been proposed, focussing on either behavioral activation or behavioral inhibition (e.g., Newman and Wallace 1993).

3. *Empathy*

There is a widespread belief that low empathy is an important personality trait that is related to offending, on the assumption that people who can appreciate and/or experience a victim's feelings are less likely to victimize someone. This belief also underlies cognitive-behavioral-skills training programs that aim to increase empathy. However, its empirical basis is not very impressive. There are inconsistent results in this field, and measures of empathy are not well validated or widely accepted. There appears to be no systematic comparison of different empathy measures analogous to the research of White et al. (1994) with impulsiveness, and no longitudinal study relating early measures of empathy to later offending.

A distinction has often been made between cognitive empathy (understanding or appreciating other people's feelings) and emotional empathy (actually experiencing other people's feelings). Eisenberg et al. (1998) further distinguished between sympathy (feeling concern for another) and experiencing personal distress. Empathy can be measured in many different ways, including self-completed questionnaires, peer ratings and systematic observation. Unfortunately, results vary according to the method used (Miller and Eisenberg 1988).

The best studies of the 1990s that have related empathy to offending in relatively large samples are by Mak (1991), Kaukiainen et al. (1999) and especially Luengo et al. (1994). In Australia, Mak (1991) found that delinquent females had lower emotional empathy than nondelinquent females, but that there were no significant differences for males. In Finland, Kaukiainen et al. (1999) reported that empathy (cognitive and emotional combined) was negatively correlated with aggression (both measured by peer ratings). In Spain, Luengo et al. (1994) carried out the first project that related cognitive and emotional empathy separately to (self-reported) offending, and found that both were negatively correlated.

Some results obtained in research on psychopathy are relevant. The Psychopathy Checklist includes two correlated factors, the first measuring personality and the second measuring an antisocial life-style (Harpur et al. 1988). The first factor measures a cold, callous personality, lacking empathy, affect, guilt and remorse, and engaging in habitual lying and manipulation. This constellation of personality traits is correlated with conduct problems and delinquency. Further research on it is warranted, especially in prospective longitudinal studies.

4. The Eysenck Theory

Before 1990, the best-known research on personality and crime was probably that inspired by Eysenck's theory and personality questionnaires (Eysenck 1996). He viewed offending as natural and even rational, on the assumption that human beings were hedonistic, sought pleasure, and avoided pain. He assumed that delinquent acts, such as theft, violence, and vandalism were essentially pleasurable or beneficial to the offender. In order to explain why everyone was not a criminal, Eysenck suggested that the hedonistic tendency to commit crimes was opposed by the conscience, which was viewed as a fear response built up from childhood in a conditioning process.

On the Eysenck theory, the people who commit offenses are those who have not built up strong consciences, mainly because they have inherently poor conditionability. Poor conditionability is linked to Eysenck's three dimensions of personality, Extraversion (E), Neuroticism (N), and Psychoticism (P).

People who are high on E build up conditioned responses less well, because they have low levels of cortical arousal. People who are high on N also condition less well, because their high resting level of anxiety interferes with their conditioning. Also, since N acts as a drive, reinforcing existing behavioral tendencies, neurotic extraverts should be particularly criminal. Eysenck also predicted that people who are high on P would tend to be offenders, because the traits included in his definition of psychoticism (emotional coldness, low empathy, high hostility, and inhumanity) were typical of criminals. However, the meaning of the P scale is unclear, and it might perhaps be more accurately labeled as psychopathy. Zuckerman (1989) suggested that it should be termed 'impulsive unsocialized sensation-seeking.'

A review of studies relating Eysenck's personality dimensions to official and self-reported offending concluded that high N (but not E) was related to official offending, while high E (but not N) was related to self-reported offending (Farrington et al. 1982). High P was related to both, but this could have been a tautological result, since many of the items on the P scale are connected with antisocial behavior, or were selected in light of their ability to discriminate between prisoners and nonprisoners. In the prospective longitudinal study of London boys, those high on both E and N tended to be juvenile self-reported offenders, adult official offenders and adult self-reported offenders, but not juvenile official offenders. These relationships held independently of other criminogenic risk factors, such as low family income, low intelligence, and poor parental child-rearing behavior. However, when individual items of the personality questionnaire were studied, it was clear that the significant relationships were caused by the items measuring impulsiveness (e.g., doing things quickly without stopping to think). Hence, it seems likely that research inspired by the Eysenck theory essentially confirms the link between impulsiveness and offending.

5. The 'Big Five'

Since 1990, the most widely accepted personality system has been the 'Big Five' or five-factor model (McCrae and Costa 1997). This suggests that there are five key dimensions of personality: Neuroticism (N), Extraversion (E), Openness (O), Agreeableness (A) and Conscientiousness (C). Openness means originality and openness to new ideas, Agreeableness includes nurturance and altruism, and Conscientiousness includes planning and the will to achieve. Openness and Conscientiousness seem related to intelligence, or at least to social or emotional intelligence. These dimensions are measured using a personality inventory called the NEO-PI. Controversially, McCrae et al. (2000) argued that these personality dimensions are biologically based tendencies that follow intrinsic

developmental pathways independently of environmental influences.

Because of its newness, the 'Big Five' personality model has rarely been related to offending. In Canada, Hart and Hare (1994) found that psychopathy was most strongly (negatively) correlated with Agreeableness and Conscientiousness. Similarly, in an Australian study Heaven (1996) showed that Agreeableness and Conscientiousness were most strongly (negatively) correlated with self-reported delinquency. Much the same results were obtained in the Pittsburgh Youth Study when the five dimensions were measured using mothers' ratings (John et al. 1994). It seems likely, therefore, that this pattern of results is replicable.

6. Future Directions

More research is needed on the theoretical linkages between personality constructs and offending and on their biological bases. Prospective longitudinal studies are especially needed in which personality traits are measured early in life and related to later offending, to investigate the relative importance of biological, individual, family, peer, and school influences on personality and crime. Different studies should be compared to investigate how far results are replicable across genders, races, and cultures (Caspi et al. 1994).

A key issue that has rarely been addressed is the relationship between personality dimensions and different types of offenders. It might be expected, for example, that violent offenders would be more impulsive than non-violent offenders. It would be useful to use personality traits to define types of people and hence move from a variable-based to a person-based approach (Krueger et al. 1994).

The main policy implication of research on personality and crime is the use of cognitive-behavioral skills training techniques to try to counteract high impulsivity and low empathy. Research on childhood temperament suggests that potential offenders can be identified early in life and that efforts should be made to change children with uninhibited and uncontrolled temperaments. Research on personality and crime should also be useful in matching types of people to types of interventions. It seems likely that different personality types would interact differently with different criminological risk factors, and hence that interventions should be tailored to personalities. This is the new frontier.

See also: Adulthood: Prosocial Behavior and Empathy; Aggression in Adulthood, Psychology of; Anti-social Behavior in Childhood and Adolescence; Crime and Gender; Crime: Sociological Aspects; Crime, Sociology of; Neuroticism; Personality Theory and Psychopathology; Prosocial Behavior and Empathy;

Developmental Processes; Temperament and Human Development; Violence and Effects on Children

Bibliography

- Bates J E 1989 Applications of temperament concepts. In: Kohnstamm G A, Bates J E, Rothbart M K (eds.) *Temperament in Childhood*. Wiley, Chichester, UK, pp. 321–55
- Blackburn R 1993 *The Psychology of Criminal Conduct*. Wiley, Chichester, UK
- Caspi A 2000 The child is father of the man: Personality continuities from childhood to adulthood. *Journal of Personality and Social Psychology* **78**: 158–72
- Caspi A, Moffitt T E, Silva P A, Stouthamer-Loeber M, Krueger R F, Schmutte P S 1994 Are some people crime-prone? Replications of the personality–crime relationship across countries, genders, races, and methods. *Criminology* **32**: 163–95
- Chess S, Thomas A 1984 *Origins and Evolution of Behavior Disorders: From Infancy to Early Adult Life*. Brunner/Mazel, New York
- Eisenberg N, Fabes R A, Shepard S A, Murphy B C, Jones S, Guthrie I K 1998 Contemporaneous and longitudinal prediction of children's sympathy from dispositional regulation and emotionality. *Developmental Psychology* **34**: 910–24
- Eysenck H J 1996 Personality and crime: Where do we stand? *Psychology, Crime and Law* **2**: 143–52
- Farrington D P 1992 Juvenile delinquency. In: Coleman J C (ed.) *The School Years*, 2nd edn. Routledge, London, pp. 123–63
- Farrington D P, Biron L, LeBlanc M 1982 Personality and delinquency in London and Montreal. In: Gunn J, Farrington D P (eds.) *Abnormal Offenders, Delinquency, and the Criminal Justice System*. Wiley, Chichester, UK, pp. 153–201
- Guerin D W, Gottfried A W, Thomas C W 1997 Difficult temperament and behavior problems: A longitudinal study from 1.5 to 12 years. *International Journal of Behavioral Development* **21**: 71–90
- Harpur T J, Hakstian A R, Hare R D 1988 Factor structure of the psychopathy checklist. *Journal of Consulting and Clinical Psychology* **56**: 741–7
- Hart S D, Hare R D 1994 Psychopathy and the Big Five: Correlations between observers' ratings of normal and pathological personality. *Journal of Personality Disorders* **8**: 32–40
- Heaven P C L 1996 Personality and self-reported delinquency: Analysis of the 'Big Five' personality dimensions. *Personality and Individual Differences* **20**: 47–54
- John O P, Caspi A, Robins R W, Moffitt T E, Stouthamer-Loeber M 1994 The 'Little Five': Exploring the nomological network of the five-factor model of personality in adolescent boys. *Child Development* **65**: 160–78
- Kagan J, Reznick J S, Snidman N 1988 Biological bases of childhood shyness. *Science* **240**: 167–71
- Kaukiainen A, Bjorkvist K, Lagerspetz K, Osterman K, Salmivalli C, Rothberg S, Ahlborn A 1999 The relationships between social intelligence, empathy, and three types of aggression. *Aggressive Behavior* **25**: 81–9
- Krueger R F, Schmutte P S, Caspi A, Moffitt T E, Campbell K, Silva P A 1994 Personality traits are linked to crime among men and women: Evidence from a birth cohort. *Journal of Abnormal Psychology* **103**: 328–38

- Luengo M A, Otero J M, Carrillo-de-la-Pena M T, Miron L 1994 Dimensions of antisocial behavior in juvenile delinquency: A study of personality variables. *Psychology, Crime and Law* 1: 27–37
- Lynam D R, Moffitt T E 1995 Delinquency and impulsivity and IQ: A reply to Block. *Journal of Abnormal Psychology* 104: 399–401
- Mak A S 1991 Psychosocial control characteristics of delinquents and nondelinquents. *Criminal Justice and Behavior* 18: 287–303
- McCrae R R, Costa P T 1997 Personality trait structure as a human universal. *American Psychologist* 52: 509–16
- McCrae R R, Ostendorf F, Angleitner A, Hrebickova M, Avia M D, Sanz J, Sanchez-Bernardos M L, Kusdil M E, Woodfield R, Saunders P R, Smith P B 2000 Nature over nurture: Temperament, personality, and life span development. *Journal of Personality and Social Psychology* 78: 173–86
- Miller P A, Eisenberg N 1988 The relation of empathy to aggressive and externalizing/antisocial behavior. *Psychological Bulletin* 103: 324–44
- Moffitt T E 1990 The neuropsychology of juvenile delinquency: A critical review. In: Tonry M, Morris N (eds.) *Crime and Justice*. University of Chicago Press, Chicago, Vol. 12, pp. 99–169
- Newman J P, Wallace J F 1993 Diverse pathways to deficient self-regulation: Implications for disinhibitory psychopathology in children. *Clinical Psychology Review* 13: 699–720
- Roberts B W, Del Vecchio W F 2000 The rank-order consistency of personality traits from childhood to old age: A quantitative review of longitudinal studies. *Psychological Bulletin* 126: 3–25
- Sanson A, Smart D, Prior M, Oberklaid F 1993 Precursors of hyperactivity and aggression. *Journal of the American Academy of Child and Adolescent Psychiatry* 32: 1207–16
- Schwartz C E, Snidman N, Kagan J 1996 Early childhood temperament as a determinant of externalizing behavior in adolescence. *Development and Psychopathology* 8: 527–37
- White J L, Moffitt T E, Caspi A, Bartusch D J, Needles D J, Stouthamer-Loeber M 1994 Measuring impulsivity and examining its relationship to delinquency. *Journal of Abnormal Psychology* 103: 192–205
- Zuckerman M 1989 Personality in the third dimension: A psychobiological approach. *Personality and Individual Differences* 10: 391–418

D. P. Farrington and D. Jolliffe

Personality and Health

1. Associations between Personality and Health: Background

It is common folk wisdom that people who worry too much are likely to develop headaches, that nervous people are prone to ulcers, that excitable people are likely to develop skin rashes, and that overly sensitive people may be prone to nausea or fainting. Popular wisdom also has it that workers who strive compulsively for long hours at frustrating job tasks are more likely to have heart attacks, and that people who

bottle up all their feelings inside them have an increased risk of cancer. Yet the scientific evidence suggests a different and more complex picture. Further, although there are indeed associations between personality and health, the causal links are not always what they first appear to be. This is important because the causal linkages tell us which interventions are helpful and which are worthless or even possibly harmful.

A significant ethical danger in studying the relations of personality and health is that individuals may be unfairly blamed for their health problems. Health and illness are complex states, usually influenced by a wide variety of biological, environmental, and psychosocial factors. Personality—the traits, motives, emotions, and behavioral patterns unique to each individual—is sometimes associated with illness, but it would be a grave oversimplification to think that the individual has direct responsibility for his or her health. On the other hand, individual behavior patterns sometimes do affect health-relevant processes, and so the person, the family, the environment, and the culture can sometimes promote the likelihood of better health.

1.1 Roots in Ancient Greece

The idea that personality and health might be related dates back thousands of years and clearly appears in the writings of Hippocrates, Galen, and their followers. The ancient Greeks, keen observers of the human condition, saw four elements, the so-called bodily humors, as key to both individuality and health. People with a healthy supply of blood might turn out to be sanguine and ruddy. Black bile (or melancholy) might lead to depression and degenerative diseases or cancer. Yellow bile (or choler) could produce a bitter, angry personality and associated diseases, and phlegm is characteristic of a phlegmatic, cold apathy. It is interesting that this scheme of four emotional aspects of personality and health was sufficiently perceptive (or ambiguous) to influence the practice of medicine for 2,000 years. Of course, notions of bodily humors have been discarded. Yet the idea that individuals can be categorized as sanguine, depressed, hostile, and repressed continues to stay with us. Modern personality and health psychology takes these chronic psycho-emotional states and combines them with a more sophisticated understanding of traits to study the associations among personality and health.

1.2 Type A Behavior

Much of the current research interest in personality and health began with theory and research on the Type A Behavior Pattern (Houston and Snyder 1988). Early in the twentieth century, a number of prominent physicians including William Osler began noting that

certain men seemed especially prone to angina and coronary heart disease. The at-risk characteristics supposedly included being ambitious, hard-working, dominant, and aggressive. In the 1930s and 1940s, the influence of psychoanalysis on psychosomatic medicine led to numerous speculations about the role of expression and repression of hidden conflicts and anger on the development of heart disease. But it was not until the mid-century writings on Type A behavior that empirical research on this topic began in earnest.

The term 'Type A' was chosen by its inventors Meyer Friedman and Ray Rosenman (e.g., Rosenman et al. 1975) because it was a nonconnotative, 'objective' designation for a so-called coronary-prone medical syndrome. Type A people were seen as struggling rapidly to accomplish many things in a short period of time, always with a deadline to meet. They were viewed as very competitive, active, and quite hostile or aggressive. Although any person can act this way from time to time, Type A people were seen as constantly driven. Individuals who did not fit this profile were labeled, by default, as Type B. That is, there was thus no independent definition of what a healthy personality style might be.

Although many hundreds of studies have been conducted on Type A behavior, the results have often been mixed and confusing. Although the initial aim was to simplify the concept by choosing a neutral term ('Type A') and avoiding related psychological concepts and theories, the characteristics of individuals cannot be adequately explained in such a sterile manner. That is, psychologists soon turned to trying to understand the trait correlates of Type A behavior, the components of Type A behavior, the basis of Type A behavior, the consequences of Type A behavior, and the health aspects of Type A behavior, even beyond coronary disease (Friedman 1992, Houston and Snyder 1988, Miller et al. 1996).

It is now clear that the fast speed and the job-involvement aspects are not important predictors of coronary disease. However, excessive competitiveness and constant hostility seem, for some people in some circumstances, to increase the likelihood of coronary heart disease. It is not yet clear the extent to which these characteristics also raise the risk of all-cause mortality, since there has been insufficient attention to construct validity.

The idea of Type A behavior proved to be an important heuristic one, leading to much empirical research since the 1970s about whether, when, why, and how personality plays a role in health. There are, however, a variety of possible mediating mechanisms.

2. Ways that Personality is Related to Health: Causal Links

There are many different reasons why personality is associated with health. Understanding the causal

pathways is very important, because an interest in this subject usually derives from a desire to improve or promote health. However, it turns out that personality can affect health, that health can affect personality, or that the association between them can be the result of several sorts of artifacts.

2.1 Symptom Reporting and Medical Usage

It is well documented that personality is related to symptom reporting, symptom sensation, and entering the medical care system. Certain people, especially people who are neurotic, with anxiety and stress, are more likely to report symptoms including headache, gastrointestinal distress, and other assorted aches and pains. In other words, if questioned closely, other, non-neurotic people would be experiencing the same symptoms but would not be as likely to talk about them or to report them to their physicians.

It is also the case that people with anxiety and stress are more likely to feel pain and other symptoms. That is, they have a lower threshold for experiencing pain and a lower tolerance for pain. They also may be especially vigilant about bodily sensations (Friedman 1991/2000, Pennebaker 1982).

People with certain psychological traits are also more likely to enter the health care system. These people are generally characterized by such traits as anxiety, depression, dependency, hostility, and moodiness. Women are also generally more likely to seek health care than men. In the extreme, there are people with factitious disorders who seek health care for the rewards that they receive there, including attention, drugs, and shelter (Costa and McCrae 1987, Friedman 1990).

In all of these cases, there are associations between personality and ill health, but the association results from a greater tendency to report and perceive symptoms and seek treatment rather than from a greater incidence of organic disease. In other words, in many cases, certain people are more likely to report being 'ill,' but they do not differ much from other people in their incidence of clear organic diseases like heart disease or cancer.

2.2 Disease-caused Personality Changes

Many associations between personality and health are the result of disease processes that affect health. Certainly any disease with a significant impact on the brain, including Alzheimer's, Parkinson's, strokes, syphilis, AIDS, and brain tumors can produce associations between personality and health. Since the changes in personality often become apparent long before the changes in organic brain function are examined and diagnosed, it may appear that the personality is predicting or even causing the disease

when the reality is that the disease is changing the personality (Woodward 1998).

In many people, the processes are likely even more subtle and complex. Various illnesses affect the blood supply to the brain, the neurotransmitters in the brain, and communication from other parts of the body to the brain (Woodward 1998). This may result in various associations between personality and health, especially involving depression or other mood changes. For example, low levels of thyroid hormones or lessened oxygen in the blood can cause depression. Importantly, many if not most prescription and non-prescription drugs, being used to treat illness, can affect the brain and thus affect mood and personality, sometimes dramatically, sometimes subtly.

2.3 Temperamental Underlying Third Variables

Some of the associations between personality and health result from a constitutional predisposition underlying certain aspects of personality and certain threats to health. For example, some people's cardiovascular systems are hyper-reactive, meaning that their blood pressure and heart rate may show marked increases when the individual is under stress. As part of this biology, such people are more likely to be introverted and neurotic (aspects of personality) and are also more likely to suffer diseases such as hypertension. In such cases of an underlying third variable, it is not necessarily the case that efforts to change personality (such as helping an introvert become more sociable) will have any effect on health risk if the underlying physiology is not altered (Eysenck 1991, Friedman 1990, Zuckerman 1991).

Similarly, a hostile temperament, which may result from a combination of genetics, early hormone exposures, and early experiences, could lead to an association between a choleric personality and health risks. Hans Eysenck has argued that such psychobiological systems produce many associations between individual characteristics and disease, but also that the psychobiological characteristics are modifiable, thus producing changes in both personality and disease risk.

2.4 Personality-caused Illness

Probably the most interesting link between personality and health arises when personality has a fairly direct, causal influence on health. For example, chronic worrying and distress somehow may increase the likelihood of headaches. Since we cannot randomly assign people to personality and do conclusive experiments, inferences about how personality affects health must necessarily be inexact. It is important to employ a wide variety of studies ranging from epidemiological studies of psychological patterns and health down to

biological studies of the physiological correlates of personality. The findings must then be pieced together.

3. Explanations of Personality-caused Illness

In those cases in which personality affects health, it is useful to distinguish among several primary causal pathways. However, given that these influences occur across time, it is often the case that these mechanisms do not function independently; rather they feed back on and interact with each other (Antonovsky 1979, Bolger and Zuckerman 1995, Friedman 1991/2000).

3.1 Poor Coping

Personality is related to coping, the ways in which an individual contends with stress. New thoughts and behaviors are constantly necessary to adapt to new challenges. With successful coping, the individual interprets and understands the challenge, draws on helpers to address it, and takes appropriate actions. Depending on the challenge, individuals who are depressed, introverted, repressed, unconscientious, vain, or conflicted may be less successful in bringing to bear necessary psychological, social, and behavioral resources (Aspinwall and Taylor 1997, Friedman 1991/2000).

A common issue involves relying on unproductive psycho-emotional ways of coping when a behavioral action could be used to solve the problem or source of stress. For example, a student might go party rather than study to deal with the challenge of an upcoming major exam. On the other hand, an individual might attempt to change the unchangeable. For example, after injury or loss, a person may need to find psychological ways of dealing with his or her new situation in order to cope effectively. Poor coping can lead either to unhealthy behaviors or unhealthy physiological reactions, or both (Maddi and Kobasa 1984, Pennebaker 1995).

3.2 Unhealthy Behaviors

It is well documented that certain personalities in certain situations are more likely to engage in unhealthy behaviors such as smoking, drinking, drug abuse, fast driving, fighting, and so on. Certain people are also less likely to engage in prophylactic measures such as using sunscreen, wearing seat belts, using condoms, brushing teeth, regularly washing hands, and so on. Some individuals are more or less likely to visit physicians, keep follow-up appointments, follow medical regimens, or have screening tests. Since these associations usually depend on at least several factors about the individual, the situation, and the unhealthy

activity, it is not usually possible to offer simple, broad generalizations.

3.3 Stress and Physiological Reactivity

Psychophysiological mechanisms often come to mind when one first considers links between personality and health. This is the idea that personality and stress directly impact internal physiological processes, thus causing disease. A more general view here is that the disruption upsets bodily homeostasis, thus allowing disease to take hold.

Much of this work dates back to Walter Cannon's writings on the 'fight-or-flight' response in which catecholamines are released and affect organs and systems throughout the body. Catecholamines (such as epinephrine) can affect processes from heart rate and blood pressure to the metabolism of fats. More recently, it has been established that the immune system is likewise affected by stress hormones. It is well documented that certain people are more likely than others to react to challenge with significant psychophysiological stress reactions (Denollet 1998, Friedman 1992, Kiecolt-Glaser et al. 1998, Zuckerman 1991).

3.4 Tropisms: Seeking Unhealthy Situations

One of the least understood but most intriguing sources of the link between personality and disease involves unhealthy situations. Some individuals gravitate towards healthier and health-promoting psychosocial environments, while other individuals find themselves pulled towards unhealthy environments. In these instances, it is not the case that personality leads directly to unhealthy behaviors or to unhealthy psychophysiological reactions. Rather, the personality facilitates entering the unhealthy situation.

Some of this motivation is derived from genetics, temperament, and early experiences. Other motivation arises from family experiences and consistent peer pressure. People who are impulsive, hostile, psychologically unstable, and have related difficulties in self-regulation are more likely to find themselves in environments that promote smoking, alcoholism, drug abuse, interpersonal violence, promiscuity, and risk-taking hobbies. Furthermore, neuroticism tends to predict to negative life events; in other words, neurotic people are more likely to experience objectively more stressful events (Taylor et al. 1997). Others, who are well socialized, conscientious, and agreeable, are more likely to wind up well educated and surrounded by mature adults. Note that someone who was genetically predisposed to alcoholism or other addiction is very unlikely to develop the health-impairing problem if raised in and exposed to very healthy psychosocial environments.

4. Disease-prone Traits

Although studies rarely assess a variety of personality traits, a variety of likely mediating mechanisms, and a variety of health outcomes, efforts to piece together the results of different types of research yield a sense of which characteristics are more or less likely to lead to good health.

4.1 Extraversion

Various theories suggest that extraversion and its components of dominance, sociability, and talkativeness should be related to good health, but the evidence is sparse. It is certainly the case that people who are well integrated into their communities, have friends, and have good relations with others are more likely to be healthier. Such social support is not, however, necessarily derived from being highly extroverted. Closer analysis reveals many people who are not especially extroverted but who have important, stable social relationships.

4.2 Optimism

Derived from popular notions of the importance of 'positive thinking,' a good deal of research has examined whether optimism is related to health. Optimism is the general expectation that good things will happen. To the extent that optimists may try harder when challenged, take better care of themselves, and cope better with the adversity of illness, optimism may sometimes be a characteristic of good health. For example, dispositional optimists tend to have a faster rate of recovery from coronary bypass surgery (Aspinwall and Taylor 1997, Scheier et al. 1989). Faster recovery with a positive outlook may also be a result of improved immune system functioning as stress is reduced (Kiecolt-Glaser et al. 1998).

However, there are also many reasons why optimism is not health protective and may be health damaging. The other side of the coin of optimism is 'optimistic bias' or unrealistic optimism (Weinstein 1984). This bias involves the faulty belief that negative events are less likely to happen to the optimist than to others. For example, one might agree that 'people' who smoke and who do not eat fruits and vegetables are more likely to get cancer, but not worry at all whether the relationship applies to oneself. Importantly, someone who is too optimistic may not take precautions or may not cooperate fully with medical treatment (Peterson et al. 1998, Tennen and Affleck 1987).

4.3 Neuroticism, Hostility, Depression

It is very well established that individuals who are more anxious, hostile, or depressed are more likely to

report being ill and seek medical care. Much research controversy has arisen over the extent to which and the circumstances under which these are causal mechanisms, with these personality traits causing disease. As noted, such individuals are more likely to feel, report, and treat symptoms. It is also often the case, however, that these symptoms result from true organic disease, promoted by psychological causes.

The best evidence concerns hostility, which was first investigated in an attempt to better understand coronary proneness. Chronic hostility (and chronic arousal of the sympathetic nervous system) raises the risk of disease and interferes with recovery (Miller et al. 1996).

There is also mounting evidence of the role of chronic depression in increasing the likelihood of development of various illnesses. Depression is associated with cortisol dysregulation, which is also linked to disruptions in lipid metabolism. It is not yet clear whether this is closely tied to hostility or is a different mechanism.

The aspect of neuroticism that is characterized by alienation, anxiety, and pessimism can produce various unhealthy effects. Such people can turn to substance abuse or overeating in an attempt to regulate their anxious states. They may develop insomnia. They may avoid interpersonal assistance and medical assistance. This may lead to a cycle of increasing anxiety, isolation, and health problems. On the other hand, someone whose neuroticism leads to a vigilance about avoiding threats, attending to symptoms, and cooperating with treatment may gain some health benefits.

Some researchers focus more generally on chronic negative affect, which includes both emotional distress like anxiety and depression as well as social interaction such as introversion, alienation, and the inability to express oneself to others. This distressed personality is sometimes referred to as 'Type D' (Denollet 1998).

4.4 Conscientiousness and Impulsivity

With significant attention to variables such as chronic hostility and Type A competition, it is perhaps not surprising that certain key aspects of a healthy personality have not been uncovered until recently. There is, however, rapidly growing evidence that conscientiousness and a lack of impulsivity is a key characteristic of healthy individuals. Conscientious people, who are prudent, planning, persistent, and dependable, tend to succeed in various life tasks, including staying healthy.

One major study of individuals who have been followed from childhood until their deaths in adulthood, conducted by Howard Friedman and his colleagues, has repeatedly found evidence that conscientiousness is the key personality predictor of longevity. Interestingly, the protective effect of con-

scientiousness does lessen the likelihood of death from accident and injury, but it is also protective against death from diseases such as cardiovascular disease and cancer. Conscientiousness does not seem to have its effects simply through one or two behavioral pathways, but rather it seems to affect a whole host of threats to health (Friedman et al. 1995). This finding of the importance of conscientiousness has been confirmed by subsequent studies by others which show that conscientiousness is associated with healthy behaviors, better self-care, better cooperation with medical treatment, and risk avoidance.

4.5 Repression, Lack of Verbalization, and Lack of Disclosure

It has long been recognized that among the many people who seem disease-prone, not all are hostile, impulsive, or neurotic. There are some who are too stoic, apathetic, and unemotional. These phlegmatic people, who look cool, calm, and sluggish, may be consciously or unconsciously suppressing or repressing inner conflict or feelings. This phenomenon is not well understood, but such people may appear tense or anxious upon close observation, although they assert friendliness and cooperation. Such people are often alienated or feel powerless.

A behavior pattern involving a repressive or denying coping style, lack of emotional expression, stress and anxiety, and helplessness/hopelessness is sometimes referred to as 'Type C.' A similar pattern in which individuals are unable to express emotion and seem at higher risk of disease progression is sometimes studied under the rubric of 'alexithymia.' There is speculation that such conditions promote the development or the progression of some cancers, but the relationship of this constellation of traits to other illnesses has not been much studied.

James Pennebaker has proposed that inhibition is a form of physiological work that increases the risk of illnesses. On the other hand, people who can talk about emotional experiences seem to set in motion processes to better health. Verbalizing may contribute to cognitive assimilation of stress and/or help people understand and organize their experiences. It may also often lead to increases in levels of social support.

5. Self-healing Personality

Although most theory and research has focused on disease-prone personalities, it is equally important to examine the characteristics that are associated with good health and longevity. These characteristics are not necessarily simply the absence of the disease-prone traits.

The 'Self-healing personality' refers to a healing emotional style involving a match between the in-

dividual and the environment, which maintains a physiological and psychosocial homeostasis, and through which good mental health promotes good physical health. The term was proposed by psychologist Howard S. Friedman in his 1991 book of the same name (Friedman 1991/2000). Self-healing, emotionally balanced people may be calm and conscientious but are responsive, energetic, curious, secure, and constructive. They are also people one likes to be around. People with self-healing personalities have better odds of a longer, healthier life.

Self-healing individuals often appear enthusiastic, and they may move in smooth, synchronized styles. Their relative lack of emotional conflict and anxiety may result in fewer speech disturbances, modulated voice tones, and an emotional consistency. Contrary to some common wisdom, a self-healing style does not involve a lack of challenge. In fact, self-healing individuals are often found in challenging, socially valued careers. They may enjoy stimulating travel because they are curious and enjoy exploration.

Interestingly, self-healing personalities are often similar to the mentally healthy orientations that have been described by humanistic psychologists (Antonovsky 1979, Csikszentmihalyi 1991, Maddi and Kobasa 1984). Humanistic psychologists such as Abraham Maslow thought that they were describing mental health when they wrote about self-actualized people who are concerned with issues of beauty, ethics, and justice. But it turns out that this spontaneous, positive, growth orientation is also associated with physical health. Indeed, it is a false dichotomy to divide physical and mental health in these ways. This positive process is sometimes also referred to by Mihaly Csikszentmihalyi as 'flow.'

5.1 Efficacy and Hardiness

Work by Salvador Maddi and Suzanne Ouellette Kobasa helped provide a framework for thinking about staying healthy in the face of challenge. They call these personality attributes hardiness. First, they suggest that a healthy personality maintains a sense of control. This is not necessarily a wild sense of optimism, but rather a sense that one can control one's own behaviors. Second, there is a commitment to something that is important and meaningful in their lives. This may also involve values and goals. Third, hardy people welcome challenge. For example, they may view change as an exciting challenge to their growth and development. Efficacy and hardiness generally involve a productive orientation, a sense of mastery, and a zest for life (Bandura 1997).

5.2 Salutogenesis

The medical sociologist Aron Antonovsky (1979) proposed an approach termed 'salutogenesis,' a theory

of how and why certain people stay healthy. What is interesting about this approach is that it does not postulate a sense of control as a primary or even necessary characteristic of a healthy personality. Rather, Antonovsky refers to a 'sense of coherence' which is central to successful coping with challenge. Coherence means that the world is understandable, manageable, and meaningful. In this scheme, someone who believes that nature or a divine force is in charge of the world could have a very healthy orientation, even though the individual does not have a personal sense of control. There is so far only scattered evidence to support this promising approach.

6. Conclusion

In sum, certain individuals are predisposed to be vulnerable to physiological, psychosocial, and interpersonal disturbances and stresses due to a combination of genetics, emotional temperament, and early experiences. When such individuals seek out or are placed in environments that bring out, create, or exacerbate these disturbances, chronic stress patterns may result. These may be accompanied by unhealthy behaviors such as substance abuse. When these individuals then encounter microbes, toxins, or excesses such as high-fat diets, the likelihood of illness increases significantly. On the other hand, other individuals have developed processes of self-healing, which build on themselves as healthier emotional styles, physiological reactions, behaviors, and social interactions become mutually reinforcing.

Overall, there is good evidence that individual characteristics are sometimes tied to health through the biological, psychological, and social aspects of personality, and there is some evidence that focused interventions can have short-term salutary effects on health. But relatively little is understood about the precise causal inter-relations of physiology, coping, social influence, and health behavior, and little is known about the extent to which a focus on personality alone can have substantial effects on various significant health outcomes. It will be important in the future to unravel how these key processes come together and affect each other, both across short time periods and across the lifespan.

See also: Biopsychology and Health; Cancer-prone Personality, Type C; Control Beliefs: Health Perspectives; Depression, Hopelessness, Optimism, and Health; Emotions and Health; Explanatory Style and Health; Health: Self-regulation; Personality Assessment; Personality Psychology; Personality Theory and Psychopathology; Self-efficacy and Health

Bibliography

Antonovsky A 1979 *Health, Stress and Coping*. Jossey-Bass, San Francisco

- Aspinwall L G, Taylor S E 1997 A stitch in time: Self-regulation and proactive coping. *Psychological Bulletin* **121**: 417–36
- Bandura A 1997 *Self-efficacy: The Exercise of Control*. Freeman and Co., New York
- Bolger N, Zuckerman A 1995 A framework for studying personality in the stress process. *Journal of Personality and Social Psychology* **69**(5): 890–902
- Costa P T, McCrae R R 1987 Neuroticism, somatic complaints, and disease: Is the bark worse than the bite? *Journal of Personality* **55**: 299–316
- Csikszentmihalyi M 1990 *Flow: The Psychology of Optimal Experience*. HarperCollins, New York
- Denollet J 1998 Personality and coronary heart disease: The Type-D Scale-16 (DS16). *Annals of Behavioral Medicine* **20**(3): 209–15
- Eysenck H J 1991 *Smoking, Personality, and Stress: Psychosocial Factors in the Prevention of Cancer and Coronary Heart Disease*. Springer-Verlag, New York
- Friedman H S (ed.) 1990 *Personality and Disease*. Wiley and Sons, New York
- Friedman H S 1991/2000 *Self-healing Personality: Why Some People Achieve Health and Others Succumb to Illness*. Henry Holt, New York/(www.iuniverse.com), 2000
- Friedman H S (ed.) 1992 *Hostility, Coping, and Health*. American Psychological Association, Washington, DC
- Friedman H S, Booth-Kewley S 1987 The ‘disease-prone personality’: A meta-analytic view of the construct. *American Psychologist* **42**: 539–55
- Friedman H S, Tucker J S, Schwartz J E, Tomlinson-Keasey C, Martin L R, Wingard D L, Criqui M H 1995 Psychosocial and behavioral predictors of longevity: The aging and death of the ‘Termites.’ *American Psychologist* **50**: 69–78
- Houston B K, Snyder C R 1988 *Type A Behavior Pattern: Research, Theory and Intervention*. Wiley, New York
- Kiecolt-Glaser J K, Page G G, Marucha P T, MacCallum R C, Glaser R 1998 Psychological influences on surgical recovery. *American Psychologist* **53**: 1209–18
- Maddi Salvatore R, Kobasa S C 1984 *The Hardy Executive: Health Under Stress*. Dow Jones-Irwin, Homewood, IL
- Miller T Q, Smith T W, Turner C W, Guijarro M L, Hallet A J 1996 Meta-analytic review of research on hostility and physical health. *Psychological Bulletin* **119**: 322–48
- Pennebaker J W (ed.) 1995 *Emotion, Disclosure, and Healing*. American Psychological Association, Washington, DC
- Pennebaker J W 1982 *The Psychology of Physical Symptoms*. Springer-Verlag, New York
- Peterson C, Seligman M E P, Yurko K H, Martin L R, Friedman H S 1998 Catastrophizing and untimely death. *Psychological Science* **9**: 127–30
- Rosenman R H, Brand R J, Jenkins C D, Friedman M, Straus R, Wurm M 1975 Coronary heart disease in the Western Collaborative Group Study: Final follow-up experience of 81/2 years. *Journal of the American Medical Association* **233**: 872–7
- Scheier M F, Magovern G J, Abbott R A., et al 1989 Dispositional optimism and recovery from coronary artery bypass surgery: The beneficial effects on physical and psychological well-being. *Journal of Personality and Social Psychology* **57**: 1024–40
- Taylor S E, Repetti R L, Seeman T 1997 Health psychology: What is an unhealthy environment and how does it get under the skin? *Annual Review of Psychology* **48**: 411–47
- Tennen H, Affleck G 1987 The costs and benefits of optimistic explanations and dispositional optimism. *Journal of Personality* **55**: 377–93
- Weinstein N D 1984 Why it won’t happen to me: Perceptions of risk factors and susceptibility. *Health Psychology* **3**: 431–57
- Wiebe D J, Smith T W 1997 Personality and health: Progress and problems in psychosomatics. In: Hogan R, Johnson J, Briggs S (eds.) *Handbook of Personality Psychology*. Academic Press, San Diego, CA, pp. 891–918
- Woodward J L 1998 Dementia. In: Friedman H S (ed.) *Encyclopedia of Mental Health*. Academic Press, San Diego, Vol. 1, pp. 693–713
- Zuckerman M 1991 *Psychobiology of Personality*. Cambridge University Press, New York

H. S. Friedman

Personality and Marriage

The success or failure of a marriage is likely to depend, at least in part, on the personality of the individual spouses, broadly defined as the set of enduring traits and characteristics that each spouse brings to the relationship. Three questions have motivated most research on the role of personality in marriage. First, what personality traits are most relevant to predicting marital outcomes? Second, does similarity or dissimilarity between partners’ personalities affect marital outcomes? Third, through what mechanisms does the personality of each spouse affect the relationship? This article reviews each of these areas of research in turn.

1. Which Personality Traits Affect Marital Outcomes?

The earliest efforts to predict marital outcomes were guided by the belief that ‘a large proportion of incompatible marriages are so because of a predisposition to unhappiness in one or both of the spouses’ (Terman 1938, p. 110). To explore the support for this belief, the first marital researchers conducted two kinds of studies: cross-sectional studies, in which lengthy personality inventories were administered to samples that included happily married and distressed couples, and longitudinal studies, in which personality inventories were administered to engaged couples whose marital quality was assessed several years later. Early marital researchers were more interested in determining whether spouses’ personalities accounted for marital outcomes than in identifying which aspects of personality were associated with marital outcomes. Thus, the results of these initial studies were frequently reported in terms of a global personality score rather than divided into specific personality traits. Nevertheless, when Burgess and Wallin (1953), in the first review of this literature, concluded that success and failure in marriage could be predicted by the person-

alities of each spouse, they noted that personality traits associated with neuroticism (e.g., irritability, moodiness, ill-temper) appeared to be especially predictive. Spouses who reported an enduring tendency to experience negative affect across a variety of contexts seemed to be at greater risk of developing marital distress.

As these results accumulated, the same researchers noticed that other enduring characteristics of the spouses (e.g., childhood family environment, sexual history, demographic variables), were also significant predictors of marital outcomes. Once statistical techniques for conducting multiple regression became widely available, a second wave of research examined the ability of personality traits to account for marital outcomes relative to these other sources of influence. Perhaps the most impressive of these studies is that of Kelly and Conley (1987), who followed a sample of 300 couples for over 40 years. At the initial phase of data collection, engaged partners provided self-reports of their early family environment, attitudes towards marriage, sexual history, and stressful life events. In addition, the researchers contacted five acquaintances of each partner and asked these individuals to rate the personality of the subject, a substantial improvement over the self-report inventories relied on by most prior (and subsequent) studies of personality. When all of these variables were examined at the same time, only the personality ratings emerged as significant predictors of marital stability. Specifically, compared to marriages that remained intact, couples whose marriages dissolved over the course of the study were characterized by higher levels of neuroticism in both spouses and higher levels of impulsivity in husbands.

In the late 1980s, the development of the five-factor model of personality briefly promised to expand the focus of marital research beyond the negative aspects of spouses' personalities. According to this model, the many traits that had been studied in prior research on personality could be adequately summarized by five relatively independent dimensions: neuroticism, impulsivity, extraversion, agreeableness, and conscientiousness. Emerging support for this model lead marital researchers to ask whether dimensions of the 'Big Five' besides neuroticism and impulsivity may also be associated with marital outcomes. Thus, for example, Kurdek (1993), in a five-year study of newlywed couples, examined the ability of each of the five personality dimensions to predict the quality and stability of marriages over time. This study did find that wives' conscientiousness was weakly but positively associated with the stability of the marriage over five years. However, echoing the results of previous research, far larger effects were obtained for the neuroticism of both spouses. Karney and Bradbury (1995), in a meta-analysis of longitudinal research on personality and marriage, reported that of all of the Big Five personality dimensions, only neuroticism had demonstrated consistent effects on marriage across

multiple studies. As the first marital researchers suggested, a spouse's 'predisposition to unhappiness' appears to predict decline and instability in marriage more reliably than other personality traits.

2. Similarity and Complementarity

As research on personality in marriage has developed, some have argued that the content of each spouse's personality may be less important to understanding marital outcomes than the degree of similarity between spouses' personalities. The assumption of this perspective is that spouses who share similar personalities should experience fewer opportunities for conflict and thus decreased risk of marital dissolution. Bentler and Newcomb (1978) offered some support for this view in a longitudinal study of newlywed couples. Although each spouse's level of extraversion had no independent association with their marital outcomes, the correlation between husbands' and wives' extraversion scores was significantly higher among couples who remained married than among couples who divorced.

Despite the intuitive appeal of this perspective, the accumulation of further support has been limited by a number of difficulties in assessing the effects of similarity in couples. First, the effects of similarity on a given trait are rarely independent of the effects of the trait itself. For example, Kurdek (1993) found that the difference between husbands' and wives' levels of neuroticism was greater in marriages that dissolved than in marriages that remained intact. This finding may support the role of personality similarity in marriage, but it may also result from the fact that differences between spouses are likely to be larger in couples where levels of neuroticism in either partner are higher overall. Estimating the unique effects of similarity between spouses requires that the effects of each spouse's personality be accounted for first. This is a control that has rarely been implemented.

A second difficulty with this kind of research is the fact that, as Kenny and Acitelli (1994) pointed out, 'A couple may appear to be very similar in their responses, yet they may be no more similar than a man and a woman who are totally unacquainted' (p. 419). Because there may be a typical way that most people tend to respond to personality inventories, the personalities of randomly matched people are likely to be similar to some degree. Estimating the unique degree of similarity in a given marriage requires that the 'typical cultural response' be controlled, but again this is a control that has rarely been implemented.

When similarity between spouses has been estimated accurately, support for beneficial effects of similar personalities has been sparse. For example, Russell and Wells (1991) reported that after controlling for the independent effects of each spouse's responses to a personality inventory, the association between marital satisfaction and the differences between spouses'

responses was rarely significant. Furthermore, the overall association between spouses' personalities tends to be weak (Buss 1984), even though spouses tend to be highly similar to each other on many other dimensions. In general, this research suggests that it may be spouses' similarity on those other dimensions, and in particular the degree to which spouses are similar in their values and attitudes, that may be more relevant to account for marital outcomes. With respect to personality, the nature of each spouse's personality appears to be more important than the degree of similarity between them.

In contrast to the similarity perspective, a competing view suggests that it is the degree of complementarity in spouses' personalities that is key to the successful marriage. For example, Winch (1958) argued that spouses select each other to resolve inadequacies in their own characters. Thus, extraverted people should be happiest with introverted partners and domineering people should be happiest with submissive partners. Although this is also an intuitively appealing idea, to date all attempts to obtain empirical support for the complementarity hypothesis have failed.

3. How Does Personality Affect Marital Outcomes?

Having established that personality traits, especially those associated with neuroticism, predict marital outcomes, the current focus of research on personality and marriage is to explain how this association comes about. Caspi et al. (1989), in their work on the effects of personality throughout the life course, proposed a broad framework that may be applied to understanding this issue. These researchers described two kinds of processes through which personality may affect development throughout the life course. 'Cumulative continuity' refers to all of the ways that individuals' personalities channel them towards certain environments and opportunities. 'Interactional continuity' refers to all of the ways that individuals' personalities affect their behavior towards and cognitions about other people. There is substantial evidence that both of these processes may account for the effects of spouses' personalities on their marriages.

3.1 Cumulative Continuity in Marriage

From early childhood, the personality of an individual affects not only the environments selected by the individual but also the opportunities available to the individual. Thus, for example, children who are prone to negative moods may choose career paths that are compatible with their temperament, but their choices may also be restricted by the behavioral consequences of their temperament (e.g., poorer educational at-

tainment, lower professional accomplishment). In this way, individuals' personalities may have cumulative consequences that greatly shape the context of their adult lives. The nature of their adult environments may in turn affect the quality of individuals' marital relationships. For example, women who experience chronic depression as adolescents tend to marry at a younger age compared to women who do not experience chronic depression (Gotlib et al. 1998). Age at marriage is associated with risks of divorce, such that people who marry at a younger age are at higher risk for divorce than people who marry later in life. Thus, the stable tendency to experience negative affect in childhood appears to be associated, at least for women, with a life course that is detrimental to marriage.

In a similar vein, the personality characteristics of each spouse may give rise to stressful events and circumstances that impact the relationship. For example, individuals who score higher on measures of neuroticism describe events in their lives as more stressful than individuals scoring lower on such measures (Marco and Suls 1993). Furthermore, controlling for these reporting effects, higher levels of neuroticism appear to be associated with greater numbers of stressful life events (Poulton and Andrews 1992). To explain such findings, some researchers have suggested that individuals who are prone to experiencing negative affect frequently generate stressful circumstances that then serve to maintain their negative orientations. As a result of these processes, the marriages of individuals scoring higher in neuroticism may develop in more challenging circumstances than the marriages of less neurotic individuals. Because challenging circumstances, such as unemployment and financial difficulty, have been shown to predict divorce, this may be another route through which the cumulative consequences of each spouse's personality contribute to marital success and failure.

3.2 Interactional Continuity in Marriage

In addition to its role in shaping the circumstances of the marriage, personality may also affect interpersonal processes occurring within the marriage. Spouses' expectations for each other, the behaviors they exchange during their interactions, and the way they interpret those behaviors, have all been associated with the stable dispositions of each spouse. Through these mechanisms, the personality of each member of the couple may give rise to self-fulfilling prophecies, such that stable personality traits predict spouses' feelings towards each other and thus lead spouses to act in ways that confirm or enhance those feelings.

Support for these ideas has been especially strong with regard to negative affectivity, that is, neuroticism. For example, spouses who are higher in negative affectivity tend to make less charitable attributions for

their partners' negative behaviors. That is, these spouses are more likely to hold their partners responsible for negative behaviors, whereas spouses who are lower in negative affectivity are more likely to excuse their partners for such behaviors (Karney et al. 1994). Perhaps as a consequence, among newlywed couples, negative affectivity is associated with marital satisfaction, such that spouses who are higher in negative affectivity are likely to evaluate the marriage as a whole more poorly than spouses who are lower in negative affectivity (Karney and Bradbury 1997). The stable tendency to experience negative affect thus appears to affect spouses' interpretations of the marriage at multiple levels, leading to partners that are perceived as less sensitive and marriages that are perceived as less rewarding.

The effects of each spouse's personality are not merely intrapsychic, however. In established marriages, the negative affectivity of husbands is associated with the marital satisfaction of their wives, such that the wives of husbands who are higher in negative affectivity tend to be less satisfied with the marriage, controlling for their own level of negative affectivity (Karney et al. 1994). One way that the personality of one spouse may affect the satisfaction of the other is through the effects of personality on the behaviors that spouses exchange during their interactions with each other. Indeed, in established marriages, neuroticism is associated with couples' problem-solving behavior such that more neurotic spouses report feeling more negative than less neurotic spouses immediately after an attempt to resolve a marital problem (Geist and Gilbert 1996). Furthermore, observational coding of spouses' behavior indicates that more neurotic wives express more negative emotions during conflict discussions. Given that the quality of couples' problem-solving behaviors predicts stability and change in the quality of the marriage over time (Karney and Bradbury 1997), the link between neuroticism and marital interactions may be one way that the personality of an individual may influence the development of marital quality for a couple.

4. Future Directions

Although research on the interactional consequences of personality has focused almost exclusively on behavior during marital problem-solving interactions, personality may play an even more prominent role in the way spouses solicit and provide each other with social support. To the extent that personality, and in particular negative affectivity, affects the way spouses treat each other and the nature of the stressful events they encounter, then research on the provision of social support in response to stressful events would be a direction for future research that promises to

integrate the cumulative and interactional consequences of personality in marriage.

See also: Adult Psychological Development: Attachment; Attributional Processes: Psychological; Lifespan Development, Theory of; Marriage; Neuroticism; Partner Selection across Culture, Psychology of; Personality and Adaptive Behaviors; Personality and Social Behavior; Personality Theories; Self-fulfilling Prophecies; Temperament and Human Development

Bibliography

- Bentler P M, Newcomb M D 1978 Longitudinal study of marital success and failure. *Journal of Consulting and Clinical Psychology* **46**: 1053–70
- Burgess E W, Wallin P 1953 *Engagement and Marriage*. Lippincott, Chicago
- Buss D M 1984 Marital assortment for personality dispositions: Assessment with three different data sources. *Behavior Genetics* **14**: 111–23
- Caspi A, Bem D J, Elder G H 1989 Continuities and consequences of interactional styles across the life course. *Journal of Personality* **57**: 376–406
- Geist R L, Gilbert D G 1996 Correlates of expressed and felt emotion during marital conflict: satisfaction, personality, process, and outcome. *Personality and Individual Differences* **21**: 49–60
- Gotlib I H, Lewinsohn P M, Seeley J R 1998 Consequences of depression during adolescence: Marital status and marital functioning in early adulthood. *Journal of Abnormal Psychology* **107**: 686–90
- Karney B R, Bradbury T N 1995 The longitudinal course of marital quality and stability: A review of theory, method, and research. *Psychological Bulletin* **118**: 3–34
- Karney B R, Bradbury T N 1997 Neuroticism, marital interaction, and the trajectory of marital satisfaction. *Journal of Personality and Social Psychology* **72**: 1075–92
- Karney B R, Bradbury T N, Fincham F D, Sullivan K T 1994 The role of negative affectivity in the association between attributions and marital satisfaction. *Journal of Personality and Social Psychology* **66**: 413–24
- Kelly E L, Conley J J 1987 Personality and compatibility: A prospective analysis of marital stability and marital satisfaction. *Journal of Personality and Social Psychology* **52**: 27–40
- Kenny D A, Acitelli L K 1994 Measuring similarity in couples. *Journal of Family Psychology* **8**: 417–31
- Kurdek L A 1993 Predicting marital dissolution: A five-year prospective longitudinal study of newlywed couples. *Journal of Personality and Social Psychology* **64**: 221–42
- Marco C A, Suls J 1993 Daily stress and the trajectory of mood: Spillover, response assimilation contrast, and chronic negative affectivity. *Journal of Personality and Social Psychology* **64**: 1053–63
- Poulton R G, Andrews G 1992 Personality as a cause of adverse life events. *Acta Psychiatrica Scandinavica* **85**: 35–8
- Russell R J H, Wells P A 1991 Personality similarity and quality of marriage. *Personality and Individual Differences* **12**: 407–12

- Terman L M 1938 *Psychological Factors in Marital Happiness*. McGraw-Hill, New York
- Winch R F 1958 *Mate Selection: A Study of Complementary Needs*. Harper and Brothers, New York

B. R. Karney

Personality and Risk Taking

There is a long-standing and persistent belief that risk taking is a stable personality trait, often referred to as risk attitude. The belief implies that a given individual will take similar risks across a range of situations and that some people will be more risk-averse (or more risk-seeking) across situations than others. The article reviews different definitions of risk attitude that show cross-situational consistency to varying degrees. Sect. 1 shows that risk attitude defined within the expected utility (EU) framework varies greatly across situations as a function of decision content and outcome framing. Sect. 2 describes a more promising conceptualization of risk taking, within a risk-value framework. It models risk taking as a function of (a) decision makers' perception of the riskiness and value or return of different courses of action, and (b) their attitude towards perceived risk, i.e., their willingness to trade off (perceived) risk for return. Two individuals might differ in their recreational pursuits, for example, either because they assess the relative risks of skydiving, bungee jumping, and playing poker very differently (based on past experience, person A may perceive playing poker to be riskier than skydiving and thus choose to go skydiving out of risk aversion—a negative attitude toward risk, while person B may perceive playing poker as the less risky option and engage in it, also out of risk aversion) or, whether their risk perceptions agree or not, because they have different attitudes toward risk as they see it (with person C and D agreeing on the greater risk posed by skydiving, but person C being attracted by this risk and thus taking it on, and person D being repelled by it and thus choosing to play poker instead).

When modeled within this framework (as described in Sect. 3), situational differences in risk taking turn out to result from differences in the perception of risk in different situations rather than differences in willingness to take on (perceived) risk, thus restoring credibility to the notion of attitude towards perceived risk (PRA) as a stable trait. Individual differences in PRA exist, but are smaller and less systematic than individual and group differences in risk perception. While the determinants of risk perception are relatively well known at this point (see *Risk: Empirical Studies on Decision and Choice*; *Risk: Theories of Decision and Choice*) not much is known about the determinants of PRA. Personality differences in variables known to be

related to risk taking seem to have their effect via differences in risk perception, as described in Sect. 4.

1. Domain and Framing Effects on Risk Taking

In the EU framework, the dominant normative model of risky decision making, the term 'risk taking' is used to characterize choice patterns. Choice of a sure outcome over a lottery with equal expected value is modeled by a concave utility function and described as risk-averse; choice of the lottery is modeled by a convex utility function and described as risk-seeking. Despite the fact that risk taking simply describes the curvature of the utility function that is derived from a series of choices, 'those who coined the term *risk aversion* had in mind the psychological interpretation that someone who prefers the expected value of a gamble over playing the gamble *does not like to take risks*' (von Winterfeldt and Edwards 1986, p. 256). Popular as well as managerial folklore tends to interpret risk taking as a personality trait.

Bromiley and Curley (1992) review the evidence for risk taking as a personality trait, i.e., as a preference for risk that is stable across situations, and find it lacking. Risk taking seems to be influenced jointly by the situation and characteristics of the decision maker. Decision domains in which the same person often shows different degrees of risk taking include games of chance/gambling, financial investing, business decisions, health decisions, recreational choices, social choices, and ethical decisions (MacCrimmon and Wehrung 1986, Weber et al. 2000).

Modeling risk taking within EU theory and defining risk attitude as the curvature of a utility function thus is clearly problematic for the notion of risk attitude as a personality trait. Attempts to restore cross-situational consistency to the construct of risk attitude by factoring differences in marginal value (e.g., the incremental value of an additional dollar or an additional life saved) out of the utility function (see *Risk: Theories of Decision and Choice*) were not successful (Keller 1985, Weber and Milliman 1997).

Prospect theory (Kahneman and Tversky 1979) generalizes EU by postulating different utility functions for outcomes that are framed as gains as opposed to losses. When outcomes are framed as gains, choices tend to be risk-averse; when the same outcomes are framed as losses (relative to a higher reference point), choices tend to be risk-seeking, further complicating the interpretation of risk taking in the EU sense as a stable trait.

2. Risk Taking and Risk Perception

In the risk-value framework (see *Risk: Theories of Decision and Choice*), risk taking is a compromise between greed (value) and fear (risk). Risk-value

models in finance equate 'value' with the expected value of a risky option and 'risk' with its variance. Generalized risk-value models allow for a broader range of risk measures. *Risk: Empirical Studies on Decision and Choice* reviews evidence that risk is perceived differently by different individuals, cultures, or subcultures. Situational differences such as outcome framing also result in different risk perception (Mellers et al. 1997). As a result, apparent differences in risk taking may be the result of differences in the perception of the riskiness of the choice options, and not of differences in attitude towards (perceived) risk. Cooper et al. (1988) report, for example, that—contrary to managerial folklore—the characteristic that differentiates entrepreneurs from other managers is not a more positive attitude towards risk, but instead an overly optimistic perception of the risks involved. For an outside observer who perceives risk more realistically, entrepreneurs will thus appear to take great risks. However, when differences in risk perception are factored out, entrepreneurs—just as other managers—demonstrate a preference for tasks that they see as only moderate in risk (Brockhaus 1982).

3. Perceived-risk Attitude as a Stable Trait

PRA is a measure of the degree to which individuals find perceived risk attractive (or unattractive) and therefore will choose alternatives that carry greater (or less) risk, all other things being equal. Weber and Milliman (1997) examined its cross-situational consistency by asking commuters to choose between pairs of trains that had risky arrival times (that depended on making a connection that had a stated probability) and to judge which of the two trains was the riskier one. The two trains in each pair had arrival times with equal expected value but different variance. Some pairs of trains had only positive arrival times (faster or equal to current travel times), others had only negative arrival times (slower or equal to the status quo). There was little consistency in people's risk taking across the gain and the loss domain when risk taking was defined in the EU sense. Few commuters had preferences that resulted in utility functions that were either both risk-seeking (convex) or both risk-averse (concave). However, consistency across the two domains was very high when PRAs were compared. The majority of commuters were risk-averse in both domains, i.e., consistently chose the train in a given pair that they had judged to be the less risky of the two.

In another study, MBA students participated in two sessions of an investment game where they had to pick one of six stocks (described by standard financial indicators) in each of 10 investment periods, and had to rate the riskiness of the stocks at different points throughout each session (Weber and Milliman 1997).

In one session, participants lost money in most of the 10 periods, whereas in the other session they mostly made money. Choices were very different across sessions (with more switching between stocks in the failure session), as were the ratings of the riskiness of the six stocks. However, over 80 percent of investors had the same PRA in both sessions, with three-quarters consistently investing in stocks that they perceived to be less risky and one-quarter consistently investing in stocks that they perceived to be more risky.

In a cross-national study, Weber and Hsee (1998) obtained risk judgments as well as minimum buying prices for risky financial investment options from respondents in the USA, Germany, the People's Republic of China, and Poland. Both risk judgments and buying prices showed significant cross-national differences, with Americans perceiving the most risk and Chinese paying the highest prices. However, after differences in risk perception were taken into consideration, the proportion of individuals who were perceived risk-averse or perceived risk-seeking were not significantly different in the four countries, with the majority again being perceived risk-averse, and only a small percentage in each country being perceived risk-seeking.

4. Personality, Risk Perception, and Perceived-risk Attitude

Some psychologists have questioned the assumption of finance models that people will and should strive to minimize risk, arguing instead that people's ideal point for risk or uncertainty could differ, either as a personality difference (Lopes 1987) or as a situational difference (Weber and Kirsner 1997). Ideal-point models (Coombs 1975) assume a person will perceive the riskiness of an alternative as the deviation between the alternative's level of uncertainty or unpredictability and the person's ideal point on the uncertainty continuum. Perceived risk of an alternative with a high objective level of uncertainty would be high for a person with a low ideal point, but low for a person with a high ideal point. Individual differences in ideal points for risk and uncertainty have been measured by the construct of sensation seeking (Zuckerman 1979), which seem to have some biological basis (Zuckerman et al. 1988) and vary with age and gender (see *Sensation Seeking: Behavioral Expressions and Biosocial Bases*). Bromiley and Curley (1992) report evidence linking sensation seeking to behavioral correlates that include greater risk taking, especially in the health/safety and recreational domain. Weber et al. (2000) also report high positive correlations between sensation seeking and its subscales in several content domains, with especially high correlations between the thrill-and-adventure-seeking subscale and recreational risk

taking and the disinhibition subscale and ethical risk taking. Consistent with the predictions of ideal-point models, the path by which differences in sensation seeking seem to affect risk taking appears to be differences in the perceptions of risk, rather than differences in attitude towards perceived risk. In other words, groups known for high levels of sensation seeking (e.g., teenage boys) seem to take large risks because they perceive the levels of risk to be smaller than other groups, and not because they cherish (perceived) risk to a greater extent.

5. Summary, Caveats, and Future Directions

The current research consensus suggests an interactional model of risk taking (e.g., Sitkin and Weingart 1995) in which situational characteristics as well as person-centered characteristics jointly influence risk taking. Situational constraints include the content domain of the risky decision as well as contextual variables such as outcome framing and aspiration levels (Lopes 1987). Person-centered characteristics include age, gender, culture, and personality. These variables influence risk taking mostly by changing people's perception of the riskiness of decision alternatives, rather than by affecting their willingness to take on more or less risk.

Because of the domain specificity of risk taking, measures of risk attitude that employ choice situations across a range of content domains (e.g., the Choice Dilemmas Questionnaire of Kogan and Wallach 1964) have little predictive validity. Domain-specific scales of risk taking, that help to diagnose apparent differences in risk taking into differences in either risk perception and/or PRA have recently developed (Weber et al. 2000). Future research will provide additional insights into the complex interactions between personality and situation that have been explored for a range of other traits (Mischel 1999) with respect to risk taking. A combination of task analysis and theory about the reasons for risk taking and its cognitive and emotional constraints should lead to the development of gender-, culture-, and domain-specific risk taking profiles, that predict level of risk taking in a situation- and person-contingent fashion.

See also: Personality Psychology; Risk: Empirical Studies on Decision and Choice; Risk: Theories of Decision and Choice; Sensation Seeking: Behavioral Expressions and Biosocial Bases

Bibliography

Brockhaus R H 1982 The psychology of the entrepreneur. In: Kent C A, Sexton D L, Vesper K G (eds.) *The Encyclopedia of Entrepreneurship*. Prentice Hall, Englewood Cliffs, NJ, pp. 92–4

- Bromiley P, Curley S P 1992 Individual differences in risk-taking. In: Yates J F (ed.) *Risk-taking Behavior*. Wiley, New York, pp. 87–132
- Coombs C H 1975 Portfolio theory and the measurement of risk. In: Kaplan M F, Schwartz S (eds.) *Human Judgment and Decision*. Academic Press, New York, pp. 63–8
- Cooper A C, Woo C Y, Dunkelberg W C 1988 Entrepreneurs' perceived chances for success. *Journal of Business Venturing* 3: 97–108
- Kahneman D, Tversky A 1979 Prospect theory: An analysis of decision under risk. *Econometrica* 47: 263–91
- Keller L R 1985 An empirical investigation of relative risk aversion. *IEEE Transactions on Systems, Man, and Cybernetics* 15: 475–82
- Kogan N, Wallach M A 1964 *Risk-taking: A Study in Cognition and Personality*. Holt, New York
- Lopes L L 1987 Between hope and fear: The psychology of risk. *Advances in Experimental Social Psychology* 20: 255–95
- MacCrimmon K R, Wehrung D A 1986 *Taking Risks: The Management of Uncertainty*. Free Press, New York
- March J G, Shapira Z 1992 Variable risk preferences and focus of attention. *Psychological Review* 99: 172–83
- Mellers B A, Schwartz A, Weber E U 1997 Do risk attitudes reflect in the eye of the beholder? In: Marley A A J (ed.) *Choice, Decision, and Measurement*. Erlbaum, Hillsdale, NJ, pp. 57–71
- Mischel W 1999 Implications of person–situation interaction: Getting over the field's borderline personality disorder. *European Journal of Personality* 5: 455–61
- Sitkin S B, Weingart L R 1995 Determinants of risky decision making behavior: A test of the mediating role of risk perceptions and risk propensity. *Academy of Management Journal* 38: 1573–92
- von Winterfeldt D, Edwards W 1986 *Decision Analysis and Behavioral Research*. Cambridge University Press, Cambridge, UK
- Weber E U, Blais A R, Betz N 2000 A domain-specific risk-attitude scale: Measuring risk perceptions and risk behavior. Working Paper, Center for the Decision Sciences. Columbia University, New York
- Weber E U, Hsee C K 1998 Cross-cultural differences in risk perception but cross-cultural similarities in attitudes towards risk. *Management Science* 44: 1205–17
- Weber E U, Kirsner B 1997 Reasons for rank-dependent utility evaluation. *Journal of Risk and Uncertainty* 14: 41–61
- Weber E U, Milliman R A 1997 Perceived risk attitudes: Relating risk perception to risky choice. *Management Science* 43: 122–44
- Zuckerman M 1979 *Sensation Seeking: Beyond the Optimal Level of Arousal*. Erlbaum, Hillsdale, NJ
- Zuckerman M, Simons R F, Como P G 1988 Sensation seeking and stimulus intensity as modulators of cortical, cardiovascular, and electrodermal response: A cross-modality study. *Personality and Individual Differences* 9: 361–72

E. U. Weber

Personality and Social Behavior

The link between personality and social behavior can be approached in two ways. The first question is to what extent social behavior is caused by personality

factors. The second question is what effects social behavior has on personality. Both questions have proven surprisingly difficult for social scientists to answer.

Personality refers to the stable traits and dispositions that characterize individual people. In simplest terms, social behavior has two possible causes: personality and the situation. Although it is tempting to assign everything inside the person to personality and everything outside to the situation, the reality is somewhat more complex, because external factors (such as being insulted by others, an emergency, or an opportunity) may create temporary internal states. Personality is thus usually reserved to the relatively stable aspects of the person, as opposed to temporary states. One handy test is what aspects of the person remain the same when he or she moves from one situation to another. Personality consists of those aspects.

1. Mischel's Challenge to Personality

The view that social behavior derives from personality is deeply rooted in common sense and traditional views. People are widely believed to have stable personality traits that account for readily discernible patterns in their behavior. Indeed, the very concept of personality is usually understood in terms of consistency. That is, people are supposedly consistent in their motivations and actions.

Walter Mischel (1968) challenged this alleged consistency of personality in an influential book. Mischel proposed that if personality were composed of a stable, consistent set of traits, then psychological measures of traits should effectively predict social behavior quite strongly. From a review of published research studies, however, he concluded that personality traits have not been shown to have strong effects on behavior. Instead, he found that the correlations between traits and behaviors were usually small: Typically the correlation was found to be 0.20 to 0.30 out of a maximum possible 1.00. He used the term 'personality coefficient' to refer to the small correlations that were typically found.

Mischel therefore concluded that the situation is far more powerful than personality as a cause of social behavior. Following the standard statistical practice, he pointed out that a correlation of 0.30 accounts for only 9 percent of the variation in behavior, leaving the other 91 percent to the situation. Put simply, the situation was 10 times as powerful as personality in determining behavior.

This view was immediately recognized as challenging the view that personality is an important cause of social behavior. Psychologists were not the only ones to have believed in personality, for the general public has also had a long and firm belief in personality.

Mischel's colleague Lee Ross (1977) coined the term 'fundamental attribution error' to refer to people's tendency to interpret someone's behavior as caused by that person's personality traits and to underestimate the importance of the situation in causing behavior. For example, people who observe an angry outburst will tend to conclude that the angry person has a hostile or aggressive personality while in fact the true cause of the angry outburst is likely to be a bad day, a provocation, a frustration, or some other cause rooted in the immediate situation. A well-known work by Jones and Nisbett (1971) pointed out that people tend to attribute their own behavior as caused by their situation (e.g., 'I was only following orders' or 'I was just reacting to what you did'), whereas they interpret other people's behavior in terms of personality traits.

Mischel's book and his striking conclusions had a major impact on psychology. One implication was that psychologists should focus on analyzing the situation rather than the individual personality in order to be able to understand and predict behavior. Social psychologists, who specialized in studying the power of situations, embraced this conclusion to the detriment of personality psychologists. This attack on personality coincided with a period of rapid growth in US psychology departments, and as a result many universities hired social psychologists rather than personality psychologists, with the result that social psychology continues to be a much larger field than personality even several decades later.

A recent analysis of the controversy by Funder (1997) summarized Mischel's argument as having three essential parts. First, a review of the literature shows that there is an upper limit to how well one can predict someone's behavior from one situation to the next, and this upper limit is small (a correlation of around 0.20 to 0.30; Nisbett 1980 later revised this upward to 0.40). Second, situations are more important than traits in explaining behavior. Third, it is no use to measure personality, and the common and traditional view of people as consistent, personality-driven beings is wrong.

2. Defending Personality

Personality psychologists gradually came forward to defend their field of study and the value of their work against Mischel's critique. One important point was that the correlations of 0.20 to 0.40 are not trivial and can in fact contribute useful knowledge. A correlation of 0.40 means that a prediction of behavior would be right 70 percent of the time (compared to 50 percent by chance).

Moreover, the conclusion that situations are much more powerful than personality traits does not stand up under scrutiny. Funder and Ozer (1983) analyzed the power of situational variables to predict behavior,

using several classic social psychology experiments on cognitive dissonance and bystander reactions to emergencies. Funder and Ozer (1983) found that the well-known effects of these situational factors were about the same size (between 0.30 and 0.40, when converted to correlations) as Mischel's personality coefficient. In other words, trait and situation have about equally powerful effects; it was simply not true that the situation was 10 times as powerful as personality.

Another important point was that the seeming weakness of personality traits for predicting behavior was partly due to the way psychologists did their research. Typically they would use a standard questionnaire to assess some general trait, such as helpfulness, and then they would correlate it with how helpful the person was in a specific situation (such as how the person responded to a request to donate blood). The problem with such designs is that only a single, context-specific behavior was measured, which might not match up well with a broad trait. In contrast to that procedure, some researchers began aggregating multiple behaviors. Thus, instead of only measuring whether the person acceded to one request to donate blood, one could combine many different helpful behaviors, such as taking care of sick friends, giving directions to strangers, volunteering to help on a hotline, donating money, befriending disadvantaged children, helping friends with chores, and assisting someone with schoolwork. Epstein (1979a, 1979b) found that aggregating behaviors in that way enabled traits to predict behavior at a level of over 0.80 in some cases, thus far more successfully than the 0.3 coefficient derived by Mischel.

The success of aggregation is revealing about how personality is linked to social behavior. It shows that personality traits can predict behavior in general with good success. They are, however, much less successful at predicting single behaviors. To know someone's personality is therefore to be able to predict how that person will behave in many situations, but it remains much harder to say how a given person will react on a particular occasion.

3. Traits and Metatraits

Another possible reason for the seeming weakness of the link between personality and behavior is that not all personalities are composed of the same kinds of traits. Bem and Allen (1974) proposed that some people are more consistent than others with respect to certain traits. Including inconsistent people in a research project will inevitably weaken the results. For example, Bem and Allen (1974) asked people whether they were consistent or inconsistent on the dimension of sociability. One person might be consistently high in sociability, or consistently medium, or consistently low. Someone else might be inconsistent, in the sense

that the person acts in a highly sociable manner one day but in an unsociable or even antisocial way the next. Bem and Allen computed the correlation between trait sociability score and a behavioral measure of sociability twice: once for the people who said they were consistent, and once for the inconsistent ones. The correlation was much higher for the people who were consistent.

The implications of Bem and Allen's idea are potentially profound. For some people, the trait of being sociable versus unsociable is a stable, consistent aspect of personality, but for others it is not. One theme in personality psychology is the attempt to measure traits, but perhaps some people should not be measured on some traits. The question 'How sociable are you?' may be relatively meaningless for some people, because the answer would fluctuate from day to day and from one situation to another. For the latter, apparently, their personality lacks anything that could be described as a trait level of sociability. Baumeister and Tice (1988) used the term 'metatrait' to describe this pattern. A metatrait is literally the trait of having (or not having) a trait. For traited people, behavior will tend to be consistent. For untraited people, behavior may vary widely even within seemingly similar situations.

The idea of metatraits poses a serious challenge to both the measurement and the conceptualization of personality. Several methods have been proposed to ascertain whether people are traited on a given trait or not. There is ample evidence that the very same measures yield much stronger results for traited as opposed to untraited people (Baumeister and Tice (1988, see also Britt and Shepperd 1999). Yet other experts have proposed that these measures could potentially have statistical problems (e.g., Tellegen 1988), and there is not yet a general consensus on the best way to deal with these.

Still, the idea that people differ in their level of consistency is promising, and eventually it may turn out to be a crucial aspect of the link between personality and social behavior. The implication is that personality contains something that makes the person's behavior consistent, and that some personalities do not have whatever that is. Although Bem and Allen (1974) and most subsequent researchers have assumed that different people are traited on different traits, Snyder (1974) proposed that some people are simply more consistent than others across all traits. He developed a measure of self-monitoring that sought to assess how much each person changes his or her behavior from one situation to another. The term 'self-monitoring' refers to the monitoring of situational cues and altering one's own behavior accordingly. People who are high in self-monitoring are defined as being alert to circumstances and quick to change their behavior so as to do what might work best in any given situation. In contrast, people who score low in self-monitoring tend to act based on their inner feelings

and values regardless of the situation. Hence low self-monitors will be more consistent across different situations, whereas high self-monitors will tailor their behavior to each situation and exhibit less cross-situational consistency.

4. Different Situations

Two other ways of understanding the link between personality and social behavior focus on differences between situations. One of these emphasizes simply that different situations vary in their power to dictate behavior. A strong situation can override personality traits, so that nearly everyone will act the same way. In contrast, a relatively weak situation will allow behavior to be driven by personality.

Research on behavior in emergencies has provided a good illustration of this approach. Darley and Latané (1968) investigated the question of whether bystanders will come to the aid of a victim in an emergency. They found that when bystanders are alone and believe themselves to be the only ones who can help the victim, they are quite likely to act. In contrast, when there are many bystanders, each is likely to think that it is someone else's responsibility to take action, and so each will tend to do nothing. Across a series of studies, they measured an assortment of personality traits, but none of these traits predicted behavior. They concluded from their research that personality is largely irrelevant to how people respond in emergencies (see Latané and Darley 1970). They suggested that emergency situations are both powerful and unusual, and so few people develop habits or traits about how to deal with them.

The conclusion that personality is irrelevant to emergencies came under criticism from personality researchers. They proposed that the original researchers had simply used extremely powerful versions of the emergency situation, with either no one else available to help or a great many people. What about when there are only one or two other people who might help? When the emergencies were set up with just one or two other bystanders, the response rate moved to the intermediate level of 50 percent, and personality traits were quite effective at predicting which people would respond (Siem and Spence 1986, Tice and Baumeister 1985).

In short, situations do vary in how powerfully they exert pressure on people to respond in a particular way. When this power is high, there is little room for personality to make a difference. In contrast, when the power of the situation is low, personality can be decisive. This conclusion complements the metatrait view that some personalities are more strongly geared toward consistent behavior than others. An untraited person will generally respond based on the situation, whereas a traited person will act based on his or her

personality. Conversely, a powerful situation will dictate behavior, whereas a weak one will leave behavior to be determined by inner traits.

The second situation-based approach was put forward by Snyder and Cantor (1998). In this theory, situations are the primary determinant of behavior — but personality dictates which situations the person enters. For example, introverted people may avoid loud parties, and so their behavior does not come under the influence of the party atmosphere. Thus, personality and situations may operate in sequence rather than in direct competition for determining social behavior. Personality involves the grander role of how people set up their lives and how they make broad choices, whereas situations exert the immediate and short-term influence over how the person acts.

5. Behavior Shapes Personality

Although most research has focused on the question of whether personality determines social behavior, there has been some work on the complementary question of whether behavior shapes personality. To be sure, personality is defined as stable and consistent, so such change will be slower, smaller, and more gradual than changes in behavior. Yet few people believe that personality is utterly immune to change. A compilation of works by many experts on the topic of 'Can personality change?' yielded an emphatic answer of 'yes,' even though change was acknowledged to be difficult and slow in many cases (Heatherton and Weinberger 1994).

One of the possible models of personality change involves internalization. For example, young women do not magically develop maternal personalities, nor are these traits always visible during their youth prior to motherhood, but upon assuming the role of mother a woman may begin to change her behavior, and over time these new patterns of maternal behavior can become firmly established to the point that the woman's personality can be said to have changed to fit the role.

How does the internalization process operate? One important theory was proposed by Jones et al. (1981). These authors suggested that people's self-concepts contain a great deal of information that is not necessarily consistent or well-organized. When people are induced to act in a certain way, they begin to recall other behaviors that are consistent (and to ignore previous acts that would not be consistent). Eventually the self-concept shifts to accommodate these patterns of behavior, and then the personality itself follows suit. Fazio et al. (1981) asked people a series of loaded questions that were designed to get them to think about themselves as either extraverted or introverted (e.g., 'What things would you do to liven up a dull party?' versus 'What do you dislike about loud,

crowded parties?"). Their self-concepts shifted accordingly. Later, when they were left alone with a stranger, their behavior followed from the newly induced self-concept, which implies that there was at least some lasting change in their personality. Specifically, the people who had been led to think of themselves as introverts sat quietly and did not talk with the stranger, whereas the people who had been led to think of themselves as extraverts were more likely to strike up a conversation with the stranger.

The importance of the social dimension of behavior emerged in subsequent work. Tice (1992) showed that people only internalize behaviors that are witnessed by other people. That is, when people gave their answers to another person in an interpersonal setting where another person heard them, they showed evidence of personality change. In contrast, when they gave their answers confidentially and anonymously by speaking into a tape recorder while alone, they showed no sign of internalization. The implication is that secret and private behaviors do not lead to internalization as much as do public, interpersonal behaviors. The social context—specifically, whether other people are there to witness the behavior—is decisive as to whether behaviors lead to inner change.

The practical side of this can be seen in the practices used by cults and other organizations that seek to 'brainwash' or otherwise bring about fundamental changes in people's values, motives, and other aspects of personality. Attempts to effect such change by bombarding the person with messages or influence while the person sits there passively have not proven effective. Nor is it apparently enough to get the person to comply in a private or anonymous fashion. Rather, success at brainwashing depends on inducing the person to comply actively and publicly. New recruits to a religious cult may, for example, be induced to speak out in front of the group about their new commitment to the cult or even to approach strangers to tell them about the cult, ask for donations, distribute literature, and the like. Such public actions help cement the person's commitment to the cult and bring about the inner change that the group seeks.

6. *Interactionism*

Although the intellectual debate has been cast in terms of whether the trait or the situation is the main cause of behavior, this is merely a matter of emphasis: Most experts believe in interactionism, which is the doctrine that behavior is an interactive product of the person and the situation (see *Interactionism and Personality*). Put another way, hardly anyone believes that traits are so powerful that people will always behave the same way regardless of situation, nor does anyone really believe that there are no meaningful differences among individuals.

The interaction between person and situation is implicit in several of the formulations already noted, such as the idea that personality dictates the selection of particular situations, and then the situations guide actual behavior. An explicit interactionism would assert that any approach is doomed if it focuses on merely the person or merely the situation. A crowded, noisy party will make some people mingle happily with the crowd and enjoy shouted discussions with groups of people, while others retreat to search for a quiet place where they can talk at length with a single person. Likewise, a major upcoming test will strike some people as a threat, making them so anxious that they avoid studying, whereas others will embrace the test as a challenge and prepare to do their best. Thus, seemingly identical situations elicit very different behaviors from different people, depending on their personalities.

Modern statistical techniques allow the observations of human behavior to be divided into three parts. Trait and situation each claim a share, and the interaction (between trait and situation) gets the third share. The trait share refers to how much behavior is affected directly by personality, independent of the situation (and thus consistently across different situations). The situation share involves how much all different people will give essentially the same response to the situation. Last, the interaction encompasses ways that the same situation affects different people differently.

There is unfortunately no general principle or agreement as to which of these three shares is generally the largest. It appears that some situations are quite powerful in affecting everyone the same way. Other situations are relatively weak (such as unstructured situations), and traits exert considerable influence there. Interactions do not always occur, but they are common and often generate the most interest, insofar as different people react in reliably different ways.

7. *Future Directions*

At present, it appears that personality psychology has successfully survived the crisis provoked by Mischel's (1968) critique, although some effects linger (such as the smaller size of personality psychology faculty). Instead of looking at personality and situation as competing against each other to explain behavior, the current emphasis is to understand how they interact and complement each other. A priority for the next generation of research is to assess the several models (described above) for possible interactions, such as differential traitedness, personality dictating entry into situations and situations being the proximal causes of behavior, and differential power of situations.

Another reason to expect increasing rapprochement between research programs focusing on personality

and programs focusing on situational causes is contained in the logic of scientific development. As in most fields, broad generalizations and sweeping general principles are likely to be identified first, whereas later generations have to study more fine-grained and subtle effects. The behavioral principles that are generally true for everyone were likely identified relatively early in psychology's scientific development. Subsequent generations of researchers will therefore have to focus more on principles that apply only to certain people, and these attempts to understand how situations affect particular groups will require a greater sensitivity to individual differences and a greater integration between personality and situational factors.

The study of how situations and specific experiences produce lasting changes in personality remains wide open for new ideas and creative research. The question of how personality changes is likely to be of increasing importance in the coming decades for many reasons, including the wish to facilitate child development toward desired personality outcomes, the increasingly recognized need to offset the consequences of harmful parenting or other nonoptimal upbringing patterns, and in particular the wish to enable clinical psychology to bring about the desired changes in a reliable and lasting manner.

See also: Interactionism and Personality; Personality and Adaptive Behaviors; Personality Psychology; Personality Theories

Bibliography

- Baumeister R F, Tice D M 1988 Metatraits. *Journal of Personality* **56**: 571–98
- Bem D J, Allen A 1974 On predicting some of the people some of the time: The search for cross-situational consistencies in behavior. *Psychological Review* **81**: 506–20
- Britt T A, Shepperd J A 1999 Trait relevance and trait assessment. *Personality and Social Psychology Review* **3**: 108–22
- Darley J M, Latané B 1968 Bystander intervention in emergencies: Diffusion of responsibility. *Journal of Personality and Social Psychology* **8**: 377–83
- Epstein S 1979a The stability of behavior: I. On predicting most of the people much of the time. *Journal of Personality and Social Psychology* **37**: 1097–126
- Epstein S 1979b The stability of behavior: II. Implications for psychological research. *American Psychologist*. **35**: 790–806
- Fazio R H, Effrein E A, Falender V J 1981 Self-perceptions following social interactions. *Journal of Personality and Social Psychology* **41**: 232–42
- Funder D C, Ozer D J 1983 Behavior as a function of the situation. *Journal of Personality and Social Psychology* **44**: 107–12
- Funder D C 1997 *The Personality Puzzle*. W W Norton, New York

- Heatherton T F, Weinberger J (eds.) 1994 *Can Personality Change?* American Psychological Association, Washington, DC
- Jones E E, Nisbett R E 1971 *The Actor and the Observer: Divergent Perceptions of the Causes of Behavior*. General Learning Press, New York
- Jones E E, Rhodewalt F, Berglas S C, Skelton A 1981 Effects of strategic self-presentation on subsequent self-esteem. *Journal of Personality and Social Psychology* **41**: 407–21
- Latané B, Darley J M 1970 *The Unresponsive Bystander: Why Doesn't He Help?* Appleton-Century Crofts, New York
- Mischel W 1968 *Personality and Assessment*. Wiley, New York
- Nisbett R E 1980 The trait construct in lay and professional psychology. In: Festinger L (ed.) *Retrospections on Social Psychology*. Oxford University Press, New York, pp. 109–30
- Ross L 1977 The intuitive psychologist and his shortcomings: Distortions in the attribution process. In: Berkowitz L (ed.) *Advances in Experimental Social Psychology* (Vol. 10, pp. 174–221). Academic Press, New York
- Siem F M, Spence J T 1986 Gender-related traits and helping behaviors. *Journal of Personality and Social Psychology* **51**: 615–21
- Snyder M 1974 The self-monitoring of expressive behavior. *Journal of Personality and Social Psychology* **30**: 526–37
- Snyder M, Cantor N 1998 Understanding personality and social behavior: A functionalist strategy. In: Gilbert D, Fiske S, Lindzey G (eds.) *Handbook of Social Psychology*, 4th edn. McGraw-Hill, Boston, pp. 639–79
- Tellegen A 1988 The analysis of consistency in personality assessment. *Journal of Personality* **56**: 621–63
- Tice D M 1992 Self-presentation and self-concept change: The looking glass self as magnifying glass. *Journal of Personality and Social Psychology* **63**: 435–51
- Tice D M, Baumeister R F 1985 Masculinity inhibits helping in emergencies: Personality does predict the bystander effect. *Journal of Personality and Social Psychology* **49**: 420–8

R. F. Baumeister and J. M. Twenge

Personality Assessment

To assess personality is to characterize the enduring psychological qualities that contribute to the coherence and uniqueness of the individual person and to the differences among individuals. The task of personality assessment commonly is construed in terms of the measurement of quantifiable personality variables. However, personality assessment may include both quantitative and qualitative techniques.

The set of methods that psychologists have employed to assess personality is highly diverse. There are two main reasons for this. One is that assessors have held varying beliefs about the nature of personality itself. The twentieth century witnessed a variety of formal theories of personality. These theoretical positions have guided assessors' decisions about the aspects of personality to assess and the procedures to use to assess them. Diversity in personality theory thus has fostered a corresponding diversity in assessment techniques. The second reason is that investi-

gators may assess personality with different goals in mind. One may aim, for example, to describe normal variations in personality in the population at large; to diagnose psychopathology suffered by a small subset of the population; or to obtain a detailed portrait of the psychological structures and dynamics that characterize a particular individual. Different goals commonly require different assessment methods. The methods that contemporary investigators employ are reviewed below, following a history of the field.

1. History

In a general sense, to assess personality is to evaluate the individual in a formal manner. Personality assessment, then, is as old as is the development of formal schemes for characterizing persons. Two early schemes are found in ancient Greece. The *Corpus Hippocraticum* introduced the notion that variations in four bodily humors (blood, phlegm, yellow bile, and black bile) determine one's predominant psychological temperament (sanguine, melancholic, choleric, and phlegmatic). Theophrastus, a disciple of Aristotle, provided a series of character sketches of common personality types, each defined according to a moral attribute (the liar, the adulterer, etc.). The differences between these approaches—one examining biological constitution and providing a simple system for characterizing individual differences, and the other focusing on social behavior in the moral domain and providing a more differentiated set of descriptors—interestingly foreshadows variations within the contemporary field.

The more recent centuries witnessed attempts to link personality to biological anatomy. Investigators measured anatomical features in an effort to identify underlying causes of overt psychological qualities. These efforts included the eighteenth to nineteenth century work of Lavater, who interpreted facial features; the 'organology' of Gall and 'phrenology' of Spurzheim, both of whom assessed variations in the anatomy of the skull; and the work of Lombroso, who assessed physical features that purportedly were markers of criminality. These efforts, of course, proved to be of little lasting value.

A more sophisticated conception of assessment was found in the efforts of Charcot, Janet, Freud, and Jung to diagnose the causes of psychopathology. Freud's free association technique, for example, was not only a therapeutic treatment but an assessment tool. Whatever the merits of these efforts, however, they failed to meet applied psychologists' need for efficient assessments that could be administered easily to relatively large populations.

Modern approaches to personality assessment can be traced to the late-nineteenth and early-twentieth century efforts of Galton and Binet to assess intelligence. Their standardized, paper-and-pencil assessments of individual differences in intellectual abilities

provided a model for the standardized assessment of individual differences in personal styles, preferences, and behavioral tendencies. The first paper-and-pencil standardized personality inventory was the Woodworth Personal Data Sheet, which was developed during World War I to screen from the US armed forces individuals suffering from psychological disorders. Thanks heavily to their ease of administration, standardized self-report tests proved to be the most commonly employed personality assessment method throughout the remainder of the twentieth century.

A noteworthy alternative to paper-and-pencil assessment was developed during World War II by psychologists at the US Office of Strategic Services, working under the direction of Harvard psychologist Henry Murray. To screen candidates for participation in risky missions, this group assessed candidates' behavior in artificially constructed situations that contained the same types of challenges (e.g., solving practical engineering problems, displaying leadership in a group) that one might face on an actual mission. Such *in vivo* behavioral assessments can be of much predictive value. However, they are much more cumbersome and costly than paper-and-pencil self-reports, and thus are used much less frequently.

Finally, two theoretical developments in the 1950s and 1960s proved to be of historical importance to the field. One was the development of the concept of *construct validity* (Cronbach and Meehl 1955). Construct validity is concerned with the degree to which an assessment tool adequately measures a hypothesized psychological construct, as that construct is understood within a given theoretical framework. Determining the validity of a test generally involves gauging the degree to which that test predicts external criteria to which the associated construct should be related. A valid self-report test of conscientiousness, for example, would be one that adequately samples the domain of conscientious behaviors and predicts external criteria involving conscientious actions such as adhering to social norms and controlling impulses.

The second development was a paradigm crisis in the field of personality assessment in the late 1960s. Although the crisis had many sources, a critique by Mischel (1968) was particularly influential. Mischel questioned whether assessments of decontextualized or 'global' personality constructs predict specific behaviors to which they are conceptually related. His review indicated that global personality measures commonly either fail to predict behavioral criteria or predict so little of the variance that they are of little practical use. His critique prompted divergent reactions. Some tried to improve the measurement of global psychological characteristics. Others developed personality theories and associated assessment tools that were not based on global trait variables. These theoretical alternatives were, instead, grounded in the analysis of basic cognitive and affective processes, and the social contexts that activate these psycholo-

gical mechanisms. These ‘social–cognitive’ theories aimed to account for both trait-like consistency and situation-to-situation variability in behavior (see *Personality Theory and Psychopathology*).

2. Distinctions among Contemporary Approaches to Personality Assessment

The diversity of contemporary assessment tools can be organized according to a set of distinctions that differentiate alternative assessment efforts.

2.1 Targets of Assessment

One distinction concerns the target of personality assessment; that is, what it is that one is trying to assess. The most basic differentiation separates (a) the assessment of overt, surface-level tendencies in experience and action from (b) the assessment of internal personality structures and dynamics. Phrased simply, one may assess either phenotypic tendencies or genotypic psychological systems that causally contribute to those tendencies.

Whether one is assessing surface-level tendencies or internal psychological systems, one still must choose the precise tendencies or systems to target. These decisions generally are guided by broad theoretical principles about the nature of personality. With respect to surface-level phenotypes, the most common choice is to assess mean-level behavioral tendencies. Assessors commonly compute people’s average tendency to display a particular type of behavior by aggregating together behavioral reports from diverse social contexts. This choice reflects the thinking of trait theories of personality, which view average tendencies in action as the defining expression of personality. In contrast, others caution that an assessment of mean-level tendencies is a necessary but insufficient step. It is insufficient because other features, such as variation in action across contexts (Mischel and Shoda 1995) or time (Larsen 1989), also distinguish individuals from one another. The assessment of contextual variation in action is consistent with social–cognitive theories of personality (Bandura 1986, Cervone and Shoda 1999), as detailed in Section 3.2.

Regarding the underlying genotypes, different theories again suggest different targets of assessment. Psychodynamic views explain behavior in terms of internal psychic conflicts that involve material of which persons are not consciously aware. Assessment thus must target these nonconscious psychological systems; projective tests such as the Rorschach inkblot test and the Thematic Apperception Test are designed for this purpose (see *Projective Methods in Psychology*). Social–cognitive theories contend that the core features of personality are cognitive capabilities

through which people acquire social skills, reflect upon themselves, evaluate their actions, and regulate their behavior and emotional experience in accord with personal goals and standards for performance (Bandura 1986) (see *Social Cognitive Theory and Clinical Psychology*). This view dictates that assessments target an interacting system of cognitive and affective mechanisms (Cervone et al. 2001). Both psychodynamic and social–cognitive theories anticipate that there may be complex, nonlinear relations between underlying psychological systems and overt behavioral patterns. In contrast, some trait theories of personality propose that internal personality structures and overt behavior are related in a relatively direct, linear manner. Personality is construed as a small set of inherited trait structures (e.g., conscientiousness, extraversion), each of which fosters a characteristic mean level of phenotypic thoughts, feelings, and actions (e.g., conscientious acts, extraverted acts; McCrae and Costa 1996). Finally, some theorists study the role in personality functioning of personal narratives that individuals construct over the course of life (McAdams 1996). Their work suggests that personality assessment should include procedures (e.g., oral interviews or written narratives) to assess individuals’ life stories.

2.2 Nomothetic vs. Idiographic Assessment

A second distinction involves overarching strategies of assessment. One strategy is to focus on an individual person and determine the potentially unique constellation of psychological variables that best characterizes that individual. Another is to focus first on the variables, and to seek a set of personality variables and associated assessment procedures that can be used to characterize the personality of any and all individuals. The former approach is an ‘idiographic’ strategy (from the Greek *idios*, referring to personal, private, and distinct characteristics), whereas the latter approach is ‘nomothetic’ (from the Greek for law, *nomos*, used here to refer to the search for universal scientific laws). A typical nomothetic strategy might use standardized tests to determine individuals’ relative standing on one or more individual-difference variables. Idiographic assessment might involve interviews, unstructured self-descriptions, or biographical analyses of an individual’s life.

Personality assessment may blend nomothetic and idiographic procedures. For example, one may have the same assessment goals and employ the same general testing procedures with all individuals. However, the exact content of the test items one employs may vary from person to person. Kelly’s (1955) Role Construct Repertory test is a classic example of a test that combines nomothetic procedures with idiographic content. The assessor’s goal is always to identify the ideas, or constructs, that people use to understand their world. The testing procedure is

always one in which test takers are asked to indicate how a set of target persons is similar or different from one another. The exact content of the test items, however, varies idiographically. Each test taker provides a personal list of individuals who are of importance to him or her. This personalized list then comprises the list of target persons employed in the test. Contemporary assessments of belief systems and social-cognitive processes commonly mix idiographic and nomothetic methods in this manner (Caprara and Cervone 2000).

In practice, personality assessment has been dominated by the use of nomothetic procedures. The most commonly employed assessment instruments, such as the Cattell 16 P.F. Inventory, the Eysenck Personality Questionnaire (EPQ), the Minnesota Multiphasic Personality Inventory (MMPI), or the revised NEO personality inventory (NEO-PI-R), are nomothetic devices (see *Minnesota Multiphasic Personality Inventory (MMPI)*). Assessment consists of the administration of a standardized test that is used to rank individuals on a set of individual-difference dimensions.

2.3 Sources of Data

Whether one is employing a nomothetic or idiographic strategy, another question faced by the personality assessor is the source of data to draw upon. Block (1968) has drawn useful distinctions among alternative data sources; the following section modifies and elaborates upon his distinctions in light of contemporary research developments.

One can delineate at least seven sources of data that may be used in personality assessment

(a) *Behavioral Observations in Naturalistic Settings* are direct observations of a person's everyday behaviors, or the analysis of biographical records (e.g., school attendance records, records of legal infractions) that provide an objective index of such behaviors.

(b) *Behavioral Observations in Experimentally Constructed Settings* are observations of a person's behavior in situations designed by a researcher to constitute a context that is relevant to a personality construct or constructs under study. The test taker's reactions in that context are interpreted as indicators of the construct. Although the use of experimentally constructed settings has the advantage of greater experimental control, it also has a significant disadvantage. Personality characteristics often reveal themselves in the way people select environments; that is, in their choices to enter some settings and avoid others. Assessment procedures that confront all individuals with a fixed set of circumstances are insensitive to the influence of personality factors on the selection of environments.

(c) *Explicit Self-reports* ask people to describe their typical behaviors, thoughts, feelings, or personal

characteristics. This is most commonly done through standardized multi-item tests, although explicit self-reports also may involve less structured techniques such as the writing of self-descriptive narratives.

(d) *Implicit Self-reports* are designed to tap beliefs or self-evaluations of which individuals may not be explicitly aware. For example, reaction-time techniques can reveal the degree to which individuals associate concepts with one another; an implicit test of self-esteem might use reaction-time measures to determine how strongly the concept of 'self' is linked to concepts of positive vs. negative emotional tone (Greenwald and Banaji 1995).

(e) *Observer Reports* are assessment procedures in which peers, family members, or other persons familiar with an individual are asked to evaluate their typical behaviors, feelings, or personal characteristics.

(f) *Psychophysiological Indices* include any of a wide variety of indicators of neural or biochemical systems that directly bear upon personality functioning and psychological differences among individuals. For example, electrophysiological measures are used to assess individual differences in cortical brain activity that may be a biological basis of overt individual differences in extraversion (Geen 1997).

(g) *Cognitive Indices* are assessments of the content or organization of a person's knowledge about themselves and the world. Many of the most important differences among individuals involve differences in knowledge and beliefs systems. A diverse set of tools, many deriving from the study of social cognition, have been employed to assess enduring individual differences in the content, complexity, and organization of knowledge structures that underlie personality functioning and individual differences (Caprara and Cervone 2000).

2.4 Comprehensiveness of Assessment

A final distinction concerns the comprehensiveness of personality assessment. Assessors sometimes wish to obtain a relatively complete portrait of the dispositional tendencies or underlying psychological dynamics that characterize an individual. Alternatively, one may desire merely to tap one or two personality variables of interest. The term 'personality assessment' commonly is applied to both comprehensive assessment efforts and measures of individual personality variables.

3. Illustrative Assessment Strategies: Five-factor Theory and Social-cognitive Theory

The conceptual distinctions drawn above are somewhat abstract. A more concrete understanding of alternative personality assessment strategies can be obtained by considering some examples of theory-

driven assessment. Two examples are considered here; namely, the personality assessment procedures that derive from five-factor theory (McCrae and Costa 1996) and social-cognitive theory (e.g., Bandura 1986, Mischel and Shoda 1995, reviewed in Cervone and Shoda 1999).

3.1 *Assessing Individual Differences in Global Dispositional Tendencies; The 'Big Five' or 'Five-factor' Model*

As noted above, the most common approach to personality assessment is to rank individuals on nomothetic individual-difference dimensions. The dimensions commonly represent global surface-level tendencies; that is, average tendencies to display behaviors that are representative of a given dispositional category. In a comprehensive assessment effort, individuals are ranked on a system of n individual-difference dimensions. The individual's personality is then represented as a point in n -dimensional space.

A primary question that arises is: What are the n dimensions? What set of dimensions is necessary and reasonably sufficient to capture individual differences in the population? A corollary question is methodological: How can one identify these dimensions? On the latter question, there has long been consensus. Investigators conduct factor analyses of the dispositional tendencies of large populations of individuals. The resulting factors then constitute the dimensions that form the core of personality assessment. Despite consensus on methods, on the former question there historically has been disagreement. Investigators in the mid-twentieth century proposed factor-analytic structures containing as many as 16 and as few as two dimensions. A major development in the 1980s and 1990s is that investigators achieved substantial consensus on the number and the nature of the factors required to assess phenotypic individual differences. Consensus centered on a five-dimensional trait taxonomy known as the big five (John and Srivastava 1999) or five-factor (McCrae and Costa 1996) model. The factors represent global individual differences in (a) extraversion or energy, (b) emotional stability or neuroticism, (c) conscientiousness, (d) agreeableness, and (e) openness to experience, particularly novel intellectual or cultural experiences. (Some interpret the fifth factor as being closer to intellect than to openness.) This five-dimensional structure has been identified in both self-reports and observer-reports, in assessment using both trait adjectives and items from professional personality inventories, and in both English (the model's language of origin) and other Indo-European and non-Indo-European languages.

The five-factor model provides an appealingly simple solution to the problem of assessing phenotypic

tendencies. One merely plots people's average tendencies to exhibit actions indicative of each of five dispositional categories. Assessors also may seek more detailed information, such as people's tendency to display behaviors that are representative of narrower trait dimensions that are hierarchically related to the primary factors (in the way that sociability, for example, is hierarchically related to extraversion). Nonetheless, measuring the big five remains the core assessment task.

To some, the model also provides an appealingly simple solution to the problem of assessing causal genotypic structures. Although many proponents of the big five structure treat it merely as a descriptive model (Goldberg 1993), some suggest that it also is explanatory. In McCrae and Costa's (1996) five-factor theory, the factors are construed as 'dimensions of individual differences in tendencies to show consistent patterns of thoughts, feelings, and actions ... (and) also ... a property of an individual that accounts for his or her placement along this trait dimension' (McCrae and Costa 1996, p. 235). 'Agreeableness,' for example, refers simultaneously to a person's tendency to exhibit the behaviors one calls 'agreeable' and to an internal structure that corresponds to, and purportedly explains, this overt tendency.

This theoretical position has an interesting implication for assessment. It dissolves the distinction between the assessment of overt tendencies and internal psychological systems (Sect. 2.1, above). Five-factor theorists assess an individual's average behavioral tendency and interpret the measure as an index of both what the person does (an overt behavioral disposition) and what the person has (an internal psychological structure).

Although the simplicity of this approach may be appealing, its underlying logic can be questioned. Two concerns are of particular note. First, the five factors are identified in analyses of populations, yet are assumed to exist in the psychological make-up of each and every individual in the population (McCrae and Costa 1996). On purely statistical grounds, there is no reason to expect that group-level statistical parameters will replicate themselves at the level of each member of the group. (As a simple example, a group may average 40 years of age, yet no one in the group may be 40 years old.) Second, five-factor theory explains overt dispositional tendencies (e.g., a person's consistently agreeable acts) by inferring hypothetical causal constructs that contain the very properties that one is supposed to explain (e.g., agreeableness). This approach violates a standard principle of scientific explanation. As Hanson (1961, pp. 120–1) explained, 'What requires explanation cannot itself figure in the explanation,' for example, 'if the colors of objects are to be explained by atoms, then atoms cannot be colored' (also see Nozick 1981). This principle of explanation suggests that one should not explain dispositional tendencies by inferring psychological

constructs that directly embody those tendencies. It follows, then, that the assessment of surface-level tendencies and of internal psychological systems that explain those tendencies should be construed as separate tasks (Cervone et al. 2001).

3.2 Assessing Cognitive and Affective Systems and Persons-in-context: Social-cognitive Theory

An approach to personality and assessment that differs markedly from five-factor theory is that of social-cognitive theory (reviewed in Cervone and Shoda 1999). Social-cognitive theory rests on three main principles. Persons and sociocultural environments are viewed as reciprocally interacting systems. The core person variables are cognitive and affective mechanisms through which people interpret the world, reflect upon themselves, and self-regulate their behavior. Finally, personality is viewed as a complex, dynamic system; social-cognitive and affective mechanisms function in concert as interacting systems that underlie the coherence of personality functioning (Cervone and Shoda 1999).

This theoretical perspective has significant implications for assessment. One implication is that assessments of surface-level behavioral tendencies and of underlying psychological systems must be kept distinct. If overt tendencies reflect complex interactions among multiple underlying social-cognitive mechanisms, then there may be no simple relation between an overt behavioral pattern and any individual social-cognitive variable. The assessment of overt tendencies then cannot substitute for the assessment of internal structures. A second implication is that personality assessment must be contextualized. People's self-regulatory abilities enable them to vary their behavior strategically in accord with the perceived opportunities and demands of environmental settings. One must attend to these contextual variations to appreciate fully the distinctive features of an individual's personality (Mischel and Shoda 1995). A third implication is that personality assessment should include more than just an assessment of current dispositional tendencies. People possess potentials that may not be reflected in their typical daily activities. A person may, for example, possess a potential for leadership or for parenthood that only manifests itself when relevant circumstances arise (Caprara and Cervone 2000). Social-cognitive assessors commonly target skills and belief systems through which individuals contribute to their own development over the course of time (Bandura 1986). Finally, social-cognitive theory and research indicates that the content of personal beliefs, and the interconnections among elements of personal and social knowledge, may vary idiosyncratically from one person to another (Cantor and Kihlstrom 1987). This implies that assessments must be highly sensitive to individual idiosyncrasy. Social-

cognitive investigators commonly assess personal belief systems through idiographic methods (Cervone et al. 2001).

4. General Comments

The divergent conceptions of personality assessment that derive from five-factor theory as compared to social-cognitive theory underscore the more general point made earlier. To assess personality, one must rely upon a theory of personality. Questions of personality assessment and personality theory are inevitably intertwined.

In recent years, basic research on biological, affective, and cognitive processes increasingly has informed the study of molar behavioral patterns and differences among individuals. These advances have begun to yield an integrative psychology of personality in which diverse sources of knowledge are brought to bear upon questions of personality functioning and individual differences (Caprara and Cervone 2000). An ongoing challenge for personality assessors, then, is further to develop assessment techniques that do not merely describe surface-level behavioral patterns, but also shed light on specific underlying psychological processes that contribute to the consistency and coherence of personality.

See also: Interactionism and Personality; Minnesota Multiphasic Personality Inventory (MMPI); Personality and Conceptions of the Self; Personality Development and Temperament; Personality Development in Adulthood; Personality Development in Childhood; Personality Psychology; Personality Psychology: Methods; Personality Structure; Personality Theories; Personality Theory and Psychopathology; Projective Methods in Psychology; Social Cognitive Theory and Clinical Psychology; Social Learning, Cognition, and Personality Development

Bibliography

- Bandura A 1986 *Social Foundations of Thought and Action: A Social Cognitive Theory*. Prentice Hall, Englewood Cliffs, NJ
- Block J 1968 Personality measurement: Overview. In: Sills D L (ed.) *International Encyclopedia of the Social Sciences*, Vol. 12. Macmillan, New York, pp. 30–7
- Cantor N, Kihlstrom J F 1987 *Personality and Social Intelligence*. Prentice-Hall, Englewood Cliffs, NJ
- Caprara G V, Cervone D 2000 *Personality: Determinants, Dynamics, and Potentials*. Cambridge University Press, New York
- Cervone D, Shadel W G, Jencius S 2001 Social-cognitive theory of personality assessment. *Personality and Social Psychology Review* 5: 33–51
- Cervone D, Shoda Y (eds.) 1999 *The Coherence of Personality: Social-cognitive Bases of Consistency, Variability, and Organization*. Guilford, New York

- Cronbach L J, Meehl P E 1955 Construct validity in psychological tests. *Psychological Bulletin* **52**: 281–302
- Geen R G 1997 Psychophysiological approaches to personality. In: Hogan R, Johnson J A, Briggs S R (eds.) *Handbook of Personality Psychology*. Academic Press, San Diego, pp. 387–414
- Goldberg L 1993 The structure of phenotypic personality traits. *American Psychologist* **48**: 26–34
- Greenwald A G, Banaji M R 1995 Implicit social cognition: Attitudes, self-esteem, and stereotypes. *Psychological Review* **102**: 4–27
- Hanson N R 1961 *Patterns of Discovery: An Inquiry into the Conceptual Foundations of Science*. Cambridge University Press, Cambridge, UK
- John O P, Srivastava S 1999 The big-five factor taxonomy: History, measurement, and theoretical perspectives. In: Pervin L A, John O P (eds.) *Handbook of Personality: Theory and Research*. Guilford, New York, pp. 102–138
- Kelly G 1955 *The Psychology of Personal Constructs*. Norton, New York
- Larsen R J 1989 A process approach to personality psychology: Utilizing time as a facet of data. In: Buss D M, Cantor N (eds.) *Personality Psychology: Recent Trends and Emerging Directions*. Springer-Verlag, New York, pp. 177–93
- McAdams D P 1996 Personality, modernity, and the storied self: A contemporary framework for studying persons. *Psychological Inquiry* **7**: 295–321
- McCrae R R, Costa P T 1996 Toward a new generation of personality theories: Theoretical contexts for the five-factor model. In: Wiggins J S (ed.) *The Five-factor Model of Personality. Theoretical Perspectives*. Guilford, New York, pp. 51–87
- McCrae R R, Costa P T 1995 Trait explanations in personality psychology. *European Journal of Personality* **9**: 231–252
- Mischel W 1968 *Personality and Assessment*. Wiley, New York
- Mischel W, Shoda Y 1995 A cognitive–affective system theory of personality: Reconceptualizing situations, dispositions, dynamics, and invariance in personality structure. *Psychological Review* **102**: 246–86
- Nozick R 1981 *Philosophical Explanations*. Belknap Press of Harvard University Press, Cambridge, MA

D. Cervone and G. V. Caprara

Personality Development and Temperament

Personality refers to the profiles of stable beliefs, moods, and behaviors that differentiate among individuals in a particular society. The exact form of expression and the number of profiles will vary with culture, as the number and variety of animal species vary with the ecological setting. New England Puritans would have regarded variation in piety as a major psychological trait, and citizens in Confucian China would have treated loyalty to the father as a central personality characteristic. Neither piety nor filial loyalty are regarded as primary personality traits by contemporary psychologists.

The personality types believed to be most prevalent in contemporary industrialized societies refer to usual styles of social interaction; degree of adherence to the ethical standards of the community; vulnerability to the emotions of anxiety, guilt, and anger; and receptivity to new ideas, beliefs, and attitudes. However, among individuals living in ancient agricultural societies, composed of small, stable geographically separate villages, variation in sociability would be less critical for adaptation than this trait is in the large cities of contemporary, industrialized nations.

The term temperament, on the other hand, refers to the profiles of mood, behavior, and physiology, believed to be under some genetic control, that appear relatively early in life and, together with experience, become the major personality traits of individuals in a particular society (Bates 1987). Temperamental qualities in children believed to be important at the start of the twenty-first century refer to ease of arousal to stimulation; the form and efficiency with which that arousal is regulated; activity level; irritability; form of reaction to unfamiliar people, objects, and situations; and capacity for sustained attention.

Scholars who believe that temperamental categories are applicable to adults have nominated avoidance of danger, seeking of novelty, dependence on social rewards, emotionality, activity, and sociability as primary temperamental characteristics (Buss and Plomin 1984).

1. Temperamental Contributions to Personality

The assumption that a young child's temperament makes a contribution to the older individual's personality has a long history. Hippocrates and Galen, two physicians living in ancient societies who were the founders of the temperamental theory, believed that the relative concentrations of four body humors—blood, phlegm, and yellow and black bile—created within each person a combination of the opposed dimensional qualities of warm versus cool and dry versus moist. The four temperamental types, called melancholic, sanguine, choleric, and phlegmatic, were the products of a particular combination of these two dimensions which were influenced, in part, by local climate and the person's diet. Melancholics were high on the qualities of cool and dry because they possessed an excess of black bile and that is why they were tense, anxious, and depressed. Sanguine individuals were high on the warm and moist qualities because of an excess of blood and were outgoing and optimistic. The choleric was high on the qualities warm and dry because of an excess of yellow bile, and, as a result, was easily angered. The phlegmatic was high on the qualities of cool and moist because of an excess of phlegm and was low in energy and initiative. These concepts remained popular among Europeans and Americans until the end of the nineteenth century.

Interest in the contributions of temperament to

personality receded during the first half of the twentieth century when political forces led to a rejection of the idea that biological processes contributed to personality. However, the contribution of temperament returned when the psychiatrists Alexander Thomas and Stella Chess (1977) nominated nine primary temperamental dimensions in children, along with three synthetic types that they called difficult child, easy child, and the child who is slow to warm up to unfamiliar situations. The difficult child, which was the least frequent category and comprised about 10 percent of the study group, was rather more likely than the other two to develop psychiatric symptoms in later childhood.

2. Reactivity to Stimulation in Infants

Ease of behavioral and physiological arousal to varied forms of stimulation, and the form of regulation of that arousal, are primary temperamental qualities in infants and young children (Rothbart 1989, Kagan 1994). Some infants become easily aroused by sight, sounds, and smells and thrash, cry, babble, or smile at these events. Other infants show minimal signs of arousal to the same stimulation. It is believed that about 20 percent of healthy, four-month-old infants react to varied forms of stimulation with frequent distress and vigorous motor activity. One-fourth of these infants, who are called high reactive, become very subdued, shy, fearful children in the second and third years and are more likely than others to develop anxious symptoms by school age. By contrast, about 40 percent of healthy infants show the opposite profile of a relaxed and minimally distressed reaction to stimulation. About one-fourth of these infants, who are called low reactive, become sociable, bold, and affectively spontaneous preschool children and minimally anxious seven year olds (Kagan 1994). A temperamental vulnerability to anxiety helps to explain why only a minority of children react to a particular stressful event with an anxious symptom, whether the traumatic event is an earthquake, kidnapping, divorce, or abuse. Only a fairly small proportion of children, usually less than 40 percent, react to the traumatic event with a fearful symptom or chronic anxiety.

Social experiences, especially experiences in the family and with peers, determine whether high reactive infants will develop a fearful personality and whether low reactive infants become bold extraverts. Hence, only about 15 percent of children who are extremely shy and anxious in the preschool years become introverted adolescents or adults (Caspi et al. 1988), and both the anxious-shy and the bold-sociable types are under modest genetic control.

There is disagreement as to whether temperament should be viewed as a continuous dimension or as a qualitative category. For example, some regard the

tendency to approach or to withdraw from unfamiliar people and situations as a continuous dimension, while others believe that the child who usually approaches unfamiliar people and events and the child who usually avoids these events belong to separate, discrete categories. Because only about 15 percent of young children are consistently shy, avoidant, and timid in a large number of unfamiliar situations, while most children are subdued in only one type of situation, it is likely that the former child belongs to a qualitatively discrete group.

Most scientists believe that each of the many temperamental types possesses a unique set of inherited physiological features. Although most of these biological features are not yet discovered, it is known that fearful-avoidant children show greater cortical activation of the right frontal, compared with the left frontal, area as measured by absence of alpha activity in the EEG on the right side (Davidson 1994, Fox et al. 1994).

The brain contains over 150 different chemical substances, each of which influences the excitability of particular ensembles of neurons and brain circuits. It is assumed that children inherit different concentrations of these molecules, as well as different densities of the relevant receptors present on neuronal surfaces. As a result children are disposed to different mood and behavioral reactions. One example involves an enzyme called dopamine beta-hydroxylase. This enzyme, which is under genetic control, is required for the final step of the synthesis of norepinephrine. Children who have low levels of this enzyme will necessarily have less norepinephrine in their brain. Because brain norepinephrine renders a child vigilant and may increase the probability of a state of fear to novelty, children with low levels of the enzyme would be expected to be less fearful. Boys who are very aggressive or antisocial, often called conduct disorder, seem to be unafraid of punishment. These children have low levels of this enzyme (Rogeness et al. 1988). Dopamine, another essential brain neurotransmitter, affects motor behavior and cognitive processes mediated by the frontal cortex. Because the frontal cortex mediates planning and control of impulsivity, it is possible that children who inherit low levels of dopamine in the frontal area will be unusually impulsive. The rationale for administering the drug Ritalin to children who have been diagnosed as impulsive or hyperactive is that this drug increases the level of dopamine in the frontal cortex and, therefore, helps children control impulsive and potentially inappropriate behavior.

3. The Role of Environment

Children are exposed to different environments over the course of childhood and these environments influence temperament in different ways. Therefore, there will be a large number of different personality

profiles emerging from the combination of early inherited temperament and experience. Some of the important experiences include the socialization practices of the parents, the relationship to brothers and sisters, success or failure in school, and the quality of relationships with peers. The social class of the child's family always exerts continuous influence on the personality that develops. A large birth cohort of children born on the island of Kauai followed from birth to the fourth decade of life revealed that the social class of the child's family and early temperament were the two most important influences on the adult personality. Minimally irritable infants who were born to highly educated families were least likely to encounter school failure, display delinquent behavior, or to develop symptoms requiring referral to a physician or psychologist (Werner 1993).

3.1 Types of Challenges

The most frequent personality profiles that develop in a particular culture are a function of the challenges to which the children in that culture must accommodate. Children in most settings deal with three classes of challenge: (a) unfamiliarity, especially unfamiliar people, tasks, and situations; (b) requests by legitimate authority for conformity to and acceptance of their standards, especially standards for competent task performance; and (c) domination by and/or attack from peers. In addition, all children must learn to control two families of emotions: anxiety, fear and guilt, on the one hand, and, on the other, anger, jealousy, and resentment.

4. Temperament and Moral Emotions

Variation in the intensity of the moral emotions of shame, guilt, and anxiety can be influenced by the child's temperament and family experience. Shy, fearful children raised by mothers who used reasoning to socialize the child showed clearer signs of a strict conscience than most other children (Kochanska 1993). Sensory information from the body ascends in the spinal column and projects to the ventromedial surface of the prefrontal lobe. Children who experience consciously more intense bodily stimulation might be more vulnerable to the moral emotions than those who experience less stimulation. School-age boys who were low reactive infants and fearless children often have low sympathetic tone in the cardiovascular system. Boys who are members of this temperamental group growing up in homes with nurturant parents who socialize achievement and the control of aggression are likely to become popular group leaders. The same type of child raised by parents who did not socialize asocial behavior and who played with peers who committed asocial acts are at slightly higher risk for becoming delinquents (Caspi et al.

1995). It is believed that a small number of criminals who commit violent crimes (probably fewer than 5 percent of all criminals) were born with a special temperament characterized by impulsivity and difficulty in controlling behavior.

5. Ethnicity and Temperament

Temperamental variation among ethnic groups is a delicate issue because of the ethnic strife that exists in many parts of the world. The average genetic distances for 100 different alleles monitoring physiological function measured in geographically separate populations were largest when Asians, Africans, and European-Caucasians were compared with each other. It is likely that some of these alleles make a modest contribution to variation in temperamental qualities. Asian-American infants, compared with European-Americans, are calmer, less labile, less vocal, and more easily consoled when distressed (Kagan et al. 1994). It is relevant that Asian-American adult patients with a psychiatric diagnosis require a lower dose of therapeutic drug than European-American patients (Lin et al. 1986).

6. Personal Control

The renewed interest in temperament has generated discussion over the degree of responsibility each person should have for his or her behavior. Western society has, in the past, been Puritan in its commitment to the belief that all individuals have sufficient will power to control their behavior in most situations. This view has been eroded since the 1950s as more citizens have become willing to excuse some asocial actions as due to temperamental factors that were not within the agent's sphere of control. It is not obvious that this permissiveness is more adaptive for a society than the traditional assumption that one of the significant products of human evolution is the capacity to monitor one's behavior.

See also: Personality Assessment; Personality Psychology: Methods; Temperament and Human Development

Bibliography

- Bates J E 1987 Temperament and infancy. In: Osofsky J E (ed.) *Handbook of Infant Development*, 2nd edn. Wiley, New York, pp. 1101-49
- Buss A H, Plomin R 1984 *Temperament—Early Developing Traits*. Erlbaum, Hillsdale, NJ
- Caspi A, Elder G H, Bem D J 1988 Moving away from the world. *Developmental Psychology* **24**: 824-31
- Caspi A, Henry B, McGee R O, Moffitt T, Silva P A 1995 Temperamental origins of child and adolescent behavior problems from age 3 to 15. *Child Development* **66**: 55-68

- Davidson R J 1994 Asymmetric brain function and affective style in psychopathology. *Development and Psychopathology* 6: 741–58
- Fox N A, Calkins S D, Bell M A 1994 Neuroplasticity and development in the first two years of life. *Development and Psychopathology* 6: 677–96
- Kagan J 1994 *Galen's Prophecy*. Basic Books, New York
- Kagan J, Arcus D, Snidman N, Yu-feng W, Hendler J, Greene S 1994 Reactivity in infancy. *Developmental Psychology* 30: 342–5
- Kochanska G 1993 Toward a synthesis of parental socialization and child temperament in early development of conscience. *Child Development* 64: 325–47
- Lin K M, Poland R E, Lesser I N 1986 Ethnicity and psychopharmacology. *Culture, Medicine, and Psychiatry* 10: 151–65
- Rogeness G A, Maas J W, Javors M A, Masedo C A, Harris W R, Hoppe S K 1988 Diagnoses, catecholamines, and plasma dopamine-beta-hydroxylase. *Journal of the American Academy of Child and Adolescent Psychiatry* 27: 121–5
- Rothbart M K 1989 Temperament in childhood: A framework. In: Kohnstamm G A, Bates J E, Rothbart M K (eds.) *Temperament in Childhood*. Wiley, New York, pp. 59–76
- Thomas A, Chess S 1977 *Temperament and Development*. Brunner-Mazel, New York
- Werner E E 1993 Risk, resilience and recovery. *Development and Psychopathology* 5: 503–15

J. S. Kagan

Personality Development in Adulthood

The study of personality is arguably the broadest subdiscipline in psychology in that it aims to understand how the whole person functions in the world (see *Personality Psychology*). Historically, personality psychology has concerned itself with grand theories about human behavior, questions about character, identity, and morality. Most of the empirical research on personality published over the past two decades fails to consider potential development in adulthood, reflecting, in part, an assumption that personality changes little in adulthood. Nevertheless there has been a long-standing interest in whether—and, if so, how—people change in systematic ways in the later years of life. The following sections summarize the research traditions and some of the central findings generated in the field of adult personality development.

1. Defining Personality Development

In part because personality draws so broadly from so many areas of psychology, including cognition, emotion, psychopathology, and motivation, consensus in the field over basic definitions is difficult to obtain (see *Personality Theories*). Even the preferred units of study (e.g., traits, behaviors, or psychological processes)

continue to be debated. Most recently, questions are being raised about the degree to which personality is bounded within the individual or is better represented in transactions between individuals and broader collective units. Importantly, different conceptual approaches to the study of personality lead to different predictions about and evidence for personality change in adulthood. Whereas personality development is considered by some to unfold naturally and unidirectionally from temperaments inherited at birth, other personality psychologists emphasize the need to consider a complex interplay of factors that contribute to personality development, including temperamental inheritance but more importantly exposure to different types of environments, acquired beliefs and expectations and the capacity for self-regulation (Bandura 1989).

In part, evidence for stability or change in personality in adulthood reflects which facet of human functioning is studied. To be clear, there is no dispute that people change in adulthood. They do. Adults are inevitably changed in idiosyncratic ways by the life experiences they encounter, including major life events, such as becoming a parent, or less dramatic but persistent experiences associated with, for example, the pursuit of a particular career and the consequent development of a particular type of expertise. However, change in adulthood is not automatically considered *personality development*. Rather, changes must be enduring, systematic (i.e., nonrandom), and predictable by age or life stage.

2. Major Approaches to Adult Personality Development

Conceptions of adult personality development have evolved out of two very different traditions in psychology; clinical psychology and life-span developmental psychology. The approaches and the findings about systematic change in adulthood reflect these different paradigmatic approaches to the study of lives. Due to considerable overlap between studies of individual differences and clinical psychology, many of the oldest and most influential theories of personality, most notably psychoanalytic psychology but also ego psychology and interpersonal psychology, were developed based on clinical observations of patients. Although strikingly different in its basic tenets, social cognitive theory of personality also evolved out of close connections between clinical and personality psychology, essentially addressing differences between normal and abnormal processing involved in basic psychological functioning. Thus, the oldest approaches to personality were tied closely to understanding psychopathology.

In contrast to the traditional individual difference approach, life-span psychology was born only in about

the 1970s and reflects the view that human development is a continuous adaptive process (Baltes and Goulet 1970). Life-span psychology aims to identify and illuminate normal developmental changes in all areas of psychological functioning, including but not limited to personality, from birth until death. Perhaps most notably, life-span psychology is distinguished by the presumption that human growth is at no time during the life course complete. Consequently, the different presumptions inherent in the two approaches direct attention to different research foci. Whereas traditional adult personality psychologists ask whether traits acquired in childhood persist across adulthood, whether particular personality crises present themselves at particular stages of life, or how personality disorders are transformed in later life, life-span psychologists are more likely to target specific age-related issues, like whether people grow wiser with age, whether conceptions of the self grow more or less complex over time, and whether self-regulatory processes change in systematic ways over time. As noted above, whether one finds evidence for stability or change depends importantly on where one looks and the particular questions one poses.

The next sections include a birdseye view of the earliest approaches to personality development and brief synopses of research on personality development deriving from the trait approach to personality development and from life-span developmental psychology.

3. Early Stage Approaches to Adult Personality Development

Following the tradition established by child developmentalists, early thinking about adult personality development was rooted in stage theories. Whereas Freud's (see *Freud, Sigmund (1856–1939)*) psychosexual stage model of personality suggested that personality was formed in early childhood and that, barring long-term psychotherapy, was highly resistant to change, his skeptical follower, Carl Jung (see *Jung, Carl Gustav (1875–1961)*), argued that the most interesting aspects of personality do not develop fully until middle age. Jung believed that only after basic biological and reproductive issues are resolved in early adulthood, are people freed to engage in more psychic pursuits. Jung posited that whereas in early adulthood biological imperatives predominate and demand adherence to gender roles, in mid-life feminine and masculine sides of people grow more integrated and spiritual concerns grow more salient. As people age, feelings and intuitions come to dominate thoughts and sensations. Thus, albeit in rather unspecified ways, Jung advanced the position that people developed in adulthood and that spirituality played an increasingly central role. Jung wrote far less about advanced old

age, but suggested that people predictably turned inward and deteriorated psychologically as they approached the end of life and wrestled with fears about death.

In the 1930s and 1940s, in Europe and the US, several stage theories, such as that offered by Charlotte Bühler, concretized thinking about adult personality development and allowed for empirical tests of theoretical predictions. Of these, Erik Erikson's (see *Erikson, Erik Homburger (1902–94)*) stage theory had the most enduring influence. Like Freud and Jung, Erikson was a classic ego psychologist and his theory was grounded in the psychoanalytic tradition. However, rather than focus on psychosexual needs, Erikson argued that human needs for psychological intimacy fueled systematic development that continued into old age. According to this epigenetic theory, people pass through a fixed sequence of stages during life each of which requires the successful resolution of a central psychic crisis. In early adulthood, the central issue involves the establishment of intimacy. In middle age, generativity, namely the passing on of power and knowledge to younger generations, is achieved or failed. In old age ego-integrity vs. despair (self-acceptance of a life lived or regret and dismay) is the focal crisis in life.

In the 1960s and 1970s, major research projects aimed at profiling adult development were undertaken at Yale (Levinson 1978) and Harvard (Vaillant 1977), which continued the stage theory tradition in the US and longitudinal studies undertaken in the 1930s began to come of age. As research participants in the Stanford Terman study of 'gifted' children, for example, entered adulthood, researchers began to examine connections between childhood and adult personality. The Child Guidance and Oakland Growth Studies undertaken at the University of California at Berkeley offered resources by which to examine predictable change. At the University of Chicago, a group of social scientists, including Bernice Neugarten, Morton Lieberman, and David Guttman, formalized the study of life-span personality development.

In the end, however, empirical support for stage theories failed to withstand empirical tests. Although interesting patterns were displayed in some very homogenous samples, identified developmental patterns failed to generalize broadly. Critics of stage theories claimed that the patterns that did emerge reflected the influence of consistent surrounding social structures around highly selected groups of people, not human development. Virtually all of the longitudinal studies included predominately (if not exclusively) white, middle-class individuals, often only males. In the 1980s, Costa and McCrae (1990) essentially waged a war against stage theories declaring that personality does not change systematically in adulthood. In a decade dominated by the *trait* approach to personality, the central and reliable finding

of the 1980s was that personality changes little after the age of 30 years.

4. *The Trait Approach to Adult Personality Development*

Traits are continuous variables represented by broadly encompassing lexical terms that account for individual differences (John 1990). Traits—such as shy, lively, outgoing, anxious, and intelligent—are conceptualized as predispositions within individuals to behave in certain ways manifest across a wide variety of situations. Gordon Allport (see *Allport, Gordon W (1897–1967)*) argued that cardinal traits are those around which a person organizes life (self-sacrifice). Central traits (e.g., honesty) represent major features and secondary traits are specific traits that help to predict behavior more than underlying personality (e.g., dress type, food preferences). Allport's definition is compatible with modern trait and temperament approaches to personality which attempt to describe people in terms of one or more central features of the person.

Personality psychologists in the trait tradition seek to identify the traits along which people differ and to explore the degree to which these traits predict behavior. Many taxonomies of traits have been offered over the years, but unquestionably the five-factor model is most widely accepted today (see also *Personality Structure*). Based on factor analysis of self-descriptions, the five traits that emerge reliably across many studies of Europeans and Americans are: (a) openness to experience, (b) conscientiousness, (c) extraversion, (d) agreeableness, and (e) neuroticism.

Traits and temperaments appear to be relatively stable through the second half of life (Costa and McCrae 1990). It appears that beyond the age of 30, extraverts remain extraverts and neurotics remain neurotics. Trait theorists have found reliable evidence for stability in personality well into old age. This finding emerges whether researchers ask individuals to describe themselves repeatedly over time or, alternatively, ask significant others, like spouses, to describe those same individuals repeatedly (Costa and McCrae 1990). It should be noted that even though persistent rank-order differences remain the same, there is some recent evidence that modest mean level changes may appear, with older adults scoring slightly higher than younger adults on agreeableness and conscientiousness and slightly lower on neuroticism, extraversion, and openness to experience (McCrae et al. 1999). Importantly, similar findings come from studies sampling Asian and European populations. However, identified changes are quite small. Overall, there is remarkable consistency in the characteristics that distinguish individuals from one another over time. There is some evidence that the core set of traits that differentiate people are genetically based and exert

their influence throughout the life course (Gatz 1992). Genetic influence is as strong in old age as early adulthood.

In summary, researchers adopting a trait approach to personality development find that along at least some of the important dimensions of personality, there is little change in personality well into old age. Critics of a trait approach, however, argue that traits communicate little about how people manage their lives in day-to-day life and because of the broadband focus exaggerate the consistency of behavior across time and situations. They criticize the trait approach for failing to better predict behavior and redirect focus to specific strategies (e.g., how an individual cognitively appraises a situation; expectancies, subjective values, self-regulatory systems, and competencies). Life-span approaches—influenced strongly by the social cognitive theory of personality (Bandura 1989)—view individuals as agentic creatures who shape their own environments (see also *Interactionism and Personality; Self-regulation in Adulthood*).

5. *Life-span Approaches to Adult Personality Development*

Rather than focus on taxonomies of personality, life-span developmental psychologists view development as a dynamic process aimed at adaptation (Baltes 1987). Two principal stays of life-span theory speak directly to personality. The first states that adaption is always time and space bound. In other words, behavioral adjustment must occur within a particular environmental and social niche. In life cycle context, what is adaptive in infancy and early childhood may not be adaptive in adolescence. Stranger anxiety, for example, may serve a highly adaptive function in infancy because it motivates dependent creatures to stay in close proximity to caregivers. It may also facilitate attachment to primary adult figures, a key developmental task of early life. Yet, stranger anxiety among adults is clearly maladaptive. Similarly, it can be argued that pursuing multiple prospective mates is adaptive in adolescence and early adulthood as people 'practice' intimate relationships but less so in middle and old age at which point emotional investment in a select few may hold greater gains than the continual exploration of all possible mates.

The second stay of life-span theory is that development inevitably demands selection (Baltes and Baltes 1990). In order for specialized (i.e., effective) adaptation to occur within a particular social, historical and physical niche, active and passive selections must be made. As people age, they come to have greater choice in the selection of environments and select environments that support their self-views. Throughout adulthood, people actively construct skills and hone environments to meet selected goals. There is good evidence that people narrow their social

spheres with age, for example, forming increasingly well contoured social convoys that accompany them throughout life (Carstensen et al. 1999). Caspi and Herbener (1990) found that people tend to choose spouses similar to themselves, and further show that people who have spouses similar to themselves are less likely than people with dissimilar spouses to display personality change in adulthood. Thus, it may be that stability is maintained across the life course because people actively create environments that maintain stability.

Finally, life-span theory holds that development is never fully adaptive because adaptation to one set of circumstances inevitably reduces flexibility to adapt to another. In this way development always entails gains and losses. Subsequently, life-span theory obviates the presumption that antecedent losses are the only or even the primary reasons for changes that occur with age and examines how people's relative strengths and weaknesses at different points in the life cycle influence successful adaptation.

5.1 Personality Development from a Motivational Perspective

Life-span developmental approaches, because they are rooted in adaptation, lead naturally to consideration of the ways that goals and goal attainment may change throughout the life course (Baltes and Baltes 1990, Brandtstädter et al. 1999, Carstensen et al. 1999) (see also *Adulthood: Developmental Tasks and Critical Life Events*). Motivational psychologists presume that there is continuity in basic human needs for competence, relatedness and autonomy across the life course. There is every reason to expect, for example, that regardless of age, people seek to control their worlds (see also *Control Behavior: Psychological Perspectives*).

Researchers who take a goal-focused approach have brought a different set of findings to bear on discussions of personality and aging, showing that goals and preferences do change with age and influence behavior. Carstensen and co-workers, for example, have shown that the perception of time left in life importantly influences social goals. Because aging is inextricably and positively associated with limitations on future time, older people and younger people differ in the goals they pursue (Carstensen et al. 1999). Older people are more likely to pursue emotionally meaningful goals whereas younger people are more likely to pursue goals that expand their horizons or generate new social contacts. Brandtstädter et al. (1999) argue that people adjust goal strivings to accommodate external and internal constraints placed on goal achievement at different points in the life cycle; a central finding coming out of this line of work, for example, is that older people respond to the loss of resources in advanced age by downgrading the importance of some previously desirable goals.

5.2 Emotion and Personality

Another way of conceptualizing personality, which is particularly conducive to life-span approaches, places emotional experience and emotional regulation at the core. In this view, emotions are not related simply to personality, they are the essence of personality (Carstensen et al. in press, Rothbart and Derryberry 1981) (see also *Adulthood: Emotional Development*). The emotions people feel when they face challenges, and the effectiveness with which they learn to regulate their emotions, are the cardinal components of personality development, forming the basis of individual differences in persistent tendencies to behave, think and feel in day-to-day life. Individual differences in the propensity to experience specific emotions influence not only the psychological and biological reactivity of the person in the moment, but come to influence conscious choices about preferred environments, behavioral styles, and also determine the social partners to which people are drawn.

Ryff (1995) has taken a differentiated approach to understanding emotions and well-being across the life-span. Rather than calculating global positive and negative affect as indicators of psychological well-being, Ryff conceptualizes well-being in terms of self-acceptance, environmental mastery, purpose in life, personal growth, positives relations with others, and autonomy. The dimensions appear to have differential relationships with age, with older adults scoring higher than younger adults on Environmental Mastery and Autonomy, but lower on Purpose in Life and Personal Growth. There also appear to be life-span developmental trajectories concerning the relationship between people's conception of their present status and their ideal selves along these dimensions. Older people tend to have less distance between their actual and ideal selves than do younger adults.

An emotion approach to personality has particularly intriguing implications for adult development because in adulthood emotions appear to grow more complex (Carstensen et al. in press) and emotion regulation appears to improve (Gross et al. 1997). With advancing age, the emotion-cognition-personality system also appears to become more-differentiated, with emotions becoming linked to an ever-increasing array of cognitions. To the extent that such changes influence motivation (e.g., Izard and Ackerman 1998), modify thinking and reasoning (Labouvie-Vief et al. 1989), or result in qualitative changes in subjective well-being (Ryff 1995), personality is importantly influenced.

5.3 Wisdom and Resilience

As noted above, the focus on adaptive development inherent in life-span approaches generates questions about the ways in which aging people adjust to changing resources and changing contexts. There is a

delicate balance between gains and losses that occurs in the second half of life that have important implications for personality. As people enter advanced ages, nearly inevitably they encounter increasingly difficult challenges, including the deaths of friends and loved ones, assaults on physical health, and threats to social status (see also *Coping across the Lifespan*). At the same time as experience in life increases, perspectives change. In some ironic way, the familiarity of losses may even make losses easier to take. Considerable focus in lifespan psychology, thus, has been on the ways that people effectively adjust in later adulthood. Resilience (Staudinger et al. 1995) and wisdom (Staudinger 1999) have been particular targets of interest because they involve the use of age-based experience to compensate for losses in circumscribed domains. Studies of wisdom, for example, show that contrary to popular lore, wisdom is unrelated to age in adulthood (Staudinger 1999). Even though experience-based knowledge does increase with age, wisdom requires a complex array of abilities that draw on multiple functions, some of which decline. Under optimal conditions, old age may be the time in life for wisdom to best emerge, but it does not do so normatively.

6. Integration of Approaches and Findings about Adult Personality Development

Does continuity or change in personality best characterize adulthood? The answer is a qualified 'yes' to both continuity and change. Along some basic descriptive dimensions, such as openness to experience and extraversion, people remain remarkably consistent in adulthood. However, in other domains just as central to personality, such as motivation and adaption, there is evidence for systematic change across adulthood. Goals change predictably with age, emotion regulation appears to improve, and well-being takes on different qualities.

7. Future Directions

The fundamental challenge that confronted personality researchers a century ago remains largely the same today: predicting and understanding individual differences in behavior. Students of adult personality development face the additional challenge of understanding how the differences that distinguish one person from another may change systematically over time. Although considerable progress has been made, the bulk of empirical findings generated simply show that prevalent assumptions about adult personality development in the twentieth century were wrong. For example, people do not appear to pass normatively through a fixed series of stages; and along broadband

dimensions characterized as basic traits, people change very little in the second half of life. Approaches that focus on motivation and emotion are newer, but initial findings suggest that they may shed considerable light on ways that individuals change in middle and old age. Finally, at the time of this writing, the human genome project very recently was declared complete. Few, if any, scientists expect that genetic findings alone will shed much light on personality. However, they may well help to do away with customary language and algorithms (such as 'heritability coefficients') that have given credence tacitly to the idea that environmental and biological influences can be cleanly separated. Whereas in past decades, substantial discussion has centered around whether biology or environment was most influential in personality development, personality researchers will now begin to address the more interesting and more important puzzle which lies in the interaction between the two.

See also: Adult Development, Psychology of; Ego Development in Adulthood; Personality Development and Temperament; Personality Development in Childhood; Personality Psychology; Personality Theories

Bibliography

- Baltes P B 1987 Theoretical propositions of life-span developmental psychology: On the dynamics between growth and decline. *Developmental Psychology* **23**: 611–26
- Baltes P B, Baltes M M 1990 Psychological perspectives on successful aging: The model of selective optimization with compensation. In: Baltes P B, Baltes M M (eds.) *Successful Aging: Perspectives from the Behavioral Sciences*. Cambridge University Press, New York, pp. 1–34
- Baltes P B, Goulet L R 1970 Status and issues of life-span developmental psychology. In: Goulet L R, Baltes P B (eds.) *Life-span Developmental Psychology: Research and Theory*. Academic Press, New York, pp. 4–21
- Bandura A 1986 *Social Foundations of Thought and Action: A Social Cognitive Theory*. Prentice Hall, Englewood Cliffs, NJ
- Bandura A 1989 Human agency in social cognitive theory. *American Psychologist* **44**: 1175–84
- Brandtstädter J, Wentura D, Rothermund K 1999 Intentional self-development through adulthood and later life: Tenacious pursuit and flexible adjustment of goals. In: Brandtstädter J, Lerner R (eds.) *Action and Self Development*. Sage, Thousand Oaks, CA
- Carstensen L L, Charles S T, Isaacowitz D, Kennedy Q (in press) Emotion and life-span personality development. In: Davidson R, Scherer K (eds.) *Handbook of Affective Science*. Oxford University Press, Oxford, UK
- Carstensen L L, Isaacowitz D M, Charles S T 1999 Taking time seriously: A theory of socioemotional selectivity. *American Psychologist* **54**: 165–81
- Carstensen L L, Pasupathi M, Mayr U, Nesselroade J 2000 Emotional experience in everyday life across the adult life span. *Journal of Personality and Social Psychology* **79**: 644–55

- Caspi A, Herbener E S 1990 Continuity and change: Assortative marriage and the consistency of personality in adulthood. *Journal of Personality and Social Psychology* **58**: 250–8
- Costa P T Jr., McCrae R R 1990 *Personality in Adulthood*. Guilford Press, New York
- Gatz M, Pederson N, Plomin R, Nesselroade J 1992 Importance of shared genes and shared environments for symptoms of depression in older adults. *Journal of Abnormal Psychology* **101**: 701–8
- Gross J, Carstensen L L, Pasupathi M, Tsai J, Götestam Skorpen C, Hsu A 1997 Emotion and aging: Experience, expression and control. *Psychology and Aging* **12**: 590–9
- Izard C, Ackerman B P 1998 Emotions and self-concept across the life span. In: Schaie K W, Lawton M P (eds.) *Annual Review of Gerontology and Geriatrics: Focus on Emotion and Adult Development*. Springer, New York, Vol. 17, pp. 1–26
- John O 1990 The big-five factor taxonomy: Dimensions of personality in the natural language and questionnaires. In: Pervin L A (ed.) *Handbook of Personality: Theory and Research*. Guilford Press, New York, pp. 66–100
- Labouvie-Vief G, DeVoe M, Bulka D 1989 Speaking about feelings: Conceptions of emotion across the life span. *Psychology and Aging* **4**: 425–37
- Levinson D 1978 *The Seasons of a Man's Life*. Knopf, New York
- McCrae R R, Costa P T Jr., de Lima M P, Simoes A, Ostendorf F, Angleitner A, Marusic I, Bratko D, Caprara G V, Barbaranelli C, Chae J-H, Piedmont R L 1999 Age differences in personality across the adult life-span. *Developmental Psychology* **35**: 466–77
- Rothbart M, Derryberry D 1981 Development of individual differences in temperament. In: Lamb M E, Brown A L (eds.) *Advances in Developmental Psychology* Vol. 1, Erlbaum, Hillsdale, NJ, pp. 37–86
- Ryff C 1995 Psychological well-being in adult life. *Current Directions in Psychological Science* **4**: 99–104
- Staudinger U M 1999 Older and wiser? Integrating results from a psychological approach to the study of wisdom. *International Journal of Behavioral Development* **23**: 641–4
- Staudinger U M, Marsiske M, Baltes P B 1995 Resilience and reserve capacity in later adulthood: Potentials and limits of development across the life span. In: Cicchetti D, Cohen D (eds.) *Developmental Psychopathology. Vol. 2: Risk, Disorder, and Adaptation*. Wiley, New York, pp. 801–47
- Vaillant G E 1977 *Adaptation to Life* 1st edn. Little, Brown, Boston

L. L. Carstensen

Personality Development in Childhood

To understand personality development in childhood, we first define personality, and how it is structured. Then we discuss how the definition is typically translated into research, noting some curious omissions and quirks in the literature. Next, we consider the raw materials of personality and how they are changed over time. Once this is completed, we discuss outcomes of research on personality development in children.

1. Personality Development as the Organization of Individual Differences

First, perhaps the most basic question is what is personality? Common personality characteristics can be shared by groups of people—say, women, Russians, three-year-old boys—but usually we think of personality in terms of defining aspects of individual persons. In modern theory, personality is concerned with the organization of each person's unique configuration of individual differences. When personality is defined this way, rather than one difference at a time (e.g., introversion), it is possible to recognize that over time conflict within a person may occur due to competing and sometimes incompatible needs (see *Personality Structure*). The self-organization of the diverse elements that live under a common skin is motivated by the need to resolve intrapersonal conflicts and to adapt, or adjust, to the world beyond the individual. For example, we would expect an intelligent, introverted child to make compromises in the service of social adjustment different from those made by a similar child who is intelligent and extraverted. Personality is concerned with the accommodations and compromises each individual must make among the competing demands of these differences. The adjustment processes occur over time, and involve genetics, maturation, and learning. In this light, personality development in childhood can be seen as one segment in a life-long process that leads to the formation of a stable adult personality structure (see *Personality Development in Adulthood*). Common but untested assumptions are that processes of personality development are more dynamic in childhood than in adulthood, and that personality structure is more fluid and open to change in the former than in the latter. Personality development does not necessarily stop after childhood, but that period is regarded as a time of special sensitivity for the formation of the broad structure.

Second, we need to recognize how the definition of personality is typically translated into research. There are historical traditions and precedents in the personality development literature that influence the way professionals engage in research in the area. For example, the personality literature does not usually regard individual differences in ability in general, and intelligence in particular, as part of its domain, despite the illustration given in the preceding paragraph and the obvious importance of intelligence to adaptation. Similarly, differences in attachment are not well integrated into the rest of the personality development literature, despite the importance of attachment processes for adaptation. For another (but related) example, readers will sometimes encounter a distinction between 'personality development' and 'emotional development.' Some early writers used the term emotional development to focus on attachment and the psychodynamic approach to personality devel-

opment, and to separate it from other kinds of personality development. Detailed historical analyses make some of these oddities understandable, but such analyses are beyond the scope of this review (see *Infancy and Childhood: Emotional Development*). The point here is that definitions must be translated into measures and procedures, and historical precedents can lead to potential biases in the translation.

Third, developmental processes for each individual occur within a specific historical time, kinship system, social learning environment, and cultural context. Some aspects of personality are tied directly to such contextual factors (e.g., parental divorce), and leave residues. These life experience-based residues represent a legitimate aspect of personality. In the past, residues of experience relevant to personality have been conceptualized mainly in terms of 'learning,' or 'socialization' (see *Social Learning, Cognition, and Personality Development*). More recently, another approach has emerged. Residues of experience relevant to personality can be conceptualized in cognitive and information-processing terms. For example, cultures differ in the ways they teach children skills. In some cultures, mixed-age peer groups engage in competitions, with older peers teaching younger children. In those cultures, substantive knowledge acquired during peer interactions in certain domains is fused to specific beliefs about peer relations and social interaction that is missing in other cultures.

Fourth, given the potentially powerful influence of social learning and sociohistorical contexts, it is a major challenge for personality development theory to identify panhuman substantive dispositions. That is, in the past the majority of researchers have assumed that (a) personality development in all human children can be characterized by individual differences in a small set of enduring characteristics; and that (b) these characteristics induce the children to respond in consistent ways over time and across settings. These two assumptions may not be justified, and at the least they need to be evaluated empirically. Phrased more positively, personality development theory needs to integrate findings of cultural diversity and variation in children's personality with the assumptions about the nature and assumed roles for panhuman dispositions. To accomplish this task, researchers must select a small number of dispositions and enduring characteristics from the vast array of human differences that will prove to be important for theory or for life outcomes. Ideally, the ones chosen for attention in childhood will be broad in scope, related to competence in adulthood, and likely to become part of later personality structure. On an *a priori* basis, we might expect all cultures to be concerned with the development characteristics associated with success in the two major domains of 'love' (sex roles, reproduction, kin relations) and 'work' (reliability, rule compliance). It is possible, of course, that each culture will encourage the development of somewhat different configurations of characteristics

within these two major domains, based on the ecological and social structures within which they live (see *Gender Differences in Personality and Social Behavior*).

2. Four Perspectives on the Substance of Personality and Development

To investigate the substantive nature of personality development, some further 'simplifying' assumptions are required. There have been four major perspectives on the identification of substantive characteristics in personality development (see Caspi 1998). The first major perspective is the most direct approach to childhood personality because it accepts it as a biological 'given.' It involves temperament, and focuses on childhood and infancy. Temperament, an important concept in developmental psychology, refers to early appearing, stable individual differences derived from the constitutional or biologically determined makeup of individuals (see *Temperament and Human Development*). It represents the presumed biological, emotional core around which later personality coalesces. For example, children show important individual differences in fearfulness or motor activity, and these may have heritable components. These are the kinds of differences that warrant attention developmentally because in theory they could constrain ontogenetic processes, and the kinds of experiences individuals accumulate as they move through life. Thomas and Chess (1977) stimulated modern interest in childhood temperament when they identified nine categories of differences based on clinical experiences with very young children. Examples are 'activity level' and 'persistence.' Second-order constructs were proposed to reduce the nine dimensions to three ('easy,' 'difficult,' and 'slow to warm up') based on factor analyses. Subsequent research showed that the original nine dimensions could not be recovered from item-level factor analyses. Instead, five to seven factors appeared. We will return to this perspective later.

Second is the social-cognitive process perspective of Mischel (1999). Researchers could identify individual differences in an enduring skill or tendency in children that could shape the ways children interact with their physical and social environments. These skills need not necessarily be conceptualized as traits (see *Personality and Social Behavior*). The cumulating effect of the skill on the child's outcomes could in theory have important long-range consequences (e.g., Kochanska et al. 1998). For example, the skills associated with delay of gratification are potentially relevant to both 'love' and 'work,' so they represent an especially good substantive characteristic to investigate as an aspect of personality development. Consistent with this perspective, Mischel et al., (1988) found that preschool children who are skilled in delay of gratification appear to interact systematically with the environment in

ways that allow them to achieve tasks more efficiently than their peers. Over time and settings, these skilled children accumulate accomplishments that allow them as adolescents to have higher social competencies and educational levels, and eventually as adults to achieve higher socioeconomic standing than other adults.

Third is the 'developmental targets' perspective (Caspi 1998), which is perhaps a meta-perspective because it can include both dimensional and typological theories under its umbrella. The goal here is to identify developmental antecedents of dispositions that are parts of stable adult personality structure. For example, extraversion and neuroticism are two major dimensions in adult personality structure, and represent developmental targets. It is important to know how these dimensions are formed developmentally. High levels of fearfulness or motor activity in childhood may be antecedents of adult neuroticism or extraversion, respectively. Once an empirical relation is demonstrated, then theoretical process mechanisms could be offered to explain how motor activity may be transformed or channeled into the targeted difference (e.g., extraversion). In effect, adult personality structure provides the target end states, and research is directed to finding its precursors. This approach requires knowledge of major, reliably appearing, and enduring adult personality structures (e.g., extraversion, neuroticism), but also empirically established antecedents that are not associated uniquely with configurations of sociohistorical events or cultural practices. (Of course, researchers (e.g., Elder 1999, McCrae et al. 1999) are interested in these socio-historical and cultural patterns as a special class of personality phenomena.) Until recently, there was no consensus on the nature of adult personality structure, much less antecedents, that could meet these criteria, so personality development based on developmental targets was inhibited.

The fourth perspective could be regarded as a specialized typological version of the case outlined in the preceding paragraph. It is the person-centered, categorical approach of Block (1971). Unlike the researchers who assume personality is structured in terms of continuous underlying dimensions, the typological approach assumes that persons are better characterized as belonging in discrete categories depending on the configuration of attributes. Based on extensive empirical data on personality descriptions of persons as they moved from early adolescence to adulthood, Block had trained clinical judges sort descriptions into piles (Q-sorts), using a continuum from least characteristic to most characteristic. This Q-sort technique is person centered because each attribute is compared with other attributes in the same person. Statistical analyses can then be used to identify clusters of persons with similar Q-sort configurations. At least three adult types were found, and can be described as (a) ego-resilients, who are well functioning in all major domains; (b) overcontrollers, who were

shy and lacked many interpersonal skills; and (c) undercontrollers, who were hostile, disagreeable, and showed little concern for others. These types could be regarded as developmental targets for developmental research. Longitudinal research with children (Hart et al. 1997) showed that three personality types could be replicated, suggesting avenues for theories of personality development. The largest category (ego-resilients) represents persons who are well adjusted; the other two smaller categories represent persons whose adjustment suffers from either overcontrol or undercontrol. At the least, such outcomes suggests that personality development should focus substantively on resilience and control. What is less clear is whether these types represent good developmental targets, in the sense of discrete natural classes, or are unique configurations of persons along some other underlying dimensions. It also is not clear whether these types represent the personality development of females as well as they do males (Caspi 1998).

3. Personality Development and the 'Big Five'

Let us move our discussion from perspectives to further research. In a series of pioneering papers in personality development, Digman and his colleagues in Hawaii began investigating the structure of child personality in teachers' ratings. Digman selected his variables to cast a wide net, and in some studies included assessments of intelligence, peer ratings, and various aptitude and personality batteries. He appeared to have no overriding theoretical agenda. At the time much of his data were collected, Digman (1994) suggested that he was an advocate for a model of child personality that included no less than ten factors. When his data were analyzed, however, a five-factor solution was reliably found (e.g., Digman and Inouye 1986). Digman recognized that the pattern he found in the child personality data was consistent with a similar pattern found in adults (e.g., Digman 1990). Perhaps the developmental targets approach could be useful after all. The time was ripe for a union between personality development researchers and researchers working with adult personality structure.

The number five seems to have a certain magical quality for personality researchers. The five-factor model, or the more descriptively accurate label 'Big Five' (BF) approach, currently occupies center stage in adult personality research (see *Personality Theories*). When people are given enough time and freedom to express their evaluations of others, five broad dimensions reappear reliably. These five dimensions are abstract 'supertraits' that summarize lower level patterns of individual differences. A useful mnemonic for remembering the dimensions of the BF is the acronym OCEAN: openness, conscientiousness, extraversion, agreeableness, and neuroticism. (The

memory aid does not convey the relative size or importance of these dimensions: extraversion and agreeableness are probably larger factors than the others.)

Two leading temperament researchers, Halverson and Kohnstamm recognized the need to bring child personality development researchers together with leading researchers working on the structure of adult personality, and they organized a conference. Papers presented at this meeting, and a few others, were published in a subsequent milestone volume entitled *The Developing Structure of Temperament and Personality from Infancy to Adulthood*. The chapters in Halverson et al. (1994) suggested possible links among temperament differences in infants and toddlers and later personality structure in children and adults. Contributors to the Halverson et al. (1994) volume showed how aspects of temperament might be linked to BF personality structure. For the personality development researchers, temperament was the focal concept because in developmental psychology, it refers to early-appearing, stable individual differences derived from the constitutional or biologically determined makeup of individuals.

Prior to the Halverson et al. (1994) syntheses, the temperament researchers and the BF researchers were like two separate, isolated tribes. Temperament research was conducted largely as a kind of developmental psychology. Temperament was conceptualized as biobehavioral individual differences in young children, and not as personality structure, *per se* (see *Personality Development and Temperament*). Because young children cannot provide verbal self-reports, temperament researchers were usually forced to rely on objective measures and ratings by expert observers. In contrast, BF research was conducted largely as personality psychology, conceptualized as the structure of individual differences in adults. Because adults can provide verbal self-reports, BF researchers rely much more on self-reports than on objective measures or expert ratings for their data. Some BF researchers seemed mildly interested in the temperamental antecedents of the structure, but others most emphatically were not (see Hofstee 1991, p. 185). Many of the contributors to the Halverson et al. (1994) volume were internationally recognized authorities on temperament, and their suggestions about links between early-appearing, emotional dispositions and later BF personality structure gave credibility to the general line of thought.

One idea reinforced at this conference was that some aspects of temperament provide the developmental substrates for later, major dimensions of personality. One route is to link temperamental characteristics in infants and young children to each of the major BF dimensions. In theory, temperamental differences in activity at infancy could be linked positively to extraversion, and negatively to conscientiousness (e.g., Caspi 1998). Another route is to

suggest ways that temperaments lay a functional foundation for later personality dimensions. For example, infants show temperament-related differences in reactions to frustration. Some infants become so upset when frustrated that they appear to be unable to respond adaptively to the frustrating event, whereas other infants seem to be modulate their emotional reactions and deploy attention in adaptive ways. Ahadi and Rothbart (1994) suggested that a kind of superordinate temperament called 'effortful control' (EC) might regulate reactions to frustration. Furthermore, processes underlying EC might lay a functional foundation for both conscientiousness and agreeableness. This provocative idea suggests that (a) temperament-based processes like EC are related to perceptual processes of attention; (b) seemingly different dimensions of adult personality may be related through their common functional connection to the control of frustration; and (c) two of the major dimensions in the BF may have a common developmental substrate, and may differentiate over time into functionally separate self-regulatory systems. The implications for the socialization of emotions, achievement, and social skills are great (e.g., Kochanska et al. 1998).

In 1998, Kohnstamm et al. delivered some hard evidence in *Parental Descriptions of Child Personality: Developmental Antecedents of the Big Five?* (Kohnstamm et al. 1998). They offered a seven-country study of parents' description of their children's personality. The seven countries involved were Belgium, China, Germany, Greece, The Netherlands, USA, and Poland. The book's title expresses the continuity beginning with early temperament research, on to Digman's pioneering work, through the Halverson et al. (1994) meeting of the minds, to the massive cross-cultural data set presented in this book. Rather than relying on top-down, theorist-imposed descriptors in existing temperament or personality measures, contributors to the Kohnstamm et al. volume asked parents to use their own words to characterize their children. The explicit assumption was that people in all seven countries frequently mention those characteristics that they think are most important or basic. In all seven countries, parents described children between the ages of two and 13 years. The descriptions were collected during face-to-face interviews by teams native to each country's language and culture. Using elaborate coding manuals, descriptions were categorized into units, which were inspired by the BF, with subcategories or facets derived mostly inductively. In addition, responses were coded as positive or negative. For example, in coding the extraversion descriptions, Kohnstamm et al. used 1A (sociability), 1B (dominance, leadership, assertiveness) and 1C (activity, pace, tempo). In this system the positive-negative poles are added so that 1A+ would include 'enthusiastic,' whereas 1A- would include 'tendency to shut self off.' After training, coders' agreement over the main categories was between 80 and 90 percent.

In theory, most cultures have a stake in supporting positive relations among people and controlling conflicts. It is reasonable then to expect persons in all cultures to recognize differences associated with harmonious relations, and to socialize children accordingly (see *Self-regulation in Childhood*). It is telling, perhaps, that the B5 dimensions that focus on interpersonal relations, extraversion and agreeableness, together account for approximately half of all parental descriptions of children. None of the other three dimensions rival either of the two interpersonal dimensions in terms of proportions of descriptions. Extraversion is one of the core concepts of trait psychology, and some form of it is measured in all widely used personality measures. In every country except Greece, extraversion was the single most frequently mentioned category at 28 percent of the descriptors (In Greece, agreeableness and extraversion were equal in importance.)

As a distinct aspect of personality, agreeableness is less well understood than extraversion. In part this may be due to an undifferentiated understanding of social behavior, which includes both social impact (extraversion) and motives to maintain positive relations with others (agreeableness). Eysenck (1994) emphasized the former aspect of social behavior in his trait model. As a consequence, personality researchers interested in the dispositional foundations of social behavior focused on extraversion at the expense of agreeableness. If agreeableness is related to 'love,' and to motives to maintain positive relations, then it is not difficult to see potential developmental continuities, particularly links to interpersonal conflict, and to childhood problems like difficult or disruptive behavior and antisocial tendencies. Data from the Kohnstamm et al. book show that in parental description, agreeableness is a large dimension, accounting for from 17 percent (China) to 26 percent (Greece) of all descriptions.

Because the Kohnstamm et al. work is massive, it is possible to give only a flavor of the findings for some of the other 'lesser' personality dimensions. Of the remaining dimensions, openness seems to be the largest category in free descriptions, but with major cultural differences. Data from the Kohnstamm et al. book show that parents in the USA make use of openness descriptors almost twice as often as do parents in all other countries except Germany. Further data focus on emotional stability (positive pole) and neuroticism (negative pole). Like extraversion, neuroticism is easily linked conceptually to early appearing temperaments (e.g., 'adaptability,' 'quality of mood'). Given the long history of research and theory on the dimension, and the applied importance of this dimension for mental health and child guidance clinics, many professional psychologists will be surprised to learn that across all seven countries, fewer than 10 percent of all parental descriptors fell into this category. One conclusion is that on a worldwide basis

most parents are not concerned about negative emotionality, anxiety, or self-confidence in their children. These qualities may be taken for granted, which is not the same as being unimportant.

Previously, we suggested that all cultures should be concerned with the development of characteristics associated with success in the domain of 'work,' and this concern would be related to evaluations on conscientiousness. Nevertheless, if there is an enigmatic domain of personality development (at least as it appears in parental descriptions of children), it is conscientiousness. As a total category, proportions of parental descriptions were low but highly variable by culture, ranging from 7 percent (USA) to an outlier of 19 percent (China). Chinese parents used almost three times as many descriptors of diligence, and twice as many descriptors of carefulness, as did US parents. Closer inspection shows that Chinese parents' descriptors were drawn from the low end of the dimension (e.g., not diligent), after the child was older than three years of age. What is enigmatic, perhaps, is the low proportion of conscientiousness descriptions relative to the increasing importance of the dimension with age of child in five of the seven countries studied. The outcomes for conscientiousness may reflect cultural differences in the value of education, and in the values of modesty and humility. Chinese parents seem to stress achievement motivation more than parents in the USA, and as a result may have been more critical. Describing one's own children in very positive terms may be especially unseemly for Chinese parents.

4. Future Directions

In the broad-brush overview presented here, it is possible to see the recent advances in theory and research on personality development in childhood. The advances are due in part to (a) efforts to find points of convergence, and to synthesize different perspectives on personality development; (b) the discovery of certain empirical regularities, both within cultures and across them; and (c) more sophisticated research techniques that permit the consideration of several variables at a time. In particular, we see increasing awareness that both culture-based, social learning environments and dispositional characteristics contribute to the personality development process. Important questions and issues remain, however, for the future.

We began this entry with a modern definition of personality as a structural integration of individual differences. Personality development theory and research have only begun to deal with the issue of organizational structures. Theoretical accounts of the development of personality structure are rare, but rarer still as systematic empirical evaluations of such theoretical accounts. So far, the best available research has demonstrated empirical links of variables taken

one at a time, which is no small achievement. This is some distance, however, from the theoretical description of personality as a structured integration. Furthermore, it could be argued that much of the personality development literature has yet to move from the assessment of reliability to the assessment of validity. Still missing from the personality development literature are systematic demonstrations of predictive validity to theoretically important external criteria. Mean differences in frequencies in personality descriptions across cultural groups or sex do not guarantee corresponding differences in overt behavior, or the prediction of overt behavior by personality. For example, Greek parents may use more agreeableness words to describe their children than do US parents, but do agreeableness differences predict choice of conflict strategies in a more differentiated way for Greek children than for US children? In a related vein, linking infant and toddler characteristics to later childhood personality is also promising, but it is not clear how (or even whether) early characteristics are foundational developmentally for later personality structure (but see Kochanska et al. 1998, Rothbart and Bates 1998). That is, early appearing temperaments and dispositions, and life experiences like parental divorce, may precede later personality, and may even predict later personality patterns. Such connections may not be causal or developmental in nature (e.g., Lewis et al. 2000). Is childhood in fact an especially formative period for personality development, or just another way station for the unfolding of a maturation-based personality system? Alternatively, is childhood just another social ecology to which each child must adapt, but with no long-term consequences for personality development? These questions go to the heart of personality development because they address the question of the origins and maintenance of continuity in persons across time. On an optimistic note, these limitations are less obstacles than incentives for future work on personality development in childhood.

See also: Gender Differences in Personality and Social Behavior; Infancy and Childhood: Emotional Development; Personality and Conceptions of the Self; Personality and Social Behavior; Personality Assessment; Personality Development and Temperament; Personality Development in Adulthood; Personality Psychology; Personality Psychology: Methods; Personality Structure; Personality Theories; Self-regulation in Childhood; Social Learning, Cognition, and Personality Development; Temperament and Human Development

Bibliography

- Ahadi S A, Rothbart M K 1994 Temperament, development, and the Big Five. In: Halverson C F Jr, Kohnstamm G A, Martin R P (eds.) *The Developing Structure of Temperament and Personality from Infancy to Adulthood*. Lawrence Erlbaum Associates, Hillsdale, NJ, pp. 189–207
- Block J 1971 *Lives Through Time*. Bancroft, Berkeley, CA
- Caspi A 1998 Personality development across the life course. In: Damon W, Eisenberg N (eds.) *Handbook of Child Psychology: Social, Emotional, and Personality Development*, 5th edn. Wiley, New York, Vol. 3, pp. 311–88
- Digman J M 1989 Five robust trait dimensions: Development, stability, and utility. *Journal of Personality* **57**: 195–214
- Digman J M 1990 Personality structure: Emergence of the five-factor model. *Annual Review of Psychology* **41**: 417–40
- Digman J M 1994 Historical antecedents of the five-factor model. In: Costa P T, Widiger T A et al. (eds.) *Personality Disorders and the Five-factor Model of Personality*. American Psychological Association, Washington, DC, pp. 13–18
- Elder G H Jr 1999 *Children of the Great Depression: Social Change in Life Experience*. Westview Press, Boulder, CO
- Eysenck H J 1994 The Big Five or giant three: Criteria for a paradigm. In: Halverson C F Jr, Kohnstamm G A, Martin R P (eds.) *The Developing Structure of Temperament and Personality from Infancy to Adulthood*. Lawrence Erlbaum Associates, Hillsdale, NJ, pp. 37–51
- Halverson C F Jr, Kohnstamm G A, Martin R P 1994 *The Developing Structure of Temperament and Personality from Infancy to Adulthood*. Lawrence Erlbaum Associates, Hillsdale, NJ
- Hart D, Hofman V, Edelstein W, Keller M 1997 The relation of childhood personality types to adolescent behavior and development: A longitudinal study of Icelandic children. *Developmental Psychology* **33**: 195–205
- Hofstee W K B 1991 The concept of personality and temperament. In: Strelau J, Angleitner A (eds.) *Explorations in Temperament: International Perspectives on Theory and Measurement*. Plenum, London, pp. 177–88
- Kochanska G, Tjebkes T L, Forman D R 1998 Children's emerging regulation of conduct: Restraint, compliance, and internalization from infancy to the second year. *Child Development* **69**: 1378–89
- Kohnstamm G A, Halverson C F Jr, Mervielde I, Havill V L (eds.) 1998 *Parental Descriptions of Child Personality: Developmental Antecedents of the Big Five?* Erlbaum, Mahwah, NJ
- Lewis M, Feiring C, Rosenthal S 2000 Attachment Over Time. *Child Development* **71**: 707–20
- McCrae R R, Costa P T Jr, de Lima M P, Simoes A, Ostendorf F, Angleitner A, Marusic I, Bratko D, Caprara G V, Barbaranelli C, Chae J-H, Piedmont R L 1999 Age differences in personality across the adult life span: Parallels in five cultures. *Developmental Psychology* **35**: 466–77
- Mischel W 1999 Personality coherence and dispositions in a cognitive-affective personality (CAPS) approach. In: Cervone D, Shoda Y (eds.) *The Coherence of Personality: Social-Cognitive Bases of Consistency, Variability, and Organization*. The Guilford Press, New York, pp. 37–60
- Mischel W, Shoda Y, Peake P K 1988 The nature of adolescent competencies predicted by preschool delay of gratification. *Journal of Personality and Social Psychology* **54**: 687–99
- Rothbart M K, Bates J 1998 Temperament. In: Damon W, Eisenberg N (eds.) *Handbook of Child Psychology: Social, Emotional, and Personality Development*, 5th edn. Wiley, New York, Vol. 3, pp. 105–76
- Thomas A, Chess S 1977 *Temperament and Development*. Brunner/Mazel, New York

W. Graziano

Personality Disorders

1. General Considerations

Until the 1980s, both the scientific and the therapeutic aspect of personality disorders constituted a rather underdeveloped field despite their great significance for clinic and practice. Since the introduction of separate Axis II personality registration used in DSM-III (APA 1980) and the subsequent editions, DSM-III-R and DSM-IV (APA 1987, 1994), research and publication worldwide have been considerably stepped up, leading to the inception of professional journals and scientific societies devoted solely to the field of personality disorders. As of the beginning of the twenty-first century there is a veritable deluge of theoretical concepts and therapeutic methods.

More than in any other area of psychiatry, one still sees many inaccurate terms and concepts, and hard data on nosology, etiology, and pathogenesis of the various personality disorders remain rare. To be sure, there is a wealth of isolated pieces of etiological and pathogenetic knowledge on factors influencing the formation and progress of accentuated personality traits, but a comprehensive theoretical or empirical model remains as elusive as ever. We can therefore only address certain points and make some qualified references, as the exceedingly complex nature of the subject still does not allow any more definite statements.

'Personality disorder' is a value-free term which encompasses all those deviations from a normal personality development which require treatment. There are, necessarily, fluid borders to mental health on the one hand and mental disorder on the other.

Concepts with meaning similar or identical to personality disorder are psychopathy, abnormal personality, psychopathic development, psychopathic or dissocial personality, sociopathy, etc. In Sect. 3, the author will give an overview of the historical roots of

these concepts while also addressing recent diagnostic and therapeutic developments.

2. Terms and Definitions

The most comprehensive term in this field must surely be 'personality.' One of the most complex terms in our language, it can be defined in a variety of ways.

Popularly, personality is often used as an ethical-pedagogical value judgment to mean character or temper, as when referring to someone as a strong personality or a person of character. Moral judgments are likewise made, as when we say that all human beings are persons, but we do not mean that everyone is a personality, which is characterized by a conscious and sustained exercise of thought and will.

The current psychological and psychiatric definition of personality is: the sum of all mental and behavioural characteristics which make each of us a unique individual.

On the other hand, a personality disorder occurs if, due to degree and/or a particular combination of psychopathologically relevant features in any of these areas, there is considerable suffering and/or lasting impairment of social adaptiveness (Saß 1987). According to DSM-IV, only when personality traits are inflexible and maladaptive and cause either significant functional impairment or subjective distress do they constitute personality disorders (Table 1).

3. A History of the Concepts of Abnormal Personality

The concept of psychopathy arises from a confluence of views entertained by the French, the German, and the Anglo-American psychiatric traditions. Sociocultural factors caused these conceptions of psychopathy to develop more or less independently well into the twentieth century. The following section deals with all three traditions.

Table 1

General diagnostic criteria for personality disorder

A. A lasting pattern of inner experiences and behavior which deviates markedly from the expectations of the culture the individual lives in. This pattern is manifested in two or more of the following areas:

1. cognition (i.e., ways of perceiving and interpreting oneself, others, and events)
2. emotiveness (i.e., range, intensity, stability, and appropriateness of emotional response)
3. interpersonal functioning
4. impulse control

B. The pattern is inflexible and pervasive across a broad range of personal and social situations.

C. The pattern leads to clinically significant distress or impairment in social, occupational, or other important areas.

D. The pattern is stable and longstanding, and its onset can be traced back to adolescence or early adulthood.

E. The pattern is not better accounted for as a manifestation or consequence of another mental disorder.

F. The pattern is not due to direct physiological effects of a substance (e.g., drug or medication) or a general medical condition (e.g., head trauma).

3.1 French Concepts

Pinel's (1809) concept of *manie sans délire* includes instances of 'deranged personality' and can be looked upon as the beginning of the scientific study of personality disorder as a nosological entity. In the eighteenth century, all mental diseases were regarded as fundamental disturbances of the intellect. Pinel was one of the first to point out that in some disorders, the emotions were primarily involved. Nevertheless, the early nineteenth-century definitions of madness remained centered mostly around the intellect. Indeed, to this day, psychiatric phenomenology neglects the disorders of affect.

Esquirol (1838) developed the concept of monomania, a diagnostic category based on the partial, primary—and independent—involvement of the intellectual, emotional, and/or volitional functions of the mind, including states where a single behavioral disturbance became the sole diagnostic criterion for the condition (e.g., pyromania or kleptomania). Not surprisingly, monomania was one of the sources of the concept of psychopathy. Since its inception, the concept of monomania has been criticized on clinical and medicolegal grounds both in France and abroad.

Morel's (1857) idea of degeneration was a pseudo-biological account strongly tinged with the idea of original sin. He proposed that: (a) degenerative alterations are pathological deviations from the normal; (b) mental illness is mostly hereditary; and (c) degeneration is both quantitative and qualitative, i.e., new disorders may appear. According to Morel's model, all types of mental and neurological disorders can be traced back to one common hereditary origin. His idea of progressive and polymorphic degeneration was accepted as an explanation of mental illness.

3.2 Anglo-American Concepts

Prichard (1835) defined moral insanity as 'madness consisting of a morbid perversion of the natural feelings, affections, inclinations, temper, habits, moral dispositions and natural impulses without any remarkable disorder or defect of the interest or the reasoning faculties, and particularly without any illusion or hallucinations.' During the early nineteenth century, the word 'moral' had many uses, in the psychological sense chiefly to denote the affective and the conative (rather than the purely intellectual) functions.

The British concept of psychopathy was shaped by D. K. Henderson (1939), who saw 'psychopathic states' as a condition of 'constitutional abnormality.' In contrast to other (especially German) psychiatrists, he thought of constitution as deriving from both heredity and environment. There were three psychopathic states, the predominantly aggressive, the inadequate, and the creative. The former two were

characterized by antisocial traits and soon became part of the Anglo-American concept of personality disorder.

Rush (1812) was the first Anglo-American psychiatrist to study individuals whose disturbances were primarily characterized by irresponsibility and aggressiveness; who showed, as he put it, a 'moral alienation of the mind.' He believed that these reprehensible acts were the manifestations of mental disease, that they were unmotivated and driven by 'a kind of involuntary power.'

Partridge (1930) defined the concept of psychopathic personality as a persistent maladjustment which cannot be corrected by ordinary methods of education or by punishment. From Partridge on, the emphasis has been on description, and etiological speculation has taken the back seat. To this day, a view of psychopathy as sociopathy dominates Anglo-American psychiatry. A parallel concept, that of the antisocial personality disorder, appeared in DSM-III (APA 1980) and was kept in DSM-III-R (APA 1987) and DSM-IV (APA 1994).

3.3 German Concepts

In Germany, the concept of psychopathy embraces most forms of abnormal personality. The current German definition can be traced back to Koch, whose monograph *Psychopathische Minderwertigkeiten* [Psychopathic Inferiorities] (1891–93) was as decisive in shaping the concept of abnormal personality in Germany as Pinel's work was in France, Rush's in the USA, and Prichard's in Great Britain. German ideas of psychopathy also influenced French and Anglo-American views, especially after the 1930s, when many German-speaking psychiatrists and psychoanalysts emigrated to these countries.

Kraepelin's concept of psychopathy, influenced by the French theory of degeneration, formed the basis of Kurt Schneider's typology, and via the latter, of today's German doctrine of psychopathy. In successive editions of his textbook, Kraepelin elaborated on his concept of 'psychopathic states' as abnormal personality. He employed the term 'psychopathic personalities' in a predominantly social sense, including also innate delinquents, unstable liars, swindlers, and pseudo-querulous individuals.

During this period, there also appeared various 'systematic' typologies to draw 'psychopathic' categories from prototypic personality theories. Foremost amongst these is Kretschmer's *konstitutionstypologisches Modell* (1919). He suggested a specific correlation between body type and personality. Kretschmer (1921) and Ewald (1924) also introduced the concept of *Reaktionstypologie* to account for specific styles of dealing with experience. Systematic typologies lost influence after the publication of Schneider's monograph in 1923.

Kurt Schneider is known especially for his famous monograph *Die psychopathischen Persönlichkeiten* [The Psychopathic Personalities] (1923). He used a 'typological approach' to personality types and tried to avoid value judgments by not including antisocial forms of behavior. Schneider defines abnormal personalities as statistical deviations from an estimated average norm, although the concept of norm is poorly formulated in his work. In his model, eminently creative and intelligent individuals are also abnormal; hence, not all abnormal personalities could be said to have psychiatric implications. Schneider defined 'psychopathic personalities [as] those abnormal personalities that suffer from their abnormality or whose abnormality causes society to suffer.' It is very important to stress that Schneider did not consider psychopathy a form of mental illness, which by definition must be associated with a brain lesion or a recognized disease process. In this, he opposed Kretschmer and Bleuler, who believed that psychosis and psychopathy were just different degrees on the same spectrum of derangement. Schneider's doctrine influenced all subsequent typologies, and current classification systems include essential parts of his concept of psychopathy. The appendix to DSM-IV even includes the 'depressive type' to encourage further research on this subject.

4. Epidemiology, Course of Illness, and Prognosis

According to German and American studies, 3–10 percent of the general population meet the diagnostic criteria of a personality disorder. Compared to numbers in earlier manuals, these are rather high. However, simply meeting these criteria need not imply that the individuals in question are so dysfunctional and impaired as to require treatment.

Prevalence rates are much higher among unselected psychiatric patients. After the first classification systems appeared, clinical studies found surprisingly high frequencies for personality disorders—50–80 percent—while more recent studies reported prevalence rates of 40–60 percent on the average. Forensic-psychiatric random samples yielded prevalence rates of up to 80 percent. In a large-scale international WHO study (Loranger et al. 1994), 39.5 percent of 716 psychiatric patients examined showed at least one personality disorder according to ICD-10, individual prevalence rates falling between 15.2 percent (anxious personality disorder) and 1.8 percent (schizoid personality disorder).

Clinical experience has shown that increasing age and decreasing vitality tend to attenuate 'sharp' personality traits, especially those which seriously impair social functioning, such as inconstancy, antisocial behavior, and impulsiveness. Other traits can become sharper with advancing age, above all obstinacy and rigidity.

The prognosis depends on the particular type of personality disorder, on any eventual comorbidity, and on the degree of severity. Further prognostic factors are psychostructural maturity as well as the level of psychological and social functioning. Prognostically favorable characteristics are motivation, trust in others, flexibility, and insight into one's own role in difficulties with interpersonal contact. Cases are complicated by concomitant illness, especially addiction and affective disorders. Thus, the mortality rate for patients with personality disorders and substance abuse is three times that of patients with a single personality disorder (Bohus et al. 1999).

Generally speaking, the risk of suicide is three times higher for individuals with a personality disorder than for the general population, with borderline, narcissistic, and antisocial personality disorders showing the highest incidence. These groups also show the highest degree of psychosocial impairment, with deviant actions, decreased capacity for work, and deficient skills at establishing dependable interpersonal relations.

5. Diagnostic Problems

In view of the complexity of the concept of personality, it becomes clear that difficulties exist in distinguishing between a normal range of varying personality traits (as many as 1800 features have been identified which might aid in characterizing possible personality traits) and personality disorders which might be of psychiatric relevance. Towards the end of the twentieth century there has been a trend towards reducing the myriad personality traits to 3, 5, or 7 essential dimensions from which the individual variations can be derived (Table 2).

Interestingly, the factor structure of behavior and its abnormalities seems to be essentially the same in the general population and in clinical groups of behaviourally abnormal patients. The differences lie mainly in the degree and the particular combination of the various dimensions, not in a fundamental difference of personality traits.

Categorical classifications are based on somatic models where the various diseases can be clearly separated from each other; personality disorders, however, show fluid borders, both between the various types and towards normality. Dimensional personality models, developed especially for scientific research, measure degree—or severity—by means of trait factors and thus assign it a position corresponding to its extent along various dimensional axes. It must be kept in mind that there is no inherent qualitative difference in personality traits between the general population and clinical groups of patients with personality disorders, only a difference in degree or severity, or eventually in a particular combination of traits. By means of factor analyses, these personality models

Table 2

Factor models of dimensional personality description

3-factor personality models	5-factor personality models	7-factor personality model
Eysenck (1977) neuroticism extroversion psychoticism	Costa and McCrae (1990) extroversion neuroticism agreeability conscientiousness	Cloninger (1994) harm avoidance novelty seeking reward dependence persistence
Millon (1981) joy/pain self/other-centeredness activity/passivity	openness to new experiences	self-directedness cooperativeness self-transcendence
Widiger et al. (1994) free-floating anxiety/eccentric behavior assertiveness/dominance social integration	von Zerssen (1988) extroversion neuroticism aggressivity conscientiousness openness (devoutness)	
DSM-IV (1994) eccentricity dramatic affect fear		

Table 3

Classification of personality disorders

DSM-IV	ICD-10
paranoid	paranoid
schizoid	schizoid
schizotypal	—
antisocial	dissocial
borderline	impulsive
histrionic	histrionic
narcissistic	—
avoidant	anxious
obsessive-compulsive	obsessive-compulsive
not otherwise specified	not otherwise specified and others

reduce the myriad personality traits to a few essential personality dimensions, independent of culture, from which the individual variants may then be derived. However, given the clinical usefulness of the categorical approach, and hence the widespread hesitancy in abandoning it, recent efforts have tended more towards a synthesis of the categorical and dimensional elements.

The classification systems used today, which increasingly determine personality diagnosis, are the *Diagnostic and Statistical Manual of Mental Disorders, 4th Edition* (DSM-IV), published by the American Psychiatric Association (APA 1994), and the *International Classification of Diseases, 10th Edition* (ICD-10), published by the World Health Organization (WHO 1992). Both are used for making diagnoses, thus forcing clinicians, practitioners, and researchers

alike to opt for one of the two systems. While the two systems resemble each other in many aspects, there are others where they completely disagree. Table 3 gives an overview of the most important discrepancies (see *Differential Diagnosis in Psychiatry*).

6. Therapy of Personality Disorders

Following a thorough anamnesis, there are two main options: psychotherapy and therapy with psychopharmaceuticals. In detail:

6.1 Thorough Initial Examination

The general procedure in treating personality disorders follows the usual rules of psychiatry. It is important that a thorough initial examination be done, with an exact medical and psychiatric anamnesis. Due to the lacking self-dystonia, the patient's self-perception may be lower than in cases showing acute psychopathological symptoms. Hence, anamnesis by significant others can also be helpful in assessing social conflict when dealing with personality disorders.

6.2 Psychotherapy of Personality Disorders

In the psychotherapy of personality disorders, the choice of the therapeutic procedure will depend on the particular form of personality disorder, its severity, as well as any concomitant mental disorders. Currently accepted forms of therapy include: cognitive/

behavioral therapeutic and supportive techniques, depth psychology, disorder-oriented and interdisciplinary therapies with treatments for specific problem areas of a personality, pedagogic or psychogogic therapies, sociotherapy, and dialogue sessions for couples or families. An important element of every therapy is that disorder-related behavior and interaction abnormalities also be dealt with in the patient–therapist relation, so that a big step in the therapy will consist in opening up new self-awareness and possibilities of change to the patient via the therapeutic interaction itself.

Psychotherapy of personality disorders usually requires a long time, since it involves a gradual reshaping of long-lasting characteristics in the areas of experiencing, feeling, and social behavior. Short-term crisis intervention can also be helpful if the problematic personality traits bring situations to a head and cause social conflicts. Building up a regular, trusting relation is of great importance. The frequency usually chosen is one to two hours per week. Aside from discussing current life conflicts, exploring past issues is also important as it can outline the development of the personality traits which are at the root of the trouble.

Group therapy is well suited for many personalities (provided they are not seriously deranged) since observing other patients allows learning by example. The positive feedback or criticism from the other group members can reinforce desirable behavior and put undesirable behavior in perspective (see *Cognitive and Interpersonal Therapy: Psychiatric Aspects*).

6.3 Pharmacotherapy

The second important therapy for personality disorders, one which in recent years has been increasingly developed, consists in the use of psychopharmaceuticals. It rests on the assumption that personality disorders can also have a biological cause which may be either constitutional or due to functional anomalies which were acquired later.

The aim is to reduce the vulnerability to affective and cognitive dysfunctions neurochemically and to modify certain behavioral reaction patterns, the target symptoms or syndromes being:

- (a) the characteristics of the personality disorder (e.g., cognitive deficits, impulsive behavior, strong mood swings);
- (b) complications due to the personality disorder (e.g., suicidal tendencies, aggressivity towards others, deficient social skills); and
- (c) the axis I disorders associated with these (e.g., depressive or anxiety syndromes, compulsion, eating disorders).

The pharmacological treatment of personality disorders is by no means an alternative to psychotherapy. Rather, it is used in support and preparation for psychotherapy, as well as for crisis intervention (especially with suicidal tendencies).

Due to the dangers of side effects and habituation or dependence associated with it, long-term pharmaceutical therapy of personality disorders must always be viewed critically. Likewise, the therapist must be aware of the danger that the patient may start believing in external control or attribute therapeutic success solely to the action of the psychopharmaceuticals.

Moreover, there are special counterindications: alcohol and medicine abuse, social unreliability, bad compliance (incoming to therapy or in taking the medication). Due to the danger of late dyskinesias following long-term use, neuroleptics should be employed judiciously and substances of low or medium potency be considered. With benzodiazepines there is a particularly high risk of addiction, which is why they are generally not indicated for the treatment of personality disorders. Medication can also be tricky with some patients, who may harbor unrealistic expectations or fear to lose control. Due to these difficulties, such individuals may have to be admitted as inpatients before psychopharmaceutical therapy can be initiated (see *Psychotherapy and Pharmacotherapy, Combined*).

7. The Most Important Personality Disorders According to DSM-IV

This section offers classification of personality disorders, followed by an overview of the most important forms of abnormal personalities in DSM-IV. Please consult the manual for a complete list of the diagnostic criteria for the following personality disorders. The differential diagnosis refers solely to the Axis-I disorders (because of limited space).

7.1 Paranoid Personality Disorder

According to DSM-IV, persons with a paranoid personality disorder show a pattern of distrust and suspicion, reading malevolence into the motives of others without sufficient basis for that supposition. They often doubt the loyalty of their friends, partners, or associates and are reluctant to confide in others for fear that the information might be used against them. Persons with a paranoid personality disorder quickly react to perceived insults with excessive force, counterattacks, and/or long-lasting enmity. Mistakes are usually blamed on others. Persons with this disorder usually have recurrent suspicions regarding the fidelity of their spouse or partner. These persons are easily insulted, emotionally rigid, and persevering while at the same time appearing humorless and restrictive in their expression of warm emotions.

7.2 Schizoid Personality Disorder

The chief characteristics of the schizoid personality disorder are an inability to develop close emotional

attachments to others, seclusiveness, and a reduced capacity for emotional experience and expression in interpersonal settings. Schizoid personalities neither desire nor enjoy close relationships, have little, if any, interest in sexual experiences, and do not have friends or confidants. Persons with a schizoid personality disorder are shy, show emotional coldness and detachment, and are seemingly indifferent to praise or criticism. They take pleasure in few activities and distinctly prefer solitary ones. They may function adequately at work, especially if their job requires little social contact.

7.3 Schizotypal Personality Disorder

The schizotypal personality disorder is characterized by strong deficits in establishing interpersonal contact. Persons with schizotypal personality disorder show a pronounced fear of social situations and actively avoid them. They do not have close friends or confidants. Their interpersonal deficits are marked by acute discomfort with close relationships, constricted affect, cognitive and perceptual distortions, and eccentricities of behavior, as in their choice of dress and movement. Persons with this disorder often develop magical thinking and a belief in the occult, and sometimes ideas of reference or paranoid ideation may influence their behavior. Regarding language, there are unclear, strange, or stereotyped expressions and incorrect use of words, though not to the point of associative loosening and incoherence.

Phenomenological, biological, genetic, and outcome data (e.g., the Danish adoption studies) show a relation between schizotypal personality disorder and schizophrenia. Therefore, it is often considered a schizophrenia-spectrum disorder.

7.4 Antisocial Personality Disorder

The main characteristic of the antisocial personality disorder is a permanent and deep-seated tendency to violate and abuse the rights of others occurring from the age 15 years on. These persons show little introspection and self-criticism, lack empathy, show coldness, egotism, an exaggerated sense of entitlement, a paradoxical idea of adaptation, and weak or faulty social norms. Their behavior is marked by impulsiveness, unreliability, weak commitment, and absence of guilt feelings. They are practically beyond therapy and prognosis is generally unfavorable.

7.5 Borderline Personality Disorder

The borderline personality disorder shows a pattern of interpersonal and affective instability which often

leads to sudden, short-lived bursts of extreme mood swings. The affective outbursts usually take place under conditions which are perceived as threatening, such as real or imagined abandonment or rejection.

Especially characteristic for borderline personality disorder is an alternating lifelong pattern of impulsively inflicting harm on oneself or others, including self-inflicted wounds, bulimic binge-and-purge attacks, periods of excessive alcohol consumption, or fights. Repetitive suicide threats and attempts are common.

Unlike antisocial personalities, most patients try to restrain or suppress their impulses, although these attempts often fail. This leads to unpredictable swings between a tense holding back of affective impulses on the one hand and sudden outbursts on the other. Predominant emotions are dysphoria, anxiety, anger, and chronic feelings of emptiness.

Borderline personality disorder patients are furthermore characterized by a highly unstable image and perception of themselves, which can also include aspects of gender identity, deficient orientation and plans for the future, as well as indiscriminate choice of social groups or partners. In their unstable and intense interpersonal relationships they often alternate between extremes of idealization and devaluation (splitting). A last and important area are the transient, stress-related dissociative or (pseudo)psychotic symptoms or paranoid ideas (see *Borderline Personality Disorder*).

7.6 Histrionic Personality Disorder

The main characteristics of this personality disorder are a strong need to be at the center of attention and to gain recognition. Persons with histrionic personality disorder show a pattern of excessive emotionality with self-dramatization, theatricality, and coquetry that may be reflected in their style of speech, which is impressionistic and lacking in detail.

There is a tendency for affective instability and superficiality, displaying rapidly shifting emotions. Most patients are largely suggestible and unable to maintain a steadfast pursuit of goals or value orientation. They are thus inconstant, especially in relationships. Persons with histrionic personality disorder often interact with others in an inappropriate, sexually seductive or provocative manner, using their physical appearance to draw attention to themselves. Relationships are often considered to be more intimate than they actually are.

Also clinically relevant, though uncommon, are the sometimes drastic 'pseudohysterical' cases showing aggravation, conversion, dramaticism, and improper behavior.

Manic states may be accompanied by exaggerated expression of emotion and impressionistic behavior,

but can be distinguished from histrionic personality disorder by the temporary nature of these symptoms and, of course, the presence of other specific/typical symptoms of mania.

7.7 Narcissistic Personality Disorder

According to DSM-IV, individuals with a narcissistic personality disorder have an exaggerated sense of self-importance, are preoccupied with fantasies of grandiosity or superiority, and require excessive admiration. They believe that they are 'special' and can only be understood by other special people.

Patients with narcissistic personality disorder tend to exploit others involuntarily, thinking that their own qualities and abilities entitle them to special treatment. There is a lack of empathy which becomes obvious in the unwillingness to identify with the feelings and needs of others. Further symptoms are a basically fragile self-esteem and feelings of envy and distrust towards others. With a clearly increased self-awareness and egotism, social discomfort and fear of negative opinions predominate. A particular problem is a tendency for depressive crises and resolute suicidality following an imagined insult.

7.8 Avoidant Personality Disorder

The avoidant-insecure personality disorder is characterized by a pervasive pattern of low self-esteem and hypersensitivity to negative evaluation. Despite their strong wish for affection, persons with avoidant-insecure personality disorder avoid social relations, being insecure, shy, tense, and anxious. Their feelings of inferiority and inadequacy in social contact lead to a severe restriction of their social skills and roles, causing them to be reluctant to take personal risks or to engage in any new activities because they may prove embarrassing. Patients with avoidant personality disorder often show restraint with intimate relationships out of a strong fear of shame.

7.9 Dependent Personality Disorder

The dependent personality disorder is characterized by an overpowering feeling of not being able to conduct one's own life. With a self-image of weakness and helplessness, patients will seek support from others in all situations, especially from their partners. Persons with a dependent personality disorder need others to assume responsibility in most major areas of their life. They have difficulties doing things on their own or making everyday decisions without advice from others, and fear that expressing disagreement

might result in a loss of support or approval. This leads to submissive behavior, sometimes to the point of being self-effacing and obsequious. In a relationship, these patients experience a constant fear of loss and abandonment, and urgently seek out somebody for support and care when a close relationship ends. They further show a cognitive distortion known as catastrophizing, which is a fearful and exaggerated estimate of the worst possible consequences of the relation ending.

7.10 Obsessive-compulsive Personality Disorder

The main characteristics of the obsessive-compulsive personality disorder are conscientiousness, perfectionism, conformism, and devotedness to work, which can be overvalued to the point of adversely affecting professional productivity and interpersonal relationships. They dislike teamwork or delegating tasks unless the others conform exactly to their way of thinking or acting. These persons show a severity and rigidity both with themselves and with others which often interferes with social functioning. They are overconscientious, inflexible, and rule-minded about matters of morality or ethics.

Obsessive-compulsive personalities are unable to separate the important from the unimportant (e.g., they cannot bring themselves to throw away old, worthless objects even when these have no sentimental value) and they have an inability to make decisions. Where money is concerned, they are miserly both with themselves and with others.

Of great importance are the interactions between the obsessive-compulsive personality disorder and depression. On the one hand, obsessive-compulsive personality traits can intensify during depression or first become disturbing, as in the form of depressive insecurity or difficulties in making decisions; on the other hand, obsessive-compulsive behavior can lead to difficulties, and hence to reactive depression, where obsessive-compulsive personality traits and depressive symptoms are closely interwoven.

7.11 Personality Disorders not Otherwise Specified

This category applies to personality disorders which are not designated by a DSM-IV diagnosis, but which cause clinically significant distress or impairment. It may also be used when the features of more than one Axis II disorder are present but the full criteria for any one disorder are not met. It also includes the depressive and passive-aggressive (negativistic) personality disorder, which are actually found in the appendix to DSM-IV, being currently under research to determine whether they should be included in DSM-V or not.

8. Concluding Remarks

Concluding, it should be noted that that all typologically arranged classifications of personality disorders are merely conventions. By giving a good description of the reality of life, they may acquire a certain plausibility, self-evidence, and clinical usefulness. A nosological diagnosis and therapy of personality disorders can only be expected for special forms, and even then only as long as it can be shown that they were caused by illnesses.

Most personality disorders, however, lie at the fringe of mental health. Instead of pathological processes, one has to deal with difficult developments in life due to a strained constitution and biography. First and foremost, these people need psychotherapeutic and psychogogic support. Far more than in mental illness, the manifestations in personality disorders are in great part shaped by such phenomena as will, decision-making, self-control, temper, intentionality, bearing, and character. Even more than mental illness, personality disorders require the biological and the psychopathological planes to be complemented by an anthropological dimension.

See also: Borderline Personality Disorder; Obsessive–Compulsive Disorder; Personality and Health; Personality Assessment; Personality Psychology; Personality Theory and Psychopathology

Bibliography

- American Psychiatric Association 1980 *Diagnostic and Statistical Manual of Mental Disorders*, 3rd edn. (DSM-III). American Psychiatric Association, Washington, DC
- American Psychiatric Association 1987 *Diagnostic and Statistical Manual of Mental Disorders*, 3rd edn. revised (DSM-III-R). American Psychiatric Association, Washington, DC
- American Psychiatric Association 1994 *Diagnostic and Statistical Manual of Mental Disorders*, 4th edn. (DSM-IV). American Psychiatric Association, Washington DC
- Beck A T, Freeman A, Pretzer J, David D D, Fleming B, Ottaviani R, Beck J, Simon K M, Padesky C, Meyer J, Trexter L 1993 *Kognitive Therapie bei Persönlichkeitsstörungen*. Psychologie Verlags Union, Weinheim, Germany
- Bohus M, Stieglitz R D, Fiedler P, Berger M 1999 Persönlichkeitsstörungen. In: Berger M (ed.) *Psychiatrie und Psychotherapie*. Urban und Schwarzenberg, Munich, pp. 771–846
- Cloninger C R, Pryzbeck T R, Surakic D M et al. 1994 *The Temperament and Character Inventory (TCI): A Guide to Its Development and Use*. Washington University, St Louis, MO
- Costa P T, McCrae R 1990 Personality disorders and the five-factor model of personality. *Journal of Personality Disorders* 4: 362–71
- Eysenck S B, Eysenck H J 1977 The place of impulsiveness in a dimensional system of personality description. *British Journal of Social and Clinical Psychology* 16(1): 57–68
- Esquirol E 1838 *Die Geisteskrankheiten in Beziehung zu Medizin und Staatsarzneikunde*. Voss, Berlin
- Ewald G 1924 *Temperament und Charakter*. Springer, Berlin

- Frances A J, Hales R E (eds.) 1986 *Review of Psychiatry, Vol 5*. American Psychiatric Press, Washington, DC
- Henderson D 1939 *Psychopathic States*. Norton, New York
- Jang K L, Livesley W J, Vernon P A, Jackson D N 1996 Heritability of personality disorder traits: A twin study. *Acta Psychiatrica Scandinavica* 94: 438–44
- Koch J L A 1891–3 *Die psychopathischen Minderwertigkeiten*. Maier, Ravensburg, Germany
- Kraepelin E 1903–4 *Psychiatrie. Ein Lehrbuch für Studierende und Ärzte*. Barth, Leipzig, Germany
- Kretschmer E 1921 *Körperbau und Charakter*. Springer, Berlin
- Linehan M 1993 *Cognitive-Behavioral Treatment of Borderline Personality Disorder*. Guilford, New York
- Loranger A W, Sartorius N, Andreoli A, Berger P, Buchheim P, Channabasauanna S M, Coid B, Dahl A, Diekstra R F, Ferguson B 1994 The international personality disorders examination. *Archives of General Psychiatry* 51: 215–24
- Millon T 1981 *Disorders of Personality; DSM-III, Axis II*. Wiley, New York
- Morel B A 1857 *Traité des dégénérescences physiques, intellectuelles et morales de l'espèce humaine et des causes qui produisent ces variétés malades*. Baillière, Paris
- Partridge G E 1930 Current conceptions of psychopathic personality. *American Journal of Psychiatry* 10: 53–99
- Pinel P 1809 *Traité médico-philosophique sur l'aliénation mentale*. 2nd edn. Brosson, Paris
- Prichard J C 1835 *A Treatise on Insanity and Other Disorders Affecting the Mind*. Sherwood, Gilbert & Piper, London
- Rush B 1812 *Medical Inquiries and Observations Upon the Diseases of the Mind*. Kimber & Richardson, Philadelphia
- Saß H 1987 *Psychopathie—Soziopathie—Dissozialität. Zur Differenzialtypologie der Persönlichkeitsstörungen*. Springer, Berlin
- Saß H 1988 Persönlichkeit und Persönlichkeitsstörungen. In: Janzarik W (ed.) *Persönlichkeit und Psychose*. Enke, Stuttgart, Germany
- Saß H 2000 Persönlichkeitsstörungen. In: Helmchen H, Henn F, Lauter M, Sartorius N (eds.) *Psychiatrie der Gegenwart. Bd.6*. Springer, Heidelberg, Germany
- Schneider K 1923 *Die psychopathischen Persönlichkeiten*. Deuticke, Vienna, Austria (1. Aufl. 1923, Thieme, Leipzig Germany 1950)
- Spitzer R L, Williams J B 1985 *Structured Clinical Interview for DSM-III, Patient Version*. Biometrics Research Department, New York State Psychiatric Institute, New York
- Widiger T A, Costa P T 1994 Personality and personality disorders. *Journal of Abnormal Psychology* 103: 78–91
- World Health Organization (WHO) 1992 *International Statistical Classification of Diseases and Related Health Problems, 10th revision (ICD-10)*. World Health Organization, Geneva

H. Saß

Personality Psychology

Personality psychology is the scientific study of the whole person. The goal of personality inquiry is to provide a scientifically credible account of human individuality. Such an account considers aspects of human nature (how a person is like all other persons), individual differences between people (how a person is like some other persons), and individual uniqueness

(how a person is like no other person). A full account of human individuality, furthermore, must situate a person's life in the biological, social, cultural, and historical contexts that give that life its meaning. This article will present a brief history of the field of personality psychology, a review of recurrent issues and controversies that have preoccupied personality psychologists, and a consideration of current and projected future research and theorizing in the field.

1. History

The roots of personality psychology can be traced back to the ancient Greeks. A pupil of Aristotle, Theophrastus (c. 300 BC) composed character sketches such as 'the flatterer' and 'the penurious man,' each of which personified (often in a humorous way) a particular personality trait. The Greek physician Galen (AD 130–200) is often credited with proposing the first taxonomy of personality traits, distinguishing among the sanguine, choleric, phlegmatic, and melancholic temperaments. It was not until the late nineteenth and early twentieth centuries, however, that scientists developed reliable methods for assessing individual differences in traits. Building on the pioneering work of Francis Galton and Alfred Binet on mental testing and spurred by the mobilization of large military forces in World War I, psychologists began to invent self-report, multi-item tests to assess individual differences in personality functioning. At about the same time, Sigmund Freud, Carl Jung, Alfred Adler, and other psychoanalytic theorists offered a strong alternative perspective on human individuality, underscoring the unconscious motivational dynamics that structure human lives and make for neurosis and conflict in human affairs.

The psychometricians and the psychoanalysts represented two very different intellectual traditions for personality psychology, both rooted in the nineteenth century. In the spirit of natural science inquiry (*Naturwissenschaften*), the former emphasized the precise quantification of common traits, readily assessed through conscious self-reports and manifest in observable behavior across many different persons. By contrast, Freud and the psychoanalysts offered a more holistic, clinically informed, and intuitive approach, consistent with the *Geisteswissenschaften* (human sciences) movement in the nineteenth century, proposing that the person is a complex and uniquely patterned whole. In the first authoritative textbook for personality psychology, Gordon Allport (1937) proposed that the two intellectual traditions might be reconciled within a single science of persons. Laying out an ambitious agenda for personality inquiry, Henry Murray (1938) published a landmark volume that drew heavily on the psychoanalytic tradition while proposing new methods for measuring personality traits, needs, and other features of human

individuality. The writings of Allport and Murray helped to establish personality psychology as an identifiable discipline in the social and behavioral sciences in the 1930s.

1.1 The Development of Personality Theories

By the mid-1950s, the field of personality psychology was dominated by a number of competing grand theories. Each theory offered a comprehensive portrayal of human individuality, specifying fundamental motivations, a structure for understanding individual differences, principles of personality development, and in some cases a corresponding clinical technique for personality change (e.g., a brand of psychotherapy). Beginning with Hall and Lindzey (1957), personality textbooks typically organized the field in terms of these grand theoretical rubrics, each of which offered its own agenda for personality research.

The largest number of theories came out of the psychoanalytic tradition. Freud's seminal theory, articulated mainly between the years 1895 and 1930, proposed that all human behavior is determined by unconscious forces over which the individual has little control, forces that typically stem from sexual and aggressive drives. Rooted in instinct and early childhood experience, unconscious wishes, urges, and drives seek expression in a social environment that is predicated on the suppression of human urgings. The conflicts between internal urges and external constraints make for anxiety, defense, and a panoply of personal adaptations that define human individuality. Variations on these themes were played out in a number of other psychoanalytic theories, beginning with Jung's and Adler's rival approaches. While Jung proposed that individuality stems from each person's exploration of an evolutionarily rooted collective unconscious, Adler emphasized conscious aspects of human functioning and the prime motives of power and social interest. Around the time of World War II, neo-Freudian theorists like Karen Horney and Eric Fromm sought to link individual development to cultural factors, while Erik Erikson broadened Freud's psychosexual focus to include psychosocial issues that confront the person across the human lifespan. Moving in a different direction, postwar object-relations theorists such as W.R.D. Fairbairn and psychoanalytic self-theorists such as Heinz Kohut redoubled efforts to explore the most primitive intrapsychic factors laid down in the first two years of life and played out most importantly in the mother–infant bond.

Outside the psychoanalytic tradition, grand personality theorists came in behaviorist, trait, and humanistic/phenomenological varieties. Seeking to operationalize Freudian concepts in terms amenable to midcentury behaviorism, Neal Miller and James Dollard developed an influential social learning theory

of personality, which was followed by more cognitively based theories offered by Julian Rotter and Albert Bandura. Employing the method of factor analysis of personality trait scales, Raymond B. Cattell and Hans Eysenck conceived of personality in terms of a finite set of objectively defined traits. Eysenck's three-trait scheme—highlighting introversion–extraversion, neuroticism, and psychoticism—proved especially influential. Finally, a wide range of grand theories were built around a humanistic vision of the person as a more-or-less rational being, endowed with rich conscious experience and motivated to actualize his or her own potential. Included here were Allport's own theory, the humanistic approaches offered by Carl Rogers and Abraham Maslow, George Kelly's personal construct theory, and existential personality theories offered by Rollo May, Victor Frankl, and others.

1.2 The Elaboration of Constructs

Postwar personality psychologists borrowed liberally from the grand theories to identify individual-difference variables—or 'constructs'—that could be reliably measured through self-report scales, ratings, behavioral observations, and other techniques, and then used in hypothesis-testing research. Popular constructs of the 1950s and 1960s included anxiety, achievement motivation, authoritarianism, extraversion, ego resiliency, field independence, and locus of control. Investigators designed programs of study to examine the antecedents, consequences, and correlates of these individual differences in personality. Around each construct, then, developed a nomological network of findings—a body of research that elaborated the construct and defined its role and significance in personality functioning. With increasing specialization and the emphasis on discrete constructs, the field of personality psychology began to turn away from the grand theories to focus on problems in the measurement of individual differences. Omnibus personality inventories, such as the Minnesota Multiphasic Personality Inventory (MMPI) and the California Psychological Inventory (CPI), measured many different variables at once. But many postwar researchers preferred to focus on one construct at a time, refining measures and articulating in great detail that construct's nomological net.

The field of personality psychology experienced a crisis in confidence in the late 1960s and 1970s. Some critics lamented the decay of interest in integrative theory and the obsessive focus on constructs over persons. Others, most notably Walter Mischel (1968), argued that personality constructs themselves were not good predictors of behavior. In Mischel's view, social situations account for vastly more variance in human behavior than do individual differences in personality traits. What was often called the 'trait versus situation debate' preoccupied the field in the

1970s and early 1980s. As the debate died down, proponents of both sides seemed to agree that personality theories and research need to explicitly embody interactionism—the idea that behavior is always a function of the interaction of the person (e.g., traits) and the environment (e.g., situations). In the 1980s and 1990s, furthermore, strong evidence was adduced for the efficacy of personality constructs. First, studies increasingly showed that traits do predict general trends in behavior over time and across different situations. Second, longitudinal studies showed that individual differences in traits are often remarkably stable over long periods of time. Third, evidence from behavior genetics, especially studies with twins, showed that most personality traits exhibit at least moderate heritability. And fourth, some consensus in the field began to emerge concerning the number and kind of basic personality traits. Based on extensive factor-analytic studies, the popular Big Five conception proposes that traits cluster into five different factors, which may be labeled extraversion, neuroticism, conscientiousness, agreeableness, and openness to experience.

2. Perennial Issues and Controversies

In his first textbook, Allport (1937) foresaw a number of issues that were destined to stimulate recurrent debate in the field of personality psychology. The one that most preoccupied Allport himself was the tension between nomothetic and idiographic approaches to personality inquiry. While nomothetic approaches seek to establish general laws of behavior that apply across persons, idiographic approaches, as embodied in the case study, focus on the unique or characteristic patterning of an individual person. The vast majority of published research in personality psychology is nomothetic, typically involving the testing of hypotheses about personality constructs and processes. But if the field itself is supposed to be concerned with human individuality, Allport argued, then some form of idiographic inquiry must be included. Skeptics have countered that the results of single-case studies cannot be generalized, and thus have little scientific value. But proponents of idiographic approaches maintain that case studies are often excellent arenas for hypothesis discovery, for applying general theories, and for illuminating complex personality organization. Along with Allport and Murray, Robert White (1952) championed the intensive study of individual lives. Recent years have witnessed a resurgence of interest in idiographic approaches and considerable optimism about integrating them with more conventional nomothetic methods.

A forerunner to the 'trait versus situation debate' of the 1970s was Allport's identification of the problem of generality versus specificity in behavior. To what extent is a person's behavior generally consistent across situations, as opposed to being specific to the

vagaries of particular situations themselves? Mischel (1968) argued that Allport and most other personality psychologists overplayed the generality idea, expecting their constructs to predict general trends in behavior across many different situations. In Mischel's (1968) view, the empirical data were much more supportive of a specificity position. Although trait constructs have regained their currency in recent years, many personality psychologists have retained a healthy skepticism about cross-situational generality, and some have proposed that some personality constructs themselves need to be defined in contingent, situational terms.

A third issue concerns measurement. The most popular personality measures have always been self-report questionnaires. But many critics have argued that such measures are unable to assess especially subtle, implicit, or unconscious aspects of human individuality. As an alternative, some have championed projective techniques, wherein the person responds freely to ambiguous cues (e.g., inkblots, story scenes). For example, David McClelland (1961) built a highly successful research program around the assessment of achievement motivation in imaginative stories told to picture cues (the Thematic Apperception Test, or TAT). Others, most notably Jack Block (1971), refined Q-sort rating procedures that bypassed self-report for the evaluations of expert judges. While a plethora of measurement techniques may be seen in the field today, the self-report questionnaire, nonetheless, remains the coin of the realm.

A fourth controversy is the often-observed disconnect between grand personality theories and construct-based personality research. While some argue that a good deal of personality research has been directly or indirectly inspired by the grand theories, others contend that the grand theories should be dismissed as historical artifacts. The controversy is especially acute with respect to psychoanalytic theories. Many Freudian ideas, for example, have proven resistant to empirical scrutiny (e.g., the Oedipus complex) or have been jettisoned as outdated or just plain wrong (e.g., the death instinct). By contrast, some ideas that have traditionally been associated with psychoanalytic approaches have become incorporated into mainstream psychological research. Of most importance in this regard is the now generally accepted notion that a good deal of human information processing occurs in an automatic, implicit, and nonconscious manner. Thus, while psychoanalytic theories have exerted a strong impact on Western thinking more generally, their current and future status in personality psychology appears ambiguous at best.

3. Current and Future Trends

The wide range of research and theorizing that currently characterizes the field of personality psychology may be organized into three different levels or

arenas of human individuality: dispositional traits, characteristic adaptations, and integrative stories. The first domain considers individuality from the standpoint of general, cross-situational tendencies; the second examines more situationally and contextually grounded aspects of persons; and the third encompasses efforts to understand individuality in the more idiographic sense of how the person understands his or her life in toto.

3.1 Dispositional Traits

Dispositional traits are those relatively general and nonconditional aspects of human individuality that go by such names as 'extraversion,' 'friendliness,' 'depressiveness,' and 'conscientiousness.' Each trait is viewed as a bipolar, linear dimension upon which individuals can be said to differ. For example, research has shown that people scoring at the high end of the extraversion trait are viewed by others as especially sociable and fun-loving, typically attend more social events, are more sexually active, and report higher levels of positive affect in everyday life, compared to people scoring at the low (introverted) end of the trait (e.g., Eysenck 1967). Evidence for the predictive efficacy of trait scores is growing, with the strongest data coming for aspects of extraversion (sometimes reconceptualized as positive affectivity) and neuroticism (sometimes reconceptualized as negative affectivity).

Psychophysiological experiments have begun to document possible linkages between neural systems and individual differences in traits. For example, Jeffrey Gray and others have suggested that individual differences in extraversion/positive-affectivity may be linked to a behavioral activation system in the brain, responsible for motivating behavior aimed at achieving goals and obtaining positive emotional rewards. These kinds of advances, along with growing evidence for longitudinal consistency and heritability of broad traits, help to underscore the field's conviction that dispositional traits are fundamental to a scientific account of human individuality (see McCrae and Costa 1990, Wiggins 1996).

Increasingly, research on personality traits has become subsumed within the Big Five framework. While most of this research focuses on individual differences among well-functioning adults, recent efforts have also been made to reconfigure psychopathology in terms of (extreme) scores on personality traits. Individual differences in infant and childhood temperament, assumed to be strongly driven by genetic differences, have also been conceptually linked to the five-factor taxonomy. And some theorists have speculated about how individual differences in Big Five traits reflect varying strategies for solving fundamental social tasks that humans have perennially faced over the course of evolution. The integration of trait psychology with evolutionary theory would appear to be a promising direction for the future.

3.2 Characteristic Adaptations

Despite substantial progress in trait research in recent years, there would appear to be considerably more to personality than traits. While traits are useful for predicting and understanding consistent, cross-situational differences in people, they are less effective in the explanation of more particular, contextually anchored behavior. Accordingly, personality psychologists have proposed a host of motivational, developmental, and strategic constructs that are contextualized in time, place, or social role. These include particular motives and goals (e.g., the power motive), life tasks (e.g., the intimacy life task), relational styles (e.g., secure attachment style), defense mechanisms (e.g., projection), coping strategies (e.g., emotion-focused coping), values (e.g., dogmatism), developmental issues (e.g., generativity), self-schemas (e.g., possible selves, self-with-other representations), and the like, all of which may be grouped under the generic rubric of characteristic adaptations. If traits sketch an outline of human individuality, characteristic adaptations begin to fill in some of the details. More contingent and delimited than dispositional traits, characteristic adaptations speak to what a person wants or fears, often during particular periods in life or within particular domains of action, and what life methods the person may use to get what is wanted or to avoid getting what is not wanted at a particular time, in particular places, and/or with respect to particular roles.

While trait approaches have their historical roots in the grand factor theories offered by Eysenck and Cattell, studies of characteristic adaptations connect more readily to theories emphasizing human cognition, motivation, and development. In that they address especially concrete issues that can be observed in everyday adaptation to stress and challenge, moreover, characteristic adaptations speak a language that is compatible with the concerns of many clinicians. They also appear to shed light on proximal personality processes by specifying what cognitions or emotional states may give rise to particular behaviors. Recent years have witnessed considerable progress in the study of a wide range of characteristic adaptations (Pervin and John 1999), but the field has yet to offer a comprehensive framework, akin to the Big Five, for organizing these far-flung constructs and research programs. In a general sense, though, most adaptations link directly or indirectly to the two great domains of social life—love and work—or what David Bakan termed ‘communion’ and ‘agency.’

3.3 Integrative Stories

Since the time of Allport, some personality psychologists have maintained that while conventional personality constructs such as traits and characteristic

adaptations are useful for predicting behavior and accounting for individual differences, they are unable to account for the integrative nature of personality functioning and the sense of a person’s wholeness. Broader concepts such as self, ego, and character structure have often been invoked to refer to this integrative, holistic quality. For example, Jane Loevinger (1976) proposed that the ego is the master synthesizer of subjective experience, and she developed a rigorous program of research to operationalize stages of ego development through the coding of sentence-completion tests. More recently, a number of theorists and researchers have proposed that people synthesize their own experience into personal stories and that such internalized and evolving narratives of the self provide human lives with unity, purpose, and meaning. Silvan Tomkins, Dan McAdams (2001), and Hubert Hermans have all developed narrative theories of personality that attempt to account for human individuality in terms of the different kinds of life stories that people construct. This movement recalls Allport’s emphasis on the idiographic, and the study of lives tradition championed by White and others. But it adds a more nuanced understanding of the psycholiterary aspects of social life, especially under conditions of cultural modernity and postmodernity, and it proposes agendas for nomothetic research whereby individual differences in life narrative form and content can be operationalized and subjected to hypothesis-testing research.

In sum, current research and theorizing in personality psychology suggest that personality itself may be viewed as a patterning of dispositional traits, characteristic adaptations, and integrative life stories, evolving over time and complexly situated in a cultural and historical milieu.

See also: Gender Differences in Personality and Social Behavior; Genetic Studies of Personality; Interactionism and Personality; Personality and Adaptive Behaviors; Personality and Conceptions of the Self; Personality and Crime; Personality and Health; Personality and Marriage; Personality and Risk Taking; Personality and Social Behavior; Personality Development and Temperament; Personality Development in Adulthood; Personality Development in Childhood; Personality Disorders; Personality Psychology; Personality Psychology: Methods; Personality Structure; Personality Theories

Bibliography

- Allport G W 1937 *Personality: A Psychological Interpretation*. H. Holt, Rinehart & Winston, New York
- Block J 1971 *Lives Through Time*. Bancroft Books, Berkeley, CA
- Eysenck H J 1967 *The Biological Basis of Personality*. Thomas, Springfield, IL

- Freud S [1916] 1961 Introductory lectures on psychoanalysis. In: J. Strachey (ed.), *The Standard Edition of the Complete Psychological Works of Sigmund Freud*. Hogarth, London, Vols. 15–16
- Hall C S, Lindzey G 1957 *Theories of Personality*. Wiley, New York
- Hogan R, Johnson J, Briggs S 1997 *Handbook of Personality Psychology*. Academic Press, San Diego, CA
- Loevinger J 1976 *Ego Development: Conceptions and Theories*. Jossey-Bass, San Francisco
- McAdams D P 2001 *The Person: An Integrated Introduction to Personality Psychology*, 3rd edn. Harcourt College Publishers, Fort Worth, TX
- McClelland D C 1961 *The Achieving Society*. D. Van Nostrand, Princeton, NJ
- McCrae R R, Costa Jr. P T 1990 *Personality and Adulthood*. Guilford Press, New York
- Mischel W 1968 *Personality and Assessment*. Wiley, New York
- Murray H A 1938 *Explorations in Personality: A Clinical and Experimental Study of Fifty Men of College Age*. Oxford University Press, New York
- Pervin L A, John O P 1999 (eds.) *Handbook of Personality Theory and Research*, 2nd edn. Guilford Press, New York
- White R W 1952 *Lives in Progress: A Study of the Natural Growth of Personality*. Dryden Press, New York
- Wiggins J S 1996 (ed.) *The Five-factor Model in Personality: Theoretical Perspectives*. Guilford, New York

D. P. McAdams

Personality Psychology: Methods

All empirical research methods in psychology are concerned with the measurement of variation and covariation. Three methods for studying three types of variation and covariation can be identified. Experimental methods discern how behavior and experience vary across different environments. Developmental methods describe how behavior and experience vary over time. Finally, differential methods measure relatively enduring differences in characteristics of persons and the covariation among these differences. Although personality psychology occasionally employs experimental and developmental methods, its primary use of differential methods defines its distinctive position in psychology (Cronbach 1957). The specific methods used by personality psychologists can be further defined by (a) the type of individual differences under study, (b) the source of information about these differences, and (c) the purpose of assessing individuals' distinguishing attributes.

1. Types of Differences Measured

Gordon Allport's (1937) definitive etymological analysis of the term 'personality' finds that *persona*, the Latin root of 'personality,' originally referred to a mask worn by actors, and thus connoted a false or superficial appearance. Before long, however, *persona*

also came to designate more substantial attributes that qualify a person to perform particular roles. Today, 'external' and 'internal' aspects of personality continue to be distinguished. Ordinary-language descriptions of personality often refer to external social impressions. Trait terms representing these impressions tend to be broad and evaluative (e.g., nice, charming, nasty, irritating), and sometimes metaphorical (cold, prickly, warm, slick). Social impressions are explained by observable behavioral traits. For example, to impress others as 'warm,' a person must consistently smile, make eye-contact, talk tenderly, express affection, and show kindness and sympathy. External (social and behavioral) aspects of personality are typically measured by observer judgment (see Sect. 2.2).

Behavioral consistencies, in turn, are explained by 'inner' traits that guide and motivate behavior. Inner traits that refer to persistent thoughts, expectations, attitudes, thinking styles, and mental abilities are called 'cognitive traits.' Inner traits involving recurring feelings and emotional dispositions are called 'emotional' or 'motivational traits.' Personality psychologists often rely on the self-report method (see Sect. 2.3) to assess inner traits. Cognitive styles and competencies (e.g., creativity), are typically assessed by performance tasks (see Sect. 2.4).

2. Types of Individual Differences Data

Differential methods can be organized according to the types of the data one can use to study individual differences. The acronym LOST conveniently summarizes the four major types of individual differences data: life events, observer judgments, self-reports, and tests, or L-data, O-data, S-data, and T-data (see Block 1977).

2.1 L-data: Life Events

'Life events' refer to relatively objective facts about people that are often a matter of public record. Some examples are birth date, number of years of education and degrees obtained, marital status and number of children, occupational record, church attendance, hospitalizations, membership in societies, criminal convictions, leadership positions, and property ownership. Although life event information is often actually obtained by self-report methods (Sect. 2.3) its factuality can be independently verified. Research indicates a very high correlation between self-reported and independently verified life events, even under highly evaluative conditions such as applying for employment.

In both applied (see Sects. 3.1 and 3.2) and basic (see Sect. 3.3) personality research, life events are sometimes used as criteria to be predicted by other types of personality data and sometimes as predictors themselves. When gathered by self-report, L-data is interpreted according to self-report methods (Sect. 2.3).

2.2 O-data: Observer Judgments

Observer judgments are a fundamental source of information about external traits for two reasons (Block 1978, Hofstee 1994). First, an observer's judgment of external traits is direct and noninferential. Second, idiosyncratic biases and errors of judgment tend to cancel out when judgments are averaged across multiple observers. This means that even inner traits may be more validly assessed by observer judgments than self-reports, despite an individual's privileged access to his or her inner thoughts and feelings (Hofstee 1994).

2.2.1 Retrospective, integrative judgments versus one-time, direct observation. Observer judgments are often made by acquaintances who use their past knowledge to summarize their perceptions of someone's personality. Concern about potential distortions and memory limitations involved in making retrospective judgments lead some psychologists to favor direct behavioral observation. Despite potential distortions, retrospective judgments are more likely than single encounters to produce the representative sampling necessary for accurately assessing relatively enduring characteristics of individuals (Kenrick and Funder 1988).

2.2.2 Normative and ipsative frames of reference for observer judgments. For observer judgments to indicate what is distinctive about someone's personality, the judgments must be compared to some reference point. Take, for example, a study in which observers record from behind one-way glass instances of different types of behavior in a nursery school classroom over a period of one year. At the end of the year, a particular behavior count by itself—say 42 acts of aggression—is insufficient for describing a child as relatively aggressive or nonaggressive. A normative frame of reference would compare the number of aggressive acts for that child to the average number of aggressive acts across all children. An ipsative frame of reference compares the number of aggressive acts for the child to all other types of acts recorded for that child.

In retrospective observer judgments, judges are often instructed to compare a person to people in general with a rating scale. The middle point (e.g., '3' on a 1–5 rating scale) represents a theoretical normative reference point (i.e., people in general), while greater numbers represent higher levels of the trait and lesser numbers lower levels. Unipolar rating scales are anchored by a single trait word at the high end, while bipolar scales are also anchored with the psychological opposite at the low end (e.g., 'thoughtful' vs. 'inconsiderate'). Defining the anchors with concrete, descriptive phrases instead of abstract adjectives improves measurement validity. Often scores on several

related rating scales will be averaged to produce an overall score on a trait. For example, broad extraversion versus introversion might be assessed by the average of ratings on more specific scales such as talkative vs. silent, outgoing vs. reserved, and so forth.

When interpreting rating scores, psychologists need to consider whether to take the numerical score at face value. For example, a rating of 4 on a 1–5 scale of thoughtfulness is above the theoretical norm of 3, but if the computed average rating for a large group of people is 4.5, a rating of 4 could be interpreted as a relatively low level of thoughtfulness. This problem is further complicated when some judges restrict their ratings to one portion of the scale. A 5 might actually indicate a much higher value from a judge who assigns mostly 3s than from a judge who assigns mostly 4s and 5s.

Psychologists unwilling to accept scores at face value will recalibrate all of a judge's ratings with respect to the mean of all ratings made by that judge. This process is called 'ipsatizing' scores. Some rating procedures expressly call for ipsative assessment in the act of judgment itself. For example, judges using the California Q-set (Block 1978) are required to sort 100 personality descriptions into nine categories following a specified distribution (five descriptions in each of the extremely characteristic and uncharacteristic categories, eight descriptions in each of the quite characteristic and uncharacteristic categories, and so forth, to 18 characteristics in the relative neutral or unimportant category).

Formats other than rating scales for observer judgment of personality include questionnaires and adjective checklists. Items in some personality questionnaires are phrased in ways that allow either self-report (Sect. 2.3) or observer judgment. Adjective check lists contain a set of adjectives that judges check if they believe them to apply to the person being judged. Scores on adjective check lists are computed by counting the number of judges who check a particular adjective and/or by summing the checks for a group of adjectives considered to measure the same trait.

Factors that potentially limit the validity of observer judgments include misleading communications, stereotypes, and gossip about those being judged; unfairly positive or negative attitudes from judges who like or dislike the targets; and insufficient knowledge due to judges knowing the target only in certain roles and settings. These limitations and methods for overcoming them are discussed by Block (1978), Hofstee (1994) and Kenrick and Funder (1988).

2.3 S-data: Self-Reports

The two basic types of self-report instruments are projective tests and objective questionnaires. In projective testing, respondents are asked to finish incomplete phrases or construct stories about intentionally ambi-

guous images. Following guidelines developed by the test author and community of test users, psychologists score the respondent's protocol for psychological themes. Proponents of projective tests claim that these instruments are able to tap deep, unconscious needs and motives; critics insist that scoring projective protocols is too subjective and unreliable. Research indicates that carefully designed projective tests can be as reliable and valid as objective measures.

Personality questionnaires (Angleitner and Wiggins 1986) rarely consist of questions anymore. Instead, questionnaire items are statements about one's self and other people. Respondents express how much they agree with each statement or the degree to which they think the statement applies to them. The most comprehensive personality questionnaires contain several hundred items. Any subset of items within a personality questionnaire that is scored for a particular trait is called a 'personality scale.' The major personality questionnaires contain as many as several dozen different scales. Items are collected into scales and responses to items are scored according to one of four strategies outlined below: empirical, rational-intuitive, theoretical, or factor-analytic.

2.3.1 Empirical scales. Paul Meehl (1945) argued that psychologists would be naively optimistic to take questionnaire item responses at face value or to attempt to judge their hidden psychological meanings expertly. More prudent, he suggested, would be to treat each item response as a bit of behavior whose meaning must be determined by its empirical correlates. Empirical scales are constructed by locating all items on a questionnaire that tends to be answered differently by two groups known by other methods to differ on a particular personality trait. A person's score on an empirical scale is defined by the number of responses that match the responses given by one of the groups used in original scale construction. That is, if a person answers many items the same way as a group of people known to be aggressive, that person is considered likely to be aggressive also.

2.3.2 Rational-intuitive scales. The rational-intuitive approach to personality scales suggests personality traits can be assessed straightforwardly by items whose content, according to common sense, seems relevant to the trait (Wolfe 1993). Thus, a person who endorses items such as 'I am an aggressive person' and disagrees with items such as 'I never get in fights' would receive points on a rational-intuitive personality scale of aggressiveness. The obviousness of rational-intuitive scales perennially raises concerns about self-enhancement (exaggerating socially desirable traits and denying undesirable traits), but research indicates that respondents self-enhance on personality questionnaires no more than they do in everyday life. Furthermore, research indicates that

rational-intuitive scales validly predict relevant criteria as well as scales constructed by any other method.

2.3.3 Theoretical scales. Like rational-intuitive scales, theoretical scales are comprised of items whose content is judged to be relevant to the personality characteristic being assessed. The difference is that the relevance is not apparent to common sense and can only be seen by professionals versed in a particular theory. For example, the theoretical items 'I am fascinated by fire,' 'I would do anything on a dare,' 'My father and I always fought,' 'I own a gun,' and 'Women find me charming' seem unrelated to common sense, but, to a Freudian, these are items likely to be endorsed by a man with a phallic personality character resulting from an unresolved Oedipal conflict.

2.3.4 Factor-analytic scales. Factor-analysis is a statistical method for identifying clusters of items that tend to be answered the same way. This method, like the empirical method, begins with a large set of items that are administered to a group of respondents. If respondents who agree with item 'A' also tend to agree with items 'B,' 'C,' 'D,' and so forth, these items are deemed to measure the same psychological trait (Briggs and Cheek 1986). The nature of the trait is normally determined by rational-intuitive inspection of the content of the items. Factor analysis can be applied to scales as well as items, and factor analytic research has repeatedly indicated that much of the content of personality falls into five broad factor domains: extraversion, agreeableness, conscientiousness, emotional stability, and a fifth factor variously called intellect, imagination, or openness to experience. Many psychologists regard the 'Big Five' or 'Five-factor model' (Wiggins 1996) as a major integrating focus for future research.

2.4 T-data: Laboratory Tests

Assessing personality by testing respondents in laboratories or other controlled conditions is motivated by two interests. The first is establishing objective, replicable procedures that cannot be affected by biases and errors potentially found in observer judgments or self-reports. In particular, the measurement of involuntary reactions such as changes in electrodermal conductivity or heart rate to assess anxiety, or pupil dilation to assess interest, is seen as a way of circumventing dissembling that can occur with self-reports. Unfortunately, as Block (1977, 1978) points out, T-data are always indirect measures of personality, and laboratory tests that might seem to be reasonable measures of personality have a record of failing unpredictably.

The second motivation for using laboratory tests is the particular suitability of such procedures for measuring certain traits. Laboratory tests are particularly apt for assessing cognitive performance variables. For example, a personality theory built around the mental accessibility of different concepts would naturally be tested by measuring reaction time to words representing different concepts. Cognitive styles are almost invariably measured by performance tasks.

Personality theories that attempt to explain mental or behavioral differences in terms of underlying biological differences also require laboratory facilities to assess those differences. In addition to traditional psychophysiological recordings, new laboratory tests for measuring the physical basis of personality include biochemical assaying, positron emission tomography and functional magnetic resonance imaging. Details of these methods can be found in sources such as Davidson (1999), and Pickering and Gray (1999).

3. Three Purposes of Measurement

The three major purposes for measuring individual differences are: (a) making decisions about people, (b) helping people make decisions, and (c) conducting basic research. All three cases involve prediction (for further information see Wiggins 1973). In decision-making, personality assessments are used to predict how an individual person will think, feel, behave, or be perceived by others if various courses of action are pursued. In research, predictions are made about how different forms of individual differences are related to one another.

3.1 Making Decisions About People

In clinical and counseling psychology, personality scores are used to predict what course of therapy will best help persons with psychological problems. In personnel psychology, personality scores are used to predict which individuals will perform best if hired and placed in particular jobs. Some applied psychologists endorse what they call an 'assessment' approach to these decisions, in which the decision maker intuitively weighs all the information gathered about the person. In contrast, a 'statistical' approach inserts personality scores into a mathematical equation developed from past empirical research. Studies indicate that statistical predictions are almost invariably more accurate than assessment predictions.

3.2 Helping People Make Decisions

Methods for helping people make decisions with personality measures differ from making decisions

about people only in terms of who is making the decisions. Individuals seeking greater satisfaction in personal relationships, for example, may complete a personality questionnaire to increase self-insight, much as they might read a self-help psychology book. Likewise, individuals uncertain about career choice can complete questionnaires that predict which careers would be most satisfying. Traditionally, psychologists interpret and discuss personality scores with clients, but some self-help personality measures are completely self-administered.

3.3 Basic Research: Uncovering Covariation Among Variables

The usefulness of all applied personality methods depends upon the ability of researchers to construct valid personality measures and to ascertain reliable covariation among these measures. This process, called 'construct validation,' is identical with any other type of scientific hypothesis testing (Hogan and Nicholson 1988). In the typical case, construct validation takes the following form. A researcher's theory predicts that individual differences in a particular personality trait (say, conscientiousness) will covary with differences in some L-data (say, job performance). A method is devised to measure the personality trait with O-, S-, or T-data and the life event with L-data. Successful prediction supports both the validity of the measures and the theory that led to the prediction. Predictive failure means either the hypothesis was incorrect, a procedural error occurred (inappropriate research sampling, administration, or scoring), or one or both of the measures lack validity. Progress in personality research occurs when many successful predictions leads to the acceptance of a measure as 'well-validated.' Careful research with well-validated measures always advances knowledge because even predictive failures indicate the need to revise hypotheses.

Because valid measurement is crucial to the entire personality enterprise, a significant amount of personality research is directed at improving measurement methods. Some of this research aims to clarify the dynamics of the measurement process, that is, the psychological processes that occur during observer judgments and self-reports. A second line of research employs computers to administer, score, and interpret personality tests. When computer programs for analyzing personality data are combined with artificial intelligence, 'observers' with artificially constructed personalities will some day make 'observer judgments' of personality worldwide over the Internet.

See also: Personality Assessment; Personality Psychology; Personality Structure; Personality Theories; Projective Methods in Psychology

Bibliography

- Allport G W 1937 *Personality: A Psychological Interpretation*. H Holt, New York
- Angleitner A, Wiggins J S 1986 *Personality Assessment via Questionnaires*. Springer-Verlag, Berlin
- Block J 1977 Advancing the psychology of personality: Paradigmatic shift or improving the quality of research? In: Magnusson D, Endler N S (eds.) *Personality at the Crossroads: Current Issues in Interactional Psychology*. Lawrence Erlbaum Associates, Hillsdale, NJ, pp. 37–64
- Block J 1978 *The Q-Sort Method in Personality Assessment and Psychiatric Research*. Consulting Psychologists Press, Palo Alto, CA
- Briggs S R, Cheek J M 1986 The role of factor analysis in the development and evaluation of personality scales. *Journal of Personality* **54**: 106–48
- Cronbach L J 1957 The two disciplines of scientific psychology. *American Psychologist* **12**: 671–84
- Davidson R J 1999 Biological bases of personality. In: Derlega V J, Winstead B A, Jones W H (eds.) *Personality: Contemporary Theory and Research*, 2nd edn. Nelson Hall, Chicago, pp. 101–25
- Hofstee W K B 1994 Who should own the definition of personality? *European Journal of Personality* **8**: 149–62
- Hogan R, Nicholson R A 1988 The meaning of personality test scores. *American Psychologist* **43**: 621–6
- Kenrick D T, Funder D C 1988 Profiting from controversy: Lessons from the person-situation debate. *American Psychologist* **43**: 23–34
- Meehl P E 1945 The ‘dynamics’ of structured personality tests. *Journal of Clinical Psychology* **1**: 296–303
- Pickering A D, Gray A D 1999 The neuroscience of personality. In: Pervin L A, John O P (eds.) *Handbook of Personality: Theory and Research*, 2nd edn. Guilford Press, New York, pp. 277–99
- Wiggins J S 1973 *Personality and Prediction: Principles of Personality Assessment*. Addison-Wesley Pub Co, Reading, MA
- Wiggins J S (ed.) 1996 *The Five-Factor Model of Personality*. Guilford Press, New York
- Wolfe R N 1993 A commonsense approach to personality measurement. In: Craik K H, Hogan R, Wolfe R N (eds.) *Fifty Years of Personality Psychology*. Plenum Press, New York, pp. 269–90

J. A. Johnson

Personality Structure

Personality structure is the branch of personality psychology that is concerned with quantitative representations of the ways in which persons consistently differ from one another in their characteristic adaptations to social environments (see *Personality and Adaptive Behaviors*). The basic unit of study is the ‘personality trait’ which is considered to be a moderately heritable disposition to behave consistently across different situations. Such traits are thought to

be manifest in the individual differences in thinking, feeling, and behaving reported in responses to personality questionnaires or to lists of trait-descriptive adjectives. Current research in the field of personality structure emphasizes two quite different structural representations of personality traits: the Five Factor Model (FFM) of personality traits and the Interpersonal Circumplex (IPC) model of interpersonal dispositions. These two models originated in different historical contexts, are based on different assumptions concerning the nature of traits, and employ different statistical models for representing dimensions of personality. Nevertheless, the FFM and the IPC models of personality structure may be viewed as complementary rather than competitive, and as being applicable to different but equally important realms of knowledge.

1. Origins of the Five Factor Model of Personality Structure

The idea that variations in human characteristics have an evolutionary origin was suggested at the end of the eighteenth century by Erasmus Darwin (1731–1801) and it later bore fruit in the empirical researches of two of his grandsons, Charles Darwin (1809–82) and Francis Galton (1822–1911). That Galton was uniquely qualified to conduct research on ‘hereditary genius’ is suggested by the accomplishments and eminence of the Darwin family and by the fact that Galton was, by any criteria, a genius himself. Galton coined the term ‘eugenics’ (Greek *eugenes* ‘well-born’) for the field which studied the inheritance of desirable characteristics, especially intelligence. In 1904, Galton endowed a permanent chair in eugenics at University College, London and appointed the mathematician, Karl Pearson (1857–1936) as its first professor. Pearson developed the correlation coefficient to express the relation between the intelligence of parents and the intelligence of their offspring. His colleague, Charles Spearman (1863–1945) developed the method of factor analysis (see *Personality Psychology: Methods*) which, when applied to mental ability data, suggested a general factor of intelligence (‘g’) in addition to more specific factors. Raymond B. Cattell (1905–98), the founder of modern personality structure research, received his doctorate under Spearman and was the first to suggest extending the applications of factor analysis to the study of individual differences in personality characteristics.

Cattell (1967) had a master plan for the study of personality structure, which he and his associates implemented over a period of 65 years: (a) determine the totality of trait-descriptive terms in a given language that denote individual differences in thinking, feeling, and behaving (‘surface traits’); (b) reduce correlated clusters of these surface trait terms by

factor analysis to their underlying determinants ('source traits'); and (c) confirm, by factor analysis and experimentation, the extent to which these source traits, as measured by personality questionnaires, may also be found in different media of observation such as ratings by observers, self-ratings, and behavioral responses to controlled test stimuli. Cattell's master plan set the stage for two distinct, but conceptually related, approaches to the study of personality structure that have only recently converged.

2. Two Traditions of Five Factor Model Research

2.1 The Lexical Tradition

The totality of surface traits in a given language may be defined with reference to the words that are used to mark individual differences in thinking, feeling, and behaving found in an unabridged dictionary of that language. Within the lexical tradition, representative words (typically adjectives) that have been selected from dictionaries are administered in self-report and observer-report format to groups of respondents who indicate the extent to which each word is descriptive of themselves or of someone they know well. Such data are then analyzed by first calculating the extent of co-occurrences (correlations) among words, and then determining the underlying 'structure' of the entire set of words by factor-analytic techniques.

Lewis R. Goldberg (1981) maintained that: (a) the frequency with which an attribute is represented in a language corresponds to the general importance of that attribute in a given culture and (b) the most important dimensions of personality are those 'universal dimensions' that are common to all languages. Goldberg and his colleagues have studied personality trait-descriptors in English and 12 other languages (e.g., German, Italian, and Turkish) and have concluded that there are *five factors* (which they called the 'Big Five') that may be considered universal dimensions of personality surface trait structure (see *Cross-cultural Psychology*). The Big Five dimensions of personality that have been identified in many different language groups are: (a) Extraversion (talkative, assertive, verbal); (b) Agreeableness (kind, cooperative, warm); (c) Conscientiousness (organized, systematic, practical); (d) Neuroticism (anxious, moody, emotional); and (e) Intellect/Openness (intellectual, imaginative, complex).

2.2 The Multivariate-trait Tradition

The second part of Cattell's master plan involved the development of personality questionnaires that would capture the basic dimensions of personality inherent in language. A personality questionnaire consists of items such as 'I like to attend parties and other social

gatherings where I can meet new people' to which the test-taker is required to respond 'true' or 'false'; or more typically, to rate the degree to which the statement is self-descriptive on a scale ranging from 'not at all' to 'highly descriptive.' Most questionnaires are designed to measure many dimensions or variables, and hence are called 'multivariate.' Because responses to questions ('true vs. 'false'; 'descriptive' vs. 'non-descriptive') can be scored with complete reliability, the tests are referred to as 'objective.'

Cattell's suggestion that Spearman's statistical method of factor analysis could be applied to personality data was indeed a fruitful one, but the new field of personality structure also inherited some of the problems that had been associated with the use of factor analysis to determine the number and nature of mental abilities. Foremost among these problems was the issue of determining the *number* of factors 'underlying' the intercorrelations among personality scales. For example, the two most prominent figures in the field of personality structure during the twentieth century, R. B. Cattell in the United States and H. J. Eysenck in England, strongly disagreed about the number of factors that were required for the comprehensive measurement of personality structure. Cattell argued that at least 16 factors were necessary and he included these factors in his Sixteen Personality Factor Questionnaire (16 PF). Eysenck held that there were, at most, the three factors included in the Eysenck Personality Questionnaire (EPQ). A 1953 survey of the literature by psychologists at the Educational Testing Service identified and categorized the approximately 450 factors that had been reported up to that date. By the 1960s, the field of personality structure was being severely criticized, not only for its failure to agree on a common set of factors, but for its failure to demonstrate any substantial 'real world' correlates of such factors. At the same time, however, it was argued that the 'Big Two' dimensions of personality, Extraversion (see *Self-esteem in Adulthood*) and Neuroticism (see *Neuroticism*) were large, ubiquitous, and almost unavoidable dimensions of personality for which substantial and important correlates had been reported (Wiggins 1968). These dimensions had figured prominently not only in the research programs of Cattell and Eysenck, but in many other research contexts as well.

2.3 Reconciliation of the Two Traditions

P. T. Costa, Jr., and R. R. McCrae (1994) developed a personality questionnaire that included scales for the dimensions of Neuroticism and Extraversion suggested by Cattell and Eysenck, as well as a measure of Openness to Experience that was suggested by an analysis of Cattell's questionnaire. Their instrument, the NEO-Personality Inventory (NEO-PI), was later expanded to include the additional dimensions of Agreeableness and Conscientiousness that had been

identified in the lexical work of Goldberg and others. During the 1980s and 1990s, Costa and McCrae conducted an unprecedented number of empirical studies with the NEO-PI that established the utility of this instrument, as well as the comprehensiveness of what came to be known as the Five Factor Model (FFM) of personality structure. In addition to revitalizing the field of personality structure, the work of Costa and McCrae reaffirmed the basic logic of Cattell's master plan which called for an integrative framework for research that included both the lexical and multivariate-trait traditions.

3. Research with the NEO Personality Inventory

The NEO-PI has been an especially useful measure in the 'applied' contexts of personnel selection (Barrick and Mount 1991) and psychiatric diagnosis (Costa and Widiger 1992). In addition, the validity of the NEO-PI as a conceptually-justified measure of personality traits has been demonstrated with reference to the following criteria that would be required of such a measure.

3.1 Heritability and Stability

All five of the NEO-PI dimensions are moderately heritable (see *Behavioral Genetics: Psychological Perspectives*) and remarkably stable over the adult life of an individual. After the age of 30, these traits, as William James put it, appear to be 'set like plaster.' Costa and McCrae (1994) reported retest correlations ranging from 0.6 to 0.8 even after an interval of 30 years.

3.2 Consensual Validation and Cross-cultural Invariance

There are reasons to be skeptical of the 'truthfulness' of individuals' self-reports to personality inventories and for that reason, it is desirable to obtain ratings by knowledgeable observers of the respondent's personality characteristics on the same items (see *Person Perception, Accuracy of*). In many of the empirical studies conducted by Costa, McCrae, and associates, an Observer Form of the NEO-PI was completed by spouses, peers, or other knowledgeable informants. In general, substantial agreement has been obtained between self- and observer-ratings, indicating a satisfactory degree of 'consensual validity.' The NEO-PI has been translated into more than 30 languages, including non-Indo-European languages such as Thai and Shona. The success of these translations in producing the expected five-factor structure in other cultures attests to the cultural generalizability of the Five-Factor Model.

4. Origins of the Interpersonal Circumplex Model

Circular representations of human nature can be traced from ancient astrology to the drawings of Leonardo Da Vinci to the modern interpersonal circumplex tradition, which began with Harry Stack Sullivan. Sullivan defined psychiatry as the study of *interpersonal relations* and considered 'personality' to reside in an individual's patterns of relationships with significant others over time. Timothy Leary (1957) and his associates attempted to operationalize Sullivanian concepts in terms of concrete measurement procedures. Their analysis of clinicians' observations of the things that patients did to each other (and to themselves) in group psychotherapy led to a taxonomy of interpersonal variables that was empirically well captured by a *circular* arrangement of 16 (later 8) variables organized around the two coordinates of dominance (vs. submission) and affiliation (vs. hostility).

4.1 The Structure of Interpersonal Behavior

The Five Factor Model and Interpersonal Circumplex are based on different assumptions about the nature of personality and employ quite different structural models. The tradition that led to the current FFM focused on the *individual* and attempted to provide a comprehensive 'list' of the factors necessary for describing the individual's most important personality traits. The tradition that led to the current IPC focused on *dyadic relationships* and attempted a detailed description of a two-dimensional 'interpersonal space' within which interpersonal actions and reactions could be classified. The IPC tradition assumes that a person's characteristic interpersonal behavior may be expressed as a *combination* of scores on the two bipolar coordinates of Dominance (vs. Submission) and Affiliation (vs. Disaffiliation).

Interpersonal space is interpreted, quite literally, in terms of the geometry of the circle. The two underlying coordinates of this circle are Dominance (90°) vs. Submission (270°) and Affiliation (0°) vs. Disaffiliation (180°). A number of theoretical rationales have been given for the assumption that these coordinates are the 'basic' reference points for understanding interpersonal behavior. For example, it has been suggested that David Bakan's (1966) distinction between agency (being a differentiated individual, striving for mastery and power) and communion (being part of a larger entity, striving for intimacy and union) provides a metatheoretical basis for interpreting Dominance and Affiliation as the basic axes of social life. Within this conceptual framework, Dominance vs. Submission and Affiliation vs. Disaffiliation are the nodal reference points on the IPC and all other behaviors may be interpreted as 'blends' of these two coordinates. For example, Extraversion (45°) is an

equal blend of Dominance and Affiliation; Introversion (225°) is a combination of submissive and disaffiliative tendencies.

4.2 The Nature of Interpersonal Space

Unlike the competing inventories of personality structure found within the multivariate-trait tradition (e.g., Cattell, Eysenck, Costa, and McCrae), different IPC measures share the same structure, but differ in their domains of application. For example, IPC models of personality and emotions have coexisted comfortably for many years and are likely to continue to do so (Plutchik and Conte 1997). Four contemporary IPC measures have quite different areas of application: The Interpersonal Adjective Scales provide a measure of relatively normal interpersonal *traits* or dispositions. The Inventory of Interpersonal Problems measures *problems* of living as perceived by respondents or others. The Impact Message Inventory measures the *covert reactions* of respondents to different target persons corresponding to different locations on the IPC. The Supportive Actions Scale measures tendencies to provide various kinds of *social support* to others in need of assistance. It is clear from the approximately 1,000 references that appeared in Donald Kiesler's (1996) survey of the IPC literature that both the utility and generalizability of the IPC structure are well established.

5. Current Status

Examination of the relevant literature of the past decade reveals that the two major models of personality structure today are the Five-Factor Model or 'Big Five' and the Interpersonal Circumplex model. An early rapprochement between these two models occurred when proponents of each of the models agreed that the Extraversion and Agreeableness dimensions of the FFM were rotational variants of the interpersonal dimensions of Dominance and Affiliation and that a full description of personality structure cannot ignore the additional dimensions of Conscientiousness, Neuroticism, and Openness/Intellect. It has also become increasingly apparent that the FFM, which had been characterized by some as an 'atheoretical' model, lends itself to interpretations from a variety of different theoretical perspectives (Wiggins 1996).

See also: Extraversion; Genetic Studies of Personality; Neuroticism; Personality and Conceptions of the Self; Personality Assessment; Personality Development and Temperament; Personality Development in Adulthood; Personality Development in Childhood; Personality Psychology; Personality Theories

Bibliography

- Bakan D 1966 *The Duality of Human Existence: An Essay on Psychology and Religion*. Rand McNally, Chicago
- Barrick M R, Mount M K 1991 The Big Five personality dimensions and job performance: A meta-analysis. *Personnel Psychology* **44**: 1–26
- Cattell R B 1967 *The Scientific Analysis of Personality*. Penguin, Harmondsworth, UK
- Costa P T, Jr., McCrae R R 1994 'Set like plaster'? Evidence for the stability of adult personality. In: Heatherton T, Weinberger J L (eds.) *Can Personality Change?*. American Psychological Association, Washington, DC, pp. 21–40
- Costa P T, Widiger T A Jr. (eds.) 1994 *Personality Disorders and the Five-Factor Model of Personality*. American Psychological Association, Washington, DC
- Goldberg L R 1981 Language and individual differences: The search for universals in personality lexicons. In: Wheeler L (ed.) *Review of Personality and Social Psychology*. Sage, Beverly Hills, CA, Vol. 2, pp. 141–65
- Kiesler D J 1996 *Contemporary Interpersonal Theory and Research: Personality, Psychopathology, and Psychotherapy*. Wiley, New York
- Leary T F 1957 *Interpersonal Diagnosis of Personality*. Ronald Press Co., New York
- Plutchik R, Conte H R (eds.) 1997 *Circumplex Models of Personality and Emotions*. American Psychological Association, Washington, DC
- Wiggins J S 1968 Personality structure. *Annual Review of Psychology* **19**: 293–350
- Wiggins J S (ed.) 1996 *The Five-factor Model of Personality: Theoretical Perspectives*. Guilford Press, New York

J. S. Wiggins

Personality Theories

Within psychology, personality refers to the individual's distinctive and enduring characteristics, including stable patterns of thoughts, feelings, emotions, and behavioral tendencies. No other area of psychology covers as much territory as the field of personality does; it is at the intersection of the study of human development, abnormality and deviance, temperament, emotions and thought, learning, social relations, and even the biological foundations of human behavior. Although the traditional focus is on stable individual differences in basic tendencies or dispositions (e.g., extraversion, conscientiousness), recent approaches also emphasize the psychological processes (such as learning, motivation, emotion, and thinking) that underlie them. Increasing attention is also given to the possible biological-genetic roots of these tendencies and processes, and their interactions with social-environmental influences.

Personality theorists try to answer such questions as: How stable and enduring are particular psychological qualities? What is the role of inheritance in the expression of personality, and what is acquired through experience with the environment? Does what

we do and think and feel characteristically depend mostly on the individual or on the situation? How do different types of people interact with different types of situations? Historically, there have been five main approaches to personality that address these questions: the psychodynamic, the behavioral, the phenomenological, the trait and biological, and the social cognitive. All are concerned with uncovering basic, general principles that can account for the wide range of behavior that people are capable of engaging in, and each proposes different units for conceptualizing and studying people. Each approach is considered in turn.

1. Psychodynamic Approaches

The psychodynamic approach to personality was founded by the Viennese physician Sigmund Freud; (see *Freud, Sigmund (1856–1939)*). Rejecting the earlier belief that people's behavior is under their conscious and rational control, Freud believed that behavior is psychologically determined by underlying unconscious causes and motives. Freud saw patients who displayed strange symptoms that had no discernible physical cause, such as a young woman who appeared to be blind although tests of her vision showed that her eyes and visual system were undamaged. To explain such puzzling cases, Freud proposed an 'anatomy' of the mind consisting of three institutions or agencies of mental operation—the id, the ego, and the superego. The id—the primary, instinctual core—obeys the 'pleasure principle,' seeking immediate gratification of basic biological sexual and aggressive impulses, regardless of reality considerations. The superego, on the other hand, represents the moral standards of a person's society, obtained through the internalization of parental values, rules, and characteristics in the course of socialization. The ego mediates between the instinctual demands of the id and the outer world of reality, localizing the appropriate objects for gratification in the environment so that the impulses of the id can be satisfied. It operates by means of logical thinking and rational planning. Its tasks include: (a) the control of unacceptable impulses or desires from the id; (b) the avoidance of pain produced by internal conflict in the efforts to control and master those unacceptable impulses; and (c) the attainment of a harmonious integration among the needs of both the id and the superego.

According to Freud ([1915] 1957), the id, the ego, and the superego are always in dynamic conflict (thus the term psychodynamics). Freud believed that in this conflict, a person's basic impulses persist and press for discharge, but the people or 'objects' at which they are directed, and the manner in which they are expressed, are transformed and disguised unconsciously to make them more acceptable and to reduce conflict and

anxiety. When the young child becomes anxious about his or her own impulses (e.g., because of fear that they will lead to parental punishment), attempts are made to repress them or otherwise disguise and redirect them. The main defense mechanism that Freud proposed—repression—massively inhibits threatening impulses or events, making them unconscious. In projection, one's unacceptable impulses or aspects are attributed to other people.

Modern psychodynamic theories tend to place much less emphasis on biological drives and highlight the individual's social circumstances and relationships with significant others. Attachment theorists, for example, emphasize the quality and varieties of early mother-child attachment relations and their consequences in the course of development (Ainsworth and Bowlby 1991). Based on experiences in this relationship, the child develops internal working models (mental representations) of others, of the self, and of relationships, which guide subsequent experience and behavior. Children who have positive, gratifying experiences with significant others tend to develop internal working models of others as responsive and giving, and of themselves as worthy of attention; those who have had painful or unsatisfying experiences are likely to develop internal models that reflect those troubled relationships. Early attachment styles may enduringly influence relationships throughout life, including the way one parents one's own children (Hazan and Shaver 1987; see *Attachment Theory: Psychological*; Bowlby, John (1907–90)).

Many of the ideas introduced by the psychodynamic approach, notably that much mental activity occurs outside of mental awareness, also are being reconceptualized in light of current theory and findings on how thought and memory work. For example, when cognitive (mental) representations of significant others are activated by a newly encountered person, the feelings and attributes associated with this representation may be applied to that new person in making inferences about him or her (Andersen and Berk 1998). This is consistent with Freud's concept of transference, said to occur when the patient responds to the therapist as if he or she were the patient's father, mother, or some other childhood figure. This modern approach, however, views transference in information-processing terms rather than as a reflection of the psychosexual impulses and conflicts hypothesized in classical psychodynamic theory (Westen 1998; see *Mental Representation of Persons, Psychology of*).

2. Behavioral Approaches

Although many psychologists were intrigued by the insights of Freud and his followers, they were primarily dedicated to developing a more scientific, rigorous approach to personality that would be objectively testable, preferably by laboratory tech-

niques. One of the pioneers in the behavioral approach, the US psychologist B. F. Skinner (1904–90), criticized many motivational concepts as being no more than labels, and even fictions, that were unhelpful for explaining what people actually do. Skinner and other behaviorally oriented psychologists analyzed problematic behaviors (e.g., aggressiveness, bizarre speech patterns, smoking, fear responses) in terms of the observable events and conditions that seem to vary with them. They then tried to discover the external events that strengthened their future likelihood and that maintained or changed the behavior of concern. Behavioral approaches have led to innovations for therapy by attempting to modify problematic behavior not through insight, awareness, or the uncovering of unconscious motivations, but rather by addressing the behavior itself and modifying it directly. Such behavior therapies emphasize the individual's current maladaptive behaviors (rather than their historical origins) and assume that they can be understood and changed by learning principles that deal with the acquisition and extinction of behavior patterns.

Although the systematic rigor of behavioral approaches was widely appreciated, the relevance of the approach for understanding the complex lives of people under the often unpredictable social conditions of life remained in question. Thus, a number of theorists began to make learning and behavior theories more 'social,' not limiting themselves to simple reinforcement principles and increasingly relying on mental or 'cognitive' processes in their account of the development of complex social behavior. In Albert Bandura's (1986) social learning approach, for example, personality is seen as the outcome of social learning processes through which distinctive beliefs, self-efficacy expectations, as well as personal standards that guide characteristic behavior patterns, are acquired. He emphasized that much social learning, from table manners and interpersonal relations to working habits to coping patterns, occurs through observation of the behavior of social models without any direct rewards or reinforcement administered to the learner (see *Social Learning, Cognition, and Personality Development*).

In another direction, the assumption that the 'laws' of conditioning are universal and apply broadly has been challenged by convincing evidence that organisms (including humans) seem biologically prepared to learn some associations or pairings more readily than others. Prewired dispositions in the brain seem to make humans distinctively prepared to learn diverse high-level mental activities, from language acquisition to mathematical skills to space perception (Pinker 1997). In yet another direction, technological advances in brain imaging now allow researchers to use methods like Magnetic Resonance Imaging (MRI) to observe areas within the brain that become activated in the course of mental activity, such as the emotional centers that may be especially important bases for individual

differences (Ledoux 1996). Advances in technology have therefore made it possible to objectively study mental events, such as emotions, memories, and attention, going far beyond the early behaviorism that confined itself to overt behavior.

3. Phenomenological Approaches

In the middle of the twentieth century, phenomenological approaches arose, in part, as a humanistic protest against the earlier psychodynamic and behavioristic views. Phenomenologically oriented theorists argued that personality is not merely passively molded by internal motivational or external situational forces that 'shape' what the person becomes. Instead, people are active agents in the world and have a measure of control over their environment and their own lives. In this view, people are considered capable of knowing themselves and of being their own best experts. Self-knowledge and self-awareness become the route to discovering one's personality and genuine self.

Phenomenological approaches to personality (sometimes called self theories, construct theories, and humanistic theories), tend to reject many of the motivational concepts of psychodynamic theories and most of the environmental determinism of behavioral theories. Instead, their focus is on the development of an active 'self': People develop self-concepts and goals that guide their choices and their life course. Understanding personality, as well as the person's goals and choices, requires attention to how the individual characteristically perceives, thinks, interprets, and experiences or even 'constructs' the personal world.

George Kelly's (1905–67) theory of personal constructs, for example, emphasized people's subjective perceptions as the determinants of their behavior. Kelly believed that, just like scientists, people generate constructs and hypotheses both about themselves and about how the world works; they use these constructs to anticipate, understand, and control events in their lives. Therefore to understand people, one has to understand their constructs, or personal theories. Problems develop when the constructs people generate don't work well for them, when they are 'bad scientists' and fail to 'test' their constructs or hypotheses against the realities of the environment, or when they see themselves as helpless victims of their own personalities or life situations. Kelly's principle of 'constructive alternativism' held that all events in the world, including one's own behavior and characteristics, can be construed in multiple, alternative ways. While it is not always possible to change these events, one can always construe them differently, thus influencing how one is affected by them and how one reacts to them.

Carl Rogers (1902–87), another pioneer of the phenomenological approach, proposed two systems:

the organism and the self (or self-concept). The organism is the locus of all experience, which includes everything potentially available for awareness. The self is that portion of the perceptual field that is composed of perceptions of characteristics of the 'I' or the 'me.' It develops from experiences and interactions with the environment, and also shows a tendency towards actualization. Rogers maintained that the central force in the human organism is the tendency to actualize itself—to move constructively in the direction of fulfillment and enhancement. The self may be in opposition or in harmony with the organism. When the self is in opposition or incongruence with the experiences of the organism (e.g., when the self tries to be what others want it to be instead of what it really is), the person may become anxious, defensive, and rigid. However, when the self is open and accepting of all of the organism's experiences without threat or anxiety, the person is genuinely psychologically adjusted, for the self and the organism are one.

In contemporary work, the 'self' is seen as multifaceted and dynamic, consisting of multiple self-concepts that encode different aspects of the person (e.g., self as lover, self as father, the 'ideal' self, the 'actual' self) and become differentially salient depending on context (Markus and Nurius 1986). According to Higgins (1987) for example, a perceived discrepancy between the mental representation of the person one would ideally like to be (the ideal self) and the representation of who one actually is (the actual self) makes one more vulnerable to feelings of dejection, such as disappointment or dissatisfaction. In contrast, a discrepancy between one's representation of who one ought to be (the ought self) and the actual self can lead to feelings of agitation such as fear and worry. Motivation for behavior change arises from the conflicts each individual feels among his or her various representations of the self. For instance, upon receiving a low grade on an exam, an undergraduate may subsequently study very hard to relieve the guilt of not living up to what she herself perceives to be her responsibility as an exemplary student. Alternatively, she may re-evaluate her negative interpretation of past events, thinking about all of the good grades she has got in other classes and the myriad of other activities she is involved in (see *Personality and Conceptions of the Self*).

4. Trait and Biological Approaches

In everyday life, people readily characterize each other in terms of personality characteristics: he or she is friendly, assertive, submissive, conscientious, and so on. The essence of the trait approach, whose fundamental premises date back to the ancient Greeks, is the assumption that behavior is primarily determined by a number of stable, generalized personality traits that express themselves in many contexts. Guided by

this assumption, advocates of this approach try to identify and measure individuals' traits and to discover the most fundamental traits on which people can be compared.

A principal focus of research on traits is on measurement—that is, the development of quantitative ways of finding and describing important stable individual differences. Traits are inferred from questionnaires, ratings, and other reports about the person's dispositions. Usually, the person's self-reports (or someone else's reports about the person) are taken as direct signs of the relevant traits. For example, the more one rates oneself as aggressive, the more one is assumed to have an aggressive disposition. The trait approach recognizes that behavior can vary depending on the situation but has focused on individual differences in the overall response tendency averaged across many situations.

Some consensus has grown among many researchers to focus on five large factors or dimensions of personality that have emerged from statistical analyses of traits. These factors, often referred to as the 'Big Five,' comprise openness to new experience, conscientiousness, extraversion (or outgoingness), agreeableness, and neuroticism. Considerable stability has been demonstrated on trait ratings and questionnaires related to the Big Five, particularly during the adult years (McCrae and Costa 1990; see *Extraversion; Neuroticism*).

In a different direction, the British psychologist Hans Eysenck (1916–97) and his associates have led the way in connecting psychological dispositions to their biological foundations. According to Eysenck, introverts need only a small amount of stimulation to overstimulate their central nervous system (CNS) which then leads them to become withdrawn in their behavior. In extraverts, by contrast, the CNS is not easily stimulated, leading them to seek activities that will increase stimulation levels, for example, by socializing more actively and by seeking activities such as parties. In support of his theory, Eysenck (1971) found that extraverts reported earlier, more frequent, and more varied sexual experiences. In another study, introverts showed greater changes in their brain wave activity in response to low frequency tones (Stelmack and Michaud-Achiron 1985), indicating their lower threshold for stimulation to the CNS.

4.1 Genetic Roots of Personality

The rapidly developing field of behavioral genetics studies the role of inheritance in personality, both in terms of dimensional traits (such as extraversion–introversion) and temperaments (such as general levels of emotionality, sociability, and activity; see *Temperament and Human Development*). Behavioral genetics most often uses the 'twin method' to assess genetic influence, comparing the degree of similarity on trait measures obtained for genetically identical (mono-

zygotic) twins as opposed to twins that are fraternal (dizygotic) and are only 50 percent similar genetically. To the degree that genetic factors affect a trait, it follows that identical twins must be more similar than fraternal twins with regard to that trait. Estimates of genetic influence vary across studies and measurement strategies. For example, they tend to be higher for self-reports of personality than for observational studies (Miles and Carey 1997). Nevertheless, the general conclusion that identical twins are more similar than fraternal twins has received considerable support, and indicates an important role of heritability in personality (Plomin et al. 1997).

Even highly heritable dispositions, however, can be constrained and limited in their full expression, as when a person's growth and ultimate height are affected by nutrition, disease, or development. Environmental influences also can change the hard wiring of the brain—the neuronal structures themselves—and thus produce stable changes within the person at the organic level. For example, stress can shrink the size of the hippocampus, a brain structure basic for higher order mental functions (Sapolsky 1996). Thus, although the social environment cannot affect the structure of DNA (barring biochemical or radiation exposure), it can influence their expression, the brain, and the person's personality (see *Behavioral Genetics: Psychological Perspectives*; *Stress: Psychological Perspectives*).

Unquestionably, one's genetic endowment has extensive influence on one's life and personality development. Just as noteworthy, however, the same findings also point to the importance of experiences and the environment for personality (Plomin et al. 1997). The expressions of genetic influences and the situations and events the person experiences are in continuous interaction, and it may be this interplay that is most important in personality development.

5. Social Cognitive Approaches

In the 1960s, a 'cognitive revolution' took place in psychology, as attention turned from behavior to the person's mental (cognitive) processes and structures. Although behavioral approaches had asserted that stimuli control behavior, evidence accumulated to suggest that the perceivers' mental representations and cognitive transformations of stimuli can determine and even reverse their impact. Such transformations were illustrated in research on the determinants of people's ability to forgo gratifying their immediate impulses for the sake of more highly desired but temporally distant goals. This kind of 'willpower' has been studied extensively in the delay of gratification paradigm, in which preschoolers are faced with the dilemma of being able to have a small, less desired reward now (e.g., one little pretzel or marshmallow) or having to wait instead for a larger, more desired

reward (e.g., two little pretzels or two marshmallows). The number of seconds that preschoolers are able to delay predicts a variety of positive developmental outcomes, including their SAT scores and their social competence in adolescence (Mischel et al. 1989; see *Self-regulation in Childhood*; *Self-regulation in Adulthood*). The duration of delay time itself depends importantly on how the children represent rewards mentally and selectively attend to different features of the situation. For example, if the young child focuses cognitively on the consummatory qualities of the reward objects, such as the pretzel's salty, crunchy taste, he or she tends to be able to wait only a short time. By thinking about the stick pretzels as little logs or about the marshmallows as puffy clouds, however, the child may be able to wait much longer for the reward. These results indicate that what is in the children's heads—not what is physically in front of them—determines their ability to delay.

Concurrent with the cognitive revolution, questions arose about cross-situational traits and psychodynamics as basic units of personality. Walter Mischel (1968) in particular showed that what people do characteristically depends importantly on the situation and context. For example, the person who seems conscientious about work may show a very different pattern with family. These findings indicated that broad trait scores and projective psychodynamic methods do not accurately predict what a given person will do in different kinds of situations and fail to explain the seeming inconsistency or variability that was observed within individuals across those situations.

The social cognitive approach to personality emerged in the 1960s and 1970s as an attempt to understand both the stable ways in which the person cognitively processes social information and the ways in which the person's behavior varies with regard to the situation. In this approach, the basic 'units' of personality are conceptualized as a set of five relatively stable person variables (Mischel and Shoda 1995): the individual's encodings or construals (of self, other people, situations); expectancies and beliefs (about outcomes and one's own self-efficacy); subjective goals and values; competencies (for the construction and generation of social behavior) and self-regulatory strategies and plans in the pursuit of goals; and their affects (feelings, emotions, and affective responses). Furthermore, clues about the person's underlying qualities may be seen in when and where a type of behavior is manifested, not only in its overall frequency. If so, the patterns of situation-behavior relationships shown by a person might be a possible key to individuality and personality coherence, rather than a source of error to be eliminated systematically (see *Interactionism and Personality*).

Evidence for the existence and meaningfulness of the stable situation-behavior relations predicted by the social cognitive approach came from an extensive

observational study conducted in a residential summer camp setting for children (Shoda et al. 1994). In this study, children's behavior was recorded specifically in relation to five camp-relevant situations (e.g., being teased, threatened, or provoked by peers; being praised by an adult). The children's social behavior was observed on selected dimensions (e.g., aggression, withdrawal, friendly behavior) as it occurred in relation to each of the interpersonal situations. The individuals' situation-behavior, 'if ... then ...' personality 'signatures' (e.g., if situation X, then they do A, but if situation Y, then they do B) were found to be both distinctive and stable. For example, one child was consistently more verbally aggressive than others when warned by an adult, but showed less aggression than others on average when approached positively by a peer. In contrast, another child was most verbally aggressive when approached positively by peers, but not particularly aggressive when warned by an adult.

To account for such findings, a cognitive-affective personality system (CAPS) theory has been proposed (Mischel and Shoda 1995). In this theory, the individual is characterized not only by the particular subset of goals, ways of encoding the world, and self-regulatory competencies that may be potentially accessible to him or her, but also by the distinctive and stable organization of relationships among the person variables available in the personality system—i.e., the person's distinctive 'network.' When a person encounters a particular situation, the CAPS network is sensitive to particular features of situations, which become encoded and activate situation-relevant cognitions and affects (thoughts and feelings) within the system. These units, in turn, make other cognitions and affects accessible while inhibiting others. The organization of relations within the person's network remains relatively stable and invariant across situations. However, as the individual moves across situations that contain different psychological features, different mediating units and their characteristic interrelationships become activated in relation to these psychological conditions.

Thus the person's unique 'network'—which is distinctive both in the types of mediating units available within the system as well as in the relationships among these units—guides and constrains the activation of the specific cognitions, affects, and potential behaviors when an individual processes situational features. It constitutes the basic structure of personality and reflects and underlies the individual's uniqueness. When the *ifs* posed by the situation change, so do the *thens* generated by the personality system, but the relationship between them is stable. This type of system is intrinsically interactionist, and has been shown to generate both overall mean differences in a given behavior as well as the specific *if ... then ...* profiles that are a person's behavioral 'signature.'

To illustrate such a system in action, suppose a person is especially sensitive to rejection and is

disposed to expect it, to look for it, and to react intensely to it (Downey and Feldman 1996). Such 'rejection sensitive' people may see even innocent or ambiguous behavior from a significant other as intentional rejection, triggering such thoughts as 'she doesn't love me,' which activate further thoughts and feelings of rejection and potential betrayal and abandonment. In turn, the person's scripts for coercive or controlling behaviors may become activated, leading to angry or hostile and even abusive reactions. Over time such hostility is likely to lead to actual rejection even when there was none before, further strengthening the cycle that characterizes this personality type.

Thus, the defining *if ... then ...* profile of rejection sensitivity—its behavioral signature—may include both being more prone than others to anger, disapproval, and coerciveness in certain types of situations in intimate relationships, as well as being more supportive, caring, and romantic than most people in other situations (e.g., in initial encounters with potential partners who are not yet committed to them). This example illustrates how the personality system remains stable, although the surface behaviors it generates change in relation to the situation. As in a musical piece, the notes played at any moment change, but they do so in an organized pattern that reflects the structure of the composition.

6. Future Directions: Toward a Cumulative Science of Personality

Historically, the field of personality has been characterized by many alternative approaches and conceptions of personality which competed against each other. Different approaches and theorists claimed to offer a comprehensive, unitary view of personality, to the exclusion of alternative viewpoints or approaches. However, the different approaches at least in part asked different questions, usually dealing only with selected aspects of the diverse phenomena subsumed under the construct of personality.

Personality psychologists are increasingly recognizing the importance of multiple approaches to understanding the person and many are trying to find ways to integrate them within a broader theoretical framework. Current psychodynamic approaches are beginning to incorporate insights from research on memory to test the fundamental tenets of psychodynamic theory. Similarly, some behavioral approaches are integrating findings on social learning and cognitive processes, aided by current technology to assess interactions between biological variables and learning. The phenomenological approach, with its emphasis on the person's subjective internal reality, continues to inform work on self-conceptions and how these influence the individual's choices and life course. Trait approaches are beginning to specify the boundary conditions within which traits will be selectively

activated and visible in behavior. Finally, cognitive-social approaches are incorporating the contributions of cognitive science and social psychology into the same framework used to understand the individual's characteristic cognitive-affective processes and dynamics. If these trends continue, personality psychology may be evolving into a more unified field that addresses the distinctive characteristics that people have and the psychological processes that underlie them, conceptualized within one overarching theoretical framework.

See also: Freud, Sigmund (1856–1939); Genetic Studies of Personality; Infant and Child Development, Theories of; Interactionism and Personality; Personality and Adaptive Behaviors; Personality and Conceptions of the Self; Personality and Crime; Personality and Marriage; Personality and Risk Taking; Personality and Social Behavior; Personality Assessment; Personality Development and Temperament; Personality Development in Adulthood; Personality Development in Childhood; Personality Psychology; Personality Psychology: Methods; Personality Structure; Psychological Development: Ethological and Evolutionary Approaches; Social Learning, Cognition, and Personality Development

Bibliography

- Ainsworth M D S, Bowlby J 1991 An ethological approach to personality development. *American Psychologist* **46**: 331–41
- Andersen S M, Berk M S 1998 Transference in everyday experience. Implications of experimental research for relevant clinical phenomena. *Review of General Psychology* **2**: 81–120
- Bandura A 1986 *Social Foundations of Thought and Action: A Social Cognitive Theory*. Prentice Hall, Englewood Cliffs, NJ
- Downey G, Feldman S I 1996 Implications of rejection sensitivity for intimate relationships. *Journal of Personality and Social Psychology* **70**: 1327–43
- Eysenck H J 1971 Introverts, extroverts and sex. *Psychology Today* **4**: 48–51
- Freud S [1915] 1957 *Instincts and their Vicissitudes*. *Standard edition*, Vol. 14. Hogarth, London
- Freud S 1959 *Collected Papers*. Basic Books, New York, Vols. 1–5
- Hazan C, Shaver P 1987 Romantic love conceptualized as an attachment process. *Journal of Personality and Social Psychology* **52**: 511–24
- Higgins E T 1987 Self-discrepancy: A theory relating self and affect. *Psychological Review* **94**: 319–40
- Kernberg O 1976 *Object Relations Theory and Clinical Psychoanalysis*. Jason Aronson, New York
- Kohut H 1980 *Advances in Self Psychology*. International Universities Press, New York
- Ledoux J 1996 *The Emotional Brain*. Simon and Schuster, New York
- Markus H, Nurius P 1986 Possible selves. *American Psychologist* **41**: 954–69

- McCrae R R, Costa P T Jr. 1990 *Personality in Adulthood*. Guilford Press, New York
- Miles D R, Carey G 1997 Genetic and environmental architecture of human aggression. *Journal of Personality and Social Psychology* **72**: 207–17
- Mischel W 1968 *Personality and Assessment*. Wiley, New York
- Mischel W, Shoda Y 1995 A cognitive-affective system theory of personality: Reconceptualizing situations, dispositions, dynamics, and invariance in personality structure. *Psychological Review* **102**(2): 246–68
- Mischel W, Shoda Y, Rodriguez M L 1989 Delay of gratification in children. *Science* **244**: 933–8
- Pinker S 1997 *How the Mind Works*. Norton, New York
- Plomin R, DeFries J C, McClearn G E, Rutter M 1997 *Behavioral Genetics*, 3rd edn. W. H. Freeman, New York
- Rogers C R 1980 *A Way of Being*. Houghton Mifflin, Boston
- Sapolsky R M 1996 Why stress is bad for your brain. *Science* **273**: 749–50
- Skinner B F 1974 *About Behaviorism*. Knopf, New York
- Shoda Y, Mischel W, Wright J C 1994 Intra-individual stability in the organization and patterning of behavior. Incorporating psychological situations into the idiographic analysis of personality. *Journal of Personality and Social Psychology* **65**: 1023–35
- Stelmack R M, Michaud-Achorn A 1985 Extraversion, attention, and habituation of the auditory evoked response. *Journal of Research in Personality* **19**: 416–28
- Westen D 1998 The scientific legacy of Sigmund Freud: Toward a psychodynamically informed psychological science. *Psychological Bulletin* **124**: 333–71

W. Mischel and R. Mendoza-Denton

Personality Theory and Psychopathology

The idea that enduring personality traits are closely related to various disease states or behavioral disorders can be traced back to initial speculation by the early Greeks. Hippocrates believed that all disease stemmed from imbalances in four bodily humors: yellow bile, black bile, blood, and phlegm. While Hippocrates' early ideas bear a rudimentary resemblance to some recent approaches to personality, most contemporary theories of the relationship between personality and psychopathology are considerably more complex, and sensitive to both endogenous and environmental variables. Before the personality–psychopathology relationship can be understood clearly, the concept of personality must be defined.

1. The Concept of Personality

The word personality is derived from the Greek *persona*, which referred to masks used in the early theater. Over time, the term has come to refer not only

to the exterior presentation of an individual, but the integrated internal systems which allow individuals to adapt to their environment. For example, Allport (1937) defined personality as 'the dynamic organization within the individual of those psychophysical systems that determine his unique adjustments to his environment.' This definition highlights the idea that personality reflects both psychological and biological processes that are used in coping with the everyday world. These points are made more clearly by Millon (1996, p. 4), who defines personality as 'a complex pattern of deeply embedded psychological characteristics that are largely nonconscious and not easily altered, expressing themselves automatically in almost every facet of functioning. Intrinsic and pervasive, these traits emerge from a completed matrix of biological dispositions and experiential learnings and ultimately comprise the individuals distinctive pattern of perceiving, feeling, thinking, coping, and behaving.'

Another important characteristic of personality is that it pertains to long-standing patterns of behavior which are relatively stable over time and are typically referred to as traits. The American Psychiatric Association has defined personality in the *Diagnostic and Statistical Manual of Mental Disorders* (4th edn.) as 'enduring patterns of perceiving, relating to, and thinking about the environment and oneself.' Taken together, the many definitions of personality which have emerged over the years highlight several consistent features: the integration of biological and psychological systems, which results in trait-like behavioral stability and function to promote environmental adaptation.

2. Conceptual Debates Related to Personality

Numerous conceptual debates have characterized the evolution of the study of personality. Pervin (1990) outlines several of these issues: 'Is personality primarily governed by internal factors of the person or environmental effects?'; 'Is personality determined primarily by nature or nurture?'; 'Is personality influenced by past experiences, present factors, or future expectations?'; and 'To what degree is personality modifiable?'

While each of these topics is significant in and of itself, this article will highlight several other key issues regarding the nature of personality which have generated considerable discussion.

3. Trait-like Stability vs. Situational Determinism

As noted previously, personality has historically referred to trait-like behavioral patterns which are stable across situations and time. However, a series of empirical studies in the 1920s and 1930s raised serious

doubts about the trait-like stability of behavior and highlighted the significance of situational factors in determining variability in behavior. Classic studies by Columbia University Teachers College (1928) suggested that children's level of honesty did not follow a trait-like pattern of predictability, but was significantly influenced by the child's immediate environmental situation. Furthermore, numerous other studies emerged suggesting that specific behaviors measured across situations and time were quite variable, further calling into question the idea of trait like stability of behavior (Mischel 1968). Consequently, this era was characterized by a polemic atmosphere in which person vs. situation debates were common.

Ultimately, this criticism of trait constructs was followed by a period of greater integration of situational and trait theories of personality. For example, work by Epstein highlighted that personality traits may not be accurate predictors of specific individual acts, but may be excellent predictors of aggregated behavior across time and situations (Epstein 1979, Epstein and O'Brien 1985). In other words, traits may be good predictors of a person's behavior 'over the long haul.' Furthermore, theories emphasizing person-situation interactionism evolved and brought into greater focus the ways in which personality traits may interact with particular situational variables. The interactionist solution to the person-situation debate was to consider which personality and situational factors interact to produce consistent behaviors vs. those which interact to produce variable behavior (Ekehammer 1974). Behavior within this approach is considered to flow from transactions between a person and the environment, each of which influences the other.

4. Should Disturbances in Personality be Conceptualized as Categories or Dimensions?

Personality researchers frequently debate about whether personality can best be understood in categorical or dimensional terms. That is, is it more useful to determine what 'type' of personality an individual displays rather than what profile of various personality 'traits' best describes the individual? This becomes particularly significant for the study of disorders of personality, given that the primary classification schemes for psychiatric disorders (e.g., *Diagnostic and Statistical Manual of Mental Disorders*, *International Classification of Diseases*) rely on categorical models (see *Syndromal Diagnosis vs. Dimensional Assessment*, *Clinical Psychology* of). Categorical models parallel traditional, medical disease-based practices in which a person is considered either to have a given disease entity or not. Such an approach has a long history in psychiatry and psychology and offers several advantages: (a) clinicians are generally familiar with

the descriptive and clinical terminology and can apply it easily; (b) predictions can be made about a person based on the description of the category; for example, if a person is categorized as sociopathic, we may assume that rule violation or criminal activity is likely; (c) personality types have clear and vivid descriptions which facilitate communication; (d) a disease based approach is appropriate to the extent that personality disturbances are discrete and homogeneous diagnostic entities with minimal overlap between categories.

Unfortunately, the empirical literature has not supported this idea. People generally do not have a particular personality type that can be easily categorized. Rather, personality appears to be better described as an amalgam of traits, and most empirical studies suggest that personality is best conceptualized in such terms. Categorical models of personality have also been criticized because different measures of the same personality category show relatively low concordance or agreement (i.e., poor convergent validity) and also individuals who meet diagnostic criteria for a given personality type often meet criteria for several other types (i.e., poor discriminant validity). Indeed, if personality could be best understood in terms of unique and nonoverlapping categories, such problems with construct validity would not be so prominent. Furthermore, setting the minimum threshold for determining the presence of a personality disorder is a fairly arbitrary procedure and confusion regarding normality vs. abnormality is common. Finally, many personality disorder categories have not shown substantial stability over time, which would be anticipated with a true trait oriented personality disorder.

In contrast, the dimensional personality trait perspective has received considerable support in empirical studies. For example, when individuals with personality disorders are compared with individuals with normal personality, the distribution of personality scores does not show an 'either-or' pattern, suggesting that personality disordered individuals differ from normals only in the *magnitude* of certain key personality traits. Furthermore, a large number of factor analytic studies of personality traits in both normal samples and personality disordered samples yield the same basic factor structure for each population, most often implying that four or five broad personality traits, or factors, may account for variations in both normal and disordered personality. These five factors are frequently referred to as the Five Factor Model (FFM) and they are traditionally labeled in the following manner: Factor 1, surgency or extroversion; Factor 2, agreeableness; Factor 3, conscientiousness; Factor 4, emotional stability or neuroticism; and Factor 5, culture or openness to experience.

From a clinical perspective, the lack of empirical support for categorical distinctions between normal and abnormal personality and different personality types has led some researchers to suggest that personality disorders do not fit the medical model, and

consequently personality 'diagnoses' should not be made. Rather, personality should be 'assessed' along core dimensions, such as the FFM or other clinical dimensions, (e.g., Minnesota Multiphasic Personality Inventory, see Helmes and Reddon (1993) for a review). Livesley et al. (1994) suggest that the simple presence of a personality disorder be determined on the basis of a behavioral pattern that results in personal or interpersonal *dysfunction*, but is then assessed, or described, on the basis of a dimensional system (e.g., FFM). Such an approach acknowledges that certain personality styles indeed cause distress (i.e., disorder), but retains the comprehensive assessment afforded by the dimensional approach. Whether or not such a system will ever be employed in psychiatric or psychological practice is likely to be determined more by political than by scientific forces.

5. *Relationship Between Personality and Psychopathology*

The idea that personality displays an important relationship to various forms of psychopathology has a long-standing history in psychopathology research. Conceptual understandings of the nature of this relationship have varied and generally represent one of several perspectives (Klein et al. 1993, Lyons et al. 1997, see Fig. 1). First, the *predispositional hypothesis* implies that personality variables precede and increase the risk of developing psychopathology. Also, personality and psychopathology are considered to be independent entities in terms of underlying causes or pathophysiology. An example of this perspective can be seen in early speculation that particular personality traits such as compliance, perfectionism, and dependence increase the risk of developing anorexia nervosa. Similarly, one may speculate that in some people, dependency may precede and increase the chances of developing depression. A second perspective, frequently called the *complication* or *scar hypothesis*, implies the opposite relationship between personality and psychopathology. From this view, psychopathology precedes and influences the development of particular personality traits which are considered the residue of the more active psychopathological syndrome. For example, research suggests that personality variables such as interpersonal dependence or obsessionality may increase after the onset of a major psychiatric syndrome such as depression, but then decrease as the syndrome resolves. The 'personality trait' may not be a true trait, but instead a complication of the depression.

A third perspective posits that both personality and psychopathology arise from the same underlying causal process. From such a *spectrum* or *common cause hypothesis*, a third variable is thought to increase the risk of both entities. From a spectrum perspective,

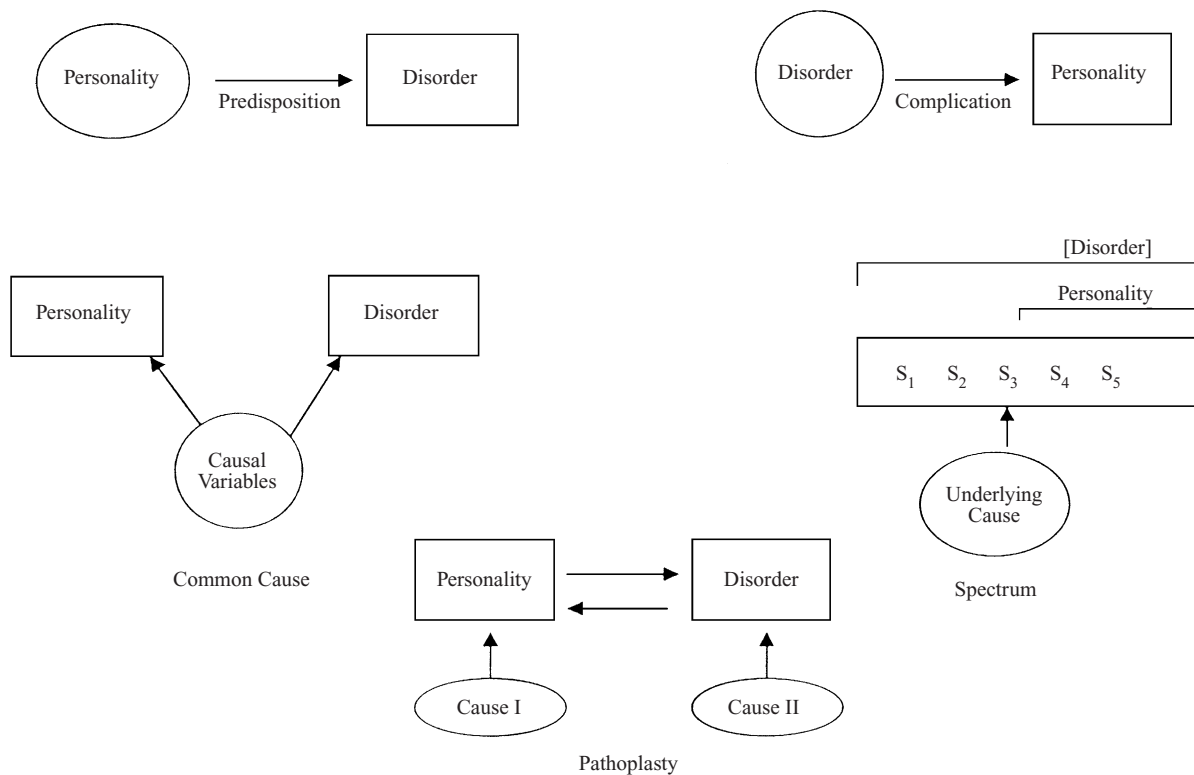


Figure 1

Models of relationship of personality and psychopathological disorders; S_1 – S_5 = symptoms

personality and psychopathology are considered to be quantitative variations of the same underlying process. For example, it could be hypothesized that schizotypal personality traits and schizophrenia each are ‘caused,’ in part, by a particular genetic pattern. Here, schizotypal personality is not thought to cause, or be caused, by schizophrenia. Rather, it reflects variations of the same fundamental disturbance. This ‘spectrum disorder’ may fluctuate from periods of schizotypal personality to more active schizophrenic episodes. Common cause models are similar to spectrum approaches in specifying a common causal variable, but differ because they suggest that personality traits and specific forms of psychopathology are independent entities and not quantitative variations of one particular disorder or psychopathology. For example, obsessive personality traits and certain forms of depression may be posited to be independent entities which both arise from the same early experiences in a family, but they would not be considered variations of the same disorder. Finally, the *pathoplasty hypothesis* refers to the effect that personality may have on the clinical presentation and course of a particular form of psychopathology. This model does not stipulate any specific causal significance to personality, but simply

implies that variations in personality will affect the psychopathology *after* it emerges. For example, an individual with high degrees of the personality trait of extroversion may behave differently when depressed than an individual with very low degrees of extroversion (i.e., introversion). Here personality ‘colors’ the way the disorder is manifest.

Unfortunately, there is a general lack of empirical data to clarify which of these models best characterizes the complex relationship between various forms of psychopathology and a myriad of personality traits. However, it is likely that specific personality traits will show different types of relationships to different forms of psychopathology and furthermore that a given personality trait may relate to psychopathology in a variety of ways. For example, in the study of neuroticism and depression there is evidence to suggest that neuroticism predisposes an individual to develop depression (predisposition hypothesis), predicts the course of depression (pathoplasty hypothesis), and may be exacerbated by the presence of depression (complication hypothesis). Clearly, the study of the relationship between personality and psychopathology is in its infancy and additional empirical studies are needed.

6. Methodological Considerations

Disentangling the relationship between specific personality traits and various forms of psychopathology will require complex research designs and statistical procedures. Clearly, longitudinal studies of individuals will be necessary to examine which facets of personality increase the risk of which disorders. Such research will also clarify which personality dimensions will reflect complications of specific psychopathologies. Additionally, family history studies and behavior genetic designs employing twin studies or adoption paradigms may help to determine more precisely how various forms of personality and psychopathology are transmitted in families and the extent to which they represent variations of the same process.

7. Summary

The study of personality has revealed that in spite of the power of environmental situations to determine behavior, people tend to be relatively consistent across time and situations in terms of their general patterns of behavior. Furthermore, it seems that people tend to differ along a few primary personality dimensions, or factors, which underlie the concept of personality. While personality traits may increase the risk of developing certain disorders or illnesses, the experience of an illness or disorder may also modify personality. Furthermore, people's personality styles will influence the way in which their disorders develop, change over time, and also respond to treatment. Powerful scientific studies which follow individuals over time will ultimately help us to clarify how personality increases the risk of disorder and furthermore how disorder may modify personality.

See also: Personality Assessment; Personality Disorders; Personality Theories

Bibliography

- Allport G W 1937 *Personality: A Psychological Interpretation*. Holt, New York
- Columbia University Teachers College 1928 *Studies in the Nature of Character, Studies in Deceit*. Macmillan, New York, Vol. 1
- Ekehammer B 1974 Interactionism in personality from a historical perspective. *Psychological Bulletin* **81**: 1026–48
- Epstein S 1979 The stability of behavior: I. On predicting most of the people much of the time. *Journal of Personality and Social Psychology* **37**: 1097–126
- Epstein S, O'Brien E J 1985 The person–situation debate in historical and current perspective. *Psychological Bulletin* **98**: 513–37
- Helmes E, Reddon J R 1993 A perspective on developments in assessing psychopathology: A critical review of the MMPI and MMPI-2. *Psychological Bulletin* **113**: 453–71

- Klein M H, Wonderlich S A, Shea T 1993 Models of relationships between personality and depression: toward a framework for theory and research. In: Klein M H, Kupfer D J, Shea M T (eds.) *Personality and Depression*. Guilford Press, New York pp. 1–54
- Livesley W J, Schroeder M L, Jackson D N, Jang K L 1994 Categorical distinctions in the study of personality disorder: implications for classification. *Journal of Abnormal Psychology* **103**: 6–17
- Lyons M J, Tyrer P, Gunderson J, Tohen M 1997 Heuristic models of comorbidity of axis I and axis II disorders. *Journal of Personality Disorders* **11**: 260–9
- Millon T 1996 *Disorders of Personality: DSM-IV and Beyond*, 2nd edn. Wiley, New York
- Mischel W 1968 *Learning Theory and Personality Dynamics*. Ronald Press, New York
- Pervin L A 1990 A brief history of modern personality theory. In: Pervin L A (ed.) *Handbook of Personality Theory and Research*. Guilford Press, New York, pp. 3–18

S. Wonderlich

Person-centered Psychotherapy

Client-centered therapy is an approach to psychotherapy based on trust in the self-directive capacities of the individual. In this respect, it contrasts with other therapeutic orientations where the therapist characteristically acts as an expert.

1. Historical Overview

Psychologist Carl R. Rogers first described this new approach to psychotherapy in a talk in 1940, not realizing the impact it was going to have. This was followed by a full-length book on psychotherapy containing a clearly stated theory of therapy together with a verbatim account of an eight-interview case. This made up approximately two-fifths of the book and was a pioneering method of case presentation at the time.

As client-centered theory and practice developed in the 1940s and 1950s at Ohio State University and the University of Chicago, a far-reaching body of research on a new hypothesis grew up: that if the therapist offered, and the client experienced, a particular kind of relationship characterized by genuineness, unconditional positive regard, and empathy, a self-directed process of growth would follow. Moving to the University of Wisconsin from 1957 to 1963, Rogers and his associates undertook a major research project which tested the client-centered hypothesis with schizophrenic patients. In 1964, he moved to La Jolla, California, using the approach in the United States and abroad, in small and large groups, school systems, workshops, and conflict resolution. The broader application of the principles of client-centered therapy

became known as the person-centered approach. Rogers died in February 1987. The movement he fathered but did not wish to dominate is carried on by a diverse and dedicated international community.

2. Basic Therapeutic Concepts

The following are the fundamental concepts of person-centered psychotherapy:

(a) An actualizing tendency which is present in every living organism, expressed in human beings as movement toward the realization of an individual's full potential.

(b) A formative tendency of movement toward greater order, complexity, and inter-relatedness that can be observed in stars, crystals, and microorganisms, as well as human beings.

(c) Trust that individuals and groups can set their own goals and monitor their progress toward these goals. Individuals are seen as capable of choosing their therapists and deciding on the frequency and length of therapy. Groups are trusted to develop processes that are right for them and to resolve conflicts within the group.

(d) Trust in the therapist's inner, intuitive self.

(e) The therapist-offered conditions of congruence, unconditional positive regard, and empathy:

(i) Congruence has to do with the correspondence between the thoughts and behavior of the therapist, who is genuine and does not put up a professional front.

(ii) Unconditional positive regard, also identified as 'caring,' 'prizing,' and 'nonpossessive warmth,' is not dependent on specific attributes or behaviors of the client.

(iii) Empathy reflects an attitude of profound interest in the client's world of feelings and meanings, conveying appreciation and understanding of whatever the client wishes to share with the therapist.

(f) Self-concept, locus-of-evaluation, and experiencing are basic constructs which emerge from the client's own interaction with the world.

(i) The self-concept is made up of the person's perceptions and feelings about self. Self-regard or self-esteem is a major component of the self-concept.

(ii) Locus-of-evaluation refers to whether the person's values and standards depend on the judgments and expectations of others, or rely upon his or her own experience.

(iii) Experiencing has to do with whether the person, in interacting with the world, is open and flexible or rigid and guarded.

(g) The Internal Frame of Reference (IFR) is the perceptual field of the individual, the way the world appears, and the meanings attached to experience and feeling. It is the belief of person-centered therapists that the IFR provides the fullest understanding of why people behave as they do, superior to external judgments of behavior, attitudes, and personality.

3. Theory of Psychotherapy

The basic theory of person-centered therapy is that if therapists offer, and clients experience, a particular kind of relationship characterized by genuineness, unconditional positive regard, and empathy, they will respond with greater self-direction and self-expression, an increased openness to inner and outer experiencing, more mature behavior and ability to deal with stress, and a concept of self that is more positive and more realistic.

4. Evolution of Theory and Practice

After Rogers assumed a professorship at Ohio State University at the beginning of 1940, he received an invitation to address Psi Chi, the psychological honor society, at the University of Minnesota, on December 11, 1940, the date often cited as the time when client-centered therapy was born. Rogers described a newer therapy which had the aim of helping individuals not only to solve their present problems but to grow in the capacity to solve future problems as well in a more integrated way, that took advantage of a general drive toward health, growth, and adjustment, that stressed emotional elements more than intellectual aspects, that emphasized the immediate situation rather than the past, and that viewed the therapeutic relationship itself as a growth experience.

The talk generated an intense reaction, both favorable and unfavorable. Rogers expanded his thinking into the book, *Counseling and Psychotherapy* (Rogers 1942), almost two-fifths of which was made up of 'The Case of Herbert Bryan,' consisting of the typescripts of eight phonographically recorded verbatim interviews. This kind of presentation was revolutionary, standing in sharp contrast to the subjective accounts of therapy being published at the time, and provided objective research data for the study of the therapeutic process.

Rogers and his students at Ohio State developed methods of classifying client statements and counselor responses, and of measuring self-regarding attitudes. The concept of self emerged as a central construct of personality organization, and great progress was made in the objective study of personality change and the behavior in therapy of therapist and client.

Research on person-centered psychotherapy advanced in a major way at the University of Chicago Counseling Center during Rogers' tenure there from 1945 to 1957. An entire issue of the *Journal of Consulting Psychology* in 1949 was devoted to a report of the 'parallel studies' project, comprising six investigations of the same group of 10 completely recorded and transcribed cases, with pre- and post-tests. The studies included measures of feelings regarding self and others, the pattern of client content (e.g., statement of problem, understanding or insight, discussion of plans), defensiveness, and maturity of behavior.

There was an evaluation of outcome using the Rorschach test, an analysis of the relationships among all these, and counselor ratings of improvement on a 1 to 9 scale. Two key findings were that (a) the measures applied to each interview provided a meaningful picture of the client's adjustment at the beginning, during, and end of therapy, and (b) there was a significant relationship between counselor estimates of success and ratings based on the interview-analysis measures. Significant relationships were not found between Rorschach results and those of the five interview measures. It was also concluded that the research methodology was applicable to therapeutic approaches generally, and that such efforts would help put psychotherapy on a scientific basis.

The 'parallel studies' project was succeeded five years later by a larger study (Rogers and Dymond 1954), which contained complete data on 29 clients seen by 16 therapists, as well as on a matched control group. It measured changes in self using the Q-technique of British psychologist William Stephenson, a sophisticated method of quantifying the way people viewed themselves at present and ideally, leading to the findings that the self-concept of clients in therapy improved significantly, and to a significantly greater degree than control group subjects, and that therapy achieved a significant increase in congruence between self and ideal. Qualitative analysis showed clients in therapy growing in their feelings of self-confidence, self-reliance, self-understanding, inner comfort, and comfortable relationships with others, with a decrease in negative feelings about themselves.

Person-centered and general psychotherapeutic theory advanced substantially with the introduction of the concept of 'necessary and sufficient conditions of therapeutic personality change' (Rogers 1957). Empathy, congruence, and unconditional positive regard were clearly defined. Capable of quantification, they stimulated hundreds of research projects. The Relationship Inventory (Barrett-Lennard 1998), an instrument measuring these conditions, has also been used in a vast number of research projects in psychotherapy and other human relations applications such as parent-child, student-teacher, and worker-employer relationships (see *Therapist-Patient Relationship*).

The most rigorous exposition of person-centered thinking, 'A theory of therapy, personality, and interpersonal relationships,' was published in Sigmund Koch's *Psychology: A Study of a Science* (Rogers 1959). A pioneering research project on therapy with schizophrenic patients and normal adults at the University of Wisconsin broke new ground in working with and doing research on this population (Rogers et al. 1967).

On Becoming a Person (Rogers 1961) included a summary of the personality and behavioral changes in therapy supported by research, an objective description of the fully-functioning person, the empirical

evidence for the conditions which facilitate psychological growth, and a description of the author's struggle to reconcile the role of scientific investigator with that of therapist operating at the height of personal subjectivity. A unique, decades-long, research effort in person-centered psychotherapy which supports Rogers' description of the fully-functioning person focuses on the understanding of personality integration, which involves physiological, perceptual, cognitive, and interpersonal subsystems (Seeman 1983). Seeman has gone outside of the usual person-centered sources, drawing on the work of analysts Erik Erikson and Heinz Hartmann, and ego development specialist Jane Loevinger, among others. His conclusions are similar to Rogers, that high-functioning persons are healthier, more efficient in their perception of reality, have superior environmental contact, high self-esteem, confidence and trust in themselves, and possess a sense of autonomy that facilitates the development of caring and generative relationships.

5. Evaluation

Orlinsky and Howard (1978) concluded that the results of relevant studies were too variable to support the Rogerian hypothesis. But they were impressed by the evidence of 13 outcome studies of patients' perceptions of therapist attributes such as nonpossessive warmth, positive regard, and acceptance; all 13 investigations yielded a significant positive correlation between such perceptions and good therapeutic outcome.

Additionally, in 1978, Orlinsky and Howard reviewed 15 studies of the relationship between outcome and patients' perceptions of their therapists as empathically understanding, and found such an association preponderantly positive, with only two or three presenting evidence to the contrary. Fourteen of 20 investigations yielded a significant positive association between therapist self-congruence and outcome, with six showing null and marginally mixed results. Nine of 10 studies focusing on the 'process' or 'experiencing' levels conceptualized by Rogers in 1957 found significant positive correlations between experiencing and good therapeutic outcome.

Orlinsky and Howard (1986) came to these conclusions about the relation of outcome in psychotherapy to aspects of therapeutic process which bear upon the conditions offered by person-centered therapists. (a) Role-investment, empathic resonance, mutual affirmation, and the overall quality of the relationship were consistently related to patient outcome in 50-80 percent of the large number of findings surveyed in this area. (b) Therapist genuineness, particularly as perceived by patients, was often but not consistently associated with better outcomes. (c) Patients' perceptions of therapist empathy were very consistently related to good outcome. (d) Therapists' warmth or acceptance toward their patients, especially

but not only as observed by patients, was quite consistently related to good outcome in therapy.

Lambert et al. (1986), analyzing scores of studies on therapy outcome, concluded that therapist personal factors such as trust, warmth, acceptance, and wisdom were crucial ingredients, even in the more technical therapies. Similar conclusions were reached by Patterson (1984) and more recently in an examination of psychotherapy outcome research studies by Bozarth (1998) with these conclusions: successful psychotherapy depends primarily on the therapist–client relationship and the client’s internal and external resources; the type of therapy is largely unrelated to outcome; the training and experience of therapists are irrelevant to successful therapy; clients who receive therapy improve more than those who do not; the evidence is weak that there are specific treatments for particular disabilities; empathy, genuineness, and unconditional positive regard are the relationship variables that correlate most consistently with effective therapy.

Watson (1984), following a careful review of the relevant research, concludes that because of methodological inadequacies, there is insufficient evidence either to support or refute the effectiveness of the ‘necessary and sufficient conditions.’ (see *Psychological Treatment, Effectiveness of*).

6. Concluding Comments

Person-centered psychotherapy has been evaluated assiduously, internally and externally, since its formulation by Carl Rogers in 1940. Its concepts and methods continue to be controversial in a field that is dominated by orientations that advocate guidance by experts. Its uniqueness in the depth of belief in the self-directive capacity of individuals and groups assures its future as a vital alternative. It continues to thrive internationally as part of a broader person-centered approach with implications for group process, education, and conflict resolution. Contributing to the health of the movement is the diversity of interpretation and implementation of Rogerian principles; it includes ‘purists’ and those who believe that a directive dimension is sometimes required. Different organizations and conferences afford the opportunity to explore these differences.

See also: Clinical Psychology in Europe, History of; Clinical Psychology in North America, History of; Interpersonal Psychotherapy; Multimodal Therapy in Clinical Psychology; Psychological Treatment, Effectiveness of; Psychological Treatments, Empirically Supported; Psychotherapy, History of; Psychiatric Aspects; Psychotherapy Integration; Therapist–Patient Relationship

Bibliography

- Barrett-Lennard G T 1998 *Carl Rogers' Helping System*. Sage Publications, London
- Bozarth J 1998 *Person-centered Therapy: A Revolutionary Paradigm*. PCCS Books, Ross-on-Wye, UK
- Kirschenbaum H, Henderson V L 1989 *The Carl Rogers Reader*. Houghton Mifflin, Boston
- Lambert M L, Shapiro D A, Bergin A E 1986 The effectiveness of psychotherapy. In: Garfield S L, Bergin A E (eds.) *Handbook of Psychotherapy and Behavior Change*, 3rd edn. Wiley, New York
- Orlinsky D E, Howard K L 1978 The relation of process to outcome in psychotherapy. In: Garfield S L, Bergin A E (eds.) *Handbook of Psychotherapy and Behavior Change*, 2nd edn. Wiley, New York
- Orlinsky D E, Howard K L 1986 Process and outcome in psychotherapy. In: Garfield S L, Bergin A E (eds.) *Handbook of Psychotherapy and Behavior Change*, 3rd edn. Wiley, New York
- Patterson C H 1984 Empathy, warmth, and genuineness in psychotherapy: A review of reviews. *Psychotherapy* **21**: 431–8
- Rogers C R 1942 *Counseling and Psychotherapy*. Houghton Mifflin, Boston
- Rogers C R 1957 The necessary and sufficient conditions of therapeutic personality change. *Journal of Consulting Psychology* **21**(2): 95–103
- Rogers C R 1959 A theory of therapy, personality, and interpersonal relationships, as developed in the client-centered framework. In: Koch S (ed.) *Psychology. A Study of Science. Vol. III: Formulations of the Person and the Social Context*. McGraw-Hill, New York, pp. 184–256
- Rogers C R 1961 *On Becoming a Person*. Houghton Mifflin, Boston
- Rogers C R 1980 *A Way of Being*. Houghton Mifflin, Boston
- Rogers C R, Dymond R F (eds.) 1954 *Psychotherapy and Personality Change*. University of Chicago Press, Chicago
- Rogers C R, Gendlin E T, Kiesler D J, Truax C B (eds.) 1967 *The Therapeutic Relationship and its Impact: A Study of Psychotherapy with Schizophrenics*. University of Wisconsin Press, Madison, WI
- Seeman J 1983 *Personality Integration*. Human Sciences Press, New York
- Watson N 1984 The empirical status of Rogers’ hypotheses of the necessary and sufficient conditions for effective psychotherapy. In: Levant R E, Shlien J M (eds.) *Client-centered Therapy and the Person-centered Approach*. Praeger, New York

N. J. Raskin

Person-centered Research

Person-centered research here means research that focuses on the person as a functioning totality within the domain studied. The person then becomes the main conceptual unit and also often the main analytical unit. In many methodological realizations of this approach, individual patterns of values in the

variables under study become the main analytical units and subjected to, for instance, classification analysis or other types of pattern analyses. Person-centered research can be contrasted to variable-centered research where the focus is on the variable as the main conceptual and analytical unit. Variable-centered research is far more common but is not the topic of this article.

Person-centered research does not need to be quantitative and can in certain situations be carried out by case-oriented research and by using a qualitative approach. Obviously, the study of the single individual, studied quantitatively using the p-technique, is in one way person-centered, but is not normally so according to the definition given above (since the focus is usually then on relationships between variables within the individual). P-technique and time-series analysis is also discussed elsewhere in the encyclopedia. The perspective given in this article is of carrying out quantitative person-centered research on a sample of persons aiming at explaining and understanding inter-individual and/or intra-individual differences.

A very short historical overview is given of the emergence of person-centered research, followed by a presentation of the theoretical foundation of the modern person-approach. Finally an overview of common methods for carrying out person-centered research is presented.

1. The Emergence of Person-centered Research

A major type of person-centered research is the typological approach by which individuals are categorized into different types. It has ancient roots and is already reflected in the classical categorization of individuals into four basic temperaments: sanguine, phlegmatic, melancholic, and choleric. The typological view can be regarded as a reflection of a view that people are functioning wholes and that there are only a limited number of typical ways in which a functioning system can be organized. From this perspective, finding and understanding these types are important scientific goals. Creating typologies is also a reflection of a basic tendency in man of categorizing encountered phenomena.

The typological approach has been, and is still, stronger in the life sciences (especially taxonomy in biology and diagnosis in medicine) than it is in the social and behavioral sciences (for an historical overview, see Misiak and Sexton 1966). Gangestad and Snyder (1985) discussed the importance of a categorical approach in personality research and pointed to the emergence of types, each sharing a common influence (like an infectious disease caused by a specific germ). The typologies presented by Jung, Kretschmer, Sheldon, and others are still of interest today and psychiatric diagnoses in the Kraepelin tradition are

influential in clinical practice. There is also a concern with typological research in sociology, including, for instance, the search for 'ideal types' (Bailey 1994).

The word 'typology' has many meanings—see Cattell 1957 for an overview of 45 different meanings of the word. It has led to much confusion and even to resentment. This resentment was often directed against the connotation of 'innate' implied in many earlier typologies and against their subjectivity. Waller and Meehl (1998) make a number of thoughtful distinctions with regard to type and typology, preferring the use of the terms taxon for, roughly speaking, a meaningful type, and taxometrics as a generic name for (their) procedures for finding taxa. It should be pointed out that neither of these limitations apply to modern typological research within the person-approach described.

One of the first proponents of a comparatively modern person-centered research strategy was William Stern who already in the beginning of the twentieth century discussed an approach in which individual patterns in many traits were the units of analysis. Other early proponents of the importance of considering persons as functioning wholes were Kurt Lewin and Gordon Allport. A systematic approach to the person-centered study of personality development was undertaken by Jack Block based on longitudinal data. He presented an empirically based typology of longitudinal personality types and was, as far as we know, the first one to use the term 'person-approach' (Block 1971). Examples of research looking for basic categories are given by the work of Lars Bergman and David Magnusson studying stability and change in patterns of extrinsic adjustment problems, by Sigrid Gustafson studying a psychopathy-linked pattern called aberrant self-promotion, by Lea Pulkkinen, studying personality styles in a developmental perspective, by Richard Robins, Oliver Johan and Avshalom Caspi, searching for personality types, and by Ed Seidman relating typical experiential neighborhood profiles to antisocial behavior.

Perhaps the most forceful modern development of person-centered research has taken place within the new developmental science (Cairns et al. 1996). There, a holistic-interactionistic paradigm has grown strong and a new type of person-centered research has emerged (Magnusson 1999). It is focused on the study of individual development and has been called 'person-approach' (for an overview, see Magnusson 1998). Its theoretical and research strategic fundaments are described in Sect. 3.

2. Theoretical and Research Strategic Fundaments of the Person-approach

The person-approach refers to a holistic view on individual functioning and development. Magnusson and Allen (1983) summarized the essence of a person-

approach in the following way: 'The person oriented approach to research (in contrast to the variable centered approach) takes a holistic and dynamic view; the person is conceptualized as an integrated totality rather than as a summation of variables' (p. 372). A basic proposition of a holistic perspective, with consequences for the application of a person-approach, is that individual functioning and development can be described in terms of dynamic, complex, and adaptive processes in which mental, biological, and behavioral factors in the individual, and social, cultural, and physical factors in the environment are involved.

The person-approach is equally applicable to research on the current functioning of individuals and to research on individual development. Thus, it has implications for research in any field of psychological inquiry. Hitherto it has mainly been used in developmental and personality research.

For a long time the holistic view was considered to be too vague and lacking of specific content to serve as the basis for strong theories for understanding and explaining individual functioning and development. However, since 1970 the holistic model has been filled with content to the extent that it now forms a scientific basis for planning, implementation, and interpretation of empirical studies on specific aspects of individual functioning and development.

Substantively, contributions come from psychological research on mental factors and behavior and from biological and medical research on the functioning of the brain and the role of physiological factors in the total functioning and development of individuals. New findings in these areas have helped to enrich the 'black box' in S-R models with substantive contents and have helped in closing the gap between different explanations of behavior in terms of mental, biological, and environmental factors.

Contributions to the effectiveness of a holistic model as a theoretical framework for empirical research on specific phenomena, also derive from the presentation of modern models for dynamic, complex processes, particularly the general systems theory. Three aspects of these models are important for the discussion here of a person-approach.

(a) From a holistic perspective, mental, behavioral and biological aspects of individual functioning, and social, cultural, and physical factors of the environment are incorporated into one integrated theoretical framework. Thus, the role of a single variable cannot be finally investigated and understood in isolation from its context.

(b) A central principle in dynamic, complex processes is the principle of dynamic interaction (continuously ongoing reciprocal influences) as contrasted to statistical interactions in data. The new models for dynamic, complex processes provide a theoretical framework for investigating and understanding the dynamic processes of interaction of operating factors within the individual, and the continuous reciprocal

interaction between the individual and the environment in the person-environment system.

(c) For the relations among operating factors at the level of the individual, mutual dependencies characterized by dynamic interactions and non-linearities can be a characteristic feature. The same holds true for the interaction of a single individual with his/her environment. For instance, individuals' psychological and physiological stress reactions to increasing stimulation from the environment are often nonlinear. In fact, most common variable-oriented research methods are not suited for handling nonlinear relations and dynamic interactions. This is exemplified by the fact that many such methods use the correlation matrix as the data to be analyzed, a matrix which mainly reflects linear relations, not nonlinear relations and interactions.

The process of organization of current and developmental processes which takes its start at conception is guided by the principle of self-organization. Self-organizing ability is a characteristic of open systems and refers to a process by which structures emerge without 'prescriptions from the outside.' Within subsystems, the operating components organize themselves to maximize the functioning of that subsystem with respect to its purpose in the total system. At a higher level subsystems organize themselves in order to fulfill their role in the functioning of the totality. Within a specific system, say the cardiovascular system, each of the operating factors (e.g., systolic blood pressure, diastolic blood pressure, and heart rate) do not function and develop independently of the others. The specific role of each operating factor is determined by the role it plays in the system. The operating factors are organized and function in terms of functional configurations, in what will be referred to in the following text as patterns. From this perspective, the important individual differences are not to be found in differences in any single variable taken out of its context of other, simultaneously operating variables. Instead, they are to be found in differences in the patterning of operating variables in the system under investigation. This applies to all levels of organization.

For the discussion of the implications of the person-approach for an effective measurement model and methodological approach, two aspects of the organization of mental, biological, and behavioral structures and processes are fundamental:

(a) Within subsystems individuals differ to some extent in the way in which operational factors are organized and function.

(b) Only a limited number of states are functional for each subsystem and for the totality.

The number of ways is restricted in which operating factors in a certain subsystem can be organized in patterns, in order to allow the subsystem to play its functional role in the totality, and the number of ways is also restricted in which subsystems can be organized

to form the total pattern for the total organism (cf. Bergman and Magnusson 1997). This view implies, among other things, that the studied complex systems have inherent restrictions which lead to (a) certain states being frequent or typical and (b) other states not occurring. The first aspect has been much studied and provides a motivation for the search for types. The second aspect has been much less studied but can also be of importance: What does not occur but, in principle, could occur contains information about how the system under study functions (cf. the concept of 'white spots' as discussed by Bergman and Magnusson 1997).

As emphasized above, the person-approach is a theoretical perspective. As such it forms the theoretical framework for the planning, implementation, and interpretation of specific substantive issues. This implies, among other things, that it has to be distinguished from the methods applied for treatment of data within the perspective of a person-approach. However, in most cases, pattern-based methods emerge as the most natural method choices. A brief overview of a selection of such methods is given in Sect. 4.

3. A Selection of Methods for Carrying out Person-centered Research

3.1 Classification and Cluster Analysis

It was mentioned in the introductory section that classification is a basic approach for carrying out person-centered research. Before discussing different types of classification analysis a few basic properties of such an analysis should be mentioned. In most forms of classification analysis the basic data are contained in the similarity or dissimilarity matrix between all pairs of subjects. Similar subjects are then grouped together in the classification analysis. The (dis)similarity between each pair of subjects could be a subjective rating but is most commonly calculated according to a formula that takes into account the (dis)similarity of the profiles of values in the variables for the two subjects. One common formula that takes into account differences in both level and form of the value profiles is the averaged squared Euclidean distance. It should be pointed out that the choice of (dis)similarity coefficient usually has important consequences for the results obtained in the classification analysis and that this choice should always be made on the basis of considerations in the specific case. The measurement characteristics of the different variables must be considered since the results of many methods are not invariant under linear transformations of the involved variables.

A classification can be achieved in different ways:

(a) It can be strictly theoretically derived as in the construction of ideal types.

(b) It can be totally empirically driven as in cluster analysis where subjects are classified together in clusters on the basis of their profile similarity. This type of classification approach is the most common one and is described.

(c) Some kind of model-based quantitative approach can also be used.

Examples of (c) are latent structure analysis and latent class analysis, originally proposed by Paul Lazarsfeld, where a number of latent classes is assumed and then the fit to the empirical data is tested and parameters are estimated (Goodman 1974). If the model holds, all relationships between variables within a latent class should disappear (the assumption of local independence).

3.1.1 Cluster analysis. In cluster analysis, a large number of methods are available for classifying objects on the basis of their (dis)similarities. Major types of cluster analysis are hierarchical methods (agglomerative or divisive), partitioning methods, and methods that allow overlapping clusters. Within each type of methods a variety of specific methods and algorithms exist. Perhaps the most common form of analysis is the agglomerative hierarchical cluster analysis. This group of methods starts with each of the n subjects being its own cluster. In Step 1 the two most similar subjects are joined to form one cluster giving in all $n-1$ clusters. In Step 2 the two most similar clusters are joined to form one cluster, giving in all $n-2$ clusters. The process is repeated until every subject is in one and the same cluster that occurs at Step $n-1$. The result is a hierarchical classification tree. Often the researcher concentrates on one cluster solution with a fairly small number of clusters and each cluster is described by its profile of means in the variables studied (called the cluster centroid). All members in that cluster should have profiles similar to the centroid. Different specific methods of hierarchical agglomerative cluster analysis have different rules for how to decide which two clusters are most similar. For instance, in the single linkage (= nearest neighbor) method the similarity between two clusters is given by the (dis)similarity of the two subjects, one from each of the two clusters, that are most similar. What method of cluster analysis is most appropriate of course, depends on the specific situation. Evaluations of the sensitivity of different clustering algorithms to the effects of errors of measurement and of the ability to recover a known cluster structure indicate that, as expected, no method appears to be generally superior to the others. Methods that often see to perform well include Ward's minimum variance method and average linkage cluster analysis (two hierarchical methods), and k-means relocation analysis based on a reasonable start classification (Morey et al. 1983). More recently, methods based on so called Beta-

flexible clustering have been suggested. For an overview of methods for cluster analysis see Gordon 1981.

3.1.2 Should all subjects be classified? It has long been recognized that multivariate outliers may disturb the results of cluster analysis and it has been suggested that in some situations the coverage must be less than 100%, i.e., not everybody can be classified. Bergman (1988) drew attention to the fact that there often exist a small number of 'unique' individuals, not similar to any other subjects that should not be forced into a cluster. He presented a procedure, RESIDAN, for *a priori* identifying and analyzing separately what he calls a residue of unclassified subjects. In a subsequent residue analysis, rare or non-existent patterns are studied which may be of theoretical significance (cf. the discussion about 'white spots' in Sect. 3).

3.1.3 Classification in the study of individual development. Important and intriguing methodological problems for person-centered research occur when studying individual development. A variety of approaches are available.

(a) Directly analyzing longitudinal patterns of values in variable profiles have often proved disappointing. If only profile form is studied, a direct longitudinal classification might work better than otherwise, since the amount of information that has to be summarized by the cluster membership variable is drastically reduced. An example of the usefulness of a longitudinal classification strategy in this situation is given in Block's (1971) study of longitudinal personality types.

(b) The analysis of cross-sectional patterns followed by linking over time is simpler and also often more robust than most other methods. One standard method of this type is called Linking of Clusters after Removal of a Residue (LICUR, Bergman 1998). LICUR is suitable in situations where both form and level of the profile is considered to be of importance. It includes procedures for removing a residue before the cluster analyses that are performed at each age separately and procedures for deciding the number of clusters. The results of the cluster analyses are linked between adjoining ages by cross-tabulating the classifications obtained at the two ages and it is tested for over frequented and under frequented cells (i.e., cluster membership combinations occurring more or less frequently than expected by chance). LICUR is applicable in a large number of settings, for instance when different variables are measured at the different measurement occasions and during periods of dramatic developmental shifts. Model-based alternatives to LICUR are provided by longitudinal extensions of latent class analysis, for instance latent transition analysis developed by Collins and Wugalter (1992).

3.2 Some other Approaches for Person-centered Analyses

3.2.1 Identifying a set of typical classes. Sometimes the interest is not in a complete classification but rather in identifying homogenous subgroups of subjects ('dense points' in a multivariate space) believed to indicate system states that are in some way 'important.' It is then believed that many subjects have entered a stable 'functioning' system state with a characteristic value pattern and that there are only a limited number of such states. Frequent typical patterns are regarded as indicators of such optimal states. One may also be interested in recognizing patterns that *a priori* are considered as 'important.' For overviews of various techniques relating to these issues see Young and Fu (1986).

3.2.2 Studying all possible patterns using configural frequency analysis. Configural Frequency Analysis (CFA) is a set of methods for studying all possible value patterns for a set of studied variables. The involved variables have to be discrete and are often dichotomized to make the number of value patterns to be examined manageable. CFA was originally suggested by Gustav Lienert. Lienert and his co-workers have developed CFA in a number of ways and in Germany it has become a research tradition. For a basic introduction see Krauth and Lienert (1982). A more recent overview is given by von Eye (1990) who also has contributed to various newer developments of CFA. The idea of types in a psychiatric setting was elaborated by Joachim Krauth who pointed to the importance of identifying combinations of variable values that emerge above chance level. His line of reasoning relates to the discussion about optimal systems states and designs as being more frequent than other states and designs. For further information on CFA, see von Eye (1990).

3.2.3 Analyzing specific properties of patterns. Instead of considering the complete value patterns in a pattern analysis, specific aspects of the patterns can be highlighted. Two such analyses are the following:

(a) In some cases the maximum and minimum scores of the subject's profile of scores are at focus. The variable taking the maximum value and the variable taking the minimum value may be seen as the essential features of the profile with the other scores providing the background. Of course, this implies that the different variables are scaled in a way that allows for comparisons between variables.

(b) The degree to which a subject's profile is even or uneven can be studied in different ways and an obvious measure of profile scatter is its variance around the mean of the scores in all variables constituting the profile.

3.2.4 Abstracting group memberships from qualitative information. A good example of this approach is the work done by Singer et al. in linking life histories to mental health outcomes. Richly detailed descriptions of individual lives form the input data for a procedure for discerning generalizable features of aggregates of multiple lives.

4. Some Final Comments

We believe that the theoretical perspective included in the person-approach presented above provides person-centered research with a useful meta-theoretical framework. But to reiterate: a person-approach should not be confused with a methodological approach using pattern analysis although such methods often are natural within a holistic orientation.

Sometimes reality may not be continuous but may rather operate to produce more or less discrete types. Or put differently: only certain configurations of system states are in some way optimal and become stable and often observed. A discussion of this issue was given in Sect. 3. It can then be argued that in person-centered research, methods for pattern-based analysis are often more useful than standard variable-based methods. Many common variable-based methods do not handle interactions well and in many situations pattern-based methods are more naturally used for this purpose. In fact, taking interactions seriously, such an approach, focusing on variables, tends to become very complicated. Lee Cronbach even made the metaphor of entering a hall of mirrors when one pursues such a goal.

It has sometimes been claimed that results from many methods used in person-centered research, for instance cluster analysis, are untrustworthy. In order to evaluate the validity of results from such studies two aspects should be discussed separately: (a) the technical aspects of the methods applied and (b) the appropriateness of the application of these methods in different settings. There is nothing wrong technically with any of the major methods used to carry out person-centered analysis. The problem arises when a method is inappropriately applied. Inconsistent results are often caused by the use of a clustering procedure that does not match the combined requirements of the problem under study and the data available. Important considerations for obtaining useful results from a cluster analysis are: (a) That the variable profile under study adequately summarizes the information *Gestalt* of interest; (b) that the values in the different variables are comparable and an appropriate (dis-)similarity coefficient has been chosen; (c) that a sound clustering algorithm has been chosen; (d) that only variables are included in the value profile which have a reasonably high reliability and, if finding hom-

ogenous clusters is the focus of interest, the profile is constituted by only a limited number of variables; and (e) that in many cases not all subjects should be classified. A number of validation procedures are available for studying the results of a classification analysis.

Two issues that are sometimes confused in person-centered research are the question about identifying generic classes ('natural clusters') and the question about ascribing the subjects in the sample to the appropriate class. It is a general experience that the second purpose tends to be more difficult to achieve. If the purpose is to identify typical value profiles that frequently emerge in different settings, one way of validating these typical profiles is to compare centroids between different samples/split-halves. Those that replicate might be regarded as types. These typical profiles need not, of course, together comprise a total classificatory system that encompasses everybody in a specific sample. On the contrary, the usefulness of a partial typology should be stressed since it can often be a more realistic goal. For instance, methods for determining how well the (dis)similarity matrix is represented by the classification solution are sometimes erroneously used for evaluating to what extent the first purpose has been achieved. Such methods are mainly relevant when the study of individual class membership is at focus.

Applying variable-oriented methods it is often assumed that (the same) linear relations approximately hold over individuals and that the essential features of a multivariate data set are captured by, for instance, the correlation matrix. This enables the researcher to use modern powerful statistical methods to construct models of data that are testable within the confinements of these basic assumptions. However, as was pointed out in Sect. 3, in settings when these assumptions cannot be assumed to hold a person-centered approach emerges as an alternative. It is natural that the recognition of multivariate complexity and higher-order interactions that follows with this perspective also makes it extremely difficult to formulate a coherent testable model of the data. Of course, this does not mean that person-centered research needs to be explorative in its methods. Theoretical considerations will lead to expectations about, for instance, types and antitypes and about typical developmental streams which are testable using various methods.

Finally, during the last decade person-centered research has received increased attention. For this we believe there are sound motives, as explicated in the section about the person-approach. In some situations one can trace a disappointment in the meager outcome in the understanding of a studied process given by even sophisticated variable-oriented methods. Applied with good judgment, person-centered research may then offer a deeper insight in how a system works. This applies to systems at the level of the individual as well as at other levels.

See also: Classification: Conceptions in the Social Sciences; Configurational Analysis; Single-subject Designs: Methodology; Statistical Clustering; Time Series: General

Bibliography

- Bailey K D 1994 *Typologies and Taxonomies*. Sage, New York
- Bergman L R 1988 You can't classify all of the people all of the time. *Multivariate Behavioral Research* 23: 425–41
- Bergman L R 1998 A pattern-oriented approach to studying individual development: Snapshots and processes. In: Cairns R B, Bergman L R, Kagan J (eds.) *Methods and Models for Studying the Individual*. Sage, Thousand Oaks, CA, pp. 83–121
- Bergman L R, Magnusson D 1997 A person-oriented approach in research on developmental psychopathology. *Development and Psychopathology* 9: 291–319
- Block J 1971 *Lives Through Time*. Bancroft Books, Berkeley, CA
- Cairns R B, Elder G H Jr, Costello E J 1996 *Developmental Science*. Cambridge University Press, Cambridge
- Cattell R B 1957 *Personality and Motivation Structure and Measurement*. World Book, New York
- Collins L M, Wugalter S E 1992 Latent class models for stage-sequential dynamic latent variables. *Multivariate Behavioral Research* 27: 131–57
- Eye A von 1990 *Introduction to Configural Frequency Analysis. The Search for Types and Antitypes in Cross-Classifications*. Cambridge University Press, Cambridge, UK
- Gangestad S, Snyder M 1985 To carve nature at its joints: On the existence of discrete classes in personality. *Psychological Review* 92: 317–49
- Goodman L A 1974 Exploratory latent structure analysis using both identifiable and unidentifiable models. *Biometrika* 61: 215–31
- Gordon A D 1981 *Classification: Methods for the Exploratory Analysis of Multivariate Data*. Chapman and Hall, London
- Krauth J, Lienert G A 1982 Fundamentals and modifications of configural frequency analysis (CFA). *Interdisciplinaria* 3, Issue 1
- Magnusson D 1998 The logic and implications of a person approach. In: Cairns R B, Bergman L R, Kagan J (eds.) *Methods and Models for Studying the Individual*. Sage, Thousand Oaks, CA, pp. 33–63
- Magnusson D 1999 Holistic interactionism—a perspective for research on personality development. In: Pervin L A, John O P (eds.) *Handbook of Personality*, 2nd edn. Guilford Press, New York, pp. 219–47
- Magnusson D, Allen V L 1983 Implications and applications of an interactional perspective for human development. In: Magnusson D, Allen V L (eds.) *Human Development: An Interactional Perspective*. Academic Press, New York, pp. 369–87
- Misiak H, Sexton V 1966 *History of Psychology*. Grune & Stratton, New York
- Morey L C, Blashfield R K, Skinner H A 1983 A comparison of cluster analysis techniques within a sequential validation framework. *Multivariate Behavioral Research* 18: 309–29
- Waller N G, Meehl P E 1998 *Multivariate Taxometric Procedures. Distinguishing Types from Continua*. Sage, Thousand Oaks, CA
- Young T Y, Fu K S 1986 *Handbook of Pattern Recognition and Image Processing*. Academic Press, Orlando, FL

L. R. Bergman and D. Magnusson

Personhood, Anthropology of

'The Anthropology of Personhood' compasses the definition and study of three conceptual terms: 'person,' 'self,' and 'individual.' It is the exploration of the nature of the identity of the individual actor and the relationship between that identity and the symbolic forms and material practices of different sociocultural milieux.

On neither the meaning nor the implication of the above conceptual terms, however, can significant disciplinary consensus be found; a number of writers have addressed definitional issues directly (Harris 1989), but large disagreements remain. The dissension is perhaps understandable when one considers that what is fundamentally at issue is the nature of individual consciousness, and its manifestation—emergence, development, and construal—in different times and places. What is the relationship between 'nature' and 'nurture' in this context: between consciousness as a biological given and as sociocultural construct; between consciousness as privately experienced and as publicly validated; between consciousness as uniquely embodied and as collectively elicited and shared?

The position taken in this article is that an anthropological study of personhood must be careful to distinguish between what may be termed 'individualism' and 'individuality': roughly, between the conventional conceptualization of individual actors in particular settings on the one hand and their personal experience on the other. Much of the conceptual difficulty which the discipline has experienced arises from a conflating of these two, a confusion which goes back to Durkheim.

1. Social Structure and System

1.1 The Durkheimian Individual

Durkheim conceived of human beings as *homo duplex*; on one side there was the biological and personal (comprising the individual body with its material senses and appetites) and on the other the social and moral (the conceptual and conscientious). The individual thus led a double existence: one rooted in the physical organism and one (morally, intellectually, spiritually superior) in a social organism of collectively and uniformly held ideas and practices. Between the two there was ongoing antagonism and tension, but through inculcation into a public language and culture, humankind was capable of rising above mean (animal) 'individuality' and becoming part of a collective conscience in which the (sacred) traditions of a society were enshrined. If individuals were indeed conscious of themselves as individuals, then this was

equally a product of their socialization in a collective conscience; 'individualism' was a social product like all moralities and all religions.

From Durkheim's (structuralist) descendants a collectivist narrative was elaborated which conceptually subsumed the individual actor within sociocultural contexts. Mauss (1985) took it upon himself to detail how experience was constituted by cultural categories, and thus how society exerted its force on the physiological individual. He outlined a purported evolution in individual consciousness which could be tied to the particular forms of social structuration. First, comes the tribal stage of *personnage*, where individuals are conceived of as ephemeral bearers of a fixed stock of names, roles, and souls in clan possession; having no existence independently of the clan, individuals possess no inner conscience. Next comes the Classical stage of *persona* where individuals are conceived of as independent and autonomous citizens of a state; they are responsible, legal persons but still they possess no inner life or identity beyond the civic. With the rise of Christianity comes the stage of *personne*; conceived of as indivisible and rational, possessing a conscience, indeed, a unique sacred soul, the individual now serves as the foundation of all political, economic, and legal institutions. Finally comes the peculiar Western stage of *moi*: the individual as a 'self,' with self-interestedness and self-knowledge, as validated by modern institutions of psychological science.

Beginning with the assumption that sociocultural holism represents the paradigmatic form of human consciousness and experience, Dumont (1986) set out to plot more precisely the origination and progression of the singular Western idea of the autonomous individual as bearer of supreme moral value by comparing it with the Oriental 'archetype.' The Christian *personne*, he suggests, is reminiscent of the Hindu figure of the 'world-renouncer,' someone who seeks ultimate truth by forgoing the world in favour of his own individual spiritual progress. In the world-renouncer one finds a Western-like individualist for whom society has lost its absolute reality, and who, despite the constraining interdependence Indian society ubiquitously imposes on its members, has become self-conscious and self-sufficient. The evolution of the Western individual and the 'substitution' of self for society as absolute value marks the triumph of a (Christian) religious ideology in which worldly renunciation becomes the collective norm. Christ's teaching that man possesses a soul which absolutely transcends the world of social institutions and powers (which is absolutely individual in relation to God and meets others' only in God) engenders a community of 'outworldly' individuals who meet on earth but have their hearts in heaven. Nonetheless, Dumont concludes, the evolution need not end there; in the wake of such movements as multiculturalism, nationalism, fascism, and Islamic fundamentalism, the cultural

future of individualism is, to say the least, unpredictable.

Characterizing the above, 'Durkheimian,' line of thought, in short, is the idea that the individual actor of Western society is the result of a recent and exceptional historico-cultural development. The concept of the 'individual' and its moral and social significance is absent elsewhere, with no *a priori* differentiation being made between individual and role, between self and society. Learned in this paradigm, moreover, it is not surprising that in much anthropological reportage on the person one finds a denial of individualism and little discussion of individuality. Inasmuch as 'the individual' exists, it is as a particular manifestation of 'the person'; it is not individuals who are seen to be diverse so much as the working parts of the complex social systems of which they are components and conduits. Individuals, in short, become collective constructs within the contexts of specific cultural ideologies and social practices, their psyches emanations of certain pre-given and prestructured life-worlds of socialization.

For instance, Strathern (1988) describes Melanesian personhood not in terms of individuals—distinct actors with discrete emotions, awareness and agency—but 'dividuals': beings constituted by properties, goods, and substances as these are ongoingly exchanged with others. The Gahuku-Gama, Read (1955) elaborates, are conscious of themselves only as members of common categories of relations; rather than, say, 'friendships' between unique individuals there are only relationships between socially defined positions. Similarly, Myers (1986) records Australian Aboriginal 'emotions' not as pertaining to deep-psychological experiences of the individual but as cultural evaluations of people and circumstance which persons must make in order to live in fluxional social settings. Geertz (1973, pp. 360), finally, relativizes Western conceptions of the autonomous individual by emphasizing the collective role-playing fundamental to the Balinese *habitus* of the person. Culture, Geertz suggests, translates into symbolic 'control mechanisms' by which the breadth and indeterminateness of actors' potential social lives are programmatically reduced to the specificity and narrowness of their actual ones.

In sum, a Durkheim-inspired anthropology of personhood has eventuated in comparative enquiries into the diversity of sociocultural conceptions of the human individual, how these conceptions connect with other sociocultural 'institutions' (family, law, exchange), and how these different ways of conceptualizing determine differences in experience. 'Personhood,' on this view, amounts to 'those culturally constituted and socially conferred attributes, capacities and signs that mark a moral career and its jural entitlements in a particular society' (Poole 1998, p. 842). 'Selfhood' refers to a human being as a locus of experience, as this experience is culturally allocated

and defined—usually as relational, socially embedded, and sociocentric. ‘Individual’ compasses an actor’s uniqueness to the extent that this latter is realized in certain collectively structured life-worlds.

1.2 The Non-Durkheimian Individual

There have been exceptions to the Durkheimian line, nevertheless: ethnographies and analyses which deny the priority (ontological, developmental, historical) of the societal in its causing and conditioning of individual experience. Tracing the ‘origins of English individualism,’ Macfarlane (1978) sets out specifically to refute those theorists who would see individualism as a recent socio-cultural development—whether epiphenomenal upon religious or Renaissance or Enlightenment or capitalistic or industrialistic thinking. For:

the majority of ordinary people in England from at least the thirteenth century were rampant individualists, highly mobile both geographically and socially, economically ‘rational,’ market-oriented and acquisitive, ego-centred in kinship and social life (1978, p. 163).

In New Guinea, meanwhile, Burridge (1979) describes how most people are ‘individuals’ and ‘persons’ in different respects and at different times, where ‘persons’ may be understood as those who realise given sociocultural categories, prescribed by traditional morality and the social order, while ‘individuals’ are those who use their perception and intuition to create anew. If persons are products of current sociocultural conditions, then individuals exist in spite of them. Each ‘spatially bounded organism’ is able to switch between these two modalities, Burridge concludes, such individuality representing an existential imperative which pre-exists culture. This also accords with Wikan’s (1990) reappraisal of the Balinese, where the institution of cultural norms of personal composure, grace and placidity do not obviate the individual emotional effort and will be recognized as necessary to effect and sustain their appearance.

An emphasis on individual actors, their personal experience and agency, similarly found expression in early anthropological theorists of consciousness (such as Gregory Bateson and Anthony Wallace), in the transactionalism with which Fredrik Barth (and also F. G. Bailey and Robert Paine) developed Leachian insights into instinctually ‘interested’ individuals, in the work of network analysts (such as John Barnes and Ulf Hannerz), and in the burgeoning of studies within interpretive anthropology more generally which focus on the constructions of experience made by particular individuals in interaction. These accord more with Weberian sentiments that collectivities must be approached solely as the resultants and modes of organization of the particular acts of individual persons, since it is they who remain the upper limit and the sole carrier of meaningful conduct.

Notwithstanding, structuralist and poststructuralist discourses still denigrate an anthropological sensitivity of this kind as ‘methodological individualism’: as an erroneous couching of explanation in terms of individuals’ characteristics, their behaviors, and interests, such that the ‘foundational’ conditions of sociocultural reality are obscured. The center ground of anthropology, it may be true to say, remains the preserve of ‘methodological collectivism,’ positing social phenomena as determined by factors which bypass individual rationality, and hence envisaging cultural development quite independently of individual consciousness. There is a continuing insistence, in other words, that the distinction between the individual and the societal is specific to the West and must be collapsed in favor of the latter—or at least of ‘social relations’—for anthropology persuasively to encounter cultural others.

On the other hand, there is a growing insistence that it is a peculiar ethnocentrism for anthropologists to fail to allow in the ‘others’ we study the personal complexity which we recognize in ourselves: as both individuals and persons, role-players and rebels, conventional and creative. The anthropology which has sought most deliberately to keep this truth in view may be termed ‘existential’ or ‘phenomenological,’ after the philosophical traditions associated with such writers as Kierkegaard, Nietzsche, Husserl, and Sartre (Rapport 1997). It sets out with very different tenets and intentions from the ‘sociological’ orientation towards social structures and systems predominant above, privileging instead individual existence and experience.

2. Existence and Experience

To become human is to become individual, Geertz (1973, p. 52) has admitted, adding that we become individual in an environment of sociocultural forms and patterns in whose terms we give form, order, direction, and point to our lives. The important question which this raises is the precise relationship between sociocultural forms and the individual lives lived by them. Becoming human and individual in an environment of common sociocultural forms need not eventuate in becoming ‘the same’ as others, or even necessarily comparable, nor becoming after a deterministic or directly influenced fashion. For in this becoming, energy, agency, intention, and interpretation can be seen to remain properties of self-conscious subjects; indeed, it is individual agents who are responsible for the creation, animation, deployment, and development of systems of forms in their usage of them.

An intrinsic dichotomy between individual and world is often eschewed as a peculiarity of ‘Western’ sociocultural milieux, as we have heard, and hence as methodologically and analytically inapplicable. This is hardly defensible. Human beings may be

socioculturally situated, but they are also, always, interpretively autonomous—responsible for a ‘personalization’ of the world—and inexorably and inevitably (by) themselves. It is not good enough simply to say that only Western culture valorizes the concept of the individual—as ‘individualism’—and therefore only in Western society do individuals act distinctively—with ‘individuality.’ For, while the former might be said to describe a particular socio-cultural form of behavior (the pursuit of self-distinguishment), the latter concerns a condition which is a human universal: by virtue of a unique consciousness, each of us perforce engages with the world as other and is possessed of distinct perspectives upon it.

Individualism, more precisely, pertains to a specific historico-cultural conceptualization of the person which includes such notions as: the ultimate value and dignity of human individuals, their moral and intellectual autonomy, their rationality and self-knowledge, spirituality, right to privacy, self-sovereignty, and self-development, and their voluntary contracting into a society, market, and polity. Individuality, by contrast, refers to that unique awareness, and awareness of awareness, which is the mark of human embodiment. It is the universal nature of existence that human beings engage with others by virtue of discrete sense-making apparatuses (nervous systems and brains); they possess discrete centers of perceptual activity in discrete bodies. Furthermore, the human body (and in particular the brain) gives rise to knowledge of the world, to a perspective upon the world, which is inherently individual: human beings come to know themselves within the world by way of cognitions and perceptions, thoughts, feelings, and imaginings, which are unique to them (Edelman 1992).

Not only is there an individuality intrinsic to human consciousness but also to human agency. For, it is in terms of his perceptions that an individual comes to act, each individual becoming a discrete centre of motivation and intention. Not only is the individual’s being-in-the-world universally mediated by very particular interpretative prisms which distance him from it, then, but while intrinsically ‘of the world,’ the individual inexorably comes to know the world, and act towards it, as ‘other.’ Individuals experience and interpret (and experience themselves interpreting) and therefore they are.

An individuality of consciousness and agency is current whatever the provenance of individualism as a cultural norm. Individuality is the human *a priori*, the physical-psychical basis on which all human knowledge of the world, all creativity within it and all activity upon it, rests (including the creation and representation of sociocultural milieux). Individuality remains consequential, that is, whether or not individual consciousness is an item of collective discourse, whether or not individual reflection is publicly eschewed, and whether or not individual distinctive-

ness is disparaged through the institutionalization of a fund of common behaviours (concerning, for instance, notions of the person). The experience of self is essentially distinct from sociocultural representations of self (cf. Spiro 1993).

Finally, an existential anthropology of personhood can hope to ‘decolonize’ (Cohen 1994) the individual human subject both from overdetermining cultural conditions and overweening social institutions (discourse, language-game, collective representation, social relationship, *habitus*, praxis), and from their holistic and hegemonically minded social-scientific commentators. It can take into account both the individual agency which brings sociocultural milieux to life and also the common sociocultural forms and practices by which individuals coordinate their activities and world-views within these milieux. Retaining respect for individual cognitive processes it apprehends the ambiguous interface between convention and individuality, and portrays sociocultural milieux as comprising, constituted by, individual difference: human individuals making a diversity of meaningful worlds by way of a commonality of cultural forms. In this way, an existential appreciation is promised of the ‘individuals’ behind the ‘persons,’ and of that consciousness of ‘self’ by which the disjunction is mediated.

See also: Individual/Society: History of the Concept; Knowledge, Anthropology of; Person and Self: Philosophical Aspects; Phenomenology: Philosophical Aspects; Self: History of the Concept; Self-knowledge: Philosophical Aspects; Self: Philosophical Aspects

Bibliography

- Burridge K 1979 *Someone, No One*. Princeton University Press, Princeton, NJ
- Cohen A P 1994 *Self Consciousness*. Routledge, London
- Dumont L 1986 *Essays on Individualism*. University of Chicago Press, Chicago
- Edelman G M 1992 *Bright Air, Brilliant Fire*. Basic Books, New York
- Geertz 1973 *The Interpretation of Cultures: Selected Essays*. Basic Books, New York
- Harris G G 1989 Concepts of individual, self and person in description and analysis. *American Anthropologist* 91: 599–612
- Macfarlane A 1978 *The Origins of English Individualism*. Blackwell, Oxford, UK
- Mauss M 1985 A category of the human mind: The notion of person, the notion of self. In: Carrithers M, Collins S, Lukes S (eds.) *The Category of the Person*. Cambridge University Press, Cambridge, UK
- Myers F R 1986 *Pintupi Country, Pintupi Self*. Smithsonian Institution Press, Washington, DC
- Poole F 1998 Socialization, enculturation and the development of personal identity. In: Ingold T (ed.) *Companion Encyclopedia of Anthropology*. Routledge, London
- Rapport N 1997 *Transcendent Individual: Towards a Literary and Liberal Anthropology*. Routledge, London

- Read K E 1955 Morality and the concept of the person among the Gahuku Gama. *Oceania* **25**(4): 233–82
- Spiro M E 1993 Is the Western conception of the self peculiar within the context of the world cultures? *Ethos* **21**: 107–53
- Strathern M 1988 *The Gender of the Gift*. University of California Press, Berkeley, CA
- Wikan U 1990 *Managing Turbulent Hearts*. University of Chicago Press, Chicago

N. J. Rapport

Personnel Selection, Psychology of

The term personnel selection has a scientific as well as a practical meaning. Scientifically, personnel selection indicates the methodology of developing and evaluating psychological methods to assess task-related qualifications which permit predictions of occupational success, job-related classification, and application of means of personnel development, including models and tools of personnel decision making. Thus, it should be referred to as personnel selection research. In organizational practice, personnel selection means the endeavor of choosing those individuals from the group of applicants who best meet the demands of the job and the organization. Although all kinds of job-related consulting for individual occupational and organizational choice are served as well by the same psychological methods, in English terminology the term personnel selection generally is used for the whole bundle of research, development, and application for all purposes in this context. Following this, this article will discuss the most important aspects relevant to these issues.

1. Comparison of Person and Job

Occupational or job-related qualification, understood as the probability of success, basically means 'who is suitable and what for?' That is, the target groups or individuals are to be defined just as well as the target

positions. The requirements of the tasks or positions are to be established, and the necessary characteristics of the job-holders have to be derived on that basis. For very specific requirements, diagnostic methods are available or will be developed, or adapted, in order to measure and to compare job-relevant traits or abilities. Besides performance criteria, other indicators of occupational success can be formulated as objectives of the diagnosis, namely job and work satisfaction, health and well-being, or the tenure in the organization. Consequently, in addition to task analysis, the potential for satisfaction of the job in question has to be determined and compared with the interests and needs of the individuals.

Aside from determining the present requirements, an attempt can be made to estimate what kinds of changes in the requirements are to be anticipated; furthermore, an undeterminable amount of change has to be expected. The required development potential of a person should correspond to the foreseeable changes and offer the prospect to be up to future developments of an uncertain nature. Predictions are difficult in this area; however, some characteristics turned out to be relevant to success in most occupational contexts. Based on these general findings, it is necessary to compare person and job on three different levels (Fig. 1).

Different methods of task and job analysis are available for the investigation of occupational requirements. For our context, the most important kinds of requirements—not always clearly separable—are formulated as trait requirements (e.g., abilities and interests), behavioral requirements (e.g., skills and habits), qualification requirements (e.g., knowledge and skills), or as outcome requirements (e.g., problem solving and quality standards). In most cases the assessment of requirements is conducted by observation or by means of interviews or questionnaires.

2. Job-relevant Characteristics

The required characteristics of the job candidate are investigated, as far as feasible and appropriate, as they apply to the job-specific requirements. However,

Occupation/job	Individual
Requirements	Abilities, skills, and knowledge
Satisfaction potential	Interests, needs, and values
Changes	Development potential and general success-relevant characteristics

Figure 1
Comparison of person and job on three different levels

several characteristics also proved to be relevant predictors of occupational success. Preceding all is general intelligence, for which meta-analytical methods (Hunter and Schmidt 1990) provided evidence which showed that it is a valid predictor of achievement for practically all occupational fields (the effect of intelligence on work performance is mainly due to the ability to acquire job knowledge). As opposed to this, global personality traits such as extraversion and openness to experience have not shown to possess general validity. However, restricted variance in applicant samples may be responsible for this finding, adding to the problem that personality inventories usually lack job orientation. Among the global personality characteristics, the trait 'conscientiousness' seems to be of general relevance. In a meta-analysis using only European samples, Salgado (1997) found emotional stability (neuroticism) also to be a valid predictor for all jobs and criteria. While validity coefficients for these general traits turned out to be rather low, some narrower and less well-researched personality constructs showed higher validities at least for several occupational groups. Among these traits are: achievement motivation, self-confidence, and dominance. The highest predictive validity for a non-cognitive trait has been found for 'integrity' (Ones et al. 1993), while its construct validity is still a controversial issue. Approaches to determine the interaction of several characteristics—especially those belonging to different diagnostic taxonomies—are only in the beginning. The application of structural equation modeling and related statistical methods should lead to substantial progress (Schmitt and Chan 1998).

3. The Trimodal Approach to Personnel Selection

The distinction of three general methodical approaches is helpful for the classification of diagnostic instruments: the *trait approach*, the *simulation approach*, and the *biographical approach* (Schuler 2001). Each one of these approaches pursues a partially independent validation logic and corresponds to specific methods of personnel assessment. With the trait or construct approach, characteristics are measured which are assumed to be relatively stable, e.g., conscientiousness and verbal intelligence. In order to assess these traits as homogeneous constructs, psychological tests are being used as typical instruments. In this case, construct validity is at the core of successful measurement. The objective of the simulation approach is the assessment of behavior similar to that required at the workplace; the appropriate validation strategy is the one which aims at representativeness or *content validity*. A typical form of measurement is the *work sample*. The third diagnostic principle is the biographical approach, although it cannot always be

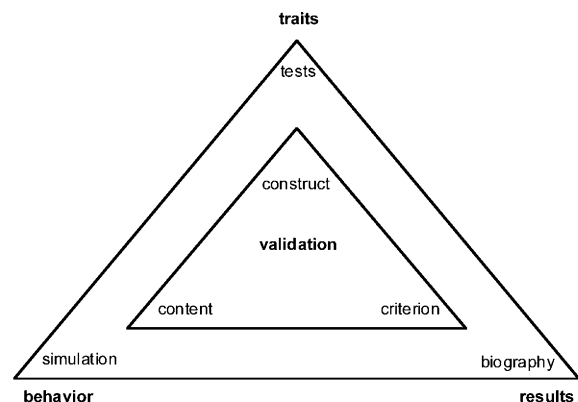


Figure 2

The trimodal approach to personnel selection

sharply distinguished from the other approaches. Typical assessment methods are biographical questions which can be presented in paper-and-pencil instruments (questionnaires), or in verbal form during an interview. Currently not in use, but a potentially fruitful method is the ideographical analysis of biographies (Fig. 2). For practical personnel selection, this means that in most cases multiple procedures are superior to cover complex requirements situations.

4. Instruments for Personnel Selection

First, the section discusses different types of single procedures, subsequently more complex procedures are presented (for a more detailed report, refer to Schmitt and Chan 1998, Schuler 2001).

4.1 Tests

Psychological tests are standardized routinely used methods for the measurement of individual behavioral characteristics, from which conclusions can be drawn pertaining to the individuals' traits or their behavior in other situations.

In scientifically controlled aptitude measurement, tests are the most commonly used instruments. They are applied to practically all occupational fields and job-relevant abilities. The most important kinds of tests used in personnel selection are:

- (a) Tests of general cognitive ability (intelligence);
- (b) Tests of specific cognitive abilities;
- (c) Tests of attention and concentration;
- (d) Tests of the sensory and motor abilities;
- (e) Other achievement tests;
- (f) General personality tests;
- (g) Specific personality tests;
- (h) Tests of attitudes, motivation, and interests.

Numerous tests have been published and are available from the test publishers. However, large organizations (such as the Federal Labor Agency, industrial companies, the military) often employ methods specifically designed for their purposes. The number of applications of these methods depend on the relevant occupational group. Figure 3 (data from 1990) shows that the extensiveness with which large companies use tests differs substantially between European countries.

4.2 Work Sample Tests

Work sample tests, being the prototype of simulation-oriented diagnostic tools, are standardized tasks which require assessees to carry out success-relevant occupational behavior. A high similarity of predictors and criteria is intended. The construction of work samples generally follows the same principles as that of psychological tests. A substantial difference is that it is largely done without 'translating' the occupational tasks into trait requirements. Rather than inferring a predisposition (trait) from a 'sign' (test behavior), conclusions are drawn from a sample of behavior for similar future behavior (Robertson and Kandola 1982). Examples are the preparation of a workpiece by a mechanic or a 'test lecture' by a teacher. A new class of work samples has been created by using video technology and virtual reality.

The use of computer-based procedures for psychological testing is rapidly growing (Booth 1998). The simulation of complex and dynamic problem solving tasks in scenarios requires the application of computers. These types of problems, originally developed in cognitive psychology, confront the applicant with a virtual environment where he/she has to control a system so that certain objectives will be reached. Hereby, the outcome variables (e.g., sales) as well as process variables (e.g., decisions), which are characteristic of a person's working behavior, can be observed.

There is also a special type of work samples that can be described as a kind of hidden intelligence tests, e.g., *job knowledge tests*. As construct validation research showed, cognitive ability is the primary determinant of knowledge acquisition. An extensive meta-analysis by Dye et al. (1993) proved job knowledge tests to be among the most powerful predictors of occupational success, especially in cases of high similarity between jobs and tests, and high job complexity.

4.3 Biographical Information

The basic principle of the biographical approach in personnel selection is the direct prediction of future behavior from past behavior. Application documents are being evaluated primarily with respect to that sort of information, although other (e.g., formal) aspects also require attention. Essential biographical elements are the curriculum vitae and references that contain

information about job experience and achievements, as well as school and university grades. Methods of analysis to gain a maximum of information out of application documents have not been developed yet.

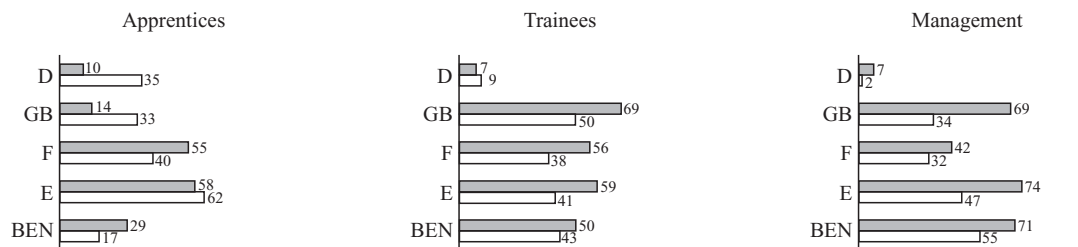
Prototypical for the biographical approach is the biographical questionnaire. These instruments are often difficult to distinguish from personality tests. In comparison to personality items, biographical items are generally related to the past, and they typically concern behavior and verifiable events. Classical principle of questionnaire construction is the validation of each item or even of each response alternative within a multiple-choice item with respect to external criteria. This results in a task-specific and often also organization-specific item selection—and as a consequence in limited generalization. For this reason, the 'blind empirical' approach is often combined with a rational approach which is requirement- and construct-related (Stokes et al. 1994). In most cases the most straightforward sources of performance-related biographical information are the level of education, school, and college grades. While the level of education is of high predictive value, school grades provide highly valid information for later educational performance, but the validity for predicting job performance is only moderate (Roth et al. 1996). Yet, in the 1980s and 1990s, predictive validity has even decreased for High School grades (Schuler 2001).

4.4 Interviews

Typical modes of employment interviews vary from a totally free conversation over partly structured to fully structured variants using standardized procedures, questions, and response evaluation. Ensuring interview validity can best be achieved by constructing the interview tool closely adapted to job requirements, standardizing questions, providing examples of response evaluation on behaviorally anchored rating scales, and combining several types of questions. The construction of interview questions can follow the same psychometric principles as test construction. Among the more elaborated types of interviews are the *Situational Interview* (Latham et al. 1980) and the *Multimodal Interview* (Schuler and Funke 1989). By means of the latter, it could be demonstrated that both construct- and simulation-oriented principles of assessment can be realized in an interview, as well as the biographical principle. Moreover, a sequence of structured and unstructured interview components takes account of a candidate's preferences for selection instruments that allow behavioral control, i.e., situations of low structure.

4.5 Multiple Procedures

The purpose of multimodal assessment procedures is to take into account a heterogeneous requirement



Note: Germany (D), Great Britain (GB), France (F), Spain (E), and the Benelux (BEN). Contrasted are the use of personality tests (■) vs. ability and intelligence tests (□) in percent of companies.

Figure 3

Application of psychological tests in several European countries

constellation by combining different methods and diagnostic principles. Basically, all types of single procedures can be combined; however, meta-analytical calculations show that noncognitive predictors can provide substantial supplements only in a few cases, i.e., incremental validity, in addition to intelligence tests. The most important of these predictors are work samples, specific personality tests, and structured interviews. As multiple selection procedures for supervisors and trainees, *assessment centers* are often used (Thornton 1992). They usually consist of work-samples such as group discussions, organizational tasks, presentations, and dyadic role-plays. Characteristic for an assessment center is that several applicants participate at the same time, and that the evaluation is carried out by several independent assessors. As far as the procedure succeeds to grasp the entirety of a person's potential for future development, we speak of *potential analysis*. Where reliable criteria are available, single components of multimodal procedures can be weighted for the calculation of a total score (e.g., by multiple regression).

5. Evaluation of Selection Instruments and Procedures

The evaluation of personnel selection procedures has to examine test theoretical criteria, especially objectivity, reliability, and validity. In addition to these, instruments have to be assessed according to organizational efficiency or advisory efficiency, to social validity as well as to further ethical and legal aspects.

5.1 Validity

Among the criteria of psychometric quality, validity is the most important, the others are necessary, but not sufficient conditions for validity. Among the variants of validity (or strategies of validation), *predictive validity* plays a crucial role, since the objective usually is the prediction of future occupational success. The following diagnostic procedures demonstrated good

or sufficient validity (Salgado 1999, Schmidt and Hunter 1998): tests of general intelligence, work samples, job-related structured interviews, biographical questionnaires, some specific personality tests (integrity, conscientiousness, achievement motivation), and multimodal procedures (assessment center, potential analysis); tests of job knowledge and assessment of job performance are in a similar range. It must be noted that there are some moderating conditions (e.g., the validity of biographical questionnaires for young persons is low) and that validation leads to different coefficients for different measures of success—i.e., supervisory assessment, position level, income, assessment of potential. The aspect of *content validity* is important especially during the steps of instrument construction, that of *construct validity* in the examination of the psychological meaning of test scores.

5.2 Organizational Efficiency

Organizational efficiency is operationalized by aspects of practicability; these are, above all, the *expenditure* of the procedure, the required competence for its application, and its availability. Above all, *economic utility* is a core element of organizational efficiency. It depends on several parameters, especially validity, selection rate, and base rate as well as the variance in the performance criteria (that means, selection utility grows with the size of subsequent achievement differences). Supplemented by different economic indicators, the expected benefit can be calculated. Such calculations frequently lead to high utility estimates for the application of personnel selection instruments. *Advisory efficiency* may be understood as usefulness of diagnostic procedures to support individual occupational and organizational choice. Due to the relatively small general relationship between interests and abilities, such benefit is usually assumed. Advisory efficiency is important primarily in job counseling at labor agencies, but also in the context of organizational personnel development.

5.3 Applicants' Reactions

The applicants' *acceptance* of the selection process and their reactions to personnel selection methods is considered an independent quality criterion. Procedures which are better accepted are distinguished by information about job requirements as well as by transparency, feedback, and the possibility to control the situation. These requirements are better fulfilled in interactive procedures (i.e., interview) than in paper-and-pencil-tests or in biographical documentation (i.e., school grades).

5.4 Ethical and Legal Aspects

Ethical, professional, as well as legal aspects of personnel selection have received increasing attention. Examples for concrete ethical problems are: use of undue diagnostic methods, negligence of job requirements, application of intended distress during selection procedures, unjust invasion of privacy, lack of confidentiality of personal data, conflict of interest in relation to the employers and the clients. In this context, endeavors have been intensified to organize selection processes as fair and socially acceptable procedures (Schuler et al. 1993).

The appropriate legal basis—in Germany it is the general individual law as well as the labor law—are subject to substantial international differences. In Europe, an adaptation process takes place at present.

6. Concluding Comments

The adequacy of the application of personnel selection instruments in specific cases depends on a number

of factors. If we disregard all required differentiations, the value of diagnostic information for the most relevant occupational groups can be judged as is presented in Fig. 4, which gives a summary of all is discussed above, that is, validity, efficiency, social validity (acceptability and fairness) as well as ethical and legal aspects. We have to get used to the idea that all these aspects will be important for the use of personnel selection methods in the future.

As a result of innovative instruments of personnel selection being vividly required by practitioners, the chance for the development and application of new approaches to assessment has increased. However, our insight into the incremental validity of additional methods or the value of combined procedures has not improved in the same way. Once science has moved towards that direction, we could approach a project that at this point of time still has the character of a vision—to draw diagnostic information not only from single, specifically constructed procedures, but from a variety of human characteristics. Performance results may be combined with self-appraisals and other assessments, evidence on abilities may be drawn from preferences and personal values, and biographical incidents, in connection with physiological data and even characteristics of the physique or other tabooed indicators, may have their place in highly valid predictor combinations.

However, one prerequisite for such a multi-indicator diagnosis are models that go beyond the available multivariate statistics, although current structural equation models may represent one step into that direction. One possible attempt is to apply the principles of fuzzy logic to validity estimation of such combined conclusions. On the other hand, the perspective of the employee or, as we call it, the *social validity* of these procedures must not be overlooked by

Types of Employment	Apprentices	Blue Collar	Employees w/o supervisory functions	Trainees	Management		
					Lower	Middle	Top
Selection method							
Application documents	+	±	+	+	+	+	+
Additional references	–	±	±	±	±	+	+
Structured, job-related interviews	+	±	+	+	+	+	+
Academic grades	+	±	±	+	±	–	–
Biographical questionnaire	–	–	±	±	+	+	–
Specific, job-related personality tests	+	±	+	+	+	+	±
Achievement and knowledge tests	+	+	+	+	+	±	±
General and specific intelligence tests	+	±	±	±	±	–	–
Work samples	+	+	+	+	+	+	+
Assessment center/analysis of potential	±	–	±	+	+	+	±

Note: + appropriate; ± relatively appropriate; – inappropriate or not required

Figure 4

Adequacy of selection instruments for groups of employees

concentrating on such 'technical' innovations (Schuler et al. 1993). To keep the further development of diagnostic techniques in balance with these requirements will be an important challenge for personnel selection research in the future.

See also: Job Analysis and Work Roles, Psychology of; Personality Assessment

Bibliography

- Booth J F 1998 Uses of PC technology in selection and assessment. *International Journal of Selection and Assessment* **6**: 57–60
- Dye D A, Reck M, McDaniel M A 1993 The validity of job knowledge measures. *International Journal of Selection and Assessment* **1**: 153–7
- Hunter J E, Schmidt F L 1990 *Methods of Meta-analysis*. Sage, Newbury Park, CA
- Latham G P, Saari L M, Pursell E D, Campion M A 1980 The situational interview. *Journal of Applied Psychology* **65**: 422–7
- Ones D S, Viswesvaran C, Schmidt F L 1993 Comprehensive meta-analysis of integrity test validities: Findings and implications for personnel selection and theories of job performance [Monograph]. *Journal of Applied Psychology* **78**: 679–704
- Robertson I T, Kandola R S 1982 Work sample tests: Validity, adverse impact and applicant reaction. *Journal of Occupational Psychology* **55**: 171–83
- Roth P L, BeVier C A, Switzer II F S, Schippmann J S 1996 Meta-analyzing the relationship between grades and job performance. *Journal of Applied Psychology* **81**: 548–56
- Salgado J F 1997 The five factor model of personality and job performance in the European Community. *Journal of Applied Psychology* **82**: 30–43
- Salgado J F 1999 Personnel selection methods. In: Cooper C L, Robertson I (eds.) *International Review of Industrial and Organizational Psychology*. Wiley, London, UK pp. 1–54
- Schmidt F L, Hunter J E 1998 The validity and utility of selection methods in personnel psychology: Practical and theoretical implications of 85 years of research findings. *Psychological Bulletin* **2**: 262–74
- Schmitt N, Chan D 1998 *Personnel Selection, A Theoretical Approach*. Sage, Thousand Oaks, CA
- Schuler H 2001 *Lehrbuch der Personalpsychologie* [Personnel Psychology]. Hogrefe, Goettingen Country
- Schuler H, Farr J L, Smith M (eds.) 1993 *Personnel Selection and Assessment: Individual and Organizational Perspectives*. Erlbaum, Hillsdale, NJ
- Schuler H, Funke U 1989 The interview as a multimodal procedure. In: Eder R W, Ferris G R (eds.) *The Employment Interview—Theory, Research, and Practice*. Sage, Newbury Park, CA pp. 183–192
- Stokes G, Mumford M, Owens E (eds.) 1994 *Biodata Handbook: Theory, Research and Use of Biographical Information in Selection and Performance Prediction*. Consulting Psychologists Press, Palo Alto, CA
- Thornton G C 1992 *Assessment Centers in Human Resource Management*. Addison-Wesley, Reading, MA

H. Schuler

Copyright © 2001 Elsevier Science Ltd.
All rights reserved.

Pes

Pestalozzi, Johann Heinrich (1746–1827)

1. Introduction

Johann Heinrich Pestalozzi was one of the most distinctive writers of his time, and was already regarded as one of the most famous pedagogs in Europe in his own lifetime. He owed his reputation as a writer essentially to his best-selling novel *Lienhard and Gertrud* (1781), and owed his fame as a pedagog to his work in Stans (1799), Burgdorf (1800–1803) and in Yverdon (1804–1825) where for 21 years he directed an institute in which children from all over Europe were educated.

Johann Heinrich Pestalozzi's work deals with a wealth of subjects in the fields of general politics, social policy, constitutional law, jurisprudence, literature, anthropology, philosophy, pedagogy, and social pedagogy, and it is documented in 30 volumes in the Critical Edition *Sämtliche Werke und Briefe*, 15 of which have hitherto appeared (also available on CD-ROM).

Pestalozzi was, in both his personal and literary life, actively opposed to social injustice, oppression, exploitation, dependence, and lack of political freedom. His ambition was to free the poor and their children from powerlessness, to help them to be able to secure their own individual subsistence, as well as to abolish the very political conditions which caused this hardship. He constantly gave central importance in his work to social justice, the satisfaction of needs, ownership and work, domestic order, and basic needs for all, including children. The subjects he addressed were rooted in social reality. Pestalozzi reflected upon and analyzed the everyday life of the poor masses, the everyday life of children, and the everyday political life of his time.

The special form of this relation to reality Michel Soetard (1981) has convincingly called the 'Principle Reality.'

2. Life and Work

2.1 1746–1799: Childhood, Youth, and Early Career

2.1.1 Family origins and childhood. Johann Heinrich Pestalozzi was born in Zurich on the 12

January 1746, the son of the surgeon Johann Baptist Pestalozzi and his wife Susanna, née Hotz. At that time, Zurich was a center of European culture, a city state with an exclusive merchant and entrepreneurial aristocracy. The town was regarded as rich, with a small number of old families holding power in the city council. Neither his father's family nor his mother's—who originally came from the Zurich area where her father was a parish priest—belonged to this group. The married couple produced seven children in 9 years of marriage from which only three grew to maturity. When Pestalozzi—one of seven children—was 5 years old, his father died at the age of 33, leaving his mother to take care of four small children. In order to lead a bourgeois life in spite of the financial predicament, economizing and making-do became a staple of their dismal everyday life. Pestalozzi had therefore to endure poverty. In the country at his grandfather's house in Höngg, where his grandfather was a parish priest, he became acquainted early on with the miserable social condition of the provincial population which was to occupy him for a lifetime.

2.1.2 School and studies. In Zurich, Pestalozzi—who had the civic rights—grew up with the sons of the patrician families whose carefree material situation was permanently laid out before his eyes. He enjoyed the schooling of a city-dweller, and went to the city elementary school from 1751 to 1754; he was then a pupil from 1754 to 1763 at the Schola Abbatissana in Zurich, the Schola Carolina, and the Collegium humanitatis (also in Zurich), the latter being a Latin school. After some initial difficulties he was a good pupil, though because of his financial situation he was always an outsider because of his 'extraordinary carelessness and uncleanness.' From 1763 to 1765 he studied at Zurich's theological college, the Carolinum. The appearance of an anonymous pamphlet he had co-written caused the abrupt end to his studies. It was clear to him in any case that he did not wish to become a priest.

In Zurich there were—following the example of other European countries—patriotic and physiocratic societies. From 1764, Pestalozzi belonged to the circle of patriots of the Helvetic Society. The Zurich scholars Bodmer and Breitinger in particular had their circle of influence here. The much-admired model was Rousseau. Republican works were read and written. It

was in this context that Pestalozzi's first publications *Agis* (1766) and *Wünsche* (1766) appeared. In 1767, Pestalozzi began a short agricultural apprenticeship with the Bern physiocrat Tschiffeli.

2.1.3 Livelihood. In February 1769, Pestalozzi became a farmer in Müllingen and ran into heavy debt. On 30 September 1769, he married *Anna Schulthess* (1738–1815), and on the 13 August 1770 their only child Hans Jakob, called Jaqueli (1770–1801), was born. Pestalozzi built a manor house, the 'Neuhof auf dem Birrfeld' which he moved into with his young family in 1771. Crop failure, bad purchases of land, and other mistakes led to failure and almost to financial ruin. In 1794, he founded an institution for the poor at the Neuhof. His initial interest was more entrepreneurial than pedagogical; in his *Neuhof-Schriften* (published in 1777), Pestalozzi accompanied his essays with a journalistic advertising campaign for support for his project. In his *Bitte an Menschenfreunde*, he successfully enlisted the support of, amongst others, Niklaus Emanuel Tschärner (1727–1794), the chairman of the Bern Society for Economics. Pestalozzi discovered his interest in all things pedagogical during his time at the Neuhof, motivated by his interaction with the poor children of the Neuhof: 'I saw how the overwhelming cruelty of self-interest constantly destroys the body and soul of almost—and I mean almost—all of these children.' Pestalozzi's institution failed financially, but he was later to make use of his experiences in the field of national education.

In the middle of the Neuhof crisis, the political work about Zurich *Freyheit meiner Vaterstadt* (1799)—which remained unpublished—was written in which he lamented the wealth of the city and its devastating effects on the constitution and political culture of Zurich.

2.2 From 1780 to 1799: Livelihood as an Independent Writer

Out of the necessity of having to finance himself, Pestalozzi became a writer. In 1780 he published *Abendstunde eines Einsiedlers* whose theme is cultivating people: 'Cultivating people towards the truth means cultivating their being and their nature towards a reassuring wisdom.' Then he wrote the first part of *Lienhard und Gertrud*, which appeared in 1781. *Lienhard und Gertrud* is a pedagogical novel which its author broadened into a novelistic pedagogical text. The theme of the novel is the story of the education and cultivation of an entire village. With *Lienhard und Gertrud*, Pestalozzi wrote about a fictitious village which he called Bonnal in a novel showing both the development and the education of the characters. The subject of the novel is bound together by a kal-

eidoscope of multifaceted and fictitious stories of education, cultivation, and socialization.

In the second (1783) and third (1785) books, it becomes clear how many individual fates have been embroiled in the deplorable state of affairs in Bonnal for many years, and that it is thereby not only a question of individual error, but at the same time of failings which lie more deeply in the political structure. Only in the fourth book (1787) does it become possible to continue telling the story of Bonnal and to organize it structurally as a story of reform, with, of course, further novelistic twists and turns and subplots. The village as the scene of the action is increasingly abandoned. Paragraph by paragraph Pestalozzi outlines his ideas for new legislation, the distribution of land, jurisdiction, and the culture of debate and justice. The novel is therefore increasingly interrupted on the level of action and plot, becomes more difficult to read, and thus loses its original novelistic character. The literary code transforms into a politically discursive text. *Lienhard und Gertrud* became Pestalozzi's greatest literary success. Pestalozzi was soon to become known far beyond the Swiss borders, first of all not as a pedagogue, but as a writer. The novel brought him the title of Honorary Citizen during the French Revolution.

In *Gesetzgebung und Kindermord* (written in 1780 and printed in 1783), Pestalozzi declared the prevailing legal practice throughout Europe of executing child-murderesses—mostly after extreme humiliation and torture—as inhuman. He mounted an emotive defence of a large number of child-murderesses, whom he viewed as innocent. To him, poor girls and women had the fundamental right to sexuality and marriage, to their children, to protection and support, and to recognition and help. The work was a plea against capital punishment.

In 1782, Pestalozzi published *Christoph und Else*, a novel of instruction, even a 'novel' for the novel *Lienhard und Gertrud*. It was to be used as a guide to how adults should read *Lienhard und Gertrud*. A peasant family talk together over 30 evenings about the plot, the theme, and the constellation of characters in *Lienhard und Gertrud*. On 26 August 1792, Pestalozzi was appointed an Honorary French Citizen and never renounced the honorary title in spite of all the criticism of the Terror, an action which aroused suspicion in conservative circles. The outcome of Pestalozzi's examination of the revolution was the work *Ja oder Nein?* (1793) in which he made the Ancien Régime responsible for the devastating outrages in Europe and thus justified the revolution, though at the same time he sharply denounced the revolutionary Terror.

In 1797, Pestalozzi's main philosophical work *Meine Nachforschungen über den Gang der Natur in der Entwicklung des Menschengeschlechts* appeared. The initial question was 'What am I, and what is mankind?' Pestalozzi expounded a natural state, a social state,

and a moral state of humanity, and went on to discuss the individual human being as a ‘work of himself,’ a ‘work of nature,’ and a ‘work of morality.’

In 1798, the revolution also reached Switzerland. For a short time, a Helvetic Republic was established. Pestalozzi became the editor of the ‘*Helvetisches Volksblatt*’—the publication organ of the new Helvetic government—and at first worked on behalf of the government. He wrote texts, appeals, and pamphlets, as well as *Abhandlung über die Natur des Helvetischen Zehnden*. Whereas the typical pamphlets on both sides at the time of the revolution constituted obvious works of political agitation in favor of one side or the other, Pestalozzi was a mediator who urged the often very hostile opposing groups towards agreement. He became a soother of passions in his pamphlets. Pestalozzi wanted to promote the unity of the Swiss in the crisis year of 1798, the unity between the poor and rich population, the unity of town and country, and the unity of the Inner-Swiss (Catholic) cantons and the majority Protestant cantons.

2.3 1799–1827 The Pedagog and Institute Director: From Stans to Yverdon

2.3.1 Stans. At the age of 53, Pestalozzi received a government commission to run an orphanage in Stans. From 7 December 1798 to 8 June 1799, he lived there together with a housekeeper and as many as 80 children by the end. His pedagogical work and his pedagogical writing were now supported by two pillars. The first was a love of children and a passion for his work. Pestalozzi made himself into the ‘father’ and ‘mother’ of all the children; he tried to transpose the notion of the great importance of satisfying basic needs. The second was that the fundamental principle of the ‘method’—which was on the one hand an almost technocratic syllabic reading method—was at the same time elevated by the idea of an elementary method of moral education. From then on, Pestalozzi’s elementary education and elementary method were central themes. The institute, however, came to a sudden end. Pestalozzi had to hand over the building to French soldiers returning from the Russian campaign for use as a hospital and leave the institute, an action which made him fall into deep depression and despair. During this time, he wrote his strongly influential work, the *Stanser Brief* (1799), which contained a passionate plea for the malleability of human beings, including, and indeed especially, for the malleability of poor, neglected children.

2.3.2 Burgdorf. Pestalozzi became a teacher in Burgdorf on 23 July 1799. The announcement of a

schoolmasters’ conference in Burgdorf Palace followed on the 24 October 1800. The appearance of *Wie Gertrud ihre Kinder lehrt* followed in 1801. In the first part, Pestalozzi reported on Stans in autobiographical form, on his first endeavors in Burgdorf, on his young colleagues, some of whom accompanied him for a considerably long time, and on how he conceived method, the unity of knowledge, ability, and feeling. Pestalozzi developed a research program which he continued in Yverdon and which aimed at developing suitable learning aids in the three directions which he soon named ‘HeadHeartHand.’ The principles of the new teaching method consisted of the ability to visualize, simplification, logical progression, and individualization. The method was aimed at developing the autonomy of the child, at creating an independent personality which could secure its own subsistence. From November 1802 to February 1803, Pestalozzi traveled to Paris as a member of the Consulta. He was, however, unable to push through his concept for national education there. In June 1804, the institute at Burgdorf was moved to Müncbuchsee, and at the end of August the opening of a subsidiary institute followed in Yverdon, a town in French-speaking Switzerland, and on the 6 July 1805, Müncbuchsee was closed.

From 1804 to 1825, Pestalozzi ran one of the largest European educational institutes of the time in the rooms of the palace at Yverdon in French Switzerland. Yverdon was an experimental school on the level of its teaching, education, and teacher training. In order to be able to disseminate the new knowledge acquired from the everyday pedagogical research, a printing works was set up specifically for this purpose. Pestalozzi, who until that time had always written his works alone, now published journals, teaching books, teaching material, discourses on teaching methods, appeals, and speeches with his colleagues. At the center of these works stood Pestalozzi’s concept of an elementary education of people which could be applied to different teaching disciplines.

At the height of its impact in 1809, the institute had 150 pupils from 6 to 16 years old, 31 teachers and assistant teachers, and 40 young people as observers from all over Europe who wished to study Pestalozzi’s method. Pestalozzi’s concept of elementary teaching was already being followed ever more closely in Prussia, Spain, and many other countries. Pestalozzi’s most important work from the Yverdon period is *An die Unschuld, den Ernst und den Edelmut meines Zeitalters und meines Vaterlands* (1815), a treatise in the form of an address in which Pestalozzi set out his program for life: ‘The beginning and the end of my policy is education.’ Pestalozzi analyzed the approach of the industrial age with its phenomena of concentration, mechanization, and standardization. Only in education did he see the means to prepare people for the Industrial Age. The work caused a great sensation and was a literary success.

The size of the staff and the differences of interest and opinion within the teaching body increasingly led to endless quarrels which finally caused the downfall of the institute. Pestalozzi left Yverdon embittered in 1825, and returned to the Neuhof.

In the last 2 years of his life, he wrote the works *Meine Lebensschicksale als Vorsteher meiner Erziehungsinstitute in Burgdorf und Iferten* (1826), the *Schwanengesang* (1826), and finally the *Langenthaler Rede*. During the crisis which industrialization and the reorganization of the states into large administrative states gave rise to, Pestalozzi saw a great opportunity for man to return to himself, and for his inner renewal by means of a comprehensive elementary education.

After a short illness, Pestalozzi died on 17 February 1827. On 19 February, he was buried near the Neuhof in the small village of Birr.

3. Critical Reception and Research Developments

The research literature on Johann Heinrich Pestalozzi's life and work, amply documented by the research papers and synopses of the Pestalozzi literature of the past 200 years, at first glance appears immensely extensive. This lies in the fact that Pestalozzi, in the history of (German-speaking) pedagogy and educational sciences, has without doubt experienced a considerable *impact* history which has admittedly only to a certain degree coincided with the *reception* history of his numerous written works. Above all, Pestalozzi's teaching method is prescribed in many countries around the world and his best-known works have been translated into many languages. Should one follow the numerous publications up to the present, it becomes apparent that, whereas in the first half of the twentieth century the 'whole Pestalozzi' became the subject, individual or partial aspects began, however, to dominate more and more markedly within the framework of the differentiation of educational science and its adjacent sciences. In this way, the paradigm exchanges in educational science are mirrored in the respective cognition interests of analyses. If the Pestalozzi from the history of ideas took center stage in scientific literature in the middle of the twentieth century, the research controversy about the *political Pestalozzi* (Rang 1967) was sparked off in the 1960s and 1970s. In the course of the 1970s with its far-reaching education and school reform, interest was then concentrated rather more on the practitioner and reformer. Within the research history, it furthermore became clear that, after a very intensive, bibliographically broad and verifiable acceptance of Pestalozzi the politically reforming pedagog, the research in the 1980s stagnated; indeed the number of papers and studies about Pestalozzi fell off considerably. That only changed in 1996. Since then, the trend has been moving in the direction of an examination of the research into context and reception (the impact history) alongside biographical research using new

scientific methods, aimed at producing a differentiated situating of the work as regards history and content. An important impetus comes from the work of the Pestalozzi Research Center which, dividing the work between the Pestalozzianum and the University of Zurich, is pushing forward with the Critical Edition, and publishes the *Neue Pestalozzi-Studien* and the *Neue Pestalozzi-Blätter*.

Pestalozzi's journalistic work is so diverse that it not only touches on areas of social and cultural practice as regards content, or even focuses on them explicitly as a central theme, but also represents groups of different discourse methods. In order to understand the work, interdisciplinary research is required e.g., into rhetoric, eloquence and discourse, the socioaesthetic culture of sentimentalism, social history, the relationship between the written and the spoken, the problematic nature of text selection, the relationship between truth and fiction, the theory of femininity and the ensuing mother cult of the eighteenth century.

See also: Education: Phenomena, Concepts, and Theories; Educational Institutions, History of; Educational Learning Theory; Educational Research and School Reform; Pedagogical Reform Movement, History of

Bibliography

- Hager F P, Tröhler D 1996 *Pestalozzi—wirkungsgeschichtliche Aspekte. Dokumentationsband zum Pestalozzi-Symposium 1996*. Haupt, Bern/Stuttgart/Wien
- Korte P 2001 *Pädagogisches Schreiben um 1800*. Der Status von Schriftkultur, Rhetorik und Poetik bei Johann Heinrich Pestalozzi
- Kraft V 1966 *Pestalozzi oder das Pädagogische Selbst*. Eine Studie zur Psychoanalyse pädagogischen Denkens. Julius Klinkhardt, Bad Heilbrunn
- Liedtke M 1992 *Johann Heinrich Pestalozzi mit Selbstzeugnissen und Bilddokumenten*. Rowohlt, Reinbek bei Hamburg
- Pestalozzi J H 1946ff *Sämtliche Briefe*. Kritische Ausgabe. NZZ, Berlin/Zürich
- Pestalozzi J H 1994 *Sämtliche Werke und Briefe auf CD-ROM*. Pestalozzianum, Zürich
- Oelkers J, Osterwalder F 1995 *Pestalozzi—Umfeld und Rezeption. Studien zur Historisierung einer Legende*. Beltz, Weinheim und Basel
- Osterwalder F 1996 *Pestalozzi—ein pädagogischer Kult. Pestalozzis Wirkungsgeschichte in der Herausbildung der modernen Pädagogik*. Beltz, Weinheim und Basel
- Rang A 1967 *Der politische Pestalozzi*. Frankfurt/Main
- Soetard M 1981 *Pestalozzi ou la naissance de l'éducateur (1746–1827)*. Peter Lang, Berne/Frankfurt/Las Vegas
- Stadler P 1988 *Pestalozzi. Geschichtliche Biographie. Band 1. Von der alten Ordnung zur Revolution*. Verlag Neue Zürcher Zeitung, Zürich
- Stadler P 1993 *Pestalozzi. Geschichtliche Biographie. Band 2. Von der Umwälzung zur Restauration*. Verlag Neue Zürcher Zeitung, Zürich

Tröhler D 1997ff *Neue Pestalozzi-Studien*. Haupt, Bern/
Stuttgart/Wien

P. Korte

Phantom Limbs

Many patients awake from the anesthetic after an amputation feeling certain that the operation has not been performed. They feel the lost limb so vividly that only when they reach out to touch it, or peer under the bedsheets to see it, do they realize it has been cut off. This startling realization does nothing to subdue the reality of the limb they experience, and may even intensify the sensations that define it. The first detailed description of this phenomenon was provided by the American Civil War surgeon, Silas Weir Mitchell (1871), who dubbed it the 'phantom limb.' The term so accurately captures the nature and subjective experience of the phenomenon that it has been adopted ever since. This entry reviews the nature of phantom limb pain and nonpainful phantom limb sensations which arise after limb amputation.

1. Phantom Limb Pain

For many amputees, a distressing problem is phantom limb pain. While the distinction between a painful and nonpainful phantom is not always clear-cut, it is useful to separate them for purposes of description, classification, and treatment. Painful phantom experiences vary in a number of dimensions. These include intensity, quality, location, frequency, and duration. Many amputees report a painful intensification of the tingling, 'pins and needles' or *paresthetic* quality (i.e., dysesthesias) that defines the nonpainful phantom limb. Some sufferers describe bouts of paroxysmal shooting pain that travel up and down the limb. Others report the phantom to be in a cramped or otherwise unnatural posture that gives rise to excruciating pain. Many amputees describe the pain in the phantom limb as indistinguishable from the pain they experienced in the limb before amputation. In still others, the phantom may be immobile or paralyzed so that attempts to move it generate pain. Finally, the phantom is often the seat of an intense burning pain, as if the hand or foot were held too close to an open flame. Frequently amputees suffer from several types of pain. Phantom limb pain, it appears, is equally prevalent among men and women, and is not related to side, level, age, or cause of amputation (Jensen and Nikolajsen 1999).

Surveys based on several thousand amputees reveal that between 78 percent and 85 percent of patients continue to experience significant amounts of phantom limb pain more than 25 years after amputation (Sherman et al. 1997). Prospective studies of the

incidence of phantom limb pain are consistent with these estimates (Jensen et al. 1985, Nikolajsen et al. 1997). Between 50 percent and 71 percent of amputees report phantom limb pain two years after amputation. There is a reduction in the frequency and duration of attacks over time. Equally striking is the low success rate of treatments for phantom limb pain: in the long term only 7 percent of patients are helped by the more than 50 types of therapy used to treat phantom limb pain (Sherman et al. 1997).

The different qualities of phantom limb pain, its high prevalence and incidence, and its low treatment success rate point to the complexity of the problem. Phantom limb pain is not a unitary syndrome, but a symptom class, with each class subserved by different etiologic mechanisms. There is not just one cause of phantom limb pain. The development and severity of this disorder are determined by the interaction of multiple factors that vary across individuals. These include ectopic activity from peripheral nerves that were transected during amputation and have formed a neuroma, contributions from the sympathetic nervous system, sensitization of central nervous system structures due to the effects of amputation, inflammation and ectopic activity, alterations in the functional organization of sensory nuclei, past experience with pain, and the immediate physiological, motivational, and psychological states of the individual.

1.1 Pain Memories in Phantom Limbs

A striking property of phantom limb pain is the presence of a pain that existed in a limb prior to its amputation (Katz and Melzack 1990). This class of phantom limb pain is characterized by the persistence or recurrence of a previous pain, has the same qualities of sensation, and is experienced in the same region of the limb as the pre-amputation pain. Cases studies of amputees have revealed pain 'memories' of painful diabetic foot ulcers, bedsores, gangrene, corns, blisters, ingrown toenails, cuts and deep tissue injuries, and damage to joints and bony structures. In addition, the phantom limb may assume the same painful posture as that of the real limb prior to amputation, especially if the arm or leg has been immobilized for a prolonged period.

The proportion of amputees who report similar pain before and after amputation may be as high as 79 percent (Katz and Melzack 1990), although according to prospective studies, the incidence is approximately 10 percent at one year after amputation (Jensen et al. 1985, Nikolajsen et al. 1997). Pain also persists in patients with deafferentation that does not involve amputation. In these conditions, the involved body part is still present but it is devoid of sensibility due to an interruption in the supply of sensory (afferent) information (i.e., deafferentation). Brachial plexus avulsions, in which the sensory nerve roots supplying the arm and hand are torn from the spinal cord, often

produce pain that is felt in the deafferented and anesthetic region. Similarly, patients with spinal cord injuries may complain of pain referred to body parts below the level of the transection, including, for example, a patient who continued to feel the pain of an ingrown toenail after a complete spinal cord break.

Painful and nonpainful sensations also persist or recur after surgical removal or deafferentation of body structures other than the limbs, such as breasts, teeth, and internal and special sense organs. Ulcer pain has been reported to persist after subtotal gastrectomy with removal of the ulcer. Patients have reported labor pain and menstrual cramps after total hysterectomy, rectal pain and hemorrhoids after removal of the rectum and anus, the burning pain of cystitis after complete removal of the bladder, and the pain of a severely ulcerated cornea after enucleation of an eye.

As noted above, not all phantom limb memories are of painful experiences. Nonpainful preamputation sensations do recur, but they are not as common and tend to include the sensation of objects that were once worn on the limb. These *superadded* sensations vary in complexity from such simple somatosensory qualities as the sensation of bandages that once wrapped a wound, a plaster cast, finger rings, and wristwatches to the multimodal, perceptually integrated phenomenon of a phantom foot clad in a sock and a shoe of specific type and color (Katz and Melzack 1990).

Taken together, these case reports and studies of amputees reveal that pain memories are not merely images or cognitive recollections; they are direct experiences of pain that resemble an earlier pain in location and quality. They are perceptually complex experiences that may even involve information from multiple sensory modalities including visual, olfactory, tactile, and motor components that had accompanied the original experience. The precise details of the experiences of pain involve localization, discrimination, affect, and evaluation—that is, all the dimensions of perceptual experience—and these properties are a function of integrated brain activity. It is likely that the outputs of sensitized spinal cells activate the neural structures in the brain that subserve memories of earlier events.

2. The Nonpainful Phantom Limb

The most salient property of the nonpainful phantom is the tingling, paresthetic quality, but sensations of temperature, posture, length, volume, and movement are also very common (Jensen and Nikolajsen 1999). The incidence of the nonpainful phantom is estimated to be between 80 percent and 100 percent. In fact, the presence of a nonpainful phantom is so common that it is considered a 'normal' sequela of amputation. It is standard practice to prepare the patient preoperatively for its subsequent occurrence.

The frequency with which a painless phantom occurs does not appear to be related to the extremity,

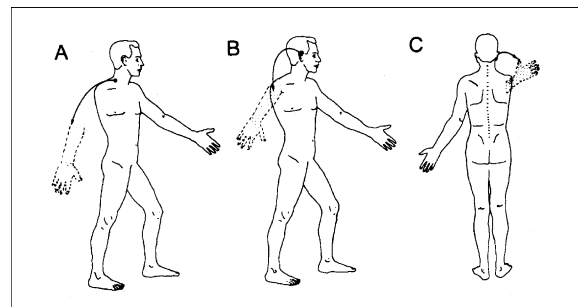


Figure 1

Illustrations of a patient who sustained a complete forequarter amputation of his right shoulder in a work accident and later developed phantom limb pain. The figurines depict the process of telescoping following a series of treatments with transcutaneous electrical nerve stimulation. Dots represent points where stimulation was applied. Arrows indicate the sites of referred phantom limb sensations

Source: Adapted from Katz and Melzack 1987, with permission

A The patient reported a phantom limb of normal length that consisted of only the lower arm and hand, with a gap between the shoulder stump and phantom elbow. During stimulation applied at a point above the clavicle, the patient reported that his phantom hand began to swell and become warm.

B In response to stimulation applied at a point on his right ear, he remarked that his phantom arm had 'telescoped' into the stump so that the elbow protruded from the stump.

C Later, stimulation of the same point resulted in a further retraction of the phantom, leaving only the wrist and hand attached to the stump

side, level of amputation, age, gender, or reason for amputation (Jensen and Nikolajsen 1999). There is considerably more variability in the estimates of the duration of the nonpainful phantom. In some amputees, the phantom may disappear within days or weeks of the amputation. In others, it remains unchanged for up to 60 years.

2.1 Fading, Telescoping, and Shrinking of the Phantom Limb

In many cases, the normal phantom undergoes a number of predictable changes within several years of amputation (Katz 2000). These include fading, telescoping, and shrinking. Immediately after amputation the phantom limb usually feels of normal length and size. Over time, the proximal portions begin to fade and soon disappear (Fig. 1). At this point, the phantom limb consists of only the extremity and major joints. These are felt at a normal distance from the body. The intervening parts have disappeared so that the phan-

tom hand or foot feels as if it were hanging, unattached, in empty space. In the case of an amputation performed at the shoulder, the phantom limb may consist of only the lower arm and hand, with a gap between the shoulder stump and the phantom elbow. In approximately one-third of amputees, a process of 'telescoping' begins. Among upper extremity amputees, the phantom is perceived gradually to approach the stump so that the hand is located in phenomenal space on a level with the elbow of the other arm. As the process continues, the gap between the stump and phantom continues to close; the amputee may find that the hand is protruding from, or attached to, the end of the stump. Later, the phantom hand may retract into the stump so that only the tips of the fingers jut out. In some cases, eventually even these may disappear completely and permanently into the stump.

There are times when a completely telescoped phantom will temporarily lengthen so that the hand or foot is once again felt at a normal distance from the stump. Less commonly, the converse also occurs; a phantom which is already telescoped may temporarily retract completely into the stump, or one which is of normal, or shortened length may become even shorter. The circumstances and stimuli which bring about these changes are extremely varied and range from peripheral stimulation of the stump and other regions of the body to more central influences.

The process of telescoping is sometimes accompanied by a marked shrinking of the phantom to the dimensions of a child's limb. This gradual diminution in perceived size or volume occurs in both upper- and lower-limb amputees. The extent of shrinking that occurs is variable. Adult amputees have reported the phantom to be the size of a doll's or baby's hand, a silver dollar, or even as small as a postage stamp (Katz 2000).

2.2 Cortical Maps and Phantom Limbs

Factors governing the fading, telescoping, and shrinking of the phantom are not well understood. The gradual fading of the phantom is thought to represent a perceptual correlate of the re-establishment of inhibitory control over hyperactive or spontaneously active cells which subserve the phantom. The gradual telescoping and shrinking have been hypothesized to depend upon input from intact peripheral fibers located in the stump. These experiences appear to be perceptual markers of the short- and long-term re-organizational changes that occur after amputation in central nervous system structures including the somatosensory cortex (Katz 2000, Merzenich et al. 1984). The perceived distance between the phantom and the stump may be a function of the distance separating their respective representations in the cortical map.

2.3 Psychophysical Correlates of Phantom Limb Sensations

Although a normal phantom occurs whenever nerve impulses from the periphery are blocked or otherwise removed, it is also true that direct stimulation of the amputation stump frequently exaggerates the tingling or paresthetic quality of sensation typical of the painless phantom limb. Careful questioning of amputees reveals that the nonpainful phantom limb is not perceived as a static phenomenon. The paresthetic quality of sensation, which defines the phantom limb percept, is in a constant state of flux, with changes occurring in intensity, body part, or both.

One mechanism that accounts for the paresthetic component of the phantom limb is a cycle of sympathetic-efferent somatic-afferent activity (Katz 2000). Changes in the intensity of phantom limb paresthesias reflect the joint activity of cholinergic (sudomotor) and noradrenergic (vasomotor) postganglionic sympathetic fibers on primary afferents located in the stump and stump neuromas. Release of acetylcholine and norepinephrine from postganglionic sympathetic fibers produces transient vasoconstriction and heightened skin conductance responses. In addition, neurotransmitter release onto apposing peripheral fibers trapped in stump neuromas increases primary afferent discharge. This information is transmitted rostrally, where it gives rise to referred phantom sensations upon reaching central structures subserving the amputated parts of the limb. The moment-to-moment fluctuations in the intensity of phantom limb paresthesias reported by many amputees may, in part, reflect a cycle of sympathetic-efferent somatic-afferent activity. Increases in the intensity of phantom limb paresthesias would follow bursts of sympathetic activity, and decreases would correspond to periods of relative sympathetic inactivity. If central sensitization has also developed, through prior injury, trauma during amputation, or peripheral inflammation, or if the sympathetic-sensory coupling involves nociceptors, the sensation may be one of dysesthesia.

2.4 Psychological and Emotional Processes Influence the Phantom Limb

The idea that emotional and psychological processes can cause pain traditionally has been tied to the notion of psychopathology. However, it is becoming increasingly clear that under certain circumstances pain may be triggered by these processes in psychologically healthy individuals as well. The aggravation or alleviation of pain referred to phantom body parts may be mediated in part by psychological processes that alter anxiety levels. Phantom breast pain after mastectomy is provoked by emotional distress in 6 percent of women three weeks after surgery, and in 29 percent one year later. Of lower-extremity amputees, 50 percent report that attacks of phantom limb pain are

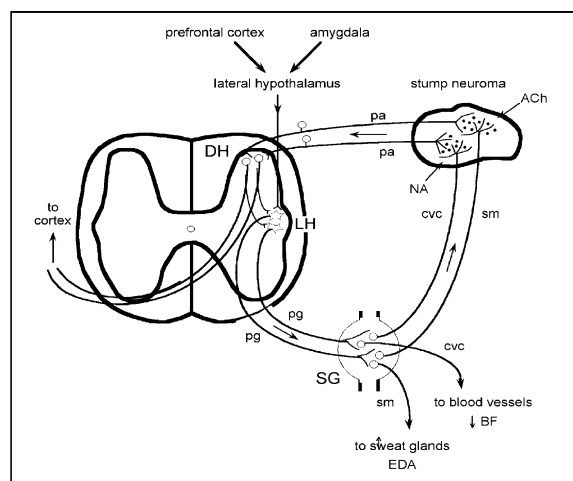


Figure 2

Schematic diagram illustrating a mechanism of sympathetically generated phantom limb paresthesias. Spontaneous sympathetic activity or excitatory inputs descending from cortex (e.g., due to the perception of a salient event, loud noise, thought, feeling, etc.) increase the discharge rate of preganglionic (pg) sympathetic neurons with cell bodies in the lateral horn (LH) of the spinal cord and terminals in the sympathetic ganglion (SG). These neurons excite postganglionic noradrenergic (NA) cutaneous vasoconstrictor (cvc) and cholinergic (ACh) sudomotor (sm) fibers that impinge on effector organs (vascular smooth muscle and sweat glands) in the stump and on sprouts from large diameter primary afferent (pa) fibers that have been trapped in a neuroma. The release of ACh and NA on effector organs results in increased electrodermal activity (EDA) and decreased blood flow (BF) to the stump. Release of these chemicals in the neuroma activates primary afferents that project to spinal cord dorsal horn (DH) cells subserving the amputated parts of the limb. These neurons, in turn, feed back to the preganglionic sympathetic neurons and project rostrally where the impulses contribute to the perception of phantom limb paresthesias. If DH cells have been sensitized due to injury, or nociceptive primary afferents are activated, then the perception may be one of dysesthesias

Source: Adapted from Fields 1987, with permission

triggered by emotional distress as long as seven years after amputation. A combination of progressive relaxation training and Electromyographic (EMG) bio-feedback of stump and forehead muscles produces significant reductions of phantom limb pain and anxiety, which are sustained for up to three years.

Psychological or emotional processes can bring about transient but profound alterations in the quality and intensity of phantom limb sensations. These

processes include concentration, distraction, relaxation, fright, forceful reminders of the events that led to amputation, and witnessing cruel and violent acts. One amputee, interviewed by the present writer, described his reaction to an accident involving his wife by reporting 'goose bumps and cold shivering down the phantom [leg]. It went through me. Everything emotional will get you that.' Another amputee stated, 'It's like everything I feel goes there—the good and the bad' (Katz 2000).

2.5 A Centrally-triggered Sympathetic-efferent Somatic-afferent Mechanism

The model schematically represented in Fig. 2 outlines a mechanism through which cognitive and affective processes associated with higher cortical and limbic centers may alter phantom limb sensations. The reciprocal connections between cortical, limbic, and lateral hypothalamic structures are well documented. The lateral hypothalamus is involved in the control and integration of neural activity associated with affectively charged behavior, and has direct projections to the lateral horn of the spinal cord. The intensity of phantom limb paresthesias and dysesthesias may thus be modulated by higher brain centers involved in cognitive and affective processes via a multisynaptic network of descending inputs, which impinges on preganglionic sympathetic neurons, producing diffuse peripheral autonomic discharge and activation of primary afferent fibers located in stump neuromas.

3. Summary and Conclusions

The phantom limb is not perceived as a static entity but as a frequently changing perceptual experience dependent upon the current sensory input, emotional state, and past experience of the individual amputee. Phantom limb phenomena range from simple, diffuse sensations of tingling to perceptually complex experiences of pains and lesions that originally were felt in the limb prior to amputation. While phantom pains and other sensations frequently are triggered by the perception of salient events, thoughts, and feelings, there is no evidence that the painful or painless phantom limb is a symptom of a psychological disorder. The sympathetic nervous system may provide an important link between higher brain centers involved in cognitive and affective processes and phantom limb sensations through its peripheral actions on primary afferents located in stump neuromas.

See also: Chronic Pain: Models and Treatment Approaches; Pain, Neural Basis of; Parietal Lobe; Somatosensation

Bibliography

- Fields H L 1987 *Pain*. McGraw-Hill, New York
- Jensen T S, Nikolajsen L 1999 Phantom pain and other phenomena after amputation. In: Wall P D, Melzack R (eds.) *Textbook of Pain*, 4th edn. Churchill Livingstone, Edinburgh, UK, pp. 799–814
- Jensen T S, Krebs B, Nielsen J, Rasmussen P 1985 Immediate and long-term phantom pain in amputees: Incidence, clinical characteristics and relationship to pre-amputation limb pain. *Pain* 21: 268–78
- Katz J 2000 Individual differences in the consciousness of phantom limbs. In: Kunzendorf R G, Wallace B (eds.) *Individual Differences in Conscious Experience*. John Benjamins, Amsterdam, pp. 45–97
- Katz J, Melzack R 1987 Referred sensations in chronic pain patients. *Pain* 28: 51–9
- Katz J, Melzack R 1990 Pain ‘memories’ in phantom limbs: Review and clinical observations. *Pain* 43: 319–36
- Merzenich M M, Nelson R J, Stryker M P, Cynader M S, Schoppmann A, Zook J M 1984 Somatosensory cortical map changes following digit amputation in adult monkeys. *Journal of Comparative Neurology* 224: 591–605
- Mitchell S W 1871 Phantom limbs. *Lippincott’s Magazine of Popular Literature and Science* 8: 563–9
- Nikolajsen L, Ilkjaer S, Krøner K, Christensen J H, Jensen T S 1997 The influence of preamputation pain on postamputation stump and phantom pain. *Pain* 72: 393–405
- Sherman R A, Devor M, Jones D E C, Katz J, Marbach J J 1997 *Phantom Pain*. Plenum Press, New York

J. Katz

Phenomenology in Human Science

Conceiving of phenomenology as a methodology of the human sciences imposes two types of constraint on the treatment of phenomenology. Since both types have formal as well as thematic implications they deserve a few introductory words of clarification.

1. The Asynchronism of Phenomenology and the Human Sciences

Relating phenomenology with the social and behavioral sciences involves a many-sided problem of *asynchrony*. When phenomenology emerged as a philosophical enterprise at the end of the nineteenth and the beginning of the twentieth century there were no social and behavioral sciences. At best there were currents at thought, hardly disciplines, that only very much later, after significant modifications and paradigmatic shifts, were categorized as social or behavioral sciences. There was, however, for instance psychology, known, criticized, adapted, or rejected by early phenomenologists, and those attitudes were reciprocated (Spiegelberg 1960). Above all, psy-

chology was not (and never became) a unitary science, but a set of competing (and changing) schools of thought. A similar variety and changeability holds for early sociology. Also phenomenology changed and has kept changing, sometimes with one and the same author, Husserl himself being the most notorious example (Embree et al. 1997).

Hence it happened that the psychology to which the ‘early Husserl’ referred was quite different from the ones the ‘middle’ or even the ‘late’ Husserl was confronted with. This variety increases considerably if different phenomenologists are considered, such as Husserl, Pfänder, Scheler, Heidegger, Merleau-Ponty, Sartre, and Schütz. Excepting Husserl, their names indicate that these philosophers were not only, at least not always, phenomenologists. If one defines, as it is done within the scope of this article, phenomenology as a methodological attitude the latter can be adopted for a specific purpose, but does not necessarily make its user a phenomenologist for good. Historically seen, ‘doing phenomenology’ (Spiegelberg 1960) is—for philosophers as well as human scientists—a much more appropriate characteristic than ‘being a phenomenologist.’ After all phenomenology has, according to Spiegelberg (1960), been a movement rather than a school. Variety has always been greater than unity.

Also in the sciences with which phenomenology got related there has been considerable variance. The early psychology, for example, of associationism and elementism had so little in common with gestalt psychology which, in turn, was so different from behaviorism or from modern cognitivism that for each of them the relationship with phenomenology has carefully to be redefined, in fact from both ends.

In sociology, whose relationship with phenomenology has been more intensive and durable, it is important to identify carefully the theoretical position with which phenomenology is related. More than is the case in other social sciences sociology provides approaches and specialties engendered or influenced by phenomenological thinking, such as ethno-methodology and conversation analysis and, above all, there is an academically established phenomenological sociology which, mainly thanks to the oeuvre of Schütz (1962/66), has been considered the prototype of phenomenology in the context of social sciences. ‘Phenomenological Psychology,’ however, as it was introduced by Husserl (1925/1977) will have to be disregarded in this article, as far as it is the name for a branch of transcendental philosophy. The prominent exception remains Gurwitsch’s (1964, 1966) creative synthesis of gestalt theoretical psychology and phenomenology.

At the end of the twentieth century we register many unilateral as well as mutual influences between phenomenology and the human sciences—the latter term, by the way, popularized by phenomenologists like Merleau-Ponty, Strasser, and Ricoeur. Among these sciences there are many of very recent origin. The

phenomenology with which they relate may be of any vintage (idealistic, realistic, existential, hermeneutic), sometimes treated in an eclectic manner. The responsibility, however, for the resulting phenomenological eclecticism rests with the human scientists.

To summarize the references to the first type of constraint: whoever is interested in the relationship between phenomenology and the human sciences must keep in mind its asynchronous historicity and multiplicity.

2. Phenomenology as Methodology

The second type of constraint is the limitation to *methodology*. On the one hand, this should be self-evident if one considers the number and variety of the sciences involved (Natanson 1973). An encyclopedia of phenomenology (Embree et al. 1997) lists phenomenological approaches and tendencies in about two dozen nonphilosophical disciplines, most of which belong to the human sciences. It is their inequality in terms of age, formalization, homogeneity, empiricism, and affinity with natural or cultural sciences that renders their comparability with phenomenology so complicated.

On the other hand, the limitation to methodology should not be too constraining; phenomenology itself is widely considered a specific methodological approach. Even if occasionally phenomenology is ontologically demarcated and referred to as the *systematic study of human consciousness or experience*, this is not much of a constraint if one accounts for *intentionality* with which phenomenologically (almost) all mental processes (perceiving, thinking, judging, remembering, feeling, desiring, willing) are endowed, namely, as meaning-something-as-something in being aware of it. Both this perspectival meaning-relatedness and the corresponding 'horizon of all horizons,' the *lifeworld*, in which 'things appear in terms of their experiential qualities, values, and uses' (Welton 1997), have become criteria of a phenomenological approach or orientation in the human sciences. The terms of 'approach' and 'orientation' are given preference to the totalizing (and hence overdrawn) designation of a phenomenological psychology, sociology, etc.

3. Performing the Epoché

Before these criteria can be exemplified a brief look into the central (Husserlian) method of *epoché* may be permitted although its proper field of application, transcendental phenomenology, is beyond the scope of this article. Yet the basic idea of *epoché* and of the related method of *reduction* provides the master-key to phenomenological methodology.

As agents in their everyday life human beings experience the world, including their own acting,

perceiving, feeling selves, as really existing. Husserl (1977) considered this universal belief in the actuality (*Wirklichkeit*) of the world the general thesis of the *natural attitude*. For the purpose of a phenomenological analysis, this belief has to be suspended (as far as possible). Suspension means neither negating nor doubting the existence of world and self. The belief, normally pervading all acts and states of consciousness, is merely 'bracketed,' that is, put out of operation. These metaphors are to signify the transition from the natural to the phenomenological attitude. By means of this transition world and self are transformed into, that is, reduced to, mere phenomena.

4. Criteria of a Phenomenological Approach

While the transcendental-phenomenological purpose of performing the *epoché* is of little or no interest in the human sciences and even phenomenologists like Merleau-Ponty (1962) have done their phenomenological analyses without performing *epoché*, the suspension of beliefs is an important methodological tool which is operative in what traditionally have been called the functions of a phenomenological approach in human science: (a) the critical, (b) the descriptive, and (c) the interpretive function. Although critique, description, and interpretation are distinct methods which for many purposes must be kept apart, under phenomenological premises they turn out to be inter-related.

4.1 The Critical Function

Given the fundamental postulate 'to the things themselves,' the primordial method is some kind of *epoché*. Critical is here the practice of disbelief with respect to all explicit and silent assumptions one has held so far about the matter to be studied, the theories, hypotheses, and methods, provided by the pertinent science, but also a self-critical avoidance of personal habits and predilections of thinking that may bias or distort the required 'fresh look' at the phenomenon in question. Equally critical the phenomenological observer has to be with respect to the language in which the unbiased observation has to be described. The theories that are to be suspended are often implied in the concepts that are intended to be descriptive, but actually are interpretive. A special difficulty arises here for the psychologist since the language for mental states and processes is predominantly metaphorical.

4.2 The Descriptive Function

As anticipated these critical caveats are also an introduction into the problem of phenomenological description in the human sciences. Although as old as

phenomenological methodology one caution has to be reiterated, since its disregard has turned out to be one of the most obstinate prejudices, namely that a phenomenological (-psychological) approach amounts to a method of introspection. The opposite is closer to the truth: phenomenological or 'intentional(ist)' description focuses on persons, selves, or others, as they are encountered in everyday situations. To clarify what this means a synthesis will be presented of conceptual-methodological recommendations that can be gathered from the writings of Husserl, Gurwitsch, Löwith, Merleau-Ponty, and from the phenomenological approaches in human science, mainly in psychology, psychiatry, and sociology.

The methodological foundation of phenomenological description was laid by Husserl who, in his 'principle of all principles,' demanded to describe things as they are experienced, strictly within the limits, but also to the full extent in which they are experienced. A merely consequent, but historically influential specification of this postulate was given by Gurwitsch who, in his doctoral dissertation of 1929 (a study of relations between phenomenology and gestalt theory) had adopted the *field concept* (see *Lewin, Kurt (1890–1947)*) and applied it to Husserl's concept of noema: a 'theme' in the center of what one is aware of is always experienced within a 'thematic field' and further 'marginal' objects of consciousness (Gurwitsch 1964). His study of the relationship between theme and thematic field explicates what Husserl since his 'Ideas' has developed as the *horizontal structure of consciousness*.

Closely connected was Gurwitsch's thematization of *perspectivity (avant la lettre)*: an object, perceived from different angles and distances, that is, from varying viewpoints, presents itself as identical in a system of noemata. Both conceptions, perspective, and horizon, have become essential guidelines for phenomenological description:

(a) Whatever is experienced (perceived, cognized, remembered, imagined, anticipated), is experienced from a given viewpoint (an epistemological insight going back to Leibniz) and as the same from varying (i.e., successive) viewpoints.

(b) Whatever is experienced from a given viewpoint is experienced in an horizontal structure that corresponds to and changes with the point of view or perspective. Horizon means, as the everyday usage suggests, both limitation and openness of experience. Moving one's 'standpoint' leads to a shift of horizons, which holds for physical, bodily, as well as mental locomotion.

(The fact that human beings are inevitably *en route* to the horizon without ever reaching it, only the places where it was, has been made a topic of phenomenological anthropology by van Peursen 1954.)

Both perspective and horizon have become interrelated structural elements of the phenomenological conception of *situation*. Whatever or whomever a

person experiences is encountered in a perspectively structured situation. The essential *situatedness of experience* entails that in everyday experience, there is no 'decontextualized' object (thema): human beings coexist with nonhuman objects and, inversely, objects, as utensils, refer to their users, as Heidegger (1962) has shown.

Karl Löwith, a student of Heidegger, was one of the first who, in his study of *The Individual in the Role of Fellow Human Being* (1928), developed a phenomenological (i.e., nonbehavioristic) concept of human behavior or comportment (*Verhalten*). He distinguishes between three ways of relating in behavior: (a) with others (fellow human beings), (b) with nonhuman objects (e.g., utensils), and (c) with oneself. He demonstrates that and how each of these intentional directions involves the other two. For example, relating with a fellow being involves not only objects belonging to him or her or to oneself, but also a modification, although not necessarily conscious, of one's self-concept.

Decades later, in a simplified version, the social psychologist T. M. Newcomb, who had developed a model of communicative acts, in which two persons communicate about something, argued that for social psychology there is no (purely) interpersonal relation in an environmental vacuum, nor any object-relation in a social vacuum.

Löwith's conception of the interdependence of the world of fellow human beings (*Mitwelt*), the world of things (*Umwelt*), and self, was further elucidated and made relevant for phenomenological description in Gurwitsch's study of the 'encounter of fellow human beings within the horizon of the natural surrounding world,' in which the world of utensils (*Zeug-Umwelt*) is a prototype of what Max Scheler—in critical opposition to Husserl's 'platonizing' phenomenology—had introduced as 'milieu' (from which Gurwitsch derived his (German) title of the *Milieuwelt*). Not unintentionally, Scheler (1973) had made use of the term 'milieu.' After its misuse in deterministic theories of descentence, Scheler reinstated milieu as a relational term: milieu is what has effects on an organism. For humans, only those objects constitute the milieu which correspond to the values of human attitudes.

'Milieu' came to be replaced by *Umwelt*. This term was introduced into biology by von Uexküll (1909) to signify the subjective environment that corresponds to the structure and state of an organism's 'inner world.' As sensory *Merkwelt* and as motor *Wirkwelt* Umwelt is species-specific. Uexküll himself generalized the concept to the meaningful ambient of persons. As the environment-as-experienced-and-acted-upon by human (and animal) beings *Umwelt* is quasisynonymous with the phenomenologically conceived situation. Beyond phenomenology, it has been adopted for *The Study of Relations in Public* by Erving Goffman (1971) and, within the 'dramaturgical model'

of social life, by Rom Harré (1979). In the context of gestalt theory K. Koffka (1935), for whom phenomenology meant 'as naive and full a description of direct experience as possible,' introduced a distinction between the 'geographical' and the 'behavioral' environment. The famous exemplification is the German legend of the man who during a snowstorm rode over the frozen Lake Constance but dropped dead when he learnt what he had just done. The 'geographical' environment was a lake, the 'behavioral' environment a snow-covered plain. In other words and phenomenologically, the horseman's behavior was 'a riding-over-a-plain, but not a riding-over-a-lake.' Koffka's example, although fictitious, once more illustrates the validity of the 'postulate of all postulates' and the full applicability of intentionality to behavior and environment.

Without regarding his own approach as phenomenological, Charles Taylor (1964) refers to Koffka (as well as to Merleau-Ponty), when he proposes the method of 'intentional description' for the analysis of 'intentional environments.'

The globalization of the environment as it is experienced and acted upon is to be seen in one of the major meanings of *lifeworld*, the world as the ensemble of what is taken for granted. Its structures have been analysed by Schütz and Luckmann (1973/1989).

4.3 From Description to Interpretation

Since the taken-for-grantedness is a significant, but normally latent feature of everyday situations its analysis goes beyond mere description. In a phenomenologically oriented psychology and psychiatry the *situated person*, conceived as body-subject in meaningful and spatiotemporally articulated environments, has since Merleau-Ponty (1962) and the Utrecht School (Kockelmans 1987) been in the focus of interest (Herzog 1992, Herzog and Graumann 1991). The situated person's intentional relatedness in its various manifestations of perceiving, acting, desiring, feeling, remembering, communicating, and sharing their situation with others, has become the topic and task of phenomenological studies in the human sciences. Phenomena, however, are, as Heidegger (1962, §7) argued, only partly and certainly not initially apparent, but hidden. Also what is taken for granted is hidden from attention. To uncover the implicit efficacy of the partly hidden, *heuristic categories* have been developed, some of which will be briefly indicated.

A subject in intentional relation to objects and/or other persons is the minimal constitutive structure of a situation. As has been argued before, the subject has a place from which the situation is perceived or acted upon which, in turn, implies that the subject is or has a body occupying that place (as viewpoint) at a given time and for a certain duration and at a distance from whatever or whomever the body-subject experiences.

Following the implications of these few situational features will reveal a host of possible conditions or determinants of the experiences to be studied. To name a few: the body-subject *per se* is of a given gender, age, state of health, mood, civil or social status, group or class or ethnic membership, plus any of a set of psychological, for example, personality characteristics, etc. Some of these features will be relevant for the experience or behavior to be studied, some will not. But none of them, however inconspicuous or hidden, must *a priori* be considered irrelevant or even 'taken for granted.'

The same careful attention is to be paid to the other elements of the situation: its spatiotemporal articulation, its historicity, and last, but not least, its sociality and communicative structure. Since the latter term covers a wide range from spoken and written language to nonverbal communication and silence, the categories developed in philosophies and the techniques provided by the various sciences of language and signs may help to identify and to elucidate the variety and the intricacies of situated discourse and communication.

5. Concluding Comments

While it is true that many researchers, mainly those from an experimental tradition, shun the time, the effort, and the expense necessarily connected with phenomenological description and hermeneutic interpretation, it is equally evident that phenomenological methodology has been established firmly in social and behavioral sciences with a long-standing exchange with phenomenology (like psychiatry). But this methodology is also gaining a foothold in sciences without this tradition (like behavioral geography, ecology, political science, and even artificial intelligence). One may speculate whether the increasing interest in phenomenological methodology is a consequence of the generally growing realization that many of the traditional natural sciences, at least, to the extent that human beings are involved in natural and technical processes, are also human sciences.

See also: Husserl, Edmund (1859–1938); Phenomenology in Sociology; Phenomenology: Philosophical Aspects

Bibliography

- Embree L, Behnke E A, Carr D, Evans J C, Huertas-Jourda J, Kockelmans J J, McKenna W R, Mickunas A, Mohanty J N, Seebom T M, Zaner R M (eds.) 1997 *Encyclopedia of Phenomenology*. Kluwer, Dordrecht, The Netherlands
Gurwitsch A 1964 *The Field of Consciousness*. Duquesne University Press, Pittsburgh, PA

- Gurwitsch A 1966 *Studies in Phenomenology and Psychology*. Northwestern University Press, Evanston, IL
- Gurwitsch A 1979 *Human Encounters in the Social World* (ed. Métraux A, trans. Kersten F). Duquesne University Press, Pittsburgh, PA
- Heidegger M 1962 *Being and Time* (trans. Macquarrie J, Robinson E). Harper & Row, New York
- Herzog M 1992 *Phänomenologische Psychologie—Grundlagen und Entwicklung* (Phenomenological Psychology—Foundations and Development). Asanger, Heidelberg, Germany
- Herzog M, Graumann C F (eds.) 1991 *Sinn und Erfahrung—Phänomenologische Methoden in den Humanwissenschaften (Meaning and Experience—Phenomenological Methods in the Human Sciences)*. Asanger, Heidelberg, Germany
- Husserl E 1977 *Phenomenological Psychology: Lectures Summer Semester 1925* (trans. Sconlon J). Nijhoff, The Hague
- Kockelmans J J (ed.) 1987 *Phenomenological Psychology: The Dutch School*. Nijhoff, Dordrecht, The Netherlands
- Merleau-Ponty M 1962 *The Phenomenology of Perception* (trans. Smith C). Humanistic Press, New York
- Merleau-Ponty M 1964 Phenomenology and the sciences of man. In: Edie J M (ed.) (trans. Wild J) *The Primacy of Perception*. Northwestern University Press, Evanston, IL, pp. 43–96
- Natanson M (ed.) 1973 *Phenomenology and the Social Sciences*, 2 vols. Northwestern University Press, Evanston, IL
- Ricoeur P 1981 *Hermeneutics and the Human Sciences* (trans. Thompson J B). Cambridge University Press, Cambridge, UK
- Scheler M 1973 *Formalism in Ethics and Non-Formal Ethics of Values: A New Attempt Toward the Foundation of an Ethic Personalism*. (Trans. Frings M S, Funk R L) Northwestern University Press, Evanston, IL
- Schütz A 1962/1966 *Collected Papers*, 3 vols. Nijhoff, The Hague
- Schütz A, Luckmann T 1973/1989 *The Structure of the Life-world*, 2 vol. (trans. Zaner R, Engelhardt T Jr). Northwestern University Press, Evanston, IL
- Spiegelberg H 1960 *The Phenomenological Movement*, 2 vols. Nijhoff, The Hague
- Strasser S 1963 *Phenomenology and the Human Sciences*. Duquesne University Press, Pittsburgh, PA
- van Peursen C A 1954 L'horizon. *Situation I*: 204–34
- von Uexküll J 1909 *Umwelt und Innenwelt der Tiere*. [Environment and the Inner Worlds of Animals.] Springer, Berlin
- Welton D 1997 World. In: Embree L et al (eds.) *Encyclopedia of Phenomenology*. Kluwer, Dordrecht, The Netherlands, pp. 736–43

C. F. Graumann

Phenomenology in Sociology

1. The Origins and Scope of Phenomenological Sociology

Phenomenological sociology is the prescientific study of social life and the process by which humans interpret, experience, and understand their individual and collective realities. The work of the social philosopher

and sociologist Alfred Schütz (1899–1959) provides the most important foundation for phenomenological sociology. Framed in the general atmosphere of the debate between scientific and antiscientific movements that arose in the late nineteenth century, phenomenology places the social sciences in the context of everyday life (Thomason 1982). Strongly influenced by Henri Bergson, Edmond Husserl and William James, Schütz argues that a prescientific understanding of everyday life provides the only means by which a science of society is possible. 'The sciences that would interpret and explain human action must begin with a description of the foundational structures of what is prescientific, the reality which seems self-evident to men remaining within the natural attitude. This reality is the everyday life-world' (Schütz and Luckmann 1973, p. 3).

Just as the natural scientist must first understand the composition and interactions of subatomic particles in order to understand the nature of chemical reactions, the sociologist, Schütz argues, must understand the common-sense world. Unlike the subatomic world, however, the world of everyday life is permeated with the understandings given by those who constitute it. The task of the phenomenological sociologist is to understand how people make sense of their own lives. According to the phenomenologist, these subjective sense-making activities in everyday life are based on taken-for-granted knowledge passed down from generation to generation. In order to understand scientifically these subjective social actions the scientist must replace common sense explanations with objective scientific constructs derived from social theory (see *Theory: Sociological*). The phenomenologist seeks to understand social action in its own subjective terms, yet to describe this world scientifically using the tools of an objective science of society. The foundation of all social science, according to the phenomenologist must be the life-world.

2. The Life-world

Schütz states that the life-world (*Lebenswelt*) can be understood as 'that province of reality which the wide-awake and normal adult simply takes for granted as common sense' (Schütz and Luckmann 1973, p. 3). The everyday life-world provides us with a sense of the 'real.' It is through our position in, and experience of the life-world that we are social beings engaged and affected by the social and natural worlds. The most important characteristic of the life-world, according to Schütz, is that it is taken-for-granted. By this he means that individuals apprehend their worlds and its problems as self-evidently real—'that's just the way it is.' As a result, most individuals, most of the time, give little thought to the 'true' nature of the world around them. 'It is the unquestioned givenness of the life-world for its denizens, including those whose business

it is, in professional terms, to analyze problems of the social world, that phenomenology finds the ground of the social scientist's activity' (Natanson 1973, p. 40).

Taken-for-grantedness arises out of the typification of the phenomenal world. That is, our perceptions are categorized from a shared stock of knowledge as 'this' or 'that' type of thing. Our typifications, however, are only of an approximate nature. Such categories are held only until further notice. If contravened by future experiences, typifications must either be abandoned or reformulated. Taken-for-grantedness is further enabled through the use of time tested recipes for social action. Of the unlimited realm of potential social action open to individuals, most potential actions are circumscribed by a taken-for-granted sense of what is possible and not possible in such typical situations.

The universe of potential recipes for social action is also part of the social stock of knowledge from which typifications are drawn. An important point to be made is that the social stock of knowledge complete with its typifications and recipes for social action is pragmatic in nature. We simply tend to do what works and to avoid what does not work. Through typification and the use of recipes for social action the world becomes unproblematic and a matter of common sense.

3. Phenomenology and Science

A phenomenological study of the social world is one that addresses how humans experience the life-world. In contemporary sociology, phenomenological research has largely been identified with any research addressing the subjective perspectives of social actors. This, however, is a misunderstanding. Properly understood, phenomenology is prescientific, an attempt to ground the social sciences in human experience, a subjective appreciation of the human condition (Embree 1988, p. 270). Schütz (1962a) argues that if we are to understand social reality all social scientific constructs must not only be linked to what people experience, but how they experience it. Scientific understandings must be connected to the experiential process through which people in their everyday lives actually experience the world. He states, 'correctly understood, the postulate of subjective interpretation as applied to economics as well as to all other social sciences means merely that we always can—and for certain purposes must—refer to the activities of the subjects within the social world and their interpretation by the actors in terms of systems of projects, available means, motives, relevances, and so on' (Schütz 1962a, p. 35).

Phenomenological sociology has sometimes been inappropriately labeled as 'anti-scientific.' This misconception stems from the anti-scientific intellectual climate from which phenomenology arose (Thomason 1982). Sensitive to these movements, Schütz indeed

understood the dehumanizing possibilities of science. His answer to this problem, however, was not to abandon science, but rather to ground the social science in the motives and realities of everyday life. Phenomenology does not necessarily attempt to replace or debunk science, but rather to provide a foundation for empirical investigations.

4. The Contribution of Phenomenology in Sociology

The possibility of a phenomenological sociology has been partially realized in two current sociological traditions: social constructionism and ethnomethodology. While both approaches address the life-world and its subjective realities, neither has fully realized the potential Schütz saw for a truly phenomenological sociology. That is, both have addressed the subjective nature of the social world, but neither has been the foundation of an objective science of society.

4.1 Social Constructionism

Phenomenological interest in the mediated and negotiated nature of all knowledge gave rise to social constructionism in sociology. Social constructionism, simply stated, is the study of the way in which people agree upon and define what reality is. Perhaps the most important work in this field is *The Social Construction of Reality: A Treatise in the Sociology of Knowledge* (Berger and Luckmann 1966) (for example, see *Sociology, Epistemology of*).

At the heart of social constructionism can be found one of phenomenology's most important concepts, the epoch. The epoch amounts to a radical skepticism not unlike Descartes' method, where belief in the 'realness' of the world is suspended. Schütz (1962c, p. 102) however, claims that Descartes was 'not radical enough,' that Descartes conclusion '*cogito, ergo sum*' ['I think therefore I am'] failed to doubt and analyze the thought itself. By methodologically setting aside or 'bracketing' the everyday world and even the nature of thought itself, the phenomenologist seeks to 'go beyond the natural attitude of man living within the world he accepts, be it reality or mere appearance' (Schütz 1962c, p. 104). The purpose of the phenomenological reduction (epoch) is to strip through the world of belief to the realm of consciousness and to examine the resulting contents.

Thomason (1982) points out that social constructionism in sociology stemming from the epoch can best be described as a methodological constructionism. Unlike ontological constructionism, methodological constructionism does not question the existence of a real world, but rather suggests that what we know of the world is always mediated and indirect. Methodological constructionism most closely resembles

Schützian phenomenology. Schütz states 'I am afraid I do not exactly know what reality is, and my only comfort in this unpleasant situation is that I share my ignorance with the greatest philosophers of all time' (Schütz 1964, p. 88, Thomason 1982, p. 4). That is, the methodological constructionist examines the agreements people make about the world, but does not question that a real world exists. Social constructionism remains an important field of inquiry in the natural and social sciences, but is of particular importance in sociology.

4.2 Ethnomethodology

Ethnomethodology is the second intellectual tradition linked to phenomenology. Ethnomethodology is also connected to the epoch, but most importantly to Schütz's commitment to the importance of the everyday life-world. However, unlike the prescientific and proscientific work of phenomenology, ethnomethodology as formulated by Garfinkel (1963) represents a radical break from the traditional models of social science with which Schütz had once tried to reconcile (Lynch 1988, p. 93). Ethnomethodology seeks to understand the method by which individuals construct, negotiate, and agree upon reality, but questions the possibility of an objective science of the subjective human condition. As a radically subjective pursuit, ethnomethodology falls short of the objective science of the life-world Schütz envisioned. Concerning such radically subjective endeavors Schütz (1962b, p. 52) maintains 'a method which would require that the individual scientific observer identify himself with the social agent observed in order to understand the motives of the later, or a method which would refer the selection of the facts observed and their interpretation to the private and subjective image in the mind of this particular observer, would merely lead to an uncontrollable private and subjective image in the mind of this particular student of human affairs, but never to a scientific theory.' While ethnomethodology remains an important influence in sociology, as currently formulated it falls short of the phenomenological sociology Schütz envisioned.

Without question, phenomenology has had a major impact upon modern sociology. Social constructionism and ethnomethodology each display a commitment to the epoch and the fundamental importance of the life-world, and therefore can directly be traced to phenomenological thinking. Both methods of analysis remain viable sociological traditions, and will no doubt continue to inform social research.

See also: Constructivism/Constructionism: Methodology; Ethnomethodology: General; Hermeneutics, History of; Hermeneutics, Including Critical Theory; Interactionism: Symbolic; Macrosociology–Microsociology; Phenomenology: Philosophical Aspects;

Schütz, Alfred (1899–1959); Social Constructivism; Social Psychology: Sociological; Sociology, History of; Sociology: Overview; Weber, Max (1864–1920)

Bibliography

- Berger P L, Luckmann T 1966 *The Social Construction of Reality: A Treatise in the Sociology of Knowledge*. Doubleday, Garden City, NY
- Embree L 1988 Schutz on science. In: Embree L (ed.) *Worldly Phenomenology: The Continuing Influence of Alfred Schutz on North American Social Science*. University Press of America, Washington, DC, pp. 251–74
- Garfinkel H 1963 *Studies in Ethnomethodology*. Prentice-Hall, Englewood Cliffs, NJ
- Lynch M 1988 Alfred Schutz and the sociology of science. In: Embree L (ed.) *Worldly Phenomenology: The Continuing Influence of Alfred Schutz on North American Social Science*. University Press of America, Washington, DC, pp. 71–100
- Natanson M 1973 Introduction. In: Natanson M (ed.) *Phenomenology and the Social Sciences*. Northwestern University Press, Evanston, IL, Vol. 1, pp. 3–46
- Schütz A 1962a Common-sense and scientific interpretation of human action. In: Natanson M (ed.) *The Problem of Social Reality: Collected Papers Volume One*. Martinus Nijhoff, Boston, pp. 3–47
- Schütz A 1962b Concept and theory formation in the social sciences. In: Natanson M (ed.) *The Problem of Social Reality: Collected Papers Volume One*. Martinus Nijhoff, Boston, pp. 48–66
- Schütz A 1962c Some leading concepts of phenomenology. In: Natanson M (ed.) *The Problem of Social Reality: Collected Papers Volume One*. Martinus Nijhoff, Boston, pp. 99–117
- Schütz A 1964 The problem of rationality in the social world. In: Brodersen A (ed.) *Collected Papers II: Studies in Social Theory*. Martinus Nijhoff, The Hague, The Netherlands, pp. 64–90
- Schütz A, Luckmann T 1973 *The Structures of the Life-World, Vol. II*. Northwestern University Press, Evanston, IL
- Thomason B C 1982 *Making Sense of Reification: Alfred Schutz and Constructionist Theory*. Humanities Press, London

J. Williams

Phenomenology: Philosophical Aspects

Phenomenology has been one of the most influential twentieth century philosophical traditions. It began with the appearance of Husserl's *Logical Investigations* in 1900–01 and of Pfänder's *Phenomenology of Willing: A Psychological Analysis* in 1900, and continued with the work of philosophers such as Johannes Daubert, Adolph Reinach, Moritz Geiger, Max Scheler, Edith Stein, Gerda Walther, Roman Ingarden, and Wilhelm Schapp. These phenomenologists were active in Munich and Göttingen and the movement to which they belonged is often called 'realist phenomenology.' Throughout the century

phenomenology has been the object of a variety of increasingly far-reaching transformations: by Husserl himself, whose ideas in 1913 marked a turn towards an idealist and transcendental form of phenomenology and then by Heidegger, Sartre, and Derrida, among many others. Among phenomenologists who remained faithful to realist phenomenology were Ortega y Gasset, in Spain, and Felix Kaufmann and Alfred Schütz in Vienna, where Brentano's teaching at the end of the nineteenth century had launched the project of an exact descriptive psychology and caught the imagination of Austrian philosophers such as Meinong, Ehrenfels, Twardowski, and Husserl. In all its forms phenomenology has continued to influence sociology and the philosophy of the social world.

Phenomenology was initially conceived of as the careful description of the mind—act phenomenology—and its objects—object phenomenology. Successful description, it was thought, would lead to an account of the main types of mental phenomena, the main types of their objects, and of the necessary and often *a priori* interconnections between these. In Husserl's *Logical Investigations* knowledge, verification and logical structures are described as though they could in principle be instantiated by a solitary subject or even by solitary mental acts. But phenomenologists quickly began to describe the social aspects of mental phenomena and to analyze the nature of social objects.

Among the most important results of the early phenomenologists are their accounts of awareness and understanding of other minds, motives, social acts, collective intentionality, the main forms of social life, and the nature of social objects. These contributions to the phenomenology of social phenomena were also employed in discussions of social and political philosophy and in the creation by Scheler of the sociology of knowledge (for partial overviews see Toulemont 1962, Theunissen 1977).

1. Empathy, Understanding, and Motivation

In what does Mary's awareness of Sam's sadness consist? Phenomenologists of nearly all stripes attempted to defend two claims about such awareness—sometimes called 'Einfühlung' or 'empathy'—which also, they thought, hold of awareness of physical objects. Awareness of other minds can be, and often is, direct and perceptual. Together the two theories constitute direct realism about other minds. Critical realism about our awareness of physical objects is the claim that we do not directly see things but infer from sensory data to their existence. Similarly, critical realism about other minds is the theory that we are not directly aware of other people or their mental states but infer to these on the basis of sensory data. On the phenomenological view, Mary can directly see Sam and his sadness and her perception need involve no

judgement or conceptualization and thus no inference. This view rests on the arguments given by early phenomenologists to show—against Kant and Helmholtz—that visual perception of things and events need involve neither inference nor concepts.

Traditionally the most popular variant of critical realism about other minds has appealed to reasoning from analogy. On the basis of observed analogies between her own behavior and mental states and observation of Sam, Mary is capable of attributing sadness to Sam. The locus classicus for criticisms of accounts of this kind is Scheler's 1913 appendix to his *On the Phenomenology and Theory of Feelings of Sympathy, 'On the other I.'* Scheler's own view contains the claim that Mary's awareness of Sam's sadness is as direct and as subject to error as her awareness of her own mental states (see on this Schütz's papers on Scheler in Schütz 1970, 1973). Less ambitious and more sober defenses of direct realism were provided by Edith Stein (1917/1980), who gives a subtle account of the relation of expression between mental states, on the one hand, and behavior and gestures, on the other hand; by Husserl (1973), who argues that Mary sees Sam's body as similar to her own body without judging this and thus without any reasoning from analogy; and by Schütz (1932). Phenomenological accounts of empathy also discuss the claim that some types of awareness of the feelings of another person involve make-believe feelings. Of the Gestalt psychologists influenced by phenomenology, Köhler (1947, Chap. 8) and Koffka (1946, Chap. 1) retain the claim that awareness of other minds need involve no judgement or inference but endorse critical realism. Bühler (1927, §9) defends both parts of the direct realism of the phenomenologists but within a naturalist and biological perspective that takes into account the dynamics of understanding other minds and the role of both steering mechanisms and criteria in bringing about 'mental contact.'

No account of empathy is complete without an account of what it is to understand other people. From the beginning, phenomenology was concerned with the analysis of the nature of motives and motivation and understanding thereof. In particular, the task was seen to be to provide an account of the relation between the internal causation of mental states and behavior, on the one hand, and motivation, on the other hand. The relation of motivation or justification was held to involve a number of distinct possible moments: the optimal case, where a mental state or disposition such as an intention, does indeed justify an action; conflict among motivations; the case where a mental state, for example, an emotion, is not justified by the subject's beliefs or perceptions. Many views were discussed and endorsed by different phenomenologists: what motivates and what is motivated are never the possible terms of a causal relation, for example, because they are ideal propositional contents; what motivates and what is motivated can be the

terms of a causal relation because they are concrete instantiations of ideal contents; the relations of causality and of motivation are always distinct but may have the same terms; there is a relation of superposition between motivation and causal relations; relations of motivation are causal relations but mental causality differs from physical causality. Finally, it was argued that motivational justification differs from both inductive and deductive justification, in particular because it often connects terms that cannot stand in the last two relations. Husserl himself thought that it was a 'basic law' of all psychological phenomena that they could stand in relations of motivation (Pfänder 1900, Husserl 1913/1969, Stein 1917, 1980).

In his 1932 *Der sinnhafte Ausbau der sozialen Welt: Eine Einleitung in die verstehende Soziologie* [The Phenomenology of the Social World. An Introduction to 'verstehende' Sociology], Alfred Schütz made use of the phenomenology of temporal experience in his analysis of motivation and understanding. 'In-order-to motives' explain how actions come about, genuine 'because-motives' explain by reference to preceding experiences how a practical project comes about. The present in which Sam is aware of his own motives, projects, and the states of affairs which are the objects of projects, provides him with a temporal perspective which is not available to Mary. This does not, Schütz thinks, amount to a denial that Sam's motivation can be grasped by Mary, only to the claim that her temporal perspective on his motivation and behavior must be segmented and discontinuous (Schütz 1932/1974, see also Grathoff 1995).

A further step towards accounting for the connection between understanding and temporal perspectives was taken by Wilhelm Schapp in his 1953/1976 book *In Geschichten verstrickt* (*Entangled in Stories*) in which such traditional phenomenological categories as action, its horizons, intentionality, and states of affairs are reworked in narratological terms.

2. Social Acts

Promises, Hume said, are the cement of society. But what is a promise? Reinach's 1913/1989 reply to this question is the most impressive piece of philosophical analysis ever penned by a phenomenologist (see Mulligan 1987). Like orders, requests, questions, and answers, promises are social acts. Such acts are typically addressed to someone, grasped by another person and are performed in the very act of speaking. They bring about changes. Thus a promise creates both an obligation and a claim. Actions of different sorts bring about states of affairs that satisfy or fulfil social acts; execution of the behavior commanded satisfies a command. Reinach warns against possible misunderstandings of the category of social acts due to dragging in the ideas to which we are accustomed. A command is neither a purely external action, nor is it

a purely inner experience, nor is it the announcing to others of such an experience (Reinach 1983, p. 19).

Reinach's account of the variety of social acts and of their effects is only the first step in his demonstration of their role in social intercourse. The second step is provided by his analysis of six possible modifications of social acts. In his analyses of such solitary mental acts as seeing or judging, Husserl had already noted that such acts could be modified. Thus one modification of seeing is visual imagining, one modification of judging is supposing. Similarly, argues Reinach, each type of social act can be modified in various ways. A promise may be a sham or pseudo promise, the promisor does not really intend to do what he promises to do. A command may be a conditional command: 'In the event that such and such occurs, I order you to do such and such.' The social act of informing someone that something is the case, however, is not subject to this sort of modification. A conditional social act is to be distinguished from a social act that has conditional content, for example a promise to do one thing if some other thing occurs. A social act may be performed by more than one person.

[We] have here to do with the case where each of the persons performs the act 'in union' with the others, where each knows of the participation of the others, lets the others participate and participates himself, we have one single act which is performed by two or more persons together (Reinach 1983, p. 24).

Similarly, a social act may be addressed to more than one person. Finally, a command or a promise, for example, may be performed in the name of another person; there are representative social acts.

3. Collective Intentionality

The intentionality of acts and states, their relation to an object or state of affairs, is always the intentionality of one subject. But whereas the intentionality of solitary acts such as judging, inferences, or seeings typically involves only a more or less explicit awareness of the self—if I judge that it is raining, I am aware that I am judging, although this is not part of what I judge—there are acts, states, and activities which involve a more or less explicit awareness of others whether these are presented as 'you,' 'she,' 'they,' 'one' or 'us.' This phenomenon, nicely expressed in German by conjoining 'Mit' (with) and many a psychological verb, noun, or adjective (e.g., 'Mitschuld' collective guilt), is now called 'collective intentionality.' One example of the phenomenon, as we have seen, is furnished by joint social acts. Reinach also points out that in the case of actions that are not social acts it is possible to speak of several performing subjects of one and the same action. There is a way of acting 'in union.' The criminal law's concept of 'complicity,' as it seems to us, has to base itself on this, and such collective actions are also important for public law,

administrative law, and international law (Reinach 1983, pp. 24–5).

But the starting point for an account of collective intentionality is not ‘Mithandeln’ (joint action, co-action) but the analysis of the mental acts and states involving affective participation such as ‘Mitgefühl,’ fellow-feeling or sympathy, those involving shared beliefs and those involving trust, confidence, or belief in someone or something (see Scheler 1966, IV B 4, Reiner 1934, pts. 1 and 2).

Scheler’s 1913 account of sympathy distinguishes between identification of emotional contagion, vicarious feeling, feeling something together with someone (‘*Mit-einanderfühlen*’), fellow-feeling (‘*Mitgefühl*’), benevolence or love of humanity, and personal love, and argues that these stand in different sorts of relation of dependence or foundation to each other (see also Walther 1922).

The phenomenologists made three important claims about collective intentionality. First, although for every type of collective intentionality implicit awareness of feeling, believing, willing, or acting together with others is to be distinguished from thoughts containing the concepts expressed by the different personal pronouns, such concepts depend on the relevant types of implicit awareness. Representations employing ‘we’ and ‘they’ presuppose presentations which help to fix the reference of our uses of the personal pronouns and cognate forms as in ‘Cosa Nostra,’ ‘Sinn Fein,’ ‘Mare Nostrum,’ and ‘Pater Noster.’ Second, collective intentionality cannot be reduced to mutual belief or knowledge. There is, Scheler claims, a type of ‘co-experiencing something’ which cannot be understood by saying that A experiences something that is experienced by B, and that both, in addition, know of their experiencing it (Scheler 1913/1973, p. 526).

Finally, the sense or feeling of togetherness may be actual or habitual, in which case it constitutes a type of habitual background (see Walther 1922). The notion of background, like that of horizon, was first explored within phenomenology in connection with perceptual experience.

4. Emotions and Explanation: Ressentiment

Of all the psychological phenomena analyzed by phenomenologists, emotions and related phenomena enjoyed by far the most attention. One reason for this is the conviction that a philosophy of mind which concentrates exclusively on beliefs, inferences, and desires can only scratch the mind’s surface and that explanation and understanding within the human sciences will suffer from any such limitation. This conviction in turn relies on the claim that emotional dispositions, tendencies, long-standing states and sentiments, both those directed towards the past and those directed towards the future, determine desires

and preferences. In order to mark the importance of emotions Scheler emphasizes that mind or spirit (‘*Geist*’) comprehends not only cognitive phenomena but also the will and affective phenomena, an emphasis also prominent in the fiction and essays of the Austrian thinker Robert Musil.

One important illustration of these claims is provided by Scheler’s analysis of resentment and subsequent interest in the topic. Building on suggestions of Nietzsche and the historical sociologist Sombart, Scheler takes resentment to be a long-lasting psychological attitude that comes into being when emotions and sentiments such as envy, rancor, revenge, resentment, and spite are repeatedly bottled up and lead to ‘value-illusions’ and false evaluations. In particular, it is a felt impotence or incapacity to realize or live up to certain values that leads the woman of resentment to false evaluations. In the (slightly modified) example from La Fontaine’s fable, the fox initially sets a high value on the grapes and changes its mind about their value solely as a result of its inability to obtain them. Scheler’s account relies on his view that value judgments are true or false—‘cognitivism.’ It is indeed essential to the mechanism described that it involves a re-evaluation, and not merely one or another type of change of preference. But it is not necessary to assume cognitivism. A weaker version of the account would simply have it that resentment involves a transition from an attribution of value to an attribution of disvalue (or vice versa) that is caused in the way described by Scheler and is thus, in the absence of any other justification, unjustified, unmotivated, or irrational. The central historical example discussed by Scheler is Nietzsche’s claim that early Christianity was characterized by pervasive resentment. One of the many other examples he discusses is the claim that the disinterested disposition to assist in the punishment of criminals and a certain sort of moral indignation together with the relevant evaluations emerge from lower-middle class envy. This claim was the object of extensive empirical investigation by the Danish sociologist Svend Ranulf (1930, 1938), who castigates Scheler’s almost complete disregard of the canons of empirical inquiry.

5. Types of Coexistence

Scheler’s taxonomy of forms of social coexistence draws on distinctions to be found in the classical writings of Tönnies, Weber, and Simmel and on his own philosophy of mind and ethical psychology. In particular, four types of coexistence are distinguished with reference to different types of collective intentionality: masses, communities, societies, and super-ordinate entities such as nation-states, state-nations, empires such as the Belgian Empire, and the Catholic Church. A mass is characterized by affective contagion, its members do not behave as autonomous

persons. The community—families, tribes, clans, peoples—displays such forms of collective intentionality as sympathy, trust, piety, loyalty, and collective responsibility rather than individual responsibility. A community forms a milieu or niche within which certain shared values are prominent. Empathy and understanding are there, typically automatic and unreflective. Philosophers have often assumed that the choice between the two accounts of empathy distinguished above is exclusive: either awareness of other minds is always immediate or it is always a process involving inference, for example analogical reasoning. Scheler rejects the assumption and argues that although the former account fits communities, the latter is typical of societal relations. Within societies self-interest, mistrust, and a lack of solidarity go together with the predominance of the artificial relations engendered by social acts such as promising and contracts which connect autonomous, individually responsible persons. The relations that constitute a community, unlike those created by promises and contracts, have no determinate temporal duration or terminus. In his account of such ‘collective persons’ as the nation—state—an account that he only sporadically attempts to make consistent with the requirements of ontological individualism—these are said combine certain traits of communities and societies. They resemble communities in that they are not typically held to enjoy a determinate duration and in the affective importance their members attach to the past. But like the autonomous persons standing in societal relations they enjoy personal responsibility. Finally, each form of coexistence is characterized by different values: the vital values of the community—its life, survival, flourishing, and death—contrast both with the values of civilization—utility, agreeableness, sociability—of society and with the spiritual values of nation—states and state—nations—culture, law, language. (On the relation between the account of values given by Scheler and by the founders of sociology, see Boudon 2000.)

These forms of coexistence are related, Scheler thinks, in the following ways. Societies are founded on communities, states on communities, and societal unities. Thus the societal relations created by promises and contracts rely on the fact that the persons standing in such relations have some familiarity with the phenomenon of trust characteristic of communities (see Scheler 1966, VI, B 4, Scheler 1963, 334ff., Stein 1970).

6. Sociology of Knowledge

Scheler’s extensive writings on sociology in the 1920’s, in particular his *Problems of a Sociology of Knowledge* and other papers collected in his 1925 *The Forms of Knowledge and Society*, display a growing appreciation for American Pragmatism. Together with the work of Karl Mannheim they established the sociology of

knowledge. Scheler employs his accounts of collective intentionality and of affective phenomena in order to set out a comprehensive formulation of the view that scientific knowledge and technological knowledge are inseparably bound up with the ‘instinctual structure’ of the bourgeoisie, the desire to dominate nature, and with the predominance of vital values. Scientific knowledge and its sociology is opposed to philosophical and cultural knowledge (*‘Bildungswissen’*) and to religious knowledge (*‘Heilswissen’*) and the sociologies of the spiritual values essential to these, all of which rest on the natural world view and natural language. His account of the link between science and mastery of nature and his distinction between scientific and a superior, non-scientific knowledge have proved remarkably influential throughout the century (see Leiss 1994). But on two points Scheler differs from the majority of his many successors. He retains a realist view of scientific knowledge. The values he distinguishes from those of science, in his philosophy as in his sociology, are spiritual: salvation is not emancipation and philosophy does not involve hearkening to the voice of being.

7. Social Objects

What sort of an entity is a social object? There are, it seems, social entities of many different kinds—enduring ‘substances’ such as judges and the unemployed, processes such as electioneering, events such as voting, states such as ownership. But, as Reinach pointed out, neither of the two venerable philosophical distinctions, physical vs. psychological, temporal vs. ideal, is of much help if we want to understand the status of, for example, the claims and obligations created by promises.

Through the act of promising something new enters the world. A claim arises in the one party and an obligation in the other. What are these curious entities? They are surely not nothing ... but they cannot be brought under any of the categories with which we are otherwise familiar. They are nothing physical; that is certain. One might rather be tempted to designate them as something psychical or mental, that is, as the experiences of the one who has the claim or the obligation. But cannot a claim or an obligation last for years without any change? Are there any such experiences? Recently one has begun to recognize again, in addition to the physical and psychical, the distinct character of ideal objects. But the essential mark of these objects, such as numbers, concepts, propositions, is their timelessness. Claims and obligations, by contrast, arise, last a definite length of time, and then disappear again. Thus they seem to be temporal objects of a special kind of which one has not yet taken notice (Reinach 1983, p. 9).

Husserl baptized such entities ‘bound idealities.’ Unlike ‘free’ idealities such as numbers, bound idealities such as word types, word meanings, and other

cultural entities such as institutions and tools come into being at a particular time and are 'earth bound ... bound to particular territories'; if there were men on Mars there would be 'Mars bound' idealities (Husserl 1973, §65). Like many free idealities, bound idealities such as word meanings, rules, and types of ceremonies are multiply instantiable. The most thorough account of such entities is Ingarden's (1931) ontology of word-meanings and of works of art.

If social entities are brought into being then it is plausible to say that they are mind-dependent entities. But since mental activities belong to the category of episodes it might seem that social entities must all have the reality of episodes. This 'actuality theory,' from the German psychologist Wilhelm Wundt, was rejected in the two most substantive phenomenological ontologies of the social world, *A Contribution to the Ontology of Social Communities*, by Gerda Walther (1922), and *Foundation of the Theory of the Social Bond* (1932) by Husserl's Japanese pupil, Tomoo Odaka. Social entities are, they argue, brought into being and sustained by the mind, but do not all belong to the category of episodes. A thorough criticism of holism about social entities within the framework of Husserl's theory of objects is given by Felix Kaufmann (1929/30, Kaufmann's (1936, 1944) philosophy of the social sciences is, however, closer to the ideas of the logical positivists than to phenomenology).

What type of mental activity is responsible for social entities? According to Walther and Odaka, collective intentionality, implicit and explicit, and in particular we-intentionality, is a necessary condition for the emergence and maintenance of social entities. What, runs Walther's influential question, makes a group of laborers working for a constructor, who interact, know this, pass bricks along and receive such social acts as orders, into a community? What are the marks of a common form of life? Among the thirteen features she puts forward is the sense of togetherness or collective intentionality, a feature also necessary for the existence of all products of a community.

8. Sociology and History of Phenomenology

A notable sociological feature of phenomenology is the role played by female philosophers. A later parallel is the role of women within analytic philosophy, phenomenology's most influential rival.

Many of the distinctions, approaches, and theses of the phenomenologists were rediscovered within analytic philosophy. Reinach's analysis of social acts anticipates John Austin's account of speech acts. The phenomenological discovery of collective intentionality anticipates the work of John Searle and Raimo Tuomela. Similarly, the claim that social entities depend for their existence on collective intentionality has been defended by Searle. Jon Elster's work on such phenomena as resentment and on emotions in general (see Elster 1985, 1999) confirms the phenom-

enologists' conviction that a desire-belief-inference psychology and sociology are inadequate. A variety of theories of motivation and of its relation to causality have been canvassed since Wittgenstein. Scheler's account of communities and their relations to other types of social formation anticipates many of the contemporary varieties of communitarianism. Thus phenomenology is now by and large of interest only to the historian.

See also: Atomism and Holism: Philosophical Aspects; Collective Behavior, Sociology of; Collective Beliefs: Sociological Explanation; Collective Identity and Expressive Forms; Collective Memory, Anthropology of; Husserl, Edmund (1859–1938); Individualism versus Collectivism: Philosophical Aspects; Joint Action; Phenomenology in Human Science; Phenomenology in Sociology; Phenomenology: Philosophical Aspects; Schtz, Alfred (1899–1959); Shared Belief; Theory of Mind

Bibliography

- Boudon R 2000 'La théorie des valeurs de Scheler vue depuis la théorie des valeurs de la sociologie classique'. Forthcoming.
- Elster J 1983 *Sour Grapes. Studies in the Subversion of Rationality*. Cambridge University Press, Cambridge, UK
- Elster J 1999 *Alchemies of the Mind. Rationality and the Emotions*. Cambridge University Press, Cambridge, UK
- Grathoff R 1995 *Milieu und Lebenswelt. Einführung in die phänomenologische Soziologie und die sozialphänomenologische Forschung*, Suhrkamp, Frankfurt, Germany.
- Husserl E (1900–1901, 1913) 1975 *Logische Untersuchungen*, ed. Holstein E. Martinus Nijhoff, The Hague, The Netherlands (1970 *Logical Investigations* [Trans. Findlay J N]. Humanities Press, New York)
- Husserl E (1913) 1969 *Ideen zu einer reinen Phänomenologie und phänomenologischen Philosophie. Erstes Buch: Allgemeine Einführung in die reine Phänomenologie*, ed. Biemel W. Martinus Nijhoff, The Hague, The Netherlands (1931 *Ideas: General Introduction to Pure Phenomenology* [Trans. Boyce Gibson W R]. Collier-Macmillan, London and New York)
- Husserl E (1931) 1973 *Cartesianische Meditationen und Pariser Vorträge*, ed. Strasser S. Martinus Nijhoff, The Hague, The Netherlands (1960 *Cartesian Meditations*. [Trans. Cairns D. Martinus Nijhoff, The Hague, The Netherlands)
- Husserl E 1939 *Erfahrung und Urteil. Untersuchungen zur Genealogie der Logik*, ed. Landrebe L. Academia, Prague (1973 *Experience and Judgment*. Routledge, London)
- Ingarden R 1931 *Das literarische Kunstwerk* (1973 *The Literary Work of Art. An Investigation on the Borderlines of Ontology, Logic and Theory of Literature*. Northwestern University Press, Evanston, IL)
- Kaufmann F 1929/30 Soziale Kollektiva. *Zeitschrift für Nationalökonomie* 1: 294–308
- Kaufmann F 1936 *Methodenlehre der Sozialwissenschaften*. Springer, Vienna, Austria
- Kaufmann F. 1944 *Methodology of the Social Sciences*. Oxford (reprinted New York 1958, this is not a translation of Kaufmann 1936)
- Köhler W 1947 *Gestalt Psychology*. Liveright, New York
- Koffka K 1946 *The Growth of the Mind. An Introduction to Child Psychology*. Kegan Paul, London

- Leiss W 1994 *The Domination of Nature*. McGill-Queen's University Press, Montreal & Kingston
- Mulligan K 1987 Promisings and other social acts: Their constituents and structures. In: Mulligan (ed.). *Speech Act and Sachverhalt: Reinach and the Foundations of Realist Phenomenology*. Hijhoff, Dordrecht, The Netherlands, pp. 29–90
- Odaka T 1932 *Grundlegung der Lehre vom sozialen Verband*. Springer, Vienna, Austria
- Pfänder A 1900 *Phänomenologie des Willens. Eine psychologische Analyse*. Johann Ambrosius Barth, Leipzig, Germany
- Ranulf S 1930 *Gudernes Misundelse og Strafferetens Oprindelse i Athen*, Det Kgl. Danske Videnskabernes Selskabs Historisk-filologiske Meddelelser XVIII, I, Copenhagen, Denmark, (1933/4 *The Jealousy of the Gods and Criminal Law at Athens*, Vols. I–II, Williams and Norgate, London)
- Ranulf S 1938 *Moral Indignation and Middle Class Psychology. A Sociological Study*. Levin & Munksgaard. Ejnar Munksgaard, Copenhagen, Denmark
- Reinach A (1913) 1989 Die apriorischen Grundlagen des bürgerlichen Rechtes. In: Schuhmann K, Smith B (eds.) *Sämtliche Werke* (1983 'The A Priori Foundations of the Civil Law,' Aletheia, Vol. III, Philosophy of Law, 1983, 1–142)
- Reiner H 1934 *Das Phänomen des Glaubens. Dargestellt im Hinblick auf das Problem seines metaphysischen Gehalts*. Max Niemeyer Verlag, Halle (Saale), Germany
- Schapp W (1953) 1976 *In Geschichten verstrickt: Zum Sein von Mensch und Ding*. B. Heymann, Wiesbaden, Germany
- Scheler M 1955 Das Ressentiment im Aufbau der Moralen, Vom Umsturz der Werte, *Gesammelte Werke* 3, pp. 33–148 (1961 Ressentiment. Trans. Holdheim W W. First edition ed. Coser L A. Free Press of Glencoe, New York, 1961; reprint: Schocken Books, New York, 1972.
- Scheler M (1924) 1960 Probleme einer Soziologie des Wissens. Die Wissensformen und die Gesellschaft, *Gesammelte Werke*, 8 (1980 *Problems of a Sociology of Knowledge* [Trans. Frings M S, ed. Stickers K W]. Routledge & Kegan Paul, London)
- Scheler M 1963 *Schriften zur Soziologie und Weltanschauungslehre. Gesammelte Werke* 6
- Scheler M (1913/1916) 1966 Der Formalismus in der Ethik und die materiale Wertethik, *Gesammelte Werke*, 2, ed. Scheler M, Francke Verlag, Bern and Munich (1973 *Formalism in Ethics and Non-Formal Ethics of Values. A New Attempt Toward A Foundation of An Ethical Personalism* [Trans. Frings M S, Funk R L]. Northwestern University Press, Evanston, IL)
- Scheler M (1913) 1973 Wesen und Formen der Sympathie. *Gesammelte Werke* 7 (1954 *The Nature of Sympathy* [Trans. Heath P Routledge & Kegan Paul], London. Reprinted, Archon Books, Hamden, CT, 1970)
- Schütz A 1970 *Studies in Phenomenological Philosophy, Collected Papers, III*, Martinus Nijhoff, The Hague, The Netherlands
- Schütz A 1973 *The Problem of Social Reality, Collected Papers I*. Martinus Nijhoff, The Hague, The Netherlands
- Schütz A (1932) 1974 *Der sinnhafte Ausbau der sozialen Welt. Eine Einleitung in die verstehende Soziologie*. Suhrkamp, Frankfurt, Germany (1967 *The Phenomenology of the Social World*. Northwestern University press, trans. Walsh G, Lehnert F)
- Stein E S (1917) 1980 *Zum Problem der Einfühlung*. Kaffke, Munich (1989 *On the Problem of Empathy* [Trans. Stein W] 3rd rev edn. *The Collected Works of Edith Stein*. Institute of Carmelite Studies, Washington, DC)
- Stein E S 1970 *Beiträge zur philosophischen Begründung der Psychologie und der Geisteswissenschaften. Eine Untersuchung über den Staat*, Niemeyer, Tübingen, Germany. This contains
- (1) (1919, 1922) "Psychische Kausalität" (Psychic causality); (2) (1920, 1922) "Individuum und Gemeinschaft" (Individual and Community); (3) (1921, 1925) "Eine Untersuchung über den Staat" (An Investigation of the State)
- Theunissen M 1977 *Der Andere. Studien zur Sozialontologie der Gegenwart*. Walter de Gruyter, Berlin & New York
- Toulemont 1962 *L'Essence de la Société selon Husserl*. Presses Universitaires de France, Paris
- Walther G 1922 Ein Beitrag zur Ontologie der sozialen Gemeinschaften mit einem Anhang zur Phänomenologie der sozialen Gemeinschaften, *Jahrbuch für Philosophie und phänomenologische Forschung*, VI (Sonderdruck). M. Niemeyer, Halle, Germany

K. Mulligan

Philanthropy, Economics of

Philanthropy is defined as benevolent behavior, usually in the form of charitable gifts, toward others in society. Charitable giving accounts for a significant fraction of income in the USA and other nations, and is often directly encouraged through government tax policy. This entry discusses the motives, causes, and influences of philanthropic behavior. Special emphasis is placed on how government policy, through the charitable deduction on income taxes and through direct grants to charities, can affect the amount of philanthropy in the economy.

1. Introduction

In the late 1990s the media tycoon Ted Turner announced a gift of one billion dollars to the United Nations, software titan Bill Gates donated five billion dollars toward health and education in the developing world, and Oseola McCarthy, an 89-year-old laundry-woman from Mississippi, gave her life savings of \$150,000 to establish scholarships for black college students in her state. These are examples of philanthropic, or charitable, behavior—individuals freely giving money to help others. Most people are, in fact, philanthropists, although not on the scale of the three mentioned above. Through the last few decades of the twentieth century, over two-thirds of Americans donated money to charity in any given year, with households giving over two percent of income on average. In 1997, those that gave to charity averaged donations of over \$1,000. In fact, people spend more on charitable giving than they do on electricity, telephones, or car insurance.

In addition, the government is an active partner with private donors in funding the charitable sector of the economy. It does this in two ways. First, are direct grants to charities. In 1994, for instance, charitable

organizations reported receiving \$51.9 billion in government grants. The second way the government spends money on charity is by excluding income spent on charitable donations from income tax. The US tax code was modified in 1917 to allow taxpayers to deduct their charitable donations from their taxable incomes, and by 1998 individuals were claiming over \$86 billion in charitable deductions.

This entry will describe how economists think about and do research on philanthropy, what economists know about charitable giving, and how public policy influences it. The focus will be primarily on gifts of money and assets by individuals, although corporations and other institutions also make charitable gifts (see Boatsman and Gupta 1996), and people also give by volunteering time (see Brown and Lankford 1992). In addition, the discussion will center on philanthropy in the USA, since a large majority of the academic research has centered on the US experience (for information on charitable organizations internationally, see Anheier and Salamon 1996). Finally, the entry will highlight what challenges remain ahead for the study of philanthropy.

2. An Economist's View of Charitable Behavior

Economics is founded on the view that people are self-interested. Economists assume that since individuals have a choice in how they behave, they must always make the choices that they think are the best at the time. This assumption, called the 'axiom of rationality,' provides the bedrock from which economic models of behavior are formed.

Applying this framework to philanthropic behavior, we can ask why would people who work hard for their money simply turn around and give it away? Does this not contradict the assumption that individuals are self-interested? Maybe. But by the axiom of rationality, we should first look for ways in which an individual believes that giving to charity is the best thing to do with that money at that time.

One possibility is that people desire more of the service provided by the charity. National Public Radio (NPR), for instance, is a charitable organization that survives largely on donations. Over a million people a year give to NPR, with most giving under \$100. Are these givers getting their money's worth? Certainly not. Anyone can listen to NPR broadcasts, even without donating, and anyone who gives \$100 is unlikely to notice any increase in broadcast quality as a result. It follows that a self-interested person is better off giving nothing and taking a 'free ride' on the donations of others. Hence, this alone cannot be an adequate explanation for why people give to charity (Andreoni 1988).

A second reason is that individuals may be getting something directly from the charity in exchange for their contributions. For instance, big donors to the opera may get better seats, or donors to a university

may get buildings named for them. While this must surely matter for large donors, most individuals only receive tokens (a coffee mug?), if anything, in exchange for donations.

A third reason could be that individuals get some internal satisfaction—a 'warm-glow'—from giving to their favored charity, and the more they give the better they feel (Andreoni 1989). This would mean that giving to charity is like buying any other good, such as ice cream. We easily accept that people have a natural taste for ice cream, so why not a natural taste for warm glow? If we accept this, then we can analyze charitable giving just like any other consumer good—when income goes up people should want more, and when the price goes up people should want less.

Scholars have examined preferences for giving from a number of different approaches, and it seems that warm glow is in fact a core economic motivation for giving (see Andreoni 1993). That is not to say it is the only reason for giving. People may also give from a sense of religious duty, from pressure at work or from friends, or as a signal of social status (Rose-Ackerman 1996). At the heart of all of these processes, however, seems to be a basic human interest in helping others or doing one's share. This, of course, raises a deeper question that may require a broader social science perspective to answer: why do people enjoy acting unselfishly? This is a question that, fortunately, we do not have to answer to continue with this article. We can take as our launching point that people get joy from giving, and we can apply our economist's tools for analyzing consumer behavior to begin studying charitable giving.

3. The Facts about Philanthropy in the USA

How do we learn about charitable giving? One source is surveys of consumers. The Independent Sector for instance, surveyed about 2,500 households by telephone. Surveys are valuable since they can obtain information on age, education levels, and other personal characteristics of respondents. A disadvantage is that individuals must rely on imprecise memories when answering questions, or may be reluctant to give true information about their incomes or donations.

Table 1
Sources of private philanthropy, 1998

Source of gifts	Billions of dollars	Percent of total
Individuals	134.8	77
Foundations	17.1	10
Bequests	13.6	8
Corporations	9.0	5
Total for all sources	174.5	100

Source: Giving 1999

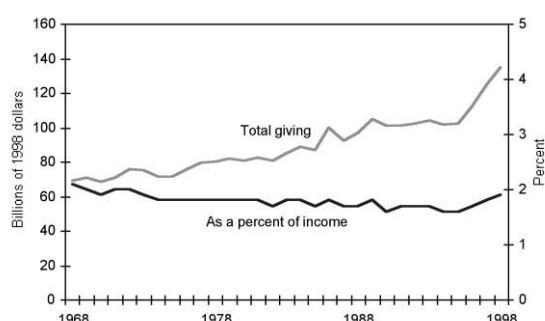


Figure 1

Giving by individuals, 1968–98^a. Dollars are inflation-adjusted to 1998 values.
Source: Giving 1999

A second important source is samples of tax returns. Since individuals who itemize their tax returns in the USA can take a charitable deduction, we can learn about donations for this sector of the economy. The advantage of tax returns is that the information on them is precise, as long as people do not cheat on their taxes. (Slemrod 1989 explored this potential problem

and found that, while there is some evidence of cheating by overstating charitable deductions, the effects are small and do not appreciably affect the analysis.) The disadvantage to this data is that tax returns contain very little information about the personal characteristics of the filers that would be helpful in explaining giving, such as education levels or religious affiliation, nor can we learn about the giving habits of those who do not itemize their tax returns. Since no data source is perfect, economists must conduct many studies on many data sources in order to uncover the ‘facts’ on charitable giving.

3.1 Sources, Totals, and Trends in Giving

Charitable donations can come from individuals, charitable foundations, corporations, or through bequests. While all are significant, by far the dominant source of giving is from individuals. Table 1 shows that in 1998 individuals gave over 134 billion dollars to charity, or 77 percent of the total dollars donated. The second biggest source, foundations, was responsible for only 10 percent of all donations.

The trends in giving over the last 30 years can be seen in Fig. 1. Total giving has been on a steady rise,

Table 2

Private philanthropy by income, age, and education of the giver, 1995

	Percent of households who give	Average amount given by those who give	Percent of household income
All contributing households	68.5	1,081	2.2
<i>Household income</i>			
Under \$10,000	47.3	324	4.8
\$10,000–\$19,999	51.1	439	2.9
\$20,000–\$29,999	64.9	594	2.3
\$30,000–\$39,999	71.8	755	2.2
\$40,000–\$49,999	75.3	573	1.3
\$50,000–\$59,999	85.5	1,040	1.9
\$60,000–\$74,999	78.5	1,360	2.0
\$75,000–\$99,999	79.7	1,688	2.0
\$100,000 and above	88.6	3,558	3.0
<i>Age of giver</i>			
18–24 years	57.1	266	0.6
25–34 years	66.9	793	1.7
35–44 years	68.5	1,398	2.6
45–54 years	78.5	979	1.8
55–64 years	71.7	2,015	3.6
65–74 years	73.0	1,023	2.9
75 years and above	58.6	902	3.1
<i>Highest education of giver</i>			
Not a high school graduate	46.6	318	1.2
High school graduate	67.2	800	1.9
Some college	74.1	1,037	2.1
College graduate or more	82.3	1,830	2.9

Source: Author's calculations, data from Independent Sector 1995

Table 3
Private philanthropy by type of charitable organization, 1995

Type of charity	Percent of households who give	Average amount given by those who give	Percent of total household contributions
Arts, culture, humanities	9.4	221	2.6
Education	20.3	335	9.0
Environment	11.5	110	1.6
Health	27.3	218	8.1
Human services	25.1	285	9.5
International	3.1	293	1.1
Private and community foundations	6.1	196	1.4
Public and societal benefit	10.3	127	1.7
Recreation	7.0	161	1.4
Religion	48.0	946	59.4
Youth Development	20.9	140	3.8
Other	2.1	160	0.3

Source: Author's calculations, data from Independent Sector 1995

with especially big jumps coming in 1996, 1997, and 1998. When measured as a percentage of income, however, giving seems much more stable. Since 1968 giving has varied from 1.6 percent to 2.1 percent of income. In the most recent year shown, 1998, giving ticked up to 2.1 percent of income, the highest in 30 years.

3.2 Who Gives?

Surveys reveal that 68.5 percent of all households gave to charity in 1995 and that the average gift among those giving was \$1,081. Table 2 shows that the more income a household has, the more likely the household is to give to charity, and the more it gives when it does donate. This table also reveals an interesting pattern typically found in charitable statistics. Those with the lowest incomes give over four percent of income to charity. As incomes grow to about \$50,000, gifts fall to 1.3 percent of income, but then rise again to 3.4 percent for the highest incomes.

What could cause this 'U-shaped' giving pattern? One explanation is that those with low incomes may be young people who know their wages will be rising, hence they feel they can afford more giving now. It may also be due to the composition of the types of charities to which people give, since lower-income people tend to give significantly more to religious causes. Hence, it will be important to account for all the factors that may explain giving before offering explanations for the averages seen in these tables.

Table 2 also illustrates that giving varies significantly with the age and educational attainment of the givers. As people get older they are typically more likely to give to charity and to give a greater fraction of their incomes. Likewise, those with more education give more often, give more dollars, and generally give a higher fraction of income. Note that the table does

not show a smooth acceleration of giving with age. Again, age, education, and income all vary with each grouping in the table and will have to be considered jointly.

3.3 What Do They Give To?

In 1997 over 45,000 charitable, religious, and other nonprofit organizations filed with the US government. Table 3 attempts to categorize these charities by the types of service they provide. This reveals that, among all types, households are most likely to give to religious organizations and to give them the most money—48 percent of all households give to religion and 59 percent of all charitable dollars go to religion.

4. Taxes and Giving

How can taxes encourage philanthropic behavior? Taxpayers in the USA who itemize their deductions can deduct their charitable giving from their taxable income (and from 1982 to 1986 the law also allowed nonitemizers a limited charitable deduction). If someone in the 15 percent marginal tax bracket gives \$100 to charity and takes a charitable deduction, this person will save \$15 in taxes, making the net cost of the gift \$85. If the person was in the 31 percent tax bracket then the \$100 gift would have a net cost of \$69. Thus, the higher one's tax rate, the lower the net cost of giving. In this way the tax rate acts as a subsidy rate to giving, and those in higher tax brackets get bigger subsidies. (Note that state income taxes often add additional subsidies. See Feenberg 1987.)

Since the tax deduction subsidizes giving, we should expect that those with higher subsidies will make larger contributions. It is also natural to expect that those with higher incomes will make larger contri-

butions. However, those with higher incomes will typically also have higher subsidies, so when we see richer people giving more how can we know whether their income or the subsidy is causing the change? If we can compare people with similar subsidy rates but different incomes, and similar incomes but different subsidy rates, we can apply statistical techniques to separate the influences of income and the subsidy.

Disentangling these is, of course, very important for policy makers. If, for instance, the subsidy does not increase giving at all, then the policy is costing the government tax dollars without benefiting the charities. If, on the other hand, the subsidy lifts giving significantly, it may be an effective tool to increase social well-being. Generally, policy makers look for the increase in contributions due to the subsidy (the benefit) to be larger than the loss in tax dollars (the cost). The next section discusses how economists explore this issue.

4.1 *The Effect of the Subsidy on Giving*

Economists began conducting detailed empirical studies of giving in the 1970s in an attempt to isolate the effect of the subsidy on giving (see Clotfelter 1985 for a review). You can imagine asking the counterfactual policy question this way: consider a proposal that would allow a taxpayer to deduct only 70 percent (or 50 percent or 20 percent, etc.) of his or her charitable contributions for tax purposes, rather than the 100 percent allowed now. This means that the net cost of a \$100 contribution would rise. For someone in the 28 percent marginal tax bracket, for instance, the cost would rise from \$72 to \$80. We would expect that this higher cost would discourage giving, but by how much?

The first generation of studies to explore this counterfactual relied on cross-sectional data, primarily from tax returns. Observing income and deductions for thousands of people in a given tax year, the researchers used the variations in income and cost to estimate how, on average, individuals would respond to such a change in policy. There were several limiting features of the data used in these studies. Of particular importance is that the samples, for reasons of confidentiality, typically did not include extremely wealthy people.

Depending on the source of the data, the years studied, the size of the sample, and the statistical approach, the results naturally varied. It has been generally agreed, however, that the prediction that best characterizes the results is that a policy that would increase the cost by 10 percent for all taxpayers—without affecting the other income taxes paid by the individual—would reduce contributions by about 13 percent (Clotfelter 1990). This suggests that behavior is quite responsive to the incentives offered by the tax deduction, and that the response is large enough to

satisfy the policy criterion noted above. In particular, if the government pays an additional 10 percent of the cost of giving and if people respond by giving 13 percent more, then the policy generates more new charity than it costs the government in tax revenue.

These studies also confirmed that giving increases with income. If after-tax incomes were to rise by 10 percent, estimates are that giving would rise by about 8 percent (Clotfelter 1990). Notice that this quantifies the left-hand part of the U-shaped giving curve discussed in Sect. 2—since giving rises by a lesser percentage than income, giving as a percentage of income (all else equal) must be declining.

Later studies were able to supplement these findings using data from surveys rather than tax returns (see Clotfelter 1990, Steinberg 1990). Although the survey results varied, they generally confirmed the policy predictions stated above. In addition, they revealed the importance of other individual characteristics in explaining giving. Most strikingly, they consistently found that giving increases with the age and the education of the giver, again as suggested in the tables above. It could be that those who are educated have more interest in charity or that interest in charity changes with age. A more plausible explanation, however, is that researchers have no information on individual wealth. Since wealth is correlated with both age and education, it is likely that the influence of wealth is being filtered through these other variables.

A second generation of studies, conducted mostly in the 1990s, has begun to shift the consensus view noted above. Unlike the earlier studies that relied on cross-sectional data, these newer studies used panel data, that is, data with observations on the same set of individuals over several years. Panel data is generally seen to be superior to cross-sectional data. The reason is that by observing the same people at different points in time the researcher is able to get a more precise measure of how behavior responds to changes in the environment.

The most influential study with panel data is by Randolph (1995). Using tax returns on about 12,000 people from 1979 to 1988, and using statistical techniques designed explicitly for panel data, he identified a much weaker effect of the subsidy and a much stronger effect of income. His results indicate that a policy that would increase the cost of giving by 10 percent would reduce giving by only about 5 percent, whereas if income were to rise by 10 percent, giving would rise by 11 percent. Others using panel data have found similar effects. Hence, at the end of the 1990s the literature on the effect of the subsidy to giving was in a state of flux. More studies with panel data are needed to reconcile the early and later findings, and to reach a new consensus on the effects of the subsidy.

Finally, these studies cannot reveal all of the subtle and complex influences on giving. Other things outside of these studies clearly matter, such as the social circles

to which people relate, the religious organizations to which they belong, and the activities of fundraisers. While many of these influences lie beyond the scope of economic analysis, when economists have examined social factors the effects seem to be small and do not appreciably change the predictions for policy (Andreoni and Scholz 1998). Still, studies that examine noneconomic influences could lead to new insights on policies toward giving.

4.2 The Effects of Tax Reforms on Giving

During the 1980s there were two major tax reforms that dramatically changed the two important variables discussed above: the cost of giving and after-tax income. Prior to 1980 there were 25 tax brackets, with the top bracket at 70 percent. This meant that a wealthy giver, after the deduction, would pay only 30 cents of each dollar donated. The first tax reform came in 1981. This created 14 tax brackets, with the top rate falling to 50 percent. The second major tax reform, the Tax Reform Act of 1986 (TRA86), was more dramatic. It reduced taxes to two rates, 15 percent and 28 percent. (Actually, it also included a 'bubble bracket' of 33 percent for those between \$71,900 and \$192,930 of annual income, before returning to 28 percent for incomes above this. A 31 percent bracket was added in 1991. Subsequent changes in the law mean that in 1999 there are five brackets: 15 percent, 28 percent, 31 percent, 36 percent, and 39.6 percent.) Hence, most people, but especially those in high tax brackets, had the net cost of their charitable contributions rise.

With the costs of giving going up, many feared that there would be calamitous decline in giving. Auten et al. (1992) and Randolph (1995) looked at the experience from the 1980s and made several interesting discoveries. First, people were sensitive to the anticipation of an increase in their cost. In 1981, for instance, wealthy people saw that the cost of a \$100 donation would rise to \$50 in 1982 from \$30 in 1981, and thus moved up a sizable share of their giving to 1981. For those earning over one million dollars (in 1991 inflation-adjusted dollars), average charitable deductions rose by \$32,000 from 1980 to 1981, but fell by \$64,000 in 1982. This same sector also anticipated the TRA86, raising deductions by \$37,500 in 1986, then reducing them by \$52,500 in 1987. Similar, although less pronounced, patterns can be seen in most other income groupings as well. In fact, Randolph (1995) shows econometrically that individuals are far more responsive to short-term (or temporary) changes in the cost of giving than they are to long-term (or permanent) changes.

How did giving change overall in response to the tax reforms? As seen in Fig. 1, dollars given tended to rise over the 1980s, despite this increase in the cost for most individuals. But this does not tell the whole story. First, even though the cost of giving went up, after-tax

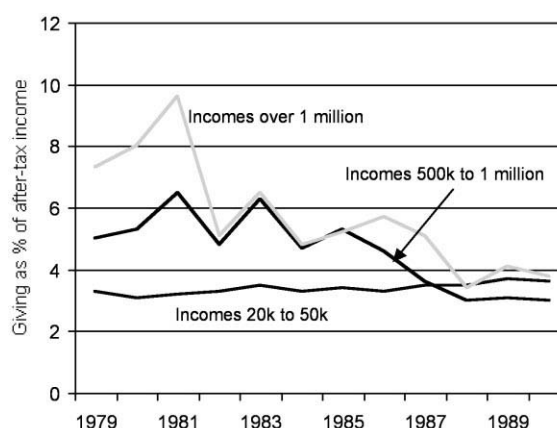


Figure 2
Changes in the Distribution of Giving: 1979–90
Source: Auten et al. 1992

income also rose for middle-income people, and this should counter the effect of raising the costs. In addition, incomes in general were rising throughout the 1980s. Also, this analysis cannot account for the behavioral response of charities that may have increased their fundraising efforts in response to the tax changes.

Fig. 2 shows one more piece of information that is masked by the averages; that is, that the distribution of giving across income groups also changed. The figure shows that giving by the middle-income group, which is by far the most populous, actually increased as a percentage of income over the 1980s. However, the giving by the wealthy fell quite substantially. To balance this, however, far more people were entering the ranks of the super-rich—the number of tax filers with incomes over one million dollars more than tripled from 1980 to 1989.

4.3 Giving by the Very Wealthy

Auten et al. (2000), in a study of giving by the wealthy, report that in 1994 the 1 percent of US households with the highest income was responsible for over 16 percent of all dollars given. As one might imagine, describing giving by the wealthy in more detail can be complicated, in part because many give through trusts and carryover deductions through several tax years. The most distinguishing characteristic of wealthy givers is perhaps that a great deal of their giving is of appreciated assets, such as art, real estate, or stocks. Donations of appreciated assets are often a heavily tax-favored form of giving (O'Neil et al. 1996). First, the giver can take a deduction on income taxes for the current market value of the asset, not the purchase price, in the year the gift is made. In addition to this, the giver does not have to pay any tax on the capital

gains made on the asset. This means gifts of appreciated property often generate two forms of tax savings. Those who had taxable incomes over \$2.5 million in 1995 had average charitable deductions of \$248,069, 44 percent of which was appreciated property. People with annual incomes of \$50,000–\$100,000, by comparison, deducted \$1,823 on average, of which only 15 percent was appreciated property.

Another distinguishing characteristic of wealthy donors is that their giving is highly variable from year to year. In any given year the median contribution of those making over \$2.5 million is only 0.7 percent of income—half the rate of those in the \$50,000–\$100,000 range. But the 95th percentile gift of the rich is almost 21 percent of income—twice the rate of those in the \$50,000–\$100,000 range. When averaged over a longer period, however, these numbers become a lot more similar. This indicates that perhaps the super-rich save up their donations over a number of years in order to give bigger gifts when they do give. Why? One reason might be to gain recognition from the charity, such as a brass plate, for being a major donor. Another reason is that large donors are often asked to serve as board members of charities, hence larger gifts can lead to greater control over the future of the organization.

5. Do Government Grants Displace Private Philanthropy?

There are several reasons to suspect that government grants to charities might depress private donations. First, if givers are aware of the grants, then they may perceive the charity as less in need of their donations, leading them to keep the money or give it to a different cause. This is known as the ‘crowding out’ hypothesis (Warr 1982, Roberts 1984, Bergstrom et al. 1986). A second reason giving may decline is that the charity may be less aggressive in fundraising after receiving a government grant. Both effects would lead to government grants to offset private philanthropy.

There are also reasons to think that government grants could have the opposite effect. Citizens who are unsure about the quality of a charity could see a grant as a ‘stamp of approval.’ Alternatively, the grant could be used as ‘seed money’ that will allow the charity to expand into a bigger operation by, for instance, constructing new buildings or opening branches in different cities (Andreoni 1998). In both of these cases, government grants could actually encourage private philanthropy.

From an empirical standpoint, the question of crowding out has been a challenge, and no study to date has been definitive. Three recent studies illustrate the difficulties of this area, and the wide array of findings. Kingma (1989) studied a special sample of givers to National Public Radio (NPR) in 1986. This data is unique in that it ties each giver directly to his or her local NPR station and the grants it receives. He

finds that if a typical station were to receive a \$10,000 grant, private giving would fall by about \$1,350, that is 13.5 percent crowding out.

Two studies using panel data on donations received by charities came to different conclusions. Khanna et al. (1995) used observations on 159 charities in the UK and found that government grants actually encouraged rather than displaced private giving. Payne (1998), using a 10-year panel of 430 US charities, applied special statistical techniques to account for the fact that government grants to charities may be caused by the same needs and interests as private giving. Without accounting for this ‘simultaneity,’ researchers may underestimate crowding out. Payne’s analysis predicts about 50 percent crowding out.

This is clearly a difficult area to study, and all of the researchers must do the best they can with incomplete data. Moreover, researchers have thus far been unable to account for how charitable fundraising, apart from givers’s willingness to donate, responds to government grants. These analyses await future study.

6. Conclusions and Future Research

Philanthropy is a significant factor in the US economy, accounting for about 2 percent of income over the past 30 years. Government and private givers are in a partnership to fund the services of over 45,000 charitable organizations. This entry has discussed the economic influences on private philanthropy and how it interacts with government policy.

The most important finding is that individuals are indeed sensitive to the charitable deduction in the US tax system. By subsidizing giving, the tax deduction has clearly increased giving by the private sector. Because policies and preferences change with time, new studies are always needed to assess the impact of the deduction. In the 1970s and 1980s, dozens of studies led to a consensus view that the subsidy was highly effective—a 10 percent reduction in the cost of giving led to a 13 percent increase in giving. In the 1990s, new data and new econometric analysis have challenged this view, suggesting that a 10 percent reduction in the cost leads to a 5 percent increase in giving. More studies are needed to clarify the picture about how responsive givers are to the charitable deduction.

The literature on the crowding out of private giving by government grants is also far from settled—some say the crowding out is substantial, at 50 percent, while others find government grants encourage rather than displace private giving. This research makes clear that the study of philanthropy needs to be expanded to consider the role of the charitable organizations more actively in the analysis. Not only do individuals respond to changes in tax policy, but fundraisers do too. How do charities adjust to changes in tax rates or government grants, and how responsive are indivi-

duals to fundraising efforts? Exploring these questions will deepen our understanding of philanthropy and help sharpen government policy.

Of course, to accomplish these important research goals, new and better data will be needed. There has not been a large national survey of individual philanthropic behavior since 1973, and another survey could yield many new insights. In addition, data is needed that ties together the actions of fundraisers, private donors, and the government. Such data would allow researchers to learn about the whole market for charitable giving—both supply and demand—and could help form policies to strengthen the partnership between the public and private sector for providing much needed charitable services.

See also: Altruism and Self-interest; Arts Funding; Income Distribution; Wealth Distribution; Welfare Programs, Economics of

Bibliography

- Andreoni J 1988 Privately provided public-goods in a large economy: The limits of altruism. *Journal of Public Economics* **35**: 57–73
- Andreoni J 1989 Giving with impure altruism: Applications to charity and Ricardian equivalence. *Journal of Political Economy* **97**: 1447–58
- Andreoni J 1993 An experimental test of the public-goods crowding-out hypothesis. *American Economic Review* **83**: 1317–27
- Andreoni J 1998 Toward a theory of charitable fund-raising. *Journal of Political Economy* **106**: 1186–213
- Andreoni J, Scholz J K 1998 An econometric analysis of charitable giving with interdependent preferences. *Economic Inquiry* **36**: 410–28
- Anheier H K, Salamon L M 1996 *The Emerging Nonprofit Sector: An Overview*. St. Martin's Press, New York
- Auten G E, Cilke J M, Randolph W C 1992 The effects of tax-reform on charitable contributions. *National Tax Journal* **45**: 267–90
- Auten G E, Clotfelter C T, Schmalbeck R L 2000 Taxes and philanthropy among the wealthy. In: Slemrod J (ed.) *Does Atlas Shrug? The Economic Consequences of Taxing the Rich*. Harvard University Press, Cambridge, MA, pp. 392–424
- Bergstrom T, Blume L, Varian H 1986 On the private provision of public goods. *Journal of Public Economics* **29**: 25–49
- Boatsman J R, Gupta S 1996 Taxes and corporate charity: Empirical evidence from microlevel panel data. *National Tax Journal* **49**: 193–213
- Brown E, Lankford H 1992 Gifts of money and gifts of time: Estimating the effects of tax prices and available time. *Journal of Public Economics* **47**: 321–41
- Clotfelter C T 1985 *Federal Tax Policy and Charitable Giving*. The University of Chicago Press, Chicago
- Clotfelter C T 1990 The impact of tax reform on charitable giving: A 1989 perspective. In: Slemrod J (ed.) *Do Taxes Matter? The Impact of the Tax Reform Act of 1986*. MIT Press, Cambridge, MA, pp. 203–42
- Feenberg D 1987 Are tax price models really identified: The case of charitable giving. *National Tax Journal* **40**: 629–33
- Giving USA 1999 *Annual Report*. American Association of Fund-Raising Counsel, New York

- Independent Sector 1995 *Giving and Volunteering in the United States*. Independent Sector, Washington, DC
- Khanna J, Posnett J, Sandler T 1995 Charity donations in the UK: New evidence based on panel-data. *Journal of Public Economics* **56**: 257–72
- Kingma B R 1989 An accurate measurement of the crowd-out effect, income effect, and price effect for charitable contributions. *Journal of Political Economy* **97**: 1197–207
- O'Neil C J, Steinberg R S, Thompson G R 1996 Reassessing the tax-favored status of the charitable deduction for gifts of appreciated assets. *National Tax Journal* **49**: 215–33
- Payne A A 1998 Does the government crowd-out private donations? New evidence from a sample of non-profit firms. *Journal of Public Economics* **63**: 323–45
- Randolph W C 1995 Dynamic income, progressive taxes, and the timing of charitable contributions. *Journal of Political Economy* **103**: 709–38
- Roberts R D 1984 A positive model of private charity and public transfers. *Journal of Political Economy* **92**: 136–48
- Rose-Ackerman S 1996 Altruism, nonprofits, and economic theory. *Journal of Economic Literature* **34**: 701–28
- Slemrod J 1989 Are estimated tax elasticities really just tax evasion elasticities? The case of charitable contributions. *Review of Economics and Statistics* **71**: 517–22
- Steinberg R 1990 Taxes and giving: New findings. *Voluntas* **1**: 61–79
- Warr P G 1982 Pareto optimal redistribution and private charity. *Journal of Public Economics* **19**: 131–8

J. Andreoni

Philosophy, Psychiatry, and Psychology

From the very beginning, scientific psychiatry has raised fundamental questions within the fields of philosophy and psychology, namely as regards the nature of the 'psyche.' The question of what it means to say that something is 'psychical' leads to speculation in psychiatry insofar as special heuristic aspects are involved.

1. Philosophical Aspects of Psychiatry

In accordance with Brentano's basic idea that the psychical can be identified with the 'intentional' (Brentano 1874), one might argue that psychiatry is so important to philosophers because the disturbances of intentionality in various domains of the mental world of humans are what might be termed 'negative images' of the psychic functions which normally operate in the human mind. The special method by which Karl Jaspers (Jaspers 1913) elaborated this view of the human mind is the 'phenomenological approach' to various kinds of disturbances of mental functioning in psychiatric patients (deficiency model of psychic dimensions in Jaspers's *General Psychopathology*).

However, this is not the only way in which the relationship between psychiatry and philosophy can

be construed. The other aspect is the issue of the potential role of psychic 'normalcy' in social life. As a fluctuating criterion for assigning people to a psychiatric context, normalcy cannot be regarded as a phenomenon of 'average' and represents merely the concept of 'ideal typology' (Blankenburg 1971). From a philosophical point of view, the subject of psychiatric 'normalcy' constitutes a major problem, since it raises the issue of the existence of an 'individual norm' in the sense of the 'alterity of the other' (Lévinas 1987). Seen in this way, psychiatric disorders are not simply 'defects' but represent variants of normal psychic life which—when exaggerated—are interpreted as disorders and treated by psychiatrists.

This concept fits in with the approach of evolutionary psychiatry which considers that psychiatric disorders have evolved because they relate to psychic functions which—in their submaximal expression—yield advantages where natural selection is concerned. Viewed in this way, psychopathology does not constitute a 'statistical, average norm' but an 'individual norm,' as has been pointed out by Blankenburg (1971). In a similar sense, Spaemann and Löw (1981) have argued that health itself contains a teleological notion: 'If 99 percent of the population suffered from headaches, headaches still would not be "normal".' Philosophical concepts in psychiatry therefore have to address the problem of 'normalcy'; they face the difficult challenge of acknowledging deviant cognitive and mental events merely as expressions of the alterity of the other.

2. Aspects of Greek Philosophy

The focus of Greek philosophy is on the individual, thinking and acting intentionally, and self-reflection is thus the cornerstone of ancient Greek thought. On the basis of the hypothesis of quasi-determinism, the Athenian philosopher Socrates (470–399 BC) asserted that the will responds to the dictates of reason. He developed the Socratic—or 'maieutic'—method which elicits truths that lie hidden in an individual's unconscious by means of a process of question and answer. This method—in a completely different context—later became the basis of Sigmund Freud's psychoanalysis and the nondirective psychotherapy of Carl R. Rogers. Socrates' disciple Plato (427–347 BC) differentiated—for example, in his *Phaedo*—between the spiritual and material worlds and applied this dichotomy to man by postulating that the psyche (soul), together with its logical functions, possesses a pre-existence before the birth of the individual and survives into immortality (Plato 1966). In *Timaios*, Plato defines madness (*manteia*) as a disturbance of the vegetative components of the soul (which are located in the liver) but in the later *Phaidros* he tells us that it can also be sent by the gods as an elevating spiritual force. The aim of Plato's philosophical

psychology is to create harmony between the different psychic activities. Whereas Plato's *idealism* defines ideas as the only true reality, his pupil Aristotle (384–322 BC) based his psychology partly on empirical investigation, thus representing a type of *realism*. In his systematic discussion *De anima* (*On the Soul*), Aristotle describes the soul as the *entelechy*, the self-contained purpose of the body. This purpose must be fulfilled in order to achieve *eudaimonia* (bliss). In Aristotelian thinking, philosophical self-reflection leads to psychological self-actualisation, and in his *Ethica Nikomacheia* Aristotle states that the purpose of ethics is not knowledge but action (Aristotle 1984).

Although Aristotle was the most systematic Greek philosopher and had an enormous influence on the development of modern science, the treatment of psychological matters in post-Aristotelian Hellenistic philosophy is often described as eclectic. Stoic and Epicurean philosophy, in particular, must be considered as practical psychology. Given that the increasingly urbanized Hellenistic world had a greater resemblance—in some ways—to modern society than the middle ages, this practical psychology is still of interest. Epicureans and Stoics tried to find ways to alleviate anxiety, but their anxiety is less fear of the gods than fear and disturbance caused by a complex but fragile society. Epicurean and especially Stoic thinking played an important role in Hellenistic culture and can thus be viewed as the first widespread popular psychological fashions.

After the fall of the Roman Empire, Christian Europe lost sight of some of Aristotle's major works. The philosophical psychology of the dominant neo-Platonism referred back to the static aspects of Platonism, shifting away from rationalism to metaphysical speculation on the relationship between a tangible and an intangible world, between the soul and God. St. Augustine (354–430) attempted to synthesize neo-Platonism with Christianity. Following on from Seneca and Plotinus, he leaned towards introspection as his philosophical and psychological method, considering that—alongside objective observation—self-observation of inner mental processes could function as the basis of knowledge. He therefore chose self-reflection as the basis of his epistemology, arguing that doubt presupposes the existence of the doubter (Augustine 1984), an idea that later became extremely influential as Descartes famous *Cogito ergo sum*.

3. Metaphysics of Middle Ages and the Mind–Body Problem

The philosophical psychology of the Middle Ages was metaphysical and—at the same time—deductive. The changeability of the soul and of world were interpreted philosophically as an inevitable but inferior property of the unchangeable, and consequently Christian dogmatism and neo-Platonism showed no interest in

the experience of psychological interaction. With the foundation of the first universities from the thirteenth century onwards, the process of philosophical debate increased and a more dynamic philosophical psychology developed as Aristotelian thinking was resuscitated through the medium of Arab philosophy. St. Thomas Aquinas (1225–74) attempted to integrate Christian dogma with elements of Greek and Arab philosophy. Aristotle differentiated between an active immortal intellect and a merely passive understanding one. Aquinas accepted this view, describing the intellect as a passive power, a *tabula rasa* which had to be filled with experience by the five exterior senses. In addition, there exists a rational power—which makes free will possible—and the soul also has vegetative components. In contrast to the neo-Platonic dichotomy of soul and body, Aquinas attempted to define man as a psycho-physical unity, as a body inspired by a soul. Taking his lead from Aristotle, Aquinas defined the soul as the form of the body and as an immortal spiritual entity. According to this theory, the body of a dead person consists of the same material or substance as that of a living person. As a result of this dichotomy, it is impossible to describe the impact of mental acts on real action without becoming entangled in serious logical difficulties. This is the *mind-body problem*: the issue of whether or not the soul or mind is an incorporeal substance, a nonmaterial entity obeying laws other than those which govern the material world. Discussion of this problem has continued until the present day.

4. Modern Philosophical Psychology

Modern philosophical discussions related to psychiatric problems already at its beginning. René Descartes (1596–1650) reflected on the fundamental nature of illusions and the constitution of reality. He posed the questions of certainty of reality and compared the reality of dreams with the consciousness in wakefulness. In his *Meditations on first Philosophy* Descartes arrived at the irreducibility of the cogito, putting an end to the infinite regress of doubt about possible fictional realities (Descartes 1996). As a consequence of this concept the modern subjectivistic understanding of consciousness has evolved, leading to the split between *cogitatio* and *extensio*, between soul and body, between the mental and the physical world, a split which represents one of the unresolvable fundamental problems in psychiatry. In this sense the dualistic understanding of the mind-body problem represents one of the central challenges, difficult to overcome in philosophical psychiatry.

Refuting earlier deductive speculation, the empirical philosophy of John Locke (1632–1704) again emphasized the Aristotelian view that the human mind is like a blank sheet of paper and that all the material of reason and knowledge comes from observation, ex-

perience, and reflection. David Hume (1711–76) described ideas as the copies of sensory impressions. Most of our beliefs go beyond current impressions, but all that we can be certain of is perception, meaning that it is impossible, for example, to comprehend the *self* as a separate identity. Radicalizing this point of view, the positivism of Auguste Comte (1798–1857) postulated that the sensible comprises the whole range of knowledge and that any knowledge that cannot be comprehended by the senses must be rejected.

Immanuel Kant (1724–1804) in his precritical anthropology speculated about constitutional aspects of psychiatric disorders. In his *Critique of Pure Reason* he tried to deduce the *a priori* conditions of the possibility of experience (Kant 1974). This type of categorical analysis of mental life has been decisive for all subsequent philosophical considerations about the formation and constitution of subjectivity and consciousness.

In this sense idealism represents one of the possible forms of reactions to Kantian philosophy. For example, Georg Friedrich Hegel (1770–1831) developed an ontological approach to intellectually reconcile all elements of existence with one another within a systematic structure in which all contradictory moments of reality appear as dialectic mediations of the universal mind (Hegel 1973). The philosophical psychologist Franz Brentano (1837–1917) elaborated a comparatively modest conceptualization of the difference between the mental and the physical, thereby developing a fundamental concept of philosophy of the psyche, namely the concept of ‘intentionality.’ In his work *Psychology from an Empirical Standpoint*, he argued: ‘Each mental phenomenon is characterised by what the scholars of the middle ages called the intentional (also probably mental) inner existence ... of an object, and which we ... would call the reference to content, direction toward an object’ (Brentano 1874).

One may argue that the main characteristics of psychiatrically ill patients are disturbances of intentionality in Brentano’s understanding of this term, that is, psychic functions are ‘intentional’ insofar as they refer to an (‘internal’) object; they can ‘mean’ something, can refer to something. If this fundamental function of mental life is disturbed, i.e., the *intentio recta* is disturbed, a psychiatric disorder can be anticipated.

The Danish philosopher Søren Kierkegaard (1813–55), deeply aware of Hegel’s philosophical concept of the absolute and at the same time vigorously writing against it, developed a concept of subjective reality, prestructuring the later existentialistic conceptualizations of subjectivity, defining it as synthesis of soul and body, carried by spirit. His pioneer work about personal freedom formulated the questions as to *The Concept of Anxiety* (Kierkegaard 1980). As has been pointed out by Hersch (1989) these basic ideas are represented by ‘existence’ and ‘moment,’ in which the

subject works through to the moment of the 'reality of freedom,' the condition of 'anxiety.' That means that 'anxiety' cannot be understood primarily as an emotional state but rather an existential condition, an intensified form of existence, leading to an autonomous subject ('choosing the self'). Kierkegaard influenced mainly the French school of philosophical psychology (Sartre, Merleau-Ponty, Camus), whereas Brentano founded, together with his student Husserl, the German school of 'Phenomenology,' essential to the history of psychiatry (Husserl 1950). In this context 'Phenomenological reduction' has to be understood as the attention to the 'nature of the condition of things' and thus the 'things as phenomena.' A great and far-reaching concept in philosophical psychology, the philosophy of the 'language game,' has furthermore been elaborated within the philosophy of language by Ludwig Wittgenstein (1898–1951). Jaspers has pointed out the importance of philosophy of language, emphasizing that language is the 'work of man.' The universals of language also play a role in the philosophy of Arnold Gehlen (1904–76) who referred to man as a 'non-determined animal,' whereby Gehlen interprets the human nature as creating its own world as an 'open world' (Gehlen 1987). Of special importance for psychiatric philosophical psychology is the concept of Helmuth Plessner (1892–1985) who developed the idea of 'positionality' of human subjects. Whereas animals show a 'centred positionality,' human beings live, according to Plessner, in an 'eccentric positionality' (Plessner 1970). This concept plays a relevant part in psychiatric psychopathology, since disorders like delusions of reference and ideas of persecution may be interpreted as a disordered eccentric positionality (Emrich 1992). In Jean-Paul Sartre's (1905–80) *Being and Nothingness*, the 'decentering of the world' plays a fundamental role in the phenomenological analysis of the 'glance' (Sartre 1996). Sartre's analysis of the interpersonal relationship ends up in an intersubject fight of 'glances.' Later Emmanuel Lévinas (1905–95) acknowledged the 'otherness of the other' in the dimension of 'face': 'The epiphany of the face is itself a word of honour in a certain way' (Lévinas 1987). This view of responsibility in the interpersonal encounter is of great importance in psychiatry; it conceptualizes an understanding of empathy in therapeutic relationships.

As a new academic discipline, psychology was obliged to assert itself against the established materialist sciences, especially medicine and biology. Positivist thinking consequently led to strongly anti-mentalistic and anti-introspectionist tendencies in psychology. The theory that it is only via *objective* observation that scientific method can formulate hypothetical laws from which it can then derive deductive arguments led to the radical behaviorist position that nothing that is the product of *subjective* introspection can be accepted. Positivism thus reduced psychology to a science that deals with behavior and

which aims to predict and control that behavior. As a result of behaviorism as represented by John Watson (1878–1958) and B. F. Skinner (1904–90), academic psychology drifted away from philosophy, becoming more practical than theoretical and more concerned with predicting and controlling behavior than with understanding it (Watson 1998, Skinner 1999). The contemporary philosophy of science, however, would claim that it is an illusion to believe that science can be reduced to a simple, *objective*, and entirely rule-bound methodology. Thomas Kuhn argued that such opinions are based on science textbooks that simply fail to tell the truth about the history of science (Kuhn 1996). In contrast to academic psychology, psychiatry has retained a more heterogeneous—even if contradictory—theoretical background. The main question at the start of the twenty-first century is whether psychology and psychiatry can only be governed by certain particular scientific methodologies or if it is possible to accept the existence of different approaches to the acquisition of knowledge. Paul Feyerabend pointed out that 'the idea that science can, and should, be run according to fixed and universal rules, is both unrealistic and pernicious' (Feyerabend 1980). In actual fact, therapeutic practitioners—in particular—find it unsatisfactory to consider introspective experience and objective behavior as mutually exclusive, as if they were distant continents with no contact between them. Modern therapeutic theories are thus becoming increasingly eclectic, and the gulf which used to exist between introspective psychoanalysis and behavior therapy, for example, is in the process of disappearing.

5. Self and Identity

Philosophical concepts of the 'self' have led to psychological consideration of the possible nature of 'identity.' Disturbances of the inner equilibrium of the components of the self which shape identity, brought about by processes of re-presenting by remembering, therefore make clear the potential for explanation inherent in a theory of the processes of identity formation based on a psychology of the self. As Lévinas in particular has shown (Lévinas 1987), however, this cannot be restricted to a theory of the structure of the 'self'; rather it must take account of interpersonality, the existence of others (the 'trace of the other' referred to in the title of Lévinas' book). Any attempt to describe the danger to processes of identity formation posed by the existence of the other from the perspective of the psychology of the self must take account of the 'ontology of the in-between' already developed by Martin Buber (1994). For, as Theunissen (1965) demonstrates in his major work on intersubjectivity, taking the other seriously in its own being involves transcending the intentionality mode.

This philosophical/psychological concept is based

on the assumption that components of a person's self can manifest themselves in others and can exist further in them.

The basis of this primary interpersonal constitution of the psyche is the mother-child dyad of early childhood. In a way, people unconsciously always retain parts of their childhood dyad 'within them,' although they nevertheless consider themselves to be ego-centered 'individuals.' In a system of ideas in which the *cogito* is identified with the empirical ego, the 'interpersonal self' cannot be conceived. This primary (but extremely repressed or internally never realized) interpersonal constitution of the self has consequences for how people deal with what is familiar and unfamiliar, both in the other and within themselves, where aggression is concerned. Christian Scharfetter, referring to Graumann's theory of subjective perspectivity, rightly refers to the fact that precisely the most intimate personal area of a person, his own unconscious, constitutes an alien entity, an unfamiliar area of sinister, threatening mystery. Scharfetter (1994) puts this as follows: 'The other is not simply an alien entity, however. The other may very well be familiar; it can be close, it can be experienced as analogous (not homologous). Indeed, the alien is able to strengthen and reinforce the self. It is through the 'You' that the 'I' recognizes itself. This is an important factor where psychotherapy is concerned. However, confrontation with the alien within the self is painful; it requires not only cognitive efforts (Hegel's 'strenuous toil of conceptual reflection'), but also psychological effort in the form of crises in identity development.

An extremely effective defense mechanism here is that of 'projective identification.' This means that the unknown, the sinister, within the subject is discovered in others by projection and combated there. Streeck-Fischer says, for example, that 'the confrontation with the alien has become a shock experience, one which hampers successful interaction with the alien (Streeck-Fischer 1994). The alien is threatened as a means of stabilizing one's own threatened self.' As Streeck-Fischer shows, this mechanism becomes particularly effective, is in a sense multiplied enormously, when one's own position is not only under pressure from the danger posed by crises in the formation of identity but is also reinforced because of social pressure. Social devaluation of specific features of the self can significantly increase projective identification in the sense of the development of aggression. Hans-Joachim Maaz, for example, shows how in East Germany 'social constraints under a repressive authoritarian system led to the self becoming alien and the alien becoming one's own.' Maaz describes the resulting aggression as follows (Maaz 1994): 'violence and hatred of other people are always evidence of internal tensions within someone, of his no longer being able to find and develop the self and the essential but of having to live a life too much of which is imposed by others.'

One of the great figures in the theory of mimesis is René Girard. In various works, notably *La violence et le sacré*, Girard developed a theory of mimetic occurrences (Girard 1977). If one accepts his thinking, the main point of the Oedipal drama, which in Freudian theory is said to mold early childhood, is not the Oedipal aspect as such—the desire for the mother in the Freudian sense—but the mimetic rivalry with the father. In an early work of 1961, *Mensonge romantique et vérité romanesque*, Girard elaborated on a discovery central to the entire development of his theory. He was able to identify a phenomenon of 'mimetic desire' in nineteenth and twentieth century novels from Stendahl to Flaubert and Proust. The novels he examined shared a peculiar basic structure: the main characters desire someone else, but their feelings are directed toward a third person, a 'mediateur.' Analyzing Stendhal, Girard says: 'in most cases of Stendhalian desire, the *mediateur* himself desires the object, or could desire it. This desire, whether real or assumed, is precisely what makes the object infinitely desirable in the eyes of the subject.'

Desires, needs, preferences, and yearnings are thus described as being something not genuinely given, something brought forth internally by drives or make-up, but are 'read,' taken on from a third party because of mimetic dependency. The basic events in the process of identification can thus be described as follows: if we can understand the other in a recreative sense only by putting ourselves into the other's intentions, into that person's world of desires, this means that we have to identify with that other. The amazing consequence of this is that processes of identification are necessary for a proper interpretation of reality. In this sense, one can say that a rational explanation of the world is possible only if there is a fundamental willingness to identify with others.

As a result, one can argue that in psychiatric patients this rational world access is hampered by their impaired interpersonal relations, that is, due to disturbed mimetic processes (Emrich 1995).

See also: Anthropology, History of; Cognitive Functions (Normal) and Neuropsychological Deficits, Models of; Mental Health and Normality; Mind-Body Dualism; Personality Theory and Psychopathology; Phenomenology in Human Science; Phenomenology: Philosophical Aspects; Psychiatry, History of; Psychology: Historical and Cultural Perspectives; Self: History of the Concept; Theory of Mind

Bibliography

- Aristotle 1984 *The Complete Works of Aristotle: The Revised Oxford Translation*. Princeton University Press, Princeton, NJ
Augustine 1984 *City of God*. Penguin, Harmondsworth, UK
Blankenburg W 1971 *Der Verlust der natürlichen Selbstver-*

- ständigkeit. Ein Beitrag zur Psychopathologie symptomarmer Schizophrenien. Ferdinand Enke, Stuttgart, Germany
- Brentano F C 1874 *Psychologie vom empirischen Standpunkt*, Vol. 1. Duncker & Humblot, Leipzig [1995 *Psychology from an Empirical Standpoint*. Routledge]
- Buber M 1994 *Ich und Du*. Lambert Schneider, Gerlingen [1978 *I and Thou*. Scribner, New York]
- Descartes R 1996 *Meditations on First Philosophy: With Selections from the Objections and Replies* (trans. Cottingham J). Cambridge University Press, Cambridge, UK
- Emrich H M 1992 Systems theory of psychosis: 'Filtering,' comparison, error correction, and its defects. In: Emrich H M, Wiegand M (eds.) *Integrative Psychiatry*. Springer, Heidelberg, Germany pp. 81–89
- Emrich H M 1995 Physiognomy of the psychological—toward a theory of mimesis. *Paragrana* 4(2): 126–43
- Feyerabend P K 1980 *Against Method*. Verso, London
- Gehlen A 1987 *Der Mensch*. Aula Verlag, Wiesbaden [1988 *Man*. Columbia University Press]
- Girard R 1961 *Mensonge romantique et vérité romanesque*. Grasset, Paris
- Girard R 1977 *Violence and the Sacred*. Johns Hopkins University Press, London
- Hegel G W F 1973 *Phänomenologie des Geistes*. Suhrkamp, Frankfurt [1979 *Phenomenology of Spirit*. Galaxy Books, New York]
- Hersch J 1989 *Das philosophische Staunen*. Piper, München, Germany
- Husserl E 1950 *Gesammelte Werke*. Kluwer, Dordrecht, The Netherlands
- Jaspers K 1913 *Allgemeine Psychopathologie*. Springer, Berlin [1998 *General Psychopathology*. Johns Hopkins University Press, Princeton, NJ]
- Kant I 1974 *Kritik der reinen Vernunft*. Suhrkamp, Frankfurt [1996 *Critique of Pure Reason*. Hackett, Indianapolis, IN]
- Kierkegaard S 1980 *The Concept of Anxiety: A Simple Psychologically Orienting Deliberation on the Dogmatic Issue of Hereditary Sin*. Princeton University Press, Princeton, NJ
- Kuhn T S 1996 *The Structure of Scientific Revolutions*. University of Chicago Press, Chicago
- Lévinas E 1987 *Totalität und Unendlichkeit*. Versuch über die Exteriorität. Alber, München [1969 *Totality and Infinity: An Essay on Exteriority*. Duquesne University Press, Pittsburgh, PA]
- Maaz H J 1994 Das Eigene und das Fremde im deutschen Vereinigungsprozeß. In: Egner H (ed.) *Das Eigene und das Fremde: Angst und Faszination*. Walter Verlag, Hamburg, pp. 13–27
- Plato 1966 *The Collected Dialogues of Plato*. Pantheon, New York
- Plessner H 1970 *Philosophische Anthropologie*. Fischer, Frankfurt
- Sartre J P 1996 *Being and Nothingness*. Pocket Books, New York
- Scharfetter C 1994 Im Fremden das Eigene erkennen. In: Egner H (ed.) *Das Eigene und das Fremde*. Walter Verlag, Hamburg, pp. 13–27
- Skinner B F 1976 *About Behaviorism*. Pimlico, London: 1999; Random House, New York: 1976
- Spaemann R, Löw R 1981 *Die Frage Wozu? Geschichte und Wiederentdeckung des teleologischen Denkens*. Piper Verlag, München, Germany
- Streek-Fischer A 1994 'Haßt du was, dann bist du was'. Über Fremdenhaß und seine selbstreparative Funktion am Beispiel jugendlicher rechtsextremer Skinheads. In: Egner H (ed.) *Das Eigene und das Fremde*. Walter Verlag, Hamburg, pp. 13–27

Theunissen M 1965 *Der Andere*. Walter de Gruyter, Berlin

Watson J B 1998 *Behaviorism*. Transaction, New Brunswick, NJ

H. M. Emrich and W. Paetzold

Phonetics: Articulatory

Articulatory phonetics is the branch of phonetics concerned with describing the speech sounds of the world's languages in terms of their articulations; that is, the movements and/or positions of the vocal organs (articulators). There were independent early Sanskrit, Arabic, and Western European traditions of articulatory description, often motivated by a desire to preserve pronunciations of an earlier stage of a language, e.g., for ritual purposes. More recently, it flourished in the nineteenth century, motivated by such goals as spelling reforms, dictionary making, language teaching, and language and dialect description, and continues in this tradition to the present.

1. Phonetic Transcription

A system for articulatory phonetic description comprises the set of vocal organs deemed to be important for speech, and the set of their movements or positions capable of characterizing the various speech sounds. A. M. Bell's *Visible Speech* (1867), a speech teaching aid for the deaf, used quasi-iconic phonetic symbols to notate positions of the lips, tongue, and soft palate. However, the most influential system of articulator-based notation has been that of the International Phonetic Association (IPA), founded in 1886. In the IPA system, the symbols are not articulatorily iconic, being instead based on Roman letters, but they encode a well-developed system of articulatory description.

1.1 International Phonetic Alphabet

The articulators specified in the IPA system are the lungs, the larynx, the two lips, the upper surface of the oral cavity from the teeth back to the uvula (divided into alveolar ridge, hard palate, and soft palate), the uvula, the pharynx and epiglottis, the tongue (divided into the tip, blade, front, back, root, and sides), and the vocal folds (or cords). These are divided conveniently according to their functions in speech as follows.

Respiratory (or initiatory): The respiratory system provides an outward pulmonic airstream with a roughly constant pressure; however, since other articulators, such as the larynx, can also be used to initiate airflow, this is sometimes also called, more generally,

CONSONANTS (PULMONIC)

Where symbols appear in pairs, the one to the right represents a voiced consonant. Shaded areas denote articulations judged impossible.

VOWELS

Where symbols appear in pairs, the one to the right represents a rounded vowel.

SUPRASEGMENTALS

kp ts

TONES AND WORD ACCENTS
LEVEL: CONTINUUM

LEVEL		CONTOUR	
or	Extra high	e	or / Rising
	High	ê	\\ Falling
	Mid	ẽ	/ rising
	Low	ẽ	/ Low rising
	Extra low	ẽ	/ Rising-falling
	Downstep	↘	Global rise
	Unstep	↘	Global fall

the airstream process. Sounds are described in terms of the source and direction of movement of their airstream.

lates the pulmonic airstream to produce a sound source called voicing; the rate of this vibration is perceived as the pitch of the voice, and the mode of vibration is perceived as voice quality. Sounds are described in terms of the presence/absence, and if

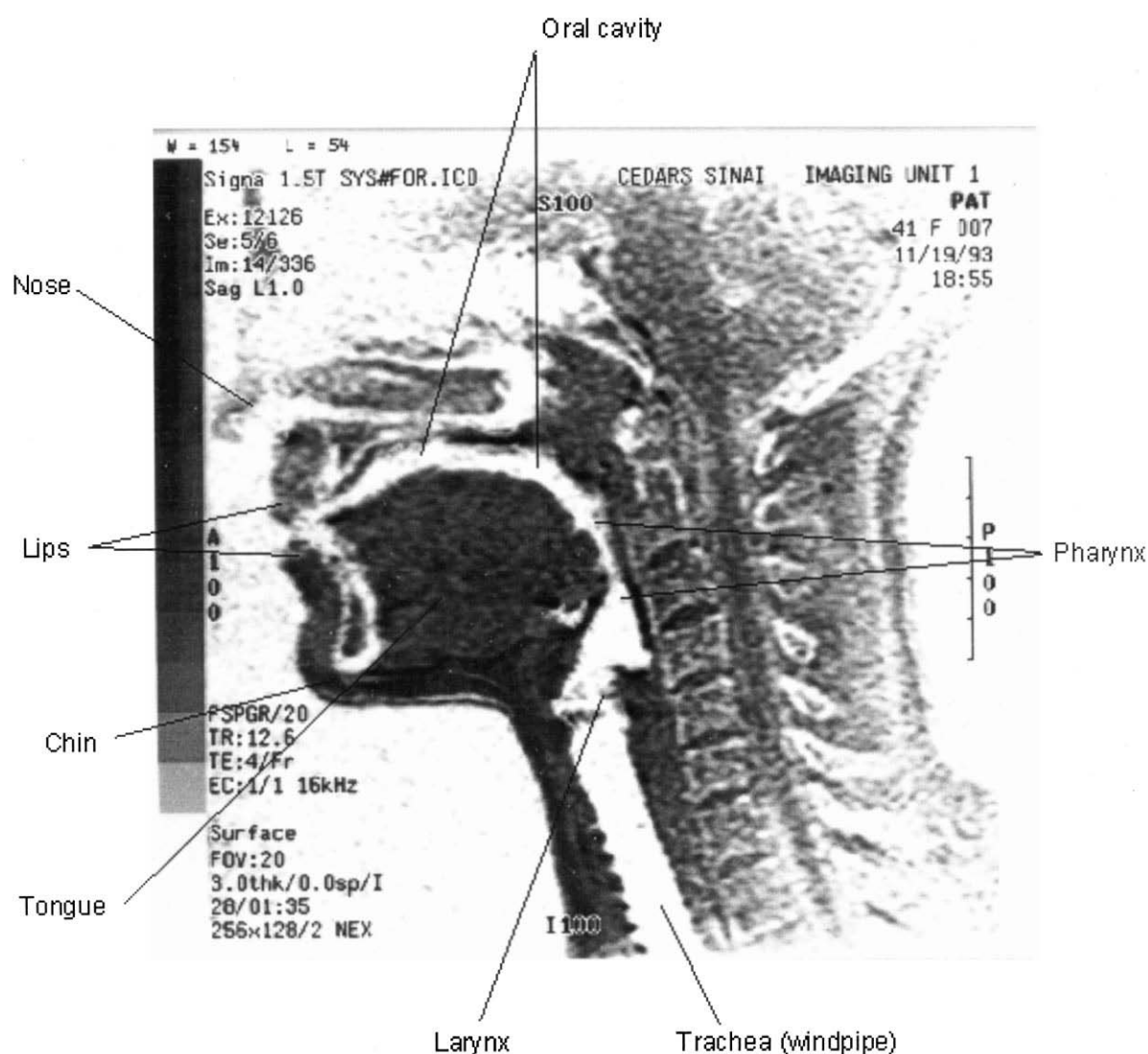


Figure 2

MR Image of midline of vocal tract during American English [s]. White portions are airspaces, black portions are structures, some of which are labeled. (Source: Fromkin 2000. Reproduced with permission)

present, the quality, of voicing; larger spans of speech are described in terms of the pitch of voicing (e.g., tones, intonation—see *Suprasegmentals*).

Articulatory (in a narrower sense): Structures above the larynx, including the pharynx walls, the tongue (and its subparts), the jaw, the velum (or soft palate—sometimes classified separately as the oro-nasal process), the uvula, and the lips, which move within the vocal tract to modify the shape, and therefore the resonances, of the airway. In addition, stationary articulators—upper teeth and hard palate (and its

subparts)—provide contact sites for these movable articulators. Within a given sound, the articulators may be classified as active vs. passive, with the active articulators moving toward the passive and often forming a constriction or narrowing along the vocal tract. (While stationary articulators are necessarily passive, movable articulators may be active or passive.) For most sounds, the extreme or target positions of the active articulators relative to the passive are taken to suffice to characterize the sounds. For example, labiodental means active lower lip to passive

upper teeth; velar means active tongue body to (presumably) passive soft palate (the velum); a front vowel has the tongue as a whole to the front, relative to the surface of the palate. These locations must be further qualified according to how close the articulators come together. For example, in a stop consonant, they touch, whereas in a vowel they are still relatively far apart. These two dimensions of the location of the active articulator are sometimes called the location or place, vs. the degree, stricture, or manner, of the constriction.

The IPA uses these articulatory dimensions (and others not presented here) both to describe speech sounds, and as a basis for defining phonetic symbols to represent those sounds. These dimensions serve as a row and column labels on the well-known IPA chart, the most recent (1996) version of which is reproduced here as Fig. 1. This chart comprises two consonant charts, a vowel chart, and lists of other symbols. Each symbol represents the combination of articulatory properties expressed by the row and column labels of its chart. Square brackets are usually used to show that a symbol is part of a phonetic alphabet. Thus the first symbol on the IPA chart, [p], represents a *bilabial* (column label—made with the two lips approaching each other) *plosive* (row label—an oral stop; i.e., made with complete closure between the active and passive articulators, and the soft palate raised to prevent nasal airflow) which is *voiceless* (leftmost symbol in its cell—made without vocal fold vibration).

The consonant chart can be seen as a grid representing the articulatory space, with sounds made in the front of the mouth located toward the left of the chart, and sounds made in the throat (pharynx) located toward the right of the chart, and the degree of openness of the vocal tract given from top to bottom of the chart. The vowel chart's layout also gives a similar articulatory grid. However, rather different articulatory descriptions are given for consonants vs. vowels, even when the same articulators are involved, as can be seen from the different row and column labels. The basic descriptive dimensions for vowels are the overall vertical and horizontal position of the active articulator (the tongue), and the position of the lips. While the position of the tongue is judged implicitly relative to passive structures, these are not referred to overtly; only the active articulator, the tongue, is discussed. Thus what might be called a 'close palatal vowel with bilabial approximation' is instead nowadays generally called a 'high (close) front rounded vowel.' Not all phonetic systems make this kind of functional differentiation of vowel vs. consonant articulations, and the issue remains controversial.

The IPA system aims to provide a symbol for every phoneme of every language. However, it is a crucial property of this system that the descriptive framework is independent of the particular symbols chosen to fill the cells; other symbols could be substituted for these



Figure 3

Tracings from MR images of midline of vocal tract during American English [ɪ] and American English sound spelled 'r.' (Sources: Narayanan and Alwan 1997, Alwan and Narayanan 1997. Reproduced with permission)

as long as they were defined appropriately, and the charts provide places for sounds that, while possible, have not been assigned symbols.

For further explication of this chart and the IPA system, including its resources for description of suprasegmental phonetics (see *Suprasegmentals*) and of the range of speech sounds found in the world's

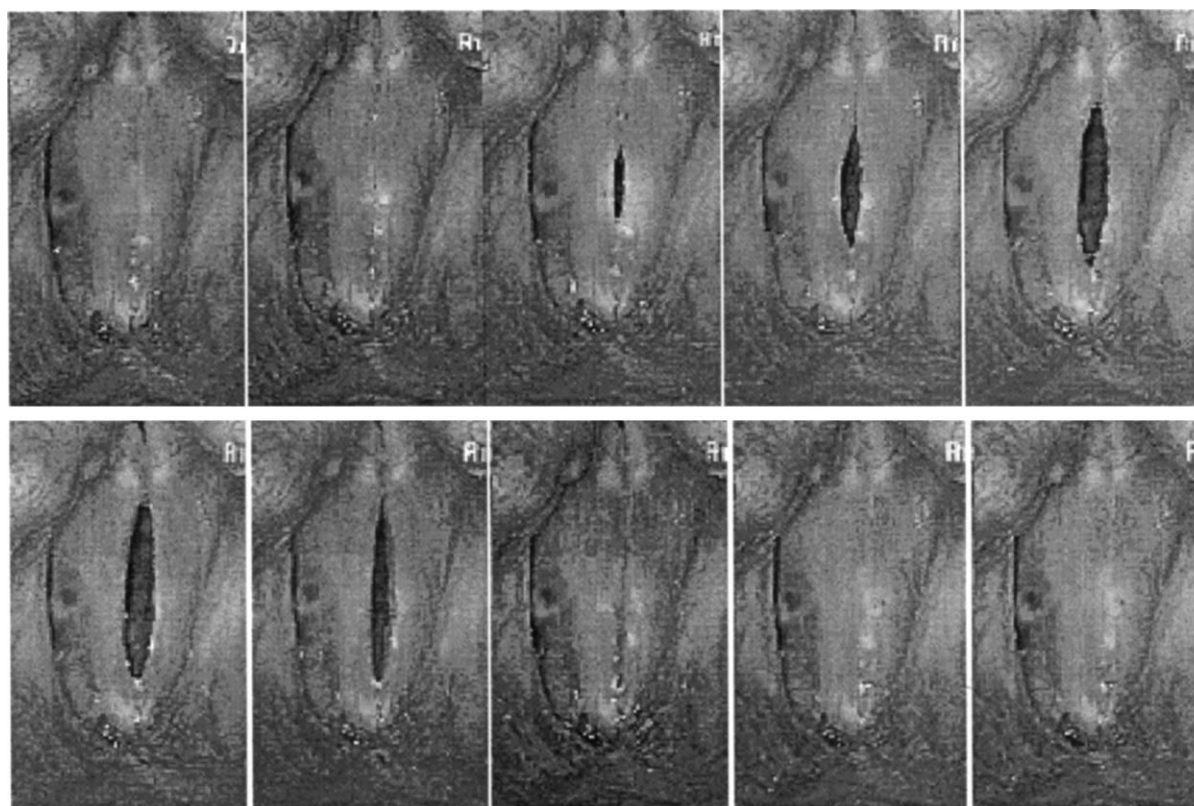


Figure 4

Frames from a video stroboscopy recording of vibrating vocal folds. (Source: Ladefoged 2001. Reproduced with permission)

languages, see the *Handbook of the International Phonetic Association* (IPA 1999), a phonetic textbook such as Ladefoged (2000), or a phonetics chapter such as that in Fromkin (2000). For more on how these and other articulatory dimensions can be used to classify speech sounds, see *Phonology*.

1.2 Nonsegmental Transcription

Speech tends to consist of alternating closings (significant constrictions) and openings of the vocal tract; these are called consonants and vowels respectively. It is a fundamental observation that the actions of the various articulators are somewhat asynchronous. That is, it is generally not the case that all the articulators move into and hold their positions for one sound, then move into and hold their positions for the next sound, and so on. The dominant traditions of phonetic description, including the IPA, assume that speech consists of sequences of separate speech segments whose diverse articulations can overlap in time. It is possible to represent much of this overlap with a phonetic alphabet (e.g., Kelly and Local 1989) but it

requires cumbersome use of diacritical markings around the alphabet symbols. In contrast, competing articulatory phonetic systems do not bundle the articulatory properties into separable but overlapping segments, nor do they assign phonetic symbols to such segments. Instead, they try to record the temporal articulatory patterns of speech by using more complex representations (see, e.g., Anderson 1985; also *Phonology, Nonlinear*).

2. Study of Articulation

In most systems for articulatory phonetic description, both the identities of the articulators in general, and their positions in particular sounds are often presented in terms of static diagrams based on side-view X-rays or other images. Figure 2 gives a Magnetic Resonance image (MRI) of the vocal tract (the part of the body used in sound production), along its midline, of a speaker of American English making the fricative consonant[s]. This scan shows the larynx, the pharynx, the tongue, and the oral cavity, along with some other landmarks. Figure 3 shows tracings of the tongue and

other relevant structures made from other MRI images of the same speaker, but in this case during the English lateral approximant consonant [l] (which looks similar to [s]) and the English central approximant consonant spelled 'r' (a sound with an unusual tongue shape found in few of the world's languages). Note that, though such pictures are static, that does not mean that the articulations are necessarily held for any appreciable length of time in real speech—think of this as the frame, from a movie, which shows the most extreme displacement. In some sounds, such as stop consonants, articulations are held; in others, such as trills, movements of articulators are crucial to sound production. Direct evidence about other aspects of speech articulation can be gained using other laboratory techniques, from those dating to the nineteenth century and earlier, to the latest medical advances. For example, Fig. 4 shows successive frames from a video recording of vocal fold vibration taken with a video stroboscope.

Much of the current effort in articulatory data collection is focused on model building. Models vary in their physicality (whether they model the movements of realistic structures controlled by muscles, or whether they perform more abstract tasks), but all seek to gain insight into the kinematic and dynamical behavior of speech, either as an instance of refined motor control, or as an important part of language.

Articulatory activity ultimately serves the acoustic ends of sound production and propagation. Each speech sound has its defining acoustic properties that result from its defining articulatory properties. The different configuration of articulators define different resonance patterns of sounds, and interruptions and modifications of the airstream differentiate sounds in other ways. Static articulatory data, especially 3-D imaging of the entire vocal tract, is used in building models of the relation of articulation of acoustics—how the articulatory positions determine the acoustic properties. Models such as those of Fant (1960), and now Stevens (1998), can be applied in articulatory synthesis of speech. There is some effort to recover articulatory phonetic information from acoustic signals so that it can be used in speech recognition systems.

See also: Linguistics: Overview; Phonology; Speech Errors, Psychology of; Speech Production, Neural Basis of; Speech Production, Psychology of; Speech Recognition and Production by Machines; Suprasegmentals

Bibliography

- Anderson S R 1985 *Phonology in the Twentieth Century*. University of Chicago Press, Chicago, IL
Bell A M 1867 *Visible Speech*. Simpkin, Marshall and Co., London

- Fant G 1960 *Acoustic Theory of Speech Production*. Mouton, The Hague, The Netherlands
Fromkin V A 2000 Phonetics. In: Fromkin V A (ed.) *Linguistics: An Introduction to Linguistic Theory*. Blackwell, Oxford, UK, Chap. 11
IPA (International Phonetic Association) 1999 *Handbook of the International Phonetic Association: A Guide to the Use of the International Phonetic Alphabet*. Cambridge University Press, Cambridge, UK
Kelly J, Local J 1989 *Doing Phonology: Observing, Recording, Interpreting*. Manchester University Press, Manchester, UK
Ladefoged P 2000 *A Course in Phonetics*, 4th edn. Harcourt Brace College Publishers, Fort Worth, TX
Stevens K N 1998 *Acoustic Phonetics*. MIT Press, Cambridge, MA

P. A. Keating

Phonology

Phonology is concerned with the sound structure of words and utterances within individual languages, the way distinctions in sound are used to differentiate linguistic items within a language, and the ways in which the sound structure of the 'same' element varies as a function of the other sounds in its context. While both phonology and phonetics are concerned with the role of sound in natural language, they differ in that phonetics deals with the articulatory, acoustic, and perceptual properties of sounds from a general, language-independent point of view (see *Phonetics: Articulatory*), while phonology studies the ways in which these properties are distributed and deployed within particular languages.

1. Goals of Phonology

The study of a language's particular sound properties has a number of aspects. First, it must be able to characterize the language's *inventory*: which phonetically possible sound types occur in utterances in the language? Second, it must characterize matters of *contrast*: which of the phonetic differences that occur in the language can serve to distinguish utterances (words, sentences, etc.) from one another? Third is the matter of *contextual limitation*: even though some property P occurs in language L, are there some environments from which P is excluded? And when P is apparently a property of (some part of) some element but that element occurs in a position from which P is excluded, what other property—if any—appears in its place? And finally, there is the description of *alternation*: when the 'same' linguistic element appears in different overt forms in different environments, what systematic differences occur? What conditions govern the range of phonetically

distinct forms that can count as the 'same' word, morpheme, etc.? Different answers to these questions yield different phonological theories.

It should be noted that the present article is limited to the sound structure of spoken languages, and ignores the expression systems of manual or signed languages (see *Sign Language*). This is misleading in important respects; research has shown that most of the basic principles of spoken language phonology are also characteristic of the organization of the expression systems of signed languages as well (Coulter 1993). Just as words are composed of sounds, and sounds of component properties, signs are also composed from structured, language-particular systems of more basic constituent elements. Units such as the syllable have close parallels in signed languages. While there are clear differences that depend on modality, these appear on further examination to be relatively superficial. A comprehensive theory of phonology as a part of the structure of natural language ought to take these broader issues into account. Until quite recently, however, the possibility of deep structural parallels between speaking and signing has not been raised, and the discussion below reflects this (undoubtedly unfortunate) limitation as well.

1.1 Some History

Prior to the early twentieth century, studies of sound in language concentrated on the ways in which sounds are made (articulatory phonetics), often confusing the letters of a language's writing system with its sounds. Toward the end of the nineteenth century, however, increasing sophistication of measurement techniques made it possible to explore a much wider range of differences among sounds, and to lay out the structure of speech in vastly greater detail. Somewhat ironically, perhaps, the explosion of data which resulted from the ability of phoneticians to measure more and more things in greater and greater detail began to convince them that they were on the wrong track, at least as far as increasing their understanding of particular languages was concerned.

Much of what was found, for example, involved the observation that speech is continuous, such that whatever is going on at any particular moment is at least a little different from what has gone on just before and what will go on immediately afterwards. A full characterization of an utterance as a physical event requires the recognition of a potentially unlimited number of distinct points in time, but it is clear that our understanding of an utterance as a *linguistic* event is hindered, rather than helped, by the recognition of this continuous character of speech. Speech normally is represented as a sequence of a small number of discrete segments, strung out in consecutive fashion like beads on a string; such a *segmental* representation vastly facilitates the discovery of regularity and coherence, but it must be emphasized that there is no

direct warrant for it in either the acoustic or the articulatory observable data of speech, and it constitutes a fairly abstract (though undoubtedly appropriate) theory of how sound is organized for linguistic purposes.

It is clear that the role of particular sound differences varies considerably from one language to another. Thus, in English, the vowel sound in the word *bad* is much longer than that in *bat* (more than half again as long), but such a difference in length is always predictable as a function of the following sound, and never serves by itself to distinguish one word from another. In Tahitian, in contrast, essentially the same difference in length is the only property distinguishing, for example, *paato* 'to pick, pluck' from *pato* 'to break out.' A theory of sound that attends only to physical properties has no way of clarifying the quite different functions these properties may have across various languages. This is not to suggest that phonetics is wrong, but rather that there is more to be said.

1.1.1 De Saussure and the 'phonemic principle.' The great Swiss linguist Saussure (1916) was the first to stress that in order to understand the role of sound in language it is necessary to focus not (just) on the positive properties of sounds, but on their differences. He suggested that in the study of individual languages, as opposed to general phonetics, utterances should be characterized in such a way that two such representations might differ only in ways that could potentially correspond to a difference between two distinct messages in the language in question. Thus, since long and short vowels never (by themselves) distinguish two distinct utterances in English, the difference should not be indicated in that language; while for Tahitian, it must be. A representation with this property will be called *phonological*; it will obviously be specific to a particular language, and the distinctive elements that appear in it can be called the *phonemes* of that language.

While de Saussure enunciated this principle quite forcefully and persuasively, he provided few specific details of just what a phonological representation should look like. There are in fact a variety of ways in which his insight could potentially be realized, and much subsequent discussion in phonology hinges on these differences of interpretation.

1.1.2 The development of phonology as a theory. Various individual investigators arrived at conclusions similar to de Saussure's about the importance of attention to language-particular sound contrasts. One of these was the Polish linguist Baudouin de Courtenay (1972), whose work actually antedated de Saussure's, but attracted little attention due to his isolation in Kazan. He developed a sophisticated view of the relation between phonetics and phonology both in individual grammars and in linguistic change. As

transmitted by his later students, Baudouin's views on the nature of the phoneme constituted an important strand in thinking about language as this developed in Russian linguistics in the early years of the twentieth century. This, in turn, provided the background from which the work associated with the Linguistic Circle of Prague grew in the 1920s and 1930s.

Two of the most prominent members of the Prague Circle were Trubetzkoy (1939) and Jakobson (1941). In their studies of Slavic languages and their histories, they stressed the notion that the collection of potentially contrastive sound types in a language was not simply an inventory, but a highly structured system. This system is organized in terms of a small number of mutually orthogonal dimensions (such as voicing, stop vs. continuant, nasality, etc.), each of which serves in parallel fashion as the basis of multiple contrasts. The notion that the fundamental terms of sound structure in language are these properties themselves and not (or at least not only) the complete sounds they characterize has remained an important component of most subsequent theorizing. The analysis of phonological structure in terms of its constituent basic contrasts, in turn, has served as a model for a variety of other disciplines in the Humanities and the Social Sciences apart from linguistics.

1.1.3 American structuralist phonology. Early thinking about sound structure in America was dominated by the anthropological interests of Franz Boas, and focused on an accurate rendering of the sound contrasts in the comparatively 'exotic' indigenous languages of the new world. Boas's student Edward Sapir, however, was concerned to place the study of language in the broader context of an understanding of the human mind and society. As such, he stressed (Sapir 1925) the notion that the elements of sound contrast in a language should be regarded as having a primarily mental reality, part of the speaker/hearer's cognitive organization rather than as external, physical events.

The rise of positivist views of science in the 1930s, and especially of behaviorist psychology, made Sapir's sort of mentalism quite marginal, and replaced it with more rigorous operational procedures for investigating notions of contrast. Especially associated with the ideas of Bloomfield (1933) and later structuralists such as Bloch and Harris (1951), the result was a theory of the phoneme based exclusively (at least in principle) on a set of mechanical manipulations of corpora of observed linguistic data, from which a set of contrasting minimal elements was to be derived. The central notion of this theory was a phonemic representation related to surface phonetic form in a way that would later be formulated explicitly as a condition of *Bi-uniqueness*: the requirement that given either a phonetic or a phonemic representation of an utterance in a given language, that could be converted uniquely

into the other (disregarding free variation) without additional information. The perceived rigor of this notion led to its being widely taken as a model not only for the study of other areas of linguistic structure, but for other sciences as well.

1.1.4 Generative phonology. The phonemic theories of American Structuralists provided a way to characterize linguistic contrasts, the inventories of sound types used in a given language, and the ways sounds can be combined into larger structures, but other aspects of sound structure were less satisfactorily accommodated within those views. In particular, questions of the ways in which unitary meaningful elements change in shape according to their sound context (or 'allomorphy': see *Morphology in Linguistics*) failed to receive systematic treatment. Since any difference among sounds that could serve to contrast linguistic elements was *ipso facto* a difference between irreducibly basic terms, there was really no way to express the notion that a single item could take a variety of forms (as in the prefixes in *inefficient*, *imprecise*, *irregular*, *illegal*, etc.) except by simply listing the variants. Such a list is undoubtedly appropriate for cases such as the forms of English to be (*am*, *are*, *is*, *was*, *were*, etc.) which are unrelated to one another in form; but in many other cases, the variation is transparently systematic and a function of the sounds in the element's environment. This sort of variation was recognized by structuralist phonologists, but relegated to marginal status.

Beginning with work of Morris Halle, a student of Jakobson, linguists began to question the centrality of surface contrasts in sound structure. The result was a new view that allowed morphophonemic regularities as well as more superficial ones to be accommodated within a phonological description. The success of this more abstract notion of sound structure in dealing with hitherto irresolvable problems in the description of stress (see *Suprasegmentals*) contributed greatly to its success, and the resulting theory of Generative Phonology as developed in the work of Halle together with Noam Chomsky rapidly became the dominant view in the field by the middle of the 1960s.

1.2 Phonology as a System of Rules and Representations

A basic insight in the development of Generative Phonology was the proposal that it is not only the representation of linguistic elements in terms of basic contrasts that matters: an adequate theory must characterize what a speaker knows about the sound system of the language, and that includes regularities of variation and alternation as well as inventories of basic elements. Combining these two aspects of phonological knowledge required the explicit recognition of

a system of rules (expressions of regular patterns of contrast and variation in sound shape) in addition to the theory of representations. Developing an adequate theory of phonological rules, in turn, necessitated a notion of phonological representation that was related to surface phonetic reality in a much more complex and indirect way than the phonemic representations of structuralist linguistics. The central problems of phonological theory came to be articulated in terms of the theory of rules, their nature, and their interaction, and the underlying phonological representations that need to be posited in order to allow them to be expressed in their full generality.

2. *Issues in Phonological Theory*

As reviewed in Sect. 1.1 above, the problem of phonological description was originally conceived as a matter of discerning the range of contrastive sound units in a language, and arranging them in a system that brings out the dimensions of their contrast. When phonology is seen as a form of knowledge, however, as generative phonologists have stressed, the sound structure of natural language takes on quite a different form and presents quite different problems. Among these are (a) the nature of underlying ('phonological' or 'phonemic') representations as well as surface ('phonetic') representations; (b) the ways in which phonological regularities serve to relate phonological to phonetic representation, including the interactions that may obtain among regularities; and (c) the relation between phonological form and other aspects of linguistic knowledge, such as word structure (see *Morphology in Linguistics*) and sentence structure (see *Syntax-Phonology Interface*). Various aspects of these questions are addressed below.

2.1 *The Abstractness of Phonological Representation*

Initial reaction to the proposals associated with generative phonology centered on its abandonment of a phonemic representation based on the condition of bi-uniqueness. Relaxing this defining characteristic of Structuralist phonemics led to phonological representations that were considerably more abstract in their relation to phonetically observable properties. The proposals of Chomsky and Halle (1968) concerning the analysis of English, for example, involved positing final 'silent *e*' in words like *ellipse*, *burlesque* (phonologically /Ellipse/, /bVrleske/), geminate consonants in words like *confetti* (/kVnfetti/), a distinction among /k, s, c/ such that *acquiesce* is phonologically /ækwiesce/, etc. Indeed, there were no constraints whatsoever on the relation between phonological and phonetic form, apart from the desire to set up underlying forms from which as much variation as

possible in the shapes of particular morphological units could be predicted. Since much of this variation is the residue of earlier sound changes that have affected the same element in different ways in different environment, many early generative analyses resembled historical accounts of the language more than they did the knowledge of current speakers.

The perception of such apparently excessive abstractness led to proposals for constraining the operation of phonological rules in grammars, and the kinds of representation that should be posited. Kiparsky (1973), in particular, suggested a variety of conditions that would have the effect of prohibiting rules of 'absolute neutralization' (by which some posited phonological distinction is eliminated uniformly in surface forms, such that it never corresponds to a phonetic distinction) and other abuses of the theory's representational freedom. This led phonologists to seek external evidence for the reality of the representations they assumed, in the form of data from poetic systems, historical change, language games, behavioral experiments, and other sources outside of the phonology itself. Such a notion would have been incoherent on the assumptions of Structuralist phonology, for which the properties of the system are a matter of its internal logic, but became important with the shift to a view of phonological form as a matter of speakers' knowledge.

It is now apparent that phonological representations, while more abstract than assumed by structuralists, are still rather constrained in the extent to which they can deviate from phonetic form. There is no real consensus in the field, however, as to what constraints are appropriate on the operation or form of grammars so as to ensure an appropriate result. In the absence of some motivated grounds for decision in general, the status of specific conditions remains somewhat arbitrary. Replacing the conception of a grammar as a set of rules by that of a set of simultaneously evaluated partially violable constraints, as discussed in a later section, severely limits the potential abstractness of underlying representations in a way that may prove to resolve these issues in a principled way.

2.2 *The Interaction of Phonological Regularities*

In the nature of a Structuralist phonemic representation, all of the regularities expressed in the grammar are mutually independent. That is, it is only necessary to know the phonemic environment of a phoneme to predict its phonetic realization; and it is only necessary to know the phonetic environment of a phonetic segment to establish its phonemic correlate. When the bi-unique relation between these levels of representation is relaxed, however, more complex possibilities arise for regularities to interact with one another.

Consider the formation of English regular plurals, for example. For the three words *cut*, *dog*, *horse*

(roughly [kæt], [dog], [hors] respectively, phonetically) and others like them, the corresponding plurals are *cats*, *dogs*, *horses* ([kæts], [dogz], [horsɪz]). Assume that the phonological representations of the nouns are essentially the same as their phonetic forms, and that the regular plural ending has a single constant phonological form: /z/ for concreteness' sake, though the point to be made is independent of this choice. Now there are two distinct regularities that are involved in determining the pronunciation of the plural ending:

(a) If the noun ends in a sibilant ([s, z, ʃ, ʒ, č, ʝ]), a vowel [ɪ] appears between the stem and the ending; and

(b) If the stem-final sound is voiceless, the ending is voiceless ([s]) as well.

Now consider how these principles interact in determining that the pronunciation of *horses* (phonologically /hors + z/) should be [horsɪz]. In this case, the conditions for both of the rules above, *b* as well as *a*, are met, and we might expect both to apply, yielding (incorrect) [horsɪs]. The relevant observation is the following: the vowel inserted as a result of rule *a* has the effect that the stem and the ending are no longer adjacent, and so rule *b* is inapplicable. That is, the regularity represented by rule *b* in this case presupposes that the set of clusters of consonants which will remain adjacent in surface forms (a matter which is potentially affected by applications of rule *a*) has been determined. This logical relation of presupposition between rules is generally expressed by saying that rule *a* applies 'before' rule *b*, though it is important to note that no claim of temporal priority is thereby asserted, but rather a purely logical notion.

Much of the abstractness of early generative phonological analyses was made possible precisely by the availability of rule ordering as a descriptive device: the possibility of specifying rule interactions such as the above as an independent parameter within grammars. Although arguments for the necessity of such ordering formed an important part of the literature, and particular accounts sometimes involved apparently crucial depths of ordering as great as twelve to fifteen, many linguists felt that (as part of the general campaign to reduce abstractness in phonology) stipulated ordering relations ought not to be permitted as a descriptive device. Despite the existence of cases apparently requiring irreducible specification of relative order (e.g., dialects with the same sets of rules and underlying forms, differing in their surface forms only as a consequence of different interactions among the rules), and the fact that all proposed formulations of principles from which observed rule interactions could supposedly be predicted had obvious and well established counter-examples, a consensus developed in the field that such 'extrinsic' rule ordering statements ought to be prohibited. This has left a number of kinds of empirically observed interactions unaccounted for, a problem which has persisted as phonologists have turned from rules to constraints as descriptive devices.

2.3 The Structure of Representations in Phonology

As already noted, the description of speech at either the phonetic or the phonological level as composed of a sequence of discrete segment-sized units is a significant abstraction from physical reality. Its justification comes not from the observable facts, but from the extent to which it allows the analyst to uncover what is orderly and coherent in linguistic structure. By the mid 1970s, however, it had become apparent that a purely segmental organization of representations impeded the description of linguistic regularity in several respects (see *Phonology, Nonlinear*).

One of the first of these problems to be discussed arose from the analysis of systems of tonal contrasts, common in the languages of China, Africa, and in fact much of the world. Careful analysis revealed two ways in which tonal properties were problematic for strictly segmental models: on the one hand, what appeared to be a single tonal specification might take as its scope more than a single segment (perhaps the vowels of several consecutive syllables, or an entire word); and on the other hand, what appeared to be a single segment (a unitary short vowel, for example) might have a tonal specification involving two or even more consecutive tonal levels. If it is assumed that each of the consecutive segmental units of which a representation is composed bears exactly one specification for each potentially distinctive property, and the specifications of distinct segments are independent of one another, both of these situations are anomalous.

As a result of these observations, the segmental view came to be replaced with an *autosegmental* notion of representation. On this picture, specifications for each potentially distinctive property (or feature) succeed one another discretely, but an additional dimension of representation is the specification of the way these are synchronized or aligned with one another. In the limiting case, where each specification for a given feature is temporally aligned with one and only one specification for each of the other features, the standard segmental picture arises, but more generally, one or more instances of many-to-one alignment may be found. Once this view was articulated, it became apparent that many properties other than those of tone were also most appropriately described in this way. Nasality, in particular, behaves similarly in many languages and the very common phenomenon of assimilation is often best treated as an alteration in the scope of some features rather than a change in their values.

On the other hand, the full flexibility of autosegmental representations predicts much more independence of specification than is actually found. Features of place of articulation typically behave as a unit, despite their logical independence from one another, as do the features characterizing laryngeal activity. This observation gave rise to the notion that phonological features are not just an unordered set, but

instead are organized into a sort of hierarchical arrangement, such that for example features of consonantal place can be referred to as a unit. There is considerable consensus on the overall form such a *feature geometry* should take, but many details and relations among properties remain unsettled.

Another representational issue concerns the existence of structural units larger than the single segment, such as the syllable. Classical generative phonology, as represented by, for instance, Chomsky and Halle (1968), makes no appeal whatever to syllables (or any other unit above the segment). This was not simply an omission on their part, but rather a systematic claim to the effect that segmental specification could always suffice to express any phonological generalization: that is, that any observation involving syllables (of which there are many in Chomsky and Halle 1968) could be satisfactorily reformulated in a way that dispensed with any units other than segments. The boldness of this claim invited immediate attack, and indeed it soon became clear that there were indeed aspects of phonological structure that required reference to syllables and other supra-segmental units. The resulting theory of the *prosodic hierarchy*, by which segments are grouped into syllables, themselves grouped together into *metrical feet*, which in turn form constituents of *phonological words* (perhaps with some additional categories, according to some writers) has become another standard assumption about phonological structure.

2.4 Interactions between Phonology and Other Areas of Grammar

The original conception of a generative phonology was as a component of the grammar that served to map lexically specified syntactic structures onto phonetic form, a representation appropriate to serve as the instructions to the language-independent mechanisms of speech production. This picture logically entails the notion that words, phrases, and sentences are already fully assembled in the input to the phonology. This assumption, however, has proven to be problematic.

The notion that phonological properties interact with the syntactic environment in which forms occur is quite a familiar one, especially in the description of accent and intonation (see *Syntax-Phonology Interface*). Word formation, too, may depend in some instances on the assignment by phonological rule of phonological properties (see *Morphology in Linguistics*) in a way that contradicts the assumption that phonology applies strictly to the output of this part of the grammar.

The most extensively elaborated picture of how phonology interacts with the rest of grammar is presented by the theory of Lexical Phonology (see papers in Hargus and Kaisse 1993 for introduction and explication). Briefly, this picture distinguishes

between lexical and post-lexical regularities, with the former applying in a way that interacts with word-formation, and the latter applying to a representation in which fully formed words appear in their syntactic context. The lexical rules, in turn, may (depending on the language) be divided into two or more *strata*, for example, rules applying to *roots*, *stems*, or *words*. A process of word formation takes some lexical unit as its input and yields an output (affixed or otherwise altered in form), with this output then subject to appropriate phonological modification. The resultant representation may serve as the input to further *cycles* of word formation plus phonology, until a complete surface word is arrived at.

These units, assembled into larger syntactic constructions, are then subject to the post-lexical rules. While various forms of lexical phonology generally constitute the background assumption of much phonological discussion, the correctness of such a view of the overall architecture of grammar has come into question. In particular, the rise of constraint based theories implying a single stage in the conversion of underlying to surface form has appeared incompatible with the sequential, derivational character of lexical phonological description. Some researchers have sought to reconcile these two trends in phonological theory, while others have attempted to augment one or the other sub-theory in such a way as to include the results of the other.

2.5 Rules vs. Constraints in Phonological Description

Formulations of phonological regularities since the 1950s have drawn on the mechanisms of automata theory: standardly, these have been expressed as rewriting rules that map an input representation onto an output, one step at a time. Developments in the theory of computation, however, have given rise to an alternative possibility. Instead of converting inputs to outputs in several stages, with each step involving the imposition of a single regularity in the computation of a new representation, one might formulate all of the regularities as elements of a single system of constraints. If these constraints are ranked with respect to one another, it is possible to say that any particular constraint may be violated in the output, provided that such a violation is motivated by the need to avoid violating a more important (higher ranking) constraint. The resulting theory provides an architecture of grammar that is (at least apparently) very different from that which has dominated most of the literature in phonology. For example, the kinds of interaction among regularities treated above as matters of rule ordering can no longer be described in this way, since all of the constraints in the grammar apply (in principle) in a single, simultaneous block. Constraints can have a language-particular ranking among themselves, but this is a matter of priority, not logical

presupposition. Other differences have been alluded to in the discussion above.

Constraint based theories along the lines of Optimality Theory (Prince and Smolensky 1993) have been shown to have a number of significant advantages over rule-based accounts, and currently constitute the most active area of investigation in phonological theory. It remains to be seen, however, what character a comprehensive theory will have once the substantial body of results achieved within earlier, apparently incompatible frameworks has been incorporated comprehensively.

See also: Phonetics: Articulatory; Phonology, Non-linear; Saussure, Ferdinand de (1857–1913); Speech Perception; Speech Production, Neural Basis of; Speech Production, Psychology of; Syntax–Phonology Interface

Bibliography

- Anderson S R 1985 *Phonology in the Twentieth Century: Theories of Rules and Theories of Representations*. University of Chicago Press, Chicago
- Baudouin de Courtenay J 1972 *Selected Writings of Baudouin de Courtenay*. Indiana University Press, Bloomington, IN
- Bloomfield L 1933 *Language*. Holt, New York
- Chomsky N, Halle M 1968 *The Sound Pattern of English*. Harper and Row, New York
- Coulter G R 1993 *Current Issues in ASL Phonology (Phonetics and Phonology 3)*. Academic Press, San Diego
- Fischer-Jørgensen E 1975 *Trends in Phonological Theory*. Akademisk Forlag, Copenhagen, Denmark
- Hargus S, Kaisse E 1993 *Studies in Lexical Phonology (Phonetics and Phonology 4)*. Academic Press, San Diego
- Harris Z 1951 *Methods in Structural Linguistics*. University of Chicago Press, Chicago
- Jakobson R 1941 Kindersprache, aphasia und allgemeine lautgesetze. In: Jakobson R (ed.) *Selected Writings, Vol. I* (English trans. 1968 Child Language, Aphasia and Phonological Universals). Mouton, The Hague, pp. 328–40
- Joos M 1957 *Readings in Linguistics*, Vol. 1. American Council of Learned Societies, Washington, DC
- Kenstowicz M 1994 *Phonology in Generative Grammar*. Blackwell, Oxford, UK
- Kiparsky P 1973 Phonological representations. In: Fulimura O (ed.) *Three Dimensions of Linguistic Theory*. The TEC Corporation, Tokyo
- Prince A, Smolensky P 1993 *Optimality Theory: Constraint Interaction in Generative Grammar*. Rutgers University, New Brunswick, NJ
- Sapir E 1925 Sound patterns in language. *Language* 1: 37–51
- de Saussure F 1916 *Cours de linguistique générale*. Payot, Paris
- Trubetzkoy N S 1939 *Grundzüge der Phonologie*. Travaux du Cercle linguistique de Prague 7

S. R. Anderson

Phonology, Nonlinear

Nonlinear phonology is a loose collection of relatively recent theories that address various aspects of the sound structure of human languages. Traditional approaches to phonology are linear, which means that they represent each word as nothing more than a string of speech sounds (called segments). Each segment is a bundle of its phonetic attributes, called distinctive features, but no further structure is assumed. Nonlinear phonology posits a richer architecture: segments are organized into a hierarchy of constituents, such as the syllable, the stress foot, and so on; and distinctive features are liberated from the segmental feature-bundle, allowing them to have domains that are larger or smaller than a single segment. This article examines some of the consequences of this hypothesis, focusing on two particular implementations of it.

1. Historical Background

Before the mid-1970s, most research in phonological theory started from the assumption that segments are linked together like beads on a string (see *Phonology*). In this view, phonological representations are very similar to alphabetic writing, with exactly one symbol standing for each segment and with little or no structure beside the linear order of the segments in a word.

The best known theory of this general type is the one in Chomsky and Halle's (1968) *The Sound Pattern of English* (SPE). The SPE theory is rigorously and consistently linear: for example, syllables play no role at all, and constituents such as the word have only a derivative status. The only departure from the alphabetic model is that each segment is assumed to be a bundle of distinctive features standing for its various attributes (such as nasal, voiced, rounded lips, and so on).

During the 1940s, two important precursors to contemporary nonlinear phonology emerged independently in the work of J. R. Firth and Zellig Harris. They shared the idea that phonological properties could have domains larger than a single segment. For example, in the South American language Terena, the meaning of 'I' or 'my' is marked by nasalizing (indicated by a tilde) a sequence of segments beginning at the left edge of the word: *owoku* 'house', *ōwōngu* 'my house'; *arine* 'sickness', *āřinē* 'my sickness'. Firth or Harris would say that the domain of nasality in Terena is (approximately) the word. Their theories did not have much influence, though, and their work was largely neglected for several decades.

2. Nonlinear Phonology

In reaction to SPE's strongly segmental orientation, the mid-1970s saw a profusion of nonlinear ap-

proaches to phonological phenomena, supplying ideas and insights that continue to be studied in the present day. These models were developed to address empirical problems that had been disregarded or treated unsatisfactorily in SPE: tone, long distance assimilation processes, syllables, the typology of stress systems, and the interaction of phonology with morphology and syntax.

Two leading ideas have emerged. One is *autosegmentalism*, the notion that distinctive features, rather than segments, are the atoms of phonological representation. (This term was coined by John Goldsmith.) From an autosegmental perspective, a distinctive feature is not an attribute of a segment, as in SPE, but a free standing entity. This status allows the feature to act independently of its segmental host, approximately as in Firth's and Harris's theories. The other leading idea is the *prosodic hierarchy* (a phrase coined by Elisabeth Selkirk), which organizes segments into successively larger constituents such as the syllable or word. The prosodic constituents are analogous to the more familiar syntactic constituents such as sentence or noun phrase (see *Constituent Structure: Syntactic*), and, as in syntax, structural requirements imposed on the constituents are an important analytic tool.

Autosegmentalism and the prosodic hierarchy are complementary, not competing, hypotheses. As a matter of logic, they are not mutually exclusive, since autosegmentalism speaks to the nature of subsegmental units (distinctive features), while the prosodic hierarchy is a claim about how segments themselves are organized into larger groupings. They have also been applied to factual domains that are mostly distinct. Autosegmentalism helps to solve problems in tone and featural phonology, such as assimilation. The prosodic hierarchy is important in understanding the phonology of syllables, stress, words, and phrases.

Autosegmentalism and the prosodic hierarchy are, ultimately, claims about how language is represented in the human mind. Though they allow for a considerable range of variation between languages, they also have a core that is claimed to be universal. Hence, the main evidence for these theories comes from observations about phenomena that recur in many languages.

The rest of this article discusses the implementation of these two aspects of nonlinear phonology, the results and controversies that emerge from them, and their present and future prospects.

3. Autosegmental Phonology

The following example illustrates one of the key arguments for autosegmentalism. Some speakers of the Bantu language Sanga use a secret language (analogous to pig Latin in English) called Kinshingelo. Words in this secret language are formed by sys-

tematically permuting the segments of Sanga words, as in the following example:

Sanga
báá.kó.lwèé bàà.dyáá má.tá.bá áá.ŋkáá.mbò
Kinshingelo
báá.lwé.kòó dyàà.báá má.bá.tá áá.mbòó.ŋká
Gloss

'The birds are eating my grandfather's corn.'

The accents mark tones: acute *ó* is high and grave *ò* is low. Double vowels are long, so *áá* stands for a long *a* vowel with a rising (i.e., low–high) tone. For expository convenience, syllable divisions have been indicated with a full stop.

A Kinshingelo word is formed by swapping the last two syllables of each Sanga word. But not precisely: the segments themselves are swapped, but their tones and length are not. Speakers of Sanga unconsciously regard *áá* as somehow divisible into three components: a part that describes the action of the tongue and lips in forming consonants and vowels, a part that describes the action of the larynx in controlling tone, and a part that describes the duration of the vowel. A useful analogy is an orchestral score, which contains distinct but coordinated instructions for many, different instruments. This is the central insight of autosegmental phonology: speech sounds consist of independent components which are coordinated in time.

Tone is the premier example of an independent component of speech, and in fact autosegmental phonology was first developed to deal with tone. Some other consequences of autosegmentalism for the analysis of tone include:

(a) *Persistence under deletion*. When a vowel deletes, its tone may remain and attach itself to some nearby vowel.

(b) *Floating tones*. When a vowel deletes, its tone may remain and express itself indirectly, by affecting following tones.

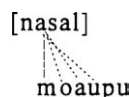
(c) *Toneless syllables*. Some syllables have no tonal specification at all. Their pitch is determined by linear interpolation between syllables that have tones.

(d) *Tone shift*. In the Bantu language Kikuyu, every tone shifts exactly one syllable to the right, domino fashion, so /tòmátómàyà/ becomes tòmátómàyà

(e) *Tone melodies*. In some languages, words are limited to a small set of fixed tone melodies, or tone melodies may mark certain morphological distinctions.

In sum, tones and segments may behave independently of one another. This is the expected result of the autosegmental hypothesis (see also *Suprasegmentals*).

Not long after it was developed, autosegmental phonology was applied to nontonal phenomena as well. The idea is that every distinctive feature—[nasal], [voiced], and so on—occupies its own separate autosegmental 'tier,' allowing it to function independently of segments or other features. The main focus of

**Figure 1**

Autosegmental representation of Warao *mōāũpu*

research in this area has been on two related problems: assimilation and harmony.

In assimilation, one segment (the target) takes on some of the features of another, nearby segment (the trigger). Harmony is a kind of long distance assimilation, where the triggering segment and its target are not necessarily adjacent. Here are some examples from the South American language Warao, and the North American language Chumash:

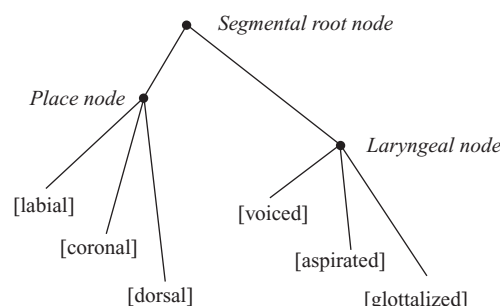
- (a) Warao /moaupu/ → *mōāũpu* 'give them to him'
- (b) Chumash /k-sunon-š/ → *kšunonš* 'I am obedient'

In Warao, the feature [nasal] assimilates, affecting a potentially unbounded sequence of vowels after a nasal consonant. In Chumash, the feature [anterior] assimilates, changing *s* to *š* (the initial consonant in *ship*) if followed anywhere in the word, at whatever distance, by another *š*.

Autosegmentalism contributes to the understanding of these processes in two ways. First, it provides a natural interpretation of assimilation. Because features and segments are independent, the domain of a feature can be larger than a single segment. Assimilation expands the domain of a feature from one segment to many. Second, autosegmentalism allows long distance harmony processes to be analyzed as local on the tier of the assimilating feature. These two key insights are usually expressed graphically, as in Fig. 1. The feature [nasal] is shown on its own autosegmental tier, separate from the rest of the segment. Temporal coordination between [nasal] and the rest of the segment is indicated by association lines. The solid association line is original, while the dotted association lines are the result of the process of assimilation. The effect of assimilation here is to expand the domain or scope of [nasal] to include the vowels *oau*.

The most radical autosegmental theory is one in which every distinctive feature is on a separate tier, with complete freedom to act independently of all other features. This view cannot be right, because there are strong functional associations among various subsets of the features. For example, the features that characterize states of the larynx have a considerable degree of cohesion in phonological systems, assimilating or neutralizing together.

Functional groupings of features are addressed in theories of feature geometry. (This term was proposed by G. N. Clements.) The idea is to posit a layer of representation between the segment and the features.

**Figure 2**

A model of feature geometry

Elements of this intermediate layer, called class nodes, are constituents corresponding to the observed functional groupings. A typical implementation, abstracting away from controversial details, appears in Fig. 2. The laryngeal features often act together in phonological processes, and so there is a constituent, the Laryngeal node, that unites them. Similarly, a single process will often affect the various features that distinguish *p* from *t* from *k*, and so there is a constituent, the Place node, to unite them as well. Feature geometry, like autosegmental phonology itself, is claimed to be universal. Research in this framework is oriented toward discovering a set of class nodes that will account for all (or at least all common) phonological processes in the world's languages.

4. The Prosodic Hierarchy

The term prosody is used by phonologists to refer to the durational, rhythmic, and phrasing aspects of speech. The study of prosody includes syllables, stress, restrictions on the size or shape of words, and many of the ways in which morphology or syntax can affect pronunciation (see *Word, Linguistics of; Linguistic Typology*).

A key insight into the nature of prosody is the idea of a small, universal set of prosodic categories arranged into a hierarchy (see Fig. 3). Certain additional assumptions give the hierarchy predictive value:

(a) *Grouping and layering*. Level *n* of the hierarchy imposes a grouping structure on the layer just below it, level *n* – 1.

(b) *Headedness*. Every constituent at level *n* designates a specific, most prominent constituent at level *n* – 1 as its head.

(c) *Structural requirements*. There are global and parochial conditions on the well-formedness of constituents. For example, languages frequently require that feet be binary, and there is also evidence of binarity conditions on other levels of the prosodic hierarchy. So, proceeding from the bottom of the hierarchy, a syllable is a grouping of segments; the head of the

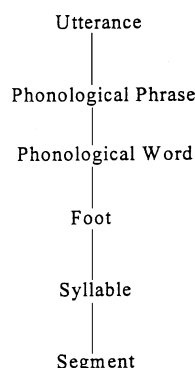


Figure 3

A prosodic hierarchy model

syllable is the peak of acoustic energy, usually a vowel. A foot is a grouping of (usually two) syllables, one of which is more prominent than the other. The more prominent syllable is the locus of linguistic stress. A phonological word is a grouping of feet, one of which is also the most prominent (the locus of main stress). A phonological phrase is a grouping of phonological words that also includes a most prominent member, and so on.

Results of surprising subtlety and generality can be derived from these modest premises. For example, many languages set a lower bound on the size of the phonological word. The Australian language Diyari is typical in this respect: it has no monosyllabic words at all. This so-called minimal word requirement follows from joint action by three different constraints: layering (the phonological word dominates the foot in the prosodic hierarchy); headedness (every word must contain at least one foot, to serve as its head); and a structural requirement that feet be binary, since they are the basic units of stress rhythm. Comparable results can be found at other levels of the prosodic hierarchy.

5. Present and Future Prospects

Nonlinear phonology has been an active area of research since the late 1970s. The basic autosegmental hypothesis enjoys wide and perhaps universal acceptance. Various details are matters of debate, however:

- Do all features have the same autosegmental status as tone?
- Is feature geometry an appropriate model of featural organization?
- If so, exactly how are the features organized?
- Are (binary) features the appropriate units of subsegmental structure?

The prosodic hierarchy is somewhat more controversial. Although notions of constituency do play a

role in most theories of prosody, there are also models that are partly or entirely constituent-free.

In recent years, much research has focused on the ways in which Optimality Theory can illuminate topics in nonlinear phonology. A particularly productive line of investigation has been the study of constraints that align the edges of prosodic constituents with morphological or syntactic ones. These constraints relate the prosodic hierarchy to the hierarchy of grammatical constituents, offering a new perspective on the interface between phonology and grammar.

See also: Phonology

Bibliography

- Chomsky N, Halle M 1968 *The Sound Pattern of English*. Harper & Row, New York
- Clements G N 1985 The geometry of phonological features. *Phonology Yearbook* 2: 225–52
- Clements G N, Keyser S J 1983 *CV Phonology: A Generative Theory of the Syllable*. MIT Press, Cambridge, MA
- Goldsmith J A 1990 *Autosegmental and Metrical Phonology*. Blackwell, Oxford, UK
- Hayes B 1995 *Metrical Stress Theory: Principles and Case Studies*. University of Chicago Press, Chicago
- van der Hulst H, Smith N 1985 The framework of nonlinear generative phonology. In: van der Hulst H, Smith N (eds.) 1985 *Advances in Nonlinear Phonology*. Foris Publications, Dordrecht, The Netherlands
- Kenstowicz M 1994 *Phonology in Generative Grammar*. Blackwell, Cambridge, MA
- Lieberman M, Prince A 1977 On stress and linguistic rhythm. *Linguistic Inquiry* 8: 249–336
- McCarthy J, Prince A 1993 Generalized alignment. In: Booij G, van Marle J (eds.) 1993 *Yearbook of Morphology*. Kluwer, Dordrecht, The Netherlands
- McCarthy J, Prince A 1995 Prosodic morphology. In: Goldsmith J A (ed.) 1995 *The Handbook of Phonological Theory*. Blackwell, Oxford, UK
- Nespor M, Vogel I 1986 *Prosodic Phonology*. Foris, Dordrecht, The Netherlands
- Odden D 1995 African tone languages. In: Goldsmith J A (ed.) 1995 *The Handbook of Phonological Theory*. Blackwell, Oxford, UK
- Pierrehumbert J, Beckman M 1988 *Japanese Tone Structure*. MIT Press, Cambridge, MA
- Selkirk E 1981 On the nature of phonological representation. In: Laver J, Meyers T, Anderson J (eds.) *The Cognitive Representation of Speech*. North-Holland, Amsterdam
- Selkirk E O 1984 *Phonology and Syntax: The Relation Between Sound and Structure*. MIT Press, Cambridge, MA

J. J. McCarthy

Photography as a Medium

For over a century and a half, photography has been the most popular and widespread means of making images. The first and primary tool of photography is the camera, and changes in the medium and its

applications are often attributed to technical developments in making, reproducing, and distributing photographs. Equally if not more significant, however, have been changes in the ways people think about photography, including ideas about what and how photographs represent, and the roles they have played in social, political, and economic life. This article examines developments and conflicts in the idea of photography within a historical framework of changing cultural practice.

1. Photography's Origins

The invention of perspective in the fifteenth century laid the foundation for the system of visual representation that became photographic. The camera obscura, used by scientists, artists, and magicians since the Renaissance, is the most important precursor to the camera. By the eighteenth century this device had evolved from a large clumsy apparatus for projecting an image onto a plane surface, to a more portable instrument, used for the curiosity and pleasure of looking at images of all sorts of objects and scenes projected by the lens.

A knowledge of light-sensitive materials was particularly available in Europe by the 1720s, and one of the riddles of photography is why nearly a hundred years passed before this knowledge was applied to fixing the camera obscura's image. Despite a broad fascination with display forms, such as the diorama, that created illusions of realistic representations of nature, the idea of the photograph as a permanent copy apparently did not exist. Suddenly, in the 1820s a range of experimenters began simultaneously to compete to be the first to successfully fix the photographic image; the desire to photograph had emerged.

Photography's invention is generally dated to 1839, when breakthroughs were announced nearly simultaneously in France and England. Louis Jacques Mandé Daguerre's induplicable monochrome on a silver-coated copper plate, and William Henry Fox Talbot's reversed image—the negative—were the first successfully publicized and patented processes. The daguerreotype enjoyed an immediate popularity, setting off a craze which by the end of the year the French press had named '*daguerreotypomanie*.' Talbot's calotype system, on the other hand, based on a negative image that could be used to print a potentially infinite number of positives in as many different sizes, laid the foundation for all subsequent developments in photography. The refinement and diffusion of these two systems, alongside many competitors, had an immediate impact on European social life and spread quickly to North America and colonial outposts.

Photographs appeared in many display forms, as cartes-de-visites and stereographs, in albums and as lantern slides, and photography surfaced as a theme in popular music, on the stage, and in artifacts ranging

from home furnishings to fashion accessories. Photography as an *idea* permeated mid-century European popular culture.

2. Early Applications of the Medium

At the same time that photographic tools and processes were being simplified, an agenda was being established for how the new medium was to be used. This included the people, places, and events represented in photographs and, equally important, who was taking the pictures and who were the intended viewers.

2.1 Photographs of People

Within months of Daguerre's announcement, professional portrait studios had sprung up in Paris, and soon studios were established in other major European cities, in New York, and in Washington. From 1860 onward, portraiture expanded to provincial towns, to less industrialized European countries, and to Australia, China, Japan, Mexico, and South America.

The photographic portrait represented a democratization of the painted portrait, and was commissioned first for private use within a family. The rising status of certain photographers led to commissions from men in public life whose positions required an official portrait. Most portraits of individuals followed the strict visual conventions of the painted portrait, a head-and-shoulders composition, with the sitter's head slightly turned to avoid a direct gaze into the camera. Portraits of family groups were common, and a custom quickly developed of photographically commemorating important events such as weddings and graduations. Photographs of deceased relatives, lying peacefully in their caskets, were also highly valued.

In the 1860s, studios began to turn out millions of inexpensive *carte-de-visites*, full- or bust-length portraits of workers, tradespeople, members of the bourgeoisie and aristocracy, native people of the colonies. These mass-produced images were often instrumental in establishing the status of a politician or celebrity, and both the French and British royal families permitted the sales of their images as *cartes*. Through its broad distribution, the *carte-de-visite* also contributed to forming Western ideals of feminine beauty and fashion, as well as ideas about people from the nonindustrialized parts of the world as exotic and other. The *carte* craze further generated a surrounding industry of albums and holders for cataloguing and exhibiting the growing image collections.

In 1888 George Eastman launched the first Kodak camera, with the slogan 'You push the button, we do the rest,' and immediately opened the practice of photography to a much broader segment of the

population. Studio portraiture declined as amateurs began to photograph family and friends. The repertoire of photographic events was also expanded as picture-taking became associated with holidays, family celebrations, and vacation trips. Photographic practice had entered the domestic domain, forming a bridge between personal memory and public history. A hundred years after the first Kodak camera, photography was fully integrated into family social life. In the mid-1980s, for example, over 90 percent of US families owned a still camera and amateurs were taking more than 11 billion photographs a year (Chalfen 1987). The majority of these pictures were of family members and friends and were rarely seen outside that social group. Nonprofessionals' private photography today accounts for the largest single application of the medium. Studies have shown that people attach enormous value to these images, and losing photographs of family and friends, for example, through fire or war, is often experienced as a greater tragedy than the loss of other material possessions.

2.2 Distant Worlds

The early years of photography coincided with the period of colonial expansion and exploration of newly won territory for the industrializing nations of Europe and North America. Wonders of nature, the dynamic new cityscape and its architecture, and the exotic cultures of distant lands held a fascination and attraction for many. The camera was an ideal tool for showing what was new or previously unknown to an audience eager to marvel over sights they could not see with their own eyes. Photography also opened up new territories for tourism by people who, having seen a place in pictures, now had the desire and means to experience it themselves. A new aesthetic of the landscape also stimulated tourism, as for example when photographs of the rocky Scandinavian coasts brought British and German tourists north in the 1870s (Rosenblum 1984, p. 112).

Photographs of popular European tourist sites, such as the city of Rome and its ruins, were marketed as stereographs. Distant lands of North Africa and the Near East became accessible to the armchair tourists of Europe as early as the 1850s through Francis Frith's photographs. Frith's *Egypt and Palestine Photographed and Described* (1858) was the first of 11 volumes and three sets of stereographs (Rosenblum 1984, p. 122). His work was motivated in part by a concern over growing French influence in the region, and must be seen in the broader context of British expansionist goals of the period. Local Egyptian photographers also opened studios and marketed their work to the growing European fascination with views of the ruins and landscape of the Near East. Similar patterns developed in India and parts of the Far East. By 1884, for example, several thousand native pho-

tographers were in business in China, serving both local clients and the Western trade.

In North America, photography was linked to Western expansion. Photographers accompanied survey teams in order to document areas for future mineral exploitation, rail routes, and civilian settlement, at the same time that they were contributing to the visual discourse of the American frontier and its native peoples. The natural wonders of the wilderness, including Niagara Falls, were immensely popular stereographic themes, used in lantern slide shows and collected in albums. This iconography and the photographers who produced it, such as Carleton Watkins and William Henry Jackson, established the grandeur of the mountain landscape as a symbol of transcendent idealism. Their work also contributed to lobbying efforts to establish a national park system in order to save the pristine wilderness from commercial exploitation by private interests.

Nor did the city escape the camera's gaze. In contrast to images of untouched wilderness and ancient unchanging ways of life, here the focus was on physical transformation. New transport systems, bridges, and buildings made possible by new technologies and materials were transforming the city into a modern metropolis. Urban expansion was a source of national pride, and both governments and private interests supported photographic documentation of the major cities of Europe and North America. Cityscapes in panorama format were especially popular, a parallel to the painted panoramas and dioramas that had emerged as mid-nineteenth century forms of entertainment.

Photography's prominent and various positions in the emergent discourse of modernity was never more evident than in the world expositions, the late-nineteenth century spectacles of science and marketing, combining knowledge and entertainment. Beginning with the Crystal Palace in London in 1851 and followed every few years in Britain, France, and the United States, these major events both symbolized and displayed industrial expansion, made possible by new technologies and materials. Displayed side by side with the symbols of modernity were the exotic products and peoples of the underdeveloped nations, the colonies. At this point in the medium's development, photography provided an unproblematic means of expanding knowledge of the visible world, and served 'the general urge in all industrialized societies to measure, describe, and picture the physical substances of all things on earth and in the heavens' (Rosenblum 1984, p. 140).

3. Concepts of the Image

Photography's popularity as a democratic form of self-documentation, intertwined with colonial interests and industrial expansion, point to the camera as a

tool that linked seeing and knowledge in new ways. Within the sciences, in the arts, and in social reform movements, distinct and often conflicting visual discourses have developed around the photograph as a source of insight and knowledge. We turn now to the ways of seeing with the camera that were generated within each of these broad fields.

3.1 Photography and Science

Within the discourses of science, photography was to be a witness to scientific observation with greater accuracy than the human eye was capable of. In medicine, photography provided close visual description of the human body, and was an aid in documentation and diagnosis of medical problems and also psychological reactions and mental aberrations. Examples are portrait photographer Professor John William Draper's book *Human Physiology* (1856), illustrated with woodcuts based on photographs and Dr. Hugh Welch Diamond, also known for his pictorial art photography, who documented women inmates of the Surrey County Asylum. Charles Darwin's *The Expression of Emotions in Man and Animals* (1872) was illustrated with photographs from a variety of sources. Microscope lenses were developed and the discovery of X-rays by Conrad Wilhelm Roentgen (the 1901 Nobel Prize recipient) led to their immediate use as camera images in medical diagnosis. Motion studies and stop-motion experimentation depended heavily on photographic documentation. Eadward Muybridge's 100,000 images analyzing movements involved in animal locomotion and a range of human activities (including walking, curtsying, and bricklaying) provided a new if contested way of seeing that also influenced the arts, in particular Cubist and Futurist painters.

Taxonomy was a unifying concept of the nineteenth century scientific paradigm, and studies of the human species were formed by the delineation and elaboration of physical and cultural distinctions. Anthropology became a dominant discourse of explanation, based on historical and biological notions of culture, and relying heavily on comparative taxonomic methods in order to systematically catalogue the world's peoples (Edwards 1997). Photography was central to the anthropometric systems developed to produce comparable data for analysis and classification. The system used by the distinguished British biologist Thomas H. Huxley for the Colonial Office, for example, involved posing individuals naked with a measuring rod in order to produce 'a systematic series of photographs of the various races of men comprehended within the British Empire' (cited in Edwards 1997, p. 56). These studies of human physiogony, used to expand evolutionary theory, contained a strong moral dimension, concerned with both creating and sustaining the complex system of social knowledge in which the images operated. Variations on the method were

applied in hospitals, prisons, and other institutions where photographs of inmates were used to establish and classify social types.

The 1870s and 1880s also brought about an awareness of the contributions photography could make to anthropological fieldwork, by Edward B. Tylor and others. Yet there are few examples of systematic photographic documentation in ethnographic field studies. Early photography of the ways of life of diverse folk groups was most often carried out by missionaries and educators or by artists working in the pictorial tradition. Edward Curtis, Adam Clark Vroman, Robert Flaherty, and Frances Flaherty documented native American tribal life in its positive aspects, romanticizing in particular the close relationship to nature. Nor was it uncommon for ethnographic photographs from the turn of the century of peasant customs made in a pictorialist style to be used to advance the cause of nationalism, as in several Eastern European countries. The most common reference to an ethnographic photography worth emulating is the classic study of *Balinese Character* (1942) by Margaret Mead and Gregory Bateson, as the 'first saturated photographic research in another culture' (Collier and Collier 1986, p. 12). Visual anthropology at mid-century was more closely associated with documentary film than with photography. Ethnographic photography survived as an *ideal*, however, despite the general absence of large-scale systematic studies. The British 'Mass-Observation' project in the mill towns of northern England in the 1930s, for example, was designed as an 'objective documentation' of working class life, in the manner of an anthropological study.

The archive has worked hand in hand with the scientific endeavors of accumulating and systematizing evidence. Seen as a repository of visual knowledge, the archive is related to the ideal of a comprehensive collection of images that together represent the whole (other examples being the family album or collection of *carte-de-visites*). The structure of the archive and its classification system was intended to mirror the physical and social world that the collection represented. Following Michel Foucault's groundbreaking archeology of the human sciences (1970), museums of natural and cultural history, as well as police, medical, and other archives, can no longer be seen as disinterested institutions that sustain and perpetuate value-free systems of knowledge, but reflect instead the structures of political and economic interests that gave rise to them. The photograph's evidentiary authority is thus inseparable from the archive as a repository of knowledge and power.

3.2 Photography and Art

Contrary to the idea of the photograph as an objective visual record, it has been argued that photography is inseparable from the Western pictorial tradition. Particularly in its relationship to landscape and nature,

photography employed a set of conventions and aesthetic standards theorized by the European concept of the 'picturesque.' Emulating themes from the fine arts, many early photographers developed techniques that undermined the sharp literalness of the lens in their aspirations to make images that were at once truthful, beautiful, and inspirational. Exhibiting esthetic photographs in an appropriate context was a paramount goal of the turn-of-the-century Secessionist movements in Europe and the United States, and was important to establishing photography's status as art. Pictorialism in photography was a synthesis of the realism of the camera image, with frequently formulaic romantic compositions of peasant or working class life.

A complex and uneasy relationship between photography and the fine arts has continued, even as the repertoire of art photography has expanded. This is generally beyond the scope of this article, apart from several developments relevant to photography's relationship to the social sciences. On the one hand, themes drawn from peasant and working class life as well as of non-Western people integrated scientific inquiry with romantic images of an exotic or primitive Other as legitimate subjects of photographic art. On the other hand, attempts in the 1920s to define and explore the characteristics which distinguish photography from other visual arts laid the foundation for a more radical 'straight' photography, called New Realism or *Neue Sachlichkeit* in Europe and North America. The nonmanipulated or spontaneous photograph became a means of exploring and exposing familiar subjects in new ways. These different schools and genres of art photography became more or less systematized as museums began to add photographs to their collections. The Museum of Modern Art in New York is a noteworthy example of a leading museum where a series of influential curators established a vocabulary of photographic practice based on the premises of modernism. By the 1940s, a photographic canon was in place, including a history of art photography with specific artists, works, and genres represented in the museum's collections (Newhall 1949). Of particular importance to the social sciences was the respect this canon established for the genre of documentary photography as a legitimate form of esthetic practice.

3.3 The Photographer as Engaged Observer

Parallel to their idealized representations of European peasants, from the 1850s onward artists began to be concerned with the social and ethical consequences of industrialization, preparing the way for documentary photography's role in social reform campaigns. 'Before' and 'after' images became a staple of reform photography in the 1890s, as evidence of the positive effect the reform school or institution had on its

residents. Photographs from the slums of New York (by Jacob Riis) and of child labor conditions (by Lewis Hine) were used to raise the fears and sway the consciences of politicians and a middle-class public. The photography of social reform was closely linked to new means of dissemination: the half-tone process for mass-producing photographs had been developed in the 1890s and a range of new periodicals arose in its wake.

Alongside the popular press' increased use of photographs after the turn of the century, social reform journals were among the first to recognize photography's ability to advocate a cause. Here the *American Charities and the Commons* is noteworthy. The *American Journal of Sociology* was also using photographs to illustrate articles, for example, on housing conditions. This visual material, often by women who held administrative positions in social action agencies, later declined under Albion Small's editorship, as the journal aligned itself with sociology's move to establish itself as a scientific discipline (Stasz 1979).

By the 1920s, documentary photography, as the genre came to be known, was recognized as an effective means of drawing attention to social issues. The best known example of such an effort was the photographers who worked for the US Farm Security Administration (FSA) under the leadership of Roy E. Stryker in the late 1930s, documenting the effects of the depression on rural Americans and the need for government aid. In other cases, photographers used the documentary form to draw attention to marginalized social groups, occasionally drawing on their own experience as members of the culture they documented. One of the clearest ideological expressions of photographers' identity was the Worker Photography movement, most active in Germany between World Wars I and II. The International Communist Party, convinced that photographs unavoidably reflect the class of the person holding the camera, mobilized workers in a variety of documentary projects and publications. In direct contradiction to a claim of impartiality, this photography based its authority on the photographer's membership in the group. It can thus be seen as a precursor to anthropology's interest in indigenous photographers' images as expressions of their specific cultural perspectives.

The authenticity of the documentary style was based on a complex set of factors. Seldom was the work based on a single photograph, but instead was built up through a series of images witnessing the photographer's proximity to the subject over an extended time. Despite many exceptions, the images also tend to follow a specific visual code. Straightforward photographs of people, taken in their everyday environment under available light conditions, are the most typical. If the subject appears to be aware of the camera, he or she is usually gazing directly into the lens. The photograph's claim to truth is based therefore on the

seemingly contradictory appeals of the photographer's simultaneous engagement with and objective distance from the subject.

4. Photography and Mass Society

Closely related to the ways photography has portrayed its subjects is the effect of images on those who see them, the audience. Photography was a mass medium almost from its inception, as the idea of the photograph and the ways it was marketed had a broad and immediate social and cultural impact. The 'very insatiability of the photographing eye' and the 'omnipresence of photographed images' position the medium in a vortex of complex moral, ethical, political, and esthetic issues. Sontag's influential book of essays *On Photography* (1973) brought many of these issues up for debate. Yet the exact nature of the photograph's appeal and its effect on its viewers has remained elusive. Despite extensive research on photography's impact on attitudes and behavior, results have been either inconclusive or too context-specific (regarding the viewer's age and background or the viewing context) to support any general conclusions. What is evident, instead, is a widespread belief in the photograph's influence which has in turn led to specific applications of the medium based on conceptions of the nature of photographic communication as a highly unstable synthesis of factual information and seductive attraction.

From the 1850s photographs began to be used alongside other illustrations in some magazines, and the widespread application of the half-tone process in the 1890s coincided with the emergence of new audiences in Europe and the United States. Photography was thought to appeal in particular to people who were either nonliterate or had not yet mastered the language of their new country. A new inexpensive illustrated press arose that used photography in often sensational ways, and photographs began to appear in advertising. Political campaigns made extensive use of photography from the 1920s onward, in documentary forms described above, but also in photomontage and posters. The reticence of many newspapers to make extensive use of photography (for example, the morning newspapers 'of record' in New York and London) is an exception which nevertheless rests on the same notion of the photograph's communicative power. The photograph, if seen as an unmediated copy of nature, does not require language or reason to be understood, but appeals instead directly to the emotions. Newspapers and other media that wished to base their appeal on rationality excluded photographs from their pages. Even if an image were manipulated through technical or other means, the message that the audience saw or 'received' was accepted as representing reality.

The model of mass communication that emerged first in the United States and dominated research for

several decades was based on the notion of communication as a successful 'transmission' of a message from an active and powerful source to a passive receiver. The origins of the model and the research it generated can also be traced to growing concern over the power of the media, particularly in the service of political propaganda and the rise of fascism in Europe after World War I. Although only sporadically applied in research on photography, the transmission model of communication served as the underlying framework of explanation of photography's influence on its audience. The power of images is assumed to be greater, for example on the relatively unformed psyches of children and youth. 'Moral panics' continue to pursue the image and its influences, giving rise to research on Hollywood films in the 1930s, on comic books in the 1940s, on television's effects on children in the 1950s and 1960s, and video games at the turn of the century.

Countering the transmission model of photography's influence on its audience is the perspective that considers photography as a cultural practice and phenomenon which participates in the construction of power and meaning through the mutual and intertextual establishment of social and political institutions and forms of visual expression. The cultural perspective views the audience not as passive receivers, but as members of a (sub)culture who employ cultural knowledge as they produce, consume, and interpret photographs. The cultural perspective does not deny the power of the media, but rather sets it in a historical, social, or political context, and examines how the media use photographs to address or position the viewer in specific relationships to media content. The perspective shifts the emphasis from an inquiry into photography's impact on individual or group attitudes and behavior to an examination of the ways photography is used in the construction of meaning, often focusing on social values and the use of stereotypes. The medium has been decisive, for example, in the production of visual representations of gender and gender relations, as revealed in Erving Goffman's study of advertising (1979). Other studies have demonstrated how media images of women not only establish standards of feminine beauty, but also link the image of the female body to products and services in a process of commodification that objectifies all women (cf. Williamson 1978, Irigaray 1985, Bordo 1993, Solomon-Godeau 1991). Other analyses have provided insights into the formation and perpetuation of stereotypes of race, in advertising but also in the news media, with groundbreaking work carried out at the Centre for Contemporary Cultural Studies in Britain in the 1970s.

Issues of media power and influence are also linked to photography through a growing concern over the image's role in surveillance. Cameras, in addition to their continually expanding position in private life, are used increasingly in both public and commercial

spaces for the ostensible purpose of increasing security and preventing crime. Yet the images taken by these cameras also raise issues of the photograph's potential for restricting, documenting, and controlling individual behavior. As technology extends the capacity to send images over global networks, the specter is raised of expanding spheres of media influence and control that transcend the nation state. At the same time, the interactive aspects of many contemporary media raise the possibility of individual and group autonomy from transnational media content, and the ability of the individual to form and reform images that are sent and received. The cultural impact of these relatively new image technologies is a current topic of debate, particularly with regard to the moral and ethical implications of their distribution and use.

5. *Technology and Revisions of Photographic Meaning*

A central question throughout photography's history is whether the meaning of the photograph is a function of the medium *per se* and its technical, chemical, and optical relationship to the subject in front of the camera, or is formed instead through the culture which surrounds it. Benjamin described how photographic reproduction and mass distribution irrevocably altered the individual's encounter with the original, following a more general argument developed within the Frankfurt School over the hegemonic power of cultural production in mass society. Yet it was not until the 1970s that a number of theorists began to address the specific problem of photographic meaning, particularly issues related to what and how the photograph represents. Roland Barthes (1977) used a semiotic analysis to unlock the problem of 'the photographic paradox': that is, the photograph represents an analogic reproduction of reality or a message without a code, and simultaneously a connotation or coded message, representing what society thinks about the subject. The major task of the denotative aspect of the image, according to Barthes, is to naturalize the cultural, that is, to transform cultural values into a message that can be 'read' as self-evident, as nature.

The deconstruction of the camera as a cultural artifact and of photography as deeply informed by culture continued on several parallel fronts. The model of the observer that dominates Western science, defined by the Cartesian perspectivalism and the direct gaze, is encoded in the optical system of the camera. The model thus defines a particular relationship between seeing and knowledge that privileges the authority of this gaze and its representations. This 'way of seeing' was critiqued most forcefully in relation to documentary photography for its dehistoricized view of marginalized social groups, a view that in turn established a visual code for representing the Other

(Bolton 1989). Because photography has been constructed as a predominantly male practice, reinforced by the observer's gaze as male, female also becomes Other. Extensions of this critique to the uses of photography in the social sciences have implicated anthropology in particular for constructing its subjects as visually exotic. Related to this phenomenon, as mentioned above, is the institutional gaze, embodied and perpetuated through the archive, the museum, or the press (Tagg 1988).

In the 1970s, anthropologists also began to criticize the use of film and photography in fieldwork. Instead of being informed by anthropological concepts, the Western anthropologist's visual records reflected 'his' own culture, coded through the practices of family photography. Within the subfields of visual anthropology and sociology in the United States and Britain and their influential journals, this insight was variously developed into a broader perspective on photography, film, and media in general as sources for studying the cultural values and perspectives of those who create the images. This in turn supported a shift from image analysis to the study of the social and cultural practices in which photography is embedded. This complemented the critique being directed at the field of anthropology as a whole for perpetuating its colonial heritage through techniques of 'writing culture' that represented its subjects as Other (Ruby 1982). Anthropology's continuing interest in the photography of indigenous peoples and other subcultural groups reflects on the one hand a recognition of photography as a significant source for cultural analysis and on the other, attempts by the field to integrate the anthropological subject's own understandings into research and reports. Current ethnographic studies of media practices frequently include investigations of how images are interpreted and used by ordinary people. The photographic practices of professionals has been largely neglected, however, apart from the many biographies of well-known photographers and analyses of their work.

Digital techniques of making, reproducing, and transmitting photographs have contributed to undermining the authority of the photograph as analog or direct copy of reality. In the mid-1980s, the digitized image was seen by many, particularly within the press, as a direct challenge to established beliefs about how the photograph represents. Gradually the technique has been integrated into professional and private practices of making and distributing images, stimulating the rise of new media networks. The immediacy of image transfer and the portability of the technology have vastly increased the number of photographs in circulation, including those depicting war and human suffering. Some analysts argue that this phenomenon results in 'compassion fatigue,' reducing the moral impact of the image (Moeller 1999). In the meantime, new applications of interactive image technologies are emerging in the arts, in youth culture, and in diasporic

communities. The digital transfer of images globally has given rise in some quarters to a utopian vision with the photograph as a cornerstone of a global visual culture. More realistic analyses point to the diversities of culture worldwide and the continued vast inequalities in access to these technologies as critical impediments to a hegemonic visual culture.

See also: Mass Communication: Normative Frameworks; Mass Communication: Technology; Mass Media: Introduction and Schools of Thought; Mass Media, Representations in; Photojournalism; Scientific Instrumentation, History and Sociology of; Vision, Psychology of; Visual Anthropology; Visual Images in the Media; Visualization: Representation in Science

Bibliography

- Barthes R 1977 *Image Music Text*. Hill and Wang, New York
- Bateson G, Mead M 1942 *Balinese Character: A Photographic Analysis*. Special Publications of the New York Academy of Sciences 2, New York
- Benjamin W 1968 The work of art in the age of mechanical reproduction. In: Arendt H (ed.) *Illuminations*. Walter Benjamin. Schocken, New York
- Bolton R (ed.) 1989 *The Contest of Meaning: Critical Histories of Photography*. MIT Press, Cambridge, MA
- Bordo S 1993 *Unbearable Weight: Feminism, Western Culture, and the Body*. University of California Press, Berkeley, CA
- Chalfen R 1987 *Snapshot Versions of Life*. Bowling Green State University Popular Press, Bowling Green, OH
- Collier J Jr, Collier M 1986 *Visual Anthropology. Photography as a Research Method*. University of New Mexico Press, Albuquerque, NM
- Edwards E (ed.) 1992 *Anthropology and Photography, 1860–1920*. Yale University Press, New Haven, CT
- Edwards E 1997 Ordering others: photography, anthropologies and taxonomies. In: Roberts R (ed.) *In Visible Light*. Museum of Modern Art, Oxford, UK
- Goffman E 1979 *Gender Advertisements*. Harper & Row, New York
- Irigaray L 1985 *This Sex which is not One*. Cornell University Press, Ithaca, NY
- Moeller S 1999 *Compassion Fatigue. How the Media Sell Disease, Famine, War and Death*. Routledge, London
- Newhall B 1949/1982 *The History of Photography: from 1839 to the Present*. Museum of Modern Art, New York
- Rosenblum N 1984 *A World History of Photography*. Abbeville Press, New York
- Ruby J (ed.) 1982 *A Crack in the Mirror. Reflexive Perspectives in Anthropology*. University of Pennsylvania Press, Philadelphia, PA
- Solomon-Godeau S 1991 *Photography at the Dock: Essays on Photographic History, Institutions, and Practices*. University of Minnesota Press, Minneapolis, MN
- Sontag S 1973 *On Photography*. Farrar, Straus and Giroux, London
- Stasz C 1979 The Early History of Visual Sociology. In: Wagner J (ed.) *Images of Information*. Sage, Beverly Hills
- Tagg J 1988 *The Burden of Representation. Essays on Photographies and Histories*. University of Massachusetts Press, Amherst, MA

Williamson J 1978 *Decoding Advertisements: Ideology and Meaning in Advertising*. Marion Boyars, London

K. Becker

Photojournalism

Photojournalism emerged in two distinct phases: during the late nineteenth century with the use of photographic images in newspapers and magazines as a new form of covering the news, and during the 1920s when photojournalism matured with the aid of technological advancements and was recognized as a form of documentary expression in most parts of the world. Unlike the spread of printing—or television later—photojournalism as a professional practice appeared almost simultaneously throughout Europe, the United States, and elsewhere. Thus, by the mid-1920s photographic expansion in Japan included the use of Leica cameras and access to the visions of *Neue Sachlichkeit* and Bauhaus practices in Germany, while similar developments in China were also directed by Western influences, although (between 1966 and the 1980s) the Cultural Revolution slowed the spread of photojournalism.

The beginnings of photojournalism occurred in the middle of the nineteenth century: photographers documented the US American-Mexican war in the 1840s; James Robertson and Felice Beato (England) photographed uprisings in India and China in the 1850s and 1860s; while Roger Fenton (England) covered the Crimean war in 1855 and Francis Frith (England) visited the Nile in 1856. Mathew Brady (USA)—with the aid of Alexander Gardner and George N. Barnard, among others—produced Civil War images in the United States between 1862 and 1864, and Jimmy Hare (USA) began working as a staff photographer for the *Illustrated American* in 1889. He later traveled to Cuba and also photographed the Russo-Japanese War, the Mexican Revolution, the Balkan War, and World War I before his retirement in 1922. Frequently, travel photographers and those accompanying expeditions, also worked as photojournalists.

The rise of photojournalism parallels the technological, social, and economic developments of the times. Technological innovations included the perfection of printing techniques to reproduce photographic images (halftones): beginning with the *New York Daily Graphic* (USA), 1873, or *Le Journal Illustré* (France), 1886, although the US press did not use halftones regularly until 1897. Also important was the development of flexible, transparent film (George Eastman, 1889), and the invention of smaller cameras (Ermanox, Leica/Germany, 1925), faster films (35mm, Kodak/USA, 1923, reversible emulsion (AGFA/Germany, 1923), and the introduction of

electronic flash equipment (Vacu-Blitz/Germany, 1929; Sashalite/England, Photoflash Lamp/USA, 1930).

The growth of photojournalism was advanced by the expansion of the reading public—with the addition of the working class—accompanied by an increased availability of inexpensive newspapers and picture magazines. Also, the public desire to ‘see’ the world—fuelled by the popularity of the movies (and newsreels, in particular) especially during the 1920s—created new demands on print media to legitimize photojournalism as a professional practice. In addition, the role of photography in education and propaganda, particularly in the Soviet Union, and its use in authoritarian societies, like Hitler Germany, energized photojournalism and strengthened its commitment to the ideals of social or political revolutions while providing increased picture coverage of the respective societies. The rise of worker-photographers in the Soviet Union and Germany during the 1920s also extended the notion of photojournalism to include the contributions of the working class to a photographic surveillance of the social environment. However, in post-1949 China, the use of photographs by the press remained restricted until much later (e.g., the first national photojournalism contest was held in 1981).

Photojournalists of the modern era found increasing numbers of professional outlets with the rise of mass circulation picture magazines, among them: *Arbeiter-Illustrierte-Zeitung*, *Berliner Illustrierte Zeitung*, *Vu*, *Paris Match* (France), *La Vanguardia* (Spain), *Picture Post* (England), *Look*, *LIFE* (USA), *Ogonyok* and *USSR in Construction* (USSR), or *China Pictorial* (China). Although photo agencies had existed since the end of the nineteenth century (e.g., Bain News Service, 1895, Underwood & Underwood, 1901, International News Photos, 1909 all in the USA) and photographers had been employed in considerable numbers by the press since the beginning of the twentieth century, increasing demands for photographic coverage by a rising number of media clients since the 1920s resulted in the establishment of new picture agencies. Photographs became commodities, and photographic agencies with their own staffs and national or international visibility produced and disseminated single images and photo stories. Among these photo agencies were: Mauritius, Weltrundschau and Dephot (Germany), Keystone, Rapho (France), Soyuzfoto, Novosti Picture Agency (Soviet Union), Acme Newspictures, Black Star, Wide-World Photos (USA), Scala (Italy), Nippon-Kobo (Japan) or—after World War II—agencies like Magnum, Viva, Gama, and Sygma (France), Contact Press Images (USA), and—since 1949 in China—the Bureau of Photojournalism, a government agency, reorganized as the Department of Photojournalism in the New China News Agency in 1952.

Together with traditional news agencies like the Associated Press, USA, or Reuters, England, these

agencies became wholesale suppliers of visual material for the world press, working as commercial sources or cooperatives while setting the professional standards of modern-day picture coverage. They were aided by the development of wire transmission (invented in 1905 by Alfred Korn in Munich, Germany) in the early 1920s (London-New York over the Bartlane Cable system). The Associated Press New Photo Service began in 1927 and delivered photographs by train and plane; its Wirephoto network (1935) and the United Press Telephoto service (1936) helped increase the traffic in photographs, but made local editorial staffs less dependent on local photo coverage. At the end of World War II, the (New York) Times Photo service offered a worldwide news photographs service. The first satellite transmission of photographs in the United States occurred in 1960, followed by the launching in 1962 of the Telstar satellite for live transmission of pictures between Europe and the United States. Laserphoto was introduced by the Associated Press (USA) in 1977 and replaced by Photostream in 1991, an innovative transmission system that uses digital technology. The Associated Press (USA) had patented a portable digital image scanner in 1983 for the transmission of photographs from remote locations. Digital cameras appeared in 1988, and digital darkrooms in 1989 (*San Francisco Examiner*). The Associated Press (USA) began distributing news photographs via the World Wide Web in 1995.

Photo agencies also became depositories of visual representations of public life during most of the twentieth century (the Hulton Deutsch Collection (England) is one of the largest picture libraries in the world with over 15 million images; the Associated Press (USA) maintains an active picture archive with over 500,000 images, accessible over the Internet). The combined organizational network of photo agencies worldwide and their archival holdings constitutes a formidable, if not dominating cultural and commercial force in the process of visual representation. To a significant extent, journalism everywhere relies on the images that photo agencies supply.

By the end of the 1920s photojournalism was firmly established worldwide in the metropolitan press with the hiring of picture editors and the use of staff photographers in the editorial process. Smaller newspapers were considerably slower in creating staff positions for photographers. Over the years, the work of photojournalists has been preserved in the public and private archives of the press around the world. Among the most prolific photojournalists of modern times are Erich Salomon, Alfred Eisenstaedt, and Wolfgang Weber (Germany), Max Alpert, Arkady Shaikhet (Soviet Union), Eugene Smith, Margaret Bourke-White, Weegee, David Douglas Duncan, Mary Ellen Mark, Susan Meiselas (USA), Werner Bischof, René Burri (Switzerland), Robert Capa (Hungary), Ihei Kimura (Japan), Wu Yinbo, Zhang

Suicheng (China), Shomei Tomatsu (Japan), Peter Magubane (South Africa), Henri Cartier-Bresson, Martine Franck, Gilles Peress (France), Sebastião Salgado (Brazil), Bert Hardy, Don McCullin (England).

Photojournalism succeeded in making the world a knowable, visual experience and in creating a sense of familiarity with people and places. The presence of photojournalism in the public sphere has a profound effect on the public staging of politics. Photojournalism, by its very nature, generates attention, and because of its impact, raises issues of privacy, sensationalism, and manipulation in a marketplace that thrives on the exploitation of the image. Unlike others, photojournalists—to be effective—must seek the proximity of unfolding events, including the hazards of social unrest, revolution, or war. Consequently, they often risk their lives and sometimes pay the ultimate price in their pursuit of the perfect picture (by 1971, for instance, 25 of 35 non-Vietnamese journalists killed during the Vietnam war were photographers or cameramen).

The practice of photojournalism, beginning in the 1920s, led to an exploration of the documentary style and produced single and multiple images to convey the news. Photographs became a regular staple of newspapers and affected their make-up and design. When photographers thought editorially, the result was a new vocabulary of journalism. But, as Evans (1978/1997, p. 255) suggests, 'photography is not photojournalism'; instead, words are needed to supply more information than are conveyed by a photograph. Words help explain, confirm, and reinforce, they also place events in time and space. According to Hicks (1952, p. 5), 'the basic unit of photojournalism is one picture with words.' Photojournalism acknowledges the importance of words, either provided in the form of outlines or stories authored by the photographer or journalist-colleagues.

The picture story (or photo essay) constitutes the long form of photojournalism; its visual narrative typically focuses on a single subject and is the outcome of careful planning and extended professional commitment to a project. Ideally, the resulting presentation—not unlike an essay—consists of a flow of images, characterized by a strong opening, a descriptive middle passage, and a decisive ending. Pioneered first in the Soviet Union and Germany during the 1920s, the photo essay became a determining characteristic of *LIFE* magazine in the United States; it challenged the imagination of photographers and editors to produce a durable visual narrative that could rely on the popularity of story telling to satisfy reader interests. There is also an affinity between photojournalism and photographic fiction; for instance, photo-novelas as social reportage constitute a form of visual narrative, which has been explored since the 1970s (as literary photography or photographic writing) by contemporary photographers.

With the advent of television in the 1950s and the demise of the mass circulation picture magazine (particularly in the United States) during the 1960s, photojournalism survives in news magazines and the daily press, in particular, where it continues to help strengthen the visual culture of the late twentieth century (aided by the ease of offset printing which improved the reproduction quality of images). The introduction (since the 1930s) of color positive and negative films (with distinctive processes developed by Agfa, Cibachrome, Kodak, and Polaroid) and their increasing use by the press (since the 1960s), added a new dimension to photojournalism; it enhanced the documentary quality of images and contributed to their visual impact (*LIFE* magazine published the first color photographs in 1952). Although black-and-white photographs maintain their abstract, documentary quality (and are used for this purpose in modern advertising), the work of contemporary photojournalism relies almost exclusively on the production and circulation of color photographs.

A distinctively different form of photojournalism emerged with the demise of picture magazines in the United States and the rising popularity of the book as an alternative medium for the dissemination of visual expressions. The book celebrates authorial prerogative and individuality of form and style; it also offers opportunities for exercising creative freedom and extended personal constructions of particular aestheticized visions of the world that favor a humanistic approach to photography and avoid claims of objectivity. The resulting emergence of a new kind of photojournalism at the end of the twentieth century borrows freely from the tradition of street photography and photographers ranging from Robert Frank, William Klein, and Garry Winogrand (USA), to Henri Cartier-Bresson and Robert Doisneau (France), or Josef Koudelka (Czech Republic). The sustained, book-length visual argument can be powerful, but when photojournalism catered more exclusively to a privileged class of bourgeois readers, experts, and connoisseurs of fine photography, it crossed into the realm of art: single news photographs appeared on museum walls and became commodified objects in the commercialized sphere of art collection (especially since the mid-1970s when photographs began to be featured regularly in art galleries).

In an image conscious world photojournalism survives and its tradition is preserved and celebrated by institutions such as the International Center of Photography (USA), or the World Press Photo Foundation, organizations such as the National News Photographers Association (USA) and educational programs, such as the University of Missouri School of Journalism (USA), all of which stand for similar organizations and activities in a number of countries.

See also: Journalism; Journalism and Journalists; News: General; Photography as a Medium

Bibliography

- Alabiso V 1998 *Flash! The Associated Press Covers the World*. Abrams, New York
- Barret A 1978 *Die ersten Photoreporter, 1848–1914*. Krüger, Frankfurt, Germany
- Carlebach M L 1997 *American Photojournalism Comes of Age*. Smithsonian Institution Press, Washington, DC
- Chapnick H 1994 *Truth Needs No Ally. Inside Photojournalism*. University of Missouri Press, Columbia, MO
- Evans H, Taylor E 1978/1997 *Pictures on a Page. Photojournalism, Graphics, and Picture Editing*. Pimlico, London
- Gidal T 1973 *Modern Photojournalism: Origin and Evolution, 1910–1933*. Macmillan, New York
- Harrison M 1998 *Young Meteors. British Photojournalism, 1957–1965*. J Cape, London
- Hicks W 1952 *Words and Pictures*. Harper and Row, New York
- Hopkinson A (ed.) 1995 *150 Years of Photojournalism*. Könnemann, Cologne, Germany, Vol. 2
- Lewis G 1991 *Photojournalism. Content and Technique*. Brown, Dubuque, IA
- Miller R 1998 *Magnum: Fifty Years at the Front Line of History: The Story of the Legendary Photo Agency*. Grove/Atlantic, New York
- Neubauer H 1998 *Black Star. 60 Years of Photo Journalism*. Könnemann, Cologne, Germany
- Tausk P 1988 *A Short History of Press Photography*. International Organization of Journalists, Prague, Czech Republic
- Yapp N (ed.) 1995 *150 Years of Photo Journalism*. Könnemann, Cologne, Germany, Vol. 1

H. Hardt

Phylogeny and Systematics

Systematics is the study of the relationships among organisms, whereas phylogenetics is the study of the evolutionary relationships among organisms. In that sense, phylogenetics is a subset of systematics. Historically, before evolution was accepted as the unifying principle in biology, phylogenetics did not exist *per se*. Instead, organisms were classified on the basis of their phenotypic similarity.

1. Defining the Questions

One of the main problems with using phenotypic similarity as a basis for classification is that organisms evolve to become adapted to the environments in which they live (their niches) so that phenotypic convergence is a fairly common occurrence. For the most part, the longer ago in time that two organisms shared their most recent common ancestor, the more phenotypically dissimilar they will have become. However, phenotypic convergence as a consequence of adaptive evolution, along with rapid divergence as organisms adapt to totally novel environments, means that any classification based on phenotypes inevitably confounds phylogenetic relationships and adaptive

evolution. Given the evolutionary lability of behavioral characteristics, analysis of the adaptive or functional significance of particular traits requires that the phylogenetic relationship is properly defined so that evolutionary changes in behavior can be mapped on to those in niche use.

The advent of cladism as a technique aimed at identifying phylogenetic groupings through phenotypic analysis was a notable first attempt to isolate phylogenetic groupings. Cladism, as initiated by Hennig (1966), is, however, based on a model of evolutionary change which excludes the possibility of phenotypic convergence because it assumes that the same character cannot evolve on more than one occasion and that once a character has evolved it cannot revert to its former state. More recently, the field of phylogenetic analysis has been revolutionized by the rapid production of gene sequence data. Realistic models of molecular evolution are incorporated into phylogenetic analyses of various regions of the genome to produce phylogenies (evolutionary trees) of ever-increasing accuracy. How those trees will be used to produce classifications is not clear.

At the moment, recognized groupings (genus, tribe, family, and so on) are being revised so that they describe monophyletic groups (containing all the contemporary descendants of a particular common ancestor). However, the nominal groupings involved differ according to historical precedence: systematists who originally produced the hierarchical classifications in use today ranged along an axis from lumpers to splitters. The former were less keen to classify species with a particular degree of morphological similarity into different higher level taxa than were the latter. As a consequence, even when species within the hierarchical groupings are rearranged to represent monophyletic groupings, the most recent common ancestor for all the species in a particular genus would have been earlier in some orders than the most recent common ancestor for two species separated at the family level in other orders.

The development of this article introduces natural groupings, explains how phylogenies are determined, and then shows how phylogenies can be used to understand the functional significance of behavioral differences among taxa.

2. Natural Groupings

It has always been clear to modern humans that there are natural groups of organisms: trees, birds, flowering plants, spiders, and tortoises would be cases in point. It has also been clear, at least in more recent times, that there is a natural hierarchy of groups: there are animals versus plants, but within animals there are vertebrates and invertebrates, and within vertebrates there are birds, reptiles, mammals, and amphibians. The embodiment of the hierarchical natural classification was produced by Linnaeus around 1800. No particular

mechanism was envisaged for why this natural hierarchical classification should exist, but one frequently stated reason for studying it was to provide insight into the mind of the creator. With the benefit of hindsight, insofar as classifications were based on functional similarity, what were recognized to be the same character could in fact have separate evolutionary origins.

The major sea change occurred in 1859 with the publication of Charles Darwin's *Origin of Species* in which he argued that life was created on one or few occasions and that subsequent speciation resulted in the natural hierarchy. Indeed, it is worth remembering that the only figure in the book is of an evolutionary tree. From that time on, with a few notable lapses, it has been accepted generally by working scientists that natural groupings are a consequence of evolution and that they should be monophyletic. In practice, however, it did not always turn out that way. For example, it is accepted that birds share a more recent common ancestor with crocodiles than either does with lizards. As a consequence, crocodiles should be classified with birds rather than with lizards. This means that the reptiles do not constitute a monophyletic group, but what is known as a paraphyletic group: a group made up of only those descendants of a common ancestor that resemble it (or each other) phenotypically.

Despite the acknowledged fact that the reptiles constitute a paraphyletic group which, if it were to be made monophyletic, should include the birds, there has been little serious effort on the part of biologists to execute this particular taxonomic revision. To do so, would in some sense apparently cost more effort than it is worth. Instead, reptiles without the birds are accepted as a paraphyletic natural group, while palaeontological and other evolutionary analyses should include birds within the group.

3. Identifying Phylogenetic Relationships

There have been many schools of systematics, and most are of merely historical rather than contemporary importance. The major division is into methods that analyze phenetic relationships vs. those that attempt to reconstruct phylogenetic relationships.

The development of phenetic methods became particularly popular in the 1960s and 1970s following the publication of Sokal and Sneath's (1963) book entitled *Principles of Numerical Taxonomy*. The book outlined apparently objective methods of character scoring and the production of phenograms (trees that summarized degrees of phenetic similarity), using the then newly and widely available computers. The fatal flaw to the whole enterprise was the absence of a model of phenetic evolution. There is an infinity of so-called clustering algorithms and distance methods, each of which is appropriate for a particular model of phenetic evolution. Unfortunately, there was no objective way

to decide which to use. It might be argued that the aim of numerical taxonomy or phenetics is not to determine evolutionary relationships but rather to describe phenetic similarity, but even then there is an infinite number of distance methods and clustering algorithms, with no principled method of deciding on appropriate ones for all circumstances. The exercise foundered.

The use of phylogenetic methods was given a major boost by Hennig's (1966) book *Phylogenetic Systematics*. Hennig realised that if the phylogeneticist was willing to make certain assumptions about character evolution, then phylogenetic trees could be constructed from data based on the distribution of character states among extant taxa. If every character evolved only once and evolutionary change was irreversible, then monophyletic groups could be defined on the basis of shared derived characters. Problems arose when different characters produced different answers so that one character placed a particular species in one group while another character placed it in a different group, which meant that at least one character did not evolve according to Hennigian principles. Common sense often prevailed among cladists to distinguish between, for example, characters that might have evolved once and been lost versus those that evolved more than once. However, once cladists began to entertain the possibility that there are different models of evolution for different characters, then the advent of a model-based statistical approach to phylogenetic reconstruction became inevitable. The problem recurs, just as it did with numerical taxonomy, that for phenetic evolution there is frequently no principled way to distinguish among alternative models of evolution.

In the years immediately before genes were sequenced at such low cost and with such speed, an informative stepping stone between phenetic and phylogenetic analysis occurred with the advent of DNA-DNA hybridization technology. The method was developed and widely applied by C.G. Sibley and co-workers. Radioactively labeled DNA strands about 500 nucleotides long from a focal species were hybridized with similar strands from a number of phylogenetic relatives. The rate of dissociation of these duplexes provided a measure of genetic similarity between the sequences, thereby excluding purely morphological convergence as a factor producing errors in phylogenetic tree reconstruction. Several surprises emerged. For example, among the birds, Australian passerines seemed to be more closely related to each other than they are to passerines from other continents, thus suggesting that previous classifications based on morphological comparisons had failed to detect considerable convergent morphological evolution. Sibley and Ahlquist obtained representative samples of many thousands of bird species, representing the large majority of tribes. They then subjected the data to DNA-DNA hybridization analysis

to produce an estimated phylogeny of the birds based simply on the rate of dissociation of DNA–DNA hybrid duplexes. That phylogeny was used by Sibley and Monroe (1990) to produce a classification of birds: categories of rates of duplex dissociation were arbitrarily defined that were associated with each particular taxonomic status. Despite the various analytical difficulties of their method, and the arbitrary nature of their classification, their procedure was claimed to be both objective and repeatable. As a first attempt at a full phylogenetic classification for a major taxon, it has probably yet to be improved upon. However, while most of the major revised phylogenetic relationships that Sibley and co-workers revealed have remained intact, more appropriate methods for phylogenetic reconstruction from gene sequence data are now available.

DNA–DNA hybridization methods provide just one measure of the distance separating a pair of sampled individuals. That distance depends on the genetic similarity between those individuals in terms of base sequences. However, the same distance measure can be achieved in many different ways: if there are 20 base differences between individuals 1 and 2, and there are 20 base differences between individuals 2 and 3, that does not tell us how many differences there are between 1 and 3—just that there are between 0 and 40. All pairwise DNA–DNA hybridizations would need to be completed, even if we were accepting that DNA diverged at a metronomic rate through evolutionary time.

The production of vast amounts of gene sequence data means that every base pair becomes a character. Once we have a base sequence from a particular gene of a particular individual, it can be compared with that for every other individual for which we have equivalent data. This allows phylogenetic reconstruction using discrete rather than distance measures. What is more, we can describe models of molecular evolution with reasonable accuracy: we may know the relative substitution rates of the first, second and third base pairs, and we often have a good idea of transition to transversion ratios (the frequency with which a substitution is between purine and purine or between pyrimidines, vs. between purine and pyrimidine). We also know that some base substitutions are effectively neutral because they result in the same amino acid being coded for at the base-substituted triplet.

Before mentioning the most commonly used methods for discrete data, it should be pointed out that distance measures can be used and are used with gene sequence data. The methods differ in whether they assume branch lengths from the root to the tips of the tree are all equal and adjust the data accordingly (e.g., unweighted group pair method with arithmetic means: UPGMA) or whether they allow different rates of evolution in different parts of the tree (e.g., the neighbor joining method which often gives a good approximation to the minimum evolution tree). De-

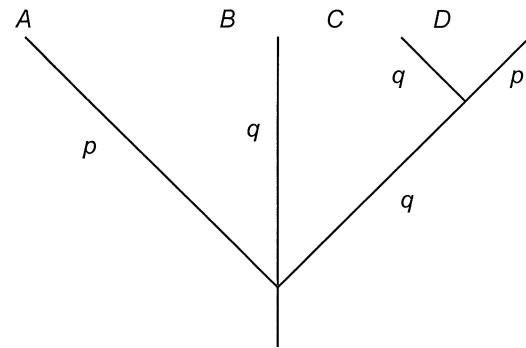


Figure 1

A phylogenetic tree that is likely to produce data that will cause parsimony or compatibility methods to give misleading results. The correct tree is shown, with p and q being probabilities of character state change along the different branches of the tree. The wrong tree is likely to be determined when p is much larger than q because then B and C are likely to be more similar than are C and D (after Felsenstein 1978, 1983)

spite the fact that information is lost (the evolution of individual base sites cannot be traced down the tree), there is also the problem that approximations are made which often result in biologically impossible branch lengths being reconstructed (see Page and Holmes 1998).

The two most common model-based discrete methods of model based phylogenetic tree reconstruction from gene sequence data involve parsimony and maximum likelihood.

Parsimony was introduced as the so-called minimum evolution distance method by Cavalli-Sforza and Edwards (1967) as a computationally rapid route for approximating maximum likelihood trees. However, as a discrete method for finding the shortest tree, it is possible to adapt strict parsimony and weight the changes in appropriate ways (e.g., a transition might be four times as frequent as a transversion) and determine that tree with the smallest total of weighted steps. There are several problems with parsimony as a procedure. The first is philosophical—parsimony is a minimum criterion by definition and it may be argued that there is no minimum principle to evolution. The second major problem is that of 'long branch attraction' (see Fig. 1). Despite these disadvantages, parsimony remains a commonly-used procedure because of its computational speed.

Maximum likelihood is a very powerful tool in phylogeny reconstruction. The method asks which phylogenetic tree is the most likely one to have given rise to the set of gene sequences that we observe in our sample from the present day, given our specified model of gene sequence evolution. Indeed, it can go further and be used to estimate the most likely parameter estimates of our evolutionary model. The likelihood

approach is increasingly commonly used for the reconstruction of molecular phylogenies, and as computers get ever faster and more powerful, it will probably become the approach of first choice.

4. Using Phylogenetic Relationships to Analyze Behavior

When behavior is mapped on to phylogenetic tree structure, at least four types of informative analysis become possible. First, behavior itself may be seen to be responsible for phylogenetic tree structure. For example, species with particular modes of behavior may be more likely to divide into new species than others. Second, by mapping behavioral characteristics on to phylogenies, it is possible to determine which traits are ancestral and which are derived, and to determine which traits have changed most frequently over evolutionary time. Third, behavioral characteristics do not evolve independently of each other or of habitat use. If we are to understand why character covariation occurs, then phylogenetic tree structure must usually be an integral part of the necessary analysis. Fourth, adaptive radiations, such as Darwin's Galapagos finches, are well known but when we map behavior on to tree structure, it sometimes becomes apparent that independent adaptive radiations have remarkably similar structures—when the tape of life is replayed, similar arrays of diversity result. In this section, examples of each of these four types of phylogenetically-based behavioral analysis are described.

4.1 Behavior and Speciation

When behavioral traits are mapped on to phylogenies, it is sometimes possible to identify so-called 'key innovations' which appear to have resulted in high rates of speciation. For example, cichlid fishes are very speciose and their radiations seem to have coincided with a change in position of a particular muscle attachment which allowed the pharyngeal bones to manipulate individual prey items while still holding them. This released other areas of the jaw, which had been used for manipulating prey, to evolve for other functions. The cichlid radiations in African lakes are indeed characterized by a remarkable variety of feeding mechanisms. However, a problem with this explanation is that there may be other unrecognized characteristics unique to cichlid fishes that may also contribute to their explosive radiations.

A more convincing demonstration of key innovations resulting in increased rates of speciation is when the same innovation appears independently in different parts of the phylogenetic tree, and is statistically associated across a number of cases with an increased rate of speciation. One of the best examples of a repeatedly evolved such innovation is the ability

of insects to feed on vascular plants, known as phytophagy. There is a wide variety of vascular plants, and whenever phytophagy evolved, there are likely to have been many vacant niches to speciate into. Mitter et al. (1988) identified 13 nodes in a phylogenetic tree of insects where all the descendant species from one branch are phytophagous, whereas all those down the other branch are not. In 11 of 13 cases there were more extant species descended from the phytophagous branch than the nonphytophagous branch, and in each of those 11 cases there were more than twice as many phytophagous than nonphytophagous species.

A final example of behavioral changes apparently causing differences in rates of speciation comes from an original suggestion made by Darwin (1871). It is known that the females of some bird species actively choose their mate on the basis of phenotype. Darwin suggested that, when populations separate, the precise criteria for mate choice might alter so that if those separated populations came back together, mate choice would have promoted speciation. The most obvious criterion for determining whether females choose their mates by phenotype is whether males differ from females in physical appearance. When passerine birds are scored as sexually dichromatic or not, then it turns out that the repeated evolution of sexual dichromatism tends to be associated with subsequently higher rates of speciation (Barracough et al. 1995). It had taken more than 100 years before the phylogeny became available so that Darwin's theory could be tested.

4.2 Character Evolution

Mapping character evolution on to phylogenetic trees can be used to test ideas about how and why characters evolve. In order to carry out this process, it is necessary to have a phylogenetic tree showing the relationships among contemporary species, character states for each species, and a model for character evolution. Since the appropriate model for the evolution of most characters is poorly understood, ancestral character states usually are inferred from either a parsimony or other appropriate minimum evolution criterion.

Gittleman (1981) set out to test the idea that when patterns of parental care evolved among fishes, transitions were gradual and did not involve change by more than one parent. Twenty-one transitions were identified using parsimony on the phylogenetic tree, all of which were between biparental and maternal care, biparental and paternal care, no care and paternal care, and no care and maternal care. No cases involved the double transitions between no care and biparental care or paternal care and maternal care, a statistically significant deficiency if all transitions were equally likely.

A more thorough and quantitative analysis of historical patterns in parental care, this time among shorebirds, was reported by Székely and Reynolds

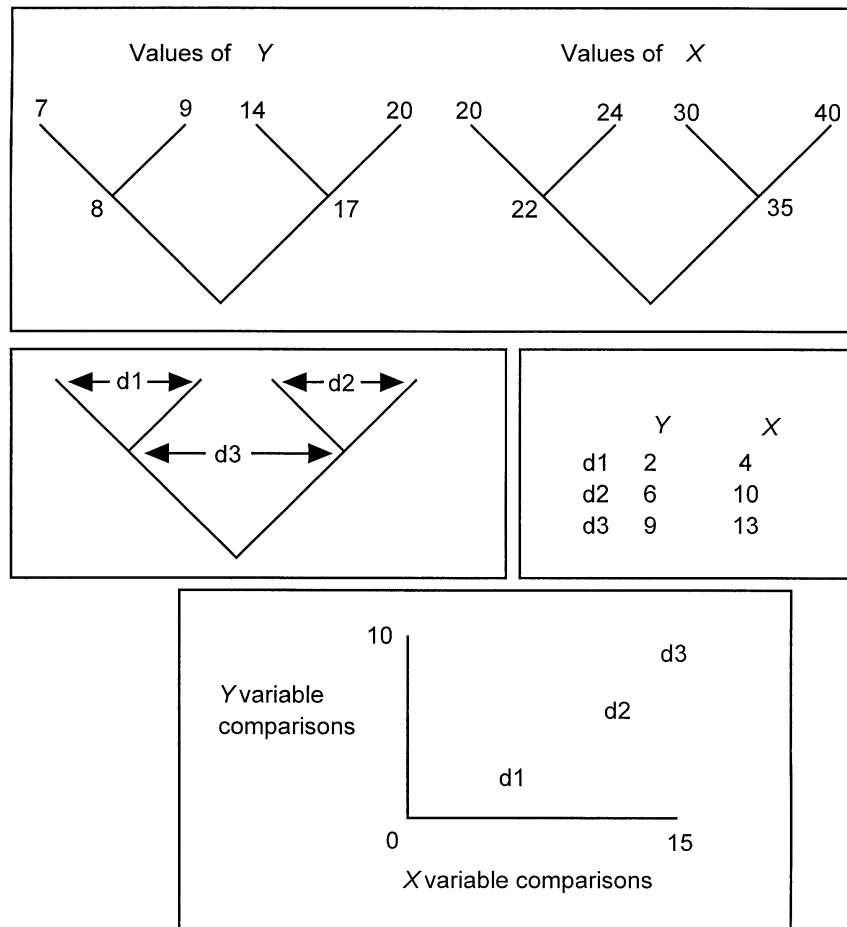


Figure 2

Comparative method of independent contrasts. Comparisons d1, d2, and d3 are independent of each other under a Brownian motion model of evolution for characters X and Y. If the two characters have shown uncorrelated evolution, values of the contrasts for one character will not be correlated with those for the other character

(1995). Again they used parsimony for ancestral character state reconstruction, but in this taxon they found that there had been an overall evolutionary reduction in paternal care, which was sometimes but not always compensated for by an increase in maternal care. The authors argued that this is a response to selection for males to partake in extra-pair matings or even to maintain multiple pair bonds.

How rapidly does behavior evolve in comparison with, say morphology and life history? Haldane (1949) suggested that one appropriate measure of evolution change was the 'darwin': a change of one logarithmic unit per million years. Gittleman et al. (1996) mapped the evolutionary change of two behaviors (group size, home range size), two life history characters (gestation length, birth weight), and two morphological characters (adult brain weight, adult body weight) using data for extant mammal species from eight taxa for which

there were reasonable phylogenies. The behaviors were more labile than the morphological characters, which were more labile than the life history characters. Such analyses may well identify characters which, for whatever reason, are the most evolutionarily constrained, leaving others more free to change.

4.3 Correlated Character Evolution

Bringing order to organic diversity is made simpler by two organizing principles: close relatives are more similar to each other than are distant relatives, and organisms share adaptations to common environments. As we have seen when attempting to reconstruct phylogenetic trees, it can be important to distinguish between these two reasons for similarity. However, the two principles cannot necessarily be

separated: because close relatives are often adapted to very similar ecological niches, they also share many adaptations in common, whether those adaptations were inherited from a common ancestor or not. If we wish to understand the extent to which similar adaptations are likely to arise in similar circumstances, it is nevertheless important to factor out identity by common descent. What is more, because close relatives share characters which may not be functionally related to each other (e.g., mammals have hair and give birth to live young, whereas birds have feathers and lay eggs), treating species values as independent points in a statistical analysis can result in functionally meaningless significant associations between characters as a consequence of pseudo-replication.

4.3.1 Independent contrast analysis. In order to tackle this problem, Felsenstein (1985) devised a method which traced the change in two or more characters on a phylogenetic tree, and defined for each character equivalent evolutionary independent contrasts which could be used for data analysis. Given character states for contemporary species, ancestral character states had to be inferred from a specified model of evolutionary change. Felsenstein used a Brownian motion model of change to estimate ancestral character states. In essence, all change that occurred between sister taxa evolved independently of change between other sister taxa. The consequence is that differences between the actual or estimated character states for sister taxa are used as the independent contrasts, which are standardized by branch lengths (variances of contrasts are larger when the time separating taxa has been longer). The direction of comparison is arbitrary, though it is conventional to order the taxa so that all contrasts for the independent variable are positive (or zero). The contrasts for the dependent variable(s) then use the same order of taxa and can take both positive and negative values. The contrasts for the dependent variable(s) are then regressed on those for the independent variables, and conventional statistical techniques (correlation or regression) are used to test for correlated evolutionary change. The null hypothesis is that change is uncorrelated, and significant departures from the null hypothesis can indicate either positive or negatively correlated character change. A hypothetical example is worked through in Fig. 2.

Although independent contrast analysis is widely used, statistical conclusions can, in fact, change when different models of character change are entertained. For example, Harvey and Purvis (1991) report a study of the evolutionary relationship between a measure of running activity and relative foreleg length (foreleg length corrected for body size) in a group of *Anolis* lizards. Three models of evolutionary change were considered (Brownian motion, punctuation, and minimum evolution), only the first of which produced a

significant relationship between contrasts for the two variables being studied ($p < 0.025$ compared with $p > 0.10$ for the other two models). Fortunately, diagnostics are available to test whether the assumptions of Brownian motion character evolution have been approximately met, given data for a set of contemporary species and a well-resolved phylogeny (Garland et al. 1992). However, when the hypothesis of Brownian motion character evolution is rejected, there remains a paucity of statistically powerful comparative methods.

4.3.2 Sister taxon comparisons. When a phylogeny is available, but there is no obvious model of character evolution, then the technique of sister taxon comparison can prove very useful. The idea is to compare pairs of contemporary species, chosen such that each 'sister pair' shares a more recent common ancestor than either member does with any member of any other pair. The result is that the difference between each pair of sister species has evolved independently of the difference between any other pair, and ancestral character states do not need to be estimated. The loss of assumptions results in fewer possible comparisons: if there are n contemporary species in a sample, so long as the phylogeny is bifurcating, then there are $n - 1$ possible independent comparisons, but at most only $n/2$ sister comparisons.

Møller and Birkhead (1992) have championed the use of sister taxon comparisons, and illustrated their use by testing the idea that high copulation rates among birds are a 'paternity insurance' mechanism. First, they asked whether rates of extra-pair copulation increased with local population density, which they did in all eight possible comparisons, which is a highly significant result (n.b. 15 potential comparisons are possible with 16 taxa when ancestral states are considered). Second, they asked whether birds living in colonies had higher rates of intra-pair copulations, which they did in 12 of 13 comparisons.

4.4 Parallel Structure of Independent Radiations

When a species invades a completely new habitat with unoccupied niches, there is the potential for an adaptive radiation to ensue. But how deterministic is evolution? Stephen Jay Gould suggested that contingency was such an important factor that, if we replayed the tape of life, we should invariably get different outcomes. That is, of course, true *sensu stricto*. However, early comparisons such as that between the placental and marsupial mammal radiations, and the very fact of evolutionary convergence, means that it is not the whole story. Recently, a number of case studies suggest a remarkable degree of similarity when the opportunity for similar radiations has been presented.

Anolis lizards in the Caribbean have invaded many small and a few large islands. On the very smallest islands, there tends to be one intermediate-sized species, but as island size increases large and small forms can coexist. Each of the largest islands supports an independently evolved *Anolis* radiation, with about seven different species occupying a variety of well-defined ecological niches. To the casual human observer there has been remarkable convergence, for example, the species living on twigs from different islands are well-nigh indistinguishable. Natural selection has led to the evolution of the same so-called ecomorphs evolving on different islands. The twig ecomorphs have a morphology that is adapted to the niche in which they live. Other ecomorphs live in the crowns of trees, on the trunks, on the ground, and so on.

While it may be easy to fool a casual human observer, would the same independently evolved ecomorphs be able to fool each other? That experiment has not been performed on *Anolis* lizards, but it has in a simpler system involving stickleback ecomorphs. At the end of the Pleistocene, after the glaciers had retreated from British Columbia, a marine stickleback invaded a number of freshwater lakes. In those lakes, parallel speciation occurred: a large benthic ecomorph evolved which feeds on invertebrates from the bottom of the lake, together with a smaller limnetic ecomorph which darts around in the open water living off plankton. There is strong assortative mating such that the benthic and limnetic ecomorphs will only mate with their own type in the wild, and have become true species. But would opposite sex individuals of the same independently evolved ecomorph from different lakes accept each other as mates? The answer is yes, in the laboratory. Rundle et al. (2000) took limnetic and benthic species from three separate lakes. Over 750 mating trials were set up in which individual females were presented with single males of the same or the alternative ecomorph, from the same or a different lake. Females were far more likely to spawn with a male of the same ecomorph as themselves, in which case it mattered not at all whether the male was from the same or a different lake.

See also: Adaptation, Fitness, and Evolution; Comparative Method, in Evolutionary Studies; Evolution, History of; Evolution: Optimization; Evolution: Self-organization Theory; Evolutionary Approaches in Archaeology; Extinction of Species; Species and Speciation

Bibliography

- Barraclough T G, Harvey P H, Nee S 1995 Sexual selection and taxonomic diversity in passerine birds. *Proceedings of the Royal Society London, Series B* **259**: 211–15
- Cavalli-Sforza L L, Edwards A W F 1967 Phylogenetic analysis:

- Models and estimation procedures. *American Journal of Human Genetics* **19**: 233–57
- Darwin C 1859 *On the Origin of Species by Means of Natural Selection, or, The Preservation of Favoured Races in the Struggle for Life*. John Murray, London
- Darwin C 1871 *The Descent of Man and Selection in Relation to Sex*. John Murray, London
- Felsenstein J 1978 Cases in which parsimony or compatibility methods will be positively misleading. *Systems in Zoology* **28**: 49–62
- Felsenstein J 1983 Parsimony in systematics: Biological and statistical issues. *Annual Review of Ecology and Systematics* **14**: 313–33
- Felsenstein J 1985 Phylogenies and the comparative method. *American Nature* **125**: 1–15
- Garland T J 1992 Rate tests for phenotypic evolution using phylogenetically independent contrasts. *American Nature* **140**: 509–19
- Gittleman J L 1981 The phylogeny of parental care in fishes. *Animal Behavior* **29**: 936–41
- Gittleman J L, Anderson C G, Kot M, Luh H-K 1996 Comparative tests of evolutionary lability and rates using molecular phylogenies. In: Harvey P H, Leigh Brown A J, Maynard Smith J, Nee S (eds.) *New Uses for New Phylogenies*. Oxford University Press, Oxford, UK
- Haldane J B S 1949 Suggestions as to quantitative measurement of rates of evolution. *Evolution* **3**: 51–6
- Harvey P H, Purvis A 1991 Comparative methods for explaining adaptation. *Nature* **351**: 619–24
- Hennig W 1966 *Phylogenetic Systematics*. University of Illinois Press, Urbana, IL
- Mitter C, Farrell B, Wiegmann B 1988 The phylogenetic study of adaptive zones: Has phytophagy promoted diversification. *American Nature* **132**: 107–28
- Møller A P, Birkhead T R 1992 A pairwise comparative method as illustrated by copulation frequency in birds. *American Nature* **139**: 644–56
- Page R M, Holmes E C 1998 *Molecular Evolution: A Phylogenetic Approach*. Blackwell Science, Oxford, UK
- Rundle H D, Nagel L, Boughman J W, Schluter D 2000 Natural selection and parallel speciation in sympatric sticklebacks. *Science* **287**: 306–8
- Sibley C G, Monroe B L 1990 *Distribution and Taxonomy of Birds of the World*. Yale University Press, New Haven, CT
- Sokal R R, Sneath P H A 1963 *Principles of Numerical Taxonomy*. Freeman, San Francisco
- Székely T, Reynolds J D 1995 Evolutionary transitions of parental care in shorebirds. *Proceedings of the Royal Society of London, Series B* **262**: 57–64

P. H. Harvey

Physical Activity and Health

The past three decades have witnessed a burgeoning of research examining the impact of physical activity on physical and mental health. This research has conceptualized physical activity as both a physiological and a psychological phenomenon. From the physiological perspective, physical activity is usually defined

as any force exerted by skeletal muscles that results in energy expenditure above the resting level. From a psychological perspective, physical activity is viewed as a behavior performed in specific social settings and controlled by motivational and volitional factors. Both perspectives complement one another, and they apply to different research questions. This article summarizes findings that refer to physiological as well as psychological aspects of the relationship between physical activity and health.

1. The Effects of Physical Activity on Physical Health

1.1 Overall Mortality

Epidemiological studies suggest that persons with moderate to high levels of regular physical activity or cardiorespiratory fitness have lower mortality rates than persons who are inactive or have low fitness levels (e.g., Sherman et al. 1999). This is supported by longitudinal research examining how changes in the level of cardiorespiratory fitness may affect subsequent mortality risks. For example, in a study by Blair et al. (1995), healthy men with an initially low level of cardiorespiratory fitness participated in two exercise tests about five years apart. Five years after the second exercise test, men in the bottom fifth of the initial fitness distribution who improved their fitness to at least a moderate level had a 44 percent lower mortality rate than men who remained in the bottom fifth. Similar results had already been found by Paffenbarger et al. (1993) in a cohort of initially sedentary Harvard alumni. Those men who started and maintained moderately intense activities extended their longevity on average by 0.72 years.

1.2 Cardiovascular Diseases (CVD)

Cardiovascular diseases, in particular coronary heart disease and stroke, are major causes of death, disability, and health care expenditures in Western societies. There is clear evidence that, within populations, the incidence of CVD is inversely related to habitual physical activity levels (Bauman and Owen 1999). Some of the studies in this field demonstrated a dose-response gradient between level of physical activity and risk of CVD outcome. Dose-response relationships were also found in those investigations in which measures of cardiovascular fitness, instead of physical activity estimates, were related to CVD mortality.

1.2.1 Coronary heart disease (CHD). Since 1953, about 40 studies analyzing the association between physical activity and risk of (CHD) have been pub-

lished. The first study in this field was conducted by Morris et al. (1953), who examined 31,000 employees of the London Transport Executive. A comparison between bus drivers and bus conductors revealed that physically active conductors clearly had a lower risk of a 'first coronary episode' than sedentary drivers. Classifying physically active and inactive people according to job title or occupational activity is typical for studies published before 1978. In the following years, level of physical activity was defined mostly by self-report measures of leisure-time activity or by indices that combined levels of occupational and leisure-time activities. More recent investigations have also emphasized the necessity to control for potentially confounding variables (e.g., age, sex, functional status) and to adjust for factors that are considered causal links between physical activity and CHD, such as serum cholesterol, blood pressure, diabetes, or body mass index.

Reviews of relevant literature have concluded consistently that physical activity is strongly related to a reduced risk of CHD. Furthermore, there is evidence in recent studies that this relationship holds true for both men and women. For example, findings from a prospective study among 72,488 female nurses who were 40–65 years old at the onset of the study suggest that sedentary women who became active in middle adulthood or later had a lower risk of coronary diseases than their counterparts who remained sedentary. Relative to women who walked infrequently, those who walked at a brisk pace three or more hours per week had a multivariate relative risk of 0.65 for coronary events (Manson et al. 1999). Physical exertion may temporarily increase the likelihood of an acute coronary event among persons with advanced coronary atherosclerosis who are not used to regular exercise. Although this risk needs to be taken seriously, it does not affect the general fact that physical activity is associated with a substantially lower overall risk for CHD.

1.2.2 Stroke. Contrary to the results for CHD, epidemiological data do not unequivocally confirm a relationship between stroke and level of physical activity. Most investigations in this field were planned as population-based cohort studies with follow-up ranges of 5–26 years. Among those studies that were designed to detect possible dose-response effects, many did not find a clear linear association. Some studies even suggested a U-shaped relationship between physical activity level and risk of stroke. By trying to explain these contradictory findings, it was assumed that the two different types of stroke, namely ischemic stroke and hemorrhagic stroke, need to be regarded separately. Since both types have specific pathophysiologicals, it may be possible that physical activity affects them in different ways. In a study by Abbott et al. (1994), inactive men had a

higher risk of hemorrhagic stroke than active men, but there was no main effect on ischemic stroke. More research is needed to further clarify whether physical activity plays a protective role against stroke.

1.2.3 Biological mechanisms. Several biological mechanisms may be responsible for the beneficial effects of physical activity on cardiovascular risks. These mechanisms involve the impact on blood pressure, atherosclerosis, plasma lipid/lipoprotein profile, availability of oxygenated blood for heart muscle needs (ischemia), thrombosis, heart rhythm disturbances, overweight, and healthier distribution of body fat.

1.3 Cancer

Cancer is one of the leading causes of death in industrialized countries. It is predicted that this proportion will increase in the next few years. Numerous studies have examined the role of physical activity in the prevention of cancer.

1.3.1 Colon cancer. Systematic literature analyses reveal that colon cancer is probably the most often-studied cancer in the field of physical exercise epidemiology. In 1996, a total of 29 studies on colon cancer were published (DHHS 1996). Of these, 18 studies considered only occupational physical activity, and 11 examined leisure-time or total physical activity. The results of the 18 studies using a measure of occupational physical activity draw a rather consistent picture: 14 studies found a significant inverse relationship between level of activity and risk of colon cancer; in four studies no such associations were observed. The 11 studies that referred to measures of leisure-time or total physical activity provided a similar pattern: eight reported an inverse association between level of physical activity and risk of colon cancer, while in three studies this was not the case.

In these studies, the level of physical activity during early adulthood was used as an effector variable. It is possible that this life period is too early to find an effect on colon cancer incidence in late adulthood. In most studies that controlled for potentially confounding dietary factors, the relationship between physical activity and risk of colon cancer remained significant. Taken together, the findings strongly suggest a significant association between level of physical activity and the occurrence of colon cancer.

1.3.2 Breast cancer. The report of the Surgeon General of 1996 reviews the results of 13 studies that examined the effects of habitual physical activity on risk of breast cancer (DHHS 1996). Of the 13 studies, four referred only to occupational physical activity;

the remaining nine studies used a leisure-time or total physical activity estimate. Of the four studies that focused on occupational physical activity, two found a significant inverse relationship between physical activity and breast cancer, but the other two did not. The nine studies considering leisure-time or total activity provide a contradictory picture: two studies found a significant inverse association, two reported a nonsignificant inverse association, three others could not observe any relationship, and the remaining two identified direct associations (i.e., the more active the person was, the higher the risk of cancer). Thus, there is only limited evidence for a protective effect of physical activity against the risk of breast cancer.

1.3.3 Conclusion. Epidemiological studies have also examined the relationship between physical activity and other cancers, such as prostate, testicular, ovarian, and endometrial cancer. Firm conclusions regarding these diseases cannot be drawn because data are often very limited or are inconsistent. The only clear association in this area seems to be the one between level of physical activity and the likelihood of developing colon cancer. To explain this finding, it has been assumed that physical activity reduces gastrointestinal transit time and thus diminishes the contact duration between the colon mucosa and potential carcinogens in the fecal stream.

1.4 Further Chronic Diseases

Level of physical activity has been related to further chronic diseases and conditions, in particular to diabetes, arthritis, osteoporosis, and obesity. Several prospective cohort studies lend strong support to the hypothesis that regular physical activity offers protection against the risk of noninsulin-dependent diabetes mellitus. Whether or not physical activity may prevent the development of osteoarthritis is still unknown. In postmenopausal women, physical exercise may impede the rapid decline in bone mass, thus protecting against osteoporosis, but this effect has not yet been unequivocally established. Reducing obesity by increased physical activity alone is a relatively slow process. Physical activity programs in combination with strategies of dietary management seem to be more effective in this regard.

2. The Effects of Physical Activity on Mental Health

There is a large amount of literature dealing with the psychological effects of physical activity. Numerous reviews and meta-analyses have dealt with this literature (Fox 1999, Morgan 1997). In studies examining the effects of physical activity on mental health, the most frequently observed outcomes include depres-

sion, anxiety, mood, self-esteem, self-efficacy, perceived stress, and cognitive functioning. On the other hand, the role of physical activity in the development and treatment of psychiatric disorders such as schizophrenia, dementia, personality disorders, or eating disorders has only seldom been investigated. The following section focuses on the association of physical activity with depression and anxiety (see *Depression, Hopelessness, Optimism, and Health*).

2.1 Depression

Research using correlational designs suggests a substantial relationship between physical activity and reduced symptoms of depression in men and women (Morgan 1997). In summarizing the existing literature, it was concluded that physically inactive persons are twice as likely to have symptoms of depression than their physically active counterparts. The association between level of physical activity and depression has been found in the general population as well as in clinical subpopulations (Fox 1999).

Correlational studies are often based on cross-sectional data, leaving the question unanswered whether physical activity has an effect on depression, or whether persons with good mental health are simply more likely to be active. In an attempt to resolve this problem of causality, cohort studies have been conducted in which the relationship between physical activity and depressive symptoms were examined in a longitudinal perspective. One of these cohort studies was conducted by Paffenbarger et al. (1994), who revealed that level of physical exercise reported at an initial interview was predictive of physician-diagnosed depression about 25 years later. In this investigation with 10,201 participants, the relative risk of depressive symptoms was 27 percent lower for men who had reported spending more than 3 hours of sports activities each week than for men who had reported no such activities. However, in other population studies, the level of physical exercise at baseline was not a risk factor for incidence of depressive symptoms years later (Weyerer 1992). Thus, the empirical evidence from cohort studies is still limited and inconsistent.

Another approach to resolve the problem of causality is intervention studies conducted in community or laboratory settings (Morgan 1997). A classical and often-cited work in this field is the experimental study by McCann and Holmes (1984). Altogether, 43 women students who scored higher than the cut-off point on the Beck Depression Inventory were randomly allocated to one of three groups: (a) aerobic exercise group, (b) placebo group, and (c) no treatment group. After five weeks, the aerobic exercise group showed a significant reduction in depressive symptoms compared with the two other groups, supporting the hypothesis that increased physical activity resulted in a reduction in depression. A more recent experimental study by King et al. (1993) demonstrated that greater

exercise participation was significantly related to fewer depressive symptoms, independent of changes in physical fitness or body weight. Other intervention studies have shown that the antidepressant effect of physical activity may be greatest in those with higher levels of initial depressive symptoms (Brown et al. 1995), suggesting that the psychological benefits of exercise are likely to be different for different people.

Several biological and psychosocial mechanisms have been proposed in order to explain the antidepressant effects of physical activity. In particular, it has been hypothesized that physical activity may induce changes in brain neuroreceptor concentrations of monoamines (norepinephrine, dopamine, or serotonin) or endogenous opiates (endorphins and enkephalins), which in turn are responsible for changes in the affective level (Morgan 1997). However, the observed depression-reducing effects of physical exercise may also be mediated by psychological factors such as exercise-induced feelings of mastery and self-efficacy, or by positive social and emotional experiences in the sport context.

2.2 Anxiety

Research has also suggested that physical activity may be linked to a reduction in anxiety. The psychological assessment of anxiety usually includes state and trait measures. State measures refer to the actual feeling of anxiety, whereas trait measures evaluate the stable personality characteristic of anxiety. In physical exercise research, state measures have been used mainly in pre-post-designs of intervention studies to explore whether a specific activity episode may have triggered changes in currently experienced anxiety. Results are relatively unequivocal in this respect. Immediately after a session of physical activity, most people report an improvement in mood, including a reduction in level of state anxiety (Berger and Owen 1998). However, this decrease in anxiety may persist for only 2–6 hours following the activity episode (Morgan 1997). In order to continuously profit from the anxiety-reducing effects of physical activity, it is obviously necessary to engage in physical activity on a daily basis. There is no clear evidence that regular physical exercise may have a general impact on trait anxiety (Schlicht 1994).

2.3 Further Psychological Effects

Physical activity has been considered to be a stress buffer in critical life situations (Fuchs and Hahn 1992). This has been explained by two different processes: physical exercise may affect the transactional stress-coping process, either by influencing an individual's appraisal of the potentially stressful situation, or acting as an emotion-focused coping strategy. However, empirical evidence supporting the stress-buffering role of physical activity is still very limited.

Another psychological construct often linked to physical exercise is self-esteem. In his review, McAuley (1994) concluded that the positive association between physical activity and self-esteem is observed both as an enduring consequence of habitual exercise training and as an immediate, short-term effect of single exercise episodes. Finally, several studies suggest that there is a direct relationship between level of physical activity and measures of cognitive performance, such as reaction time, nonverbal reasoning, and memory.

3. Conclusions

There is clear evidence that physical activity is significantly associated with reduced risk of all-cause mortality, all cardiovascular diseases combined, CHD, hypertension, colon cancer, and non-insulin dependent diabetes mellitus. Findings also strongly suggest that regular physical activity may reduce the risk of developing obesity, osteoporosis, depression, and state anxiety. However, effects on psychological factors such as cognitive functioning, self-esteem, body image, self-efficacy, control beliefs, or optimism are less clearly documented. Future research should focus on these emotional and cognitive factors that may be mediators in the relationship between physical activity and mental health. Based on data from epidemiological studies, it has been concluded that physical activity that increases the weekly energy expenditure by about 1,000 kcal (equivalent to about 90 minutes of running or 120 minutes of jogging) may be associated with substantial physical health benefits (DHHS 1996). But how much physical activity is needed to improve or maintain psychological well-being? Probably the mental health benefits of physical exercise are not only a matter of expended kilocalories (if at all), but also of exercise-related social interactions and emotional experiences. Clarifying the mechanisms by which physical activity may affect mental health dimensions is a challenge to sport and health psychology in the coming years.

See also: Depression, Hopelessness, Optimism, and Health; Health Behaviors; Physical Activity and Health Promotion; Sport and Exercise Psychology: Overview

Bibliography

- Abbott R, Rodriguez B, Burchfiel C, Curb J 1994 Physical activity in older middle-aged men and reduced risk of stroke: The Honolulu Heart Program. *American Journal of Epidemiology* **139**: 881–93
- Bauman A, Owen N 1999 Physical activity of adult Australians: Epidemiological evidence and potential strategies for health gain. *Journal of Science and Medicine in Sport* **2**: 30–41
- Berger B, Owen D 1998 Relation of low and moderate intensity exercise with acute mood change in college joggers. *Perceptual and Motor Skills* **87**: 611–21
- Blair S N, Kohl H, Barlow C, Paffenbarger R, Gibbons L, Macera C 1995 Changes in physical fitness and all-cause mortality: A prospective study of healthy and unhealthy men. *Journal of the American Medical Association* **273**: 1093–8
- Brown D R, Wang Y, Ward A, Ebbeling C, Fortlage L, Puleo E et al. 1995 Chronic psychological effects of exercise and exercise plus cognitive strategies. *Medicine and Science in Sports and Exercise* **27**: 765–75
- DHHS US Department of Health and Human Services 1996 *Physical Activity and Health: A Report of the Surgeon General*. US Department of Health and Human Services, Atlanta, GA
- Fox K 1999 The influence of physical activity on mental well-being. *Public Health and Nutrition* **2**: 411–18
- Fuchs R, Hahn A 1992 Physical exercise and anxiety as moderators of the stress-illness relationship. *Anxiety, Stress, and Coping: An International Journal* **5**: 139–49
- King A C, Taylor C, Haskell W 1993 Effects of differing intensities and formats of 12 months of exercise training on psychological outcomes in older adults. *Health Psychology* **12**: 292–300
- Manson J, Hu F, Rich-Edward J, Coldit G, Stampfer M, Willett W et al. 1999 A prospective study of walking as compared with vigorous exercise in the prevention of coronary heart disease in women. *New England Journal of Medicine* **341**: 650–58
- McAuley E 1994 Physical activity and psychosocial outcomes. In: Bouchard C et al. (eds.) *Physical Activity, Fitness, and Health: International Proceedings and Consensus Statement*. Human Kinetics, Champaign, IL pp. 551–68
- McCann L, Holmes D S 1984 Influence of aerobic exercise on depression. *Journal of Personality and Social Psychology* **46**: 1142–7
- Morgan W P (ed.) 1997 *Physical Activity and Mental Health*. Taylor & Francis, Washington, DC
- Morris J N, Heady J, Raffle P, Roberts C, Parks J 1953 Coronary heart disease and physical activity of work. *Lancet* **2**: 1111–20
- Paffenbarger R, Hyde R, Wing A, Lee I, Jung D, Kampert J 1993 The association of changes in physical activity level and other lifestyle characteristics with mortality among men. *New England Journal of Medicine* **328**: 538–45
- Paffenbarger R S, Lee I, Leung R 1994 Physical activity and personal characteristics associated with depression and suicide in American college men. *Acta Psychiatrica Scandinavica Supplementum* **377**: 16–22
- Schlicht W 1994 Does physical exercise reduce anxious emotions? A meta-analysis. *Anxiety, Stress, and Coping* **6**: 275–88
- Sherman S, D'Agostino R, Silbershatz H, Kanne W 1999 Comparison of past versus recent physical activity in the prevention of premature death and coronary artery disease. *American Heart Journal* **138**: 900–7
- Weyerer S 1992 Physical activity and depression in the community: Evidence from the Upper Bavarian Field Study. *International Journal of Sports Medicine* **13**: 492–6

R. Fuchs

Physical Activity and Health Promotion

Physical as well as mental health are stabilized and promoted by regular physical activity. The shortness of this article forces us to examine its health-promoting effects only in brief. The main reflections will conc-

entrates on two subjects: health promotion as a science-based theory to change health damaging behavior and the phenomenon that people of Western societies lead a sedentary lifestyle despite the evident benefits of activeness.

1. *Physical Activity and Health*

Disease prevention by regular physical activity has become more and more important. Its significance has been shown in various studies. Epidemiological and naturalistic as well as experimental studies demonstrate the risk-reducing and health-enhancing effects of regular physical activity. In particular, aerobic training seems to be effective in preventing cardiovascular diseases. The psychological benefits of physical activities are decreasing negative emotional states (e.g., anxiety) and increasing positive emotional states (e.g., vigor).

Whereas the effect of exercise on physical health can be explained by adaptation processes of the metabolic and cardiovascular system, the effects on mental health remain unexplained. Different kinds of hypotheses assume physiological (e.g., thermoregulation, endorphins), psychological (e.g., attention distraction, states of *flow*), or rather unspecific mechanisms (e.g., effort justification). Further research is needed to enlighten these mechanisms. However, it can be stated that physical exercise is an appropriate means for health promotion in most cases.

2. *Physical Activity and Health Promotion: Two Complementary Strategies*

Health promotion is a scientific conception to help people change their lifestyles towards a state of optimal health. Recently published theories conceive health and disease as independent dimensions. In view of such a concept, reducing cardiovascular risk by regular exercise does not lead intentionally, at the same time, to optimal physical, mental, and social wellbeing. The strategy of health promotion therefore has been extended in the past decades. Whereas in the 1980s health promotion was defined as a disease preventing strategy, it has been supplemented by a setting-based lifestyle strategy in the 1990s. Health promotion today focuses on disease prevention and health enhancement, while also aiming at changing behavior and improving the living conditions of social groups and individuals.

2.1 *Physical Activity and Health Promotion: The Behavior-oriented Strategy*

When focusing on personal behavior change without considering the environmental conditions, health pro-

motion seems to have rather limited success. About 50 percent of the population in the USA do not adopt regular exercise and 25 percent are not active at all (USCDCP 1994). Those who begin an exercise program often fail to adhere to it. Most of them relapse to inactivity after the first six months (Marcus et al. 2000). Previous preventive interventions have often used threatening contents to convince people to give up inactivity (e.g., threat of coronary artery disease).

When reviewing the literature on fear communication Leventhal (1970) described how threatening intervention practices are often provoking a negative relationship between fear and persuasion. So there is no linear relationship between threat and motivation to quit health-damaging behavior. One cause of this paradoxical effect is well documented in social psychology literature as the phenomenon of reactance (a reaction to feel a restricted liberty of behavior).

In the last decade of the twentieth century Health Psychology applies theoretical models (see *Health Behavior: Psychosocial Theories*) which define the cognitive determinants of health behavior (continuum theories) and construe behavior change as a stepwise process (stage theories). Current continuum theories are the Health-Belief-Model, the Protection-Motivation-Theory, the Theory-of-Planned-Behavior and the Subjective-Expected-Utility-Theory (see *Health Behavior: Psychosocial Theories*). Empirical tests verify at least four variables as predictive for health behavior in general (see *Health Behaviors*). A behavioral change is probable on condition that a person perceives health hazards as being likely (vulnerability) and severe (severity), and on the condition that this person is supplied with disease preventing alternatives (outcome expectancies), and has enough personal efficacy (self-efficacy) to change a risky lifestyle.

With regard to physical exercise, perceived behavior control, positive outcome expectancies, social support and severity, as well as vulnerability, define valid and reliable sociocognitive predictors determining adoption to exercise for men and women in different ways (Sallis et al. 1992). In contrast, knowledge about intervention strategies that try to foster exercise adherence or maintenance to physical activity, is at an early stage (Marcus et al. 2000).

Valid social and demographic predictors of health-related activities are mainly age, gender, neighborhood environment, school attainment, ethnicity, and income aside from sociocognitive variables. This means that women, elderly people, people with a lower education, low income, ethnic minorities, and persons living in an area with low recreational potential, are little active in sports.

Moreover health-related behavior (including health damaging activities) has proved to be quite stable (Sutton 1994). It is obvious that inactive people are not aware of the risks of leading a sedentary lifestyle. Continuum theories are scarcely suited to explain

changes from the sedentary-oriented stage to that of readiness for physical activity.

Stage theories (Transtheoretical-Model, Precaution-Adoption-Process) seem rather adequate to deal with this phenomenon (Weinstein et al. 1998). They postulate distinguishable stages in which varying processes pass, leading to different levels of change: for example, precontemplation, contemplation, active change, maintenance, and perhaps relapse. Adoption of exercise has been investigated by various authors, most of them using a cross-sectional comparison of people believed to be in different stages (e.g., Prochaska 1994). The results indicate an increase in the perceived benefits of exercise from a pre- to a postdecisional stage. More experimental and longitudinal research is needed to clarify the conditions under which people change from a sedentary to an active lifestyle.

2.2 Physical Activity and Health Promotion: The New-Public-Health Approach

Many health problems and severe diseases in modern societies are surely caused by overeating, alcohol consumption, drug abuse, and a sedentary lifestyle. Therefore, the question of how to change health-damaging lifestyles still remains important. The models and theories mentioned above help to develop efficient intervention strategies.

Physical activity is part of a health-enhancing lifestyle. The achievement of physical activity goes beyond reducing the risk of cardiovascular diseases. Contributing to social integration gives the feeling of internal control and—especially with young people—provides the conditions for learning adequate, rule-based social behavior (see Danish et al. 1996). Exercising in groups seems to promote the learning of suitable skills to mobilize social networks.

British studies especially, however (see Blaxter 1990), substantiate that particularly middle- and upper-class members of modern societies cling to a health-enhancing lifestyle. Body image and its relevance to physical activity is now a subject of growing importance for those with a modern, middle-class lifestyle. As can be suspected, inequalities in exercise engagement are not only caused by personal income, though this is a premise for taking part in physical activities such as aerobics, calisthenics, and power training in fitness studios. Causes may be found more likely in the 'social environment' of persons. Social environment construes informal as well as formal rules and norms and it provokes attitudes and values, feelings, thoughts, and actions, which then determine the actual position of a person in a given setting (e.g., workplace, school, family). From this point-of-view a certain health-damaging behavior (e.g., inactivity) may be functional in so far as it demonstrates identity with a defined social group or class. Empirical facts show that lower-class people share a mechanistic view

of the body. In general they do not expect to have personal influence on health and wellbeing. The consistent epidemiological results from USDCDP (1994) that are mentioned above, demonstrate that less physical activity participation correlates with lower educational attainment, lower income, minority status, old age, and being overweight. These factors seem to be significant barriers to exercise.

Interventions merely focusing on health-damaging behavior run the risk of ignoring the constraints and barriers of behavior change. They are liable to lead towards 'healthism.' They are likely to provoke moralistic attitudes such as blaming the victim (see Diekstra and Jansen 1990). Health promotion in the sense of New Public Health is a crossdisciplinary science that avoids such misinterpretation by preferring an ecological perspective. The *Ottawa Charter for Health Promotion* of the WHO pleads for an enlargement in the paradigm of the intervention process. Health promotion in this sense considers both personal behavior and personal environment. It includes disease prevention as well as health enhancement. Thus health promotion has the chance to look for daily life resources which are health enhancing. Those resources are, among other things, sufficient income, adequate command of work, ability to cope with daily hassles and daily uplifts, social integration and social support, self-esteem, and the feeling of internal control (Diekstra and Jansen 1990). As already mentioned, exercising in groups can contribute to the strengthening of these resources; therefore all social groups should have the opportunity to profit from these resources as well.

King et al. (1997) show that participation in an exercise program depends on its demands and type. They compared two groups of women—those who had less than 12 years of education with those who had more than 12 years. The Stanford-Sunnyvale Health-Improvement Project (SSHIP) found that the less-educated group, which was given an adequate home-based exercise program, had relatively high rates of long-term exercise adherence (two years). So, if matched to the right program, subgroups of lower classes have the chance to do quite well. Studies like the SSHIP will help to identify those variables, which can provide information on when and why people become more or less active over the course of time.

3. Conclusions

Physical activity is an appropriate means of enhancing physical as well as psychological wellbeing. Despite the growing evidence of the health benefits of regular physical activity, about 50 percent of adult people remain sedentary, especially those in lower socioeconomic status groups. Research in Health Psychology as well as in New Public Health is challenged to clarify those conditions and program parameters which are appropriate to enhance the motivation to

exercise and which bind people together in long-term regular exercise programs.

See also: Health Behavior: Psychosocial Theories; Health Behaviors; Health Behaviors, Assessment of; Physical Activity and Health; Public Health; Sport and Exercise Psychology: Overview; Sport, Anthropology of

Bibliography

- Blaxter M 1990 *Health and Life Styles*. Routledge, London
- Danish S J, Nellen V C, Owens S 1996 Teaching life skills through sport: Community based programs for adolescents. In: van Raalte J L, Brewer B W (eds.) *Exploring Sport and Exercise Psychology*. APA, Washington, DC
- Diekstra R, Jansen M A 1990 The importance of psychosocial interventions in primary health care. In: Schmidt L R, Schwenkmezger P, Maes S (eds.) *Theoretical and Applied Aspects of Health Psychology*. Harwood Academic Publishers, Chur, Switzerland, pp. 87–101
- King A C, Kiernan M, Oman R F, Kraemer H C, Hul M, Ahn D 1997 Can we identify who will adhere to long-term physical activity? Signal detection methodology as a potential aid to clinical decision making. *Health Psychology* **16**: 380–9
- Leventhal H 1970 Findings in theory in the study of fear communications. In: Berkowitz L (ed.) *Advances in Experimental Social Psychology*. Academic Press, New York, pp. 119–86
- Marcus B H, Dubbert P M, Forsyth L A, McKenzie T L, Stone E J, Dunn A L, Blair S N 2000 Physical activity behavior change: Issues in adoption and maintenance. *Health Psychology* **19**: 32–41
- Prochaska J O 1994 Strong and weak principles for progressing from precontemplation to action on the basis of twelve problem behaviors. *Health Psychology* **13**: 47–51
- Sallis J F, Hovell M F, Hofstetter C R 1992 Predictors of adoption and maintenance of vigorous physical activity in men and women. *Preventive Medicine* **21**: 237–51
- Sutton S 1994 The past predicts the future: Interpreting behaviour-relationships in social-psychological models of health behaviours. In: Rutter D R, Quine L (eds.) *Social Psychology and Health: European Perspectives*. Avebury, Aldershot, UK, pp. 71–88
- USDCDP 1994 Prevalence of sedentary lifestyle—behavioral risk factor for chronic disease by education level in racial/ethnic population—United States, 1991–1992. *Morbidity and Mortality Weekly Report* **43**: 894–9
- Weinstein N D, Rothman A J, Sutton S R 1998 Stage theories of health behavior: Conceptual and methodological issues. *Health Psychology* **17**: 290–9

W. Schlicht

Physical Sciences: History and Sociology

The twentieth century was the century of the physical sciences. Immensely successful at making sense of the material world, they are also deeply embedded in the social fabric, and have become associated with technological development, economic well-being, and

national security and prestige. The physical sciences have also provided the knowledge required to produce a vast array of weapons and weapons systems. Their history and sociology are contested terrains, in which the issues at stake are commensurate with the ambiguity of their social role.

1. Object

The physical sciences as a coherent body of practices emerged in Western Europe in the seventeenth century (Dijksterhuis 1986/1950, Shapin 1996). At the time they encompassed the fields of physics, astronomy, and mechanics. They were at the core of a radical epistemological and social project which defined nature in mechanistic terms and as subject to law-like and predictable behavior. Their aim has been to understand but also to exploit and to manipulate their objects. Their means has been a combination of ‘reason’ and ‘experience,’ both of which had to respect locally defined and shared canons of good practice. Their practitioners mobilized mathematical tools and capitalized on the rigor of mathematical demonstrations to consolidate the logical force of their truth claims. They extended the reach of the human senses by building and using increasingly complex and powerful instruments which allowed for the controlled exploration of the properties and processes of matter.

2. Scope

The scope of the physical sciences has constantly evolved to include fields as diverse as chemistry, electricity, optics, and the study of solids. Beginning in the late nineteenth century, their potential for not simply understanding but also manipulating the material world led to their institutionalization, and the formalization of their links with the state, philanthropic foundations, and large corporations. Patrons and philosophers alike were impressed by their intellectual and pragmatic success, as revealed by the panoply of technologies which they spawned and on which millions of people depended for their well-being. Characterized as a set of techniques and practices, they were seen as a model to be emulated and reproduced. Traditional forms of knowledge were appropriated and subjected to their methods. Parts of the life sciences and the human and social sciences were colonized. Existing disciplinary boundaries were dissolved and new specialisms and fields emerged in the spaces thus created (Krige and Pestre 1997).

Consider the relationship between mathematics and the physical sciences. For Galileo and those who shared his intellectual project in the seventeenth and eighteenth centuries, the ‘book of nature’ was written in mathematical symbols and the laws governing its behavior could be expressed in mathematical form. This local unity between the physical sciences and mathematics persisted in the majestic constructions of

a Newton or a Laplace (Koyré 1985). Later, however, some mathematicians began to detach their objects from the empirical and the experimental (Bottazini and Dahan 2001). Particularly in Germany they reverted to a 'pure' mathematics in opposition to, and to defend themselves against, the rising intellectual and institutional importance of a mathematically informed theoretical physics which was based on, and guided, experimental work. It was this theoretical physics which transformed our conception of the structure and properties of matter in the late nineteenth and early twentieth centuries, combining theoretical analysis with philosophical reflection.

The exigencies of World War II swept aside these cultural norms and intellectual methods. A new kind of physics emerged, motivated by a pragmatic demand to get numbers out and to solve problems efficiently using an evolving toolkit of mathematical techniques. This problem-oriented rather than discipline-based approach persisted in the postwar era. It was entrenched through summer schools and conferences, and by interdisciplinary teamwork in universities and corporations, frequently undertaken with the needs of the military and the Cold War in mind. Pure mathematics continued with a life of its own, dedicated to abstraction, and often pursued with idiosyncratic brilliance in schools like the Bourbaki group in France. Its members saw themselves as the priests of high culture and disdained the vulgar materialism of those 'mathematicians' whose work was enmeshed with the empirical and the experimental.

Proponents of the physical sciences, actively encouraged by patrons, have repeatedly sought to colonize neighboring fields, redefining their objects so as to impose their methods and practices on them. The molecularization of biology is a case in point (Abir-Am 1982, Kay 1996). In the 1930s, officers of the Rockefeller Foundation, believing that the study of life needed to be purged of what they saw as its metaphysical content, actively promoted the migration of concepts and of instruments and techniques developed in the physical sciences, along with their practitioners, into biological research. The study of macromolecules of biological origin in the 1930s, and the unraveling of the molecular structure of DNA in the 1950s consolidated this reductionist project and contributed to the emergence of a new discipline, molecular biology. It was articulated around the concept of information, made use of instruments like x-ray crystallographs and electron microscopes, and benefited from an interdisciplinary symbiosis between physicists and researchers in the life sciences.

3. Practices

The physical sciences have always been enmeshed in complex matrices with other sciences, technology, and various social actors. This process was accelerated and consolidated in World War II. The development of

radar, nuclear bombs, and long-range rockets mobilized physicists at the core of interdisciplinary teams which brought together, in addition, material scientists, chemists, mathematicians, and engineers in vast techno-scientific systems. These systems were shaped by the military and demanded unquestioning political and financial support from the state and the resources of major corporations. They also required the collaboration of hundreds and thousands of 'non-scientific' men and women who were mobilized for the war effort, either voluntarily (as in the case of women who worked around the clock in radar stations equipped to detect enemy submarines off the South African coast) or under conditions of appalling cruelty (as in the forced labor camps which built V2 rockets for the Nazi regime in the 1940s).

World War II institutionalized the hierarchical and planned organization of research in large interdisciplinary teams centered around a project or a problem (Galison 1997, Galison and Hevly 1992). The practice of some of the physical sciences after the war, notably in high-energy physics, reproduced and exemplified many of these features. It has come to be taken as the paradigmatic 'big science.' Particle physics brings together hundreds of researchers from many different institutions around major accelerators and the detectors used to explore the collision processes produced in them. It is patronized by the state as an icon of prestige and a training ground for some of the most brilliant members of a strategic elite. It relies on advanced sectors of industry for the development and supply of sophisticated technologies (in areas like high vacuum, electronics, and superconductivity).

This change of scale has demanded that physicists become adept at a variety of skills and bow to constraints imposed by various regimes. The size of the equipment built and used and the numbers who work with it, the danger of explosion with the cryogenic detectors (bubble chambers) deployed above all in the late 1950s and 1960s, and the competing demands for opportunities to perform experiments have all transformed the material culture of physics. For leaders in the field, being a physicist does not simply mean pursuing one's specialty in a multidisciplinary team. It means being adept at fundraising, at project management, and at dealing with industry. It also requires that one accepts the bureaucratization of experimental work. Committees set research priorities and allocate beam time and floor space. The laboratory directorate imposes policies intended to protect the interests of the research community which depends on the facility it manages, and of the government or organization which funds it.

4. Instruments

Instruments lie at the core of the physical sciences. They are the essential means to the production and certification of knowledge. They are also complex and

capricious (Collins 1992, Latour 1987, Shapin and Shaffer 1985). One has to be trained to use them, learning how to control their irregularities and to make sense of their output. The results obtained, and the truth-claims made, are impregnated with the vagaries of the instruments and the tacit, hands-on knowledge and experience of those who used them. Replication, preferably by others elsewhere with an instrument which is 'similar in the relevant respects,' is the ideal and idealized strategy used to transform these local truth-claims into universal truths which tell us something about the structure of the world. Such similarity is contestable, and often contested. It also presupposes not only agreed protocols of experimental manipulation but also a level of technological development congenial to the construction of standardized, precision instruments. This level was not reached in Europe until the nineteenth century. Instrument-makers, be they gifted craftsmen with tiny workshops in the streets of London or Paris a century or two ago, specialized firms developing and marketing standardized laboratory tools, or giant industries supplying heavy equipment, thus provide the essential technical and technological infrastructure demanded by the physical sciences.

The production of standardized instruments requires the integration of theories of how the instrument works into its design. Simple theories of optics were embedded in the telescopes, microscopes, and binoculars produced by the vast eyeglass industry built up by Zeiss at the end of the nineteenth century. Scientists themselves have played an active role in developing theoretically based instruments dedicated to their research agendas. Marie Curie not only amassed radium in order to pursue her research, she also promoted and financially backed the construction of measuring instruments which she could use to establish and impose standards for radioactivity (Boudia and Roqué 1997). Ernest Lawrence built his laboratory and his reputation on the construction of ever-more powerful particle accelerators based on a theoretical understanding of the behavior of charged particles in electromagnetic fields (Heilbron and Seidel 1989). Standardized, these instruments were promoted as tools to cure cancer, were used to manufacture enriched uranium for the Manhattan project during World War II, and were advocated as particle beam weapons during the Cold War. Felix Bloch's recondite research into the magnetic properties of the nucleus, for which he shared a Nobel prize for physics in 1952, suggested a new and sensitive technique for identifying the components of compound mixtures. Working along with his graduate students at Stanford University and the local firm Varian Associates, which specialized in resonance techniques, his nuclear magnetic resonance instrument became a standard tool for chemical analysis in the 1960s.

Instruments can inspire awe, and nowhere more than in the physical sciences where their complexity

and size have often been exploited to legitimate the practices in which they are embedded and to increase the social weight of those who build them and demonstrate their powers. The sixteenth-century clock in Strasbourg cathedral was a masterpiece of craftsmanship. It was also a material illustration that the mechanical view of nature, congenial to and promoted by the partisans of the 'new philosophy,' was a feasible alternative to the dominant conception which took the natural world to be throbbing with life, intelligence, and purpose. The spectacular discharges sparked by Ruhmkorff coils in public arenas in the late nineteenth century thrilled a populace fascinated by an electricity which was replete with associations to medicine, spiritualism, and technological application. Giant particle accelerators and detectors (and telescopes), regularly put on display to impress the general public, promote and seek to justify the enormous cost of these recondite fields of research. Their size and power are limited only by technological feasibility, financial constraints and, today, political will. Each successive generation of instruments is reputedly bigger and better than its predecessors, as measured in easily understood, readily quantified units of energy (or resolving power). Each demonstrates scientific and technological leadership, inflates national pride, and fuels international rivalry.

5. Patronage

Patronage has made possible the emergence, consolidation, and then vast expansion of the sphere of influence of the physical sciences, notably that which has occurred over the past 150 years. Patronage by kings, emperors, the military, national governments, and private foundations has created the social niches and institutions and supplied the financial resources the practitioners of the physical sciences have needed to promote their projects. In return they have held out the promise that the knowledge they produced would improve not only understanding, but also, through that, the prestige and power of those who supported them.

Accommodation and compromise has usually characterized the relationship between practitioner and patron (Mukerji 1990, Price 1965). Both partners have fashioned identities and ideologies, and negotiated boundaries of power and autonomy, compatible with their mutual dependence on each other. Exceptionally that dependence has been denied, or subjugated to other dominant ideologies, as in the grotesque persecution of Jewish scientists by the Nazi regime in Germany in the 1930s. More usually pragmatism prevails, as in the unseemly race between the allies to secure the services of the scientists and engineers who had developed missiles used by that selfsame regime to bombard major European cities. The pragmatism of

patrons is matched by that of those whom they support; practitioners of the physical sciences have shown themselves remarkably adept at establishing a *modus vivendi* with whatever political regime provides them with the institutional and financial resources they need to pursue their work. The 'universalism' and internationalism of science so actively promoted in the twentieth century has gone along with the depoliticization of the community and the propagation of an ideology which has decoupled scientists as producers from their patrons as users of the knowledge which they have, in fact, co-produced.

The increasingly costly and complex equipment demanded by the physical sciences has led governments, above all in Western Europe, to pool resources and to fund collectively international laboratories and organizations (Krige and Guzzetti 1997). These collaborative initiatives have generally been restricted to basic research in domains of little direct strategic importance (like high-energy physics, nuclear fusion and space science, and also molecular biology). Even here the loss of sovereignty involved can be a disincentive to the participating governments. They drain resources from national programs and shift the locus of policy to an arena in which practitioners and patrons are locked into scientific, technical, financial, and political decision-making processes over which they have limited control. Indeed, the stability of intergovernmental schemes in the physical sciences depends on practitioners successfully persuading national authorities of the scientific and technical need for the laboratory and its equipment and on patrons defining the collaborative venture as an instrument of foreign policy. This has been particularly valuable for some domains of the physical sciences in the aftermath of the Cold War. With their field no longer perceived as an icon of national prestige and global power, nuclear physicists in the United States and even more dramatically in the ex-Soviet Union have seen support for their research falter, and have had to put their skills at the disposal of new and sometimes exotic patrons (like Wall Street brokers and investment banks). Their colleagues in Western Europe, by contrast, have continued to benefit from governments' use of their science as a tool to extend the 'European family' of nations to include many from the former Soviet bloc.

6. Historiography

Currently there are, broadly speaking, two ways of writing the history of the physical sciences (Golinski 1998). One approach focuses on product. It takes the output of research as an unproblematic corpus of established truths about the world. It is progressivist, and draws attention to scientific achievements and the logical, unidirectional connections between them. It is coupled with a sociology which defines scientific know-

ledge as comprising verified truths which are independent of social determination, and a corresponding view of the scientist as the member of a self-regulating community searching for objective knowledge.

The other approach focuses on process and practice. Inspired by sociological and anthropological arguments which insist that truth-claims are underdetermined by empirical evidence, it problematizes the products of research. It explores the social practices whereby consensus is forged among scientists, the processes whereby knowledge is legitimated in different times and places, and how it is reconfigured as it circulates among various groups of social actors. If for the former approach truth is simply a relationship between statements and the world, for the latter it is also a social accomplishment.

Each of these approaches emerged in a specific sociohistorical context and was shaped by a corresponding concept of science and the social role of its history. The focus on product was intended to celebrate science and scientists as the bearers of Reason and Truth, progressive forces opposing 'irrational' political ideologies which sought to impede their development and impose constraints on their field of action. The focus on process is a *fin de siècle* reaction to the success of that project. It seeks to puncture the pretensions of a science and of scientists whose cognitive authority and social power is immense, at least as far the physical sciences are concerned in some advanced industrialized countries. By insisting that social processes are central to the production of knowledge, it seeks to tie truth claims to the local contexts in which they emerge and are contested, and to draw attention to the uncertainty, ambiguity, and messiness which pervades scientific practice at the research frontier.

Each of these approaches has its associated political agendas and neither can lay claim to universal applicability. The former serves to defend the scientific approach wherever the appeal to empirical evidence and rational argument is threatened by institutionalized intolerance, and where science and technology have not yet been bent to alleviate poverty and suffering. The latter seeks to demythologize science and scientists wherever they have become entrenched and empowered in social structures, questioning the uncritical acceptance of science as a force for progress and of scientists as experts whose truth-claims and whose power lie beyond democratic debate and control.

See also: Biomedical Sciences and Technology: History and Sociology; History of Science; History of Science: Constructivist Perspectives; History of Technology; Human Sciences: History and Sociology; Scientific Disciplines, History of; Scientific Instrumentation, History and Sociology of; Scientific Knowledge, Sociology of

Bibliography

- Abir-Am P G 1982 The discourse on physical power and biological knowledge in the 1930s: A reappraisal of the Rockefeller Foundations 'policy' in molecular biology. *Social Studies of Science* **12**: 341–82
- Bottazini U, Dahan A 2001 *Changing Images of Mathematics. From the French Revolution to the New Millennium*. Harwood Academic Publishers, Chur
- Boudia S, Roqué X 1997 Science, medicine and industry: the Curie and Joliot–Curie laboratories. *History and Technology* **13**(4): 241–354
- Collins H 1992 *Changing Order. Replication and Induction in Scientific Practice*, 2nd edn. University of Chicago Press, Chicago
- Dijksterhuis E J 1986/1950 *The Mechanization of the World Picture: Pythagoras to Newton*. Princeton University Press, Princeton, NJ
- Galison P 1997 *Image and Logic. A Material Culture of Physics*. University of Chicago Press, Chicago
- Galison P, Hevly B 1992 *Big Science. The Growth of Large-scale Research*. Stanford University Press, Stanford, CT
- Golinski J 1998 *Making Natural Knowledge. Constructivism and the History of Science*. Cambridge University Press, Cambridge, UK
- Heilbron J L, Seidel R W 1989 *Lawrence and his Laboratory. A History of the Lawrence Berkeley Laboratory*. University of California Press, Berkeley, CA
- Kay L E 1996 *The Molecular Vision of Life: Caltech, the Rockefeller Foundation and the Rise of the New Biology*. Oxford University Press, London
- Koyré A 1985 *From the Closed World to the Infinite Universe*. Chapman and Hall, London
- Krige J, Guzzetti L 1997 *History of European Scientific and Technological Cooperation*. European Commission: Science, Research, Development, Luxembourg
- Krige J, Pestre D 1997 *Science in the Twentieth Century*. Harwood Academic Press, Chur
- Latour B 1987 *Science in Action. How to Follow Scientists and Engineers Through Society*. Harvard University Press, Cambridge, MA
- Mukerji C 1990 *A Fragile Power. Scientists and the State*. Princeton University Press, Princeton, NJ
- Price D K 1965 *The Scientific Estate*. Harvard University Press, Cambridge, MA
- Shapin S 1996 *The Scientific Revolution*. University of Chicago Press, Chicago
- Shapin S, Schaffer S 1985 *Leviathan and the Air Pump: Hobbes, Boyle and the Experimental Life*. Princeton University Press, Princeton, NJ

J. Krige

Physicalism and Alternatives

Physicalism is the thesis that all there really is is whatever there has to be given the way reality is physically. Thus, if physicalism is true and there really are people, shopping malls, cities, and countries, then, given the way reality is physically, there have to be people, shopping malls, cities, and countries. The

alternatives to physicalism imply that there are actual phenomena in addition to whatever phenomena there have to be given the way reality is physically. Substance dualism is an alternative that maintains that there are actual substances—for example, immaterial souls, or entelechies, or disembodied spirits—that need not exist given the way reality is physically. Property dualism, another alternative to physicalism, maintains that there are properties—for example, irreducible mental properties or irreducible moral properties—whose actual pattern of distribution need not occur given the way reality is physically. Physicalism is a controversial thesis. Even if it is true, it might have been false; one or more of the alternative views might have been true instead. The debate over whether physicalism or one of its alternatives is true is thus an empirical debate.

1. Physicalism

1.1 Formulating Physicalism

So-called type physicalism is the doctrine that every type or kind of thing is physical. Physicalists need not be committed to type physicalism. Many physicalists allow that some kinds of things are such that is logically possible for them to have instances in, for instance, a spiritual realm that is wholly devoid of matter. For example, it seems logically possible for certain kinds of social institutions and social contracts to have instances in such a spiritual realm. Physicalists insist only that in reality, that is, in the world as it actually is (i.e., in the actual world), any instance of any kind of thing, including any kind of social institution or social contract, is at some level of composition, wholly composed of physical entities (Hellman and Thompson 1975).

Given that being wholly composed of physical entities suffices for being physical, physicalism implies token physicalism for objects, the thesis that every object that is physical. Physicalism is not, however, equivalent to this thesis. One reason is that even if every object is physical, it might be an accident that this is so, for there might be laws of nature according to which, were certain physical conditions to obtain, wholly disembodied spirits would come into existence. That is, the laws of nature might be such that a certain brew of physical ingredients would bring into existence wholly disembodied spirits, even though, by chance, the physical ingredients in question have never come together into that brew (Witmer 1997). Since physicalists deny that there are such laws it is too weak to formulate physicalism as token physicalism for objects. Another reason it is too weak is that token physicalism for objects is compatible with there being properties whose actual patterns of distribution over objects fail to be determined by physical laws and the pattern of distribution of physical properties, and,

according to physicalists, the pattern of distribution of every property is so determined.

It would be too strong, however, to formulate physicalism in a way that implies that all properties are physical properties, for physicalists typically allow that there are functional properties. A functional property is a second-order property: it is the property of having some property or other that occupies a certain lawful or counterfactual role (Putnam 1967, Block 1994). Functional properties are realized by the first-order properties that occupy the roles in question. (A functional property may be multiply realizable in that there may be more than one first-order property that can occupy the relevant role.) Many functional properties are such that it is logically possible for them to be realized by nonphysical properties, which is why they themselves do not count as physical properties. Physicalists claim, however, that physical laws and physical conditions determine what functional properties there actually are, and that, in actuality, functional properties are ultimately realizable only by physical properties. And they claim further that physical laws and physical states and events determine the occurrence of every state or event that is an instance of a functional property. Since nonphysical properties and kinds may be cited in laws of the special sciences, it would also be too strong to formulate physicalism in a way that implies that all laws are physical laws. Physicalists maintain only that all fundamental laws are physical laws, that nonphysical laws are determined by physical laws and physical conditions.

Still, physicalism is first and foremost a thesis about what there actually is. It is the thesis that all there actually is is whatever there has to be given the way the actual world is in every physical respect—in respect of what physical objects there are, what physical properties they have, what physical relations they bear to other physical objects, what physical laws govern their behavior, etc. Physicalism can allow that there are actual nonphysical kinds of objects, nonphysical properties, relations, laws, etc., provided that they must exist, given the way the actual world is physically. What physicalism disallows is that the actual world contains anything that is unnecessary for it to be exactly as it is physically.

Call any possible world (i.e., any way the world might be) that is exactly like the actual world in every physical respect, a physical duplicate of the actual world. It is, of course, trivial that the actual world is a physical duplicate of itself. Call any possible world that is a physical duplicate of the actual world and that contains nothing other than what is required to be a physical duplicate of the actual world, a minimal physical duplicate of it. It is by no means trivial that the actual world is a minimal physical duplicate of itself. For this claim has, for instance, the consequence that there are actually people, shopping malls, cities, and countries only if there have to be people, shopping malls, cities, and countries in order for the actual

world to be a physical duplicate of itself. Physicalism is equivalent to the controversial thesis that the actual world is a minimal physical duplicate of itself.

Call any possible world that is exactly like the actual world in every respect whatsoever, physical or otherwise, a duplicate simpliciter of it. A possible world could be a physical duplicate of the actual world without being a duplicate simpliciter of it. For there is nothing about the way the actual world is in physical respects that excludes there being, say, a realm of disembodied spirits, provided that this spiritual realm exerts no causal influence on the physical. Physicalism, however, disallows that there are such ‘extra entities’ in the actual world. Physicalism is equivalent to the following thesis: any minimal physical duplicate of the actual world is a duplicate simpliciter of the actual world (Jackson 1998). That is to say, all that there actually is is whatever there has to be given the way the actual world is physically if and only if any minimal physical duplicate of the actual world is a duplicate simpliciter of it.

The idea that any minimal physical duplicate of the actual world is a duplicate simpliciter of it can be vividly illustrated by a creation myth. Suppose that Mother Nature wanted to create an exact duplicate of the actual world. According to physicalism, then, all that She would have to do is create an exact physical duplicate of the actual world—a world exactly like it in every physical respect. Her work would then be done. She would have thereby created a world exactly like the actual world in every respect whatsoever.

1.2 The Term ‘Physical’

What is meant by ‘physical’ in the phrase ‘the way the actual world is in every physical respect’? The relevant notion of physicality is tied to physics. There is, however, a dilemma that any attempt to state the tie must face (Hempel 1969, Crane and Mellor 1990). The dilemma can be formulated thus: if by ‘the way the actual world is in every physical respect’ is meant the way current physics says the world is (in respect of what objects there are, what properties they have, etc.), then physicalism is no doubt false, for current physics is no doubt false. On the other hand, if by ‘the way the actual world is in every physical respect’ is meant the way the true, complete physics would say the world is, then physicalism is either trivially true or too obscure to be a subject of rational assessment.

While current physics is a vast improvement over any previous physics, it is, nonetheless, no doubt false. The first horn of the dilemma should thus be avoided. What, then, of the second horn? If the second horn were embraced, then it would be trivially true that the way the world is in every physical respect is the way the true, completed physics would say that it is, for the claim would be true by definition. It would not follow, however, that physicalism is true by definition,

or even that it is trivially true. For were 'the way the actual world is in every physical respect' to be understood in this way, then physicalism would be the thesis that all that there actually is is whatever there has to be for the actual world to be the way the true, complete physics says that it is. That thesis is by no means trivial. The true, complete physics would not, for instance, postulate a spiritual realm that is causally isolated from the realm governed by its laws. (Even if, contra physicalism, the true, complete theory of the actual world would postulate such realm, the part that counts as physics would not.) Yet it is by no means trivially true that there is no such realm.

Still physicalists do not maintain that 'the way the actual world is in every physical respect' means the way the true, complete physics would say the world is. For the true, complete physics might postulate phenomena that physicalists deny actually exist. Mechanics, the theory of motion, is a central branch of physics. A true, complete physics would thus include a true, complete mechanics. A true, complete mechanics would take into account all of the fundamental causes of motion. Suppose, then, that the actual world contains Cartesian minds, substances that are not extended in space, but that occupy intrinsic, conscious states that exert a fundamental causal influence on the motions of particles in the brains of human beings, and, thereby, a fundamental causal influence on the behavior of human beings. If so, then the true, complete physics would be a Cartesian-like physics in that it would postulate Cartesian minds. But were this the case, physicalist would regard physicalism as false. It may well seem preposterous that the true, complete physics is a Cartesian-like physics. Current physics is surely not that badly wrong. The point, however, is that it is not true by definition that the true, complete physics is not a Cartesian-like physics.

Physicalists indeed claim that the way the actual world is in every physical respect is the way the true, complete physics would say that it is. But this is not intended to be true by definition. Physicalists make the claim because they are betting that the true, complete physics will share certain features with current physics (Pettit 1993). Call any theory of physics that is an improvement over current physics and that retains the features in question (whatever they are), a successor physics to current physics. The physicalist is betting that the true, complete physics will be a successor physics to current physics. By 'the way the actual world is in every physical respect' physicalists mean the way that the true, complete successor physics to current physics says the world is. And thus physicalism is the thesis that all there actually is is whatever that has to be given the way the true, complete successor physics to current physics says that the actual world is. To steer a course between the horns of the dilemma, physicalists must identify the features that a theory of physics must share with current physics to count as a successor to it.

The place to look to discern the features in question is at the advances in physics and the special sciences of chemistry and biology that physicalists maintain lend support to physicalism. At the start of the twentieth century, not all chemical theorists accepted atomism, and even some atomists speculated that the explanation of chemical bonding might have to invoke fundamental chemical forces. By the middle of the first decade, however, Einstein's work on Brownian motion convinced the entire community of chemical theorists of the truth of atomism. Then, after two decades of work on the structure of the atom, quantum mechanics was proposed. Quantum mechanics provides a reductive explanation of chemical bonding that invokes no forces or energies other than those at work below the level of composition of the atom. For a time after the advent of quantum mechanics, there remained speculation that there may be a fundamental, emergent force at work in living organism, a vital *élan*; but with the advances of molecular biology, such as the discovery of DNA, this idea was abandoned. On the evidence, living things are wholly composed of subatomic entities (such as electrons, quarks, etc.), entities whose behavior is determined by forces and energies exerted below the level of the atom. Biological processes are wholly implemented by chemical processes, which, in turn, are wholly implemented by quantum mechanical ones (McLaughlin 1999a).

Physicalists are not committed to the truth of the whole of current physics, or even to the truth of quantum mechanics. They are, however, betting that the true, complete physics will share the following features with current physics: (a) every entity postulated by the theory will either be subatomic or, at some level of composition, wholly made up of subatomic entities; and (b) the theory will postulate no fundamental forces or energies that are not exerted below the level of the atom. Many physicalists hold that to count as a successor to current physics a physics must retain features (a) and (b). These physicalists allow for the discovery of new objects, properties, forces, fields, and laws, provided that these discoveries are consistent with (a) and (b). They maintain only that the true, complete physics will retain features (a) and (b).

Some physicalists, however, hold a more relaxed view of what features a theory of physics would have to have to be a successor to current physics. Some allow that a fundamentally new kind of macroscopic matter could be discovered that is not composed of subatomic entities, so long as all living things and everything with mental capacities is wholly composed of subatomic entities. Moreover, some physicalists allow for the discovery of emergent forces at work in living things and beings with mental capacities, so long as the forces in question are generated by spatio-temporal configurations of subatomic entities. It is logically possible that when atoms becomes so configured as to compose a cell or so configured as to compose a human brain in a certain of neural state, a

fundamental force comes into play that is not exerted by matter at lower levels of complexity. Such emergent forces would exert a fundamental 'downward' influence on the behavior of particles (Broad 1925, McLaughlin 1999a). They are ruled out by condition (b). But some physicalists allow that a successor physics to current physics could countenance such emergent forces since they would be generated by spatio-temporal configurations of atoms, and so not, for instance, by states of an entelechy or by ontologically fundamental acts of will.

Stronger and weaker notions of what counts as a successor physics to current physics can thus be distinguished, and with them stronger and weaker versions of physicalism. But even the weakest notion(s) of what counts as a successor physics to current physics would, for instance, exclude a Cartesian-like physics from so counting. The notions of a successor physics to current physics that are invoked to characterize versions of physicalism are vague since it is, for instance, somewhat vague what counts as physics. But they are not so obscure as to render any of the varieties of physicalism outside the realm of rational assessment. (The differences between the stronger and weaker versions of physicalism will not matter for what remains, and so will be ignored.)

Physicalism implies that there is a true, complete successor physics to current physics, but it is not equivalent to that claim. Nonphysicalists can accept that there is a true, complete successor physics to current physics, in even the strongest sense. Indeed, nonphysicalists can even allow that current physics may be true. For physicalism says more than that the actual world is the way that the true, complete successor physics to current physics says that it is. Physicalism is the thesis that all there actually is is whatever there has to be given the way that the true, complete successor physics to current physics says that the world is. It is this claim that nonphysicalists deny.

2. *Whether Physicalism or One of Its Alternatives*

2.1 *The Causal Priority of the Physical*

Physicalism is defended both by appeal to empirical evidence and by appeal to supraempirical considerations of overall coherence and theoretical simplicity.

Physicalists claim that physical determination is true: the objective probability of any physical event or state is wholly determined by prior physical events and states. Thus, the objective probability of the physical movements of any living organism or sentient creature, and any movement of any of its physical parts, will be wholly determined by prior physical events (perhaps those occurring on some cross-section of the backward light-cone of the physical movements in question). Thus, given physical determination, if there are actual

nonphysical states or events, then either (a) they do not causally interact with physical occurrences at all, or (b) they causally interact with physical occurrences, but not by means of any underlying causal chain of physical occurrences, or else (c) they causally interact with physical occurrences derivatively, that is, by means of some underlying chain of physical occurrences (McLaughlin 1994).

Physicalists appeal to supraempirical considerations of overall coherence and theoretical simplicity to rule out alternatives (a) and (b). In response to (a), they maintain that there is no need to postulate states or events that have no physical effects in order to explain anything that requires explanation. In response to (b), they maintain that since physical determination is true, this possibility would involve a bizarre kind of autonomous causal overdetermination of the physical by the nonphysical that can be ruled out on grounds of theoretical simplicity, and not at the expense of overall coherence (Witmer 1997). Finally, in response to (c), some physicalists argue that only physical events causally interact with physical events. However, some physicalists allow that there are nonphysical events that causally interact with physical events in a derivative way, by means of underlying causal chains of physical occurrences. For they maintain that the nature of the underlying relationship is such that in any minimal physical duplicate of the actual world, the physical occurrences in that world will underlie exactly the same sorts of nonphysical occurrences that their physical counterparts underlie in the actual world, and, so, that (c) is compatible with physicalism.

Few philosophers now challenge physical causal determination, most accept the thesis on the evidence. To answer supraempirical considerations for physicalism, nonphysicalists typically argue that there are reasons to accept that there are certain actual phenomena not countenanced by physicalism, and that those reasons outweigh the consideration that the phenomena would be either epiphenomena or autonomous overdeterminers of certain physical occurrences.

2.2 *Elimination or Entailment?*

According to physicalists, all that there actually is is whatever there has to be given the way the actual world is physically. When it is claimed that a certain phenomenon actually exists that is not required for the actual world to be exactly as it is physically, physicalists have two options. They can argue that there is no phenomenon of the sort in question, and thus be eliminativists or irrealists about it. (All physicalists are eliminativists about disembodied spirits, for instance.) Or they can argue that, despite first appearances, the phenomenon in question must exist given the way the actual world is physically.

The debate between physicalists and nonphysicalists is mainly centered on the mental and on phenomena

logically dependent on the mental. Some physicalists are eliminativists about the mental, or at least about the mental as conceived by folk or commonsense psychology (Churchland 1978, Stich 1999). But most physicalists are realists about the mental, and many are realists about the moral also, and attempt to explain moral properties reductively, in part, by appeal to mental properties.

Physicalists argue that a phenomenon must exist given the way the actual world is physically by arguing that it is plausible that its existence is entailed by the totality of physical facts (Jackson 1998). The totality of physical facts includes all particular physical facts and all general ones, including all nomological ones, and includes as well a 'that's all' physical fact, a fact to the effect that the physical facts in question are all the physical facts.

Given physicalism, the totality of physical facts, in conjunction with the contingent thesis of physicalism, entails all of the facts. Thus, if, for example, it is a fact that a certain property has a certain worldwide pattern of distribution, then the totality of physical facts in conjunction with the thesis of physicalism will imply that the property has that worldwide pattern of distribution. The totality of physical facts will, of course, imply facts about the actual worldwide pattern of distribution of physical properties. The totality of physical facts, in conjunction with the assumption of physicalism, will imply facts about the actual worldwide pattern of distribution of functional properties and any properties metaphysically necessitated by physical properties, and will imply as well that no other sorts of properties are actually possessed by anything.

One way to establish that mental facts are entailed by the totality of physical facts and the truth of physicalism alone is by showing that the relevant mental concepts are definable in terms of physical and topic-neutral concepts (Smart 1959). (Topic-neutral concepts are neither physical nor mental concepts; they include the concept of a part, the concept of causation, and logical concepts such as the concept of conjunction, the concept of negation, and the concept of universal quantification.) Both analytical behaviorism and analytical functionalism entail that there are such definitions. According to analytical behaviorism, mental concepts can be defined in terms of concepts of dispositions to peripheral behavior (Byrne 1994). According to analytical functionalism, mental concepts can be defined more broadly in terms of macro-physical and topic-neutral concepts (Armstrong 1968, Block 1994, Lycan 1994).

Analytical behaviorism is widely regarded as a failed project. It is now generally acknowledged that no mental concept can be defined in terms of concepts of dispositions to peripheral behavior. Analytical functionalism continues to have proponents (Lewis 1994), but it is controversial whether it is true of any mental concept. Analytical functionalism is most plausible for

intentional concepts such as the concepts of belief, desire, intention, and the like, since these concepts are, arguably, dispositional concepts. But it is especially controversial whether analytical functionalism is true of sensory (or phenomenal) concepts such as the concept of the felt quality of pain (Block 1995, Chalmers 1996). Many philosophers, both physicalists and nonphysicalists, claim that sensory concepts fail to have *a priori* functional analyses of any sort. They maintain that there is an explanatory gap between sensory concepts and physical and topic neutral concepts in that sensory concepts are neither definable in terms of such concepts nor are their references fixed by associated physical and topic neutral concepts (Levine 1983).

Scientific functionalism and type materialism are physicalist alternatives to these *a priori* approaches to explaining how mental facts are entailed by the totality of physical facts and the assumption of physicalism. To focus for illustration on the case of pain, scientific functionalism maintains that pain is identical with a type of functional state (and, so, a type of second-order state), albeit the identity is knowable only *a posteriori* (Loar 1997). Type materialism maintains that pain is identical with a type of abstract neurobiological state, and that the identity is knowable only *a posteriori* (Hill 1991). As Saul Kripke (1980) has demonstrated, identity statements in which the identity sign is flanked by rigid designators (terms that have the same reference in any world in which they refer) are either necessarily true or necessarily false, even if only *a posteriori*. 'Pain' is a rigid designator; it is the name of a certain kind of sensation. Statements asserting that pain is identical with a certain type of functional state or ones asserting that pain is identical with a certain type of neurobiological state will be necessarily true, if true at all. Given the necessity of identity, if pain is identical with either a type of functional state or a type of macrophysical state, then facts about the pattern of realization of the functional state or macrophysical state in question will imply facts about the pattern of realization of pain. Thus if, for instance, pain = neurobiological state *N*, then facts about the worldwide distribution of *N*-states will entail facts about the worldwide distribution of pain (and conversely). According to type materialists, the entailments will be only *a posteriori*. Scientific functionalists make an analogous claim of *a posteriori* entailment.

Some philosophers argue that physicalism must be committed to the *a priori* entailment of all the facts (save indexical ones such as, e.g., I = Brian McLaughlin) from the totality of physical facts and the assumption of physicalism, on pain of being committed to an unacceptable view of necessity and/or an unacceptable view of conceptual reference (Chalmers 1996, Jackson 1998). Also, some of these philosophers argue that since facts about sensory consciousness will not be *a priori* entailed by the

totality of physical facts and the assumption of physicalism, physicalism is false. Scientific functionalists and type materialists agree that facts about sensory consciousness are not so *a priori* entailed, but they reject the claim that the entailments must be *a priori*. They argue that while sensory concepts are not *a priori* linked to physical or functional concepts, and so do not admit of *a priori* functional analyses of any sort, sensory concepts nonetheless apply solely in virtue of the physical or functional properties to which they refer. Suffice it to note that this dispute turns on unresolved issues about the nature of necessity and the nature of conceptual reference.

See also: Causation: Physical, Mental, and Social; Physical Sciences: History and Sociology

Bibliography

- Armstrong D M 1968 *A Materialist Theory of Mind*. Routledge and Kegan Paul, London
- Block N 1994 Functionalism (2). In: Guttenplan S (ed.) *A Companion to the Philosophy of Mind*. Blackwell Reference, Oxford, UK, pp. 323–32
- Block N 1995 'Qualia'-based objections to functionalism. In: Lycan W G (ed.) 1995 *Mind and Cognition: an Anthology*. Blackwell, Malden, MA pp. 444–68
- Broad C D 1925 *The Mind and Its Place in Nature*. Harcourt Brace, New York
- Byrne A 1994 Behaviorism. In: Guttenplan S (ed.) *A Companion to the Philosophy of Mind*. Blackwell Reference, Oxford, UK, pp. 132–40
- Chalmers D 1996 *The Conscious Mind: in Search of a Fundamental Theory*. Oxford University Press, New York
- Churchland P M 1978 Eliminative materialism and the propositional attitudes. *The Journal of Philosophy* 78: 67–90. Reprinted in Lycan W (ed.) 1995 *Mind and Cognition*. Blackwell, Oxford, UK, pp. 206–23
- Crane T, Mellor D H 1990 There is no question of physicalism. *Mind* 99: 185–200
- Hempel C 1969 Reduction: ontological and linguistics facets. In: Morgenbesser, Suppes P, White M (eds.) *Philosophy, Science and Method*. St. Martin, New York, pp. 179–99
- Hellman G P, Thompson F W 1975 Physicalism: ontology, determination, and reduction. *Journal of Philosophy* 72: 551–64
- Hill C S 1991 *Sensations: A Defense of Type Materialism*. Cambridge University Press, Cambridge, UK
- Horgan T 1994 Physicalism (1). In: Guttenplan S (ed.) *A Companion to the Philosophy of Mind*. Blackwell Reference, Oxford, UK, pp. 471–9
- Jackson F 1982 Epiphenomenal qualia. *Philosophical Quarterly* 32: 127–36. Reprinted in Lycan W (ed.) 1995 *Mind and Cognition*. Blackwell, Oxford, UK, pp. 69–77
- Jackson F 1998 *From Metaphysics to Ethics: a Defense of Conceptual Analysis*. Clarendon Press, Oxford, UK
- Kim J 1999 Physicalism. In: Wilson R A, Keil F C (eds.) *The MIT Encyclopedia of the Cognitive Sciences*. MIT Press, Cambridge, MA, pp. 645–7
- Kripke S 1980 *Naming and Necessity*. Harvard University Press, Cambridge, MA
- Levine J 1983 Materialism and qualia: the explanatory gap. *Pacific Philosophical Quarterly* 64: 354–61
- Lewis D 1994 Reduction of mind. In: Guttenplan S (ed.) *A Companion to the Philosophy of Mind*. Blackwell Reference, Oxford, UK, pp. 412–31
- Loar B 1997 Phenomenal properties. In: Block N, Flanagan O, Guzeldere G (eds.) *The Nature of Consciousness: Philosophical Debates*. MIT Press, Cambridge, MA, pp. 597–616
- Lycan W 1994 Functionalism (1). In: Guttenplan S (ed.) *A Companion to the Philosophy of Mind*. Blackwell Reference, Oxford, UK, pp. 317–23
- Lycan W G (ed.) 1995 *Mind and Cognition: An Anthology*. Blackwell, Malden, MA
- McLaughlin B P 1994 Epiphenomenalism. In: Guttenplan S (ed.) *A Companion to the Philosophy of Mind*. Blackwell Reference, Oxford, UK, pp. 277–88
- McLaughlin B P 1999a Emergence. In: Wilson R A, Keil F C (eds.) *The MIT Encyclopedia of the Cognitive Sciences*. MIT Press, Cambridge, MA, pp. 267–9
- McLaughlin B P 1999b The philosophy of mind. In: Audi R (ed.) *The Cambridge Dictionary of Philosophy*. Cambridge University Press, Cambridge, UK, pp. 684–94
- Melnyk A 1997 How to keep the 'physical' in physicalism. *Journal of Philosophy* 94: 622–63
- Papineau D 1993 *Philosophical Naturalism*. Blackwell, Oxford, UK
- Pettit P 1993 A definition of physicalism. *Analysis* 53: 213–23
- Poland J S 1994 *Physicalism: The Philosophical Foundations*. Oxford University Press, Oxford, UK
- Putnam H 1967 Psychological predicates. In: Capitan W H, Merrill D D (eds.) *Art, Mind, and Religion: Proceedings*. University of Pittsburgh Press, Pittsburgh, PA, pp. 37–48
- Rosenthal D M (ed.) 1991 *The Nature of Mind*. Oxford University Press, Oxford, UK
- Smart J J C 1959 Sensations and brain processes. *Philosophical Review* 68: 141–6
- Stich S 1999 Eliminative materialism. In: Wilson R A, Keil F C (eds.) *The MIT Encyclopedia of the Cognitive Sciences*. MIT Press, Cambridge, MA, pp. 265–7
- Witmer G 1997 *Demanding Physicalism*. Ph.D. dissertation, Rutgers University, New Brunswick, NJ

B. P. McLaughlin

Piaget, Jean (1896–1980)

A most productive scholar in the area of children's intellectual development, Jean Piaget was primarily interested in epistemology or theory of knowledge. His genetic psychology (psychology of the genesis of thought) was intended to provide a scientific basis for epistemology, with the main concern to explain how knowledge can become more rigorous and more objective. Piaget's psychological explanations were inspired by several scientific disciplines and beside his studies in psychology and epistemology he published in other fields such as biology, logic, and sociology. His work derived from evolutionist approaches rooted in the nineteenth century yet it is still in line with scientific trends at the end of the twentieth century.

The core of Piaget's theory is the conception of mind as an active system which, like an organism,

assimilates external data into its structures and can adjust to these data. Special emphasis, in the knowledge process, is placed on the subject's activity and the structures through which he or she perceives and understands things. Thus, the development of knowledge is viewed as the construction of new structures (rather than the accumulation of new information) triggered and guided by internal processes during the subject's encounter with reality. Each new stage in this process opens up new possibilities. In this interactionist conception, structures are neither predetermined nor found in the environment and there are striking similarities between biological evolution, cognitive development in children, and the development of scientific thought.

1. Seven Decades of Scientific Productivity

1.1 Childhood and Adolescence (1896–1918)

Jean Piaget was born in Neuchâtel, a small city of Switzerland, on August 9, 1896. He was the son of an historian and professor of medieval literature at the university. Of his mother, he wrote in his autobiography that she was an intelligent and energetic woman, whose psychological imbalance created problems in the family. At the age of 11 years, Jean Piaget started classifying and collecting mollusks with the help of the director of the natural history museum of his hometown. In his teens, he was already an acknowledged specialist in the field and started publishing papers in scientific journals (as of 1911). This activity led him to decide to study natural sciences at the university between 1915 and 1918. At the same time, as of the age of 16 he became passionately interested in philosophy, and particularly in the theory of knowledge or epistemology. He soon decided to devote his life to the elaboration of a biological theory of knowledge, whose goal would be to explain the evolution of ideas both in individuals and in the scientific community. This project was influenced by evolutionary approaches in the field of biology (Lamarck, Darwin) and in philosophical works on issues of intelligence (Spencer, Le Dantec, Bergson). In the following years, Piaget's professor of philosophy A. Reymond convinced him of the importance of a rationalist approach making use of logic and mathematics.

In 1918 Piaget published a philosophical and autobiographical novel (*Recherche*) describing the crisis he had gone through because of the conflict he felt between religion and science, or, in other words, between value and rationality. Eventually, the search for truth, which is the goal of science, became the most important value for Piaget. *Recherche* also contained the outline of a theory of the evolution of living and cognitive organizations based on the idea of improving

forms of equilibrium between the whole and its parts. Piaget's twofold interest on the one hand in biology and the associated problems of the organization and evolution of living beings, and on the other hand in epistemology and related issues concerning the validity of knowledge and its development, forms the basis of all his work.

1.2 Complementary Studies (1918–1921)

Piaget was convinced that a theory of knowledge should be validated by psychological research, and decided to get training in scientific psychology as soon as he graduated from university with his Ph.D. thesis on mollusks in 1918. He then spent a term in Zurich, attending courses in psychopathology—where he encountered psychoanalysis—and in experimental psychology. He then studied in Paris from 1919 to 1921, where he attended lectures in psychology (specifically by Janet), philosophy (with Lalande), and history of science (by Brunschvicg who had great influence on Piaget's ideas). Courses in psychopathology provided him with a clinical interview method, which he soon used in his discussions with children. In effect, in parallel with his studies, he worked in the laboratory of Binet and Simon, standardizing intelligence tests for children. Piaget was fascinated by the reasoning of the children who did not yet answer 'correctly,' and he discovered that the study of children's thinking could be a means to validate a theory of knowledge. This program, which was to be the main focus of his work for 60 years, was very similar to the program designed by James Mark Baldwin (1861–1934), from whom Piaget borrowed several concepts and the expression 'genetic psychology.'

1.3 The Early Psychological Research (1921–1929)

In 1921, Claparède offered a position in his Institute for educational sciences (known as the Jean-Jacques Rousseau Institute) in Geneva to Piaget, who was to conduct most of his professional life in this institution. What he found there were stimulating intellectual atmosphere and students willing to assist him in his work. Together with them the young researcher started a series of studies of children's reasoning and soon published his first books: *The Language and Thought of the Child* (1923); *Judgment and Reasoning* (1924); *The Child's Conception of the World* (1926); *The Child's Conception of Physical Causality* (1927) (the titles given here and in the remainder of this text are the English translations of the French originals; the years are those of the original publications). The striking facts he discovered about the use of language, the non-differentiation of concepts and the explanations of physical phenomena in children, as well as the detailed

analyses and the synthetic categorization of these behaviors, brought Piaget almost immediate international fame. These studies showed that young children's thinking is structurally and functionally different from that of older children and adults. Young children lack a true logic of classes and relations, they have difficulties viewing reality from an objective viewpoint, and they lack moral norms, which command compliance on the basis of a feeling of necessity. Children's thinking tends to be egocentric, that is, they have difficulties in looking at things in perspective. In these early works, the passage from egocentric thought to a logical and more objective form of thinking was explained by a social factor: cooperation, or exchange between individuals who consider themselves as equals.

During this period, Piaget was involved in many activities beside his research. He taught child psychology at the J.-J. Rousseau Institute, philosophy and history of science at the University of Neuchâtel (1925–1929), was for a few months psychoanalyzed by Sabina Spielrein, a pupil of Freud, and continued to study the evolution of mollusks. In 1923 he married a student at the Geneva Institute. Piaget and his wife Valentine spent much time minutely observing their three children who were born between 1925 and 1931. These observations were the basis of several books published in the following years.

1.4 Administrative Responsibilities, Professorship and New Directions of Research (the 1930s)

In 1929, Piaget was appointed professor of history of science at the University of Geneva, where he subsequently became professor of sociology (1939) and of experimental psychology (after the death of Claparède, 1940). He also played an important administrative role at the J.-J. Rousseau Institute as associate director (1929), then as co-director (since 1932), reorganizing the Institute and establishing links between the Institute and the University. At the same time, and until the 1960s, he was director of the *Bureau International de l'éducation* and managed to convince the governments of many countries to join that organization.

Piaget's most important activities continued to be research and writing. During these years, he accumulated data on children's reasoning. However, apart from his book on moral judgment (1932)—simultaneously in French and in English—which was the sequel to the psychology books published in the 1920s, his major publications dealt with infant cognitive development. The choice of this topic stemmed from Piaget's conviction that logic has its roots in actions and also from his deep interest in the issue of the origins of knowledge. *The Origins of Intelligence* (1936) and *The Child's Construction of Reality* (1937) con-

tained extremely original observations and explanations of infant behaviors, and departed radically from Piaget's earlier writings as far as the explanatory model was concerned. Piaget discovered behaviors that had previously been overlooked, for example, the failure of babies younger than 9 months of age to look for an object hidden in front of them. Replications of these observations and experiments many years later around the world have confirmed these findings that had resulted from the observation of his three children.

In view of theory building, Piaget analyzed the elements of knowledge, the 'schemes' of actions, which become more differentiated and coordinated as development progresses. They evolve into increasingly complex and better-adapted forms of behavior, from automatic 'reflexes' to intelligent activities. This development is a special case of the relationship between the organism and the environment, in other words of biological adaptation. It results from the interaction between the subject of knowledge, who assimilates reality to his or her structures, and the object of knowledge, which may elicit an accommodation of these structures. Cognitive development is viewed as an evolving equilibrium between assimilation and accommodation. In *The Child's Construction of Reality* (1937), Piaget studied the growth of the practical knowledge of object permanence, space, time, and causality in infants. He introduced a new and abstract explanatory notion: the mathematical concept of 'group,' which several authors in the social sciences considered as an interesting model at the time. Piaget applied it to the organization of spatial knowledge in infants and was convinced that it could be used to explain the logic of school-age children. To this end, from 1937 onwards, he worked for three years on this model, expressing it in algebraic formulae.

Although Piaget was now a famous author who had received an honorary degree from Harvard University in 1936, his two books on infant cognition were not translated in English until the 1950s. They contained too many abstract considerations and microanalyses and were definitely ahead of their time. A third book, also based on the observation of Piaget's own children, was written during this period and published later: *Play, Dream and Imitation* (1945). It dealt with imitation in infants and the origins of symbolic thought in young children. Intellectual activity as well as play and imitation were defined in terms of equilibrium or disequilibrium between assimilation and accommodation.

1.5 The Mature Years: The Advent of Structural Analysis (1940s)

During this and the following period, Jean Piaget was prodigiously productive, thanks to a strict and very

regular work regime. In the mornings he would teach, discuss research with his large team of co-workers, and perform his administrative duties. Piaget would then spend the rest of the day mostly by himself. In the early afternoon, while thinking about the topics of his current writing, he would stroll or ride a bicycle in the countryside near Geneva. During the rest of the afternoon he would write, which was the best way for him to elaborate his ideas. Even when traveling or attending conferences he would not let a day go by without writing a few pages. Although he loved classical music he almost never went out in the evenings to movies, theaters or concerts. Both in summer and in winter, his love of nature led him to repeated sojourns in the mountains of the Valais, where he both hiked and worked.

The logical model of ‘grouping’ of mental operations, which he had recently developed, and which accounted for the organization of inferences in overall structures, inspired much research and writing. It permitted Piaget not only to describe the limitations of young children’s reasoning as he had done in his first books, but also to explain the sudden, stable, and rather generalized intellectual progress due to the advent of the elementary logic, of ‘concrete operations’. Between 1941 and 1948, his work resulted in six books: *The Child’s Construction of Quantities* (with B. Inhelder, 1941), *The Child’s Conception of Number* (with A. Szeminska, 1941), *The Child’s Conception of Time* (1946), *The Child’s Conception of Movement and Speed* (1946), *The Child’s Conception of Geometry* (with B. Inhelder and A. Szeminska, 1948), *The Child’s Conception of Space* (with B. Inhelder, 1948). These books study the elementary mastery of basic scientific concepts. This mastery, which usually emerges between the ages of 6 and 8 or 9 years, was explained in terms of the underlying ‘grouping’ structures that organize mental operations. This type of structure entails the reversibility of thinking, for Piaget an essential feature of natural logic. In contrast, younger children think that the number of a set of objects or the quantity of a substance change when spatial transformations occur, and that the duration of a movement corresponds to the distance covered, without taking into account the relative speed of the moving objects. These young children have difficulties comparing the number of elements in a subclass with the number in the superordinate class. As to space, geometrical capacities like inverting a spatial order, constructing a straight line with vertical sticks using a sighting method, and drawing the horizontal line of the water level in a tilted glass also usually emerge after the age of 7 years. The basic concept studied in each book was analyzed and defined in the light of children’s answers. For example, number was analyzed in terms of a synthesis between logical classification (class inclusion) and the seriation of asymmetrical relations (ordering by increasing or decreasing size); time was defined as the coordination of movements,

in which the ability to consider speed played a fundamental role, and so forth.

Piaget also published a *Treaty of Logic* (1949) and with his co-workers he studied perception in children in the laboratory of experimental psychology. These studies, which attracted Piaget’s attention to the phenomenon of regulation, were to provide the subject matter of a much later book. The same holds true for the papers in sociology which were published during this decade and collected in a book called *Sociological Studies* in 1965.

The second World War rendered contact with foreign colleagues difficult, but Piaget was invited in 1942 to give a series of lectures in Paris at the ‘*Collège de France*’ (published in *The Psychology of Intelligence*, 1946). The abstract nature of his explanations struck many listeners as much as the brilliant analyses inspired by various scientific disciplines. After the war, UNESCO entrusted Piaget with several missions.

1.6 Achieving the Ultimate Goal of his Work: The Elaboration of a Genetic Epistemology (1950s)

In the 1950s, while continuing his work in genetic psychology with the help of his co-workers, Piaget devoted much of his effort to the foundation of a genetic epistemology, i.e., the study of the basic concepts and processes of the development of scientific knowledge. This used to be a topic of philosophical study but now became an object of scientific analysis. Actually, Piaget had indeed pursued epistemological issues throughout his entire previous work while ostentatiously focusing on the explanation of children’s thinking. Now he wished to discuss questions of knowledge *per se* as they were raised in the sciences (but continued to use the study of children’s cognitive development for this purpose). In 1950, Piaget published three volumes of an *Introduction to Genetic Epistemology*. The book begins with a general introduction defining genetic epistemology and its method and stating the main issues. Piaget develops an idea he had already expressed in earlier writings, that of a ‘circle of sciences’: physical reality is explained by mathematics, mathematics is the product of mind, which is a result of biological evolution, and organisms are closely connected with physical reality. The book deals in great detail with the epistemology of mathematics and physics, and in lesser detail with the epistemology of biology and the social sciences.

To make progress in the domain of genetic epistemology, Piaget founded the International Center for Genetic Epistemology, which was a working group comprising researchers in psychology and scientists from various disciplines (mathematics, physics, logic, etc.). The Center was launched in 1955 after obtaining, with difficulty, the necessary funding. A new series of books, which contained the results of the studies

conducted at the Center for genetic epistemology, appeared from 1957 onwards. About 40 volumes were published in the series until the death of Piaget in 1980.

During the 1950s, when Piaget held a position as professor at the University of Paris (1952–1963), his psychological research in Geneva continued to progress and resulted in three books with B. Inhelder: *The Development of Elementary Logical Structures* (1959), a sequel to the studies of concrete operations; *The Origin of the Idea of Chance* (1951) and *The Growth of Logical Thinking from Childhood to Adolescence* (1955). The latter book provided a new description and definition of the highest stage of logical development: the formal operational stage achieved between the ages of 12 and 15 years. The formal model of the INRC group, derived from Klein's mathematical group, provided an account of the new possibilities of reasoning which were available to adolescents. It involved a greater number of combinations of operations than the grouping structures (in particular a 'double reversibility' of thought). Moreover, the experiments by Inhelder showed how the understanding of physical phenomena developed in children and how, through the use of an adequate interview procedure, adolescents could rediscover fundamental scientific laws on their own.

1.7 The Last Two Decades (1960 To 1980): The Pinnacle of Fame

Only in the late 1950s was Piaget's work rediscovered by English and American psychologists, although he already had been a most respected professor and researcher in the French-speaking countries, and a well-known author in many other countries. During the following two decades he was among the most cited psychologists and best known scientists in the world. Most of his books were translated from French into foreign languages (a total of 24 different languages). From the end of the 1950s onwards, he received an increasing number of honorary degrees each year. When he died, he had received over 30 honorary degrees and several prestigious awards. However, these honors and the travels they involved did not prevent Piaget from continuing his work.

Until the beginning of the 1970s, studies of genetic epistemology mainly focused on problems related to logic, language, space, time, and causality. The latter topic—strangely neglected for several decades—proved difficult to study and remained on the program of the Center for Genetic Epistemology for several years in succession. On the whole, these books, while discussing issues specific to each science (e.g., the relative importance of time and speed in judgments about duration, or the possibility of geometrical intuition) validated the constructivist conception of Piaget's epistemology. Effective forms of knowledge are neither innate nor drawn from the observation of

the environment. They derive from the interaction of active individuals with their environment. A consequence of this position (summarized in the short book *Genetic Epistemology*, 1970) is that the rate of intellectual development can be only moderately accelerated by education. The goal of education, according to the few texts devoted by Piaget to this topic (for example a long article in 1965) is to favor children's intellectual activity and their discovery of relationships and abstract concepts.

During the 1960s Piaget returned to his earliest field of interest: biology. He published a paper on the evolution of mollusks and about the reproduction of the 'sedum' plants, which he grew in his garden and observed carefully. In his book *Biology and Knowledge* (1967) he developed his ideas on biology and the parallel between biological and cognitive development. Simultaneously, he published three books on psychology. They dealt with the 'figurative' aspects of knowledge, whose function is to provide a kind of copy of things: *The Mechanisms of Perception* (1961), *The Child's Mental Imagery* (1966), and *Memory and Intelligence* (1968). These books showed that figurative aspects are progressively subordinated to the subject's operations: some optical illusions decrease with perceptual activity, and with the advent of logical thinking (concrete operations), mental images can adequately represent transformations, while memory of organized material improves. Intelligence is not captured by representation, i.e., a system of images or linguistic symbols; rather intelligence consists in operating on elements of reality or symbols and thus transforming them.

In the 1970s, after Piaget's retirement as professor at the University of Geneva (1971), he continued his work with a large team of researchers at the Center for Genetic Epistemology. Most of his publications dealt with processes of cognitive development. For over 20 years, his main aim had been to demonstrate that efficient forms of knowledge result from structures that organize mental operations, and that these structures, which become increasingly complex over time, are rooted in biological organization. The time had come to deal with the processes involved in the construction of these structures, a topic that Piaget had neglected since the 1930s, except for a few epistemological papers written in the 1950s. During the last decade of his life, between the ages of 75 and 84 years, Piaget developed concepts that had been introduced earlier but needed fuller consideration and a precise definition: equilibration, reflexive abstraction, generalization, the opening up of new possibilities, and so on. These concepts are highly interesting and very original ideas, difficult to operationalize, but with deep and general explanatory value. Unfortunately, Piaget was so eager to continue his work that he changed the focus of his research and his writing nearly every year, although these complex concepts would deserve more sustained treatment.

Reflexive abstraction is the process through which something new is abstracted from experience. What is abstracted is not a property of the external world, but derives from the organization of the subject's actions or operations. Equilibration is a self-regulation process, which is specific to living organizations. At the cognitive level, it is a reaction to an intellectual perturbation (such as a contradiction) which regulates the cognitive system and, in an attempt to compensate the perturbation, creates improved forms of knowledge. For his eightieth birthday in 1976, a kind of dissertation defense was organized: Piaget defended the theses developed in *The Equilibration of Cognitive Structures* (1975) before a 'jury' composed of famous scientists who represented different disciplines. In this last period, Piaget's work also dealt with the grasp of consciousness, contradiction, and dialectics. He explored the relevance to the explanation of cognition of the mathematical concept of morphism and of the idea of logic of meaning. The last work was conducted in collaboration with R. Garcia who was also co-author of *Psychogenesis and History of Science* (published posthumously in 1983). Piaget also dealt with problems of biology. His post-Lamarckian ideas attributed some importance to aspects of experience in the evolution of species (for example in *Behavior, a Motor of Evolution*, 1976).

1.8 Factors of Scientific Creativity

Piaget died on September 16, 1980, leaving behind a life's work of unusual scope, in terms of depth, originality, interdisciplinary relevance, and volume. The explanation of his exceptional scientific productivity is not only the conjunction of a brilliant mind with a great capacity for work, but also the stimulation of reflection by the results of constantly renewed research. It was also due to Piaget's sociability, which enabled him to easily convince others to work with him; and to a strong sense of a mission to which everything else was subordinate. According to Piaget, having a theory to contradict was a powerful stimulus. The dominant empiricist conceptions in psychology—neo-associationism—and logical positivism in philosophy played this role well into the 1960s. Since the mid-1960s, Piaget was busy developing arguments against the relatively widespread idea, prevalent since Chomsky published his work, that important aspects of knowledge are innate.

2. The Importance of Piaget's Work for the Social Sciences

Piaget's theory is exceptionally comprehensive, as it gives a detailed account of cognition from birth to adolescence, and also deals with scientific knowledge

in adults. It contains an incredible wealth of experimental data covering these periods in life, especially the age range from 5 to 9 years. It deals with fundamental problems often ignored in other psychological theories, such as the relationship between biological activity and cognition, the relative role of the subject and the environment, and the connections between different aspects of cognition such as perception, reasoning, and memory. In some respects Piaget's work was ahead of its time and it could therefore easily establish connections with emerging disciplines such as cybernetics (since the mid-1940s) and, later, artificial intelligence. The concepts of self-regulated systems and of feedback specific to cybernetics were involved, less explicitly, in the Piagetian ideas of reversible structures (present in his work at least since 1937) and of the regulation of cognitive activities (frequently referred to by him in the early 1940s). The artificial intelligence programs which expressed in symbolic language the expertise involved in the mastery of a task bear some resemblance to the Piagetian concept of repeatable action schemes and recall Piaget's attempts to formalize mental structures. Moreover, the multidisciplinary approach known as cognitive science, which emerged in the 1980s, is completely in line with Piaget's approach, with its links between biology, psychology, logic, and physics.

One aspect of Piaget's work, genetic epistemology, which is focused on scientific knowledge as investigated through studies of children, has had little impact outside of Geneva. By contrast, Piaget's psychological work has definitely influenced subsequent orientations and research in the social sciences.

Since the 1920s, Piaget's findings about the reasoning of children had considerable influence on child psychology. The influence was strong in the 1960s and 1970s, in particular regarding the four successive stages of cognitive development, (sensorimotor, pre-operational, concrete operational, and formal operational). In fact, Piaget's idea was not to define general stages of child development, but necessary steps in the integration of new forms of knowledge. His detailed findings concerning the emergence of new capacities, such as the search for hidden objects in babies, conservation of quantity in school-age children, and combinatory operations in adolescents remain useful landmarks of psychological development although they cannot be used as psychometric tests. During the same two decades, various experiments in education were directly inspired by aspects of the theory, namely the role of the child's activity and the importance attributed to general structuring abilities rather than specific learning. However, mainstream psychology and education, especially in English-speaking countries, often remained impervious to Piaget's influence, except for the growth of cognitive developmental psychology.

Cognitive developmental psychology became a productive subdiscipline of psychology, especially since

the beginning of the 1970s. Practically every developmental psychologist in the world at that time had at least some knowledge of Piaget's work and the very idea of studying cognitive development using a cross-sectional method was inspired by the work previously undertaken in Geneva. Piaget's writing often played the same role of a 'theory to contradict' that associationism or nativist conceptions had played for Piaget. Many authors in developmental psychology launched their careers by conducting experiments that were designed to prove Piaget wrong. Because his constructivist conceptions grant a limited role to social influence, and in particular to education, they were unacceptable to many. Much research has been performed especially in the USA to show that cognitive competencies existed in children well before the ages indicated by Piaget. In order to demonstrate this, the problems presented to the children were modified and compared to those presented by Piaget himself. The race for early competencies in the 1980s and 1990s gave rise to the idea that many cognitive capacities were innate. In the end, Piaget's conceptions were not followed, but thanks to his work a whole field of psychology developed.

Piaget's work paved the way for the cognitivist revolution in psychology, from the mid-1960s onwards. Psychologists had become interested in mental states, a taboo for many decades due to the predominant behaviorist conception. The new generation of psychologists could read in Piaget's work detailed studies of mental states that had been conducted over a period of several decades.

At the end of the twentieth century, Piaget's theory had indirect descendants in the fields of child psychology, cognitive science, and developmental psychology. Ideas of specialists in educational science are more directly inspired by Piaget's work, especially in French-speaking countries and in Latin America. Studies in moral development also derive directly from Piaget's research in this field. More generally, constructivist approaches which attribute importance to the activity of the subject, in applied psychology as well as in other disciplines, usually refer, albeit vaguely at times, to the work of the Swiss psychologist and epistemologist.

See also: Binet, Alfred (1857–1911); Cognitive Development; Child Education; Constructivism in Cognitive Psychology; Developmental Psychology; Infant and Child Development, Theories of; Janet, Pierre (1859–1947); Kant, Immanuel (1724–1804); Kohlberg, Lawrence (1927–87); Khler, Wolfgang (1887–1967); Lifespan Theories of Cognitive Development; Piaget's Theory of Child Development; Piaget's Theory of Human Development and Education; Psychology: Historical and Cultural Perspectives; Scientific Concepts: Development in Children; Vygotskij, Lev Semenovic (1896–1934)

Bibliography

- Bringuier J-C 1989 *Conversations with Jean Piaget*. University of Chicago Press, Chicago
- Ducret J-J 1990 *Jean Piaget: biographie et parcours intellectuel*. Delachaux Nestlé, Neuchâtel, Switzerland
- Gruber H, Vonèche J (eds.) 1995 *The Essential Piaget*. Aronson, London
- Inhelder B, Piaget J 1955 *De la logique de l'enfant à la logique de l'adolescent*. P.U.F., Paris. [English translation, 1972 *The Growth of Logical Thinking From Childhood to Adolescence*. Routledge, London]
- Montangero J, Maurice-Naville D 1992 *Piaget ou l'intelligence en marche. Vocabulaire et aperçu chronologique*. Mardaga, Liège, Belgium [English translation, 1997. *Piaget or the Advance of Knowledge. Glossary and Chronological Overview*. Erlbaum, Mahwah, NJ]
- Piaget J 1926 *La représentation du monde chez l'enfant*. Alcan, Paris [English translation 1972 *The Child's Conception of the World*. Littlefield Adams, Totowa, NJ]
- Piaget J 1932 *Le jugement moral chez l'enfant*. Alcan, Paris [English translation 1977 *The Moral Judgment of the Child*. Penguin, Harmondsworth, UK]
- Piaget J 1936 *La naissance de l'intelligence chez l'enfant*. Delachaux and Nestlé, Neuchâtel, Switzerland [English translation 1970 *The Origins of Intelligence*. Routledge, London]
- Piaget J 1945 *La formation du symbole chez l'enfant*. Delachaux and Nestlé, Neuchâtel, Switzerland [English translation 1962 *Play, Dreams and Imitation*. Norton, New York]
- Piaget J 1967 *Biologie et connaissance*. Gallimard, Paris [English translation 1971 *Biology and Knowledge*. The University of Chicago Press, Chicago]
- Piaget J 1970 *L'épistémologie génétique*. P.U.F., Paris [English translation 1972 *The Principles of Genetic Epistemology*. Routledge, London]
- Piaget J 1970 reedited *Psychologie et épistémologie*. Denoël, Paris [English translation 1971 *Psychology and Epistemology*. Grossman, New York]
- Piaget J 1972 reedited *Problèmes de psychologie génétique*. Denoël, Paris [English translation 1973 *The Child and Reality: Problems of Genetic Psychology*. Grossman, New York]
- Piaget J 1972 *Où va l'éducation*. Denoël-Gonthier, Paris [English translation 1973 *To Understand is to Invent: the Future of Education*. Grossman, New York]
- Piaget J 1975 *L'équilibration des structures cognitives*. P.U.F., Paris [English translation 1985 *The Equilibration of Cognitive Structures*. The University of Chicago Press, Chicago]
- Piaget J 1981 *Le possible et le nécessaire* 2 Vols. P.U.F., Paris [English translation 1987 *Possibility and Necessity*. University of Minneapolis Press, Minneapolis, MN]
- Piaget J, Inhelder B 1941 *Le développement des quantités chez l'enfant*. Delachaux and Nestlé, Neuchâtel, Switzerland [English translation 1974 *The Child's Construction of Quantities*. Routledge, London]
- Piaget J, Inhelder B 1948 *La représentation de l'espace*. P.U.F., Paris [English translation 1967 *The Child's Conception of Space*. Norton, New York]
- Piaget J, Inhelder B 1968 *Mémoire et intelligence*. P.U.F., Paris [English translation 1973 *Memory and Intelligence*. Basic Books, New York]
- Vidal F 1994 *Piaget before Piaget*. Harvard University Press, Cambridge, MA

J. Montangero

Copyright © 2001 Elsevier Science Ltd. 11433
All rights reserved.

Piaget's Theory of Child Development

Jean Piaget's (1896–1980) theory counts as one of the most influential and enduring theories of cognitive development, and is firmly grounded in biology and epistemology. Although the focus of this entry is on the child developmental aspect of his work, it is therefore important to review some of the key epistemological concepts that Piaget adopted. Unsolved issues as well as present and future developments of Piaget's theory are also briefly reviewed.

1. The Epistemology of Piaget

Piaget, who trained as a biologist, had as a primary objective to understand the development of knowledge in the human species, rather than to understand why and how children develop. His first originality was to address philosophical questions (such as the origin or the development of knowledge) by empirical means, in particular by relying on the study of child development (supposed to be, at the beginning of his career, a mere 'detour' of some years). He established the discipline of *genetic epistemology* (the term 'genetic' referring to the concept of genesis or evolution, as proposed by Baldwin), with the aim 'to study the roots of the various sorts of knowledge from their most elementary forms on and to follow their development in ulterior levels including scientific thinking,' and by grounding it in both historico-critical and psychogenetic approaches. Piaget's epistemology is constructivist, in the sense that knowledge is neither a mere reflection of the external world (realism), nor a projection of pre-existing structures of the mind (nativism), and the origins of knowledge are to be found in the practical and cognitive activity of the subject. Piaget's epistemology is also fundamentally interdisciplinary.

His primary hypothesis was that intelligence is a form of biological adaptation (e.g., Piaget 1967), that is, intellectual or cognitive behaviors are the products of an organism, in interaction with its environment, and the end point of biological evolution. This is not to say that psychological phenomena can be reduced to biological ones, as Piaget always opposed biological reductionism; intelligence is an organizing activity whose functioning transcends biological organization, by elaborating new structures. The main characteristic common to knowledge and to living organisms is their adaptive character. *Adaptation* itself, whether biological or cognitive, relies on two mechanisms, defined very early in Piaget's work, namely *assimilation* (incorporation of new information into an existing system) and *accommodation* (modification of existing schemes or structures by newly assimilated elements, so as to respond to the demands of the environment). By so doing, actions and operations become coordinated with one another; such coordinations define a new entity, which in turn constitutes a new object for

thought and action. This is the process that Piaget labeled *equilibration*. Contrarily to frequent, but too simplistic interpretations, equilibration does not correspond to a stable (even if only temporary) state of equilibrium, but is the process of coordination in itself.

As a consequence of the functional continuity that Piaget postulated between biological and psychological entities, a given level of knowledge is always considered to result from a reorganization of the preceding one. Reconstruction means more than a mere addition of elements of a lower type in order to attain a higher level, but implies a total reorganization and a change in scale. One direct consequence of a constructivist perspective is that an epistemologist or a psychologist studying any level of development will ask what characterized the preceding level and will stress developmental progression towards higher levels. An essential mechanism for this reconstruction is *reflecting abstraction*, that is, a process by which the subject extracts characteristics or coordinations from actions or operations; it is called 'reflecting' both because it consists of a mental reorganization and because it implies a reflection or a transfer onto a higher level of what was established at a lower level. Reflecting abstraction is contrasted with empirical or simple abstraction, which bears mainly on observable characteristics of objects or classes of objects.

2. The Psychology of Jean Piaget

Piaget's psychology of intelligence is the most universally known part of his work, even though he himself did not consider it the main part (e.g., Piaget 1947, 1970, Piaget and Inhelder 1966). His theory transformed the field of developmental psychology by providing it with a new vision of the development of the child, to the extent that almost every developmental study has some connection with a question Piaget raised. Intelligence is defined as the most general form of coordination of the actions and operations that characterize the various developmental levels, and not as a mental faculty or an entity in itself; it develops through a succession of general stages, defined by overall structures ('*structures d'ensemble*'). Piaget was not interested in individual or task-specific performances, but in the generic behavior of subjects in classes of situations.

The successive stages or levels are general structures, which appear in a fixed order of succession, but are not necessarily tied to a given age, and which are such that a given structure n integrates, as a substructure, the preceding structure $n-1$. The number of stages defined in Piaget's publications varied from three to five. Most often, he referred to four stages: (i) the *sensorimotor stage* (approximately the first two years), itself consisting of six substages, or three periods, in which practical *schemes* develop, among which is the scheme of object permanence (operationalized by the

study of the search for hidden objects), one of the most studied domains in present day infancy research; (ii) the *preoperational stage* from approximately 2 to 6 years of age, also labeled representative (prerelational and intuitive, or prelogical) period, made possible by the emergence and the development of the *symbolic function*; (iii) the *concrete operational stage* (7 to 12 years) during which invariants (conservation), logical classes and relations, the concept of number, as well as different concepts of space develop, all of which are marked by reversibility; and (iv) the *formal operational stage*, during which adolescents develop hypothetico-deductive reasoning (i.e., beyond concrete objects and actions), and experimental schemata. Piaget used abstract algebra to formalize the underlying structures: a group of displacements for the sensorimotor stage, groupings for the concrete operational stage, and a group of four operations (INRC group) for the formal operational stage. He also attempted to formalize the pre-operational stage by means of the logic of functions. Other cognitive activities such as mental imagery, perception, memory, and language were always dealt with by Piaget in relation to the development of operational structures. They were termed 'figurative' because they attempt to represent reality as it appears. Figurative activities are contrasted with (and subordinated to) 'operative' activities which attempt to transform reality (e.g., Piaget and Inhelder 1966b).

Although Piaget grew less and less interested in the description of these general stages, this is probably the feature of his theory that has been most often retained and criticized. In particular, the generality of the stages was questioned by the existence of '*decalages*,' that is, temporal lags between acquisition of concepts supposed to pertain to the same level. Readers of Piaget have often overestimated the necessity of synchronism in acquisitions across situations as implied by the theory. It should be stressed that Piaget did not postulate that the generality of structures meant strict synchronism, even though some of his writings were admittedly ambiguous with respect to this issue (e.g., Chapman 1988).

3. Factors of Development

Piaget described four factors of development: (i) organic growth, and maturation of the nervous and endocrine systems; (ii) experience, which Piaget further decomposed into a factor of physical experience (actions exerted upon objects in order to extract their properties by simple abstraction) and a factor of logico-mathematical experience (extraction of knowledge by reflecting upon actions, whether effective or mental); (iii) social interactions and transmission; and (iv) equilibration. This last factor was undoubtedly the most important for Piaget; whereas he often paid little

more than lip-service to the first three factors, he spent considerable time in demonstrating the necessity of equilibration, which he considered to account for the 'vection' or direction of development (Piaget 1975). Majoring equilibration ('*équilibration majorante*') is seen as the causal mechanism underlying the progression of cognitive development. It is a dynamic sequence of states of disequilibrium–equilibrium and embodies the organism's strive toward a more general equilibrium. Equilibration is the only mechanism, according to Piaget, that can account for the emergence of novel behavior, even though more recent work has disputed the idea that Piaget's theory can really account for the emergence of novel behaviors. Cognitive structures are put into disequilibrium by perturbations which resist the assimilatory process and lead to errors or failures of action. Re-equilibration is made possible by regulations, which entail the modification of an action as a function of feedback regarding its previous outcomes, ultimately leading to compensatory reconstructions. Equilibration is thus necessary when contradictions and cognitive conflicts arise, hence the importance given by Piaget to cognitive conflict as a source of cognitive development.

4. Unsolved Issues in the Piagetian Theory

Piagetian theory has been subjected to numerous criticisms (probably proportional to his production) on the part of developmental psychologists, only some of which are summarized here; most critics addressed, however, only isolated elements of the theory or failed to take into account the fact that Piaget's project was essentially epistemological. First, numerous decalages, which could not be predicted by the theory, have been observed between concepts supposed to belong to the same structural level; between-task correlations are relatively low, implying that all tasks do not measure the same construct. This led a number of researchers to claim that development is domain specific. Moreover, the concept of overall logico-mathematical structure proved very difficult to operationalize. Second, Piaget's portrait of the child appeared too universal and focused on the subject, independently of his/her cultural environment. Third, the theory placed too much emphasis on the development of logico-mathematical thinking, as contrasted with other forms of thinking. Fourth, Piagetian theory did not provide sufficient information on learning mechanisms and on the effects of training. Consistent with his epistemological project, Piaget was indeed interested in demonstrating the commonalities, in terms of general mechanisms, between learning and development rather than in focusing on the specificities of learning. Linked to this is the difficulty for the Piagetian theory to account for the appearance of novel behaviors. Finally, like many other, general psychological

theories, Piaget did not take into consideration individual differences; however, it has been shown that the variability in performance, whether intra- or inter-individual, is much larger than could at first be suspected and constitutes a fundamental phenomenon rather than mere noise.

5. *The Heritage of Piaget*

Despite the criticisms addressed to the theory, Piaget's theory leaves an enormous heritage, some of which is no longer even acknowledged, because it has been incorporated into developmental psychology, almost as truisms. First, the constructivist view of development that Piaget defended is presently widely held by cognitive developmentalists, as well as by educational and cognitive psychologists. Likewise, the concepts of assimilation and accommodation, even though probably too general, are adopted by a large number of theoreticians. The view of children as active thinkers has also been generally adopted. Second, although there was a period in the 1960s and 1970s during which a large number of developmental researchers argued for domain specificity in development and against the existence of general stages, there is a growing consensus that the characterization of human cognitive development needs something more general than a long list of specific concepts, even if less general than the cognitive structures defined by Piaget. In any case, most developmentalists see the need for structural analyses that transcend any particular cognitive task. Similarly, the concept of scheme (defined as an invariant abstracted across several actions, to which new actions or objects can be assimilated and which allows for generalizations) has been retained and is close to other concepts such as schemas or scripts proposed in cognitive psychology. Third, although the construct of equilibration has been criticized for being too fuzzy, similar conceptualizations are now being proposed, whether by dynamic systems theoreticians, or by neo-Piagetians.

Last, but not least, Piaget contributed enormously to the research methods used in developmental psychology, even though he was criticized for not using a standard, quantifiable methodology. He pioneered the use of a clinical method (later to be called critical method), in which the researcher probes for the child's understanding through repeated questioning. Piaget insisted that the child's incorrect responses reflect a complex, underlying cognitive system, and should not be merely recorded as failures. Also, Piaget's work was and still is a source of very ingenious empirical techniques, in many different fields; Piagetian measures such as conservation, class inclusion, spatial representation, perspective taking, object permanence, moral judgment, transitive inference, and many others, have each produced countless studies by researchers around the world. Most of these tasks

were very productive, because they produced data that were both counterintuitive and generative of hypotheses about the way children think.

6. *Beyond Piaget's Theory*

As mentioned in the previous section, the Piagetian theory is still very active, not only in developmental psychology, but in neighboring fields too; many researchers around the world continue working with Piagetian tasks, even if they do not always acknowledge it, or even know it. Also from an empirical perspective, a flourishing new direction, which is very much in the spirit of Piaget's theory, is that of the 'theory of mind.' From a theoretical point of view, the core part of the theory has been retained while giving rise to new developments, as should be the case with any sound theory. Only some concepts were really abandoned, the best example of which is the logicization of overall structures. Without pretending to be exhaustive, two directions, at present, appear to be very close to Piaget's proposals, and promising for the future, namely the neo-Piagetian models, and the application of dynamic systems modeling to developmental psychology.

The objective of the *neo-Piagetian* models, initiated by Pascual-Leone and exemplified by scholars such as Robbie Case, Kurt Fischer, Graeme Halford, and Robert Siegler (e.g., Case 1992, Case and Edelstein 1993, Pascual-Leone 1987) was to provide a response to some of the questions left open in Piagetian theory, while preserving most of its core postulates. These models have in common the search for isomorphisms or invariants across different classes of situations. As a consequence, they have maintained or rehabilitated (after the emphasis on domain specificity mentioned above) the notion of general stage; however, the structuralist features are quite different from those postulated by Piaget. First, the hypothesis of a unique, general, logical structure underlying the organization of all behaviors at a given moment of development was abandoned, and replaced by more local ones, thereby granting more importance to context. Second, isomorphisms across domains are considered to result from common constraints or 'strictures' rather than from a common structure: general limits in processing resources or attentional capacity impose a ceiling on cognitive performances while allowing for situational and individual variability. Third, the importance given to context has as corollary a much stronger emphasis on task analysis, that is, on an analysis of the characteristics of external situations together with an analysis of the strategies that children adopt when confronted with these situations and of the processing load that these strategies entail. The challenge now is to understand under which conditions characteristic forms of reasoning are most likely to emerge, and why they do not despite the presence of sufficient general

resources. Fourth, some neo-Piagetians also assign a more fundamental status to individual differences; consequently, they claim that there may exist different developmental paths for different types of subjects. This implies that development is no longer led by a single, dominant general mechanism such as equilibration, but that several general mechanisms may interact to produce development, leading to a postulate of multidimensionality and multidirectionality in development (Reuchlin et al. 1990; Rieben et al. 1990). Such a postulate has long been adopted by lifespan developmentalists (e.g., Baltes et al. 1998), but took longer to be envisioned by child developmentalists.

A new, promising way of theorization is provided by nonlinear dynamic systems theories. Such models can be used to detect, and then model, developmental transitions, that is, changes from one stable mode to another one. The emphasis is on self-organization, which can be linked to Piaget's concept of equilibration. Although some scholars have begun to use nonlinear dynamics to build specific models of processes of development, particularly in the field of motor development or in conservation tasks (e.g., Thelen and Smith 1994, Van Geert 1994), these are only the preliminaries, and are still to be taken as metaphors; moreover, they require heavy methodological designs. Nevertheless, such attempts, which may help uncover discontinuities, i.e. stages, under what appears to be continuous change, could help bring explicit modeling into developmental psychology.

See also: Cognitive Development in Childhood and Adolescence; Developmental Psychology; Infant and Child Development, Theories of; Infant Development: Physical and Social Cognition; Lifespan Theories of Cognitive Development; Piaget, Jean (1896–1980); Piaget's Theory of Human Development and Education; Theory of Mind

Bibliography

- Baltes P B, Lindenberger U, Staudinger U M 1998 Life-span theory in developmental psychology. In: Damon W (ed.) *Handbook of Child Psychology*, 5th edn., Vol. 1, Theoretical Models of Human Development. J. Wiley, New York, pp. 1027–143
- Case R 1992 Neo-Piagetian theories of intellectual development. In: Beilin H, Pufall P (eds.) *Piaget's Theory: Prospects and Possibilities*. L. Erlbaum Associates, Hillsdale, NJ, pp. 61–104
- Case R, Edelstein W 1993 *The New Structuralism in Cognitive Development: Theory and Research on Individual Pathways*. Karger, Basel
- Chapman M 1988 *Constructive Evolution: Origins and Development of Piaget's Thought*. Cambridge University Press, New York

- Pascual-Leone J 1987 Organismic processes for neo-Piagetian theories: A dialectical causal account of cognitive development. *International Journal of Psychology* **22**: 531–70
- Piaget J 1947 *La Psychologie de L'intelligence*. Colin, Paris. [1950 *The Psychology of Intelligence*. Routledge & Paul, London]
- Piaget J 1967 *Biologie et Connaissance*. Gallimard, Paris. [1971 *Biology and Knowledge*. University of Chicago Press, Chicago]
- Piaget J 1970 Piaget's theory. In: Mussen P H (ed.) *Carmichael's Manual of Child Psychology*, 3rd edn., Vol. 1. Wiley, New York [Reprinted in Mussen P H 1983 *Handbook of Child Psychology*, Vol. 1. Wiley, New York]
- Piaget J 1975 *L'équilibration des Structures Cognitives*. Presses universitaires de France, Paris. [1977 *The Development of Thought: Equilibration of Cognitive Structures*. Viking Press, New York]
- Piaget J, Inhelder B 1966 *La Psychologie de L'enfant* 2nd edn. Presses universitaires de France, Paris. [1969 *The Psychology of the Child*. Basic Books, New York]
- Piaget J, Inhelder B 1966b *L'image Mentale Chez L'enfant*. Presses universitaires de France, Paris. [1971 *Mental Imagery in the Child*. Routledge & Paul, London]
- Reuchlin M, Lautrey J, Marendaz C, Ohlmann T (Eds.) 1990 *Cognition: l'individuel et l'univers*. Presses universitaires de France, Paris
- Rieben L, de Ribaupierre A, Lautrey J 1990 Structural invariants and individual modes of processing: On the necessity of a minimally structuralist approach of development for education. *Archives de Psychologie* **58**: 29–53
- Thelen E, Smith L B 1994 *A Dynamic Systems Approach to the Development of Cognition and Action*. MIT Press, Cambridge, MA
- Van Geert P 1994 *Dynamic Systems of Development*. Harvester Wheatsheaf, New York

A. de Ribaupierre

Piaget's Theory of Human Development and Education

1. Introduction

The life and work of Jean Piaget are one: his work was his life and his life inspired his work (Vonèche 2001). His theory of genetic epistemology (or experimental developmental theory of knowledge and knowledge acquisition) resulted in an immense *œuvre* of several thousand pages. All stemmed out of the encounter of his philosophical preoccupations as an adolescent about the validity of knowledge (scientific and religious essentially) when compared with his experience as a young natural historian.

As an amateur philosopher, in his teens, he had some deep religious preoccupations and wanted to reconcile science and faith into one unique form of knowledge. Religious knowledge was, in his mind, essentially subjective and emotional, whereas scientific

knowledge was objective and strictly cognitive. The obvious solution for the young Piaget was double. On the one hand, 'science cannot prescribe its premisses to value judgment ... [they] are given in immediate awareness and are undemonstrable' (Piaget 1915, p. 80). On the other hand, thinking is the essence of human existence in a very Cartesian (*cogito ergo sum*) way and in opposition both to, Kant's *a priori*ism on the one hand, and to, on the other hand, Brunschvig's idea of a radically contingent spiritual development over time in the history of human thought. Therefore, the solution of the problem of the origin and value of knowledge lies in a sort of experimental study of its evolution and metamorphoses over that special sort of time called ontogenesis. He considered this study as a sort of embryology of knowledge yielding the same results as biological embryology for the succession of living organisms and their hierarchical order. The stages of knowledge are equivalent to the biological scale of organisms.

2. Basic Theoretical Assumptions

Thus, knowledge is biological adaptation continued by different means. Whereas biological adaptation is made possible by specific organs performing specific functions and by specific species occupying specific ecological niches—thus essentially a mechanism of differentiation and speciation—things are quite different in the case of knowledge. In intelligent adaptation there is an ever increasing separation of form and content and cognitive processes exploit the organism as a whole and no longer specific parts of it.

What remains similar in biological and cognitive adaptation is the double process of assimilation and accommodation. Biological equilibrium is assured by metabolism, cognitive equilibrium by self-regulation. Cognitive self-regulation takes three different forms: rhythms, self-regulations, and mental operations. These three structures go from the simplest to the more complex and from the most external interface between organism and milieu to the most internal.

If rhythm is self-explanatory, the two forms of self-regulations: homeostasis and homeorhesis, are not. Homeostasis is the process by which any variation in one direction from the point of equilibrium is counter-balanced by a movement in the opposite direction. These movements existed in rhythm, but were successive. Now, they are simultaneous in homeostasis and form the components of a system. They limit each other by a reciprocal process of alternances. It is simply and purely a feed-back mechanism.

Homeostasis assures the stability of a given system or organism by keeping it in a stationary state in spite of perturbations. Homeorhesis is the same for evolutionary systems: instead of keeping the system stationary, it keeps the direction of evolution going, in spite of perturbations, on its path. It is a feed-forward

mechanism. What is common thus to regulations and rhythms is the central role of action. Every action in one direction is compensated by an action in the opposite direction, either successively or simultaneously.

An operation, for Piaget, is nothing more than an internalized action. For instance, an arithmetic operation is the action of reuniting two units: $2 + 2 = 4$. But an operation is a very remarkable action, because it is strictly reversible. Rhythmic actions are not reversible, they always take place in the oriented arrow of time. Regulations do not compensate each other totally; they always function by excess or default. They are not strictly additive and totally reversible. Operations are perfectly reversible: $4 - 2 = 2$ exactly. Such a perfection is mental, never physical.

The smart thing with Piaget's theory is that the same basic process covers the full range of adaptation from the simple metabolic assimilation of food to the most refined, abstract logico-mathematical operation. The process of transformation of rhythm into regulation and of regulation into operation is double: internalization, on the one hand, and reflexive abstraction, on the other.

But, if one thinks it over, internalization is a reflexive process and thus reflexive abstraction is enough. It is a reflection of a form on a higher level.

3. Cognitive Development

The process of knowledge is the same in the child and in the species. It goes, in both, from a state of dualism between the subject and the object to successive stages of differentiation and integration. The newborn baby does not distinguish self from world at first. Then the distinction is made by the baby only in the course of action. Piaget calls this period sensory-motor.

Its main features are the recognition that objects continue to exist when out of sight because they are part of a larger container called space, that they appear in time and present some relationship of causality, as far as reality is concerned, and that means and ends can be coordinated and differentiated, as far as intelligence is concerned. By reflexive abstraction, on its external immediate action, another level of equilibrium is reached by the growing child with the emergence of the semiotic function (language, symbolic play, imitation, representational drawing, dream). This new way of functioning makes present what is absent in reality. It paves the way for concepts, classes, relations, and number that will be attained in the next stage called concrete operational in which representational self-regulations become fully equilibrated into mental operations. In turn, concrete operations will be transformed into formal ones that are independent from their material support by means of a progressive dissociation between form and content until it is total and forms a complete logical system.

Thus the young adolescent is, by now, capable of hypothetico-deductive reasoning, of examining all the possibilities as well as the internal necessity and coherence of every single thought.

4. Piaget's Constructivism

The data collected by Piaget and the School of Geneva are world-famous. They have been criticized on many grounds, but always from a reductionist epistemological point of view, either nativism or environmentalism, whereas Piaget was constructivist. This constructivist perspective has been constantly misunderstood. It should be clear that such a structure as an operation with its internal logical necessity does not correspond either to a physiological apparatus or to a societal given. No mother, no teacher teaches a baby object permanence. Nevertheless all toddlers attain this concept-in-action. Conservation of matter, weight or volume is not taught in any school or home and nobody remembers having discovered it one morning. Nevertheless, it is reached without critical period. Nor drill.

Constructivism does not deny reality, contingency or stability of scientific knowledge as it is occasionally assumed. This assumption stems out of the confusion of Piagetian constructivism with a fairly recent trend in sociology of knowledge called the social construction of science.

Unlike sociological constructivism, Piaget recognizes the remarkable stability of scientific knowledge over time. He never denied, for instance, that the successive physical theories that have appeared since Newton or even Galileo were so stable that they encompassed superseded theories as special cases of their own theory. Piaget specifically showed the striking case for theory inclusion that Galileo's system formed for Newton's and Newton's for Einstein's. He made the same remark about modern theories of force interactions in contemporary physics.

As far as contingency is concerned, he accepted without any doubt that the advance of molecular biology was made possible by the concomittant and congruent development of computer technology allowing for long and complex computations required by molecular biological methods. Brunschvig's influence had been strong enough in that sense on the young Piaget as to remain with the old man.

Reality represents a greater difficulty, because there are different ways of approaching this concept. One can consider reality as the information given by the external world to which the human mind has to conform willy-nilly by registering it or dying. Or one can consider reality as what resists the *modus operandi* of the mind, in which case an object is what objects to the mind's demands and not a mere thing, outside there, imposing upon the mind.

Reality, for a constructivist, is a constant negotiation between the subject and the object in the course of their interaction. But, once again, the term interaction has a precise meaning for Piaget. Basically, interaction can be understood in three different ways. The first meaning is a mere relationship between two separate elements. For instance, in the case of a heated iron bar that is longer than a cold one, there is an interaction between length and temperature. The second meaning implies circularity in the relationship. For instance, the circularity between the scarcity of public transportation in the countryside which is due to the excess of private cars which, in turn, is caused by the lack of public transportation. The third and last meaning is Piagetian: it supposes the unseparability between reality and mental 'schèmes.' This means that a scientific fact does not exist independently from a scientific theory.

This is true for naive theories too. Before object permanence, it is pointless for a child to search for a disappeared object, since, in the infant's naive theory, the object has been reabsorbed by the environment until it comes back by magic and not by means of a complex logico-mathematical structure called the practical group of displacements.

Thus, Piagetian constructivism should not be confused with the social construction of reality. It is a theory that drives a middle course between the extremes of empiricism and rationalism in philosophy, Darwinism and creationism in biology, Durkheim and Tarde in sociology, behaviorism and Gestalt in psychology. In addition, Piaget thought he had empirically tested the validity of his position in the course of his studies of cognitive developmental as we shall see now.

5. Cognitive Development and Education

Essentially what Piaget observed is that intelligence evolves from states of lesser knowledge to higher ones and that cognition goes from simple to complex. But, if knowledge were learned, it would go at random from some patches of learning to others according to chancy encounters with the environment. On the other hand, if knowledge were the mere unfolding of innate potentialities, why should it go from simple to complex? There is no evolutionary nor social need for such a contrived process. Thus knowledge must be constructed step by step from direct action to internalized mental operations.

At every level of this construction corresponds a mode of equilibrium between self and reality that forms the structure of a stage that Piaget defines in logico-mathematical terms. The very existence of stages is, for Piaget, the evidence that knowledge is neither the mere unfolding of innate potentialities as in nativism nor simple registration of the features of the world as in environmentalistic and social learning theories, since the notion of stage implies a total

reorganization on a higher plane of what has been attained earlier at a lower level. This is mental construction.

Such a position involves also a choice of explanatory models that has been often misunderstood. Instead of choosing a physiological model of explanation for cognitive development such as maturation or brain states and transformation as do most psychologists, Piaget selected abstract models such as mathematical groups or equilibration theory to account for psychological processes such as adaptation. By so doing he conformed his explanation to his observation that intelligence is biological adaptation pursued by new and more potent means.

This carries an educational lesson. Teaching and learning which are essential to human development should be adjusted to the child's level of understanding so that any pedagogical incentive speaks to the child's condition at any given moment but also appeals to the next higher structure by calling attention on the contradictions and lacunae of the existing developmental level. This is no spontaneism. This is no drill learning nor socialization. This is cooperation (in the strong sense of a Piagetian operation) with the growing mind from the part of educators whose noble task is to co-construct knowledge with every new baby who comes to this world.

See also: Cognitive Development: Child Education; Piaget, Jean (1896–1980); Piaget's Theory of Child Development; Vygotskij, Lev Semenovic (1896–1934); Vygotskij's Theory of Human Development and New Approaches to Education

Bibliography

- Boden M A 1979 *Piaget*. The Harvester Press, Brighton, UK
Chapman M 1988 *Constructive Evolution: Origins and Development of Piaget's Thought*. Cambridge University Press, New York
Furth H G 1969 *Piaget and Knowledge: Theoretical Foundations*. Prentice-Hall, Englewood Cliffs, NJ
Furth H G 1970 *Piaget for Teachers* 2nd edn. Prentice-Hall, Englewood Cliffs, NJ
Gruber H, Vonèche J (eds.) 1995 *The Essential Piaget*. Aronson, Northvale, NJ
Modgil S, Modgil C (eds.) 1983 *Jean Piaget: An Interdisciplinary Critique*. Routledge & Kegan Paul, London
Piaget J 1915 *La Mission de l'Idée*. La Concorde, Lausanne, Switzerland
Piaget J 1970 *Genetic Epistemology*. Columbia University Press, New York
Piaget J 1971 *Biology and Knowledge: An Essay on the Relations Between Organic Regulations and Cognitive Processes*. University of Chicago Press, Chicago (Original work published 1967)
Piaget J, Inhelder B 1969 *The Psychology of the Child*. Routledge & Kegan Paul, London (Original work published 1966)

Smith L (ed.) 1992 *Jean Piaget: Critical Assessments*, Vols 1–4. Routledge, London

Vonèche J 2001 Identity and narrative in Piaget's autobiographies. In: Brockmeier J, Carbaugh D (eds.) *Narrative and Identity: Studies in Autobiography, Self and Culture*. John Benjamins, Amsterdam

J. J. Vonèche

Pidgin and Creole Languages

1. Introduction

Most studies of pidgins and Creoles (PC) have focused on their origins, despite an undeniable increase during the 1990s in the number of works on structural features. Recently, some creolists have also addressed the question of whether, as a group, Creoles can be singled out as a structural type of languages. Space limitations make it impossible to discuss structural features in this essay, aside from the fact that there are no features that are peculiar to PCs, as explained in Sect. 2.

2. What are Pidgins and Creoles?

Strictly speaking, PCs are new language varieties, which developed out of contacts between colonial nonstandard varieties of a European language and several non-European languages around the Atlantic and in the Indian and Pacific Oceans during the seventeenth to nineteenth centuries. Pidgins typically emerged in trade colonies which developed around trade forts or along trade routes, such as on the coast of West Africa. They are reduced in structures and specialized in functions (typically trade), and initially they served as non-native *lingua francas* to users who preserved their native vernaculars for their day-to-day interactions. Some pidgins have expanded into regular vernaculars, especially in urban settings, and are called 'expanded pidgins.' Examples include *Bislama* and *Tok Pisin* (in Melanesia) and Nigerian and Cameroonian Pidgin English. Structurally, they are as complex as Creoles. The latter vernaculars developed in settlement colonies whose primary industry consisted of sugar cane plantations or rice fields, which employed massive non-European slave labor. Examples include Cape Verdean *Criolou* (lexified by Portuguese) and *Papiamentu* in the Netherlands Antilles (apparently—Portuguese-based but influenced by Spanish); Haitian, Mauritian, and Seychellois (lexified by French); Jamaican, Guyanese, and Hawaiian Creole, as well as Gullah in the USA (all lexified by English); and

Saramaccan and *Sranan* in Surinam (lexified by English, with the former heavily influenced by Portuguese and the latter by Dutch). Note that although Melanesian pidgins are associated with sugar cane plantations, they apparently originated in trade settings and were adopted on the plantations (Keesing 1988).

The terms *Creole* and *pidgin* have also been extended to some other varieties that developed during the same period out of contacts among primarily non-European languages. Examples include Delaware Pidgin, Chinook Jargon, and Mobilian in North America; Sango, (Kikongo-)Kituba, and Lingala in Central Africa, Kinubi in Southern Sudan and in Uganda; and Hiri Motu in Papua New Guinea (Holm 1989, Smith 1995). In the original, lay people's naming practice, the term *jargon* was an alternate to pidgin. However, some creolists claim that pidgins are more stable and jargons are an earlier stage in the 'life-cycle' that putatively progresses from Jargon, to Pidgin, to Creole, to Post-Creole by progressive structural expansion, stabilization, and closer approximations of the *lexifier*—the language which contributed the largest part of a Creole's lexicon.

Chaudenson (1992, p. 21) argues that Creoles developed by *basilectizing* away from the lexifier, i.e., acquiring a *basilect*, which is the variety the most different from the *acrolect*, the educated variety of the lexifier. Mufwene (2001) emphasizes that Creoles and pidgins developed in separate places, in which Europeans and non-Europeans interacted differently—sporadically in trade colonies (which produced pidgins) but regularly in the initial stages of settlement colonies (where Creoles developed). Moreover, the term *pidgin* was coined in 1807 (Baker and Mühlhäusler 1990), over two years after the term *Creole* was used in reference to a language variety. '*Pidgin English*,' apparently a distortion of '*business English*' developed in Canton, a geographical area where no large colonial plantation industry developed and no Creoles have been identified.

The term '*Creole*' was originally coined in Iberian colonies, apparently in the sixteenth century, in reference to nonindigenous people born in the American colonies. It was adopted in metropolitan Spanish, then in French and later in English by the early seventeenth century. By the second half of the same century, it was generalized to descendants of Africans or Europeans born in Romance colonies. Usage varied from one colony to another. The term was also used as an adjective to characterize plants, animals, and customs typical of the same colonies.

Creole may not have applied widely to language varieties until the late eighteenth century. Such usage may have been initiated by metropolitan Europeans to disfranchise particular colonial varieties of their languages. It is not clear how the term became associated only with vernaculars spoken primarily by descendants of non-Europeans. Nonetheless, several

speakers of Creoles (or pidgins) actually believe they speak dialects of their lexifiers.

Among the earliest claims that Creoles developed from pidgins is the following statement in Bloomfield (1933, p. 474): 'when the jargon [i.e., pidgin] has become the only language of the subject group, it is a creolized language.' Hall (1962) reinterpreted this, associating the vernacular function of Creoles with nativization. Thus, Creoles have been defined inaccurately as 'nativized pidgins,' i.e., pidgins that have acquired native speakers and have therefore expanded both their structures and functions and have stabilized. Hall then also introduced the pidgin-Creole 'life-cycle' to which DeCamp (1971) added a 'post-Creole' stage.

The first creolist to dispute this connection was Alleyne (1971). He argued that fossilized inflectional morphology in Haitian Creole (HC) and the like is evidence that Europeans did not communicate with the Africans in foreigner or baby talk, which would have fostered pidgins on the plantations. Chaudenson (1992, p. 21) argues that plantation communities were preceded by homesteads on which mesolectal approximations of European lexifiers, rather than pidgins, were spoken by earlier slaves. Like some economic historians, the socioeconomic history literature suggests that in North American colonies Creole Blacks spoke the lexifier fluently. In ads on runaway slaves in British North American colonies, bad English is typically associated with slaves imported as adults from Africa. Diachronic evidence of Creoles suggests that the basilects developed during the peak growth of plantations (in the eighteenth century for most colonies!), when infant mortality was high, life expectancy was short, the plantation populations increased primarily by massive importation of slave labor, and the proportion of fluent speakers of the earlier colonial varieties kept decreasing (Chaudenson 1992, p. 21, Mufwene 2001).

According to the life-cycle model, as a Creole continues to coexist with its lexifier, the latter exerted pressure on it to shed some of its 'Creole features.' This developmental hypothesis can be traced back to Schuchardt's (1914) explanation of why African-American English (AAE) is structurally closer to North American English than Saramaccan is to its lexifier: coexistence with it in North America and absence of such continued contact in Suriname. DeCamp (1971) and several creolists who followed suit (e.g., Rickford 1987) resurrected the position by invoking 'decreolization' ('loss of "Creole" features') to account for speech continua in Creole communities.

It is in this context that DeCamp (1971) coined the term 'post-Creole continuum,' which must be interpreted charitably. If a variety is Creole because of the particular sociohistorical ecology of its development, rather than because of its structural peculiarities, it cannot stop being a Creole even after some of the features have changed. Besides, basilectal and mesolectal features continue to coexist in these communi-

ties, suggesting that Creole has not died yet. Lalla and D'Costa (1990) present copious data against decreolization in Caribbean English Creoles.

Closely related to this issue is the common assumption that Creoles are separate languages from their lexifiers and ex-colonial varieties thereof spoken by descendants of Europeans. Thus, the nonstandard French varieties spoken in Quebec and Louisiana, as well as on the Caribbean islands of St. Barths and St. Thomas, are considered dialects of French rather than Creoles. Likewise New World nonstandard varieties of Spanish and Portuguese are not considered Creoles, despite structural similarities which they display with Creoles of the same lexifiers. Although not admitted in linguistics, the classification seems to have been associated generally with whether the majority in the largely proletarian communities speaking the new, colonial vernaculars is of European or of non-European descent. Otherwise, all colonial varieties of European languages are restructured and contact-based; the current classification of vernaculars into Creoles and non-Creoles remains question begging. The embarrassment stems largely from the absence of a yardstick for measuring structural divergence from the lexifier, especially when it is dubious that it was the same in every contact setting.

It has also been claimed that Creoles have more or less the same structural design (Bickerton 1984). This position is as disputable as the counterclaim that they are more similar in the sociohistorical ecologies of their developments, or even the more recent claim that there are Creole prototypes from which others deviate in various ways (McWhorter 1998). The very fact of resorting to a handful of prototypes for the general Creole structural category suggests that the vast majority of them do not share the putative set of defining features, hence that the features cannot be used to single them out as a unique type of language. On the other hand, variation in the structural features of Creoles (lexified by the same language) is correlated with variation in the linguistic and sociohistorical ecologies of their developments (Mufwene 2001). The notion of 'ecology' includes, among other things, the nature of the lexifier, structural features of the substrate languages, changes in the ethnolinguistic make-up of the populations that came in contact, the kinds of interactions between speakers of the lexifier and those of other languages, and rates and modes of population growth.

To date the best known Creoles have been lexified by English and French. Those of the Atlantic and Indian Ocean are, along with Hawaiian Creole, those that have informed most theorizing on the development of Creoles. While the terms 'Creole' and 'creolization' have often been applied uncritically to various contact-induced language varieties, several distinctions, which are not clearly articulated have also been proposed, for instance, between pidgin, Creole, koiné, semi-Creole, intertwined varieties,

foreign workers' varieties of European languages (e.g., *Gastarbeiter Deutsch*), and 'indigenized varieties' of European languages (e.g., Nigerian and Singaporean English). The denotations and importance of these terms deserve re-examining.

3. *The Development of Creoles*

The central question here is: how did Creoles develop? The following hypotheses are the major ones competing today: the substrate, the superstrate, and the universalist hypotheses.

Substratist positions are historically related to the 'baby talk hypothesis,' which I have traced back to nineteenth-century French creolists (e.g., Adam 1883). Putatively, the languages previously spoken by the Africans enslaved on New World and Indian Ocean plantations were the primary reason why the European lexifiers which they appropriated were restructured into Creoles. These French creolists assumed African languages to be 'primitive,' 'instinctive,' in 'natural' state, and simpler than the 'cultivated' European languages with which they came in contact. Creoles' systems were considered to be reflections of those non-European languages. The baby-talk connection is that, in order to be understood, the Europeans supposedly had to speak to the Africans like to babies, their interpretation of foreigner talk.

The revival of the substrate hypothesis (without its racist component) has been attributed to Sylvain (1936). Although she recognizes influence from French dialects, she argues that African linguistic influence, especially from the Ewe group of languages, is very significant in HC. Unfortunately, she states in the last sentence of her conclusions that this Creole is Ewe spoken with a French vocabulary. Over two decades later, Turner (1949) disputed American dialectologists' claim that there was virtually no trace of African languages in AAE and showed phonological and morphosyntactic similarities between Gullah and some West-African (especially Kwa) languages. He concluded that 'Gullah is indebted to African sources' (p. 254).

There are three main schools of the substrate hypothesis at the beginning of the twenty-first century. The first, led by Alleyne (1980) and Holm (1988) is closer to Turner's approach and is marked by what is also its main weakness: invocation of influence from diverse African languages without explaining what kinds of selection principles account for this seemingly random invocation of sources. This criticism is not *ipso facto* an invalidation of substrate of substrate influence; it is both a call for a more principled account and a reminder that the nature of such influence must be reassessed (Mufwene 2001).

The second school has been identified as the 'relexification hypothesis.' Its leading proponent Lefebvre (1998) argues that HC is a French relexifi-

cation of languages of the Ewe-Fon (or Fongbe) group. This account of the development of Creoles has been criticized for several basic shortcomings, including the following: (a) its 'comparative' approach has not taken into account several features that HC (also) shares with nonstandard varieties of French; (b) it downplays features which HC shares also with several other African languages which were represented in Haiti during the critical stages of its development; (c) it has not shown that the language appropriation strategies associated with relexification are typically used in naturalistic second language acquisition; and (d) it does not account for those cases where HC has selected structural options which are not consistent with those of Ewe-Fon. Moreover, relexificationists assume, disputably, that languages of the Ewe-Fon group are structurally identical and that no competition of influence was involved among them.

The least disputed version of the substrate hypothesis is Keesing's (1988), which shows that substrate languages may impose their structural features on a new, contact-induced variety if they are typologically homogeneous, with most of them sharing the relevant features. Thus Melanesian pidgins are like (most of) their substrates in having DUAL/PLURAL and INCLUSIVE/EXCLUSIVE distinctions and in having a transitive marker on the verb. Sankoff and Brown (1976) had shown similar influence with the bracketing of relative clauses with *ia*. However, the pidgins have not inherited all the peculiarities of Melanesian languages. For instance, they do not have their VSO major constituent order, nor do they have much of a numeral classifying system in the combination of *pela* with quantifiers. For an extensive discussion of substrate influence in Atlantic and Indian Ocean Creoles, see Muysken and Smith (1986) and Mufwene (1993).

Competing with the above genetic views has been the 'dialectologist,' or superstrate, hypothesis, according to which the primary, if not the exclusive, sources of Creoles' structural features are nonstandard varieties of their lexifiers. Speaking of AAE, Krapp (1924), for example, claimed that this variety was an archaic retention of the nonstandard speech of low-class Whites with whom the African slaves had been in contact. According to him, African substrate influence was limited to some isolated lexical items such as *goober* 'peanut,' *gumbo*, and *okra*. Since the late 1980s, Shana Poplack and her associates have shown that AAE shares many features with white nonstandard vernaculars in North America and England, thus it has not developed from an erstwhile Creole (see Poplack 1999 for a synthesis). Because some of the same features are also attested in Creoles, we come back to the question of whether most features of Creoles did not after all originate in their lexifiers.

Regarding French Creoles, the dialectologist position was first defended by Faine (1937), according to whom HC was essentially Norman French. This

position was later espoused by Hall (1958), who argues that 'the 'basic' relationship of Creole is with seventeenth-century French, with heavy carry-overs or survivals of African linguistic structure (on a more superficial structural level) from the previous language(s) of the earliest speakers of Negro Pidgin French; its 'lexical' relationship is with nineteenth- and twentieth-century French' (p. 372). Chaudenson (1992, p. 21) is more accommodating to substrate influence as a factor that accounts for the more extensive structural divergence of Creoles from their lexifiers compared with their non-Creole colonial kin.

The 'universalist hypotheses,' which stood as strong contenders in the 1980s and 1990s, have forerunners in the nineteenth century. For instance, Adolfo Coelho (1880–1886) partly anticipated Bickerton's (1984) 'language bioprogram hypothesis' in stating that Creoles' owe their origin to the operation of psychological or physiological laws that are the same everywhere, and not to the influence of the former languages of the people among whom these dialects are found. Bickerton pushed things further in claiming that children made Creoles by fixing the parameters of these new language varieties in the their unmarked, or default, settings as specified in Universal Grammar. To account for cross-Creole structural differences, Bickerton (1984, pp. 176–7) invokes a 'Pidginization Index' (PI) that includes the following factors: the proportion of the native to non-native speakers during the initial stages of colonization, the duration of the early stage, the rate of increase of the slave population after that initial stage, the kind of social contacts between the native speakers of the lexifier and the learners, and whether or not the contact between the two groups continued after the formation of the new language variety. Some nagging questions with Bickerton's position include the following: Is his intuitively sound PI consistent with his creolization qua abrupt pidgin-nativization hypothesis? Is the abrupt creolization hypothesis consistent with the social histories of the territories where classic Creoles developed? How can we explain similarities between abrupt Creoles and expanded pidgins when the stabilization and structural expansion of the latter is not necessarily associated with restructuring by children? Is there convincing evidence for assuming that adult speech is less controlled by Universal Grammar than child language is? How can we account for similarities between abrupt creolization and naturalistic second-language acquisition?

Not all creolists who have invoked universalist explanations have made children critical to the emergence of Creoles. For instance, Sankoff (1979) makes allowance for Universal Grammar to operate in adults, too.

Few creolists subscribe nowadays to one exclusive genetic account, as evidenced by the contributions to Mufwene (1993). The 'complementary hypothesis' (Corne 1999, Mufwene 2001) seems to be an adequate

alternative, provided we can articulate the ecological conditions under which the competing influences (between the substrate and superstrate languages, and within each group) may converge or prevail upon each other. This position was well anticipated by Schuchardt (1909, 1914) in his accounts of the genesis of *Lingua Franca* and of Saramaccan. More and more research is now underway uncovering the socio-historical conditions under which different Creoles have developed, for instance, Arends (1995), Chaudenson (1992, p. 21), Corne (1999), and Mufwene (2001).

Still, the future of research on the development of Creoles has some problems to overcome. So far knowledge of the nonstandard varieties of the lexifiers spoken by the European colonists remains limited. There are few comprehensive descriptions of Creoles' structures, which makes it difficult to determine globally how the competing influences interacted among them and how the features selected from diverse sources became integrated into new systems. Few structural facts have been correlated with the conclusions suggested by the sociohistorical backgrounds of individual Creoles. Other issues remain up in the air, for instance, regarding the markedness model that is the most adequate to account for the selection of features into Creoles' systems. For developmental issues on PCs, the following recently edited collections are good starting points: Muysken and Smith (1986), Mufwene (1993), and Arends et al. (1995). More specific issues may be checked in volumes of the Creole Language Library (John Benjamins) and of Amsterdam Creole Studies, in the *Journal of Pidgin and Creole Languages*, and in *Etudes Créoles*. Several issues of *Pacific Linguistics* also include publications on Melanesian Creoles.

4. *Creolistics and General Linguistics*

There is much more literature on the genesis, sociology, and morphosyntax of PCs than on their phonologies, semantics, and pragmatics. With the exception of time reference and nominal number, studies in semantics and pragmatics are scant. On the other hand, the development of quantitative sociolinguistics owes a lot to research on AAE since the mid-1960s (see, e.g., Labov 1972) and Caribbean English Creoles (e.g., Rickford 1987). Numerous publications in *American Speech*, *Language in Society*, and *Language Variation and Change* reflect this. There are also several surveys of creolistics. Holm (1988, 1989), Arends et al. (1995), and Mühlhäusler (1986, 1997), cited here, are just some of the references, which vary both in adequacy and in geographical areas of focus. It is hoped that, Corne (1999) is the beginning of a new trend of comparative studies of Creoles lexified by the same language.

Studies of structural aspects of Creoles have yet to inform general linguistics beyond the subject matters

of time reference and serial verb constructions. For instance, studies of *lectal continua* have had this potential, but little has been done by creolists to show how their findings may apply to other languages. The mixed nature of *mesolects*, those intermediate varieties combining features of both the *acrolect* and the *basilect* should have informed general linguistics against the fallacy of assuming monolithic grammatical systems. However, little has been done on the subject matter. Likewise, the debate on Creole genesis could have informed historical linguistics on the importance of varying external conditions to language change.

Although lack of consensus among creolists may be invoked as a general reason for this failure to influence general linguistics, alarming indifference from theoretical linguists, especially those engaged in theories of typology and universals, is a more important reason. Consensus cannot be expected of creolistics any more than of other subfields of linguistics or any other scientific discipline. However, in the broader context of language contact (including second-language acquisition), studies of especially Creole genesis have been inspiring. For instance, Thomason and Kaufman (1988) are cited widely in studies of indigenized English. Andersen (1983) was an important step to consolidate common interests between second-language acquisition and Creole genesis. More cross-fertilization might be expected between studies of Creole genesis and those of (child) language development (DeGraff 1999), as among diverse subfields of linguistics.

See also: Archaeology and the History of Languages; Creolization: Sociocultural Aspects; Dialectology; Evolution and Language: Overview; Language and Ethnicity; Language and Society: Cultural Concerns; Language Endangerment; Language Policy: Linguistic Perspectives; Language Policy: Public Policy Perspectives; Melanesia: Sociocultural Aspects

Bibliography

- Adam L 1883 *Les idiomes négro-aryens et malayo-aryens: essai d'hybridologielinguistique*. Maisonneuve, Paris
- Alleyne M C 1971 Acculturation and the cultural matrix of Creolization. In: *Pidginization and Creolization of Languages*. In: Hymes D (ed.) Cambridge University Press, Cambridge, UK, pp. 169–86
- Alleyne M C 1980 *Comparative Afro-American: An Historical-Comparative Study of English-based Afro-American Dialects of the New World*. Karoma, Ann Arbor, MI
- Andersen R (ed.) 1983 *Pidginization and Creolization as Language Acquisition*. Newbury House Publishers, Inc., Rowley, MA
- Arends J (ed.) 1995 *The Early Stages of Creolization*. John Benjamins, Amsterdam
- Arends J, Muysken P, Smith N (eds.) 1995 *Pidgins and Creoles: An Introduction*. John Benjamins, Amsterdam
- Baker P, Mühlhäusler P 1990 From business to pidgin. *Journal of Asian Pacific Communication* 1: 87–115

- Bickerton D 1984 The language bioprogram hypothesis. *Behavioral and Brain Sciences* 7: 173–221
- Bloomfield L 1933 *Language*. Holt, Rinehart and Winston, New York
- Chaudenson R 1992 *Des îles, des hommes, des langues: essais sur la créolisation linguistique et culturelle*. L'Harmattan, Paris. Revised, English edition, *Creolization of Language and Culture*, Routledge, London
- Coelho F A 1880–1886 Os dialectos românicos ou neolatinos na Africa, -Asia, ae America. *Bolletim da Sociedade de Geografia de Lisboa* 2: 129–96 (1880–1881), 3: 451–78 (1882), 6: 705–55 (1886)
- Corne C 1999 *From French to Creole: The Development of New Vernaculars in the French Colonial World*. University of Westminster Press, London
- DeCamp D 1971 Toward a generative analysis of a post-Creole speech continuum. In: Hymes D (ed.) *Pidginization and Creolization of Languages*. Cambridge University Press, Cambridge, UK, pp. 349–70
- DeGraff M (ed.) 1999 *Language Creation and Language Change: Creolization, Diachrony, and Development*. MIT Press, Cambridge, MA
- Faine J 1937 *Philologie créole: études historiques et étymologiques sur la langue créole d'Haïti*. Imprimerie de l'Etat, Port-au-Prince, Haiti
- Hall R A Jr. 1958 Creole languages and genetic relationships. *Word* 14: 367–73
- Hall R A Jr. 1962 The life cycle of pidgin languages. *Lingua* 11: 151–6
- Holm J 1988 *Pidgins and Creoles: Theory and Structure*. Cambridge University Press, Cambridge, UK, Vol. 1
- Holm J 1989 *Pidgins and Creoles: Reference Survey*. Cambridge University Press, Cambridge, UK, Vol. 2
- Keesing R M 1988 *Melanesian Pidgin and the Oceanic Substrate*. Stanford University Press, Stanford
- Krapp G P 1924 The English of the Negro. *The American Mercury* 2: 190–5
- Labov W 1972 *Language in the Inner City: Studies in Black English Vernacular*. University of Pennsylvania Press, Philadelphia
- Lalla B, D'Costa J 1990 *Language in Exile: Three Hundred Years of Jamaican Creole*. University of Alabama Press, Tuscaloosa, AL
- Lefebvre C 1998 *Creole Genesis and the Acquisition of Grammar: The Case of Haitian Creole*. Cambridge University Press, Cambridge, UK
- McWhorter J H 1998 Identifying the Creole prototype: Vindicating a typological class. *Language* 74: 788–818
- Mufwene S S (ed.) 1993 *Africanisms in Afro-American Language Varieties*. University of Georgia Press, Athens, GA
- Mufwene S S 2001 *The Ecology of Language Evolution*. Cambridge University Press, Cambridge, UK
- Mühlhäusler P 1986 *Pidgin and Creole Linguistics*. Basil, New York
- Muysken P, Smith N (eds.) 1986 *Substrata Versus Universals in Creolegenesis*. John Benjamins, Amsterdam
- Poplack S (ed.) 1999 *The English History of African-American English*. Blackwell, Oxford, UK
- Rickford J R 1987 *Dimensions of a Creole Continuum: History, Texts, and Linguistic Analysis of Guyanese Creole*. Stanford University Press, Stanford
- Sankoff G 1979 The genesis of a language. In: Hill K C (ed.) *The Genesis of Language*. Karoma, Ann Arbor, MI, pp. 23–47
- Sankoff G, Brown P 1976 The origins of syntax in discourse: A case study of Tok Pisin relatives. *Language* 52: 631–66
- Schuchardt H 1909 Die Lingua Franca. *Zeitschrift für Romanische Philologie* 33: 441–61
- Schuchardt H 1914 *Die Sprache der Saramakkaneger in Surinam*. Johannes Muller, Amsterdam
- Smith N 1995 An annotated list of pidgins, Creoles, and mixed languages. In: Arends J, Muysken P, Smith N (eds.), pp. 331–74
- Sylvain S 1936 *Le Créole Haïtien: Morphologie et Syntaxe*. Wetteren, Imprimerie De Meester, Belgium
- Thomason S G, Kaufman T 1988 *Language Contact, Creolization, and Genetic Linguistics*. University of California Press, Berkeley
- Turner L D 1949 *Africanisms in the Gullah Dialect*. University of Chicago Press, Chicago

S. Mufwene

Pilgrimage

Pilgrimage involves the movement of a traveler or group of travelers away from home, usually with a specific, sacred goal in mind. In most cases, the term implies a journey to a place that has been sanctified by the presence or actions of a divinity or holy figure. Certain variations on this pattern are possible. Pilgrims sometimes wander continuously without a fixed destination, or the journey may be considered to occur in a purely metaphorical sense, involving spiritual development and transformation within the person. This article focuses, however, on the most common understanding of pilgrimage: literal movement across space that may also result in the crossing of conventional cultural, social, and religious boundaries.

1. Themes in the Study of Pilgrimage

Social scientific studies of pilgrimage were relatively sparse until the 1970s, although in the 1950s Eric Wolf had published a short but influential article on pilgrimage to the Virgin of Guadalupe in Mexico (Wolf 1958). Since the early 1970s, research has expanded and covered all of the world religions. Such academic engagement reflects a continuing and even increasing interest in pilgrimage that has become evident among participants who have better access than ever before to safe and effective modes of travel.

Many earlier studies adopted Durkheimian, functionalist perspectives in analyzing the significance and effects of pilgrimage. In other words, they interpreted the process of traveling to and gathering at a sacred place as providing ritualized opportunities for social unification and moral regeneration within populations that were normally widely dispersed. For instance,

Wolf (1958) argued that the shrine of the Virgin of Guadalupe in Mexico City represented a dominant symbol for the whole nation, condensing various historical, mythical, and cultural themes and focusing them on the activity of pilgrimage to a single place.

The most influential social scientific approach to the subject has come from the anthropologist Victor Turner (see, for instance, Turner and Turner 1978). He notes that, after an original founding event such as a vision, journey, or martyrdom, pilgrimages most frequently involve the gradual organization and routinization of the sacred journey. More speculatively, Turner argues that pilgrimage in the world religions shares certain characteristics with rites of passage in tribal societies. In both forms of ritual, the person is removed from the space and time of the everyday social environment and undergoes an emotionally and physically testing experience that can have lasting effects on perceptions of self-identity. Pilgrims thus enter the realm of what Turner calls social antistructure or *communitas*. He means by these terms that pilgrimage can involve the stripping away of everyday social statuses among participants, resulting in liberation from conformity to conventional norms and hierarchies. Pilgrims ideally cultivate direct and egalitarian bonds of comradeship during the temporary, set-apart time of their journey and dwelling at a shrine. *Communitas* therefore has the potential to contain an element of social critique since it puts the taken-for-granted quality of conventional institutions in question. Representatives of orthodoxy in many of the salvation religions often feel ambivalent towards pilgrimage and attempt either to suppress or control its tendencies to foment lay frustration with religious hierarchies.

Turner notes further that pilgrimage is frequently associated with images of death and rebirth, mirroring the symbolic destruction and reconstruction of the religious traveler's identity. For instance, many who go on the major pilgrimage (or *hajj*) to Mecca will choose to be buried in the white robes they wore on the pilgrimage. It is believed that these people will thus present themselves at the Last Judgment in the garb appropriate to a rejection of worldly, profane actions.

Turner's work has proved to be so significant partly because it appears to provide a vocabulary and theoretical framework that can be generalized and applied to varied cultural contexts. However, numerous subsequent works have criticized aspects of his thesis that pilgrimage is inherently antistructural in form and significance. Authors have pointed out that such a paradigm is possibly more reflective of a Christian ideal of human fellowship than of social and political realities in religious systems in general. Many examples of pilgrimage appear actually to reinforce situations of conflict between participants. For instance, Messerschmidt and Sharma (1981) examine Hindu pilgrimage in Nepal and find that caste and class divisions are not transcended through common

engagement in the ritual. They argue further that conventional Christian ideals of egalitarianism and homogeneity contradict certain Hindu assumptions concerning the inevitable persistence of religious hierarchy.

Powerful and widely cited criticisms of the Turnerian paradigm are contained in a book that is ostensibly devoted to studies of Christian pilgrimage but presents arguments that have much wider application. Eade and Sallnow (1991) characterize pilgrimage sites as ideal locations for contesting, rather than consolidating, the sacred among religious participants. As centers of power that can provide religious legitimation for political aspirations, shrines often present powerful incentives for a particular interest group to attempt to control the behavior and beliefs of religiously motivated visitors (the point is illustrated clearly by the politically charged and often violent history of Jerusalem). Pilgrimages, according to this view, provide the material and ideological focus for competing discourses and agendas, not only between sacred and secular authorities, but also between lay pilgrims and clerics, locals and incomers, members of different denominations, and so on. Sacred sites are described by Eade and Sallnow as the equivalent of empty symbolic vessels waiting to be filled by the possibly mutually contradictory assumptions of the various people who visit them. The authors also suggest that to regard pilgrimage as a unitary institution is highly problematic, even within a single religious system, since it can contain so many historically specific behaviors and meanings. They add, however, that Christian and perhaps other forms of pilgrimage can be conceptualized as involving a basic triad of 'person,' 'place,' and 'text.'

In fact, Turner did anticipate at least some of the later criticisms that have been leveled at his model of pilgrimage. He was aware of the potential for pilgrimage sites to be co-opted by political elites, the fragility of states of *communitas*, and the capacity of pilgrimage sites to encourage widely differing interpretations from pilgrims. However, the notion that pilgrimage should be seen as an arena for contestation as much as cooperation has proved popular in recent years. Its adoption by many scholars is not surprising since it resonates with contemporary social scientific perspectives that have been influenced by postmodern understandings of culture as unstable, fragmentary, and open to multiple appropriation.

A less controversial set of issues, discussed by Turner and others, is raised by the contention that pilgrimage can encourage participants to consider themselves part of translocal religious and cultural systems. Among Hindus, for instance, a sense of participation in a sacred geography that covers the whole of the Indian subcontinent can be cultivated by travel to shrines. Some devotees choose to circumambulate the entire country, visiting sacred spots along the way. Holy places themselves may be officially

or unofficially ranked in importance, thus constituting an internally heterogeneous but widely dispersed framework of sites. Among Christians, Jerusalem (located within 'the Holy Land' itself) has often been regarded as the pre-eminent location for pilgrimage, containing as it does so many sites associated with biblical events. It is followed in importance by other locations of transnational importance, such as Rome or Compostela, and then by pilgrimage locations that have significance in national or more narrowly defined regional contexts.

2. Methods of Study

Pilgrimage raises unusual problems in relation to conventional social scientific methodologies. Any given site is located within a much wider political, economic, and religious framework that must be taken into account in analysis. As a religious practice, pilgrimage has often complemented (and incorporated) other translocal activities like trade, tourism, and exploration. Pilgrims themselves may move across geographical, social, and religious boundaries, so that the scholar is faced with a diffuse and literally shifting object of study.

Various methods can be used to encapsulate the most important aspects of the phenomenon. Pilgrims can be followed as they journey to a shrine (Gold 1988, Frey 1998). More commonly, a single pilgrimage center or shrine is focused upon, and the activities of various pilgrim groups observed as they pass through the sacred center (Dubisch 1995). At a wider, macro level of analysis, whole systems of pilgrimage within a religious tradition may be studied (Turner and Turner 1978), or comparisons made between pilgrimage systems in the various world religions (Coleman and Elsner 1995).

The fact that pilgrimage sites tend to persist over time provides a good opportunity to combine historical with sociological interests (Turner and Turner 1978), and much work in this area demonstrates the ways in which a given site can retain its identity as a sacred place despite the fact that it is appropriated by different and sometimes competing religious systems over perhaps hundreds of years. Guadalupe, for instance, provides a good example of a shrine that has probable pre-Christian as well as Christian roots. The mixture of religious associations increases its significance as a site whose symbolism and mythology can be invoked in current debates concerning the relationship between precolonial, colonial, and post-colonial identities in Mexico.

3. Future Themes of Research

Pilgrimage has the potential to be at the center of much future research that could be of importance to the social sciences in general. As a social phenomenon,

it encourages the researcher to engage with some key theoretical debates in a variety of disciplines. The current popularity of pilgrimage as an activity raises the question of its significance for predictions of secularization and/or religious revitalization in various parts of the world. Some scholars have also speculated over the extent to which apparently secular activities such as going on holiday or visiting the graves of pop stars share the same structural features as pilgrimage. Certainly, it seems likely that tourism and pilgrimage will become increasingly intertwined in many cultural contexts in the future, given the popularity and flexibility of both activities. In addition, pilgrimage studies are relevant to ongoing discussions of the significance of processes of globalization in the modern world. Studies of sacralized forms of travel, involving the transference of people, ideas, and objects across national borders, can complement other research that is concerned to understand how social, cultural, and political units are being transformed by developments in technologies of mass travel and communication. Furthermore, in a world that is characterized by the displacement of populations from traditional homelands, pilgrimage may emerge as an increasingly important institutional means through which to unite diasporic communities. Tweed (1997) has recently argued, for instance, that pilgrimage allows Cuban exiles in Florida to map the landscape and history of their country of birth on to a new environment. Finally, pilgrimage provides many possible avenues of research in the light of increased social scientific interest in the body and embodiment, ranging from a focus on beliefs in miraculous healing to analysis of forms of movement associated with travel to and around sacred shrines.

See also: Migration, Theory of; Monasticism: West and East; Traditions: Social

Bibliography

- Coleman S, Elsner J 1995 *Pilgrimage: Past and Present in the World Religions*. Harvard University Press, Cambridge, MA
- Dubisch J 1995 *In a Different Place: Pilgrimage, Gender, and Politics of a Greek Island Shrine*. Princeton University Press, Princeton, NJ
- Eade J, Sallnow M J (eds.) 1991 *Contesting the Sacred: The Anthropology of Christian Pilgrimage*. Routledge, London
- Frey N L 1998 *Pilgrim Stories: On and Off the Road to Santiago*. University of California Press, Berkeley, CA
- Gold A G 1988 *Fruitful Journeys: The Ways of Rajasthani Pilgrims*. University of California Press, Berkeley, CA
- Morinis A (ed.) 1992 *Sacred Journeys: The Anthropology of Pilgrimage*. Greenwood Press, Westport, CT
- Messerschmidt D A, Sharma J 1981 Hindu pilgrimage in the Nepal Himalayas. *Current Anthropology* 22: 571–2
- Turner V, Turner E 1978 *Image and Pilgrimage in Christian Culture*. Columbia University Press, New York
- Tweed T A 1997 *Our Lady of the Exile: Diasporic Religion at a Cuban Catholic Shrine in Miami*. Oxford University Press, Oxford, UK

Wolf E 1958 The Virgin of Guadalupe: A Mexican national symbol. *Journal of American Folklore* 7(1): 34–9

S. Coleman

Place in Geography

‘Place’ is vivid in experience, yet elusive as a concept. As a general term it embraces an aspect of the diversity of the world that is necessarily experienced though countless unique places. In spite of this ambiguity, ‘place’ in geography was considered for a long time to be more or less synonymous with ‘region’ or ‘settlement’ and so obvious that it needed no clarification. In the last half of the twentieth century this apparent obviousness has been brought into question as ideas about ‘place’ have been subjected to critical theoretical interpretation, and as changes in the character of actual ‘places’ have suppressed or distorted diversity. This article describes this sequence of change and the relationships between them.

1. ‘Place’ and ‘Places’

Aristotle, in his *Physics*, noted that ‘The question, What is Place?, presents many difficulties for analysis,’ then concluded that it refers to the precise dimensions of the space that contains something—‘place’ is a neutral container (McKeon 1941). This is a flexible but not very edifying definition since a container can range in scale from the space on my desk where I keep my dictionary, to Canary Wharf, to Paris, to Australia, to the earth as the place of humankind.

In contrast, Strabo (1917) wrote about 5 BC in the introduction to his *Geography* that ‘A knowledge of places is conducive of virtue.’ He believed that geographers, with their habits of travel and knowledge of the specific parts of the world, were qualified to distinguish good settings for human activity from bad ones, and by implication that content was more important than the container. The difficult question of ‘place’ faded into the background for almost 2,000 years, and, while it became acceptable to describe geography as ‘the study of places’ (Hartshorne 1959), this was left unexplained because ‘a place’ was apparently an unproblematic concept. The issue of scale was left, and remains, unresolved.

So ‘place’ has an enigmatic character, partly general and partly specific, midway between objective fact and subjective feeling. Entrikin (1991) calls this ‘the betweenness of place’ and proposes that it is best viewed from a point in the middle. From this perspective, ‘a place’ appears as a combination of objects and meanings that distinguishes somewhere from its surroundings, regardless of scale. For Walter (1988) it is

a ‘location of experience; the container of shapes, powers, feelings and meanings,’ and for Cresswell (1996) it is ‘a meaningful segment of space’ and a ‘container of social power’ (see *Geographic Thought, History of; Region: Geographical; Regional Identity; Scale in Geography*).

2. Experience and Sense of Place

In the 1960s a new theoretical and scientific approach to the spatial geometry of location, known as central place theory, served as a catalyst for more rigorous thinking about ‘place’ (see *Location Theory*). This approach considered towns and villages in terms of measurable properties and relative locations, and thus implicitly adopted an Aristotelian perspective that emptied places of content.

Geographers interested in specific places were dismayed by the aridity of central place theory but recognized that it was impossible to continue to use the concept of ‘place’ uncritically. Phenomenology was identified as a philosophical method that allows rigorous interpretations of place without denying diversity and meaning. The writings of M. Heidegger and M. Merleau-Ponty were used to argue that ideas of place are rooted in experiences of dwelling in specific places, and that these precede all abstractions about location, environment, or geography. Places are aspects of everyday experience, territories of symbols, memories, and associations; they are centers of being and belonging that connect people with their worlds.

Phenomenology discloses several elements of ‘place’ and place experience, including identity and sense of place, topophilia, and home. These insights have been absorbed into most subsequent discussions of place in geography, first in humanistic approaches and more recently in arguments which view place as a social construction (see *Humanistic Geography*).

2.1 Identity and Spirit of Place

Identity is what makes somewhere distinctive; it also suggests similarities, as when it is claimed that a particular fast-food restaurant is identical with one elsewhere. The term ‘identity of place,’ therefore, acknowledges that every place simultaneously differs from and has similarities with other places. This apparent contradiction requires clarification.

From an abstract perspective, it is possible to distinguish four essential components of place identity: location; an appearance of hills, rivers, buildings, and other objects; a community engaged in various social and economic activities; and a set of meanings and symbols. These are distinct and are not reducible one to the other (one cannot discuss location in terms of architecture, and so on), but these components are, in real places, always closely interwoven. Oaxaca in

southern Mexico is a colonial city that shares its architectural tradition and town plan with numerous cities in Latin America, and there is nothing unusual about its role as a center of regional government or its passion for fiestas. However, its location is unique, as are its elegant streets and buildings, and it seems unlikely that anywhere else has a annual festival—La Noche de los Rabanos—dedicated to carved radishes. In all places, even unremarkable ones such as new suburbs, some aspects are shared and others are distinctive, though the balance of these can vary widely. Even if appearance and activities are ordinary, meanings and symbols are always in some way distinctive.

Somewhere with exceptional place identity, like Oaxaca, is sometimes said to possess ‘spirit of place.’ This expression comes from the Latin *genius loci*, but the idea is derived from the ancient and widespread belief that certain places are inhabited by their own gods or spirits. In ancient Greece almost every hill-top, spring, grove, and outcrop of rock had its own guardian spirit; the Acropolis was the home of the goddess Athena. In this more secular age ‘spirit of place’ has lost its association with gods, and now refers to the immediately apparent yet rather elusive aesthetic quality of striking distinctiveness. Spirit of place is what makes particular places attractive and memorable.

2.2 Sense of Place

Sense of place is the capacity to recognize and respond to diverse identities of places. A place affects us directly through sight, smell, sound, and touch and sense of place is the synthetic faculty that combines these impressions with memory, imagination, and reason. (Note, however, that ‘sense of place’ is popularly and confusingly used to be synonymous with ‘spirit of place’; thus, it might be said that ‘Provence has a wonderful sense of place’).

Sense of place is not a uniformly shared and undifferentiated faculty. Many people are uninterested in places, some have an instrumental view of the world that regards places as resources to be exploited, while others are finely tuned to *genius loci* and place diversity. These variations have been clarified by making a distinction between an outsider’s sense of place, which looks at the world mostly in terms of objective qualities, and an insider’s sense of place, which seeks to grasp the distinctive properties of particular places in terms of their meanings for those who live and work in them. An outsider’s perspective pays scant attention to distinctiveness and is inclined to impose judgments on the basis of previously formed opinions; it is manifest, for example, in colonialism and in the propagation of placelessness (see Sect. 3.1). In contrast, an insider’s sense of place is responsive to the idiosyncracies of environment and culture. While

it comes with knowing and being known somewhere, and is most apparent in home places, it is also an ability that can be learned or enhanced by developing skills of careful observation and imagination.

2.3 Topophilia and Home

‘Topophilia’ lies at the intersection of spirit and sense of place. Literally it means ‘love of place,’ though Tuan (1974), who has given the term currency, defines it as ‘the affective bond between people and place or setting.’ There are, he suggests, ‘landscapes of persistent appeal,’ such as mountains and seashores, islands, carefully tended farms, and small towns. These are perceived through lenses of culture, familiarity, memories, even personal moods. Topophilia blends the inherent qualities of landscapes and places with the attitudes that different individuals bring to it. There are no fixed relationships in this. Sometimes a beautiful setting can be soured by ill-temper, yet at other times that same beauty may displace ill-temper.

Underlying an insider’s sense of place and topophilia is the notion and experience of ‘home.’ It is in ‘home’ where the significance of individual lives is most concentrated, where people feel they belong, where love of place is strongest. Home is a powerful emotional bond that can, for example, draw back refugees and those who have been uprooted in spite of the most adverse circumstances. Because of the intensity of these associations, in phenomenological interpretations ‘home’ is identified as the archetypal place.

2.4 Corrupted Sense of Place

Phenomenological interpretations mostly have been concerned with clarifying the subjectivity of place experience. In emphasizing roots and home they imply that ‘place’ and ‘sense of place’ are positive and provide the foundation for a deeply rooted yet balanced outlook on the world. This is not always the case. Harvey (1989) argues that consequences of a strong sense of place include sectarian politics and parochialism. This is apparent in local and neighborhood practices that resist anything which might upset racial homogeneity or promote social equity (see *Locational Conflict (NIMBY)*). There is a fine and unclear line between such intolerance and efforts to protect a place from massive redevelopment and equivalent changes that might destroy its identity, but when that line is crossed sense of place can be said to have become corrupted. A corrupted or poisoned sense of place selectively stresses local identity and home to the virtual exclusion of shared values (Relph 1997). It is always unpleasant, and at its worst, in xenophobia and ethnic cleansing, it can be brutally violent.

3. *Modernist and Postmodernist Changes in 'Places'*

At the beginning of the twentieth century most peoples' lives were, as they always had been, largely confined to within about one day's walk of where they lived. Nowhere was entirely isolated from outside contact, but with lack of mobility social changes dispersed slowly and usually adapted to specific contexts, so the local distinctiveness and geographical diversity were continually reinforced. This diversity was created under social and technological conditions dramatically different from those which now prevail. Since the mid-nineteenth century technologies of electronic communications and mass travel, coupled with rapid cultural and economic changes, have collapsed distance. The ways places are made, and the ways they are experienced, have both changed. Old places may be regarded with great affection, but they are historical fragments; geographies that are being made now consist mostly of placeless or arbitrarily fabricated settings.

3.1 *Placelessness*

Placelessness is the condition in which different places look much the same, and, more importantly, offer the same opportunities for experience. It is the erosion of geographical distinctiveness and diversity. Examples are legion: shopping centers, apartment buildings, airport terminals, expressways. Augé (1995) calls these 'nonplaces.' The new high-rise Okura Hotel in Kobe in Japan could equally well be in Houston or Mexico City and the adjacent shopping center is so dislocated from its context that much of it only has signs in English.

In part, placelessness is the expression of modernist styles of architecture and planning that were conceived in the 1920s to take advantage of new building materials of glass, steel, and concrete. These styles are functional, standardized, and international. They respond to the essential sameness of people everywhere and can be applied anywhere, regardless of climate or culture. Placelessness is also the child of transnational corporations and a global economy that strive to provide familiar goods and services to consumers in the most profitable and efficient way possible. Modernist design and corporate capitalism fit together well. The result is that the same products, the same companies, similar buildings, and similar landscapes can be found almost everywhere.

3.2 *Place Exploitation and Postmodernism*

Since about 1980 placelessness has developed a more subtle and complex character. As uniformity replaces diversity, distinctiveness becomes increasingly valuable and worth maintaining. Sometimes, as in heritage

preservation, the aim may be to protect intrinsic local identity, but the larger consequence has been to turn place identity into a commodity. Distinctiveness sells. In tourism, historical districts and unique landscape features are obviously attractive place qualities. Less obviously, in a global economy in which capital is mobile and businesses can locate anywhere with educated workers and reasonable land costs, a distinctive place identity can be decisive factor for a city trying to attract investment.

Once place identity is regarded as a commodity rather than an intrinsic property of somewhere, it can be exploited in many ways to promote tourism and attract other business. It can be selectively enhanced either by reconstructing destroyed buildings or historical districts, as in Place Royale in Quebec City, or by converting former industrial areas into festival markets of restaurants and galleries, as in Covent Garden in London. Links with celebrities or events, real or fictional, can promote marketability, so Atlanta is associated with *Gone with the Wind*; and Wigan has a Wigan Pier shopping center with an Orwell restaurant. If local identity is too bland elements of more distinctive places can be copied. Las Vegas, which excels in place plagiarism, has design fragments of Egypt, Italy, the tropics, New York, Ireland, and Japan in hotels and casinos on its main strip. More restrained copying is to be found in the restaurants, shopping centers, and entertainment districts of most cities, so gondolas ply the otherwise placeless waterfront in Toronto.

The exploitation and fabrication of place identity erodes distinctions between local history and contrived heritage, between fiction and reality, in order to allow capital to pursue its relentless aims. Harvey (1989) argues that these are processes of postmodernity. Their consequence is a placelessness achieved not through uniformity, but through what Sorkin (1992) calls 'ageographical variety.' Place identities have been uprooted, turned into transportable goods, and rearranged without regard for context.

4. *A Progressive or Critical Sense of Place*

These changes in the character of actual places raise questions about the notion of 'sense of place.' Phenomenological accounts imply that sense of place is constant across space and time and unaffected by electronic communications and increased travel. A different interpretation, based in critical social theory and political economy, proposes what Massey (1993) calls 'a progressive sense of place.' Instead of viewing places as bounded settings rooted in home and local traditions, a progressive sense of place looks outward from a specific geographical context to connections with wider processes. It sees each place as a unique point of interaction within networks of social and economic relations. From this perspective places appear as distinctive social spaces of inclusion or

exclusion, within a broad and changeable geographical context (Cresswell 1996).

A progressive or critical sense of place offers a possible foundation, that is neither nostalgic nor exclusionary, for understanding recent social and urban changes (Harvey 1996). It recognizes the global processes at work in place exploitation and, therefore, can provide insights into cultural 'displacements' and unanticipated juxtapositions of mosques, Chinatowns, and protestant churches that are the consequence of the intercontinental population movements of the last 50 years and which have turned many cities and towns into multicultural and multicentred places (see *Place, Power of*).

5. Place-making

As a concept, 'place' has evolved rapidly from a background, neutral term to one that is an important theoretical and empirical subject not only in geography, but for many of the social sciences (Walter 1988, Augé 1995, Casey 1997). A simultaneous, possibly coincidental shift from distinctiveness to placelessness and ageographical variety has occurred in the identity of actual 'places.' A repeated theme in recent discussions of these changes in 'place' and 'places,' regardless of discipline and regardless of whether the approach is theoretical, empirical, or phenomenological, is the need for individuals and local communities to be involved in making and maintaining their own places. Places have ceased to be objects for description and have come to be regarded as complex theoretical phenomena, as foundations for social resistance against remotely generated economic forces, and as centers of stability in world of change.

See also: Geography; Place, Power of; Postmodernism in Geography; Urban Geography

Bibliography

- Augé M 1995 *Non-places: Introduction to the Anthropology of Super-modernity*. (trans. Howe J) Verso, London
- Casey E S 1997 *The Fate of Place: A Philosophical History*. University of California Press, Berkeley, CA
- Cresswell T 1996 *In Place/Out of Place: Geography, Ideology and Transgression*. University of Minnesota Press, Minneapolis, MN
- Entrikin J N 1991 *The Betweenness of Place: Towards a Geography of Modernity*. The Johns Hopkins University Press, Baltimore, MD
- Hartshorne R 1959 *Perspectives on the Nature of Geography*. Rand McNally, Chicago
- Harvey D 1989 *The Condition of Postmodernity*. Blackwell, Oxford, UK
- Harvey D 1996 *Justice, Nature and the Geography of Difference*. Blackwell, Oxford, UK
- Massey D 1993 Power-geometry and a progressive sense of place. In: Bird J, Curtis B, Putnam T, Robertson G, Tickner L

- (eds.) *Mapping the Futures: Local Cultures, Global Change*. Routledge, London and New York
- McKeon R (ed.) 1941 *Aristotle: Basic Works*. Random House, New York
- Relf E 1976 *Place and Placelessness*. Pion, London
- Relf E 1997 Sense of place. In: Hanson S (ed.) *Ten Geographical Ideas that Changed the World*. Rutgers University Press, New Brunswick, NJ
- Sorkin M (ed.) 1992 *Variations on a Theme Park: The New American City and the End of Public Space*. Hill and Wang, New York
- Strabo 1917 *The Geography of Strabo*, 8 Vols. (trans. Jones H L). Heinemann, London
- Tuan Y-F 1974 *Topophilia*. Prentice-Hall, Englewood Cliffs, NJ
- Walter E V 1988 *Placeways: A Theory of the Human Environment*. University of North Carolina Press, Chapel Hill, NC

E. Relf

Place, Power of

The 'power of place' refers to the spatial form of landscapes, towns, and buildings as an important aspect of social and political life. The phrase is used in various contexts: analyzing social life in existing places, advocating the preservation of historic places, and formulating goals for the design of new places. As a phrase, 'power of place' combines the meanings of location, locale, and 'sense of place' (Agnew and Duncan 1989). It involves an approach to aesthetics drawn from work dealing with 'local distinctiveness' (Clifford and King 1990) in architecture, landscape architecture, cultural geography, and environmental psychology, with an approach to politics drawn from work on space in the social sciences (Hayden 1995). Its practical applications in urban planning include preservation and urban design.

1. Definitions of Place and Power

'Place' is one of the trickiest words in the English language. It carries the resonance of homestead, location, and open space in the city as well as a position in a social hierarchy. The authors of books on architecture, photography, cultural geography, poetry, and travel rely on 'sense of place' as an aesthetic concept but often settle for 'town character,' 'local distinctiveness,' or 'the personality of a location' as a way of defining it. In the nineteenth century and earlier, place also carried a sense of the right of a person to own a piece of land or to be a part of a social world, and in this older sense, place carries more political history with it. Phrases like 'knowing one's place' or 'a woman's place' still carry some territorial meanings and imply political power relationships (see *Community Power Structure*).

1.1 Place and Social Life

People experience places in ways that are critical to their well-being or distress. An individual's sense of place is both a biological response to the surrounding physical environment and a cultural creation. From childhood, humans come to know places through engaging all five senses: sight, sound, smell, taste, and touch. Research on perception shows the simultaneous engagement of several senses in orientation and wayfinding. Children show an interest in landmarks at three or earlier, and by age five or six can read aerial maps with great accuracy and confidence, as an example of the human ability to perceive and remember places (Tuan 1977).

As social relationships are intertwined with spatial perception, the subject of human attachments to places, as well as the ability to conceptualize places, is a fascinating one that defies simplistic research models in any field. Environmental psychologists define 'place attachment' as having social, material, and ideological dimensions (Altman and Low 1992). Some sociological studies of urban renewal, especially those which study the process of mourning for a lost neighborhood, have utilized attachment theory as well (Marris 1986). Many sociologists speak of 'locale' as part of fieldwork; anthropologists deal with the design of settlements and living space as part of the material culture of any group. Studies in the cognitive mapping of places often reveal the social dimensions of knowledge of the city. If place does provide an overload of possible meanings and methods of inquiry for the researcher, philosophers have argued that place's very same assault on all ways of knowing (sight, sound, smell, touch, and taste) make it powerful as a way of structuring social attachments and as a container of memory. Thus the concept of *lieux de memoire* (sites of memory, or places of memory) popularized by Pierre Nora in France, and expanded as processes of remembrance (Hayden 1998).

1.2 Place as Cultural Landscape

Cultural landscape studies, as geographer Carl Sauer and the 'Berkeley School' developed them from the 1940s on, focused on the evolution of places and included the 'combination of natural and man-made elements that comprises, at any given time, the essential character of a place' (see *Landscape Architecture*). Cultural landscape studies emphasized the connection of natural environments and built interventions. Yet cultural landscape methods for studying places, and people's shaping of them, tended to stress the physical and not the political dimensions of places, leaning to the study of rural, pre-industrial landscapes, vernacular house types, and patterns of cultivation, considering ecology but avoiding issues of political contestation. Among the noted essayists of this type was John Brinckerhoff Jackson (1997).

1.3 Places and Power Struggles

As the landscape is more densely inhabited, economic and social forces affecting it are more complex, change is rapid, layers proliferate, and often abrupt spatial discontinuities result. The shaping of urban places involves gaining political access to territory as well as owning, planning, designing, constructing, using, and demolishing various kinds of built space. One can't simply turn to work on location in economic geography (or any other kind of quantitative analysis) because there the human experience of place is an essential topic that is often lost. Rather, the cultural landscape model has been more firmly connected to urban sociology, anthropology, and political science, retaining the physical analysis necessary to convey the shape of urban places, while adding more focused analysis of social and economic contestation (Duncan and Ley 1993, Groth and Bressi 1997). Historians are also claiming this subject matter: 'The story of how different peoples have lived and used the natural world will become one of the most basic and fundamental narratives in all of history, without which no understanding of the past could be complete' (Cronon 1992).

It has often proved easier to study either the natural or the built components of an urban cultural landscape than to wrestle with the combination of the two, fusing analyses of place and power. For many environmental historians, the deployment of land and natural resources has been a central preoccupation. Urban historians have focused more on power struggles; architectural historians on styles and patterns of building. The phrase, 'power of place,' suggests that all three areas are intertwined in some of the spatial struggles of urban development. The aesthetic qualities of architecture or of the built environment cannot be divorced from those of the natural environment without strange dislocations of experience: expensive houses swathed in smog; skyscrapers built on unstable soils, freeways blocking river fronts.

1.4 Avoidance of 'Place'

In the last decades of the twentieth century, social scientists frequently avoided 'place' as a concept, and thus sidetracked the sensory, aesthetic, and environmental components of the urbanized world, in favor of more quantifiable research with fewer epistemological problems (Agnew and Duncan 1989). Melvin Webber used 'non-place urban realm' to refer to the decline of face-to-face activities and the rise of telephone and television. Some analysts have extended this by arguing for the importance of an increasingly 'placeless world' (Relph 1976) or 'non-places' (Augé 1995) dominated by 'organization space' (Easterling 1999). Others prefer to retain the concept of place, and critique unpleasant, unlivable places (Hayden 1995). A paved-over neighborhood, restructured as freeway

lanes, parking lots, and mall, should still be considered a place, if only to explain why the new space lacks the pedestrian scale, lively texture and local meanings of the older place.

2. *Connections between Place and Space*

Sociologists' analyses of the 'production of space' provide a framework that can be used to relate the 'power of place' to the political economy. Lefebvre (1991) argues that every society in history has shaped a distinctive social space that meets its intertwined requirements for economic production and social reproduction. In terms of production, this would be close to cultural geography in identifying spaces or landscapes shaped for mining, manufacturing, commerce, or real estate speculation. Most original is his analysis of the space of social reproduction, which ranges over different scales, including the space in and around the body (biological reproduction), the space of housing (the reproduction of the labor force), and the public space of the city (the reproduction of social relations). Here he links the physical to the social in decisive ways. Cultural critics praised Lefebvre: he 'called for a new kind of spatial imagination capable of confronting the past in a new way and reading its less tangible secrets off the template of its spatial structures—body, cosmos, city' (Jameson 1991). His analysis also fueled the 'non-place' debate: a quality of late capitalist space is the creation of many identical units—similar but not 'placeless' places—by the large commercial real estate market which has become, in itself, a distinguishing feature of the global economy.

The cultural costs in terms of identity, history, and meaning have also been weighed, and here the phrase 'power of place' is used again. In a world increasingly given over to global corporations and their large, box-like buildings unrelated to local landscapes, the importance of local places as icons of resistance increases. Place-based social movements may resist demolition, preserve historic places, or demand place-sensitive urban design and architecture.

3. *Place-based Social Movements*

One of the consistent ways to limit the economic and political rights of groups has been to constrain social life through limiting access to places. The interplay between the social and the spatial is constant. The women's rights movement of the late nineteenth and early twentieth centuries was a struggle against spatial segregation based on gender. Women were not allowed to walk freely in public, obtain education, or vote. So too, the civil rights movement of the 1960s emphasized overcoming spatial segregation. Jim Crow laws restricted the African American community: streets, neighborhoods, schools, colleges, hotels, stores, fire

stations, swimming pools, and cemeteries were all segregated. To understand the intersecting struggles of race, class, and gender, the spatial dimensions of traditional 'woman's sphere' would have to be studied in combination with the spatial limits imposed by race or class. In the twentieth century, urban renewal often attempted to target poor neighborhoods and those inhabited by people of color. Resistance to demolition was a place-based campaign.

3.1 *The Power of Place, Los Angeles*

In Los Angeles, The Power of Place, a non-profit corporation, was founded by urban historian and architect Dolores Hayden. Operating from 1984–91, it documented and interpreted the political dimensions of place as historical experience, with a special concern for gender, class, and ethnicity. One premise of this project was that architects, planners, and artists, as well as specialists in preservation, have an important role to play in making the entire city more comprehensible, and its history more visible, in the minds of citizens. The group traced an itinerary of work places in downtown connected to the growth of the city: citrus farms, flower farms, oil wells, garment factories, and other economic activities between 1780 and 1940. On this itinerary, buildings were preserved and interpreted, and new public art and open space designs were created (Hayden 1995). Similar projects were organized in London by a group called Common Ground (Clifford and King 1990) and continue to be developed in many other American cities.

3.2 *Civil Rights Museums*

It is possible to identify historic urban places which have special significance to certain populations fighting spatial segregation of different kinds: a church where major civil rights meetings were held; a local newspaper that crusaded for fair housing; the first place in a city where women tried to vote; the first place where new immigrants were allowed to own land; the first integrated primary school. The Women's Rights National Park in Seneca Falls, New York is one such museum. There are also sites of assassinations, lynchings, massacres, and riots. The motel where Martin Luther King was shot, the Lorraine Motel in Memphis, Tennessee, is now a national civil rights museum, but every city and town has similar landmarks where territorial struggles have been waged.

3.3 *The Lower East Side Tenement Museum*

Another kind of museum to exploit the power of place is the Lower East Side Tenement Museum in New York. Ordinary dwellings contain the physical evi-

dence necessary to analyze the material conditions of ordinary workers' lives in earlier times. Standing inside a tenement apartment—perhaps four hundred square feet of living space for an entire family, minimal plumbing, only one or two exterior windows—leaves a visitor gasping for air and looking for light. The claustrophobic experiences of immigrants living for decades in crowded, unhealthy space are conveyed by the building in a way that a text or a chart can never match. Because a tenement was typical of hundreds of thousands of tenements all over the city, its implications for the historic urban landscape are broad. One building tells the bitter story of thousands of developers and millions of tenants. When this museum was acquired by the National Trust for Historic Preservation in the late 1990s, it signaled new prestige for social history museums everywhere.

3.4 *Parades and the Power of Place*

Festivals and parades also help to define the power of place by staking out routes in the urban cultural landscape. They intermix vernacular arts traditions (in their costumes, floats, music, dance, and performance) with spatial history (sites where parades begin, march, and end). African American jazz funerals marching through the streets of New Orleans, Chinese New Year parades, saint's day processions in Irish American or Italian American Catholic communities, and graveyard ceremonies for the Day of the Dead in Mexican American communities are just a few examples. Political parades have been representing communities and their causes for a long time, from workers with the tools of their trades to suffragettes. Since the 1960s, civil rights marches in Southern cities, women's marches to 'Take Back the Night,' and Gay and Lesbian Pride Parades in major cities have also been part of campaigns to achieve greater political representation.

4. *The Power of Place—Applications in Urban Design*

Some recent books proclaim the imminent end of urban public places. Advances in computer technology have affected many parts of daily life and often dispersed neighborhood activities. 'Disneyfication' fills theme parks and casinos with entertainments built to resemble older towns. Shopping malls take up historicist built forms as well. The global migration of capital, and the resultant production of 'organization space' emphasizes freeway interchanges or airports as locational determinants, eroding the importance of older central cities. Virtual reality promises to simulate more and more spatial experiences.

Nevertheless, human beings still situate their lives and work in material spaces of different scales. There

are many signs that places are more important, and more contested, than ever. Numerous older cities and towns have been protected by a growing movement for historic preservation. New town design is frequently done in accordance with the architectural movement promoting pedestrian scale and known as the 'New Urbanism.' Similarly, planners involved with campaigns for 'smart growth' have been heartened to hear that most Americans ranked urban and suburban sprawl ('placelessness') as an important political issue in 2000.

Studies exploring the 'power of place' are rapidly increasing in academic research, although the authors are scattered in half a dozen fields—history, geography, architecture, urban planning, landscape architecture, art history, anthropology, sociology, and others. In the 1980s, 'place' was not quantitative enough for the mainstream of social science, or narrowly aesthetic enough for the mainstream of art history and design. At the start of the twenty-first century, scholarly work on the broad cultural and political importance of places is part of many interdisciplinary programs such as cultural studies or American studies, as well as all of the separate disciplines. Work on the 'power of place' also has much to offer citizens and local public officials. Scholars and practitioners take discussion of 'the power of place' and other public arts to public forums in cities and towns (Imagining America 2000). Understanding the creation of places as a social and political process can lead to new projects to extend public meaning in the urban cultural landscape.

See also: Cultural Landscape in Environmental Studies; Cultural Landscape in Geography; Land Use and Cover Change; Location Theory; Locational Conflict (NIMBY); Political Geography; Power in Society; Social Movements, Geography of; Spatial Equity; Spatial Interaction; Spatial Interaction Models; Spatial Optimization Models; Spatial Pattern, Analysis of

Bibliography

- Agnew J A, Duncan J S 1989 *The Power of Place: Bringing Together Geographical and Sociological Imaginations*. Unwin Hyman, Boston, MA
Altman I, Low S (eds.) 1992 *Place Attachment*. Plenum Publishing, New York
Augé M 1995 *Non-places: Introduction to an Anthropology of Supermodernity*. Verso, London
Conzen M (ed.) 1990 *The Making of the American Landscape*. Unwin Hyman, Boston, MA
Clifford S, King A 1990 *Local Distinctiveness*. Common Ground, London
Cronon W 1992 A place for stories: Nature, history, and narrative. *Journal of American History* 78: 1347–76
Duncan J, Ley D (eds.) 1993 *Place/Culture/Representation*. Routledge, London

- Easterling K 1999 *Organization Space*. MIT Press, Cambridge, MA
- Groth P, Bressi T (eds.) 1997 *Understanding Ordinary Landscapes*. Yale University Press, New Haven, CT
- Hayden D 1995 *The Power of Place: Urban Landscapes as Public History*. MIT Press, Cambridge, MA
- Hayden D 1998 Landscapes of loss and remembrance: The case of Little Tokyo in Los Angeles. In: Winter J, Sivan E (eds.) *War and Remembrance*. Cambridge University Press, Cambridge, UK
- Hayden D, Dubrow G, Flynn C 1984 *The Power of Place: Los Angeles*. The Power of Place, Los Angeles
- Imagining America: Artists and Scholars in Public Life* 2000 <http://www.ia.umich.edu/>
- Jackson J B 1997 *Landscape in Sight: Looking at America*. Yale University Press, New Haven, CT
- Jameson F 1991 *Postmodernism, or, The Cultural Logic of Late Capitalism*. Duke University Press, Durham, NC
- Lefebvre H 1991 *The Production of Space*. Basil Blackwell, Oxford, UK
- Marris P H 1986 *Loss and Change*. 2nd edn. Routledge & Kegan Paul, London
- Ralph E 1976 *Place and Placelessness*. Pion, London
- Tuan Y-F 1977 *Space and Place: The Perspective of Experience*. University of Minnesota Press, Minneapolis, MN

D. Hayden

Placebo Studies (Double-blind Studies)

The story of placebo and of the double-blind method is part of the greater saga of the randomized clinical trial. When regulatory agencies mandated in the 1960s and 1970s that clinical trials of new treatments conform to research methodology, including randomization, double-blind design, and placebo controls, they affirmed a shift that had taken place in American medicine. For four millennia, and well into the 1950s, medical practice had been governed by expert opinion, and treatment was mainly placebo, at worst 'filling the cemeteries' (Shapiro and Shapiro 1997).

In clinical trials, patients are the subjects of all treatment modes, including drugs, radiation, surgery, psychotherapy. This article describes how trials have been conducted and the results interpreted. Clinical trials are termed 'open label' when patients and researchers know that the active drug is being given; 'single-blind' when the patient does not know what he or she is receiving; and 'double-blind' when neither the patient nor the research team knows who is receiving what. 'Randomization' denotes the assignment of experimental subjects through a table of random numbers, dice, or other statistical tools. 'Placebo' was first defined for medical use in 1785 as a substance that was pharmacologically inert, given to reinforce a patient's expectation to get well. Glantz's definition of placebo terms the placebo effect 'a response at-

tributable to therapy *per se*, as opposed to the therapy's specific effects' (Glantz 1997). Current usage differentiates active placebo that has pharmacologic effect from inert placebo.

Thomas Kuhn, the historian of science, defines normal science as the prevalent system of beliefs and practice. He likens practitioners to chess players, who test alternative moves but do not challenge the rules of the game. The game is challenged through discovery or a new hypothesis. Practitioners of normal science combat the skeptics, solve the problem, postpone solution, or attempt to reconstruct the discussion from new fundamentals. Kuhn quotes Max Planck: 'A new scientific truth can often prevail only when its opponents die and the new generation grows up.'

Kuhn underlines that resistance to change evolves not from wrong-headedness but from the impossibility of measuring or comparing different standards. Scientists and laymen could not accept Einstein's theory of relativity that proposed that space was curved, because Newtonian physics had proven that space was flat, homogeneous, and unaffected by the presence of matter. In the 1960s and 1970s depression was thought to be caused by psychological factors until challenged by evidence that a functional deficit of neurotransmitters, norepinephrine, dopamine, or serotonin or abnormality of their receptors is the cause. While opposition to this remains, most researchers now are refining particulars of neurotransmitter metabolism.

The acceptance of the double-blind method and placebo controls are among the most important developments in science and illustrate how new discoveries drive new theories that are eventually accepted in place of incompatible, older theories.

1. History of Placebo

Placebo concept and the word derives from the Hebrew 'I will walk before,' interpreted to mean 'to please.' It gained a secular connotation in the fourteenth century, to become derisive meaning one who 'servily echoes another' (Shapiro and Shapiro 1997).

The first medical definition of placebo appeared in a medical dictionary in 1785. In the late eighteenth and early nineteenth century, it was defined as medication given to please. Eventually, it meant 'a common method or medicine' and referred to treatments as well as medicaments not yet classified as inert. References to placebo as medications were retained in the medical dictionaries until the 1950s.

The use of placebos as controls in clinical research emerged from their long history as therapeutic agents. For four millennia, medical practitioners utilized thousands of substances, 0.03 to 0.04 percent of which are pharmacologically active. These physicians did not discriminate among active and inactive concoctions, and it is not known whether active drugs were used in

effective doses. In the nineteenth and twentieth centuries, some physicians' rationalization for the use of placebo was mobilizing hope.

Efforts to introduce placebo controls into research designs had been confounded in the 1930s by the growing low esteem of placebo as treatment on the part of the medical establishment. But because of self-delusion or hypocrisy while placebos were considered quack treatments prescribed by quacks, most community physicians also used placebos. They gave vitamins without acknowledging their lack of activity for the conditions for which they were prescribed, for example. Jerome Frank, the American psychiatrist, has said that psychoanalysis might be the major placebo of the twentieth century.

Reformers from different generations concerned with the practices of their mainstream clinical colleagues advocated alertness to self-deception as to the efficacy of treatments. Roueché said that 'results from 10 well-controlled cases were superior to haphazard impressions from 1,000 cases' (Marks 1997).

2. Medical Theories

In primitive supernaturalism, illness was attributed to actions of the gods and treated with magic administered by shamans. Greek medicine combined rational and irrational explanations and treatments, using miracles, divine healing, and natural remedies. Dream interpretation was a precursor of the confession and psychotherapy. Galenic medicine of ancient Rome was predicated on Galen's principle of opposites: if a drug caused fever, it would cure chills. Galenic theories had an answer for every question and maintained a tenacious hold on medical practice well into the nineteenth century. For example, Galen had observed pus in wounds and erroneously concluded that it was essential to healing rather than a symptom and made no attempts to relieve it. Because of this belief, thousands died of infection.

Theriac, a concoction that was the mainstay of Galenic medicine, had between 33 to 100+ substances in it, eventually being prescribed as a remedy for almost every human ailment. Theriac contained opium, which alleviated pain but was addictive. The rest of the ingredients were mainly placebos; many were harmful.

The laying on of hands was a hallmark of Christianity's return to mystical explanations of illnesses. Charles II of England used therapeutic touch on almost 100,000 persons to cure scrofula, glandular tuberculosis. Skeptics attributed cures to effects of the imagination, incorrect diagnoses to start with, and spontaneous remissions. In the nineteenth century, Claude Bernard asserted that science had to be 'founded on observation and proved by experience.' He influenced the development of studies on pathology, physiology, and bacteriology, but pharmacology continued to lag behind. However, criticisms of

'normal' medicine and its treatments were voiced, notably by Oliver Wendel Holmes, who urged the retention only of quinine for malaria, mercury for syphilis, colchicum for gout, iodine for goiter, and ipecac for dysentery. If all medications 'could be sunk to the bottom of the sea, it would be all the better for mankind and all the worse for the fishes,' he said in 1860 in a much-quoted statement. Often paradigm shifts are spurred by knowledge acquired elsewhere.

Laboratory science was making major inroads by the end of the nineteenth and early twentieth century, when therapeutic chemical compounds began to be synthesized and replaced some of the naturally occurring medicaments. This was made possible by progress in laboratory sciences, biochemistry, physiology, genetics, and immunology, which are in turn based on chemistry and physics. Animal experiments where variables could be well controlled became standard.

While the concept of the scientific method was beginning to impact medicine in the nineteenth century, theories of therapeutics based on *a priori* reasoning, were still entertained. Homeopathy was a success because of placebo effects, but also due to its theoretical base that less is more. Miniscule dosing probably protected patients from ill affects of noxious substances dispensed in more robust doses by non-homeopathic physicians.

3. Beginnings of the Scientific Method

From antiquity, 'open label' experiments were conducted but results were explained by 'medical ideas (that) were speculative... and medical hypotheses that relied on a priori reasoning (Shapiro and Shapiro 1997). In the seventeenth century, Sir Kenneth Digby, an English Renaissance man, carried out a single blind experiment on a single patient. He attributed the positive results to the popular concept of sympathetic powder (Shapiro and Shapiro 1997).

In 1872, John Tyndall, an English physician, suggested that the community pray for from 3–5 years for patients ill with fatal diseases to see if differences emerged between them and others not prayed for. If or how the suggestion was carried out was not clear.

Towards the end of the seventeenth century through the early eighteenth century, some British physicians called for comparative trials using 'simple arithmetic,' i.e., 'open label' trials. These reformers were generally Scots, provincials, without access to established centers of medical influence and power, and thus, more open to new ideas. Harry M. Marks, the historian, warns that the historical record of the clinical trial is incomplete and informs that Western European and American scientists influenced each other. He points out that among medical reformers, two parallel viewpoints were emerging: medicine can be a science if it rigorously applies research that is

subsequently channeled into clinical practice, and clinical medicine can be as scientific as laboratory sciences if scientific methods are applied to judging medical treatment.

Marks further subdivides the history of the clinical trials in the US into organizational and methodological reform. The first Food and Drug Act was adopted in 1906. The American Medical Association Council on Pharmacy and Chemistry was established in 1912. The standards of the AMA Council on Pharmacy and Chemistry set by expert opinion were adopted in the Food and Drug Act of 1938.

In 1958, the AMA Council on drugs authorized the publication of a paper by Walter Modell and Raymond Houde and 'approved the use of the double-blind method, controlled comparisons, randomization, (and) statistics...' (Shapiro and Shapiro 1997). By 1970 the Food and Drug Administration was empowered to monitor the safety and effectiveness of new drugs based on adequate, well-controlled investigations using appropriate statistical methods, the result of the reforms in methodology that influenced organizational reform.

In the US, Marks credits successive generations of challengers to the methodological status quo, nineteenth-century therapeutic nihilists, and also followers of Claude Bernard's rational therapeutics that stressed knowledge of etiology before treatment. The therapeutic reformers of the twentieth century included Torald Sollmann, George H. Simmons, Thomas C. Chalmers, and Morris Fishbein. Simmons, editor of *Journal of the American Medical Association*, attacked manufacturers for flooding the market with exaggerated claims for unproven nostrums. Fishbein, another editor of *JAMA*, 'unrelentingly described the fads and follies of healing cults and medical follies' (Shapiro and Shapiro 1997). Chalmers urged in 1975 that the first patient put on any new drug be part of a clinical trial.

Charles S. Pierce and Joseph Jastrow are credited with the first use of randomization in 1885. Rivers in 1908 used the double-blind design and placebo (without using the terminology) in an experiment testing the influence of alcohol and other drugs on fatigue. He noted, 'The action of suggestion can never be excluded from any medical treatment' (Shapiro and Shapiro 1997). In 1917, John E. Coover first used the terms 'controlled experiment' and 'randomization' (Shapiro and Shapiro 1997), which only became standard through Sir Ronald Fisher's book, *The Design of Experiments* in 1935. The number of experimental studies increased steadily through the work of, among others, Louis Pasteur, Robert Koch, and Rudolph Virchow. Torald Sollmann, professor and chair of the Department of Pharmacology at the Medical School of Western Reserve University, headed for 55 years the AMA Council on Pharmacy and Chemistry. Sollmann urged that it was 'essential that clinical experimentation should follow the canons of other

scientific experimentation' (Shapiro and Shapiro 1997). He coined the term 'comparative method.' Publications of the Council made reference to 'the blind test' and published what was probably the first study of the 'double-blind procedure.' In 1930, Sollmann recommended the use of the 'blind test,' used the term 'placebo' for the first time in research, and dropped the term 'comparative method.' But his news had little impact on practice, and the AMA Council on Pharmacy and Chemistry continued to rely on reviews by its member experts rather than on controlled studies.

The experts were researchers of unquestionable skill and integrity, but they rarely had experimental data and focused on confronting the tactics of manufacturers, their aggressiveness and exaggeration of claims, and on manufacturer-supplied laboratory evidence. Their influence led to a 'seemingly deluded faith in experts,' from 1900–1930 i.e., themselves and colleagues of the same ilk (Marks 1997).

A major milestone was a single-blind study in 1932, when Harry Gold tested the use of xanthine against placebo (lactose) for cardiac pain. He realized that physicians were asking the patients leading questions and prejudicing answers, and thereafter, tried to blind the physicians, as well. In 1935, Hediger and Gold compared two forms of ether in a 'blind test' and legitimized the use of placebos. Kwit summarized their insights: 'Towards the end of the study we realized that the doctor must not know what he gives, the patient... what he receives, and the questioner must not know what was given' (Shapiro and Shapiro 1997).

Like Sollmann earlier, Gold in the 1930s failed to have an immediate effect on medical practice. However, throughout the next two decades, researchers flocked to work with him and the new methodology, and he supervised many double-blind, placebo controlled studies. In a 1950s interview with Shapiro, Gold said, 'The placebo and the double-blind were companions after our 1937 study.'

A 1948 joint study by the US Public Health Service and the British Research Council is generally considered the first clinical trial. It tested the effectiveness of streptomycin in the treatment of tuberculosis. Patient selection was randomized. Evaluators of the patient X-rays did not know which patients had received streptomycin and which placebo. The treatment was not blind, because streptomycin-treated patients received up to four injections per day. Since the controls were bed-rest only patients, it would have been unethical to subject them to injections of placebo.

4. New Paradigm I

It was not until the 1950s, when antibiotics were discovered and mechanisms of several metabolic diseases understood that modern medicine began, and

clinical research could continue to make inroads. Through the 1960s and 1970s, the scientific method superseded authority and tradition. Treatments had to show sensitivity, specificity, and predictability through statistically sound techniques, randomization, the double-blind method, and placebo controls.

Current clinical standards for research are reflected in the US Preventive Services Task Force recommendations for treatment. Guidelines of 1989 (revised 1996), rank in descending order of reliability and ascending order of bias and misinterpretation randomized controlled trials, nonrandomized controlled trials, cohort studies, case control studies, comparisons between times and places, uncontrolled experiments, descriptive studies, and expert opinion.

5. *New Paradigm II*

The new scientific canon of clinical research encountered challenges and doubts almost as soon as it became established. Questions of method were raised about blindability in double-blind studies, as well as the precise definition, activity, and validity of placebo. Critical voices questioned ethical aspects of clinical trials that deprived half the patients of a medication that might be effective. The ensuing debate polarized ethicists and physicians.

6. *Blindability*

Extensive reviews of blindability indicated that studies took its validity for granted. It was suspect in the few double-blind trials that tried to carry out such tests of validity. Basoglu, in review of the literature found that only three of 23 double-blind psychopharmacology studies maintained blindability in patients, physicians, and assessors. Side effects, explanations in informed consent forms, past experience with similar drugs, and improvement with the active drug but not placebo sabotaged blindability. In his own comparison study of blindability of alprazolam, an anti-anxiety agent, with behavior therapy for panic disorder, Basoglu found that patients, treating therapists, and assessors were able to guess correctly, to a statistically significant extent, the patient's category. Patients treated with the active drug were more accurate than those on placebo, and active drug accuracy was highest at late stage of therapy, when optimal dosing and maximal efficacy had presumably been achieved.

Morin et al. (1995), recognizing that judgments carried out at the end of a study as to what was drug, what placebo could be influenced by treatment response to active drug, compared Temazepam, a hypnotic, and an inactive placebo. They carried out three assessments over 8 weeks. All patients received cognitive or behavioral therapy. Both patient and physician groups showed a statistically significant

high degree of accuracy in guessing at 1 and 8 weeks with no better than chance guessing at week 4.

From these and other studies, it appeared that certainty ratings increased with time, that experienced patients showed greater accuracy in penetrating blinding, and that patients on active drugs guessed better than those on placebo. In summary, neither the patients nor the administering physicians were blinded to any significant degree.

7. *Placebo*

Harry Gold had accepted earlier twentieth-century definitions of placebo, such as Rivers' that placebo gives an inactive imitation of the substance being tested and Sollmann's that placebo furnishes the comparative check of the natural course of the disease, and did not explore the nature of placebo and placebo effects. Current theories on the nature of placebo and placebo effects include spontaneous remission due to biological diversity, conditioning, mobilization of endogenous endorphins, and the statistical phenomenon of regression to the mean. In 1955, H. K. Beecher calculated that 35.2 percent improvement in 15 clinical trials involving different diseases was due to placebo.

McDonald and Mazzuca (1983) concede the evidence that patients improve during placebo treatment, but point out that that is not proof that placebo treatment is the reason. The percentage of improvement observed in placebo treated patients is identical to that predicted from a statistical phenomenon known as regression to the mean where both objective and subjective measures of abnormality tend to improve and move towards the mean with successive measurements. McDonald and Mazzuca (1983) illustrate this in a review of 30 randomly selected placebo controlled clinical trials. Placebo effects and regression to the mean are both 'not uniform, constant, or predictable.' (Regression is by definition a random phenomenon.) McDonald and Mazzuca do not negate the usefulness of placebo. In comparisons of known drug with unknown, placebo controls can clarify ambiguous results. Placebo controls are also useful in conditions that fluctuate.

McDonald and Mazzuca note that placebo effects can be created through conditioning, citing the work of Stewart Wolf on the effects of conditioning and suggestion on measurable physiologic events in human beings.

8. *Placebo Amplification*

Richard Thomson believed that fewer studies would show a significant difference between placebo and drug if an active placebo were used as control. He reviewed 75 studies of tricyclic antidepressants com-

pared with active placebo (atropine) or with an inert placebo (lactose). Indeed, as postulated, antidepressant efficacy was ranked as superior in only one of the seven atropine/placebo studies available, while the tricyclic antidepressant was ranked superior in 43 of 68 studies controlled by an inert placebo. One explanation might be that atropine might have an antidepressant effect of its own. However, Thomson concluded that placebo amplification, a situation where physiological effects of taking a drug convince subjects already committed to treatment that the treatment is, in fact, effective, was the phenomenon being observed.

Roger P. Greenberg compared a tricyclic antidepressant with a new drug being tested, as well as an inactive placebo. He postulated that in a study focused on a new drug's effectiveness, there would be less interest in proving the efficacy of the older drug. Indeed, compared with claims in earlier literature, the older drug was much less effective.

Shapiro and Shapiro (1997) reviewed 27 studies of guessing of placebo and drug allocation and found that overall, both patients and physicians made correct identifications in 60 to 80 percent of tests. Eighty percent of drug guesses, but only 50 percent of placebo guesses were correct in their own study.

Fisher and Greenberg (1995) found in a review of world literature of studies of psychotropic drugs prior to 1993 that 'double-blind is vulnerable to penetration and hence, to bias,' placing all previous psychopharmacological research in question.

To try to rectify previous research on the efficacy of psychotropic drugs, Fisher and Greenberg suggest greater complexity in research design, the use of active placebo, various crossover designs, 'conventional drug' and 'drug disguised' groups, dose ranging, and the testing of drugs in opposing contexts, i.e., in an atmosphere of optimism or pessimism. They concede that current regulations preclude a drug disguise scheme for example, and agree that a completely foolproof design is not likely ever to be contrived. Personalities of patients and researchers were a factor in correct guessing of placebo or drug. While overall studies showed agreement among evaluators, rater differences were noted. Patient noncompliance was another confounding variable.

9. *Ethics*

As early as in the 1960s, ethical considerations about clinical trials involving randomization, double-blind methodology, and placebo controls were raised. Benevolent paternalism had been a prevalent ethical guideline since antiquity. At worst, it led to egregious exploitation of patients. In the late 1950s and 1960s, H. K. Beecher, in several watershed papers, cited instances of unethical and unscientific proceedings by

researchers. He recommended that patients must give 'informed consent.' The US Public Health Service responded by proposing the establishment of Institutional Review Boards. The National Institutes of Health mandated Internal Review Boards, and the Food and Drug Administration, published in 1970, rules for 'adequate and well controlled clinical evaluation' (Shapiro and Shapiro 1997).

Beecher's papers helped move medicine from its tradition of paternalism to informed consent and patient autonomy. However, many in the medical establishment complained that the patient-physician dyad was being fractured by a mass of lawyers, judges, journalists, philosophers, and legislators. And that inevitably, bureaucracy was stifling the practice of medicine.

Ethicists had entered discussions in the 1970s, questioning placebos, the double-blind procedure, and the randomized clinical trial. Sissela Bok and Beth Simmons took strong stands opposing the use of placebos without informed consent. Others considered the use of placebos to be unethical in practice but permissible in research, to the point of not informing subjects of the research design. Some proposed that it was unethical not to carry out a placebo controlled, double-blind study. Still others concluded that since placebos have shown positive therapeutic effects using a placebo as control does not deprive patients of all treatment.

Questions have also been raised about how informed is informed consent in general, and especially if the subjects are cognitively or psychologically impaired? Attempts to resolve some of the issues are also being made. Where an effective treatment exists but a new one is being tried, the potency of the established treatment, as well as the severity of the patient's condition, are considered. As part of the experimental design, patients are moved from experimental or placebo cohorts to conventional treatment if their conditions deteriorate.

In 1978, while the rest of the Western scientific community moved towards research criteria of randomization, double-blind, and placebo controls in clinical trials, Germany passed laws substituting proof of efficacy through clinical open trials or expert opinion as qualifications for acceptance and registration of drugs. Controlled studies were no longer considered as essential to proof of efficacy.

In the introduction to the Physician's Desk Reference for Herbal Medicines, the editors say that in the absence of the 'rigors of FDA review' of these preparations, the editors are basing their guidelines on 'the closest available analog to FDA approved labeling—the findings of the German Regulatory Authority's watchdog agency commonly called Commission E'...its conclusions represent the best expert consensus...'

Statisticians contend that uncontrolled studies tend to favor a positive outcome.

10. Summary

Scientific thinking, defined by uncertainty that drives observation, hypothesis formation, and further search, was a late arrival to medicine, which was still laboring under few restrictions, guided by elaborate theories, mysticism, superstition, tradition, and authority.

Scientific thinking met much resistance in the medical establishment, but open label clinical trials, clinical trials using placebo controls, and double-blind methods began and continued into the twentieth century. In the 1950s when it became possible for the first time to cure people, statisticians became essential partners in the research team (Glantz 1997). Methodological and ethical questions began to be raised in subsequent decades about contemporary scientific criteria. Patient activism had also contributed to the erosion of paternalism that afflicted research design, as well as clinical practice.

The future may render clinical trials obsolete. We may well have a database atlas of the brain and body that will enable the researchers to calculate and theorize about effects without the use of human subjects.

The reader is referred to Marks (1997) and Shapiro and Shapiro (1997) for more complete historical surveys, detail, and documentation.

See also: Alzheimer's Disease; Antidementive Drugs; Antidepressant Drugs; Schizophrenia, Treatment of

Bibliography

- Basoglu M 1997 Double-blindness procedures, rater blindness, and ratings outcome: Observations from a controlled trial. *Archives of General Psychiatry* **54**: 744–8
- Beecher H K 1955 The powerful placebo. *Journal of American Medical Association* **159**: 1602–6
- Bystritsky A, Waikar S V 1994 Inert placebo vs. active medication: Patient blindability in clinical pharmacological trials. *Journal of Nervous and Mental Disease* **182**: 485–7
- Cohen A, Posner J 1995 *A Guide to Clinical Drug Research*. Kluwer Academic Publishers, Dordrecht, The Netherlands
- Fisher S, Greenberg R P, Riter J A 1995 Placebo washout is not a meaningful part of antidepressant drug trials. *Perpetual and Motor Skills* **81**: 688–90
- Frank J D, Frank J B 1993 *Persuasion and Healing*. Johns Hopkins paperback edn. The Johns Hopkins University Press, Baltimore and London
- Freston J W 1986 Dose ranging in clinic trials: Rationale and proposed use with placebo positive controls. *American Journal of Gastroenterology* **81**: 307–11
- Glantz S A 1997 *Primer of Biostatistics*, 4th edn. McGraw Hill, New York
- Kuhn T S 1996 *The Structure of Scientific Revolutions*, 3rd edn. The University of Chicago Press, Chicago
- Marks H M 1997 *The Progress of Experiment: Science and Therapeutic Reform in the United States, 1900–1990*. Cambridge University Press, Cambridge, MA
- McDonald C J, Mazzucco S A 1983 How much of the placebo

effect is really statistical regression? *Statistics in Medicine* **2**: 417–27

- Morin C, Calecchi C, Brink D 1995 How 'blind' are double-blind placebo-controlled trials of benzodiazepine hypnotics? *Sleep* **18**: 240–5
- Niklson I A 1997 Factors that influence the outcome of placebo-controlled antidepressant clinical trials. *Psychopharmacology Bulletin* **33**: 41–51
- Report of the US Preventive Services Task Force 1996 *Guide to Clinical Preventive Services*, 2nd edn. Williams and Wilkins, Baltimore, MD
- Shapiro A K, Shapiro E 1997 *The Powerful Placebo: From Ancient Priest to Modern Physician*. The Johns Hopkins University Press, Baltimore, MD
- Thomson R 1982 Side effects and placebo amplification. *British Journal of Psychiatry* **140**: 64–8

E. Shapiro

Placebos, Use of: Ethical Aspects

A placebo is a therapy, or an intervention designed to simulate therapy, that has unspecified activity for the condition being treated. Placebo components exist in all treatments that produce an expectation of good effects, including competent and caring professional behavior of those delivering the therapy, and status or attractiveness of the institution. Placebos effectively relieve symptoms of most medical and psychological conditions. The underlying mechanisms are increasingly a topic of theory, speculation, and research. The ethics of using placebos depends on context, theory, and purpose. This article examines evolving theories and uses of placebos and associated ethical issues.

1. History

Before modern medicine, many substances thought to have curative powers, e.g., arsenic, produced feelings of wellness, before perhaps killing the patient. Such placebo effects, due to ignorance of the physician and patient, were primary active ingredients of ancient medical practice. By the mid-1800s, Koch, Pasteur, and others demonstrated that specific bacteria could cause specific diseases, giving rise to specific pharmacologic and surgical interventions replacing non-specific healing therapies. Placebo effects became 'nuisance variables' to be controlled for, and ethics considered when it was acceptable to use placebos instead of accepted treatment modes. Finally, placebo effects are being studied as interesting phenomena in their own right.

Their therapeutic value is increasingly recognized, and the ethics of failing to use placebos to produce beneficial results is being questioned. Today's profes-

sionals produce placebo effects unwittingly, incidentally, or deliberately in research and practice in many fields including medicine, psychotherapy, and teaching. Today's ideal professional would appropriately and deliberately combine the power of modern science and technology with the healing power of placebos. Unfortunately, impersonal cost-cutting programs may inadvertently eliminate such effective strategies.

2. Definitions and Uses

The ethics of placebo use depends on which definition or theory of placebos is employed.

2.1 The Pharmacological Paradigm

The term placebo, from Latin *placere*, 'to please,' was coined to mean a pharmacologically inert substance, e.g., a sugar tablet, deliberately disguised as medication. Thus defined, the placebo is used to satisfy complaining or malingering patients; placebo effectiveness was until recently considered proof of patient psychopathology or feigned illness. Today's physician still employs placebos when the requested medication is inappropriate, but to satisfy concerns about lying to patients, may use very weak medications rather than inert substances.

Research on new treatments may yield 'success' simply as placebo effects. To control for this 'noise' researchers randomly assign both subjects and researchers to placebo and therapy groups. Double blind designs are used because researcher belief also influences results. The therapy is deemed successful if it yields better results than the placebo. (Without a placebo group, recovery due to treatment could not be distinguished from recovery due to spontaneous natural healing, as well as from expectancy effects.) To satisfy ethical requirements, researchers:

- Inform subjects that they may receive placebos;

- Specify the chance of receiving a placebo;

- Use placebos only when genuinely uncertain about treatment effectiveness;

- When inert substances might do harm, compare experimental treatment to baseline clinical data or current effective treatment;

- Monitor subjects regularly for adverse effects and remove immediately from placebos if indicated;

- Provide the placebo group with effective therapy later;

- In double-blind studies, ensure that the code can be broken quickly to discover whether a placebo is causing a worsening condition;

- Do not use placebos with highly vulnerable populations;

- Use placebos resembling the active preparation so that patient and researcher cannot detect the difference.

Use of placebo control groups has, unfortunately, become such a hallmark of publishable research that placebo controls may be used when other approaches, e.g., comparison with current effective treatments, are as rigorous, often more appropriate, and put the subject's interests first. Placebo control groups in psychiatric research on such vulnerable populations as persons suffering severe depression or schizophrenia have come under considerable scrutiny. Full comprehension and consent of the patient and next-of-kin concerning the possible use of a placebo, and careful surveillance of the patient, are crucial to ethical use of placebos when suicide or other forms of harm are plausible risks.

The pharmacological paradigm ignores sociocultural and interpersonal variables, expectations, learning, neurophysiology, and real therapeutic value of placebos. The malingering patient notion implies personality or demographic characteristics of placebo responders, but such characteristics have not been found.

2.2 Remembered Wellness

Newer conceptions and uses of placebos are highlighted in Benson and Friedman's (1996) article. Under the right circumstances anyone might respond to a placebo. Placebos are not absence of treatment. What persons believe can diminish pain and objective symptoms of disease. The components of 'remembered wellness'—the patient's and physician's positive beliefs and expectations, and a good relationship between the two parties—are one of medicine's most powerful assets. Impersonal medical care and drug or surgical interventions are not effective in treating most diseases. Most visits to physicians stem from conditions related to stress and other psychosocial factors (Kroenke and Mangelsdorff 1989) which are responsive to placebo effects.

Such placebo effects based on 'remembered wellness,' accompanying standard therapies comprise effective care. For example, in an experiment using five standard therapies for angina pectoris, patients who believed in the therapy experienced about 80 percent relief from chest pain, reduced nitroglycerine usage, improved electrocardiogram, and increased exercise tolerance; skeptical patients experienced about half as much therapeutic benefit (Benson and McCallie 1977). In Canner (1980) among patients given placebos, less compliant patients had a 28 percent mortality rate after five years, while compliant patients had a 15 percent mortality rate.

Placebo use has evolved to entail warm, caring, sympathetic relationships that foster positive beliefs, combined with appropriate standard specific therapies. Thus defined, placebos are devoid of deception, and integral to quality of life, health, sound professional conduct, and empowerment and self-help.

Under this conceptualization, it is reductionist scientific medicine coupled with impersonal managed care that is ethically questionable.

3. Theories of Why Placebos Work

Three fundamental theories of placebo mechanisms are: conditioning (learning), hope and expectancy (cognition), and neurobiology (subjective experience). These three mechanisms are not mutually exclusive. Each provides ethical insights into placebo use.

3.1 Learning

Classical conditioning involves pairing an unconditioned stimulus (UCS), e.g., a charismatic, enthusiastic teacher who emphasizes the excitement and importance of a topic, with a neutral stimulus (a not-particularly exciting curriculum) which becomes the conditioned stimulus (CS) and evokes enthusiasm and dedication in the student. Pairing morphine (UCS) given to patients in great pain, causes the nurse and syringe to become conditioned stimuli (CS) for pain reduction. Placebo effects persist when only the CS is present and, following principles of partial reinforcement, continue undiminished if the UCS is reintroduced periodically, but eventually extinguish if the UCS is never reintroduced. For example, if a painkiller (UCS) having severe side-effects is interspersed with a placebo in an identical capsule (CS) and intermittent reintroduction of the UCS, the effect may be a continuously effective analgesic with diminished side-effects. There is a powerful relationship between magnitude of UCS and response to the CS. For example, a massive dose of morphine (UCS) will result in a massive placebo (CS) effect. These principles of conditioned analgesia have been found with animals as well as humans.

Ethics requires disclosure of use of conditioning; e.g., following explanation of learning theory, the doctor might state that the UCS would be intermittent and that the client's own relevant powers will serve to reduce pain. Conditioning produces expectancy, and memory of UCS effectiveness, as well as measurable neurological change.

3.2 Desire and Expectancy

Need or desire (e.g., for relief from pain) and expectation that a given therapy is effective will mediate placebo analgesia. These cognitive factors become associated with the CS and extend and support the role of learning. Authoritative professionals can create powerful hopes and expectations which can produce a placebo response without prior exposure to an objective UCS, though such exposure (combining conditioning with expectancy) increases expectancy.

Similarly, expectancy can result from learning about an agent's therapeutic value, especially when one strongly desires that effect. Thus, persons desperate for relief are highly vulnerable to quackery, fraud, superstition, or cults, which may produce significant temporary relief through placebo effects. Conversely, anything expected to cause distress, sickness, or death (e.g., feared surgery, awkward physicians, institutions with poor reputations) can have a *nocebo* effect (Hahn 1997). Culturally based expectations, and good or bad relevant prior information or experiences, as well as aspects of the current therapeutic situation will create expectations and have placebo or nocebo effects. Positive expectations result in far superior outcomes of standard medication, surgery, education, and other interventions. The ethical professional uses expectancy effects to good ends, not to deceive people.

3.3 Neurobiology

If placebo effects are solely products of imagination, the ethics of placebo use concerns fooling people. However, as scientists discover how subjective 'mind' acts on the physical body, the ethics of placebo use changes. The functioning of neurons is involved in mental representation (e.g., the self, an image); subjective experience (e.g., pain, feeling of wellness); and bodily change (e.g., hormone secretion, healing); and the neural network, itself, is an objective physical entity. Thus, for example, injury causes pain, morphine reduces that pain, and placebos also alter neural representation of the injury (pain) in proportion to the preceding amount of morphine. Research has demonstrated neural substrates of placebo effects, especially analgesic effects, and imaging technology holds promise of rapid advancement in this area (see Fields and Price 1997).

Neurological models of placebo effects provide a framework for research on the role of language, culture, and environmental context in helping to reduce stress and pain, reversing disease processes, and healing injury. Future discoveries in this domain may lead to rewriting the ethics of placebo use.

4. Summary

Placebos have evolved from deception, and 'noise' that researchers control for, to phenomena studied in their own right. Research has shown that what persons think can affect objective (structurally or functionally measurable) changes in physiological functioning. Further theory and methodological developments promise the possibility of effective self-administered placebos—a new kind of self-help, and of radically new ethical considerations regarding placebo use.

See also: Assessment and Application of Therapeutic Effectiveness, Ethical Implications of; Clinical Treat-

ment Outcome Research: Control and Comparison Groups; Ethics for Biomedical Research Involving Humans: International Codes; Medical Experiments: Ethical Aspects; Placebo Studies (Double-blind Studies); Placebos, Use of: Ethical Aspects

Bibliography

- Benson H, Friedman R 1996 Harnessing the power of the placebo effect and renaming it 'remembered wellness.' *Annual Review of Medicine* 47: 193–9
- Benson H, McCallie D P 1977 Angina pectoris and the placebo effect. *New England Journal of Medicine* 300: 1424–9
- Canner P L 1980 Influence of adherence to treatment and response of cholesterol and mortality in the coronary drug project. *New England Journal of Medicine* 303: 1038–41
- Fields H L, Price D D 1997 Toward a neurobiology of placebo analgesia. In: Harrington A (ed.) *The Placebo Effect: An Interdisciplinary Exploration*. Harvard University Press, Cambridge, MA, pp. 93–116
- Fish J M 1973 *Placebo Therapy*. Jossey-Bass, San Francisco, CA
- Frankenhaeuser M, Jearpe G, Svan H, Wrangsj a B 1963 Psychophysiological reactions to two different placebo treatments. *Scandinavian Journal of Psychology* 4: 245–50
- Hahn R A 1997 The nocebo phenomenon: Scope and Foundations. In: Harrington A (ed.) *The Placebo Effect: An Interdisciplinary Exploration*. Harvard University Press, Cambridge, MA, pp. 56–76
- Harrington A 1997 *The Placebo Effect: An Interdisciplinary Exploration*. Harvard University Press, Cambridge, MA, pp. 56–76, 93–250
- Kroenke K, Mangelsdorff A D 1989 Common symptoms in ambulatory care: incidence, evaluation, therapy, and outcome. *American Journal of Medicine* 86: 262–6
- Peck C, Coleman G 1991 Implications of placebo theory for clinical research and practice in pain management. *Theoretical Medicine* 12: 247–70
- Rossi E L 1993 *The Psychobiology of Mind-Body Healing*. Norton, New York
- Smith T 1999 *Ethics in Medical Research*. Cambridge University Press, Cambridge, UK
- Spiro H M 1986 *Doctors, Patients, and Placebos*. Yale University Press, New Haven, CT
- Wickramasekera I E 1988 *Clinical Behavioral Medicine*. Plenum, New York, Chap. 5

J. E. Sieber

Plagues and Diseases in History

Disease is the sting of death and death is the necessary victory of the grave. Disease is a mechanism of adaptation—that is, of attack, equilibrium, and defense—of the living being to its habitat. Human development and evolution, in its social, cultural, and ecological setting, give rise to alterations in the human organism, both in body and mind, against which it

must struggle. The evolution of the species, society, and culture takes place within a given environment and this environment may alter the pathological spectrum. If according to Bichat life is resistance to death, it may be said that life is resistance to disease. The human being seeks Hippocratic and Aristotelian equilibrium but, in doing so, he or she is frequently subjected to changes and movements that endanger this quest.

These changes, quite frequently depend on the human being, on his or her customs, social groupings, or on settlements that lead him or her to avoid malaria by settling in mosquito free zones or to control smallpox by vaccination. On other occasions, such changes do not depend on the human being but arise in the environment or within the diseases themselves that may appear or disappear of their own accord, as was the case with the 'English sweate.' On yet other occasions, these changes are merely due to cultural appreciations, such as is the case where certain sexual practices such as masturbation or homosexuality are stigmatized as diseases. On the other hand, disease is not always considered harmful: thus it may be considered a distinction of the gods, as was the case of epilepsy in the time of the classic authors. This same disease was associated with the devil by the Christians, and historically, it had always been a marked distinction of great personages from Caesar to Napoleon. Disease can also be considered a way towards perfection or transformation, leading to the creation of art or the salvation of the soul. It may also have political repercussions: it is claimed that George III's madness gave rise to problems for the British crown while Roosevelt's fragile health was considered to have placed him in a weak position in the Yalta negotiations. Socially, disease has sometimes been considered the stigma of groups or races. Medicine and politics have endeavored to preempt social changes by means of eugenics which began with recommendations or orders to avoid marriages entailing risks (real or imaginary) and went on to such processes as sterilization and abortion. Hereditary problems are thus intertwined with cultural, social, religious, and moral ones. In contrast with the stark determination of former times, there is today an attempt to draw new optimism from the possibilities of genetic modification. Although, in this latter respect, there is fear of a future where such a formidable force could rest in the hands of the rich and powerful. Thus, individual disease or illness is rooted in society, being tied up with the history of great endemics and epidemics.

1. Ancient and Middle Times

Paleopathology enables us to discover the high incidence of diseases in the early inhabitants of the

Earth, especially in the cases where such diseases left their mark on the bone. It is thus possible to detect remnants of tumors, infections, and osteoporosis as also traumatism, dental alterations, rheumatism, and osteopathy. From the first settlements in the Fertile Crescent, changes have taken place in the relationship of man and animals with their environment and it is possible to establish an extensive catalogue of diseases which largely subsist down to this very day. Particularly significant are the forms of settlement, the crops, and the contact of man with cattle, pets, and parasites. Thus in the fourth and third centuries BC, malaria had already set in the early civilizations in marshy areas where the water and the crops provided the conditions for the mosquito to infect man. Evidence for the existence of seasonal and intermittent fevers is already to be found in Chinese, Indian, Sumerian, Assyrian, Babylonian, and Egyptian cultures. The link between splenomegaly and fever is mentioned in the Papyrus of Ebers, dating from 1570 BC. The connection between fever and marshy areas appears in the Hippocratic text *On Airs, Waters and Places*. Deforestation, agriculture, temperature, and a greater population density, as well as military and commercial movements favor the spread of the disease. In the third century BC in Italy, the disease becomes acute and in the first century BC Varron coincides with Lucretius in pointing to small animals as the cause of the disease and recommended that these be either abandoned or sold and that people should dwell on high ground, this latter being traditional advice and here also defense reasons counted. The spread of malaria got worse in Rome owing to negligence and flooding and it reached its zenith at the time of the fall of the empire, a cataclysm to which it may have contributed.

The bubonic plague was endemic in the Near East in pre-Classical times, the buboes being described by Hippocrates. In 430 BC, Thucydides describes the plague of Athens in the war against Sparta, but it is doubtful that this case refers to the bubonic plague. The first reliable description of this disease is that given by Rufus of Ephesus in the first to the second century AD, in the epoch of Trajanus; and in the sixth century, the plague of Justinian initiates the descriptions of the first great epidemic cycles. Up until the Christian era, there is equilibrium between human beings, cities, and parasites but this equilibrium breaks down with the advent of the convoys that bring the two healthiest extremes into contact, namely, China and Rome. India and the Middle East continue to be dangerous and the situation at the extremes further deteriorates. Soldiers and merchants, animals and merchandise, slaves and prisoners all contribute to linking up the extremes of the known world. A consequence of this contact was the Antonine Plague in the second century. This plague was attributed to smallpox, something that had already existed as far back as the year 3000 BC, as certain mummies show.

In the Middle Ages, leprosy spread greatly. This biblical and present-day disease is accompanied by malignant connotations since dirtiness and overcrowding provided the conditions for the spread of the otherwise not extremely contagious *Mycobacterium leprae*. In this respect, the Route to Santiago de Compostella where pilgrims slept in groups and under conditions that left a lot to be desired provided a fertile ground for this disease. Another markedly religious character for the spread of the disease is to be found in the 'Sacred Fire,' 'St Anthony's Fire,' or ergotism, during which rye ergot was eaten. These diseases are linked to many others that owed their spread to squalor and poverty, for example, scabies and lice, consumption and fevers, anthrax and ophthalmia. In the Mediterranean bowl, malaria continued to be endemic with the population tending to resort to residing on the hills and mountains, far from marshy lands. The plague set in the incipient cities with the effects of rats and lice, making this disease an explicative example par excellence of epidemic history; see in this respect, Zinsser (1934). A second epidemic cycle commences with the Black Death in 1348. At this time, Venice establishes control over persons and ships. Commerce and the cities collapse and thus the hegemony of Castille over Aragon takes place, giving rise to the origin of the modern European Kingdoms and empires.

2. Modern Times

European expansion meant a sudden traffic of parasites towards the new worlds. The discovery of America, with its scattered population devoted to agriculture, where there were no dirty cities and where the inhabitants were devoid of immunity to the diseases of old Europe, produced an enormous demographic crisis. Certain diseases, such as smallpox ravaged the indigenous population, first in South and later in North America. Syphilis has supposedly come from America, brought back by the Spaniards and it caused havoc in Renaissance Europe. Sea voyages led to the development of scurvy, due to the lack of fresh foods. Other diseases also came from overseas, such as yellow fever, although it originated from Africa and later from the tropics, due to the commerce of slaves and merchandise. In the seventeenth century, it took root in Brazil and is still present today in the Amazon. Soon this disease spread to the Caribbean and throughout America, reaching New York and Boston and likewise Colombia, Ecuador, and Peru. It affected warm America and Europe, especially the Iberian Peninsula. Immunity and cut offs in commerce stalled the disease and today it remains endemic in tropical areas of America and Africa.

The plague arrested demographic growth in Europe since the enmity between states and trade impeded protection against the disease. The last plague of Marseille in 1720 and the relative peace of the

eighteenth century allowed all European nations to set up sea and land defenses with Austria becoming a solid bulwark in the face of the Ottoman Empire, thus sealing off all possible spread of the plague. Towards the middle of the eighteenth century, the third wave of the disease broke out and with the exception of Europe, it spread to all countries including Asia, Africa, and paradises like America and Australia, leaving remnants in those places. The eradication of the plague in Europe was a result of the advent of public hygiene, immunity to the disease, and the disappearance of rats and the old and dirty wooden buildings. The fire of London 'purified' the city, since the rebuilding was carried out in stone, eliminating animals and dirt.

The disappearance in Europe of the plague was followed by the appearance of cholera that had been endemic in India for 2000 years. This disease was described by Western travelers in the sixteenth century and its spread to Europe and America was a consequence of pilgrimages, trade, and a lack of cleanliness in water. A great pandemic broke out in 1817 and spread through Eastern Africa and Southern Asia and in a second wave it spread through Kabul and Persia reaching Russia and Poland in 1830. It then swept across Europe and had reached America by 1832. Over two centuries, terrible waves spread from the East and as late as the 1970s, the seventh outbreak of pandemic reached Europe. However, the timely and necessary cleansing of urban water supplies helped to stall its advance progressively. Nevertheless, as late as 1892, the free Hanseatic city of Hamburg suffered an outbreak of cholera epidemic while the neighboring Altona, governed by the new Prussian Reich, was able to avoid the disease, thanks to the successful filtering of its water. Cities and governments advanced in hygiene and they attempted to bring about healthier cities by means of appropriate public health measures.

Together with the major epidemics, death was continuously caused by two groups of endemic diseases. Two seasonal cycles of disease remained rooted into modern times: one in summer affecting children and due to enterocolitis, lactation problems, dysentery, and typhus, and another in winter affecting the old and due to typhus and pneumonia. The nineteenth century sees the beginning of the demographic revolution in Europe with a sharp decline in mortality, especially maternal and infant mortality. Diseases like diphtheria, chicken pox, scarlet fever, measles, and smallpox are combated with serum and later on with vaccination. Mother care also advanced both throughout pregnancy as also at childbirth and during lactation. Tertian fevers continue to be active and are combated in the Mediterranean area by clearing up the marshy areas, avoiding the mosquito by mechanical means, or by the use of cinchona bark (Peruvian bark), and later on of quinine.

The rise of great and industrial cities with low-grade outskirts and the emigration of peasants to the city

favoring the spread of maladies associated with poverty, squalor, lack of appropriate food, and exposure of people devoid of defenses. Some of the diseases are strictly occupational which is the case of miners and textile workers who were exposed to injury and to the industrial poisons that affected the proletariat. Other diseases are closely related to the harsh conditions of life with the spread of typhus and pulmonary diseases, especially tuberculosis, being due to this factor. Tuberculosis spread over the turn between the eighteenth and nineteenth centuries and while it was known as the disease in fashion, affecting notable and distinguished people, it struck the proletariat much more severely. Today, tuberculosis is again a serious threat. The same can be claimed regarding the enormous spread of syphilis as also with respect to drug abuse, beginning with alcohol and continuing with marijuana, cocaine, and morphine. At the middle of nineteenth century the opium trade was legalized in China, after the victory of the British army.

3. Contemporary Times

This last century marked the descent of the high mortality rates in countries that reached high standards of sanitary development, that is countries that invested in public health care, such as Britain or the USSR. Such advances have been due to public health services, hospitals, antibiotics, surgery, and vaccination and they have been reinforced by the developments in immunology, pharmacology, and biotechnology and with the promising future of improved genetic material. Smallpox is the first disease to be considered totally eradicated with only some samples of the virus being kept at a few laboratories for study purposes. But other viruses have taken its place. Thus the old influenza caused three heavy outbreaks of pandemic in 1918–9, 1957–8, and 1968: these were known respectively as the 'Spanish 'flu,' which originated in the USA, the 'Asian 'flu,' and the 'Hong Kong 'flu.' Poliomyelitis originated in the USA, affecting non-immune children and this disease has only been controlled in developed countries by means of vaccination. AIDS has meant a development in virus pathology and this disease has associations with the exploitation of poor countries, drug trading, and also an elite sector that has sprung rapid research in the field. The world commotion surrounding this disease has served both to reveal human altruism on the one hand, but, on the other, contempt towards those infected by the disease, since AIDS is becoming more and more the lot of the poor, such as large areas of Africa or southern Asia and of downcast or marginal groups such as the chronically ill, drug addicts, prostitutes, and homosexuals. Nevertheless, heterosexual transmission is advancing and is placing the future of all Africa in danger.

Forms of illness are changing. Development has managed to stamp out old forms of infectious morb-

idity. On the other hand, the increase in life expectation favors the development of Alzheimer and cancer. Cardiovascular accidents and traumatism, traffic and sports, factories and radiation are something that the twenty-first century will inherit. Ghettos and unemployment, and some of the old drugs such as heroin and tobacco or new synthetic ones are causing havoc even to the rich world. But in underdeveloped countries, the old morbidity continues to exist, giving rise to very low life expectation. Death is most rampant amid the least protected, especially women, children and the elderly, with the wars causing high death rates amongst men. Under such circumstances, diseases caused by deprivation and infection continue to be the most rampant.

Malaria has been eradicated from Europe thanks to drainage, improvements in crop cultivation, mechanical means, quinine, and insecticides. Today, faith is also placed on vaccination and the genetic modification of the mosquito. Yet malaria continues to be endemic in warm zones of America, Asia, and Africa. In the interwar years, the Rockefeller Foundation sponsored a campaign against the disease. In 1955, the World Health Assembly, meeting in Mexico, warned against the resistance of the mosquitoes to insecticides. Together with the WHO and the UNICEF, the US government was financing the eradication program of the Pan-American Health Organization. Today, an effort is being made to interest governments in this disease, a disease that represents a danger to 40 percent of the population. In Africa the situation is very serious due to the increase in agriculture and irrigation as well as emigration, the political and economic problems and the bad sanitary conditions following in the wake of housing expansion and deforestation. The WHO, through the 51st World Health Assembly of 1998, set up the program 'Roll Back Malaria' and, with aid from UNICEF and the World Bank, is endeavoring to bring about economic and sanitary improvements in developing countries.

Cholera continues to be endemic in Asia, between India and the Far East, and also in warm zones in America and Africa. *Yersinia* (or *Pasteurella*) *pestis*, discovered by Yersin and Kitasato, is still to be found throughout the world, except in Western Europe, carried by rodents, rabbits, and related species. Bacteria are carried by black and brown house rats, *Rattus rattus*: rat fleas are the arthropod vectors transmitting 'epizootic' plague to humans from rodent hosts, but transmission among humans is also possible. Short-term human immunity requires frequent vaccination. Almost forgotten in the civilized world, there is nevertheless always the possibility that the devil may once more send his rats to the old, rich cities.

See also: Health: Anthropological Aspects; Middle Ages, The; Mortality, Epidemiological, and Health Transitions; Mortality: The Great Historical Decline

Bibliography

- Ballester R 1977 *La historia clínica pediátrica durante el siglo XIX*. Universidad Zaragoza, Spain
- Biraben J N 1975-6 *Les hommes et la peste*. Mouton, Paris
- Campillo C 1993-4 *Paleopatología*. Uriach, Barcelona, Spain
- Corbellini G, Merzagora L 1998 *La malaria tra passato e presente*. Università La Sapienza, Università di Cassino, Rome
- Crosby A W 1972 *The Columbian Exchange*. Greenwood Press, Westport, CT
- Engelhardt H T 1974 Explanatory models in medicine: Facts, theories, and values. *Texas Reports on Biology and Medicine* 32: 225-39
- Fischer-Homberger E 1979 *Krankheit Frau*. Hans Huber, Bern, Switzerland
- Kiple K F (ed.) 1993 *The Cambridge World History of Human Disease*. Cambridge University Press, Cambridge, UK
- Lain Entralgo P 1961 *Enfermedad y pecado*. Toray, Barcelona, Spain
- McNeill W H 1977 *Plagues and People*. Oxford University Press, Oxford, UK
- Peset J L 1993 On the history of medical causality. In: Delkeskamp-Hayes C, Gardell Cutter M A (eds.) *Science, Technology, and the Art of Medicine*. Kluwer Academic, Dordrecht, The Netherlands, pp. 57-74
- Peset M, Peset J L 1972 *Muerte en España*. Seminarios y Ediciones, Madrid, Spain
- Zinsser H 1934 *Rats, Lice, and History*. Bantam Books, London

J. L. Peset

Planetary Cartography

Planetary cartography is the making or study of maps of the planets, satellites, asteroids, and comet nuclei comprising our solar system. Maps of the Sun and other stars may be included for the sake of completeness, and future astronomers may prepare maps of planets orbiting other stars. The definition of 'map' is often biased towards modern terrestrial topographic maps, so for this article a broad definition is adopted similar to that of Harley and Woodward (1987): a map is a graphic representation designed to facilitate a spatial understanding of the object or region it represents. How many worlds have been mapped as of mid-1999? Several thousand maps of the Moon and Mars have been published, but maps of Pluto are limited to simple illustrations in journals. Asteroid 29 Amphitrite is portrayed in one simple sketch map, and the only published map of asteroid 1 Ceres is an illustration in a science fiction magazine. If Amphitrite is included but Ceres is not, 60 bodies in the solar system have been mapped in a scientific sense. When should a world be said to be mapped? It is accepted for this article that composites of multiple observations or images augmented with feature names or other in-

formation can be called maps, but not primary observational records (photographs or drawings).

The history of cartography can be portrayed as a linear evolution driven by successive developments in theory, technology and exploration, with modern scientific mapping as its zenith. Planetary cartography is particularly susceptible to this approach. Recent work (e.g., Harley and Woodward 1987) recognizes maps once dismissed as irrelevant and broadens the field to consider uses or aspects of maps overlooked in purely technical histories. Here both approaches are combined, useful references identified and areas indicated that will repay further study, for no comprehensive survey of this subject has yet been undertaken.

1. Background

The International Astronomical Union (IAU) oversees body and surface feature nomenclature, the definition of north and the locations of prime meridians, and publishes its decisions in its *Transactions*. Surface feature locations are determined by establishing a control network of points measured in multiple images of a world. Greeley and Batson (1990) summarize these technical issues. Many mapped bodies are near-spherical, but where they are not their shapes are estimated from stereoscopic images and the changing outlines of the object in views from different directions.

2. Pretelescopic Planetary Cartography

Markings on the Moon have always been visible, so representations of them are the earliest planetary maps. Drawings by Leonardo da Vinci dated about 1505 were long considered the oldest, but recently other ancient candidates have been suggested. Carvings at Knowth, Ireland (about 2800 BC) may represent the lunar markings, and one at the end of an interior passage might be illuminated at times by the rising Moon. The final form of Stonehenge, a circle containing a double horseshoe, may be a schematic map of the Moon akin to those at Knowth. Its solar alignment is most celebrated, but the full Moon occasionally rises on the monument's axis. Some Egyptian depictions of the sacred eye representing the Moon resemble the lunar markings. Plutarch gives several lunar place names. Some Arab world maps around AD 1100 resemble the lunar markings in reverse, recalling the ancient belief that the Moon is a mirror reflecting an image of the Earth. Van Eyck painted the Moon with markings in 1430, though they may be merely schematic. William Gilbert drew a lunar map with place names in about 1600, just before the invention of the telescope. Of uncertain age, not

ancient but probably drawing on a long tradition, are representations of the *maria* in native American art, including rock art in the Stein Valley, British Columbia and drawings of the Barasana culture, Colombia. The only other body which could be mapped without a telescope is the Sun, since sunspots had long been known in China, but no such map has ever been described.

3. Telescopic Planetary Cartography

Following the pioneering observations of Harriott and Galileo in 1609, lunar mapping advanced rapidly until the 1800s (Kopal and Carder 1974, Whitaker 1989). Successive cartographers sought mainly to improve accuracy and detail, since the Moon is unchanging. Riccioli's nomenclature became standard and was formalized by the IAU in the 1930s. Photographs eventually replaced more stylized drawings, culminating in several superb atlases produced under the direction of Gerard Kuiper (University of Arizona) at the dawn of the space age. Frequently overlooked in histories of lunar cartography are maps prepared by and for nonastronomers, such as Homann's 1720 map comparing the nomenclatures of Hevelius and Riccioli, and numerous small charts shown as insets in celestial or world maps from roughly 1650 to 1800. Some atlases of the 1800s also contain lunar maps. Though rarely more than poor copies of astronomers' map, they may have had a wider audience than their models. No systematic study of them has yet been made.

Other worlds were more difficult targets for telescopic observers (Greeley and Batson 1990). The first map of Venus, by Bianchini in 1728, misrepresented faint markings in the planet's opaque clouds as lands and seas. Similar maps were being drawn until the 1950s. Herschel first mapped Mars in 1783. Schiaparelli in 1877 measured feature locations for positional control, described the spurious 'canals' which influenced public opinion for decades, and introduced the modern feature nomenclature. The markings of Mars are variable, so maps are made whenever it is well placed for observation (Blunck 1982). Maps of Mercury, beginning with Schiaparelli's of 1889, were compromised by the assumption of an erroneous rotation period that was not corrected until the 1960s. Markings on the four largest satellites of Jupiter are at the limit of telescopic discernability, but a map was made by Douglass in 1897, and others followed in the 1900s. Ephemeral cloud features in Jupiter's atmosphere are plotted on what are essentially meteorological maps to monitor changes. Similarly, sunspots and other ephemeral features of the solar photosphere are mapped to record the Sun's changing appearance. These 10 bodies, excluding Earth, were all that had been mapped before the Space Age.

4. Space Age Planetary Cartography

Spacecraft have visited many worlds since 1959, and like terrestrial explorers these planetary missions generate many maps. Some are used for premission planning, some for operational use during the mission, and others provide a systematic record of discoveries made or activities undertaken. The subject has not yet been described in detail and will form fertile ground for future scholarship.

4.1 Cartography for Lunar Exploration

The Soviet probe Luna 3 photographed the Moon's far side for the first time in 1959. Later missions filled remaining gaps in coverage and revealed greater detail, permitting large scale mapping of some areas. Landing craft photographed their surroundings, which were mapped at large scales. The US Geological Survey (USGS), Army Map Service, and Aeronautical Chart and Information Center produced over 1,000 maps for the Apollo missions (Kopal and Carder 1974). Astronauts used maps for site selection, training, orbital navigation, and surface activities. In the Soviet Union, MIIGAiK and other agencies produced maps for their lunar missions, including robotic landers and rovers, but never undertook a systematic large-scale mapping program. Lunar mapping in both nations almost ceased in the mid-1970s, leaving a legacy of 1:250,000 and 1:1,000,000 scale maps as well as global and regional geological maps. Public interest in Apollo resulted in numerous popular maps, atlases, and globes in the 1960s, and for a time many general world atlases included a lunar chart. Most disappeared as interest in lunar exploration waned.

4.2 Cartography for Solar System Exploration

Following Apollo, robotic spacecraft travelled to other planets, all but Pluto having been visited by 1989. In the 1990s several asteroids were explored and other bodies were observed in greater detail. Almost every mission resulted in maps, from systematic geological and topographic mapping by USGS to large-scale charts of the immediate surroundings of landers on Venus or Mars. The Soviet Union attempted no systematic mapping, except at Venus, but its landers gave the only surface views of Venus yet obtained. Mapping by USGS up to about 1994 is summarized by Greeley and Batson (1990). Forty-two worlds had been mapped from planetary spacecraft data by 1999, 20 by USGS. One significant problem concerns small irregularly shaped bodies for which geodetic methods and map projections are still being developed. Many planetary maps since 1990 take the form of digital photomosaics and maps of radar reflectivity, elevation, etc. which may be distributed on CD-ROM or via the World Wide Web, reprojected to

any desired map projection and used in Geographic Information Systems. USGS has taken the lead in this, and many such products are now available for free download via NASA's Planetary Data System. The number of maps published on paper is declining, which will compromise historical studies. Public interest in Mars and the possibility of human expeditions before about 2020 led to the publication of several popular Mars maps in the late 1990s. These may be said to include maps of the Cydonia region of Mars illustrating unconventional concepts regarding the artificiality of features such as the famous 'face,' showing place names, alignments, and architectural or geological interpretations.

4.3 Telescopic Mapping in the Space Age

New instruments and techniques have enabled mapping of many worlds not visited by spacecraft, including Pluto and the asteroid 4 Vesta. Radar images of Earth-approaching asteroids have been used to derive shape models and maps, and comet nuclei have been mapped very tentatively by observing changing patterns of material ejected from their surfaces. These maps usually appear in scientific literature and are rarely seen by a nontechnical audience. Numerous stars should be included among the mapped celestial bodies, since methods have been devised for interpreting features in their spectra in ways that lead to simple maps of phenomena in their photospheres. Interferometers operating at visible wavelengths may eventually permit the mapping of surface features on planets orbiting nearby stars. Even in this solar system, the number of mapped worlds will exceed 100 not far into the twenty-first century.

5. Science Fiction Mapping

Some science fiction books and fan magazines contain maps reflecting both the stories themselves and contemporary knowledge of their settings. Like other popular maps they help shape public awareness of the planets. Post (1979) illustrates some examples, but the subject would repay greater attention. Here are considered only fictional maps of real worlds, Mars being the most prominent example. Fictional maps once showed canals and ruined cities, but now often identify lowlands which become seas on a future Mars made more earthlike by colonists. Notable among works illustrated in this way are the 'Barsroom' (Mars) novels of Edgar Rice Burroughs and magazines relating to his work, and the novels of Kim Stanley Robinson. A different class of planetary map has evolved to support role-playing games. These maps frequently include a mixture of fictional and real surface features, and have appeared in atlas and sheet map form as well as illustrations in game manuals.

6. Conclusion

Planetary cartography is a broad and rapidly evolving field with a long history. It has not yet been examined in its entirety. Maps record both the history of astronomy and space exploration and a much broader cultural response to those endeavours. A systematic effort to collect and describe this material is needed, including an attempt to preserve the increasing number of ephemeral maps on the Internet.

Bibliography

- Blunck J 1982 *Mars and its Satellites*, 2nd rev. edn. Exposition Press, Smithtown, NY
- Greeley R, Batson R M 1990 *Planetary Mapping*. Cambridge University Press, Cambridge, UK
- Harley J B, Woodward D 1987 *The History of Cartography*. University of Chicago Press, Chicago, Vol. 1
- Kopal Z, Carder R W 1974 *Mapping of the Moon*. Reidel Publishing, Dordrecht, Holland
- Post J B 1979 *An Atlas of Fantasy*, new rev. edn. Souvenir Press, London
- Whitaker E A 1989 Selenography in the seventeenth century. In: Taton R, Wilson C (eds.) *The General History of Astronomy*. Cambridge University Press, Cambridge, UK, Vol. 2

P. Stooke

Planning, Administrative Organization of

This article is about the role of planning in organizations, with a focus on urban planning in local government. It takes the view that it is best to analyze the general before focusing on the particular. To this end, a generic model is outlined, which sets the stage for the later discussion. The article concludes with alternative strategy options for coping with some unique aspects of the urban planning situation.

1. The Nature of Organizations

1.1 Organizations and Management

A human organization is one form of an abstract concept called a system. A system is an assembly of parts that form a whole greater than the sum of the parts. Every system has some unifying principle or force that holds it together. In the case of a human organization, this unifying element is provided by the management of the organization. Management in general refers to those people in an organization who are authorized to give direction to the other members of the organization. Understanding management is the key to understanding any organization.

The field of management is large and complex, and many theories compete for attention. However, most theorists agree, that some structure of power is a necessary part of management. The wellsprings of human motivation remain somewhat mysterious, and often people will cooperate voluntarily without coercion, but some sort of power to influence behavior, either by promise of reward or threat of reprisal, seems essential to the success of any organization. Without it, managers cannot be sure their leadership will be followed.

The problem facing management is that members of organizations are self-conscious human beings with an innate desire for individual freedom and gain, and an innate dislike of restriction and loss. They naturally compete for power, and for the benefits in freedom and gain they believe come with it. Therefore, every time a human organization is formed, and a hierarchy of management power is set up, a field of competitive tension is automatically created, as surely as a magnet generates a field of electromagnetism.

1.2 Culture of Organizations

For the purpose of discussion, this field of tension will be called the 'culture' of the organization. Defined many ways by scholars, the term 'culture' is used here to refer to the degree to which the members of an organization are willing to vest power and authority in management. Three somewhat arbitrary positions along a spectrum make useful benchmarks: (a) a culture of 'consent;' (b) a culture of 'cooperation;' and (c) a culture of 'dissent.'

In (a), organization members are accustomed to accepting the centralization of power in a management hierarchy, and look to specific directions from management for guidance on a wide variety of operations. In (b), members are accustomed to working together creatively under minimum management guidance, except for specific directions on matters that most can agree require a higher vision. In (c), members are accustomed to questioning management directions of all kinds, and frequently tend to place a higher value on individual independence than on group cooperation.

Obviously, the more a culture is oriented towards dissent, the more complex is the management challenge. To understand how managers have dealt with this problem, it is useful to construct a simple model of a generic organization, and see how variations in its structure relate to differences in internal cultures and external conditions. The model presented here is based on four tasks that can be said to constitute the primary responsibilities of management.

1.3 Management Task 1: Mission

Mission is the management task of defining and communicating the basic purpose for the creation and

existence of the organization. Its role is to act as a symbol of values, an orientation for planning, and an inspiration for action.

1.4 Management Task 2: Structure

Structure is the management task of planning and implementing the form of the organization. In practice, this usually means dividing the constituent members into clusters, called administrative units, and assigning a manager to each cluster. The result is often portrayed as an 'organization chart,' a diagram showing the administrative units, with lines linking them vertically to either a single person (called here the 'executive,' or a small group (called here the 'board,' or to both, at the top of the chart. Together these two entities, plus their support staff, constitute the management 'apex.'

The vertical linkage always takes the shape of an upside down tree. This tree represents the way that all members of the organization relate to each other in terms of their authority, that is to say, their legitimate power to exercise control over subordinates. A hierarchy controlled by an apex is always necessary, because organizations need a way to reconcile disputes in a timely manner.

1.5 Management Task 3: Procedure

Procedure is the management task of planning and implementing the functions of the organization. In practice, it means controlling the behavior of individuals within the organization to ensure that plans are executed and products are produced efficiently and effectively.

Systems theory suggests that human organizations, like biological forms of life, can be described in terms of function, as linear process (with some internal feedback loops) from 'input' to 'throughput' to 'output,' and that its output can be divided into two components, one called 'product' one called 'waste.' This simple model is popular in the fields of economics and business management, where it provides a foundation for the idea of using product to measure the functional performance of organizations of all types.

1.6 Management Task 4: Product

Product is the task of designing the products that the organization creates and exports into its external environment. Product in this sense is an extension of the concept expressed in the term 'mission.' The products of an organization should represent a more focused and defined expression of the values and directions implicit in its mission statement.

The concept of product has come to play a critical role in analyzing and comparing the relative value of different organizations. Two criteria are used, one focused on the internal environment of the organization 'efficiency,' and the other on its external environment ('effectiveness'). Both criteria use the product as a surrogate by which to measure the organization as a whole. Efficiency is the ratio derived from dividing the product by the input. It is a quantitative measure. Effectiveness is the degree to which the product satisfies the mission of the organization with regard to its external environment. It is a qualitative measure.

2. Model of a Generic Organization

Summing up the key factors outlined in Sect. 1 a graphic model can be visualized. The structure of the organization is represented by a triangle, with its vertical axis reflecting the authority hierarchy of management. The procedure of the organization is represented by a horizontal line across the bottom of this triangle, reflecting the flow of work transactions through the organization from input to output. Operators comprise the bottom layer of the triangle, performing the actual transactions that create the product as it moves from input to output. They are supervised by managers in the upper layers of the triangle.

The mission and structure responsibilities of management fall along the vertical axis of the triangle, and the procedure and product tasks fall along its horizontal axis. Mission and structure, once established, are not often changed, whereas product and procedure must be changed frequently to keep the organization in balance with its changing external environment. Structure is a key factor influencing product and procedure. Three basic variations in structure design can be observed in practice, as follows.

2.1 Structure Alternative 1: The Divisional Model

In this model, the basic principle for clustering people together into administrative units is their degree of commonality with regard to technique, skill, knowledge, or activity. This creates a set of what may be called 'functional' divisions, each one centered on a different type of specialized operation. The effect of this model is to allow the apex to delegate authority over specialized operations to subordinate managers within the hierarchy, while still retaining strong direction over the design and production of the composite final product.

This model emphasizes quality of input, by setting up hierarchies of specialized knowledge and skill that can apply technology to improve the product. Power over all four management tasks is strongly concentrated at the apex. This model can be very efficient,

but depends on a culture of consent, fast internal communication linkages, and a comprehensible external environment. It can lose effectiveness whenever any of these supporting conditions break down.

2.2 *Structure Alternative 2: The Decentralized Model*

In this model, the basic principle for establishing administrative units is to include in each division all the special resources necessary to produce a complete particular product. This model creates a set of 'product' divisions that share all of the characteristics of the parent organization, with the exceptions that: (a) each division produces a different product; and (b) enough freedom is withheld by the Apex to ensure the health of the parent organization.

This model emphasizes quality of output, by setting up hierarchies of situational analysts who can adapt product designs to changing external situations. Power over product and its procedure is delegated by the apex, but control over mission and structure is retained. This model can be very effective in responding to local and rapidly changing product needs, but it depends on a culture of consent with regard to mission and structure, a greater aptitude for situational strategy by divisional leaders, and sufficient supplies of input resources to duplicate them across every product division. It can lose effectiveness whenever any of these supporting conditions breaks down.

2.3 *Management Structure 3: The Matrix Model*

In this model, the organization is divided into a set of 'functional' divisions as well as a set of 'product' divisions. The apex delegates control of both product and procedure, but avoids concentrating them entirely in the product divisions, by requiring them to share control of these tasks with the functional divisions through negotiated divisional agreements.

This model tries to achieve a balance between quality of input and quality of output, by delegating power in both directions. Power over mission and structure is retained at the apex. This model can be very effective in facing situations that call for the strengths of both the divisional model and the decentralized model. However, it heavily depends on a culture of cooperation in product and its procedure, and a culture of consent in mission and structure. Without managers and operators who value these cultures, this model can be difficult to sustain over long periods of time.

3. *The Place of Planning in the Generic Model*

3.1 *Planning vs. Implementation*

Planning is a process of choosing goals and ways to achieve them, 'before initiating action.' It is one of the

basic activities of the human mind. As long as the process of planning and the initiation of action constitute sequential events within a single human mind, there is no need to reflect on the relationship between the two events. However, as soon as the planning is performed by one person, and the action to approve and implement the planning by another, the question arises: How shall the two events be connected? How this question is answered is the key to the administrative organization of planning.

3.2 *The Generic Planning Process*

The process of planning for a client begins with 'analysis' and ends with 'synthesis.' As defined here, it involves six sequential actions, the first three of which are analytic, and second three of which are synthetic:

- (a) 'Survey' the nature of the current situation;
- (b) 'Forecast' trends, assuming no purposeful intervention by the client;
- (c) 'Compare' the trend scenario to the client's needs and desires;
- (d) 'Invent' possible alternative scenarios achievable through intervention;
- (e) 'Evaluate' the scenarios for feasibility and fitness;
- (f) 'Integrate' alternative scenario elements into one final version, 'the plan.'

The planner can adopt one of two modes while performing this process. All six steps can be performed by the planner alone, with only the sixth step, 'the plan,' passed on to the client. Call this the 'advocacy mode.' Or, the client, and other interested parties as well, can be invited to participate in some, or all, of the steps. Call this the 'participatory mode.'

3.3 *The Base Case: Planning and Implementation Combined*

The least political complications ensue when the divisional model is used in a consent culture, and planning is performed solely by the management apex. Military organizations typically use this model this way, largely because they face such threatening external environments that internal control and secrecy in planning become cardinal virtues.

Certain kinds of business organizations also use the model this way, especially those facing fierce competition analogous to the military situation. However, businesses in the modern world increasingly seem to find it more useful to foster a cooperation culture than a consent one, and, to that end, often shift to some form of the decentralized or matrix models.

Government organizations, by contrast, especially in democracies, increasingly find themselves facing cultures of dissent in the larger organizational unit (i.e., the nation or municipality) for which they function as management. Within individual govern-

ment agencies, however, management can often develop a culture consent or cooperation, because members of the agency voluntarily contract for employment knowing the management climate of the place.

In summary, evidence from the field suggests that the model to which management naturally gravitates is the divisional one, with the synthetic steps of the planning process performed directly by the apex, supported when necessary by a delegation of the analytic steps to a subsidiary unit without hierarchical authority. Some theorists suggest that this model represents the default mode to which all organizations tend to revert, especially when facing external pressures, such as strict performance evaluation, or strong technical competition at the output end, or just scarce resources at the input end.

3.4 Delegated Planning and the Administrative Planning Unit

When external conditions overburden the planning capacity of the apex in a divisional organization, the easiest approach is to expand total capacity by setting up an administrative planning unit (APU). This may be viewed as a lesser delegation of power than is involved in re-engineering the structure of the organization into a decentralized or matrix model, but it is a delegation of power nevertheless, and a power that is potentially very broad in scope. It is easy for other administrative units in the organization hierarchy to become nervous or hostile. Like management itself, every APU must work in a power struggle.

Since the apex usually wishes to retain direct control of the organization's mission and structure, the delegated planning responsibility usually only applies to product and/or its procedure. In manufacturing organizations the former is often called 'research and design,' and the latter 'quality control.' In either case, although the plans of the APU affect the entire organization, the APU's most important client, in terms of plan implementation, is always the Apex.

3.5 The APU in Decentralized and Matrix Models

In the decentralized and matrix models, the responsibility of planning for product and procedure is already delegated in large measure to the individual product divisions. If the apex wishes to exert control or influence over the product units in these divisions, it may set up an APU to fill some planning role that is intermediary between the mission and structure role of the apex and the product and procedure role of the divisions. When this happens, the APU role is typically

concentrated more on the analytic steps of the planning process than on the synthetic ones, or alternatively is focused on operational techniques for supporting a cooperation culture.

4. The Administrative Organization of Urban Planning

4.1 Urban Planning

Urban planning is a field and profession that arose after 1900, primarily in industrial nations. A search began to find methods to ameliorate the negative physical and social conditions of the industrial revolution, especially in urban areas. The concept arose that a good social reality needs the support of a good physical reality, and that the latter can be achieved by adopting a plan for the physical layout of the urban place. The idea was born that one of the major products of local government should be a spatial plan for the physical environment of its geographical area of jurisdiction.

In time, it became evident that a spatial plan by itself could not adequately deal with problems in the physical, social, and environmental systems of urban areas. Theorists concluded that other 'products' of local government (such as rules to regulate traffic, pollution, housing, and employment) were needed to supplement a spatial plan. Today, a wide typology of both spatial and operational plans are in use, as governments around the world seek to improve the form and function of their urban places for a wide variety of reasons. The challenge of getting such plans approved and implemented is made greater by one peculiar characteristic of urban planning that makes it qualitatively different from the kind of planning that all other types of human organizations engage in.

Two observations make the point. First, an organization's product is generally considered to be a creation that is extruded from the interior environment of the organization into its external environment. It is aimed at the exterior, not the interior, of the organization. Second, this product is much smaller in scale than its external environment, whether it be a battle plan for execution by a military organization, a computer or a software training program for sale by a business organization, or schools and police protection provided by a local government.

By contrast, the urban planning product is typically far larger in scale than these. More significantly, it is aimed, not externally, but internally—at the organization's own members. In fact, it is a plan to govern an important part of the lifestyle of the entire human society for which the government is the management. Because this plan is so sweeping in its internal effect, getting it approved and implemented is far more difficult than is the case for the typical

product plan in other organizations. It is more like the adoption of a labor-management agreement than the approval of a new automobile design.

4.2 Government Organizations

Other challenges to urban planning also arise because of the unique conditions under which government organizations must operate. These need some explanation. The mission of government organizations is generally to maintain a healthy and stable society. Typically, they consist of at least two tiers: national and local. National governments often face turbulent, complex, and hostile external environments, and tend to maintain a strong military function. For this and other reasons, most national governments seem to be organized structurally in the divisional model.

By contrast, local governments would seem to face less hostile challenges from their external environments, but, in fact, they must cope with some other conditions that are common to both national and local governments, and that may affect local governments more. The most significant of these are: (a) a limited managerial role within their organization (i.e., their constituency or society); (b) uncoordinated apex members; and (c) a culture of dissent. Together, these three conditions create a management situation of greater complexity than is the case for most nongovernmental organizations.

Regarding limits on governmental power, local governments are usually limited by statutes, in many more ways than is common for the management of nongovernmental organizations. This limitation of authority rests on a foundation of accepted social values, the most prominent of which are the importance given to protecting individual human rights (in democratic societies), and the importance given to protecting private economic markets (in capitalist societies).

Regarding uncoordinated apex members, democratic societies often create government structures that separate to various degrees the power of the board (the legislature) from the power of the executive (the president, governor, etc.), for the purpose of avoiding an excess concentration of power at the apex.

Regarding a culture of dissent, civil societies can often drift towards this end of the cultural spectrum for a variety of reasons, including problems created by the pace of technological change, or international relations, or changing human aspirations, or other sociological factors.

4.3 Urban Planning in Local Government

It is usual to find urban planning delegated to an APU, within a divisional local government structure. This APU inherits the same situational factors com-

mon to all APU's in divisional structures, but in addition, must deal with the unique conditions outlined above. In sum, the APU must find a way to get approval and implementation of its product in the face of: (a) the enormous social impact of its product; (b) a culture of dissent; (c) limited authority of government generally; (d) a fractured apex; and (e) competition for power with other units of government. What are some alternative approaches?

4.4 Strategies for Urban Planning APU's

The most direct approach of course, is to seek the approval of the apex, followed by the issuance of an apex command protocol throughout the organization. This approach has the merits of simplicity and efficiency, and can work well in situations where the scope of social change embraced by the plan is not too big, the cultural dissent level is not too high, the apex is not too uncoordinated, and the power struggle with other governmental units can be won. It fits well with the advocacy mode outlined earlier for the planning process method. Call this the 'apex approach' and the 'stability scenario.'

A second approach has better chances of success when the social change gets bigger, the dissent level gets higher, the apex is more uncoordinated, and the internal power struggle outcome is uncertain. In this approach, the APU needs to shift its planning process to the participatory mode, and reach out early in the process to seek the participation and counsel of all the key managers and operators within the government structure, as well as key movers and shakers within the general society. Call this the 'stakeholder approach' and the 'complexity scenario.'

A third approach may be necessary when the social change is very large, the dissent level is very high, the apex is seriously fractured, and the interagency power struggle is intense. In this approach, the APU expands the outreach scope of its participatory mode planning process to engage the support of as many individuals in the wider society as it can reach, on the theory that this third tier group may be persuaded to influence the stakeholders and the apex to approve and implement the plan produced by the APU. Call this the 'constituency approach' and the 'conflict scenario.'

4.5 Relationship of Organization to Strategy for Urban Planning APU's

Where the charter of the APU does not specify the structure of the unit, the unit's executive (usually called the planning director) can choose an organizational form. In small APU's in a stability scenario, the prevailing practice seems to mirror the tendency of other organizations to adopt the divisional form and use the advocacy mode. In a small

APU, the director is personally under a lot of external pressure, and, like a beleaguered military commander, may feel the need for internal secrecy and control, especially if the situational scenario shifts from stability to complexity or conflict.

In large APU's, especially within a complexity scenario, there are benefits to adopting the decentralized form and using the stakeholder approach, but they depend on developing an internal cooperation culture. By setting up different units to plan for subsections of the society, the scale of the planning product becomes more manageable. With responsibility for product delegated, the director is under less immediate political pressure. The weakness of this approach only surfaces when there are insufficient resources to supply all the divisions with adequate technical knowledge, and/or the need arises to make the divisional plans consistent with each other from a larger perspective.

Conflict scenarios present the most difficult situations to deal with, especially if the technical or consistency component of the plans is important. It is dangerous to use the constituency approach, because the APU may lose apex support if the elected officials feel threatened by this outreach into their electorate. However, if the director can build an internal culture of cooperation sufficient to support a matrix form, it is possible to blend a set of procedures and products adequate to the challenge, but, it is no sure thing. In the final analysis, urban planning is an art, not a science—in administrative organization as well as urban form.

See also: Administration in Organizations; Decision Making, Psychology of; Leadership in Organizations, Psychology of; Leadership in Organizations, Sociology of; Management: General; Organizational Climate; Organizational Culture; Organizational Culture, Anthropology of; Organizational Decision Making; Organizations, Sociology of; Planning Theory: Interaction with Institutional Contexts; Urban Places, Planning for the Use of; Design Guide

Bibliography

- Churchman C W 1979 *The Systems Approach*. Dell, New York
- Drucker P F 1973 *Management: Tasks, Responsibilities, Practice*. Harper and Row, New York
- Frame J D 1995 *Managing Projects in Organizations*. Jossey-Bass, San Francisco, CA
- Galbraith J R 1977 *Organization Design*. Addison-Wesley, Reading, MA
- Grove A S 1983 *High Output Management*. Random House, New York
- Hatch M J 1997 *Organization Theory*. Oxford University Press, Oxford, UK
- Kaiser E J, Godschalk D R, Chapin F S Jr 1995 *Land Use Planning*. University of Illinois Press, Urbana, IL

- Keidel R W 1995 *Seeing Organizational Patterns*. Berrett-Koehler, San Francisco, CA
- Miller J G 1978 *Living Systems*. McGraw-Hill, New York
- Mintzberg H 1989 *Mintzberg on Management*. The Free Press, New York
- So F S, Getzels J 1988 *The Practice of Local Government Planning*. International City/Country Management Association, Washington, DC

R. E. Tustian

Planning Ethics

Planning ethics refers to the integration of ideas from the fields of planning and moral philosophy, or ethics. As such, it covers a vast terrain of issues in planning theory and practice. Contrary to terms such as 'ethics in planning' or 'ethics of planning,' planning ethics does not privilege either planning or ethics; nor does it connote only a descriptive representation of values in planning. Instead, it is intended to depict the fact that planners and planning can benefit from incorporating ideas from moral philosophy into planning thought and professional endeavors. Further, ethics can also benefit by gaining insight as to how concepts in moral philosophy actually work in practice. Both disciplines may then be more effective in prescribing courses of action in human settlements (planning) and developing theoretical approaches of how persons ought to behave (ethics).

1. Ethical Content in Planning

Planning ethics may be divided according to several different analytic schemes. Wachs (1985) developed a particularly comprehensive and useful typology in this regard. As supplemented by Hendler (1995), it represents a way of discussing planning ethics with reference to five aspects of planning theory and practice. These attributes are: everyday behavior, plan making, administrative discretion, planning techniques, and normative planning theory. In each category are examples of descriptive analyses in which researchers describe and discuss the values planners actually have (e.g., Howe 1995, Gowdy and Hendler 1999) as well as normative analyses where planners advocate particular approaches to planning ethics (e.g., Beatley 1994, Harper and Stein 1993).

1.1 Everyday Behavior

Everyday behavior pertains to the fact that planners are, as persons or moral agents, subject to the same ethical guidelines as everyone else, although their professional status may give rise to particular rights

and responsibilities. Thus, planners may be judged according to the consequences of their actions (consequentialist ethics) or with reference to the inherent characteristics of the actions themselves (nonconsequentialist or deontological ethics). Consequentialist approaches to planning ethics include most prominently utilitarianism with its focus on creating the greatest amount of utility from either individual actions (act-utilitarianism) or actions arising from the application of ethical rules (rule-utilitarianism). Nonconsequentialist approaches, on the other hand, include Kantian ethics with its interest in the ethical treatment of persons as ends and not (only) means. While Kantian and utilitarian perspectives continue to attract much attention in philosophical and planning circles (e.g., Howe 1990, 1995), other consequential and nonconsequential ethical approaches have also proven attractive to planners and philosophers alike; these include, but are not limited to, communitarianism (e.g., Blanco 1995) which emphasizes community practices and virtues, environmental ethics (e.g., Beatley 1994, Jacobs 1995) which focuses on the relationships between and among humans and nonhuman components of our environments, and feminist ethics (e.g., Hendler 1994, Ritzdorf 1995) which attempts to incorporate into our considerations of moral issues the notions of care, trust, nurturing, and eradicating all forms of oppression. Communicative or discourse ethics (e.g., Benhabib 1992), too, has become an important part of contemporary ethical thinking and refers to procedural accounts of how one should make ethical decisions, with the presumption that, if one adheres to the rules guiding such a procedure, the outcome should, by definition, be good or right. This approach may be seen as underlying models of planning processes that emphasize consensus building and/or communicative rationality (e.g., Forester 1989).

Considerations of planners' everyday actions that are based on these sorts of ethical ideas include evaluations of behaviors corresponding to planners as professionals and nonprofessionals. Professionals are often seen as having added moral obligations in contrast to nonprofessionals, given their status in society. Doctors, lawyers, engineers, architects, and planners, among others, are trusted with going beyond motivations of profit and self-interest and focusing instead on the public interest. This status is accompanied by particular responsibilities for specialized knowledge not easily accessible to nonprofessionals, and self-regulation in the face of this specialization (Bickenbach and Hendler 1994). Specific behaviors addressed in this first category of planning ethics include, among others: avoiding conflict of interest; maintaining appropriate relationships with clients, employees, and colleagues; advertising one's services in certain manners; practicing within one's area and scope of competence; providing independent professional advice; and treating information appropriately

whether in terms of confidentiality or disclosure in reference to considerations of the public interest. Providing ethical guidance for planners facing decisions in these and related areas is central to this first aspect of planning ethics that focuses on everyday behavior.

1.2 Plan Making

Plan making is that part of planning having to do with the development, implementation, monitoring, and evaluation of plans, programs, and policies. Together, these sorts of normative statements have ethical implications in that certain amenities or disamenities are distributed or redistributed among citizens (and nonhuman components of our environments) in ways that benefit certain individuals or groups and impose costs on others. Assessing the implications of such decisions and actions pertains directly to the ethical theories and approaches listed above. Kantian-based ethics, for example, may be applied via a Rawlsian perspective in which resources are said to be distributed equally or, if inequity exists, so as to improve the situation of disadvantaged segments of the population in question. Examining the consequences of a planning document using a utilitarian calculus of costs and benefits is also common (see Sect. 1.4). Whichever ethical theory is used, however, is immaterial to the central idea that almost every decision made by planners potentially provides benefits and harms to individuals, groups, neighborhoods, or jurisdictions, and that this result has ethical implications associated with it. The importance of examining the (usually implicit) underlying assumptions and resulting implications of the ethical directions of plans, programs, and policies is the focus of this area of planning ethics.

1.3 Administrative Discretion

Administrative discretion refers to the ambiguity in planners' roles in their working environments. A number of roles have been proposed, ranging from advocate to technician to mediator to social learner (e.g., Gunton 1984). The fact that a planner's role is not discretely and concretely prescribed via the planner's professional organization and/or work environment means that considerable latitude may exist in the planner's determination of an appropriate or desirable way-of-being in professional planning work. Planners thus may exercise discretion in selecting for themselves a role (or roles) but this decision must be subject to ethical scrutiny. Whether, for instance, a planner may in fact explicitly advocate for a particular point of view is a controversial issue in this area of planning ethics. While, as suggested in the discussion of plan making above, any plan or program or policy does necessarily advocate for one course of action as opposed to

others, what and how such directions are advocated is sometimes seen as problematic within planning organizations.

In addition to choices among roles is the fact that administrative discretion highlights the issue of accountability in that planners are not elected officials and yet they may represent the interests of one group over another at any one point in time and in any given plan-making exercise. While such an affiliation may be couched in terms of the public interest, any value-laden connection between planning and developers on the one hand and environmentalists or social justice advocates on the other, for example, may be regarded as an improper exercise of the planner's discretion regarding that planner's role. Despite the fact that this representation of planning work is rooted in the largely outmoded conception of planning as value-free technical work, planners must acknowledge and address the fact that their discretion may be perceived as potentially troublesome.

1.4 Planning Techniques

Planning techniques are those methods or processes followed by planners in preparing or evaluating their plans, programs, and policies. Planners have a considerable array of analytic techniques at their disposal. Cost-benefit analysis, risk assessment, environmental assessment, program evaluation, participatory methods, and many others may be included in this regard. These techniques may aid a planner in deciding a course of action to follow or in assessing the effects of particular actions on different publics. Most of these techniques, and perhaps especially cost-benefit analysis, have been connected conceptually and methodologically to utilitarianism, as follows. First, utilitarianism relies upon the quantification of 'goods' and 'harms' so that both can be summed and compared. This is a starting point, too, for cost-benefit analysis in which costs and benefits are identified, measured, added, and contrasted in order to determine whether a course of action would result in more benefits than costs. Second, this sort of conclusion is clearly based on utilitarian thinking in which a maximum amount of good is sought.

The fact that planning techniques are linked to ethical theories is not particularly troubling. This kind of observation attests to the post-positivist world in which we find ourselves where the planner is not regarded as a value-free technician but as an individual who necessarily brings certain biases, values, and conceptual lenses to their work. Similar, then, to contemporary views of science, defining one's perspective and unearthing one's assumptions is an essential first step in doing research and in doing planning. While this approach to science and to planning continues to be debated, an issue arising

from the mere association of certain ethical approaches to certain planning techniques remains. This is the fact that techniques based on consequential or utilitarian thinking can be seen as over-represented in planning ethics in that methods based more on nonconsequential thinking do not have the prominence of their philosophical cousins. One branch of moral philosophy thus gets short shrift in helping planners to develop and evaluate plans—a branch that is generally regarded as just as viable and important as consequential approaches. Lake (1993) discusses one aspect of this phenomenon in the context of feminist ethics and techniques based on geographical information systems (GIS). He argues that the data-aggregation focus on GIS is incompatible with the focus on the 'concrete other' (Benhabib 1992) in feminism. The same could be said of other ethical theories that attempt to highlight individual experiences and inherent value (the 'right') as opposed to formulaic assessments of consequences (the 'good'). Uncovering assumptions and seeking a broader distribution of ethical underpinnings for planning techniques is of central importance to this area of planning ethics.

1.5 Normative Planning Theory

With its focus on the nature and intent of planning as a profession, as well as its scope or mandate, normative planning theory falls clearly within the realm of ethical analysis. What the planning profession *should* profess (Haworth 1984) is a question that demands this sort of scrutiny.

There have been many disparate suggestions regarding the ethical mandate of planning. Udy (1980) proposed that the four eternal values of goodness, beauty, truth, and love lay at the heart of planning and should be recognized as the normative direction for planning initiatives. Blanco (1995) developed an alternative list that included reason, community, environment, and democracy. Krumholz and others (e.g., Krumholz and Forester 1990) have written and practiced extensively in the area of equity planning, and Davidoff (1978, pp. 69–70) forcefully asserted that,

if a planner is not working directly for the objective of eradicating poverty and racial and sexual discrimination, then she or he is counterproductive. If the work is not specific in its redistributive aims, then it is at best inefficient. If the work is not aimed at redistribution, then a presumption stands that it is amoral. These are strong words. They must be. So long as poverty and racism exist in our society, there is an ethical imperative for a single direction in planning.

In the face of these conflicting sets of values, the field and profession of planning continue to have a normative nature that can only be called ambiguous. Each professional planning organization defines for itself its

central ethical values and then applies them in creating a regulatory structure for its practitioners. In contrast, then, to the fundamental pursuit of health in medicine and the quest for justice in law, planning cannot express its basic ethical norms as succinctly nor with as much certainty. Developing such a rubric or becoming more accepting of a diverse set of principles of moral guidance is a key part of this last aspect of planning ethics.

2. *Regulating Planning Ethics*

All of these ethical aspects of professional behavior are addressed in the codes of professions such as planning (see, e.g., Hendler 1990). Codes of conduct, especially, provide ethical guidance in the form of rules, sanctions, and encouragement towards, or away from, particular behaviors. Codes of ethics are more general and are usually seen as articulating a normative basis for the profession in the form of key values that are thought to be central to the professional community in question. In planning, these values often include environmental integrity, public participation, holistic decision making in terms of temporal, geographic, and disciplinary scope, and a respect for a social justice and a multitude of diversities. While not easily enforceable, codes of ethics provide for a unified voice of the moral intent of a profession and contribute to community building and identity formation within a profession such as planning.

Codes of conduct are used in assessing the behavior of individual professionals and, if necessary, imposing sanctions if the rules under scrutiny are found to have been violated. Suspension, financial penalty, expulsion, and education/counselling are examples of the kinds of sanctions levied by professional organizations. In that planners rarely have to be certified to practice their work, considerable debate is directed towards the licensing of planners and attendant issues regarding their discipline. While some planning institutes are content to regulate only their own members and show little concern for the behaviors of non-members, others have endeavored to expand their membership and refine the notion of planning so that only their members may refer to themselves as professional planners.

Together, codes of ethics and conduct may be seen as a bridge linking abstract moral theories and practical work such as planning in that the sorts of ethical issues discussed here are articulated precisely and resulting discussions may shed light on the implications of different ethical ideas 'on the ground.' As such, codes are growing and developing documents that have evolved significantly during the past two decades of planning ethics discussions. They themselves have also stimulated much in the way of discussion regarding the efficacy of the values they present as being good or right (e.g., Lucy 1988).

Through the processes employed in formulating a code and in subsequent discussions and revisions, planning ethics has enjoyed much in the way of constructive critique (e.g., Kaufman 1990).

Some planning organizations go beyond codes *per se* in educating their members about ethics and contributing to the interpretation of their codes by publishing ethical advisories that apply given sections of a code to a particular planning issue (e.g., AICP 1996). As the sophistication of planning ethics discourse increases, codes and planning practice can only be expected to benefit from an enhanced understanding of, and perhaps consensus about, the fundamental values that underlie planning activities.

3. *Planning Ethics in Theory, Practice, and Education*

Discussions of planning ethics have been most pervasive in planning theory where philosophical analyses of planning and planners are commonplace. It is here that arguments regarding the central tenets and values of planning have arisen, and varying normative models of planning have been developed. Planning practice, however, has only recently begun to emphasize planning ethics in the form of increasingly comprehensive and thoughtful professional codes, as well as enhanced disciplinary proceedings. Empirical analyses of planners' values and accompanying ethical approaches have assisted in our constructing snapshots of the moral terrain in which planners have placed themselves. While consensus is not apparent, certain loci of values around environmental integrity, social justice, and democratic decision making often emerge. In addition to such descriptive studies are normative assessments of planning activities that include both subfields of planning such as transportation, environment, education planning, waste management, etc. (e.g., Hendler 1995, Beatley 1994), as well as specific planning behaviors (e.g., Howe 1995).

Planning education, which usually has the intent of both training future professionals as well as being part of universities, where intellectual growth and development is a priority, has begun to 'mainstream' discussions of planning ethics in its curricula. Far from inculcating particular values, teaching about ethics in planning programs usually has the goal of helping students engage in rigorous ethical analysis as well as preparing them to practice in a profession which espouses certain principles. Thus, values are presented along with the means to examine them critically. This performs the dual function of satisfying planning institute expectations regarding the education of future professionals as well as the more academic intent of teaching students how to think clearly and coherently. Having the skill of developing and critiquing logical arguments helps students and professionals work through ethical dilemmas they will certainly

encounter in their employment and contribute to the increasingly sophisticated nature of planning ethics—theory, practice, and education.

4. Challenges in Planning Ethics

While the value-laden nature of planning endeavors is generally accepted, certain challenges and questions remain. Examples of these include the following. The similarities and differences between public and private planners continue to be analyzed in terms of their moral obligations. The fast-encroaching world of computer technologies and associated planning applications remains ill examined in terms of the fit of these technologies and the moral intentions of the planners and the planning profession they serve. As planning becomes broader substantively with focuses on land use planning but also on social planning, health planning, environmental planning, etc., it remains to be seen whether a single professional organization may be appropriate to govern planners' activities. Further, addressing planning issues in ways that not only accommodate but also celebrate racial, ethnic, sexual, and other forms of diversity continues to occupy the interests of many people who write about and practice planning. Current emphases on planning processes and public participation contribute to resolving such issues but do not address questions regarding such processes as issues of identity, representation, method, and practical considerations regarding the efficient use of planning resources. Empowerment, while often discussed in the context of planning theory, is seldom found in planning codes. Economic development, though central to the 1990s climate of planning, is not well represented in the ethical guidance supplied to planners. And so on. Clearly, planning will continue to have much to consider in the way of discussions regarding its ethical mandate and the methods to be employed in addressing the issues that arise within the context of its thought and practice.

As ethics has highlighted for planners the values in their work, and the need to address critically the values that lie at the base of their theories and their practices, so, too, has planning underscored the importance of discussing ethical issues with an eye to the application of any theoretical stance based in moral philosophy. Thus, planning ethics, as an interdisciplinary field of inquiry and one which does not rest on the laurels of either 'parent' discipline (i.e., planning or ethics), planning ethics promises to contribute to an especially critical and grounded approach to posing, and answering, the sorts of questions and challenges listed here.

See also: Environmental Conservation: Ethical Concerns; Environmental Justice: Ethical Practices, Institutional Oversight, and Enforcement; United States Perspectives; Ethics and Values; Neighborhood Re-

vitalization and Community Development; Planning Issues and Sustainable Development; Planning, Politics of

Bibliography

- American Institute of Certified Planners (AICP) 1996 *AICP Code of Ethics and Professional Conduct*. AICP, Washington, DC
- Beatley T 1994 *Land Use Ethics*. Johns Hopkins, University Press, Baltimore, MD
- Benhabib S 1992 *Situating the Self*. Routledge, New York
- Bickenbach J, Hendler S 1994 The moral mandate of the 'profession' of planning. In: Thomas H (ed.) *Values and Planning*. Avebury Technical, Aldershot, UK, pp. 162–77
- Blanco H 1995 Community and the four jewels of planning. In: Hendler S (ed.) *Planning Ethics. Readings in Planning Theory, Practice and Education*. Center for Urban Policy Research, New Brunswick, NJ, pp. 66–82
- Davidoff P 1978 The redistributive function in planning. Creating greater equity among citizens of communities. In: Burchell R, Sternlieb G (eds.) *Planning Theory for the 1980s*. Center for Urban Policy Research, New Brunswick, NJ, pp. 69–72
- Forester J 1989 *Planning in the Face of Power*. University of California Press, Berkeley, CA
- Gowdy A, Hendler S 1999 Environmental values and professionalism in planning: a case study. *Plan Canada* 39(2): 28–32
- Gunton T 1984 The role of the professional planner. *Canadian Public Administration* 27(3): 399–417
- Harper T, Stein S 1993 Normative ethical theory: is it relevant to contemporary planning practice? *Plan Canada* September: 6–12
- Haworth L 1984 Orwell, the planning profession, and autonomy. *Environments* 16(2): 10–15
- Hendler S 1990 Professional codes as bridges between planning and ethics: a case study. *Plan Canada* 30: 22–9
- Hendler S 1994 Feminist planning ethics. *Journal of Planning Literature* 9(2): 115–27
- Hendler S (ed.) 1995 *Planning Ethics. A Reader in Planning Theory, Practice and Education*. Center for Urban Policy Research, New Brunswick, NJ
- Howe E 1990 Normative ethics in planning. *Journal of Planning Literature* 5(2): 123–50
- Howe E 1995 *Acting on Ethics in City Planning*. Center for Urban Policy Research, New Brunswick, NJ
- Jacobs H 1995 Contemporary environmental philosophy and its challenge to planning theory. In: Hendler S (ed.) *Planning Ethics. A Reader in Planning Theory, Practice and Education*. Center for Urban Policy Research, New Brunswick, NJ, pp. 83–103
- Kaufman J 1990 American codes of planning ethics: content, development and after-effects. *Plan Canada* 30: 29–34
- Krumholz N, Forester J 1990 *Making Equity Planning Work*. Temple University Press, Philadelphia, PA
- Lake R 1993 Planning and applied geography: positivism, ethics, and geographic information systems. *Progress in Human Geography* 17(3): 404–13
- Lucy W 1988 American Planning Association's ethical principles include simplistic planning theories. *Journal of the American Planning Association* 54: 147–9
- Ritzdorf M 1995 Feminist contributions to ethics and planning theory. In: Hendler S (ed.) *Planning Ethics. A Reader in*

Planning, Theory, Practice and Education. Center for Urban Policy Research, New Brunswick, NJ, pp. 104–19
Udy J 1980 Why plan? Planning and the eternal values. *Plan Canada* 20(3/4): 176–83
Wachs M (ed.) 1985 *Ethics in Planning*. Center for Urban Policy Research, New Brunswick, NJ

S. Hendler

Planning Issues and Sustainable Development

‘Sustainable development’ or ‘sustainability’ for short simply means that in a global context any economic or social development should improve and not harm the environment. This concept has developed from a global political process over the last three decades of the twentieth century into one that now touches every part of society. Sustainability has cut across all disciplines and professions, and has developed many complexities in relation to how people plan for their futures (Pezzoli 1997). Its application to planning is full of interest and controversy.

Town and regional planning has its origins firmly in a desire to improve the environment and create a better future. The visions of Ebenezer Howard, Patrick Geddes, and Lewis Mumford, for example, were all rooted in the need to make a greener city. Today there are many planners who use sustainability to rationalize why they are planning the way they do. This article will try to clarify the ways that sustainability is shaping how planning is done, as well as the differences in the ways that planners relate to sustainable development.

The issues relating to sustainability that appear to be uniting planners are:

- (a) How necessary it is to use holistic planning frameworks; and
- (b) How planning processes must start with community/civil society perspectives.

The issues which appear to divide planners are:

- (a) How planners view the size of cities; and
 - (b) How planners view density and car dependence.
- Some resolution will be attempted on these issues.

1. The Concept of Sustainability

Sustainable development was defined by the Brundtland Commission in 1987 after they were given the task of resolving the apparent global conflict between development and the environment (WCED 1987). They concluded that it is possible to develop the world in a way that improves the environment if:

- (a) High income nations see their responsibility to reduce their resource consumption with every new step in the development process;

- (b) Low income nations direct development to those who need it most as only then will population growth stabilize and misuse of natural resources begin to ease;

- (c) All nations need to see development as meeting community goals, thus development must be inherently a ‘bottom up’ process; and,

- (d) All professions and disciplines need to adjust to a more holistic, integrated approach if sustainability is to be converted into real change.

The thrust of this article is that, within the sustainable development framework of thinking, there are clear differences in what planners are suggesting sustainability means for cities. However, the overall situation cannot be seen as just a series of raging controversies, and indeed there are substantial areas of agreement.

2. The Agreements: Holistic Frameworks and New Planning Processes in the Age of Sustainability

Dyck (1998) suggests that sustainability is a concept which challenges prevailing planning theory because planning has emerged primarily from a modernist social context. In this perspective, modernism has associated growth or development with increased resource use, trickle-down economics, and ‘expert-oriented’ planning. Very few planners are, however, espousing a modernist view in the new millennium, and even the World Bank seems to agree with the basic conceptions of how sustainable development should now be viewed (Leitmann 1999). In particular, the areas of emerging agreement are the need for new, more holistic planning frameworks and community-sensitive planning processes.

2.1 Holistic Frameworks

The principles of sustainability outlined above can be applied to cities, though the guidance on how this can be done was not very clear in Agenda 21 and other UN documents. It is probably true to say that the major environmental battles of the past were fought outside cities, but that awareness of the need to come back to cities is now recognized universally by environmentalists, government, and industry (United Nations Center for Human Settlements 1996).

Anders (1991), in a global review of the sustainable cities movement, pointed out that ‘The sustainable cities movement seems united in its perception that the state of the environment demands action and that cities are an appropriate forum in which to act’ (p. 17).

One of the strongest themes running through the literature on urban sustainability is that, to solve problems, planners need to view the city as an *ecosystem*. As Tjallingii (1993, p. 7) puts it: ‘The city is

(now) conceived as a dynamic and complex ecosystem. This is not a metaphor, but a concept of a real city. The social, economic and cultural systems cannot escape the rules of abiotic and biotic nature. Guidelines for action will have to be geared to these rules.'

The city is an open ecosystem, having inputs of energy and materials. The main environmental problems (and economic costs) are related to sustaining the growth of these inputs, and managing the increased outputs. By looking at the city as a whole and by analyzing the pathways along which energy and materials (and pollution) move, it is possible to begin to conceive of management systems and technologies which allow for the reintegration of natural processes, increasing the efficiency of resource use, the recycling of wastes as valuable materials, and the conservation of (and even production of) energy.

There may be ongoing academic debate about what constitutes sustainability and how it applies to cities (Slocumbe 1993, Mitlin and Satterthwaite 1994), but what is clear is that many strategies and programs around the world have begun to apply such notions both for new development and for redevelopment of existing areas. Holistic frameworks such as the Extended Metabolism Model (Newman and Kenworthy 1999), the Ecological Footprint (Wackernagel and Rees 1996), and Local Environmental Action Planning (Leitmann 1999) are all emerging as new planning frameworks.

2.2 Planning Processes

One of the strengths of the sustainability concept is that it is not a set of regulations or even highly definitive principles. Its principles need to be examined by every local community to see how they can be applied to create better long-term futures for their children. Thus planners at the local level have often found something of the reality of sustainability through the use of Local Agenda 21 Plans which have emerged as a tool for planning since the Rio conference (ICLEI 1996, 1997, Selman 1996). Over 2,000 local governments have produced such plans. Others have adopted approaches based on sustainability indicators such as Sustainable Seattle (City of Seattle 1993) or the Local Environmental Action Plan (Leitmann 1999), and many are now using ICLEI's Cities for Climate Protection system as a guiding approach.

All these new processes stress that planning has moved beyond traditional approaches such as Impact Assessment or Strategic Planning based on trend projections and functional separation. They have in common a desire to plan from a community base, integrated with the goals of the market and other levels of government, including the new global agendas generated from Rio, Kyoto, and beyond.

The stories of hope emerging at a local level are rich and diverse when examined at the grass roots level

(e.g., Pathways to Sustainability Conference 1997, ICLEI 1997). The reason for this is probably that at the local level it is possible for government more easily to make the huge steps in integrating the environmental, economic, and social, in order to make policy developments that are sustainable. They are also closer to concerned people, and more distant from the single-issue powerful lobbies such as the fossil fuel and road lobbies, who are so obvious in shaping national priorities (Korten 1990, 1995).

2.3 The Disagreements: Application of Sustainability to City Size, Urban Form, and Transportation

The disagreements that are apparent among planners, who all profess to want more sustainable cities, are reducible to issues concerning urban form and transportation. They relate to the size of cities and to their density/car dependence.

3. City Size and Sustainability

Many planners subscribe to the view that, on environmental grounds (and other criteria), cities would be better if they were smaller (Troy 1996, Morris 1982). This is an old planning debate, for example, Fischer (1976) shows how optimum size cannot be found on any social or economic variable, but the debate has new currency in light of the current move to seek 'smart growth' and growth management. The desire to keep settlements small is seen by many to be a major part of what sustainability means.

The evidence from studies of settlements shows that as cities get larger they become more efficient in energy use per person and other critical sustainability variables (Newman and Kenworthy 1999, Naess 1993, Smith et al. 1998). This appears to be related to the economies of scale and density associated with bigger cities, which applies equally to technologies seen to be appropriate for sustainable development such as transit, recycling, or the new information industries. It does not mean, however, that larger cities do not reach capacity limits in terms of their airsheds and water systems. But it does mean that simple ideas of decanting people from big cities to smaller ones would not help to create a more sustainable global future.

The evidence suggests that smaller cities find the process of achieving efficiency in economic and resource terms harder than larger cities. The strong emotional appeal of 'smallness' to our age is not, however, without its basis. E. F. Schumacher's 'Small is Beautiful' approach, with its attack on modern gigantism, is still relevant as it suggests that all technology needs to be at an appropriate scale for the community that it is meant to be serving. Community-scale technology is emerging as communities begin to

assert their roles—whether the technology is in villages in the Third World or modern large cities (Korten 1990). The thrust of the New Urbanism and the ‘urban village’ movement is that communities need to be physically designed for and given infrastructure at the scale of the community, in cities both large and small (Calthorpe 1993).

3.1 Density and Car Dependence

In cities across the world the issue of urban sprawl is firmly on the agenda. Associated with it are the issues of density and levels of car dependence, and hence traffic. In the USA in 1998, 240 local governments had antisprawl initiatives on their ballots, and reductions in traffic had become a top priority across the country. However, debate among planners on these issues shows little sign of resolution.

The divide continues to widen between those advocating more compact cities and transit-oriented reurbanization (e.g., Cervero 1999, Whitelegg 1993, Roseland 1998, Newman and Kenworthy 1999, Calthorpe 1993, Elkin et al. 1991), and those advocating reduced densities and only technological changes to cars (e.g., Gordon and Richardson 1989, Troy 1996, Wachs and Crawford 1991, Johnson 1993). Another category of planners sit on the density/car dependence fence (e.g., Haughten and Hunter 1994), while others see the problems but tend to avoid the planning implications in low density, car-dependent planning (e.g., Leitmann 1999, World Bank 1996, Stren et al. 1992).

The data on density and car dependence is very clear. All cities appear to lie on the negative exponential curve of density and gasoline per capita (Newman and Kenworthy 1999). The same relationship is seen across cities, so that as density increases, per capita energy goes down, well beyond any income factors, for example, Manhattan uses 500 percent less energy per capita than the outer suburbs of New York. The mechanism is understood easily in terms of the relative ability to use different modes as distances between land uses spread out. It suggests that density is a powerful determinant of gasoline use, and hence if cities are to develop sustainably they should be facilitating reurbanization rather than new fringe development. Other planning studies have also shown over many decades the link between low urban density and a range of environmental problems (RERC 1974, Berry et al. 1974).

Despite this kind of evidence there is a long tradition of planners seeing density as a problem (Newman and Kenworthy 1989). Those planners advocating the benefits of sprawl who are associated with land or automobile lobbies can be discounted in debates about urban sustainability (STPP 1994). However, there are many environmental planners who also adopt the benefits of low-density planning (e.g., Gordon 1990,

Todd and Tukul 1981, Coates 1981). This suggests that there may be a deeper issue in this debate; an issue that also lies behind the question of city size and sustainability.

Radberg (1995), when reviewing the form of cities required to achieve sustainability, described the rationale for a more compact city which can conserve energy. He then set out the form of a city for achieving green objectives that are ‘related to the ecological perspective’ for ‘recycling and cultivation’; this city is ‘greener, more ruralized, more spread out.’ He therefore raised the questions: Is there a conflict between the low-energy city and the green city? Is there a fundamental conflict between the need to have more urban land for the green city in order to do local ecological processing, and a low-energy city with its need to minimize travel distances and hence have a frugal use of land?

The resolution of this issue as suggested by Radberg (1995) was to enable some parts of cities to be developed more intensely for reductions in energy, and other parts to be more reduced in their density so that ‘greener’ activities could occur there. The evidence, however, is that those planners who support the reduction of density are not able to accept that any density increases are worthwhile (e.g., Troy 1996). On the other hand, studies of urban ecology innovations by Scheurer (1999), confirmed by Beatley (2000), found that there are no increases in ecological activity in low-density areas. Indeed, the best examples of urban ecology (recycling, solar design, storm water retention, permaculture ...) occurred in denser developments, mostly in inner city areas where communities were able to create innovative experiments, or in intentional rural or urban-fringe communities. In the car-dependent suburbs urban ecology innovations are rare.

Newman and Kenworthy, (1999) tried to resolve the issue of density and the green city. They suggest that green innovations can indeed occur anywhere in a city, although experiments may best be applied in rural areas, especially in places where rural population decline is occurring. However, the broader global sustainability agenda cannot be subsumed by this new rationale for density-lowering; cities need to reurbanize while conducting simultaneous improvements in urban ecology. They suggest that the difference in attitude between the two groups is as much due to confusion over where different approaches to sustainability should be employed. In simple terms, the city should be allowed to become more urban, and the countryside more rural, if the multiple goals of planning and sustainability are to be met.

A resolution of this issue may not occur while people still use the word ‘sustainable’ to describe extensions of low-density, car-based development (even if designed with New Urbanist principles). The strong emphasis on reductions in energy that is now

associated with sustainability will probably mean that car-based development increasingly will have to be described using other words that are less demanding of their global contribution.

See also: Environmental Economics; Environmental Vulnerability; Environmentalism: Preservation and Conservation; Sustainability Transition: Human–Environment Relationship; Sustainable Development; Sustainable Transportation; Transportation Geography; Transportation Planning; Transportation: Supply and Congestion; Urban Geography

Bibliography

- Anders R 1991 The sustainable cities movement. Working Paper No. 2, Institute for Resource and Security Studies, Cambridge, MA
- Beatley T 2000 *Planning for Sustainable Cities: The European Experience*. Island Press, Washington, DC
- Berry B et al. 1974 Land use, urban form and environmental quality. Research Paper No. 155, University of Chicago Press, Chicago
- Calthorpe P 1993 *The Next American Metropolis: Ecology and Urban Form*. Princeton Architectural Press, Princeton, NJ
- Cervero R 1999 *Transit Planning*. Island Press, Washington, DC
- City of Seattle 1993 Sustainable Seattle: Indicators of Sustainable Community, City of Seattle, WA
- Coates G 1981 *Resettling America: Energy Ecology and Community*. Brick House, Andover, UK
- Dyck R G 1998 Integrating planning and sustainability theory for local benefit. *Local Environment* 3(1): 27–41
- Elkin T, McLaren D, Hillman M 1991 *Reviving the City: Towards Sustainable Urban Development*. Friends of the Earth, London
- Fischer C S 1976 *The Urban Experience*. Harcourt Brace, New York
- Gordon D (ed.) 1990 *Green Cities*. Black Rose, Montreal, PQ
- Gordon P, Richardson H W 1989 Gasoline consumption and cities: A reply. *Journal of American Planning Association* 55: 342–6
- Houghton G, Hunter C 1994 *Sustainable Cities*. Taylor and Francis, Bristol, PA
- ICLEI (International Council on Local Environmental Initiatives) 1996 *The Local Agenda 21 Planning Guide: An Introduction to Sustainable Development Planning*. ICLEI, Toronto, ON
- ICLEI (International Council on Local Environmental Initiatives) 1997 *Local Agenda 21 Survey*. ICLEI, Toronto, ON
- Johnson E 1993 *Avoiding the Collision of Cities and Cars*. American Academy of Arts and Sciences with the Aspen Institute, Chicago
- Korten D C 1990 *Getting to the 21st Century: Voluntary Action and the Global Agenda*. Kumarian, Bloomfield, CT
- Korten D C 1995 *When Corporations Rule the World*. Kumarian, Bloomfield, CT
- Leitmann J 1999 *Sustaining Cities: Environmental Planning and Management in Urban Design*. McGraw-Hill, New York
- Mitlin D, Satterthwaite D 1994 *Cities and Sustainable Development*. International Institute for Environment and Development, London
- Morris D 1982 *Self Reliant Cities: Energy and the Transformation of Urban America*. Sierra Club Books, San Francisco

- Naess P 1993 Energy use for transport in 22 Nordic towns, NIBR Report No. 2, Norwegian Institute for Urban and Regional Research, Oslo, Norway
- Newman P W G, Kenworthy J R 1989 *Cities and Automobile Dependence: An International Sourcebook*. Gower, Aldershot, UK
- Newman P W G, Kenworthy 1999 *Sustainability and Cities: Overcoming Automobile Dependence*. Island Press, Washington, DC
- Pathways to Sustainability Conference 1997 Proceedings of UN Conference, City of Newcastle, New South Wales, Australia
- Pezzoli K 1997 Sustainable development: A transdisciplinary overview of the literature. *Journal of Environmental Planning and Management* 40(5): 549–74
- Radberg J 1995 Termites heap or rural villages? The problems of urban density and sustainability. In: *The European City—Sustaining Urban Quality*. Proceedings of Conference, Copenhagen, April. Ministry of Environment and Energy, Copenhagen, Denmark
- Real Estate Research Corporation 1974 *The Costs of Sprawl*, Council of Environmental Quality, Washington, DC
- Roseland M. 1998 *Towards Sustainable Communities*. New Society Press, Vancouver, BC
- Scheurer J 1999 *An Evaluation of Danish Ecological Housing and Planning*. Danish Building Research Institute, Horsholm, Denmark/ISTP, Murdoch University, Perth, WA
- Selman P 1996 *Local Sustainability: Managing and Planning Ecologically Sound Places*. Paul Chapman, London
- Slocum D S 1993 Environmental planning, ecosystem science and ecosystem approaches for integrating environment and development. *Environmental Management* 17(3): 289–303
- Smith M A F, Whitelegg J, Williams N 1998 *Greening the Built Environment*. Earthscan Publications, London
- Stren R, White R, Whitney J (eds.) 1992 *Sustainable Cities*. Westview Press, Boulder, CO
- STPP 1994 *A Citizens Guide to Transportation Planning and Livable Communities*. Surface Transportation Policy Project, New York
- Tjallingii S 1991 The responsible city. Paper presented at the International Federation for Housing and Planning, International Congress, Berlin
- Todd J, Tukul G 1981 *Rehabilitating Cities and Towns: Designing for Sustainability*. Planet Drum, San Francisco
- Troy P 1996 *The Perils of Urban Consolidation*. The Federation Press, Sydney, NSW
- United Nations Center for Human Settlements 1996 *UN Global Review of Human Settlements*. Habitat, Nairobi, Kenya
- Wachs M, Crawford M 1991 *The Car and the City: The Automobile, The Built Environment and Daily Urban Life*. Michigan University Press, Ann Arbor, MI
- Wackernagel M, Rees W 1996 *Our Ecological Footprint: Reducing Human Impact on the Earth*. New Society Publishers, Philadelphia, PA
- Walter R, Arkin L, Crenshaw R (eds.) 1992 *Sustainable Cities: Concepts and Strategies for Eco-City Development*. Eco-Home Media, Los Angeles
- World Bank 1996 *Sustainable Transport—Priorities for Policy Reform, Development in Practice*. World Bank, Washington, DC
- World Commission on Environment and Development 1987 *Our Common Future*. Oxford University Press, Oxford
- Whitelegg J 1993 *Transport for a Sustainable Future: The Case for Europe*. Belhaven Press, London

P. Newman

Planning, Politics of

Planning—here defined as current efforts to affect future outcomes—is a ubiquitous human activity in both public and private life. Individuals plan when they choose their career paths, build families, and save for their retirements. More formally, both corporate managers and government officials regularly develop detailed plans to guide the future activity of their organizations. No society or time-period has a monopoly on planning. Indeed, examples of planning can be identified in virtually every culture and epoch. Some of the best-known episodes in the ancient literature (e.g., the biblical account of Joseph's grain storage) and the most celebrated public works projects in history (e.g., the building of the pyramids, the Great Wall of China and the eighteenth century reconstruction of Paris) were, fundamentally, exercises in planning.

A major aim of planning is to cope with uncertainty. By delineating precise goals, planners seek to stabilize and shape subsequent conditions. Although (nearly) everyone plans, some do so as a career. In the USA today, planning is a recognized profession with barriers to entry, distinct graduate training programs, and particular methods, modes of discourse and internal status rewards. Many planners work for government agencies or as government contractors.

1. Planning Defined

The term 'planning' has no fixed technical meaning in the political science or public administration literatures (Meyerson and Banfield 1955). It refers to activities ranging from socialism to program budgeting to purposive efforts to manage land use, municipal zoning, and regional development. The denials of certain planning theorists notwithstanding, planning is inescapably political because it reflects and reinforces social values (Benveniste 1989). Yet planning and politics may also be in tension with one another (Banfield 1961). At its most heroic, planning seeks to pre-commit major aspects of a society in a particular long-term direction designated by central elites. Yet knowledge is always partial and social conditions inevitably change. Heavy reliance on central planning thus risks reducing opportunities for social learning and partisan mutual adjustment, processes that are essential to a well-functioning polity. To the extent that planning accommodates such vital error-correction and conflict-resolution processes, however, it becomes hard to distinguish from ordinary forms of policymaking (Wildavsky 1979).

At least three (overlapping) dimensions of government planning may be distinguished. First is whether the objective of the plan is to shape the future operations of the overall economy (economic planning) or, alternatively, primarily to affect the

government's own internal administrative operations (administrative planning). A second important dimension concerns the scope of the plan—whether the intellectual demands, and aspirations, of the planning effort are large (synoptic or comprehensive planning) or relatively modest (strategic planning). Finally, one can distinguish between plans that rely on coercion to effect change and those that seek to achieve their goals through voluntary compliance (indicative planning).

1.1 Economic Planning

Throughout the twentieth century, economic planning has been an important response to the perceived moral failings and practical defects of capitalism in both Europe and North America. Prior to the 1980s, the governments of the former USSR and its Eastern European satellites attempted economic planning on a grand scale, seeking directly to control all means of production (as well as consumption patterns) through centralized decision-making (Berliner 1957). For many, the tragic experience of central economic planning stands as a political symbol of the inherent limitations of socialism. More concretely, this experience is a legacy that the peoples of Russia and Eastern Europe are struggling to overcome.

Under Stalin and his successors, central Soviet economic planning was organized hierarchically. The top state planning agency, Gosplan, established the target rate of national economic growth and the allocation of production across industrial sectors and geographic regions. This overall national plan was then broken down by various ministries into directives assigned to local economic units. Finally, managers and engineers at the factory or farm level were required to implement particular instructions. Until the 1960s, the USSR reported strong economic growth. By the 1970s, however, it was apparent that productivity had slowed drastically and that resources were being grossly misallocated. While military facilities could generally obtain whatever they needed, the domestic sector was being starved. Living standards deteriorated. In 1987, President Mikhail Gorbachev announced a policy of *Perestroika*—restructuring—designed to resuscitate the moribund economy (Heilbroner 1993). But the system collapsed.

Critics argue that the Soviet planned economy failed for two main reasons. The first is inadequate information. Very early on the free market economists Ludwig von Mises and Friedrich Hayek insisted that a socialist system was doomed to fail because central economic planners, lacking the price signals of a market system, would be unable to obtain the information required to promote economic efficiency (Von Mises 1935, Hayek 1948). In fact, economic planners in the Communist world—as Polish economist Oskar Lange anticipated in an important 1938 analysis of socialism—were often able to gain much of

the information they needed by monitoring changes in plant inventory levels (Lange and Taylor 1938). The more important reason for the failure of socialist central planning, it seems, was that actors typically had little or no economic incentive to respond to their directives (Heilbroner 1993).

Although Soviet-style comprehensive economic planning was never accepted, the idea of economic planning has long enjoyed intellectual support in the West, even in the USA. The terrible human costs of the Great Depression weakened elite and popular faith in *laissez faire* capitalism and ultimately led to new forms of government intervention in the economy. Most Western governments assumed responsibility for planning for full employment. In France, 'indicative' planners sought to achieve a voluntary exchange of information among business and labor leaders, thereby allowing the French government to use its fiscal and regulatory tools to shape private investment decisions (Lindblom 1977). In the USA, a President's Council of Economic Advisers was established in 1946 to promote stability and growth. During the 1950s and 1960s, Keynesian fiscal planners actively sought to manipulate the level of consumer demand and macroeconomic output. The large tax cut of 1964 was explicitly justified in Keynesian terms. The era of activist Keynesian economic planning came to an end in the mid 1970s, however, with the simultaneous appearance of high inflation and slow growth (stagflation), which contradicted the assumptions of Keynesian models. While use of automatic stabilizers remained a fixture of fiscal policymaking, US politicians in particular lost confidence in the ability of economic planners to 'fine-tune' macroeconomic policy to the business cycle.

1.2 Administrative Planning

In the West, it is in the realm of public administration that one can identify genuine attempts at synoptic planning. Comprehensive government planning has long been central to normative theories of public administration. The influential French administrator Henri Fayol (1841–1925) argued that a 'good plan' requires clearly defined, written, comprehensive objectives together with the resources necessary to carry out these tasks. In the USA, broadly similar concepts found expression in the scientific and executive management movements of the Progressive and New Deal eras. The normative commitments and pragmatic understandings that emerged from these movements (e.g., 'one-best way' thinking; the emphasis on hierarchy and span of control; the separation of politics from administration) have had a significant impact on USA administrative practice and doctrine. In a classic statement in 1937, Luther Gulick, a member of the New York Bureau of Municipal Research who served on Franklin Roosevelt's famous Brownlow Committee, defined the executive function by the acronym

POSDCORB (planning, organizing, staffing, directing, coordination, reporting, and budgeting). Effective planning was thus held to be the crucial first step to the achievement of rationality and technical efficiency in government administration. In the USA, administrative planning enjoyed a revival in the 1960s (Schick 1975). Hundreds of federal and local statutes mandated the formulation of plans for allocating inter-governmental grants, designing cities, and other purposes. The centerpiece of the new administrative planning was PPBS—Planning, Programming, Budgeting Systems—which essentially demanded that federal agencies engage in synoptic planning to promote the public interest. Administrators were to (a) articulate precise program goals, (b) comprehensively list alternative ways of achieving these goals, and (c) analyze the costs and benefits of each significant option. PPBS was introduced in the Pentagon in 1961 by Defense Secretary Robert McNamara. By 1965, President Lyndon Johnson ordered PPBS to be extended to nearly all federal, civilian departments. The promise of PPBS was considerable: Planning and policy analysis would be used to increase administrative effectiveness. Defective programs would be found wanting and, it was hoped, replaced with superior alternatives. The actual results of PPBS were disappointing to say the least, however. The reporting and monitoring requirements of the planning effort created a paperwork burden that quickly overwhelmed government officials. PPBS performed especially badly where, as was the case in many domestic agencies, bureaucratic outputs were soft and hard-to-measure. Most importantly, PPBS failed to accommodate the political and organizational incentives of budget actors. The Nixon Administration formally abolished PPBS in 1971 (Knott and Miller 1987, Kettl 1992). The failure of PPBS lends support to Aaron Wildavsky's argument that a fragmented and seemingly uncoordinated budget process generally does a better job of allocating public resources than does a system in which central planners make a synoptic, integrated evaluation of budget alternatives (Wildavsky 1964). Despite the manifest failures of PPBS, similar, comprehensive-planning exercises were adopted subsequently by US governments (e.g., zero-base budgeting and management-by-objectives).

Given the severe limits of synoptic planning, why have so many governments persisted with these sweeping reform efforts? Probable reasons include: there may be a political benefit to appearing 'rational'; planning exercises may enable elected officials to give respectability to things they wish to do for other reasons; and focusing on planning procedures is often easier than making substantive decisions (Knott and Miller 1987). The enduring interest in synoptic government planning may actually rest less on science and rationality than on a secular faith in the perfectibility of man (Wildavsky 1973). Indeed, in an important account of paradigmatic failed attempts at 'state

simplification' (e.g., the Great Leap Forward in China, collectivization in Russia, compulsory villagization in Tanzania, Mozambique, and Ethiopia), political anthropologist James Scott argues that the hegemonic planning mentality is supremely unscientific in the sense of being uncritical, unskeptical, and contemptuous of the importance of practical, local knowledge and informal processes in the face of unpredictability (Scott 1998).

2. Limits and Possibilities of Planning

Although synoptism has failed everywhere it has been tried, more limited forms of government planning clearly have succeeded. The Controlled Materials Plan of the USA in World War Two, the American Social Security program, the massive interstate highway system—each of these initiatives was bureaucratically 'planned.' Long-term goals were articulated, the stated objectives were (more or less) achieved, and the political outcomes were, if hardly uncontroversial, at least defensible. Planning works best when it is highly bounded, is focused on a specific problem, and is economical of analytical talent (Lindblom 1965, 1977, see also Devons 1950). Thus, like the formal hierarchies through which it tends to be implemented, planning has its organizational place (Alexander 1992). The challenge is to keep it there—and to render the practice and results of planning compatible with processes of social interaction.

See also: Patriotism

Bibliography

- Alexander E R 1992 A transaction cost theory of planning. *Journal of the American Planning Association* 58: 190–200
- Banfield E 1961 *Political Influence*. Free Press, Glencoe, IL
- Benveniste G 1989 *Mastering the Politics of Planning*. Jossey-Bass, San Francisco
- Berliner J 1957 *Factory and Manager in the USSR*. Harvard University Press, Cambridge, MA
- Devons E 1950 *Planning in Practice*. Cambridge University Press, Cambridge, UK
- Gulick L, Urwick L 1937 *Papers on the Science of Administration*. Columbia University Press, New York
- Hayek F A 1948 Socialist economic calculation: The present state of the debate. In: Hayek F A (ed.) *Individualism and Economic Order*. University of Chicago Press, Chicago
- Heilbroner R 1993 Socialism. In: Henderson D R (ed.) *The Fortune Encyclopedia of Economics*. Warner, New York
- Kettl D F 1992 *Deficit Politics: Public Budgeting in Its Institutional and Historical Context*. Macmillan Publishing, New York
- Knott J H, Miller G J 1987 *Reforming Bureaucracy: The Politics of Institutional Choice*. Prentice-Hall, Englewood Cliffs, NJ
- Lange O, Taylor F 1938 *On the Economic Theory of Socialism*. University of Minnesota Press, Minneapolis, MN
- Lindblom C E 1965 *Intelligence of Democracy*. Free Press, New York
- Lindblom C E 1977 *Politics and Markets*. Basic Books, New York

- Meyerson M, Banfield E C 1955 *Politics, Planning and the Public Interest*. The Free Press, Glencoe, IL
- Schick A 1975 The trauma of politics: Public administration in the sixties. In: Mosher F (ed.) *American Public Administration: Past, Present, Future*. University of Alabama Press, AL
- Scott J 1998 *Seeing Like A State*. Yale University Press, New Haven, CT
- Von Mises L 1935 Economic calculation in the Socialist Commonwealth. In: Hayek F A (ed.) *Collectivist Economic Planning*. G. Routledge and Sons, London
- Wildavsky A B 1964 *The Politics of the Budgetary Process*. Little, Brown, Boston
- Wildavsky A B 1973 If planning is everything, maybe it's nothing. *Policy Sciences* 4: 127–53
- Wildavsky A B 1979 Between planning and politics: Intellect vs. interaction as analysis. In: Wildavsky A B (ed.) *Speaking Truth to Power*. Little, Brown, Boston, Chap. 5, pp. 114–40

E. M. Patashnik

Planning Theory: Interaction with Institutional Contexts

During the twentieth century, the concept and practice of what is variously called town design, spatial planning, territorial management, and town and country planning developed into a complex area of public policy. The theory informing this policy field has been the product of continual interaction between ideas about how to do planning, the experience of planning in practice, the intellectual context available to those thinking about planning, and the agenda of problems and opportunities which planning has been harnessed to address. Inherently, the planning enterprise is caught up in institutional contexts, which shape its meaning, its material expression, and its consequences. This contribution illustrates the interaction between ideas about the nature, purpose, and method of planning, the context of their production and the practices at which they have been directed.

1. Cities, Governance, and Planning

Few of the ideas about planning which evolved in the twentieth century came initially out of academia. There was no significant academic infrastructure for the field until the last quarter of the twentieth century. The origins of twentieth century planning ideas lay in the experience of rapidly urbanizing societies, and the perceptions of injustice, disorganization and threats to human health and happiness which urbanization seemed to bring. These promoted political movements, pressure group politics, expert evangelism, and attempts to transform city government to cope with the complexity of issues which arise in large cities. The focus of attention ranged from the neighborhood, the city, the region, the process of governance and the

overall organization of society. Two core ideas have informed these otherwise very diffuse endeavors: firstly, that through appropriate governance action, it is possible to 'make a difference,' especially in making cities and societies more sustainable and just; secondly, that the spatiality of social processes and the qualities of place are important dimensions of well-being.

These two ideas have been developed in very different ways as illustrated below, with the aim of highlighting both the situatedness of planning ideas and how their impacts are highly dependent on the specifics of institutional milieu.

1.1 The Good City

In the early twentieth century, planning was generally understood as focused on cities, on how to manage the congested and sprawling consequences of the urbanization arising from industrialisation and new transport technologies (Mumford 1961). Throughout the century, there was concern about the qualities of cities as living places, as aesthetic products, as business locales, and as biospheric relations. 'Theories' of urban form and the arrangement of regional space dominated the discourses of city planning until at least the 1960s (Boyer 1983, Hall 1988). Protagonists, a mixture of campaigners and specialists in urban development (architects, engineers, and surveyors), battled over the relative merits of different forms. By the 1950s, most European states had some form of urban planning and land use regulation embedded in their central and local government policy frameworks and competencies.

This success in institutionalizing the idea of city planning was soon overtaken by vigorous critique of the predominant focus on urban form. Sociologists such as Herbert Gans in the USA argued that cities should be understood through the way people lived in them rather than through their physical form. Jane Jacobs, a fine observer of the way people used urban neighborhoods, argued against planners' ideas of demolishing old neighborhoods to replace them with modern buildings and the separation of living and working spaces. Regional economists and geographers challenged assumptions about the dynamics of urban growth embodied in the urban plans produced by architects, engineers, and surveyors. The emphasis of the earlier generation of town planners on achieving urban quality of life for the mass of the people was replaced by a concern with distributive justice and the 'socially just city'. This idea emphasized that the poor should not draw the short straw in the distribution of access to urban living space. Peter Hall showed that the containment of British cities by greenbelts, far from achieving gains in quality of life for ordinary citizens, had made housing generally more expensive and led to deteriorating conditions in the UK's inner cities. By the 1970s, Marxist analysts had joined the

debate. Manuel Castells showed, in a study of a French city, that the aim of urban planning to improve the quality of life was a mask behind which developers and business interests could pursue strategies for realizing land value gains. This critical story had resonance in other parts of Europe where urbanization was proceeding apace, producing substantial development gains to landowners and developers. It was echoed in examples of urban renewal in the USA.

As a consequence of this critique, at least in Europe, coupled with a slowdown in the pace of urban growth, the idea of large-scale comprehensive planning fell into disrepute. Urban management practices focused on the fine grain, the neighborhood, the site, and the urban renewal area. Attention in the 1970s and 1980s turned to projects and mechanisms. The focus on the physical form of the city was displaced by ideas about appropriate planning processes (see below). Architects, engineers, and surveyors were replaced as a source of expertise for planning by regional economists, policy analysts, and those specially trained as 'planners'. The discussion of 'planning theory' turned away from ideas about what the city as such could be like. Influenced by postmodern ideas in urban design and postmodernism in philosophy, a critique developed which suggested that the whole idea of attempting to 'plan' such a complex phenomenon as a 'city' was oppressive to the multiple realities and possibilities of urban life.

By the end of the twentieth century, a new wave of interest in the quality of cities, rural areas, and urban regions was swelling up in policy arenas and among commentators on planning. The pressure for this came in part from an increasing awareness of new economic and social realities. The recognition of the openness of urban economies to the dynamics of global economic forces raised widespread concern about how to 'position' cities and urban regions in a landscape of competing regions. The appreciation from direct experience of lifestyle diversity, and its consequences for the ways different groups were using the space of an urban region, made people aware that traditional models of self-contained, hierarchically-ordered cities no longer captured the reality of urban life, especially as affected by new technologies (Sandercock 1998, Beauregard and Body-Gendrot 1999). Most influential of all were the ideas associated with the rise of a new environmental consciousness in the last quarter of the twentieth century. This brought with it conceptions of the limits which the natural environment placed on human actions, and of the need to consider trans-generational impacts not just short term ones.

This led to a wave of interest in the 'Sustainable City' (Satterthwaite 1999). For some, this meant minimizing adverse impacts on natural environments, linked to ideas about 'ecological footprints' of urban development, and conceptions of the 'carrying capacity' of environmental systems. For others, it emphasized keeping cities compact and preventing land

being used up in development. Others enlarged the concept to integrate the concern with social justice to that of 'environmental sustainability.' The dominant concept in public policy in much of western Europe by the turn of the twentieth century was 'sustainable development,' which combined the principles of minimizing ecological impacts with ways of continuing to promote economic development. However, it is clear that these two dynamics do not comfortably co-exist with each other. There are also tensions with notions of social justice, the multi-cultural city, and fostering innovative plurality. The reawakened emphasis on urban futures in planning theory thus opens up a complex set of issues that had hardly begun to be developed by the end of the millenium.

These ideas and debates about the quality of the 'good city' and the 'good environment' were shaped by the changing dynamics of urbanization processes, as well as changing expectations and values. In turn, concepts from planning theory have influenced the practices of urban management. Planners' ideas have molded the restructuring of existing urban areas, the design of new towns and settlements, and residential neighborhood design. They have affected the form and location of transport projects. They have shaped the evolution of urban street scenes through the impact of land use regulations. These regulations have in turn influenced the way land and property values are distributed and property development processes have evolved. Planners' ideas achieved this impact by framing the organizing principles and discourses used by governments in developing policy frameworks and organizing competencies for planning 'systems'. However, the organizing principles and discourses of one era become embedded in practices and can come to outlive their usefulness. By the turn of the millenium, in many parts of the world, 'planning' as practiced was identified as 'part of the problem' of governance processes that do not match the aspirations and challenges of contemporary cities and regions.

1.2 Good Government

For the protagonists of early twentieth century city planning, the everyday business of politics and government was 'part of the problem.' They sought to show a way forward by striding out ahead with their plans and designs, apart from this messy world. By mid-century, an alternative idea had taken root. Planning was a way of 'doing government' itself. In Europe, socialist movements promoted the role of planning as an alternative to allocating resources through market processes. In the USA, planning was promoted as a way of improving the quality of local government. Clientelist practices were well developed in many cities, with adverse consequences for both the efficiency and effectiveness of public policy. Bureaucratic management was no better, as 'playing by the rules' led to

government actions which did not adapt to the dynamics of changing situations or local specificities. The way forward, strongly advocated in the University of Chicago School of Planning in the 1950s, was to adopt contemporary principles of business management. This meant developing strategic frameworks to articulate the goals of an organization. These could then be translated into policy principles or objectives which individual officials could use as guides as they developed responses appropriate to specific situations. This powerful idea, with its focus on 'means-ends relationships,' was strongly linked to the promotion of democratic government. All the steps justifying a particular policy could be explicitly identified and implemented by logical analysis and scientific evidence. Officials could be held to account in relation to their performance in achieving an organization's goals, while the process of strategy-formation made these goals and their justification transparent to citizens (Faludi 1973, Friedmann 1987).

Initially, it was assumed that the work of 'policy planning' would take place in a world of technical experts, reporting to politicians. As Western democracies have become increasingly educated and pluralistic, distributing power and knowledge ever more widely among business groups, pressure groups, and citizen mobilization movements, this position has been challenged. However, concepts of strategy, of evaluation of policy ideas and performance in relation to objectives, and of the systematic link between policy and implementation had become deeply embedded in expectations of government behavior by the end of the twentieth century. City planning practices have been transformed by these shifts in the way government is performed. Planning theory has both influenced the design of spatial planning processes in many parts of the world and learned from their practice.

At the core of the original idea of a 'rational planning process,' as it was often called in the 1960s and 1970s, was the view that governments—and their teams of experts—needed to be able to explain why a particular strategy was appropriate in a particular context, how it would be implemented, and whether it had been implemented. It became linked to the development of ways of making strategic and policy plans, techniques of evaluating alternative options, appraising the impacts of projects, the development of indicators to evaluate performance, and attempts at modeling situations—for example land use and transport interactions—to assess the likely impacts of different courses of action. All these are still in common use in twenty-first century urban and regional government, providing a methodological kitbag for planning office staff and consultants.

However, these early ideas attracted criticism almost as soon as they were propounded. The overall approach was challenged as deeply instrumentalist—defining the purpose of government in terms of achieving specific objectives effectively. The meth-

odologies drew on welfare economics, with its utilitarian assumptions. The approach assumed that policy ideas were translated into action in a linear way, from 'top' to 'bottom' of an organization or government system. It was seen to be over-generalized, failing to consider the differences, which might arise in different policy fields and in different institutional contexts. It also handed power to the experts, as the early city planning ideas had done. Politicians, lobbyists, and citizens were pushed to the sidelines. In the 1950s, a famous study by Meyerson and Banfield in Chicago showed that local politicians were not prepared to accept the restraints on their actions, which the rational policy-planning model implied. In the 1960s and 1970s, scholars such as Charles Lindblom, Amitai Etzioni, John Dyckman, and Richard Bolan began to show how contexts might limit how far a policy-driven planning practice might go. Others criticized the approach as promoting 'corporatist' practices, entrenching the influence of local business and political elites through their power to shape and overturn strategies. An alternative was to accept the increasingly pluralistic form of local politics, and develop a local politics around competing plans, produced by different interest groups. This approach, promoted by Paul Davidoff, became known as advocacy planning (Faludi 1973). The experience of policy practice began to look less and less like a strategy-driven process. Instead, analysts noted that the content of policy seemed to be negotiated in the flow of action, rather than being set in advance.

By the late 1970s, planning theorists turned their attention to issues of policy 'failure' and implementation 'gaps,' following landmark studies by Aaron Wildavsky and others. This work looked more closely at the institutional context in which policy frameworks were articulated and the ways planners, understood as experts and government officials, went about their tasks. It identified the problems, which arose where policy ideas were not sufficiently 'owned' or understood by politicians, or by other parts of public administration which controlled significant resources or regulatory powers. There were close parallels with work undertaken in business studies. This analysis highlighted the difficulty of promoting strategy-driven governance processes where other ways of 'doing governance' controlled the culture of local government. It also showed that 'field level' actors, officials, and experts working up the details of policies, or acting at the interface between officials and the clients of public administration, developed their own ways of understanding policies and how they might be applied.

This led to a critique of the rational planning process as a 'top down' viewpoint, challenged by a 'bottom up' viewpoint of the actors 'on the ground.' (Barrett and Fudge 1981) The linear model of the policy process was increasingly challenged by conceptions of policy development as an interactive process, continuously evolving over time and drawing in different actors at

different stages. In this formulation, planning was no longer either outside government, nor imposed on governments in an attempt to transform their processes. It was caught up 'inside' politics and governance processes. In the 1980s, a significant strand of work associated with John Forester, Charles Hoch, and Howell Baum, began to address questions of the way planners committed to progressive values operated in these ethically challenging situations (Forester 1989, Campbell and Fainstein 1996).

During the 1980s, a major conceptual reorientation brought new ideas about the nature of policy processes to a central position in debates in planning theory. These built on the earlier work on 'implementation processes' in two ways. Some analysts focused on the way experts and officials made sense of the complex situations in which they found themselves. Others began to explore the way policy agendas and the relations of governance mutually constituted each other in interactive processes. Both emphasized the significance of the frames of reference and the 'social worlds' which those involved in policy arenas brought to bear on policy design and implementation. Policy development was presented less as a technical exercise in calculation and design, and more as a social process of articulating ways of thinking and ways of acting which could command support among a wide range of players in urban and regional contexts where power was likely to be diffused in complex ways rather than concentrated in the hands of a single elite. Collaborative, participative forms thus challenged top-down and 'producer-driven' approaches to public policy. Such ideas were infused with the much wider critiques of positivist science and linear models of knowledge and application developing in the 1980s in the social sciences generally. Some argued that planning was an impossible task, caught up in the discarded clothes of 'modernity.' Others, and notably John Bryson, Judith Innes, and Maarten Hajer drew on emerging interpretive approaches. These emphasized that policy meanings were actively constructed by participants through communicative processes which, if effective, could recast policy 'discourses' and redirect resource allocation processes and regulatory practices (Healy 1997, Innes 1995).

In parallel, there was an explosion of attempts in various parts of the world to develop new, more participative and collaborative forms in planning practice. These were partly promoted by the idealism of those pursuing a democratic version of the sustainability agenda, notably in the practices of producing Environmental Audits and Appraisals for localities, as encouraged by the World Environment Congress in Rio in 1992. But they were also a manifestation of a struggle to recast the forms of governance inherited from mid-century in Europe and the US. While neo-liberal strategies sought a way forward by reducing the burden of government, through strategies of 'privatization' and 'deregulation,' alternative strategies pro-

moted models of more participative democracy, which could reduce the perceived gap between the worlds of politics and government and the worlds of citizens. The objective was to refocus governance activity around the concerns of civil society and the economy, rather than those of the 'producers' of government.

By the late 1990s, ideas about consensus-building practices and collaborative strategy formation, collected and consolidated by planning theorists, were being rapidly internationally circulated to help fulfill in practical ways the ambitions of politicians and officials trying to move in new directions. In this context, the value of policy effectiveness embodied in the 'rational planning process' was supplemented by a concern with 'building democratic culture' (respectful and 'citizen-centered' styles of governance). Proponents of the new communicative planning argued that strategies that were developed interactively among affected parties (stakeholders) led to greater 'ownership' which in turn provided the social and political resources through which strategic directions could be maintained. They also argued that such processes brought civil society into a much stronger position to shape public policy agendas.

1.3 Qualities of Place and Qualities of Governance

By 2000, there were vigorous debates in the planning field about this new agenda. As in the earlier discussion of the 'rational planning model,' a key issue was the qualities of the institutional context in which such practices could flourish and how far their promotion could induce changes in the context. However, there was much less discussion of how these 'good governance' ideas might link to the qualities of a 'good city.' The two strands of thinking in planning ideas largely proceeded separately in the late twentieth century. They also potentially challenge each other, as the one stresses qualities of process and the other stresses qualities of places and their diverse relations. A governance process informed by the citizens 'of' a locality, which proceeds in a richly participative way, is likely to encounter multiple ideals about a 'good city' and a 'good living place.' What may emerge through processes of debate, struggle, and conflict resolution may be far removed from the ideals of a good city being developed in debates among professionals.

2. Planning Theories in Context

The enterprise of planning is profoundly committed to changing the ways things are and to transforming realities, whether those of cities or their governance. It is inherently normative in its value orientation, infused with the belief that deliberately designed actions today

can have beneficial impacts tomorrow. Those who reject the idea that this is desirable or possible—*anarchists, extreme free marketers and cultural post-modernists*—will reject the enterprise of planning. Yet the idea that, in democratic states, governments should have policies and should be judged on how effectively they follow them through, is one of the great inventions of the twentieth century. It is being carried forward vigorously into the twenty-first century. What are changing, however, are the conceptions of where policies should come from, how they should be focused and who gets involved in their articulation and implementation. At the beginning of the twenty-first century it is not enough for policies to be designed by experts. The quality of government is increasingly judged by the way its policy formation attends to the multiple voices of complex societies, by its skill in recognizing the complexity of economic, social, and environmental dynamics, and its appreciation of the specific realities of the way policy programs are realized.

In this context, a policy focus on 'place quality' has a new salience. Such a focus, if grounded in the specific contingencies of particular places and the multiple perceptions and meanings of those with a stake in them, can connect the issues which concern citizens and businesses in ways which make sense to them. These tendencies recreate a policy environment in which the enterprise of planning has much to contribute, provided it can release itself from its 'entrapment' in the past and reconnect the debates about the qualities of place to those on good governance. Three recurring themes in the discussion of the 'good city' and 'good governance' represent important challenges for future spatial planning and territorial management practices.

2.1 Ideas and Actions

The first challenge focuses on the relation of theory to practice. The history of planning thought shows the many ways in which planning ideas filter into the construction of concepts, goals, and norms which guide specific actions—through advocacy and campaigning which impact on popular perception and politicians' agendas; through the design of planning systems and formal norms; through the ways these are interpreted in administrative and legal arenas; through the formulation of explicit strategic frameworks and the ways these get translated into routinely used concepts and norms; and through the training of 'planners' and the professional organizations which bring 'planners' together. Through continual interaction with these worlds, planning ideas are themselves changed. The contribution of planning thought in the twentieth century has been its preparedness, however ineffectively it is often developed, to link together the

philosophy of normative action, the social science of urban and regional dynamics and the fine grain of the practices of urban and regional governance. The challenge has been to draw out the significant strategic processes of governance which affect the qualities of places and territories, as lived, experienced, and given identity in multiple social processes (Friedmann 1987).

2.2 Substance and Process

The geographer Ed Soja has remarked that twentieth-century thought has had a peculiar difficulty in perceiving the impact of 'being in a place' and of the spatiality of social relations. The autonomous individual in a 'free society,' or the social group as a 'class' in relation to other social classes, has been much more dominant. Contemporary culture, however, notices all kinds of cleavages and fractures, and stresses the complex relations individuals may have with others. Grasping the significance of place and territory in these complex representations and relations, and developing governance processes in which to give them voice, is a major challenge for urban and regional politics and policy. In planning theory, there has long been a debate about how to relate the 'substance' of policy to the 'process' of policy development, with theories being criticized for giving too much weight to one or the other.

If the predictions made in the previous section are correct, then the second challenge for planning theory is to contribute to the co-evolution of the 'substance' and 'process' of territorial management-policy frameworks. This has major implications for institutional design—who contributes what, when, and where. It also has implications for the articulation of values. There were major debates in the mid-twentieth century about whether planning was 'value-free' or 'value-full'. If 'value-full,' were the values about process (fair, just, 'rational'), or about substance (achieving specific goals, for example, aesthetically-pleasing environments, quality of life for ordinary people, redressing the unequal distribution of resources and power etc.)? By the end of the century, process values had been extended to include qualities of inclusion, empowerment, and respect to all citizens, while substance values were infused with ideas about environmental sustainability and social justice. However, if the substance and process of governance for complex territorial problems are seen to co-evolve, then it must be expected that values too co-evolve. This implies the development of critical capacities for 'reflection-in-practice' on what is evolving, what it might lead to and what values might be promoted as a result (Schon 1983). In work in planning theory, this has led to a critique of planning strategies and the design of planning systems. It has also focused on the personal dilemmas of 'doing' planning work.

2.3 System Maintenance or Transformation

The third challenge in planning theory lies in the approach to social change. Behind any idea of the 'good city' or 'good government' is some model of the nature and dynamics of social organization. For some advocates of planning ideas, it is the model itself that needs change. For others, the model needs correction to make it work better. As John Friedmann argued in the 1960s, some planning ideas sought to maintain basic systems. Others aimed to transform them. At the time, he was contrasting the management needed to trouble-shoot in western democracies with the transformations needed in illiberal Latin-American dictatorships, newly experimenting with democracy. By the 1970s, he had become much more radical, demanding 'social transformation' in over-urbanized capitalist societies, dominated by global corporations.

Throughout the twentieth century, people have been attracted to the planning field through transformative idealism. In the second part of the century in Europe, the target of transformation was a form of capitalism which dominated life chances, structured the way urban and regional space was used and damaged environments across the globe. For many, it was a shock to discover that the 'planning systems' and 'planning practices' which actually evolved, seemed deeply committed to maintaining the very practices which they had sought to challenge. Ideas about alternative ways of organizing society, however, come up against the reality of the dynamics of social change. In some periods of tumultuous social change, any kind of planning is an impossibility. In some periods of entrenched stability, similarly there are few cracks of opportunity to mobilize to do things differently. Most of the time, most situations are somewhere in between. In the contemporary world of complex relationships and diffused power, change happens both through big events (for example, the OPEC oil crisis, the collapse of communism, the 1987 property crash, the rise to power of neo-liberal governments, etc.) and through lots of small shifts and mobilization episodes. Planning processes can help these shifts along. Instead of making plans which are assumed to be predictions of what will happen, making plans becomes more of an activity of social mobilization, encouraging new ways of understanding situations and developing new ways of working. Strategy development in such conceptions takes on more of the character of collective invention and risk-taking in trying to shape futures rather than the mid-century assumption that government could command resources and control futures.

3. The Contribution of Planning Theory

These debates in the planning field provide rich resources for imagining, developing, and critiquing governance practices as they impact on urban and regional dynamics in the twenty-first century. Formal-

ized in the planning-theory literature, they range through philosophical debate, through analysis of systems and practices, and through discussion of the situation of 'doing planning work.' If the planning field is to realize its potential to contribute, imaginatively and critically, to the 'turn' towards more place-focused policy agendas, planning theory needs to maintain a close and multi-faceted engagement with these evolving practices.

There are many 'windows' through which this engagement has developed. One is that of the 'planner,' working within specific contexts trying to make things happen differently. This has been the key contribution of John Forester, whose work speaks sympathetically to planners in complex situations around the world (Forester 1989). Another is local institutional practice. In recent years, this work has been closely linked with the disciplines of urban politics and urban and regional geography, exploring questions about the relation between politicians, political regimes, experts and citizens, and the way planning strategies and tools have been used, in particular institutional milieu. A third 'window' has focused on understanding the design of governance systems. This has examined competencies, law, resource allocation systems, and appropriate expertise, the 'top-down' levers of policy change, and how these interact with 'bottom-up' interpretive practices. Much of this work has been informed by the search for ways of changing governance practices. A fourth 'window' has concentrated on ideas about the good city and the good society, especially in relation to ideas of environmental sustainability and multicultural societies and on how these have worked out in practice. The critical challenge for the development of planning as an enterprise is to recognize that these different 'windows' do not represent separate compartments of the planning field. They merely cast a more intense lens over a particular set of relations. All the others are still there, and need to be recognized as present, even if not the focus of a particular study.

The role of planning theory, as a distinctive task in this complex and dynamic institutional terrain, is to provide a critical imagination and structuring ideas for the dynamic invention processes which are continually unfolding in political communities in cities and regions. It should offer evaluative vocabularies for the recording and critiquing of these processes, both in relation to their content and their procedures. It should interpret emergent patterns and assess their consequences. It should draw in ideas from other fields and translate them in ways understandable to the various thought-worlds caught up in city governance. It should consolidate the distinctive understanding evolved by engagement with this activity and make it available to those working in other fields. This implies that the task is always an 'in-between' one. If the 'practitioners' of planning theory withdraw too far from the worlds of those engaged in 'doing planning,'

then they lose their capacity to record, shape, connect, and translate. If they stay too close to the 'doers,' they are unable to see beyond what the doers themselves see, and therefore cannot 'add value' to the work of doing. The challenge for planning theory in the twenty-first century is thus to enrich the evolution of public policy focused on place quality and to do this in a way which fosters the flow of ideas between the many arenas engaged in shaping urban and regional governance.

See also: Advocacy and Equity Planning; Development and Urbanization; Landscape Architecture; Market and Nonmarket Allocation; Metropolitan Growth and Change; International Perspectives; Neighborhood Revitalization and Community Development; Planning Ethics; Planning Issues and Sustainable Development; Planning, Politics of; Planning Theory (North American); Population Pressure, Resources, and the Environment: Industrialized World; Rural Planning: General; Strategic Planning; Sustainable Development; Urban Growth Models; Urban Places, Planning for the Use of: Design Guide; Urban Planning: Methods and Technologies; Urban Planning (Western), History of

Bibliography

- Barrett S, Fudge C 1981 *Policy and Action*. Methuen, London
 Beauregard R A, Body-Gendrot S 1999 *The Urban Moment*. Sage, Thousand Oaks, CA
 Boyer C 1983 *Dreaming the Rational City*. MIT Press, Boston, MA
 Campbell S, Fainstein S 1996 *A Reader in Urban Theory*. Blackwell, Oxford
 Faludi A 1973 *A Reader in Planning Theory*. Pergamon, Oxford
 Forester J 1989 *Planning in the face of power*. University of California Press, Berkeley, CA
 Friedmann J 1987 *Planning in the Public Domain: From Knowledge to Action*. Princeton University Press, NJ
 Hall P 1988 *Cities of Tomorrow*. Blackwell, Oxford
 Healey P 1997 *Collaborative Planning: Shaping Places in Fragmented Societies*. UBC Press, Vancouver
 Innes J 1995 Planning theory's emerging paradigm. *Journal of Planning Education and Research* 14: 183–9
 Mumford L 1961 *The City in History*. Harcourt, New York
 Sandercock L 1998 *Towards Cosmopolis*. Wiley, London
 Satterthwaite D (ed.) 1999 *Earthscan Reader in Sustainable Cities*. Earthscan, London
 Schon D 1983 *The Reflective Practitioner*. Basic Books, New York

P. Healey

Planning Theory (North American)

1. The Invention of Planning Theory

In the quarter of a century prior to World War I a small international movement centered in the North

Atlantic Basin sought to establish 'city planning' (under one or another of its names) as a professional craft and an imaginative sensibility. The movement cast itself in a heroic role. The new practice of planning, speakers at the founding conferences insisted, would correct and tame (what they represented as) the 'wild' and essentially 'unplanned' urban growth of the nineteenth century. The new practice would infuse the fragmented work of professional architects, social workers, hygienists, and civil engineers with an integrative civic vision. The image of the well-ordered city—symbolized by a plan that would guide its growth 'tomorrow'—was presented as a global icon of modernity. Chicago, New York, London, Birmingham, Paris, Berlin, and Vienna would learn from one another and tutor Tokyo, Shanghai, Cairo, Delhi, Colombo, and Rio de Janeiro (Sutcliffe 1981).

Despite their expansive aspirations, the leaders of the planning movement—both before and after World War I—built their professional careers within a conception of cities as well-bounded spaces filled with physical structures that organized 'public' activities. (The founders argued among themselves: Were houses public or private, and were they appropriately part of the planning brief?) Through the first half of the twentieth century the 'theory of town planning' was articulated largely in speculative proposals for new spatial forms, in the tacit assumptions of working plans, and in textbooks legitimating the 'rational' integration of 'architecture and urbanism.' The movement did not develop a distinctive theorizing tradition assessing the connection between the design of planning practices, processes or institutions, and the character of plans; between the character of plans and of the worlds they shaped (Choay 1997).

In the second half of the twentieth century, all of that changed—though not everywhere in quite the same way, and certainly not in an instant. The demand for professional planners grew dramatically after World War II, and universities responded by creating academic programs. Where planning education remained an adjunct of engineering or architecture, procedural theories played a relatively small role in shaping the new curricula or the public statements of the professoriate. Where, however, planning was articulated as a distinctive craft, the success or failure of 'its theory' were vital elements of the new professional identity. ('Professions,' it was widely claimed, 'needed a theory.')

That 'need' did not, of course, sweep away the intellectual difficulties that made the construction of procedural planning theories so difficult. The familiar forms of social theorizing were (and continue to be) designed to explain repeated historical behaviors, or to forecast the path of systems that we cannot, or do not choose to, change. In contrast, planning processes in their varied forms attempt to control (rather than to explain) a world that does not yet—and may never—exist. No matter how sophisticated, planners find

themselves torn repeatedly by the discovery that when, imaginatively, they unfreeze relations in order to change them it becomes harder to anticipate the outcomes of their actions, mistakes are legion, and the burdens of control grow heavy. The internal intellectual history of the attempt to construct and validate theories of planning in the last half of the twentieth century is centered in that dilemma.

The builders of the epistemic community devoted to planning theory created a full array of disciplinary artifacts: Journals, conferences, anthologies, review articles, 'fields' for doctoral students to master, careers, and stylized self-criticisms. Inevitably, favored failings were part of the community's identity. Insiders worried that their enterprise was too theoretical or not theoretical enough; that it ignored the spatial core of urban planning in order to run after foreign gods, or that it was too insular in its close attention to planning professionals.

Some of the builders—both early and late—imagined that they might in time satisfy the aspirations of practitioners and educators for a coherent and cogent theory that would guide and legitimate professional city and regional planning. At mid-century, the small Program for Education and Research in Planning at the University of Chicago, radiating confidence in the promise of rational planning, achieved an abiding place in the mythology of the planning Academy despite—or perhaps because of—its premature demise. By the end of the century, however, that promise seemed to most members of the community of theorists either misguided in principle or unachievable in practice. At its conferences and in its journals, the community acted as a loosely coupled network of distinctive and often contending groups. Planning theory—in the singular—appeared as an ironic misrepresentation of the varied products of those distinctly plural circles.

Individual theorists have not, of course, been immune from the arrogance of expertise. The procedural authority of the community of theorists has, however, been extremely limited in most political regimes. Unlike architects and engineers, theorists have never been certified by public or quasi-public authorities. Nowhere is the approval of a board of theorists required to validate procedural designs or to approve the sale of malpractice insurance. Theorists are not legally responsible for flawed organizational proposals, incoherent arguments, or the misreading of texts. They do, of course, hold each other accountable for simple foolishness, complex confusions, and the corruption of novices but their (loose) mutual discipline cannot be more demanding than the structure of the field will allow.

Perhaps because of that freedom from responsibility, the theorists have been intellectually ambitious. Undeterred by ordinary scholarly boundaries and without asking anyone's permission, they have extended the privileges of membership in their com-

munity to 'colleagues'—dead and alive—who had little direct connection with city and regional planning, and who rarely noticed the appropriation of their work. Sins of omission play a more important role in the stylized self-criticism of the community than sins of commission. A claim that a voice has been neglected commands the community's attention in the way that a broken tooth commands an inquiring tongue.

Whatever their initial aspirations, the theorists have not created a confident science of planning. To those who read, they have, however, suggested both alternative planning modes and subtle similarities in overtly different circumstances (Krieger 1981, Verma 1998). They have sustained an integrative bridge across diverse domains—land use, environmental and social policy, the design of policies and the associations of civil society, and local and regional economic development—that stretch far beyond the conventional scope of urbanism. Planning theory at the end of the twentieth century appeared as a counterbalance to the authority of local practices, legal mandates, urban regimes, technical expertise, and settled professional doctrines. Those were, however, powerful enemies. It is little wonder that theorists—offering at their best only criticism, questions, and suggestive intimations—lived at the margins of the professional planning enterprise with its characteristic and quite sensible respect for tacit knowledge and uniform practices. Within the main body of professional practitioners, theorists prospered only when they created formal protocols for representation, design, evaluation, and negotiation that could be replicated in many settings and invested with political authority. Geographic Information Systems (now inflated to Geographic Information Science) were the rage of planning agencies at the end of the century. GIS were, however, only the latest in a long series of tools—first among them, cost/benefit analysis—whose substantial philosophical baggage is characteristically left at the agency door (Pickles 1995).

2. The Platform of Planning Theory

The conception of 'architecture and urbanism' that was displaced by planning theory had integrated four assumptions:

- (a) the object of planning was the physical form of a settlement;
- (b) designers could specify the ideal, rational form for a settlement in a particular site;
- (c) rational forms had wholly predictable and appropriate outcomes; and
- (d) rational form should dominate other design preferences or proclivities.

The critical challenge to these assumptions was articulated on the one side in a direct assault on geographic determinism and the cogency of architectural conceptions of urban form; and on the other

side, by an almost casual dismissal of the 'dream of a rational city.' The texts that codified theories of spatial design and the debates of the first half of the century virtually disappeared from the 'talk' of the emergent community of theorists. Nigel Taylor, in his history, *Urban Planning Theory Since 1945* (1998) lists five mid-century texts in the 'physicalist' tradition: none of them was included in the bibliography of Andreas Faludi, *Planning Theory* (1973b), Kevin Lynch, *Good City Form* (1983), or John Friedmann, *Planning in the Public Domain* (1987). Faludi (1973a, p. x) characterized his own frequent use of physical planning examples as 'coincidental, a mere reflection of the fact that is the field where I come from and in which I teach.' Friedmann did not include 'urbanism' in his graphic presentation of distinctive planning traditions since the great revolutions of the eighteenth century. New recruits into the community of theorists should (in his account) place themselves in one of those traditions as heirs of an Enlightenment belief that scientific knowledge can and ought to be applied effectively to social improvement.

The combination of a frontal and a 'casual' challenge to the integrative conception of architecture and urbanism created an intellectual space in which it was possible to rethink the practices of architecture and urban design, and to create a platform for the construction of planning theory as a field of inquiry and reflection. The major premise of that platform was that though planning processes, settings, and outcomes were entangled, it was useful to distinguish between them. Within the framework of that heuristic distinction, theories of planning should articulate and explain the attributes of each element and illuminate the relations among them: If planners—lay or professional—act in a particular way in settings of such and such attributes, what are the likely outcomes? If they seek outcomes of a particular sort in a given setting, how should they plan? What sort of outcomes should they seek, and should their choices vary from one setting to the next? The intellectual history of planning theory is usefully understood as a tale of a community learning to live on that platform—a stage of its own making. (The best way to follow this tale is probably through the series of anthologies cited in the bibliography.)

The practitioners who shaped the early-twentieth-century planning movement embraced an ambitious conception of planning as a process—and as a profession—that was uniquely rational, scientific, or self-conscious; that looked into the future and across conventional institutional domains; and that served the public interest or other comparably admirable objectives. The emergence of the new planning academy after World War II did not dim that heroic self-image. The rationality of urban form was, however, replaced by the intellectual claims of a toolkit of protocols for representing complex systems and their cybernetic controls, for evaluating projects and

programs, for modeling urban land use and transportation patterns, and for revealing the dynamics of location, spatial differentiation, and interpersonal negotiations. Those protocols—*inter alia* systems theory, experimental design, mathematical simulation, factorial ecologies, and game theory—had a profound impact on the first postwar generation of theorists, displacing the integrative conception of urbanism but retaining an essentially rationalist justification for professional planning. The mission of theory was not to describe or explain planning processes but to shape them so that they were ethically compelling and instrumentally effective (or even efficient). Normative theorizing would embed those demanding criteria in the identities of individual planners, the practices of planning agencies, and the norms of a disciplined profession. (For exemplary statements of that rationalist ethos, see Chadwick 1971 and Branch 1990.)

Those ambitious rationalist protocols and aspirations did not control the platform for very long before they evoked a critical counterattack on the ‘new utopians’ and the pathologies of technocracy (Boguslaw 1965). Some of these attacks came from theorists who had never been true believers in an age of uniquely rational planning. Perhaps more interesting, were subtle changes in the thought of rationalists (such as Melvin Webber or Britton Harris) responding to the difficulties of managing the indeterminate complexity of ‘wicked’ problems, and of bringing knowledge to bear upon even simple planning issues when ordinary words carried so many conflicting meanings (Rittel and Webber 1973, Vettoretto 1997, Harris 1989, 1997). Both internal and external voices challenged the assertion that a corps of professional planners could articulate and defend a unique social welfare function, or that the difficulties of collective action could be resolved consensually by privileged institutions or tools. Accounts of urban and regional planning activities that dealt only with the behavior of professional planners obscured the ways in which a great many agents attempted to plan, playing both with and against one another. It may be useful to represent planning as a system, but neither the governor in that system nor the feedback loop need appear as a professional planner or agency. In many settings, governance is the work of a political regime, a market, or an interorganizational field. In Charles Lindblom’s account, the ‘intelligence of democracy’ depended on the design of interaction as well as the practice of analysis. Theories addressed only to the creation and maintenance of an expert profession of analysts could not explain adequately the dynamics of those interactive processes, institutions, and practices. Nor could they provide compelling support for their redesign (Lindblom 1965, 1990).

The critique of the rationalist toolkit was extended by a reinterpretation of reason and rationality. These ideas, circa 1950, justified claims as fitting within a

coherent world view; and as cogently linking a chain of assertions to core propositions without violating legitimate rules of discourse. By 1975—or even earlier—the implications of these criteria had been altered substantially by the recognition of multiple worlds, discursive formations, and pragmatic tests of truth claims.

Paradoxically, the shift was signaled most clearly by the simultaneous publication in 1973 of Andreas Faludi’s anthology celebrating the achievements of postwar planning theorists, and his own treatise synthesizing these achievements in an account of three polar modes in which intelligent professional agencies sought to enhance human growth: blueprint vs. process, rational comprehensive vs. disjointed incrementalist, and normative vs. functional planning. By articulating these polarities as foundations of a positive theory of planning, Faludi revealed how little the community of theorists knew of the variability of these modes over time; and of the ways in which the field of planning organizations and institutions responded to the attributes of the settings upon which it acted. The description of three polarities virtually begged for additional options that would—and that did—enrich the planning repertoire, but that also made it impossible to cultivate a complete general theory of planning. Though Faludi’s description of ‘human growth’ as the goal of planning had a distinguished intellectual pedigree, it was obscure and (in the event) quickly dominated by the intense interest of theorists in John Rawls’ (1971) conception of ‘justice as fairness.’ The theorists, though deeply engaged by the philosophic disputes surrounding *A Theory of Justice* (and subsequently by Rawls’ responses to his critics), did not, however, agree unanimously that justice was a superordinate goal that should guide all planning choices, that the conditions of the original position would yield a rational consensus on Rawlsian principles, or that the distinctions between rights and goods (and thick and thin) were compelling.

In practice, reason and rationality had always spoken with diverse voices. In the last quarter of the twentieth century, they were intractably pluralistic in theory. Pluralism implied that reason and rationality were neither permanent nor universal criteria. Encoded in planning doctrines, they were precious but temporary and local public orders. In different ways, both Richard Rorty and Jürgen Habermas—as they were coopted by the community of planning theorists—defined the planning ‘problem’ as the design of legitimate ‘conversations.’ Closer to the core of the community, the fascination with ‘participation’ as a remedy for the putative flaws of both autocratic and representative institutions privileged the voices of engaged advocates rather than heroic rationalists.

Without the taming influence of a commanding conception of rationality, the pluralist expectation encouraged an increasingly complex conception of the architecture of the intellectual field. In both positive

and normative theorizing the distinctions between process, setting, and outcome could be drawn in many different ways. A development coalition committed to promoting growth might be treated as a setting permitting a comparative study of professional planning processes in an array of urban regimes. A theory trying to explain differences in development patterns might, however, represent the coalition as the most important dimension of the urban planning process. The theorist in such a field is a designer adjusting and readjusting the three elements in order to find a process that can deal with a setting, a setting that can sustain a process, and a conception of outcomes that will not shatter a rightly cherished world. The theorist as critical reader unpacks that complex judgment to test its credibility and its delight: Is that a world worth making?

By the mid-1980s many members of the community of theorists had accepted an image of their field as a lively but not very tidy patchwork—wall *bricolage* was the fashionable descriptive image. As theorists learned to create sensitive ethnographies of practice, to test the cogency of procedural typologies, and to appreciate the disjunction of what Thomas Schelling (1979) described as ‘micromotives and macrobehavior,’ the harder it became to make sense of the wall. Even the most elaborate plans—symbols of the authority and heroic aspirations of practitioners—were understood by theorists as vulnerable to the (highly variable) interpretations of uncontrollable readers and the secret influences embedded in the everyday language of planners. In Judith Innes’ (1990) report of her own conversion, the new ‘interpretive or phenomenological view of knowledge’ was ‘more contextual, more evolutionary, and more complex’ than the positivist scientific model in which she had cast her 1975 account of social indicators.

That was not, however, the end of the tale. Whatever their view of the political roles and possibilities of professional planning, theorists worried lest the literary forms of their arguments alienate both students and experienced practitioners; that it would encourage skepticism, and an insidious delight in uncertainty. They fantasized about a postmodern ‘abyss,’ in which both practical confidence and public meaning were lost.

The community struggled with those fears in the last two decades of the twentieth century. The attempt to control a pluralism that was deeply treasured took at least four different forms—some new and some part of its initial endowment. The first responsive form, as in the complementary work of Ernest R. Alexander (1996) and Andreas Faludi (1996), accepts the difficulty of articulating and defending a complete general theory of planning. It insists, however, that the patchwork of typologies, ethical claims, and empirical findings can be integrated by a robust planning doctrine into a series of contingent protocols: Within the frame of such a doctrine, if a situation changes

from A to B, there is a shift from the now inappropriate mode to another that better fits the current setting. These protocols are more complex than a single theory but not so difficult (or so numerous) that they deny the possibility of achieving consensual agreements on the nature of processes, settings, and outcomes, and their relationships.

The second response addresses pluralism and the patchwork of theory in a quite different way. Surely, the worlds we plan are remarkably varied and often the devil is in the details of those differences. Nevertheless, the advocates of this response—such as David Harvey (1985)—insist, there are really only a small number of compelling ways to describe and explain planning processes, institutions, and practices; only a small number of compelling ethical systems and principled social conflicts. We misrepresent the condition of planning theory, they insist, when we fail to recognize the ways in which the great global systems—e.g., capitalism, statism, democracy, colonialism, racism, liberalism—are defended by the construction of the illusion of huge variations and the exaggerated promise of plasticity.

The third response divides the patchwork of theory into two parts. One—planning proper—is subject to all of the familiar perils of indeterminacy, complexity, and interpretive variety. The second part is the domain of planning support systems that tame those perils (even if they cannot eliminate them) by establishing consensual representations of ‘fact’ and ‘structure,’ allowing us to elaborate the implications of our own images and plans, and stabilizing public orders (Klosterman 1997). Serious support requires, of course, that the choice of representational modes and data sets address the protean requirements of planning processes. The support we seek when we are encouraging divergent thinking early in a planning process would be disruptive or inadequate as we fill in the final details of a fully elaborated scheme. Data sets or maps that cloak essentially contested meanings in a formal garb cannot transform (or support) planning debates: They only reproduce them. Strategic planning *may* not be well-served by the data that regulates operational management; or organizations in a turbulent environment by the information practices of a placid field (Emery and Trist 1965).

Finally—in what is often represented by its critics as the dominant school within the community (Healey 1997, Yiftachel 1999, Throgmorton 1999)—the communication strategies that have accentuated the pluralism of both planning and planning theory have been engaged to limit the perils of pluralism. Initially, the project of communicative enlargement served principally to encourage the voices of silent groups and to ‘pay attention’ to their lives, to democratize access to knowledge, and to multiple the opportunities to be heard. This project is by no means completed to the satisfaction of its partisans, but it has now taken a somewhat different turn. How is it, many theorists

(e.g., Baum 1997, Forester 1999, Mandelbaum 2000) wonder, that we sometimes manage to understand and change one another, and to resolve moral conflicts? What sorts of rhetorical practices and what sorts of settings sustain a practice of reflection, successful mediation, cooperation within and between organizations, and mutual trust? How—even more profoundly—may strangers learn to deliberate across the ordinary divides of adversarial democracies? What does ‘justice’ require of our speech and of the distribution of knowledge?

The current theorizing about communication in planning is capacious. It is not only about talk and texts, but about the ways in which symbolic meanings are imputed to actions and symbols are acts; about the ways in which arenas, forums, and courts (Bryson and Crosby 1992) shape communication patterns. The focus on communication accommodates ethnographies of planners at work, close studies of the dynamics of communities as moral associations, critical assessments of the networks that support the ‘new information technologies,’ civil societies and political regimes at every level, and the detailed protocols (Friend and Hickling 1997) that support the work of complex task forces. Inevitably, the links between these diverse intellectual forms sometimes seem very fragile, and the communication explanation and remedy tendentious. The power of the communication approach lies in the ways in which it makes sense of the diversity that theorists discover as they engage one another on the shared platform. Though theorists have not articulated a consensual set of findings or shared uncertainties, they have found a way to talk about their disagreements, to identify themselves within intellectual traditions, and to map prospective bridges across contested grounds.

See also: Architecture; Community, Social Contexts of; Development Theory in Geography; Planning Issues and Sustainable Development; Planning, Politics of; Planning Theory: Interaction with Institutional Contexts; Urban Growth Models; Urban Places, Planning for the Use of; Design Guide

Bibliography

- Alexander A 1996 After rationality: Towards a contingency theory for planning. In: Mandelbaum S J, Mazza L, Burchell R (eds.) *Explorations in Planning Theory*. Center for Urban Policy Research, New Brunswick, NJ, pp. 45–64
- Baum H S 1997 *The Organization of Hope: Communities Planning Themselves*. State University of New York Press, Albany, NY
- Boguslaw R 1965 *The New Utopians: A Study of System Design and Social Change*. Prentice Hall, Englewood Cliffs, NJ
- Branch M C 1990 *Planning: Universal Process*. Praeger, New York, NY

- Bryson J, Crosby B 1992 *Leadership for the Common Good: How to Tackle Public Problems in a Shared Power World*. Jossey Press, San Francisco, CA
- Burchell R W, Sternlieb G 1978 *Planning Theory in the 1980s*. Center for Urban Policy Research, New Brunswick, NJ
- Campbell S, Fainstein S (eds.) 1996 *Readings in Planning Theory*. Blackwell, Cambridge, MA
- Chadwick G 1971 *A Systems View of Planning: Towards a Theory of the Urban and Regional Planning Process*. Pergamon Press, Oxford, UK
- Choay F 1997 *The Rule and the Model: On the Theory of Architecture and Urbanism*. In: Bratton D (ed.) MIT Press, Cambridge, MA
- Emery F, Trist E L 1965 The causal texture of organizational environments: An approach to macroorganizational theory. *Human Relations* 27: 201–38
- Faludi A 1973a *Planning Theory*. Pergamon Press, Oxford, UK
- Faludi A (ed.) 1973b *A Reader in Planning Theory*. Pergamon Press, Oxford, UK
- Faludi A 1996 Rationality, critical rationalism, and planning doctrine. In: Mandelbaum S J, Mazza L, Burchell R (eds.) *Explorations in Planning Theory*. Center for Urban Policy Research, New Brunswick, NJ, pp. 65–82
- Forester J 1999 *The Deliberative Practitioner: Encouraging Participatory Planning Processes*. MIT Press, Cambridge, MA
- Friedmann J 1987 *Planning in the Public Domain: From Knowledge to Action*. Princeton University Press, Princeton, NJ
- Friend J, Hickling A 1997 *Planning Under Pressure: The Strategic Choice Approach*, 2nd edn. Butterworth Heinemann, Oxford, UK
- Harris B 1989 Beyond geographic information systems: Computers and the planning professional. *Journal of the American Planning Association* 55: 85–90
- Harris B 1997 The theory of planning and of its profession. *Environment and Planning B: Planning and Design* 24: 483–9
- Harvey D 1985 *The Urbanization of Capital: Studies in the History and Theory of Capitalist Urbanization*. Johns Hopkins University Press, Baltimore, MD
- Healey P (ed.) 1997 Planning theory, political economy and the interpretive turn: The debate continues! *Planning Theory* 17: 10–85
- Healey P, McDougall G, Thomas M J (eds.) 1982 *Planning Theory: Prospects for the 1980s*. Pergamon Press, Oxford, UK
- Innes J E 1990 *Knowledge and Public Policy. The Search for Meaningful Indicators* (2nd (expanded) edn of 1975 title *Social Indicators and Public Policy: Interactive Processes of Design*). Transaction Publishers, New Brunswick, NJ
- Klosterman R E 1997 Planning support systems: A new perspective on computer-aided planning. *Journal of Planning Education and Research* 17: 45–54
- Krieger M H 1981 *Advice and Planning*. Temple University Press, Philadelphia, PA
- Lindblom C 1965 *The Intelligence of Democracy*. Free Press, New York, NY
- Lindblom C 1990 *Inquiry and Change: The Troubled Attempt to Understand and Shape Society*. Yale University Press, New Haven, CT
- Lynch K 1981 *Good City Form*. MIT Press, Cambridge, MA
- Mandelbaum S J 2000 *Open Moral Communities*. MIT Press, Cambridge, MA
- Mandelbaum S J, Mazza L, Burchell R (eds.) 1996 *Explorations in Planning Theory*. Center for Urban Policy Research, New Brunswick, NJ

- Pickles J (ed.) 1995 *Ground Truth: The Social Implications of Geographic Information Systems*. Guilford Press, New York
- Rawls J 1971 *A Theory of Justice*. Harvard University Press, Cambridge, MA
- Rittel H W, Webber M 1973 Dilemmas in a general theory of planning. *Policy Sciences* 4: 155–69
- Schelling T C 1978 *Micromotives and Macrobehavior*. Norton, New York
- Sutcliffe A 1981 *Towards the Planned City: Germany, Britain, the United States, and France, 1780–1914*. St. Martin's Press, New York
- Taylor N M 1998 *Urban Planning Theory Since 1945*. Sage, Thousand Oaks, CA
- Verma N 1998 *Similarities, Connections and Systems: The Search for a New Rationality for Planning and Management*. Lexington Books, Lanham, MD
- Vettoretto L 1997 Planning theory and fragmented societies. Explorations in Melvin Webber's civic liberalism. *Planning Theory* 17: 113–41
- Throgmorton J A 1999 Learning through conflict at Oxford. *Journal of Planning Education and Research* 18: 269–70
- Yiftachel O 1999 Planning theory at a crossroad: The Third Oxford Conference. *Journal of Planning Education and Research* 18: 267–9

S. J. Mandelbaum

Plasticity in Human Behavior across the Lifespan

The concept of plasticity originated in the medical sciences and refers to the capacity of organism to be modifiable. This variability and flexibility enables human beings to adapt to a great range of new and changing environmental conditions. Starting from this origin, the concept of plasticity has been applied to a broad scope of phenomena, the plasticity of nervous systems, changes in observable behavior patterns, and life trajectories.

Neuronal plasticity research, for example, addresses the biochemical and neuronal processes underlying the development and the variability of behavior over the lifespan; that is, it refers to the mechanisms of brain development that involve changes in structure and function (brain plasticity). A key concept in this field was the 'critical period' (Wiesel and Hubel 1963) which refers to a phase in the lifespan when neuronal structures are highly malleable based on the level and type of activity exposed to. For organisms at increasing maturity levels it was expected that neurostructural plasticity decreases with age and is absent in the mature cortex. However, subsequent research revealed substantial plasticity in the mature adult cortex, for example after lesions, strokes, and nerve damage or—even more important—after learning occurs (e.g., Buonomano and Merzenich 1998, Recanzone 2000). Studies showing activity-dependent plasticity on brain structures and transmitter systems in

the aging brain provided convincing evidence that brain plasticity is not restricted to juvenile stages of development, but at least in the human species covers the entire lifespan (e.g., Cotman and Neeper 1996).

In contrast to neuronal plasticity research, the focus in lifespan psychological plasticity research lies on identifying the range and the conditions of behavioral modifiability as well as their age-related constraints across the entire life course (behavioral plasticity). In that context, the concept of plasticity generally refers to the ability of human beings to display rather different patterns of behavior as well as different life trajectories (Baltes and Willis 1982, Lerner 1984). So, plasticity can refer to variability of life course patterns (ontogenetic plasticity) or to the variability of a performance in a specific moment in time (concurrent plasticity). Moreover, the study of intraindividual behavioral variability is not restricted to cognitive functioning but extends to noncognitive domains such as motivation and personality. To understand changes in behavioral and ontogenetic plasticity across the lifespan, broader theoretical frameworks of human functioning have to be considered that take the dynamic and complex interplay between evolutionary/biology-based and culture-based influences on the nature of lifespan development into account (Baltes et al. 1998, Finch 1996).

1. Human Behavioral Plasticity as a Product of Evolution and the Foundation of Culture

Much of the behavioral and ontogenetic plasticity is based on the phylogeny of mammals, a biological strata characterized by open behavioral programs (Mayr 1974). With few exceptions, human behavior is not regulated by preprogrammed stimulus-response patterns, but entails the potential to adapt to a variety of ecologies and stimulus constellations. This plasticity opens behavioral patterns both in terms of concurrently available behavioral options and—even more important for ontogenetic change—in terms of acquiring new behavioral patterns. This holds too and is even accentuated when one views the evolution of the human mind from a perspective of evolved domain-specific behavioral and cognitive modules (Cosmides and Tooby 1994, Fodor 1983), which are open to ecological and cultural variations. In the course of human phylogeny such evolved mental modules have escaped domain-specific encapsulation (see Fodor's original conception of encapsulated modules, Fodor 1983) and thus provided the building blocks of human intellectual capacities and behavioral adaptiveness. This variability in behavior and development is as much a product of evolution as the more narrowly constrained behavioral systems of other species (e.g., Wilson 1980).

Behavioral plasticity opens the potential for systematic influences of experiences, most notably in-

fluences of socialization and instruction. Both on the part of the socializing agent and recipient, flexible regulation of behavior in the interaction allows the partners to move beyond preformatted parenting and attachment programs of behavior, and reach out for the social—and ultimately cultural—transmission of experience between generations, and more generally between experts and novices. The ability to instruct and to profit from instruction probably is the critical difference in behavioral regulation which sets *homo sapiens* apart from other primates (Tomasello et al. 1993). Only humans seem to intentionally instruct their children (or nonexpert adults) to modify behavior in accordance with a standard or goal. They correct, model subcomponents, and thus selectively modify error-prone behavior of a partner to fit a certain standard. This specific instructional potential of human interactions is the vehicle of fidelity in information transmission (Tomasello et al. 1993), the prerequisite of cultural heritage. In modern societies this instructional potential has led to the institutionalization of instruction in educational institutions, such as schools and universities. Instructional interactions between humans and the training in educational systems have in turn greatly amplified the range of plasticity in human behavior. The combination of evolutionary potential and cultural transmission can thus generate an immense, albeit not infinite, universe of behavioral plasticity in many potential developmental pathways. Indeed, this great potential for variability, plasticity, and adaptivity in human ontogeny is a cornerstone of lifespan developmental theories (Baltes 1997, Lerner 1984).

2. Plasticity of Performance Across the Lifespan

The range of plasticity in performance in diverse domains of functioning (e.g., athletic, intellectual, creative) is determined by the joint influences of either the genotype-related component of performance, that is, genetically-based interindividual differences in performance, or the phenotype-related component of performance, that is, intraindividual differences based on current performance factors (e.g., motivation, alertness), accumulated investment, and instruction. While genotype-related influences delimit the range of plasticity for any given individual, the phenotype-based influences determine whether this potential of plasticity is actually realized (Kliegl and Baltes 1987). Specifically, these four influences are the following. First, there are age-graded differences in the baseline opportunities for maximum performance. It is well known and empirically established that across the human lifespan, performance in various domains of functioning undergoes radical changes, typically reflecting an inverted U-shaped trajectory, with a maximum performance potential at a given age (Lehman

1953, Simonton 1988). Age of maximum performance is relatively early for most athletic disciplines, particularly the track-and-fields sports (Schulz and Curnow 1988), and moves further into adulthood for areas of peak performance which require extensive expertise, such as soloist musical performance (Ericsson et al. 1993), science, and poetry (Simonton 1988). Second, a great many contextual factors can influence an individual's baseline reserve capacity (Kliegl and Baltes 1987). Thus, the component of reserve capacity which influences current performance can be optimized by improving motivation, alertness, clarity of task instruction, and other contextual factors. Albeit opening up many avenues for intervention to promote improved performance, these performance factors are problematic from the perspective of basic research (Salthouse 1985). The performance factors can vary with age and thus confound age differences, so that accurate assessments of latent performance potential at different ages is compromised. Baltes, Kliegl and their colleagues have proposed the methodology of Testing-the-Limits as a research strategy to assess cognitive reserve capacity in terms of latent performance potentials. This research paradigm optimizes performance factors for all individuals and assesses their ability to profit from a training designed to construct a specific cognitive skill, for which the different age groups do not hold differential experiences (Kliegl and Baltes 1987). This latter component of instruction-related plasticity of performance is the third influencing factor, addressed next. Third, performance can be optimized by providing appropriate instruction. The influence of instruction on performance is a long-standing research topic in diverse fields of study and approaches to psychological functioning. Vygotsky conceptualized the zone of proximate development as the range of potential performance an individual can attain with the help and instruction of an expert interactive partner (Vygotsky 1978). This conception about developmental reserve capacity in socially constructive interactions has spawned much research on social facilitation in parent-child dyads (scaffolding, Wood et al. 1976; apprenticeship, Heckhausen 1987, Kaye 1982, Rogoff and Wertsch 1984), child-child interactions (Rogoff 1990), cognitive aging (development-enhancing interventions, Kliegl and Baltes 1987), and educational psychology (see zone of potential development, Brown and French 1979). Finally fourth, plasticity in performance is a function of the individual's long-term investment in a given area of skill or expertise. Research on life-course trajectories of peak performance across a wide variety of domains has demonstrated repeatedly that world-class levels of performance require extensive and high-quality expert-guided training and practice over long periods of the lifespan, particularly in the early phases of acquisition (Ericsson and Charness 1994, Simonton 1988). This powerful effect of focuses and prolonged

practice and training is itself the best witness of the impressive plasticity of human performance across the lifespan.

3. Cognitive Functioning in Old Age: A Case in Point

An extensively studied area of human behavioral plasticity is cognitive functioning in old age. Numerous studies have demonstrated declining performance in fluid intelligence performance with increasing age among older adults (Salthouse 1985). This pattern of age related decline has been challenged to be reflective of true latent performance potential, since many factors such as motivation or lack of recent practice could produce performance differences such as those found between older and younger adults in intellectual functioning. This critique prompted researchers to optimize performance conditions for older adults for instance by using ecologically valid tasks. Moreover, specifically designed training programs were administered to promote improved performance in older adults (Verhaeghen et al. 1992, Willis 1989). Examples of early training studies are among others the Penn State Adult Development and Enrichment Project (ADEPT) and the Berlin PRO-ALT project (for review see, Baltes and Lindenberger 1988). These studies demonstrated substantial benefits of older adults' fluid intelligence performance from cognitive training. For tasks that were similar to those addressed by the cognitive training, older adults improved their performance. However, the training effect showed no transfer to other fluid intelligence tasks. Additional studies comparing the benefit of guided instruction with self-regulated practice showed that self-instruction was as effective in improving cognitive performance in older adults as was guided instruction. This suggests that a large share of the old age plasticity in cognitive functioning is due to the activation of already available cognitive skills, which have been fallen into disuse over the adult lifespan.

The findings of preserved substantial learning potentials up to old age and the possibly arising optimistic view of unchanged cognitive plasticity over the entire lifespan had been put in perspective by Testing-the-Limits studies aiming to delimit maximum performance potentials (Kliegl and Baltes 1987). In this research, cognitive skills are taught to subjects using specific training programs to construct an expertise in a certain cognitive domain, for example in the domain of memory for nouns. In this way, an expertise is cognitively engineered, so that the researchers know the component abilities, and can thus specifically compare subjects at different ages in terms of specific subskills. Subjects are trained extensively until they have reached asymptotic performance levels. Once maximum performance levels are achieved, time constraints are introduced. The findings indicate that

although older adults can reach high levels of performance (e.g., in recalling nouns), their maximum (asymptotic) performance level after extensive training is much lower compared to younger adults. Moreover, their functioning breaks down when they encounter severe time constraints. Thus, at limits of functioning, age differences are magnified and reveal the potentially underlying neurobiological constraints associated with the biology of the aging brain (Baltes and Kliegl 1992). Finally, Testing-the-Limits methodologies have been successfully employed to identify elderly at risk of dementia (Baltes et al 1992) and distinguish pathological from normal cognitive aging (Reischies and Lindenberger 1996).

4. Plasticity in Personality Across the Lifespan

Processes of self-regulation, personality, and subjective well-being show much less systematic age variation than cognitive functioning. Indeed, the stability of life satisfaction, self-esteem, perceived control, and overall happiness in itself is proof of the adaptive plasticity of these processes. The age-normative shift from a predominance of perceived gains to ever increasing risks for decline (Heckhausen et al. 1989) would lead one to expect concomitant declines in indicators and components of subjective well-being. Instead, emotional balance remains resilient across the life span (Staudinger et al. 1993).

This resilience of the emotional balance to lifespan changes is one of the key achievements of the human motivational system, which has evolved to manage the challenges associated with the relative dearth of biologically-based predetermination of behavior. The relatively weak biological regulation of behavior furnishes a regulatory need to achieve and maintain selectivity in behavioral investment, and to compensate for the ever-present failure and loss of a behavioral system that is underdetermined and thus highly error-prone. Both selectivity and loss compensation, key concepts in lifespan developmental research (Baltes et al. 1998, Schulz and Heckhausen 1999, Carstensen et al. 1999), require adaptive motivational processes, which focus volitional investment into a chosen goal and buffer the potentially negative effects of aging-related losses. Typical examples of such metamotivational processes (i.e., secondary control, Heckhausen and Schulz 1995) which facilitate goal striving, are upward social comparisons with superior or advanced others, who can serve as models for self-improvement (Heckhausen and Krueger 1993), and enhanced valuation of the chosen compared to competing goals. Examples of metamotivational processes, which compensate for the negative effects of losses or failures, are self-enhancing comparisons with inferior others, self-protective attributions which take credit for gains but refuse to take blame for losses, and the adjustment of goals to fit one's real potential (for the latter example

see also Brandtstädter and Renner's conception of accommodation, Brandtstädter and Renner 1992). In general, the age-graded structure of opportunities and constraints sets the stage for the individual to use the age window of maximal opportunities to strive for a given goal. The regulation of motivational investment requires substantial plasticity in that the individual needs to enter cycles of engagement and disengagement with various developmental goals (Heckhausen 1999). Motivational engagement and disengagement with specific life goals should be contingent and flexibly adapted to changes in opportunities and constraints across the lifespan.

See also: Behaviorism; Human Behavioral Ecology; Life Course in History; Life Course: Sociological Aspects; Lifelong Learning and its Support with New Media; Cultural Concerns; Lifespan Development: Evolutionary Perspectives; Lifespan Development, Theory of; Lifespan Theories of Cognitive Development

Bibliography

- Baltes M M, Kühl K P, Sowarka D 1992 Testing for limits of cognitive reserve capacity: a promising strategy for early diagnostic of dementia? *Journal of Gerontology: Psychological Sciences* **47**: 165–7
- Baltes P B 1997 On the incomplete architecture of human ontogeny: selection, organization, and compensation as foundation of developmental theory. *American Psychologist* **52**: 366–80
- Baltes P B, Kliegl R 1992 Further testing of limits of cognitive plasticity: Negative age differences in a mnemonic skill are robust. *Developmental Psychology* **28**: 121–5
- Baltes P B, Lindenberger U 1988 On the range of cognitive plasticity in old age as a function of experience: 15 years of intervention research. *Behavior Therapy* **19**: 283–300
- Baltes P B, Lindenberger U, Staudinger U M 1998 Life-span theory in developmental psychology. In: Lerner R M (ed.) *Handbook of Child Psychology, Vol. 1: Theoretical Models of Human Development*, 5th edn. J. Wiley, New York, pp. 1029–1143
- Baltes P B, Willis S L 1982 Plasticity and enhancement of intellectual functioning in old age: Penn State's Adult Development and Enrichment Project (ADEPT). In: Craik F I M, Trehub S (eds.) *Aging and Cognitive Processes*. Plenum Press, New York, pp. 353–89
- Brandtstädter J, Renner G 1992 Coping with discrepancies between aspirations and achievements in adult development: a dual-process model. In: Montada L, Filipp S-H, Lerner M R (eds.) *Life Crises and Experiences of Loss in Adulthood*. Erlbaum, Hillsdale, NJ, pp. 301–19
- Brown A L, French L A 1979 The zone of potential development: implication for intelligence testing in the year 2000. *Intelligence* **3**: 255–77
- Buonomano D V, Merzenich M M 1998 Cortical plasticity: From synapses to maps. *Annual Review of Neuroscience* **21**: 149–86
- Carstensen L L, Isaacowitz D M, Charles S T 1999 Taking time seriously: a theory of socioemotional selectivity. *American Psychologist* **54**: 165–81
- Cosmides L, Tooby J 1994 Origins of domain-specificity: the evolution of functional organization. In: Hirschfeld L A, Gelman S A (eds.) *Mapping the Mind: Domain Specificity in Cognition and Culture*. Cambridge University Press, Cambridge, UK, pp. 85–116
- Cotman C W, Nepper S 1996 Activity-dependent plasticity and the aging brain. In: Schneider L, Rowe J W (eds.) *Handbook of the Biology of Aging*, 4th edn. Academic Press, San Diego, CA, pp. 283–99
- Ericsson K A, Krampe R T, Tesch-Römer C 1993 The role of deliberate practice in the acquisition of expert performance. *Psychology Review* **100**(3): 363–406
- Ericsson K A, Charness N 1994 Expert performance: its structure and acquisition. *American Psychologist* **49**(8): 725–47
- Finch C E 1996 Biological bases for plasticity during aging of individual life histories. In: Magnusson D (ed.) *The Life-Span Development of Individuals: Behavioral, Neurobiological and Psychosocial Perspective*. Cambridge University Press, Cambridge, UK, pp. 488–511
- Fodor J A 1983 *The Modularity of Mind*. MIT Press, Cambridge, MA
- Heckhausen J 1987 Balancing for weaknesses and challenging developmental potential: a longitudinal study of mother-infant dyads in apprenticeship interactions. *Developmental Psychology* **23**: 762–70
- Heckhausen J 1999 *Developmental Regulation in Adulthood: Age-Normative and Sociostructural Constraints as Adaptive Challenges*. Cambridge University Press, New York
- Heckhausen J, Dixon R A, Baltes P B 1989 Gains and losses in development throughout adulthood as perceived by different adult age groups. *Developmental Psychology* **25**: 109–21
- Heckhausen J, Krueger J 1993 Developmental expectations for the self and most other people: age grading in three functions of social comparison. *Developmental Psychology* **29**: 539–48
- Heckhausen J, Schulz R 1995 A life-span theory of control. *Psychological Review* **102**: 284–304
- Kaye K 1982 *The Mental and Social Life of Babies. How Parents Create Persons*. University of Chicago Press, Chicago
- Kliegl R, Baltes P B 1987 Theory-guided analysis of mechanisms of development and aging through testing-the-limits and research on expertise. In: Schooler C, Schaie K W (eds.) *Cognitive Functioning and Social Structure over the Life Course*. Ablex, Norwood, NJ, pp. 95–119
- Lehman H C 1953 *Age and Achievement*. Princeton University Press, Princeton, NJ
- Lerner R M 1984 *On the Nature of Human Plasticity*. Cambridge University Press, New York
- Mayr E 1974 Behavior programs and evolutionary strategies. *American Scientist* **62**: 650–9
- Recanzone G H 2000 Cerebral cortical plasticity: Perception and skill acquisition. In: Gazzaniga S (ed.) *The New Cognitive Neurosciences*, 2nd edn. MIT Press, Cambridge, MA, pp. 237–47
- Reischies F M, Lindenberger U 1996 Grenzen und Potentiale kognitiver Leistungs-Fähigkeit im Alter [Limits and potentials of cognitive ability in old age]. In: Mayer K U, Baltes P B (eds.) *Die Berliner Altersstudie*. Akademie Verlag, Berlin, pp. 351–78
- Rogoff B 1990 *Apprenticeship in Thinking: Cognitive Development in Social Context*. Oxford University Press, New York

- Rogoff B, Wertsch J V (eds.) 1984 Children's learning in the "zone of proximal development". Jossey-Bass, San Francisco
- Salthouse T A 1985 A Theory of Cognitive Aging. North-Holland, Amsterdam
- Schulz R, Cunow C 1988 Peak performance and age among super athletes: track and field, swimming, baseball, tennis, and golf. *Journal of Gerontology: Psychological Sciences* 43: 113–20
- Schulz R, Heckhausen J 1999 Aging, culture and control: setting a new research agenda. *Journal of Gerontology series B: Psychological Sciences* 54: 139–45
- Simonton D K 1988 Age and outstanding achievement: What do we know after a century of research? *Psychological Bulletin* 104: 251–67
- Staudinger U M, Marsiske M, Baltes P B 1993 Resilience and levels of reserve capacity in later adulthood: Perspectives from life-span theory. *Development and Psychopathology* 5: 541–66
- Tomasello M, Kruger A C, Ratner H H 1993 Cultural learning. *Behavioral and Brain Sciences* 16: 495–511
- Verhaeghen P, Marcoen A, Goossens L 1992 Improving memory performance in the aged through mnemonic training: A meta-analytic study. *Psychology and Aging* 7: 242–51
- Vygotsky L S 1978 *Mind in Society*. Harvard University Press, Cambridge, MA
- Wiesel T N, Hubel D H 1963 Single-cell responses in striate cortex of kittens deprived of vision in one eye. *Journal of Neurophysiology* 26: 1003–17
- Willis S L 1989 Improvement with cognitive training: Which dogs learn what tricks? In: Poon L W, Rubin D C, Wilson B A (eds.) *Everyday Cognition in Adulthood and Late Life*. Cambridge University Press, Cambridge, UK, pp. 545–69
- Wilson E O 1980 *Sociobiology*. Belknap Press of Harvard University Press, Cambridge, MA
- Wood D J, Bruner J S, Ross G 1976 The role of tutoring in problem solving. *Journal of Child Psychology and Psychiatry* 17: 89–100

J. Heckhausen & T. Singer

Play and Development in Children

1. Introduction

The study of play has held particular allure to social and behavioral scientists, at least since Groos (1898). Since that time it has been assumed that play serves an important role in children's development into adults. An important step in documenting the importance of play should be defining play in terms of the ways in which it is different from forms of behavior not considered playful.

2. What is Play?

No one definition of play is necessary or sufficient, yet child psychologists (Pellegrini and Smith 1998) and ethologists (Martin and Caro 1985) agree upon some

common definitional criteria: play does not seem to serve any apparent immediate purpose. This assumes that children are less concerned with the outcome of their behavior than with the behavioral processes *per se*.

Play has also been defined according to phenomena that precede (e.g., an atmosphere that is familiar, safe, and friendly, a minimally intrusive adult, and children who are free from stress) and succeed play behaviors (e.g., behavior is categorized as play fighting, not aggression, if children stay together after the conclusion of the bout; if they separate, it is defined as aggression). Caution is needed as in this case as antecedents and consequences of behaviors can be considered as elicitors and outcomes, respectively, of those behaviors, rather than as components of the behavior *per se*.

3. Forms of Play

Different forms of play occur from infancy through adolescence and they have distinct antecedents and possible functions.

3.1 Exploration

Exploration is an information-gathering venture (e.g., mouthing and simple manipulation of objects) (Belsky and Most 1983). Through exploration children come to know their environments.

Exploration dominates behavior for the first 9 months of life. By 12 months, play and exploration co-occur. By 18 months play accounts for more of the child's interactions with the environment than does exploration (Belsky and Most 1983). Exploration also precedes play to the extent that children of all ages must explore an object, or know its properties, before they can play with it. The knowledge resulting from exploration is basis for play.

3.2 Fantasy Play

Fantasy, or pretense, is the paradigm case of play during childhood. Fantasy involves an 'as if...' orientation to the world and involves actions, use of objects, and verbalizations and nonliteral meanings; often, it involves playing a distinct pretend role such as mummy. Fantasy begins during the second year of life, peaks during the late preschool years, and then declines (Fein 1981) and there are reliable gender differences: girls engage in fantasy play both more frequently and at more sophisticated levels than do boys.

Fantasy play is influenced by familial variables as well as play props and play partners. Children who are securely, compared with insecurely, attached to their

mothers engage in more sophisticated pretend play (Roggman and Langlois 1987). Regarding props, play themes generally follow the themes inherent in the props available (Pellegrini and Perlmutter 1989). Further, children's pretense is more sustained and complex when they are playing with friends, compared with acquaintances, possibly because the mutuality and emotional commitment of friends may motivate children to sustain cooperation (Howes 1992).

3.3 Locomotor Play

Locomotor play is play that is physically vigorous, e.g., chasing and rough-and-tumble play (RandT). The less 'thematic' and nonsocial forms of vigorous play, such as running and jumping, account for close to 20 percent of the activities of 3- and 4-year-olds and then declines during primary school years (Pellegrini and Smith 1998).

Vigorous play is shaped by both hormonal and social events, resulting in reliable gender differences (Pellegrini and Smith 1998). During the course of normal fetal development, androgens predispose males to be more physically active than females. Socialization practices, such as father engaging in vigorous play with sons more frequently than with daughters, reinforce the difference.

4. Functions of Play

Paradoxically, play is defined in terms of its 'purposelessness' yet seen as serving an important function in children's development. The paradox can be reconciled by supposing that the apparent lack of immediate function actually conceals a function delayed in development.

4.1 What is Meant by Function?

Function has at least two meanings in the social and behavioral sciences. First, function can be defined in terms of an evolutionary history. A behavior is functional if it has added to the reproductive success of individuals. Functions can also be defined in terms of beneficial consequences during the lifespan of any particular individual. Benefits may be accrued immediately or during the period of childhood, or deferred until adulthood.

4.2 Immediate or Deferred Functions

Most theories of play have proposed that the benefits of play are not immediate to the period of childhood but deferred until later in development. In Bateson's (1976) terms this view presents play as a 'scaffolding'

metaphor: play functions in the assembly of a skill and then is disassembled when the skill is complete. This remains an influential view, even though empirical studies have not provided support for this claim (e.g., Martin and Caro 1985, Pellegrini and Smith 1998).

Alternatively, play may not be an imperfect version of adult behavior but may be adapted and beneficial to the specific niche of childhood. Bateson's (1976) metaphor for the immediate functions of play is metamorphic (e.g., the sense of mastery associated with children's play probably relates to children experimenting with new activities).

5. Making Functional Inferences

Function can be inferred using cost-benefit analyses, consideration of design and contextual features, and experimental enhancement or deprivation of play components.

5.1 Cost-Benefit Analyses

With cost-benefit analyses we assume a correspondence between high costs and high benefits; low costs imply low or high benefits. Play can be costly in terms of time, caloric expenditure, and survivorship (e.g., danger). Basic cost data on children's play is, however, scanty. Yet there are some hints from the animal play research which suggest that play accounts for a limited portion of an animal's time and energy budgets—around 10 percent (Martin and Caro 1985, Pellegrini and Smith 1998). Consequently, we conclude that benefits are probably low and deferred as immediate benefits are more likely to be reaped given risks associated with survivorship.

5.2 Design Features

A design-features argument involves explication of similarities between aspects of play behavior and similar features in mature behavior. Design feature arguments have been made most frequently for RandT as its fighting-like features (e.g., wrestling) should relate to fighting skills and dominance. Problems associated with this design-features approach to the study of function relate to the fact that a mature behavior may develop from many different antecedents (Martin and Caro 1985).

5.3 Experimental Deprivation Studies

A few experimental play-deprivation studies have been conducted where juveniles are deprived of an aspect of play. The assumption is that deprivation of an aspect of play, if it is developmentally important, should

result in a 'rebound effect' when children are given the opportunity to play again; children compensate for the deprivation by engaging in high levels and longer durations of play. Most deprivation studies can be criticized on the grounds that more than one thing is involved when we deprive children of play (e.g., depriving children of social play simultaneously deprives them of social interaction).

The effects of depriving children of play opportunities have largely been limited to studies of physical play. In different field experiments, Smith and Hagan (1980) and Pellegrini et al. (1998) experimentally deprived children of opportunities for physical play. Across all experiments the results showed that increased deprivation led to increased levels of play when opportunities for play were resumed. We argued that children rebounded after deprivation so as to overcompensate for lost opportunities for physical exercise, which serves important training functions.

5.4 Experimental Enrichment Studies

With enrichment procedures, children are given opportunities to engage in play that vary according to theoretically relevant dimensions, such as conceptual conflict (Pellegrini 1984), and their effects are measured. Most of the criterion measures in these studies are taken in close temporal proximity to the training, thus assuming that play has immediate rather than deferred benefits.

There have been numerous studies examining the effects of and pretend play enrichment on various dimensions of children's social cognitive status, such as role taking. Most of these studies showed positive benefit effects of play; but they must be interpreted very cautiously, primarily because many play enrichment experiments have serious internal validity problems (Smith 1988). For example, many experiments confound play treatment with adults tutoring children in play and most effects are due to experimenter bias.

6. Conclusion

We have outlined some of the basic definitional and developmental tenets of play. The role of play in children's development is still a controversial and unresolved topic. Cost-benefit analyses suggest that there should be benefits; but they may be relatively small and possibly limited to the period of childhood, rather than deferred until after childhood. Suggested benefits include physical fitness (for physical play), fighting and dominance skills (for RandT), and affiliative and theory of mind skills (social and pretend play).

With this said, however, we note that much more research is needed before we can make sound conclusions regarding the developmental functions of play. Most basically, more descriptive studies of

children's play, documenting costs in terms of time, caloric expenditure, and survivorship in contexts other than school are needed. This information could be used in the service of cost-benefit analytic studies of play. Correspondingly, these features could also be used to test functional hypotheses more directly using field experimental methods.

See also: Children's Play: Educational Function; Infancy and Childhood: Emotional Development; Piaget's Theory of Child Development; Play, Anthropology of

Bibliography

- Bateson P P G 1976 Rules and reciprocity in behavioral development. In: Bateson P P G, Hinde R A (eds.) *Growing Points in Ethology*. Cambridge University Press, Cambridge, UK, pp. 401–21
- Belsky J, Most R K 1981 From exploitation to play. *Developmental Psychology* **17**: 630–9
- Fein G 1981 Pretend play: an integrative review. *Child Development* **52**: 1095–118
- Groos K 1898 *The Play of Animals*. D. Appleton, New York
- Howes C 1992 *The Collaborative Construction of Pretend*. State University of New York Press, Albany, NY
- Martin P, Caro T M 1985 On the functions of play and its role in behavioral development. *Advances in the Study of Behavior* **15**: 59–103
- Pellegrini A D 1984 Identifying causal elements in the thematic-fantasy play paradigm. *American Educational Research Journal* **21**: 691–703
- Pellegrini A D, Horvat M, Huberty P 1998 The relative costs of children's physical activity play. *Animal Behavior* **55**: 1053–61
- Pellegrini A D, Perlmuter J C 1989 Classroom contextual effects on children's play. *Developmental Psychology* **25**: 289–96
- Pellegrini A D, Smith P K 1998 Physical activity play: the nature and function of a neglected aspect of play. *Child Development* **69**: 577–98
- Roggman L A, Langlois J H, Hubbstait L 1987 Mothers, infants, and toys: social play correlates of attachment. *Infant Behavior and Development* **10**: 233–37
- Smith P K 1982 Does play matter? Functional and evolutionary aspects of animal and human play. *The Behavioral and Brain Sciences* **5**: 139–55
- Smith P K 1988 Children's play and its role in early development: a re-evaluation of the 'play ethos.' In: Pellegrini A D (ed.) *Psychological Bases for Early Education*. Wiley, Chichester, UK, pp. 207–26
- Smith P K, Hagan T 1980 Effects of deprivation on exercise play in nursery school children. *Animal Behavior* **28**: 922–8

A. D. Pellegrini and P. K. Smith

Play, Anthropology of

Malleable, flexible, fluid, uncertain, changeable—all are adjectives that are used to describe play, whether as a medium of cognition, practice, or communication. Play refers to a complex of universal phenomena

among human beings that also are found among numerous animal species (Fagen 1981). All peoples, though they may not distinguish in semantic terms between play and other behavior, likely engage in playing as they shift into a conditional or subjunctive sense of being, one that opens toward a multiplicity of possibility. The mathematician and the theoretical physicist play during the process of shaping their ideas no less than children imagine, practice, and repeat their inventions in their play. Useful overviews of play phenomena are given in Spariosu (1989), Schwartzman (1978), and Sutton-Smith (1997).

Despite its pervasive presence, play is perhaps the least concrete of phenomena and therefore one of the most difficult to define. Any useful definition of play highlights its multiplicity of potential experiences. The anthropologist, Bateson (1972) perceives play as paradoxical, in that it is and is not what it seems to be. The classicist Spariosu (1989) terms play, amphibolous, or going in two directions at once. In a succinct manner, the anthropologist, Droogers (1996) defines play as, 'the capacity to deal simultaneously and subjunctively with two or more ways of classifying reality.'

1. Play and Reality

There is an intimate relationship between play and reality because play is first and foremost experiential, whether personal or social, and reality does not exist without its experiencing. In the creation and practice of realities of play, the imagination is crucial. Without the cognitive, the capacity to imagine realities, there is no play, just as, without the practice of experience there is no reality.

Play raises to the fore issues of alternative realities, the role of these and other multiplicities in social and personal life, and their consequences for human cognition and practice. The cognitive perceptions of play enable persons to enter into alternative realities, and perhaps into the simultaneity of multiplicity. Cognition is related intrinsically to how realities of play are imagined and framed differently from other realities. The practice of play entails the sensuous embodiment of these alternative realities; and, so, the subjective experiencing of play from within these imagined contexts that offer a sense of release, freedom, exploration, and perhaps exhilaration. The communicative component suggests that these specialized contexts of play are able to comment on the social orders from which they arise, and to send these messages back as the participants in play return to the reality of nonplay. Therefore contexts of play are metacommunicative in their relationships to social order: play communicates to its participants not only *how* to play but how such playing relates to social order. That play is framed as imaginative and malleable, and that through these qualities it generates and conveys metamessages informing players as to

how to relate to aspects of social order, makes play perhaps the most fruitful of contexts of socialization, both for children (Briggs 1998) and for adults.

2. Framing Play

Given that play constitutes domains that cognize, practice, and communicate the shaping of imagination, are these domains to be conceptualized as incisively distinct from domains of serious reality, of not-play? This is the implication of Bateson's highly influential theory of the framing of play which separates play lineally from the serious, or not-play. According to Bateson, the cognitive metamessage, 'This is play,' brings into existence the play frame which cleaves between the seriousness of everyday reality and the imaginings which are practiced within the frame. One may adduce from this not only that play comes into existence as a corollary of its framing but also that play, and the practice of its imaginative multiplicities of possibility, is located within the restrictions of its framing. Viewed from outside the frame, play is make-believe activity that in fact does not exist. Experienced from within the frame, play may well be utterly engrossing in the embracing reality of its presence. These views accord well with Johann Huizinga's (1970) classic study of how civilization emerged through the plastic premises of play; although in the process, play itself became a temporary sphere of make-believe activity, secluded, limited, and bounded in time and space. Once play is framed through lineal cognition, it becomes a world unto itself, one of temporary limited perfection with its own internal teleologies (Gadamer 1988). Within itself the practice of play may well proceed in deadly earnest.

2.1 Play and the Playful

Sutton-Smith (1997) offers a useful modification to the binarism of a reality either of play or of not-play, of make-believe or of serious reality. Sutton-Smith distinguishes between play and the playful. To build on his argument that the playful plays with the very framing of play itself, one can add that the playful is a mood, an attitude, that may permeate both serious reality and play, or that suddenly may surge into presence within the mundane. In either instance, the playful is full of the impulse to perceive and feel in ways other than those offered by the immediacy of a given reality. The abrupt presence of playfulness disrupts the most routine of expectations. Playfulness injects speedy uncertainty into the most expected of social practices.

The playful attitude within mundane reality may be integral to the existence and organization of the latter; while play itself may be more 'distant,' as it were, from mundane reality. Playfulness and play complement

one another. The eruption of playful feeling and perception are a significant key to establishing the play frame in relationship to mundane reality. But the play frame itself directs the kinds of ludic or make-believe activities that take shape within it, relatively separated from mundane reality.

2.2 Play and Foundation-for-form

In order for a play frame to come into existence there seems to be a need for a form of reference, a foundation-for-form or organic model that can be altered in systematic ways and turned into someone or something else. Nonetheless, whatever is changed retains crucial similarities to the form of its foundation and so remains intimately related to it. For example, The medieval European Feast of Fools which mocked the traditional Christian Mass needed the latter so that this could be inverted and altered. The play-mass, performed in church by priests, would have had no significance for participants were it not derived from and contrasted with its everyday analogue, the sacred Mass.

To take a more mundane example, when one little American girl, aged seven, turns to her sister, aged five, and says: 'Let's play sisters,' she is establishing a play frame within which these girls can play the foundational form of their sisterhood, imaginatively and perhaps with novelty. Play changes the known signs of form into something else by altering the reified boundaries that define and characterize the phenomenon. These girls will still be sisters, but they will be able to explore, practice, and experience their sisterhood with a difference, with license, in ways perhaps closed or unimagined within their usual relationship. They can play with the strangeness of being sisters. In turn, they may carry the messages of their experience back into their ongoing relationship as sisters, changing this in minute but additive ways. Hypothetically, they playfully may enter into play-sisterhood whenever they desire.

3. Attributes of Play

Play is a way of organizing activity, rather than any particular set or content of activities (Miller 1973). Any activity may be altered playfully, and played at. But in order to do this the everyday, instrumental perception of how means lead to and produce ends must be loosened, stretched, or perhaps uncoupled entirely. Goals exist in play, but the means to attain them are given degrees of autonomy. This is why play behavior can be endlessly repetitive or follow a crooked line around artificial obstacles. The patterning for play may be suggested by how the foundations-for-form that are used in playing function in the everyday. Say that the two little sisters and their friends place chairs in a row, the back of each facing

the seat of the next, and that these children imagine these chairs to be a train. The children practice playing train, until they decide that the train is a plane, and the row sprouts wings, carrying its passengers, but then is turned into a bird flying off with children on its back, becoming a row of chairs in school and kindergarten that is also a plane. In each instance, one foundation for form (train, plane, bird) is imagined to take the place of another. Each such organic model is voluntarily turned into a foundation for imaginative practice that plays in turn with the multiplicity of these organic models, one after the other or as hybrids of one another (a row of chairs that with near simultaneity is train, plane, bird).

The perception and practice of play brings into being something that had not existed before in quite the same way. This is done by changing the shape and positioning of boundaries that categorize phenomena, thereby altering their meaning. Each shift by sisters into playing sisters seems like a repetition, yet is instead a different beginning since it imagines anew and plays with the very possibilities of sisterhood. Even a slight change in the boundaries that give shape and significance to phenomena involves some degree of destruction and recreation. In its cognition and feeling, play is associated closely with creativity, creation, and destruction.

Play is frequently implicated in the making of change. Change is the making of difference, even if this difference seems to mimic or eventually to reproduce an earlier state. Play introduces malleability into social and cultural life, taking apart the given and perhaps offering alternative visions and practices. Play often is involved in learning something new, through open-ended repetition, variation, exaggeration, and experimentation. Play is not unusual in the liminal phases of initiation rituals (Turner 1969), healing rituals (Kapferer 1983), and mortuary rituals of tribal peoples, when initiates, the ill, or the dead undergo profound transformations of being. Playfulness, in a more postmodern idiom is central to the practice of certain neo-shamanic rituals whose goal is the transformation of consciousness (Lindquist 1997). These and many other changing conditions involve shifts of cognition, practice, and communication. More significantly, change (creative or reproductive) usually involves fragmentation, destruction, or transformation of some sort, whether temporary or permanent, and these processes are often violent. The violence of change and change through play are often closely related. Some form of destruction and violence (to cognition, feeling, body and being) is involved in the shattering suddenness of the playful and in play.

4. The Truth-values of Play

The relationship between play and violence relates to how play has been conceived, and to the truth-values accorded to play. In his survey of philosophical ideas

of play, Spariosu (1989) finds two trends, one that stresses the practice of play as prerational (with its roots in ancient Greek thought and poetry), and the other that emphasizes play as rational activity. In prerational thinking, play is full of violent agon and power, strongly and pleasurably emotional, arbitrary in its appearance, experienced perhaps as ecstatic, and opening toward unlimited freedom. In rational thinking, the naked, enjoyable forces of chaos and disorder, of irrational power and violence, are suppressed in favor of play as rationally framed activity that is set apart from serious reality. Here play is unreal in its irrationality, and therefore controlled, nonviolent, and limited in its freedom. To expand on Spariosu, one may add that the eruption of pre-rational play (akin to playfulness) into everyday life is full of its own unsuppressed truth-value. The truth-value of prerational play is integral to living within realities that overlap and that are continuous with one another—realities that relate to one another as multiplicities of possibility. By contrast, the make-believe status of rational play, framed and contained in special times and places (as Huizinga, Bateson, and others argue) denies any direct or immediate truth-value to play, enabling its practice to be treated as fictive, as ‘just play,’ ‘mere play,’ ‘done only in play,’ and so forth.

4.1 Play and Cosmology

The organization of cosmology demonstrates from another perspective the importance of the truth-values attributed to play. For example, in Indian cosmogenesis, play (or more accurately, playfulness) is implicated in the very creation of cosmos. Thus, varieties of Hindu cosmology and religion embed premises of play—malleability, multiplicity, ongoing change—at a very high level of cosmic organization. In consequence, these self-same premises of play resonate throughout the levels of cosmos. The entirety of cosmos is influenced by premises of play. In this kind of cosmology, play is a top-down idea, responsible in no small degree for the multiplicities of reality that constitute Hindu cosmos and religions (Handelman 1992, Handelman and Shulman 1997).

Here play has high degrees of truth-value and is rarely referred to as make-believe. By contrast, mainstream monotheistic cosmologies, if they pay any attention at all to ideas of play, embed these at a low level of cosmic organization, distant from cosmogenesis and the ordering of religion and society. In monotheistic cosmologies, play is more of a bottom-up idea, lineally framed, rationally contained, and separated from serious reality. Bottom-up play enters cosmology and religion from their conceptual and emotional undergrowth, and therefore is perceived as threatening. Here the messages of play are often positioned in opposition to or as a negation of the order of things; and play is perceived as fantasy,

illusion, and ephemeral; yet also as subversive and as resisting the order of things. But here the practice of play has low truth-value. Thus, the truth-value of top-down play is openly full of power and violence, while these qualities are veiled, hidden, and denatured in the low truth-values accorded to bottom-up play.

5. Game and Play

The idea and practice of game is closely related to that of play and may be understood as a variation of the latter, in the following senses. Game, like play, is a way of organizing activity. Game, too, loosens the relationship between means and ends. But the existence of game depends on relatively enduring rules, rather than on foundations-for-form that are easily changeable. Therefore games constitute distinct, autonomous realities, virtualities, given their internal, rule-based organization which to a high degree determines how they are practiced. Playfulness has less of a role in the practice of games. Games are less malleable and harder to alter than play.

Yet it is precisely because of their ruledness that games can be organized to make directed change through their practice. Lévi-Strauss (1966) comments that game is a systematic way of creating inequality. Game begins with equality between its sides and produces differences which are perceived as significant. The logics of game, in fact, are often close to those of ritual which is intended to generate change through its operations. Traditional Mesoamerican ball-games had game-like aspects, yet they were ritual-like in their relationship between means and ends, perhaps ends of cosmic renewal that their practice was intended to accomplish. More generally, the presence of play and game within rites seem to contribute to the attainment of ritual goals. (This is powerfully exemplified in the roles of clowns in Native-American Pueblo rituals.)

Perhaps in more systemic ways than play, game may awaken interpretive, reflexive feelings about aspects of social and moral order. The highly agonistic Native-American Kwakiutl potlatch, a ritual full of conflict that both demonstrated and reordered social status, produced the ‘play potlatch’ (Codere 1956) in which participants laughed at their own strivings for power and hierarchy. The rigid gender divisions in traditional Iranian society produced games practiced by women in which these distinctions were called into question, mocked, and played with (Safa-Isfahani 1980).

The twentieth century has seen tremendous growth in the use of game as a metaphor for living and social order (Minnema 1998)—Wittgenstein’s language games, Von Neumann’s game theory, war games, educational games, the game of intelligence testing, games people play, the play of quantum mechanics, chaos theory, virtual realities, and so forth. In post-modernist terms, these visions are of life organized as interpenetrating games, none of which give any total

or complete vision of reality, and no certainty about the stability of forms of living. The game metaphor reflects more fluid and mobile social orders that are perceived, correctly or incorrectly, as the creations of individual agency. Television as spectacle has contributed greatly to this growth by telecasting existing competitive game forms and aiding in the invention of new ones. Sports are at the front of these developments. These multitudes of contest (in which violence is close to the social surface if not on it) continuously rank and re-rank participants over and again in hierarchies that are more kaleidoscopic, fragile, and ephemeral than they are enduring. Play as the multiplicities of possibility, and the playful as the eruption of these possibilities in the everyday, are competing with more unitary monotheistic ethos, at least in the visions of the middle-class Western world.

Current thinking on play is reconceptualizing the problem of framing—the problem of relationships among realities, social and cognitive—to open alternatives to the ongoing rupture between play as rational and as irrational. Research on Internet chat groups plays into questions of framing (Danet et al. 1998). Issues of power in and through the practice of play are also related to questions of framing. Thus, ideas of ‘dark play’ (Schechner 1988), when players do not know that they are playing, and ‘evil play,’ when the manipulator of the frame positions herself outside it and plays others, yet pretends to be within the frame and controlled by it, relate to issues of play and power. They echo Bateson’s still enduring thought that a more complex question to ask of a play phenomenon is whether it is indeed this, or whether it is pretending to be so. Imagining the pretense of play is even more powerful than play imagined as such.

See also: Game Theory; Leisure, Psychology of; Play and Development in Children; Sport, Anthropology of; Sport, Sociology of

Bibliography

- Bateson G 1972 *Steps to an Ecology of Mind*. Chandler, San Francisco
- Briggs J L 1998 *Inuit Morality Play: The Emotional Education of a Three-Year-Old*. Yale University Press, New Haven, CT
- Codere H 1956 The amiable side of Kwakiutl life: The potlatch and the play potlatch. *American Anthropologist* 58: 334–51
- Danet B, Ruedenberg L, Rosenbaum-Tamari Y 1998 Hmmm... where’s that smoke coming from? Writing, play and performance on internet relay chat. In: Sudweeks F, McLaughlin M, Rafaeli S (eds.) *Network and Netplay: Virtual Groups on the Internet*. AAAI Press, Menlo Park, CA and MIT Press, Cambridge, MA
- Droogers A 1996 Methodological ludism: beyond religionism and reductionism. In: van Harskamp A (ed.) *Conflicts in Social Science*. Routledge, London
- Fagen R 1981 *Animal Play Behavior*. Oxford University Press, New York
- Gadamer H-G 1988 *Truth and Method*. Crossroad, New York
- Handelman D 1992 Passages to play: Paradox and process. *Play & Culture* 5: 1–19
- Handelman D, Shulman D 1997 *God Inside Out: Siva’s Game of Dice*. Oxford University Press, New York
- Huizinga J 1970 *Homo Ludens: A Study of the Play Element in Culture*. Maurice Temple Smith, London
- Kapferer B 1983 *A Celebration of Demons: Exorcism and the Aesthetics of Healing in Sri Lanka*. Indiana University Press, Bloomington, IN
- Lévi-Strauss C 1966 *The Savage Mind*. Weidenfeld & Nicholson, London
- Lindquist G 1997 *Shamanic Performances on the Urban Scene: Neo-Shamanism in Contemporary Sweden* (Stockholm Studies in Social Anthropology 39). Almqvist & Wiksell International, Stockholm
- Miller S 1973 Ends, means, and galumphing: Some leitmotifs of play. *American Anthropologist* 75: 87–98
- Minnema L 1998 Play and (post)modern culture: An essay on changes in the scientific interest in the phenomenon of play. *Cultural Dynamics* 10: 21–47
- Safa-Isfahani K 1980 Female-centred world views in Iranian culture: Symbolic representations of sexuality in dramatic games. *Signs* 6: 33–53
- Schechner R 1988 Playing. *Play & Culture* 1: 3–19
- Schwartzman H B 1978 *Transformations: The Anthropology of Children’s Play*. Plenum, New York
- Spariosu M I 1989 *Dionysus Reborn: Play and the Aesthetic Dimension in Modern Philosophical and Scientific Discourse*. Cornell University Press, Ithaca, NY
- Sutton-Smith B 1997 *The Ambiguity of Play*. Harvard University Press, Cambridge, MA
- Turner V W 1969 *The Ritual Process*. Aldine, Chicago

D. Handelman

Plea Bargaining

Research into plea bargaining ‘in action’, not plea bargaining ‘on the books,’ has been conducted, for the most part, by studying US federal and state courts. Nonetheless, the central theme of this article is that the findings of this research are applicable in a comparative context and can, with some qualifications, help us to understand plea bargaining around the world. Plea bargaining may take different forms, and it might be referred to, or conceived of, as something different, but it does exist in criminal justice systems outside the USA. By examining US studies and the small amount of comparative research available, we can speculate that plea bargaining is a function of the raw material that courts process, and that if we permit an expanded definition of plea bargaining—even allowing for the particularities of various jurisdictions—we will see that this manner of case disposition is a near-universal province of courts. Space constraints preclude a systematic examination of what we know about these dynamics in a comparative sense; this article will discuss general themes and questions.

1. The Typology of Plea Bargaining

While scholars have concluded that plea bargaining was common in nineteenth century British courts (Zander 1998, pp. 323–34), most comprehensive studies have focused on the USA, and most have been conducted in the past 25 years. (For a summary of the early work on plea bargaining, as well as an agenda for future plea bargaining research at the time, see *Law and Society Review* 13 (2), 1979; for useful citations on subsequent related work see McCoy 1993, Vogel 1999.) In the early days, there was a debunking tone to this work, as scholars questioned and tested the assumption that trial courts ‘try’ cases. What was viewed as anomalous in an earlier period of US history—something which might have characterized the more political (and frequently ‘sleazy’) municipal courts—plea bargaining is now understood to be the central method of processing cases in trial courts. Plea bargaining was not an aberration of US courts; it was the norm. In what might be viewed as the epitaph for the role of trials in trial courts, a judge noted: ‘I don’t think that a trial necessarily represents a failure in the plea bargaining system. I really believe that the trial is a necessary adjunct of the system’ (Heumann 1977, p. 150).

While most scholarly attention has been focused on plea bargaining in the USA, recent comparative endeavors have produced data that afford us the opportunity to understand plea bargaining as it exists (in many forms) around the world. While no two nations’ legal systems or plea bargaining experiences are exactly alike, we are at the point where we can begin to construct typologies (explicit, policy, and implicit plea bargaining) that can and should invite future research. Notably, while different forms of plea bargaining will continue to exist, there is a kind of convergence at work, as the institutions, attitudes, and processes of nations around the world begin to resemble each other when it comes to the actual practice of plea bargaining.

1.1 Explicit Plea Bargaining

Plea bargaining in the USA in its most basic and commonly understood form, is the process of active, explicit negotiation between the defense attorney and prosecutor concerning the resolution of the defendant’s case. Most frequently the ‘deal’ revolves around the benefits that will flow to the defendant if he or she chooses to plead guilty. These might include a reduction in a charge, a dismissal of a charge or charges, or a prosecutorial sentence recommendation (or an agreement for the prosecutor not to press a particular sentence). Also, increasingly, the prosecutor might agree not to add a mandatory sentence charge to the defendant’s indictment. Any, or many, of these will be given to the defendant in return for a guilty plea. Generally the exchange is explicitly negotiated, and a

judge may or may not also be part of the negotiating process. When not involved, the judge usually goes along with the negotiations; however, if it is felt that the agreement is not acceptable to the judge, the defendant is usually allowed to withdraw the plea.

This form of plea bargaining, as indicated above, is one in which a negotiation is undertaken explicitly and the parameters of the agreement are agreed upon specifically. While the USA (especially since the late 1960s following the US Supreme Court’s approval of plea bargaining) represents the best example of a system that embraces explicit plea bargaining (indeed, the explicit variety is the one that is conjured up when comparisons with the USA are made), other nations which have experienced increases in case pressure have sought relief through increases in explicit plea bargaining. While it has been proven that this relationship is not a causal one in American courts (Heumann 1977), we can see a correlative relationship in systems such as South Africa’s, for example, which is considering the expansion of the realm of explicit bargaining (and inviting judges into the process) as a response to its increasingly overwhelmed courts (Schwikkard, and van der Merve 1999, p. 352). It is important to note that plea bargaining (involving prosecutors and defendants) was already common in South Africa, and thus was not caused by a case overload; what we are seeing now, in other words, is a development of earlier practice. Other examples of nations that employ more explicit practices include Israel, which allows on-the-record plea bargaining (Harnon and Stein 1999, pp. 239–41), and Canada, which openly encourages pleas with regard to charge and sentence (Roach 1999, p. 75).

1.2 ‘Policy’ Plea Bargaining

A second kind of plea bargaining, neither explicit nor implicit *per se*, is derived from, and relies upon, the policies a prosecutor’s office may have adopted over time. Specifically, the prosecutor may announce that his/her office will not negotiate in specific types of cases (e.g., when a firearm is involved, or certain drug matters). But the prosecutor, without explicit negotiation will, as a matter of policy, substitute some lower charge for defendants who plead guilty to a ‘policy case.’ Thus defendants (or defense attorneys) can be reasonably certain about the benefits that will attach to a plea of guilty. This form of plea bargaining seems to be unique to the USA—a product, seemingly, of the mandatory sentencing guidelines that have stripped some courtroom actors increasingly of the discretion necessary to make explicit deals.

1.3 Implicit Plea Bargaining

A third kind of plea bargaining—what we might label ‘implicit plea bargaining’—takes place whenever a

defendant pleads guilty, with the expectation that the plea will lead to a reward in terms of charge and/or sentence reduction. Unlike the first kind of plea bargaining there is no explicit negotiation; unlike the second type there is no specific policy reward for a plea. In this kind of plea bargaining, the defendant hopes that a plea—a gesture of cooperation or contrition—will yield a beneficial disposition of the case.

Plea bargaining in the USA generally took the ‘implicit’ form until the 1960s, and we can see that in Japan, for example, defendants will often expect a reward for their plea, especially as the national government continues its crackdown on organized crime. (For an interesting discussion of why ‘elite’ USA Federal judges preferred implicit plea bargaining, see Padgett 1990.) Implicit plea bargaining is, based on data we have on other countries, the most common form in the world, but it is also the most difficult to study.

Many nations deny that plea bargaining takes place, and some (the UK, for example) even make plea bargaining illegal. In the early 1970s, several official arms of the government actually tried to stop the publication of a social science study based on defendant interviews, a study which claimed that extensive coercive plea bargaining characterized the British system (Baldwin and McConville 1977). And though a subsequent study qualified the contention about the coerciveness of the process, it still maintained that a variant of what can fairly be called plea bargaining thrived in the UK (Seifer 1980, pp. 179–97).

Japan also takes a more negative attitude toward plea bargaining, declaring it to be ‘illegal’ (Johnson 2001, p. 2). Thus if one were to believe the ‘law on the books,’ one would have to conclude that Japan stands as a clear exception to the argument that plea bargaining is both central to criminal justice systems and quite common as a practice. Yet in one of the few empirical studies of Japanese law in action, Johnson demonstrates that, despite the formal law to the contrary, several modes of plea bargaining are common in the Japanese criminal justice system. An aversion to the ‘crassness’ of the American form of plea bargaining, coupled with the purported illegality of this activity, may lead to denials of plea bargaining’s existence. But these denials are simply wrong—plea bargaining, in various forms, is alive and well in Japanese courts.

We can see this version of plea bargaining at work in some cases in Japan, and in Argentina and Italy (Sanders 1966, pp. 335–9, Carrio and Garro 1999, p. 43, Van Cleave 1999, pp. 272–4). In Japan, abbreviated—very abbreviated—‘trials’ are really uncontested presentations of evidence, which frequently are called ‘slow pleas of guilty.’ Running through these alternatives and others is something like the notion of implicit plea bargaining, the idea that not contesting a claim (inquisitorial systems typically do

not admit a guilty plea) leads to an expectation of leniency in disposition. In France, to offer yet another example, defendants feel that cooperation clearly leads to sentence reductions (Frase 1999, p. 169).

One thing that tends to disguise plea bargaining in other nations (and to allow for official denials) is the existence of functional equivalents. These equivalents represent this form of negotiation in its most ‘implicit’ manner. In Germany, for example, the prosecutor can either waive cases if it seems there is no public interest in prosecuting them, or issue a ‘punitive order’ (more of an administrative than a criminal resolution of the case) when it appears that trial alternatives are not desirable. ‘The prosecutor’s offer of a conditional waiver as well as the punitive orders have been considered a German variant of plea bargaining.’ (Blankenburg 1996, pp. 284–5, see also Weigand 1999, pp. 208–9).

2. Trends in Plea Bargaining in the USA

Because plea bargaining can take place in a variety of ways—all of them by definition subtle and discrete—this manner of bargaining presents scholars with the greatest challenge; to perceive, isolate, and understand negotiations of this kind seems to require the kind of comprehensive comparative studies that have not yet been performed. The implications for defining plea bargaining in these three ways ought to be made clear. Explicit bargaining, for a long while, was viewed as illegitimate, with judges in court often concerned with demonstrating for the record that nothing had been promised in return for a plea (see, e.g., Blumberg 1967, pp. 88–94); so for much of the twentieth century, plea bargaining in the USA assumed either the ‘policy’ or ‘implicit’ form. The recognition of the centrality of explicit plea bargaining by the US Supreme Court, however, rendered this unnecessary, and enabled plea bargaining to take place more openly in the court (*Brady vs. US*, 397 US 742 (1970); *North Carolina vs. Aflord*, 400 US 25 (1970); and especially *Santobello vs. New York*, 404 US 257 (1971); see also *Bordenkircher vs. Hayes*, 434 US 357 (1978)). With the advent of mandatory sentencing guidelines, however, many jurisdictions around the country have been forced to return to the more subtle, implicit manner of bargaining—moving the USA closer to the ‘backroom’ style of bargaining that takes place around the world and exemplifying the ‘convergence’ discussed above.

Policy plea bargaining and implicit plea bargaining, on the other hand, have more subtle implications. Both allow jurisdictions—prosecutors in particular—to claim that they do not plea bargain, or that plea bargaining in a category of cases has been proscribed. The symbolic value of such claims should not be underestimated. The ‘bargaining’ image of plea negotiations makes the explicit process vulnerable to accusa-

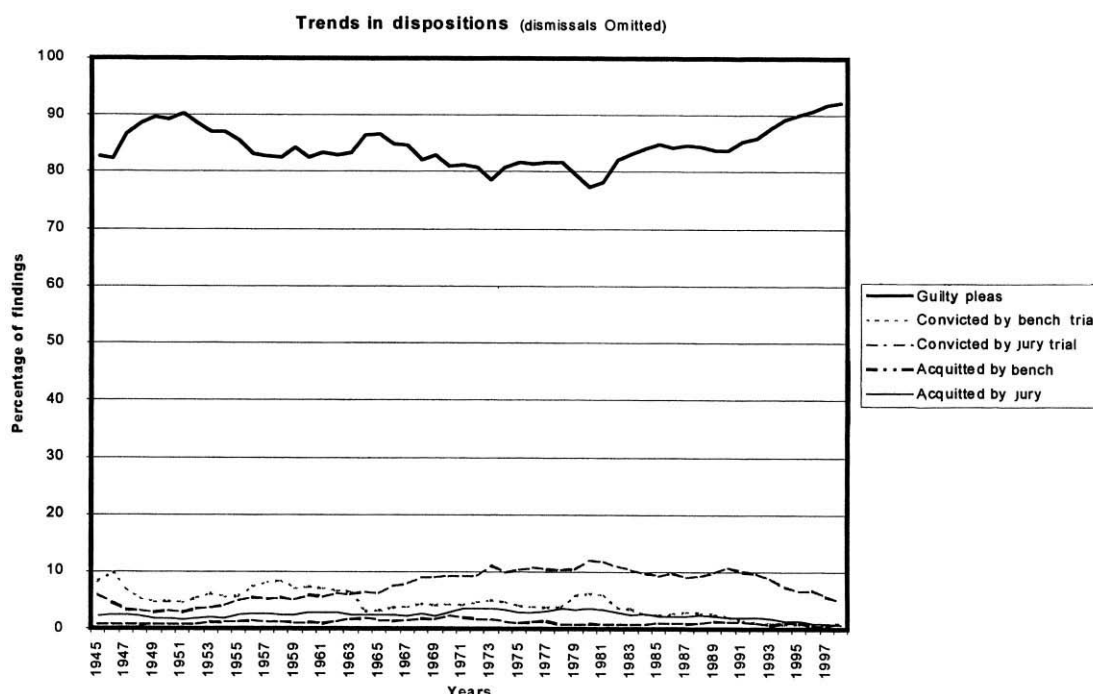


Figure 1

Guilty pleas (80–90 percent) compared with other outcomes of cases brought before the US courts 1945–97. Source: US Department of Justice, Bureau of Statistics

tions of ‘bargaining with justice.’ This, one might speculate, is why so few nations have engaged in the explicit form of plea bargaining. Even the USA, as we can posit from studies of the impact of sentencing guidelines, is beginning to move back toward the policy and implicit varieties of bargaining. (To appear to be in conformity with a ‘no plea bargaining’ policy, a court system may also make increased use of bench trials. These are often very abbreviated trials in which the defendant does not contest the case meaningfully, but nonetheless has an expectation of some benefit because he/she has waived a jury trial. See, for example, Heumann and Loftin 1979, pp. 417–21.) Yet if defendants still plead with the expectation of benefit, plea bargaining, less explicit, but plea bargaining nonetheless, can continue to flourish.

Some illustrative data on guilty plea rates are instructive. Figure 1 graphs guilty pleas, bench trials, and jury trials (convictions and acquittals) for US District Courts from 1945 through 1998 (the last year for which these data were available). The pattern is quite clear—roughly 80 percent of the defendants whose cases are not dismissed plead guilty. The very limited role of full trials in US courts is evident. Comparably, if we look at some recently reported data from three typical counties in a Midwestern state, guilty pleas (again for cases not dismissed) range from

a high of 97 percent to a ‘low’ of 87 percent (see Hannaford and Munsterman 1999, p. 66). One cannot help but be struck by the prevalence of guilty pleas in Federal and state courts in the USA. One would be hard pressed to find an American jurisdiction in which the guilty plea—or its functional equivalent (e.g., *pro forma* bench trials, or other forms of ‘slow pleas of guilty’)—does not comprise at least 70 percent of dispositions, with the usual percentage considerably higher. (Regarding these pleas in their many forms, see Mather 1979, pp. 55–6, 69–70, 78–9.)

As the 1945 data suggest, and other data demonstrate, plea bargaining is not a new phenomenon in US courts. Defendants pleading guilty for a putative reward and waiving the ‘Miranda rights’ (rights of the citizen when faced with police interrogation, such as the right to remain silent, right to legal counsel etc.) are the norm, not the exception in these courts, and this pattern has been consistent for some time (see, e.g., Heumann 1977, pp. 29–30, Padgett 1990, pp. 413–50, Zander 1998, pp. 323–4, Vogel 1999, pp. 161–246). Indeed, at least one scholar has argued that plea bargaining increased with the adversarial nature of the courts (Feeley 1997, p. 323).

Contrary to a popular belief, and indicated by its history and its prevalence in our courts at the time of writing, plea bargaining must reflect something other

than a response to caseload pressure. The common belief that plea bargaining is caused by case pressure has been disproved by studies of American courts, and, based on the research we have thus far, we can speculate reasonably that it would be disproved in a comparative context as well. While we are aware that plea bargaining takes place (in some form) in the nations that have been studied, we are still unsure as to *why* this manner of negotiation takes different forms in different nations. That is, why is it that the UK tends to practice and to view plea bargaining in one way, while Japan proceeds in another? The results and motives are often the same—prosecutors induce (or accept) a guilty plea and deliver (or end up providing) some kind of ‘reward.’ So what is it, specifically, that distinguishes nations and their experiences in a comparative context? Would we find the answer in a study of institutions? Culture? History? We both invite and eagerly await scholarship that concentrates on these important questions.

3. *Why Plea Bargaining?*

To develop a rationale for plea bargaining, we must begin with the defendant’s expectations/preferences. Though data may be unclear and variable depending on circumstances, most defendants believe that in most cases ‘it pays to plead guilty’ (Brereton and Casper 1982, pp. 45–70). Defendants often believe that their plea will be rewarded in terms of charge and/or sentence reductions. Why they think the system so rewards them varies with defendants, and seems to reflect the variety of justifications for such a reward advanced by prosecutors, judges, and defense attorneys. Some feel that the system appreciates the appearance of contrition that a plea represents—a theme frequently heard in court is that ‘pleading guilty is the first step to rehabilitation.’ Others feel that they are saving the state the time and expense of a trial, while others who provide information about associated crimes and criminals believe they should earn a reward for their contribution. Indeed, some Japanese scholars urge an increase in plea bargaining in order to obtain more information from defendants about organized crime (Johnson 2001, p. 35).

Finally, defendants feel that by pleading guilty they ensure the state some certainty—a certainty they themselves, in fact, enjoy more. In this sense defendants might be characterized as, almost naturally, in ‘plea bargaining mode,’ they seem to enter the system expecting a bargain. A study of the intentions and expectations of defendants, across nations and cultures, could shed some light on the internal dynamics of criminal justice systems around the world. Is ‘bargaining mode,’ for example, a product of a ‘bargaining’ culture—or one reputed to be interested in ‘dealing’—or do defendants naturally have this inclination without state encouragement? A study that con-

trolled for other nations’ reputations and cultures might yield answers to this question.

While defendant choices and expectations drive the system, complementary calculations by the system’s professional actors actually shape the defendants’ views. Prosecutors, as we know from studies of American courts, are influenced by saving money and time in many of their cases and (all things being equal) value defendant pleas. Also, to the prosecutor, even in the many open-and-shut cases (often known colloquially as cases which are ‘born dead’) that come through the office, there is always some uncertainty about post-trial outcomes; pleas remove any uncertainty, no matter how low the probability of acquittal (Mather 1979, pp. 41–2, 65–72. For the comparable concept of cases that are ‘born dead’ see Heumann 1977, p. 60). Without the data to predict with certainty, it seems reasonable to suggest a kind of ‘prosecutorial culture’ that might transcend national borders and even legal systems. That is, while we should be careful not to generalize too much, it seems reasonable to expect significant similarities in the attitudes (and certainly headaches) of prosecutors around the world.

Defense attorneys make similar assessments about minimizing pointless work and about ensuring certainty. Finally, judges also value the ostensible finality of pleas, and the reduction of the probability of appeal after a plea. Avoiding reversal on appeal, we know from the American experience, is one factor that is said to weigh large in the support that judges give to pleas. But does the same concern apply in other settings and legal traditions? Again, these actors (and others) are an active part of legal systems around the world. What then is the explanation for the different plea bargaining patterns we see in our initial comparative assessment?

While the above discussion outlines adequately what we know of plea bargaining up to the turn of the millennium, the standard explanations for plea bargaining ought not to obscure two very important considerations. The first we might call ‘all is not rosy in the plea bargaining world.’ The explanations we have given read as if defendants walk arm-in-arm with prosecutors, defense attorneys, and judges into the wonderful world of pleas. To some extent, many guilty defendants do take pleas, do so without coercion, and do so aware of the advantages of their choice. But all is not rosy, particularly for defendants who want to opt for trial. There is no question that in the US plea bargaining system there is generally a very strong inclination toward a plea; defendants who insist on trials are often in conflict with this clear reversal of the ‘presumption of innocence’ standard. Though we do not have good evidence for this, defendants may suffer harsher (i.e., harsher than the plea offer) dispositions upon conviction. Though these dispositions can be rationalized (there is no penalty for trial; pleas were rewarded because of the possibility that a trial might

lead to an acquittal; defendant not penitent; etc.), the specter of trials being viewed as unacceptable options in a plea bargaining world is distasteful.

Second, it is important to clarify the case-pressure/plea-bargaining argument. We have said that case pressure does not cause plea bargaining; both contemporary courts and those of the past with no realistically conceivable case pressure plea bargained most of their cases; crowded or uncrowded courts plea bargain for the reasons we have indicated. However, to say that plea bargaining is not caused by case pressure is not to say that case pressure may not affect plea bargaining; for many US courts, volume may be substantial and staffing less than optimal. In many courts, case pressure may affect dispositions, but we do not have a clear sense of how this might operate. Several possibilities come quickly to mind—rewards for pleas, which may exist in a ‘no pressure’ system, might become even greater; the reward for a plea may stay the same, but the ‘penalty’ for trial may be increased; the prosecutor may choose to screen out more cases through dismissals.

Case pressure, while not the cause of plea bargaining, can still affect the nature and form of the negotiation entailed, and we need more comprehensive comparative data on the effects of this pressure in other national settings. Recognizing that plea bargaining is fundamentally not a function of case pressure helps us to understand the centrality of pleas in US courts; acknowledging that case pressure can affect court disposition patterns helps us to understand why so many assume (incorrectly) that plea bargaining was adopted to manage heavy caseloads. We can speculate as to the relationship between plea bargaining and case pressure in other nations, but without data that considers the multiple variables involved, we are limited to our conjecture.

4. External Influences on Plea Bargaining: The USA as a Case Study

The single most dramatic change in American criminal court processes in recent years has been the increased use of mandatory sentences in particular, and adoption of sentencing guidelines in general (see *Punishment: Social and Legal Aspects*). In an effort to reduce judicial discretion in sentencing, the US Congress and individual states have circumscribed this discretion by requiring mandatory minimum sentences in certain cases, and generally requiring sentences within a narrow sentencing grid for defendants with particular background characteristics (e.g., number of prior convictions) who are convicted of particular crimes (see Tonry 1996, pp. 6–12, DuBow 1998, pp. 231–45). However, in limiting the open exercise of judicial discretion in the interests of reducing disparity in outcomes, the effect has actually been to increase (but mask) the discretion of prosecutors, while confining

these activities to the ‘backroom.’ By constraining judicial discretion, prosecutorial discretion has been enhanced; restricting sentencing ranges for judges has promoted less visible, but equally consequential, sentencing practices by prosecutors.

The significance of increased prosecutorial discretion in the USA, as suggested above, seems to be that—among other things—American plea bargaining is beginning to appear more like the negotiations that take place in other nations. This ‘convergence’ as we might call it, has tempered both the desire and capacity for actors in the American system to engage in explicit plea bargaining; ‘deals,’ therefore, increasingly must be of the ‘implicit’ or ‘policy’ variety. While this makes plea bargaining more difficult to study—relegated, as it is, to the literal and metaphorical ‘backrooms’—it presents scholars with an interesting series of questions to consider.

5. Conclusion

As we move beyond the exploration of plea bargaining itself, a number of very important matters—e.g., the relationship of plea bargaining to sentencing reform and to court caseloads, the effects of culture, traditions, and institutions on plea bargaining activities, and the appearance of a convergence of plea bargaining practices, move to the fore. These are matters that should be high on our research agendas as we begin to fashion a new, better informed, and less naive set of plea bargaining questions for the twenty-first century.

See also: Courtroom Consultation: Psychology; Courts and Adjudication; Criminal Defenses; Criminal Justice, Sociology of; Criminal Law and Crime Policy; Due Process in Law; Negotiation and Bargaining: Role of Lawyers; Punishment: Social and Legal Aspects; Trials: Cultural

Bibliography

- Baldwin J, McConville M 1977 *Negotiated Justice: Pressure to Plead Guilty*. Martin Robertson, London
- Blankenburg E 1996 Changes in political regimes and continuity of the rule of law in Germany. In: Jacob F, Blankenburg E, Kritzer H, Provine D M, Sanders J (eds.) *Courts, Law, and Politics in Comparative Perspective*. Yale University Press, New Haven, CT, pp. 249–314
- Blumberg A 1967 *Criminal Justice*. Quadrangle Books, Chicago
- Bradley C (ed.) 1999 *Criminal Procedure: A Worldwide Survey*. Carolina University Press, Durham, NC
- Brereton D, Casper J 1982 Does it pay to plead guilty? *Law and Society Review* 16: 45–70
- Carrio A, Garro A 1999 Argentina. In: Bradley C (ed.) *Criminal Procedure: A Worldwide Survey*. Carolina University Press, Durham, NC, pp. 3–52
- DuBow S S 1998 Introduction to models for sentencing reform in the United States and effects on the sanctioning process. *Law and Policy* 20(3): 231–45

- Feeley M 1997 Legal complexity and the transformation of the criminal process: The origins of plea bargaining. *Israel Law Review* 31:(183)
- Emmelman D 1996 Trial by plea bargain: Case settlement as a product of recursive decisionmaking. *Law and Society Review* 30(2)
- Frase R 1999 France. Bradley C (ed.) 1999 *Criminal Procedure: A Worldwide Survey*. Carolina University Press, Durham, NC, pp. 143–85
- Goldstein A, Marcus M 1977 The myth of judicial supervision in three inquisitorial systems: France, Italy, and Germany. *Yale Law Journal* 87: 240–83
- Hannaford P, Hans V, Munsterman T 1999 How much justice hangs in the balance. *Judicature* 83(2): 66
- Harnon E, Stein A 1999 Israel. In: Bradley C (ed.) *Criminal Procedure: A Worldwide Survey*. Carolina University Press, Durham, NC, pp. 217–44
- Heumann M 1977 *Plea Bargaining: The Experiences of Prosecutors, Judges, and Defense Attorneys*. University of Chicago Press, Chicago
- Heumann M, Loftin C 1979 Mandatory sentencing and the abolition of plea bargaining: The Michigan felony firearm statute. *Law and Society Review* 13(2): 417–21
- Jacob F et al. (eds.) 1996 *Courts, Law, and Politics in Comparative Perspective*. Yale University Press, New Haven, CT
- Johnson D 2001 Plea bargaining in Japan. In: Feeley M, Miyazawa S (eds.) *Sho Sato Collection*. Macmillan, London
- McCoy C 1993 *Politics and Plea Bargaining*. University of Pennsylvania Press, Philadelphia, PA
- Mather L 1979 *Plea Bargaining or Trial?* Lexington Books, Lexington, MA
- Murray J, Rau A, Sherman E (eds.) 1998 *Process of Dispute Resolution: The Role of Lawyers*, 2nd edn. Foundation Press, Westbury, NY
- Padgett J 1990 Plea bargaining and prohibition in the Federal courts, 1908–1934. *Law and Society Review* 24(2): 413–50
- Provine M 1996 Courts in the political process in France. In: Jacob F et al. (eds.) 1996 *Courts, Law, and Politics in Comparative Perspective*. Yale University Press, New Haven, CT, pp. 177–248
- Roach K 1999 Canada. In: Bradley C (ed.) *Criminal Procedure: A Worldwide Survey*. Carolina University Press, Durham, NC, pp. 53–80
- Sanders J 1996 Courts and law in Japan. In: Jacob F, Blankenburg E, Kritzer H, Provine D M, Sanders J (eds.) *Courts, Law, and Politics in Comparative Perspective*. Yale University Press, New Haven, CT, pp. 315–88
- Schwikkard P, van der Merve S 1999 South Africa. In: Bradley C (ed.) *Criminal Procedure: A Worldwide Survey*. Carolina University Press, Durham, North Carolina, pp. 319–60
- Seifer R 1980 Plea bargaining in England. In: McDonald W, Cramer J (eds.) *Plea Bargaining*. Lexington Books, Lexington, MA, pp. 179–97
- Stith K, Cabranes J 1998 *Fear of Judging*. University of Chicago Press, Chicago
- Tonry M 1995 *Malign Neglect—Race, Crime, and Punishment in America*. Oxford University Press, New York
- Tonry M 1996 *Sentencing Matters*. Oxford University Press, New York
- US Department of Justice 2000 *Sourcebook of Criminal Justice Statistics 1998*. Bureau of Justice Statistics, Albany, NY
- Van Cleave R 1999 Italy. In: Bradley C (ed.) *Criminal Procedure: A Worldwide Survey*. Carolina University Press, Durham, NC, pp. 245–82
- Vogel M 1999 The social origins of plea bargaining: Conflict and the law in the process of state formation, 1830–1860. *Law and Society Review* 33(1): 237–43
- Weigand T 1999 Germany. In: Bradley C (ed.) 1999 *Criminal Procedure: A Worldwide Survey*. Carolina University Press, Durham, NC, pp. 187–216
- Zander M 1998 Plea bargaining goes back a hundred years. *New Law Journal* 148(6830): 323–24

M. Heumann

Plural Societies

The idea of the ‘plural society’ was first coined by a British colonial administrator, Furnivall, on the basis of practical experience in Southeast Asia, in his classic volume, *Colonial Policy and Practice: A Comparative Study of Burma and Netherlands India* (1939), more as a descriptive device than as theory. It reflected the condition, common in colonial and later, newly independent postcolonial states, of deeply divided societies, usually along ascriptive lines of ethnicity or race (sometimes called ‘communal’), but also of culture, religion, or language. Characteristically, such societies are deeply segmented, with weak institutional integration and little consensus over ‘national’ agenda, unity, or identity. These tendencies are enhanced by the fact that the core values of each of the constituent (ethnic or other) communities are often incompatible, and that they each pursue separate goals and interests, in a fashion which Furnivall described as ‘mixing’ without ‘combining,’ and ‘meeting only in the market place,’ for basic economic needs. In the colonial context such societies were held together largely by coercion rather than consensus. Boeke, a Dutch colonial administrator in Furnivall’s tradition, developed similar ideas about segmented societies in his *Economics and Economic Policy of Dual Societies, Indonesia as Exemplified by* (1953), which he conceptualized in binary (‘dual’) terms, highlighting the colonial/indigenous cleavage, particularly in the domain of modern vs. traditional economic institutions.

One common source of confusion arises around two diametrically opposed meanings of the form ‘pluralism’ which may refer either to the *plural* society, or to its obverse, viz. a cohesive *pluralist* society integrated through a multiplicity of cross-cutting linkages, as conceptualized by such political scientists as (Dahrendorf 1959). This represents a continuation of the tradition of Alexis de Tocqueville, whose model of nineteenth-century America as knit together by webs of associations, helped to neutralize ethnic and other divisions.

Beginning in the 1960s, when most colonial states achieved independence, the term ‘plural society’ migrated into academic discourse, as anthropologists, sociologists, and political scientists attempted to inject

some theoretical rigor into the study of such weak states, held together as much by coercion as by the market. Among the pioneers were Smith and Kuper writing initially on (especially South) Africa and the West Indies. Together, they edited *Pluralism in Africa* (Kuper 1969), while Smith also authored *Stratification in Grenada* (Smith 1965b) and *The Plural Society in the British West Indies* (Smith 1965a). Smith and Kuper distinguished three types, or levels of pluralism. Cultural pluralism, as a form of extreme heterogeneity, sets the necessary, though not sufficient, preconditions for social pluralism, which is marked by the existence of relatively closed communities with separate core values and institutions and engaging in few cross-ethnic contacts. Finally, where constituent groups within a state are 'differentially incorporated,' or have unequal civil and legal status as citizens, there exists a situation of structural pluralism, a feature often found in racially divided societies. These writers, and to some extent Van den Berghe (1970), largely due to the particularities of the societies they studied, were most concerned with the type of pluralism known as ethnic stratification. Most common in 'white-settler' (hence, not strictly postcolonial) societies, social structure is envisaged in the form of a vertical hierarchy of ethnic/racial communities, which are simultaneously ranked as classes (occasionally called 'eth-classes'), the most egregious example of the era being apartheid South Africa.

In other world regions, such as Southeast Asia, another kind of plural society was identified, where the ethnic/cultural/communal sections, rather than being ranked hierarchically, live side-by-side as 'parallel,' 'pillared,' or 'segmental' communities (Benedict 1962, Morris 1967), each largely institutionally self-sufficient. Characteristically, individuals in such societies are primarily identified as members of their collectivity (or 'ethnicity'), which constrains them in most activities and life-chances, from religion, kinship practices, education, and occupation, to political party membership. In maximally plural situations (Despres 1968), there is little opportunity for autonomous individuals to act independently or to forge cross-community ties without reference to the 'basic givens' of their birth status. 'National' unity consists of little more than opportunistic alliances of multi-ethnic elites, while most other societal institutions and relations remain divided along ethnic lines. Societies of this type included Indonesia (Geertz 1963), Malaysia (Ratnam 1965, Nagata 1979), and Mauritius (Benedict 1962), where the presence of colonially-imported immigrant communities of Chinese and Indians exacerbated the cleavages and even conflict. In these places, and in such comparable plural states as Nigeria, Sri Lanka, Guyana, and Fiji, even national political elections are fought along communal lines, where the principal unity and co-operation emerges at the top, through cross-ethnic connections among elites for their mutual advantage. Political scientists describe

such electoral arrangements as 'consociational' (Lijphart 1977), and these are probably implicated in the familiar 'crony-capitalist' style of leadership of many rapidly developing states, such as Indonesia and Malaysia.

Where the impetus for linking social, economic, and political institutions and issues to ethnic identity is state policy, there is little prospect for any form of assimilation ideology or practice, with all its implications for lack of national identity or shared 'imagined community.' In face of internally contested goals and identities, departments or ministries of 'national unity' are sometimes created to impose or direct what daily living cannot, but without some 'depluralization' (Kuper and Smith 1969), both the objective structures and subjective perceptions of separateness persist. Until this happens, in the view of some political scientists of the 1960s and 1970s, societies are not yet politically modern (Inkeles 1969). What is needed to overcome these problems, according to Geertz (1963) is an 'integrative revolution,' whereby transcendent civic loyalties come to supercede particularistic ones, or turned to the benefit of a more cohesive state. In effect this involves the creation of more cross-cutting institutional and individual networks in the public domain, based on superethnic principles and issues, of a neutral political space corresponding to a pluralist or even to a 'civil society'. While it has been claimed (Eriksen 1994) for Mauritius, that such deppluralization is achievable over time, particularly under the impact of an open market economy, and arguably that even South Africa is moving in this direction, it is at best an unpredictable and oscillating process. Evidence (Eriksen 1994) does suggest that some communal cleavages can be retooled to work for, rather than against, national unity as in Mauritius and Trinidad, whereas post-Suharto Indonesia, post-Abacha Nigeria, and contemporary Sri Lanka, on the other hand, remain vulnerable to the centrifugal tendencies of their internal plural divisions. As Tambiah (1989) has suggested, the initial post-colonial euphoria of nation-building may be sufficient in the short term to paper over the plural cracks, but they do seem to re-assert themselves subsequently as a new round of internal ethnic resource and power politics, or 'retribalization' (Cohen 1974) comes into play. The language of 'ethnonationalism' covers many of the same problems, in the sense that ethnonational movements represent politically mobilized ethnic communities, which may eventually be consummated in an independent or secessionist state, in the manner of Croatia, Serbia, Macedonia, and Bosnia of the 1990s. Others are not (yet) consummated, but remain poised for secession from larger (plural) states, as in Quebec or the Tamil northeast of Sri Lanka, exemplifying the perennial problem of such polities.

In one variation of the plural society's political and socially fissiparous tendencies, the binary type is considered by some observers (Milne 1981) to be the

most fragile. In his selections of 'bipolar' plural states, notably Malaysia, Guyana, and Fiji (to which could be added Sri Lanka and Belgium), the argument is made for the presence of a zero-sum or winner/loser mentality in matters of resource and power distribution, such that A's gain is invariably construed as B's loss, and vice versa. In situations of multiple division, where three or more communities are present, there are more possibilities for cross-cutting coalitions, whereby groups make different alliances according to interest, such that one can neutralize another, and this is particularly clear in the case of political party relations. In practice, however, even in situations where three or more nominal parties or plural segments exist, the actual arrangements may operate in *de facto* binary fashion. Thus in Malaysia, where three principal ethnic-based and several other political parties play the electoral field, in most major issues the public rhetoric is framed in terms of Malays (or Muslims) vs. non-Malays (or non-Muslims), the second category consisting of a collective assortment of other ethnic and religious constituencies. In Indonesia, the plethora of indigenous ethnic groups are packaged politically and symbolically as one (*pribumi*) in face of the (relatively tiny) immigrant Chinese community. Similarly, in Canada, the array of diverse immigrant ethnic communities count as a single unit (the 'Rest of Canada/ROC') or Anglophone, in opposition to Francophone Quebec in the business of federal politics. The reduction of multi- to bipolarity significantly affects the oppositional character and abrasiveness of internal relations, at least at the maximal political level.

Plural constructions of society tend to focus exclusively on the divisiveness of ethnic, racial and cultural, or religious identities and their derivative institutions. Conspicuously missing in most of the plural literature is any reference to class stratification and its potential for the creation of cross-cutting cleavages, or, viewed from another perspective, of unity in a different direction. Where class enters the picture at all, as in the writings of Kuper and Smith, the emphasis is on the fact that each ethnic/racial selection is internally stratified by its own unique set of strata, none of which correspond with those of any other, hence their cooperative, class-action potential is neutralized. No South African black miner would find common cause with a white worker in the same sector, nor a Malay with a Chinese rubber tapper. However, the deflection of class to ethnic/racial images is often as much the product of a state worldview or agenda, whose interest it is to frame all local issues in ethnic terms and to rule by playing off one community against another. This is also intended to a void labor unrest or economic instability, and is reinforced by banning unions (or channeling them along ethnic lines), and other class-like activities. One consequence is that members of such societies spontaneously perceive all inequalities as based on ethnic/communal

or sometimes religious statuses. It is this kind of mindset which causes the Chinese in many countries to be unilaterally classified as the 'exploiters,' while conveniently ignoring the existence both of poor Chinese and rich local elites. Furnivall's 'meeting in the market place' does not necessarily seem to create the de Tocquevillian or Dahrendorfian conditions necessary for a pluralist society.

The lack of attention to the potential or existence of class expression in plural societies may be as much ideological as a reflection of the actual situation. In contrast to Marxian convictions that the fundamental determinant of social inequality is material, and that ethnicity is merely an epiphenomenon of class, plural society theorists maintain that ethnic and racial identities are *sui generis*, the cause of most social inequality and conflict.

Since the late 1970s, the term 'plural' seems to have yielded to or been superseded by the label 'multicultural' in many places. A careful examination of so-called multicultural societies today reveals a number of parallels with the older 'plural' term, in its emphasis on ethnic, racial, or cultural identities as the primary source of diversity and division, often at the expense of class. Some commentators maintain that states which promote multicultural policies, which make a virtue of ethnic difference, and even design many social activities and programs along these lines, are a cynical official device ('myths') to obscure more basic class or race inequalities and lack of opportunities for mobility or political empowerment (Peter 1981) for Canada. Proponents of this view suggest that the choice of the term 'culture' as the operating principle conveys a bland, benign, and misleading image of tolerance and equality of opportunity, turning 'difference' away from ethnic conflict and to the benefit of national unity. Given that most societies today are in the process of becoming demographically, socially, religiously, and culturally more multicultural through immigration, even the older, more stable states of Europe, once confident of their national character and identity, are now rethinking these as they come to terms with unaccustomed internal ethnic and cultural diversity. Particularly pressing are the handling of religious and legal pluralness by the host society. The actual policy responses to *de facto* multiculturalism, however, vary by country and over time, as Germany's (1999) reconsideration of the status of its largely Turkish 'guestworkers' (*Gastarbeiter*) illustrates. In the case of the US, the long-established ideology of assimilation is challenged by ideas of multiculturalism as policy (Walzer 1992, Waldron 1992). Increasingly, the emphasis is on the acceptance of difference rather than on the promotion of homogeneity. Also at issue is the question of recognition of (ethnic/cultural/racial) collectives or group rights vs. the more established privileging of the individual and of individual rights as the underpinning of western liberal democracy (Kymlicka 1995). This has implications for

such 'corrective' policies introduced by some contemporary plural states in the direction of quotas or of affirmative action, which could be construed as benign attempts to condone or perpetuate pluralness (Porter 1975), and hark back to what could be interpreted as compromising the standards of liberal democracy and a civic nation. This dilemma, of simultaneously recognizing group identities and integrity while maintaining liberal democratic ideals, is most articulately treated in Taylor (1992) dissections of Canadian multiculturalism.

A general survey of current world trends indicates a declining emphasis on assimilation as a route to nation-building, from Israel to Australia, largely as a response to accelerated and extensive transnational movements and to the existence of deterritorialized communities. This raises new problems of internal social cohesion within states, as they face the question as to whether multiple identities (and even citizenships) among their populations, (who may simultaneously also be participating in the activities of transnational communities offshore), are necessarily mutually exclusive or to the detriment of the state. The original, rather divisive forms of pluralness may, in certain circumstances, form the foundations of a new style of pluralistic politics of diversity or difference, in dialogic as much as conflictive relations, operating between, as well as within states.

See also: Ethnicity: Anthropological Aspects; Multiculturalism, Anthropology of

Bibliography

- Benedict B 1962 Stratification in plural societies. *American Anthropologist* 64(6): 1235–46
- Boeke J 1953 *Economics and Economic Policies of Dual Societies, as Exemplified by Indonesia*. Institute of Pacific Relations, New York
- Cohen A 1974 *Urban Ethnicity. ASA Monographs, No. 12*. Tavistock, London
- Dahrendorf R 1959 *Class and Class Conflict in Industrial Society*. Routledge & Kegan Paul, London
- Despres L A 1968 Anthropological theory, cultural pluralism and the study of complex societies. *Current Anthropology* 9(1): 3–26
- Despres L A 1975 *Ethnicity and Resource Competition in Plural Societies*. Aldine, Chicago
- Eriksen H 1994 Nationalism Mauritian style: Cultural unity and ethnic diversity. *Comparative Studies in Society & History* 36(3): 549–74
- Furnivall J H 1939 *Colonial Policy and Practice. A Comparative Study of Burma and the Netherlands India*. Cambridge University Press, Cambridge, UK
- Geertz C (ed.) 1963 The integrative revolution. In: Geertz C (ed.) *Old Societies and New States: The Quest for Modernity in Asia and Africa*. Free Press, Glencoe, IL, pp. 105–57
- Inkeles A 1969 Participant citizenship in six developing countries. *American Political Science Review* 57(4): 1120–41
- Kuper L, Smith M G (eds.) 1969 *Pluralism in Africa*. University of California Press, Berkeley, CA
- Kymlicka W (ed.) 1995 *The Rights of Minority Cultures*. Oxford University Press, Oxford, UK
- Lijphart A 1977 *Democracy in Plural Societies: A Comparative Explanation*. Yale University Press, New Haven, CT
- Milne R S 1981 *Political in Ethnically Bipolar States: Guyana, Malaysia, Fiji*. University of British Columbia Press, Vancouver, BC, Canada
- Morris H S 1967 Some aspects of the concept plural society. *Man* 2(2): 169–84
- Nagata J 1979 *Malaysian Mosaic: Perspectives from a Poly-ethnic Society*. University of British Columbia Press, Vancouver, BC, Canada
- Peter K 1981 The myth of multiculturalism and other fables. In: Dahlie-Fernando T (ed.) *Ethnicity, Power & Politics in Canada*. Methuen Press, Toronto, Canada, pp. 56–7
- Porter J 1965 *The Vertical Mosaic*. University of Toronto Press, Toronto
- Porter J 1972 Dilemmas & contradictions of a multi-ethnic society. *Transactions of the Royal Society of Canada* 10: 193–205
- Ratnam K J 1965 *Communalism and the Political Process in Malaya*. University of Malaya Press, Kuala Lumpur, Malaya
- Smith M G 1965a *The Plural Society in the British West Indies*. University of California Press, Berkeley, CA
- Smith M G 1965b *Stratification in Grenada*. University of California Press, Berkeley, CA
- Tambiah S 1989 The politics of ethnicity. *American Ethnologist* 16(2): 335–49
- Taylor C 1992 *Multiculturalism: Examining the Politics of Recognition*. Princeton University Press, Princeton, NJ
- Van den Berghe P L 1970 *Race and Ethnicity: Essay in Comparative Sociology*. Basic Books, New York
- Waldron J 1992 Minority cultures and the cosmopolitan alternative. *University of Michigan Journal of Law Reform* 25: 751–992
- Walzer M 1992 Comment. In: Taylor C (ed.) *Multiculturalism: Examining the Politics of Recognition*. Princeton University Press, Princeton, NJ, pp. 99–104

J. Nagata

Pluralism

The term 'pluralism' was introduced into political science by the Englishman Harold Laski in a number of essays published mostly while teaching at Harvard University between 1916 and 1920. For the term, Laski was indebted to the pragmatist philosophy of William James, whose book, *A Pluralistic Universe* (1909), Laski hailed as of 'vital significance for political theory.' In philosophy, pluralism, as distinguished from 'monism,' denoted a system of thought which recognizes more than one ultimate principle. Although Laski borrowed the word 'pluralism' from American pragmatism, the ideas he advanced under the rubric of pluralism were taken largely from a heterogeneous group of European, and in particular English, legal historians and political theorists whose work Laski encountered while a student at Oxford immediately prior to World War I. Among the most important of these were the German medievalist Otto Von Guericke;

Frederick Maitland, the preeminent legal historian in England; J. Neville Figgis; and Laski's Oxford tutor, Ernest Barker, who in 1914 coined the terms 'polyarchism' and 'the new federalism' to describe this new set of ideas.

The defining project of the English pluralists (as they would come to be known by the time Laski returned to England in 1920) was to vindicate the rights and autonomy of associations by challenging the modern doctrine of unlimited, unitary state sovereignty. Pluralists drew an invidious contrast between the 'federalistic structure of medieval society,' in which a dense network of autonomous associations flourished, and the modern notion of state sovereignty, which, Maitland lamented, reduced 'to impotence and then to nullity all that intervenes between Man and State.' But pluralism was much more than an historical description of the world we have lost; the concept of sovereignty was not just unfortunate, it was untrue to the facts of social life. 'We say,' Figgis explained, 'that the world is *not*, as a fact, composed of a few vast unities known as States, set over against crowds of isolated individuals; but is a society of societies, each and all with rights, liberty and life of their own.' Corporations were not legal fictions created by the state, but rather were natural organizations with their own autonomous existence. The idea of unlimited state sovereignty, Laski argued, was 'a fiction of logic ... out of accord with the facts of life'; the 'history of societies' showed repeatedly that organized groups, even unpopular ones like militant female suffragists, striking miners, or conscientious objectors, imposed limits on the allegedly sovereign state, forcing public officials to adopt policies to which they were opposed.

Although pluralists like Laski and Figgis frequently appealed to 'the real facts of human society,' their challenge to what they termed 'the monistic theory of the state' was more normative than empirical, though typically the 'is' and the 'ought' were woven seamlessly together. In arguing that 'the state is only one among many forms of human association,' for instance, Laski meant not only to highlight the empirical fact of the social and political importance of associations and groups, but also, and more important, to advance a radically anti-statist normative vision. For the young Laski, the state was owed no more loyalty than other groups; each group, including the state, deserved only the loyalty it earned 'by virtue of its achievement.' The 'ought,' in Laski's mind, derived from the 'is': It was the 'facts themselves' that 'dissolve ... the inherent claim of the State to obedience.' Not all pluralists, however, shared Laski's anarchic and syndicalist sympathies. Others, like Figgis and Barker, for instance, allowed that the state had an important role to play in regulating conflict between associations and thus in creating the political conditions necessary for associations to pursue their own ends. But while Figgis and Laski might disagree on what should replace the

monistic state, they both looked to pluralism less to explain the world than to change it. Their main objective was to help the groups they cared most deeply about, which, in the case of Figgis, an Anglican priest, were religious associations, and in the case of Laski, as well as the guild socialist theoretician, G. D. H. Cole, were labor and the trade unions. And herein lies an explanation for English pluralism's precipitous decline in the 1920s. For pluralists failed to persuade others, or even eventually themselves, that pluralism was the most effective means to help the groups they wanted to assist. Laski himself gradually abandoned or watered down his pluralist ideas because, along with the rest of the left, he came to believe, particularly after the Great Crash of 1929, that labor had more to gain from an active, even paternal, state than from a minimal, pluralist state.

Ironically, at the time that English pluralism as a normative vision and intellectual force was in decline, the empirical study of organized groups was picking up steam in American political science, particularly with the publication of Peter Odegard's *Pressure Politics: The Story of the Anti-Saloon League* (1928) and Pendleton Herring's *Group Representation Before Congress* (1929). These sorts of empirical studies of the role of organized groups in the American political process never interested Laski. Indeed, they only confirmed his belief that the trouble with American political science was that 'everywhere the emphasis is on description and not principle,' a diagnosis diametrically opposed to that held by most American political scientists interested in the role of groups during this era. They believed, with Arthur Bentley, that there was too much moralizing about the evils of special interest groups and too little objective empirical description and analysis of group behavior in the political process. Over the next quarter century American political scientists worked diligently to remedy these empirical shortcomings, such that by 1950, a year before the publication of David Truman's landmark book, *The Governmental Process*, (Truman 1971/1951), Merle Fainsod could conclude that 'there are few fields which American political scientists have cultivated as intensively and profitably as the analysis of the dynamic interplay of interest groups as they give shape to public policy at almost all governmental levels.' (cited in Crick 1959)

In 1950 the discipline was relatively rich in empirical studies of interest group influence but the term 'pluralism' was rarely, if ever, used, though political scientists were beginning to speak of the 'group basis of politics' or the 'group theory of politics.' Truman as well as Bertram Gross had helped to revive interest in Bentley who, in his 1908 book, *The Process of Government* (Bentley 1967/1908), had argued that groups, not laws, institutions, or constitutions, were the essential 'raw materials' of government. Truman followed Bentley in trying to understand the government process by treating groups, specifically interest

groups, as 'a primary unit of analysis.' Over the next decade this increasingly influential mode of analysis would become known by advocates and even more frequently by detractors as 'analytical pluralism' or simply 'pluralism.'

Critics were quick to charge pluralists with believing that groups were all that mattered, often pointing to Bentley's audacious remark that 'when the groups are adequately stated, everything is stated.' But most of those tagged as pluralists, including most especially Truman, never held to anything so implausible. Pluralists looked to groups (in Truman's words) as 'a major explanatory variable,' but explicitly rejected the notion that government officials merely passively registered or mirrored the sum total of organized group preferences. The relative importance of groups and government officials in the shaping of policy, Truman emphasized, was an empirical question, not one to be settled by 'preassigning weights and roles.' Analytic pluralists, to be sure, generally avoided using the term 'the state,' but this was not because their explanations were 'society-centered' or because they believed government officials lacked autonomy or importance vis-à-vis societal groups, but rather because they doubted, for epistemological as well as empirical reasons, that the myriad institutions and officials of government, particularly in the United States, were coherent enough to warrant a locution that treated the government as a unitary actor with a single will (Ellis 1992). The same doubts fed pluralists' skepticism of concepts like 'the public interest' or 'the will of the people,' since society no less than the state was a complex mix of rival interests and objectives. In this respect, the modern American pluralists were kindred spirits of the early-twentieth-century English pluralists.

Another strain of criticism of American pluralism objected not to its focus on societal groups but to its portrayal of the distribution of power within the United States as dispersed. Pluralists, it was claimed, slighted the severe inequalities in resources and power in American society, and believed mistakenly that the system was self-correcting: that the mobilization of one group would necessarily lead to the mobilization of an equal and opposite countervailing pressure. Pluralists were also said to believe that all important or legitimate interests and opinions were represented within the political system, or even that all groups had substantially equal access to the policy-making process. As the ideological dust kicked up in the 1960s has slowly settled, scholars (e.g., Smith 1990, Hirst 1987) have begun to recognize that few, if any, of those scholars commonly described as pluralists harbored such naive views of the political process. The important question, in any event, is not whether it is possible to find someone myopic enough to believe that all groups have equal, or nearly equal, power, but whether pluralists are logically committed to these propositions. How widely dispersed power is within

any given policy domain or in any given place or time is an empirical question that cannot be settled by methodological or theoretical fiat. Nor can it be answered satisfactorily without a comparative perspective: the same distribution of power may be reasonably characterized as either concentrated or dispersed depending on one's comparative referent or on one's normative benchmark. There is no inherent reason that an empirical commitment to the study of interest groups *à la* Truman need predispose the analyst to find any given distribution of political power. Nor need a normative commitment to pluralism, understood as wide dispersion of political influence and power, necessarily predispose one to find pluralism as a sociopolitical fact in any given setting.

The term 'pluralist' is a particularly confusing label because it is often applied to those who conclude that power in a given locale is dispersed rather than concentrated. Pluralism can be a useful term to describe either an empirical reality—one in which there is, for example, widespread dispersal of power, and bargaining rather than hierarchical decision-making—or a normative commitment to such dispersed power, but it does not help to call those scholars who find dispersed power 'pluralists.' When Robert Dahl (1961) finds a pattern of dispersed influence in contemporary New Haven politics, that no more makes him a pluralist than his discovery of concentrated oligarchic power in early New Haven makes him an elitist.

Dahl remains the archetypal pluralist in the minds of many, even though he is not a group theorist on the model of Bentley and Truman. Whether or not Dahl counts as a pluralist, he is certainly among the keenest student of the dilemmas of pluralism. As a graduate student at Yale in the 1930s, Dahl took a year-long seminar on recent political thought from Francis Coker, who had written the entry on pluralism for the *Encyclopedia of the Social Sciences* (1934). There Dahl encountered Laski and the English pluralists as well as Coker's ambivalent critique, in which he praised pluralists for drawing attention to the importance of autonomous groups intermediate between the individual and the state, while also criticizing pluralists for slighting the need for centralized political control. State control was necessary, Coker argued, not only to referee intergroup conflict but also to protect the 'unorganized interests of the general body of citizens' against the organized. Moreover, without active state intervention pluralist group autonomy would result in stronger interests, particularly employers abrogating the rights of weaker interests, particularly wage-workers. Dahl amplified his teachers' arguments in *Politics, Economics and Welfare* (1953). It was one thing, Dahl and his co-author Charles Lindblom said, to recognize (as did Laski) that government officials often could not 'enforce decisions by unilateral, hierarchical means' but must often bargain with

powerful and autonomous groups to negotiate 'temporary armistices.' However, it was 'quite another to turn this social fact into a prescription of the desirable, and to argue in effect that politicians should not even attempt to exercise "the last say," but should turn that power over to national organizations bargaining among themselves.' The fundamental dilemma of a pluralist democracy, in Dahl's (1982) view, is that autonomous associations, groups, or organizations are highly desirable and yet also capable of doing great harm if not controlled by a central authority.

Why is pluralism desirable? Advocates of pluralism see autonomous organizations fulfilling a variety of beneficial purposes. Some emphasize pluralism's role in facilitating individual development and self-expression, while others emphasize the importance of groups in integrating the individual into society. Associations are also seen as seedbeds of civic virtue, an enabling environment in which citizens learn the habits of self-government. A different emphasis is evident among those who stress pluralism's defensive political function, as an organizational counterweight to the coercive powers of the central government. Pluralism's vital contributions to democratic governance has long been Dahl's primary focus, a concern that has renewed resonance as scholars, politicians, and citizens attempt to understand and direct the process of democratization in the former communist countries in Eastern Europe and the Soviet Union. In the absence of vibrant organizational pluralism, the experiences of these nations suggest, it is difficult for formal democratic institutions such as elections, parliaments, and presidents to curb hierarchy and domination.

Pluralism may be essential to the functioning of large-scale democracies but that does not mean, as Dahl himself emphasizes, that pluralism is without disadvantages. To begin with, group leaders may dominate or oppress group members, particularly where the costs of exit from a group are high. Less ominously, the group's elites may fail to represent the opinions of its membership—the elites may be more conservative or more radical, more compromising or more confrontational, than the membership. The English pluralists at times obscured these problems by treating organizations, whether churches or trade unions, as if they were undifferentiated organic wholes with 'real' personalities (even while they denied such a real personality to the state), but American analytic pluralism was rarely, if ever, guilty of reifying groups. Far from treating large groups as if they were homogenous, as is sometimes alleged (Kariel 1968), Bentley and other American group theorists were keenly aware, in Odegard's words, that all groups 'are themselves almost infinitely divisible into subgroups whose interests are not always congruent or compatible with one another.'

A second problem with organizational pluralism is that all groups are not created equal: organized

interests, all other things being equal, have an advantage over unorganized interests, and some interests are easier to organize than others. Autonomy for the lambs may well mean a quick dinner for the wolves. It is the weaker groups, as Schattschneider (1960) argued, who generally have the greatest interest in 'enlarging the scope of conflict' by appealing to public authority and thereby modifying the imbalance in private power relations. Recognition that group autonomy may serve the interests of the most powerful groups is a potent argument against the radically antistatist vision of the English pluralists, but not nearly so troubling to the pluralist vision of post-war American political scientists, most of whom were New Deal Democrats. The problems of group autonomy are real, but they need to be counterbalanced with an appreciation for the equally real limits of central control. Absent the articulation of group preferences and mobilization of group interests, a central political authority would define who was a group and what were legitimate interests. Without group pressures, the state would find it difficult if not impossible to gauge the intensity with which preferences are held. Group autonomy without central control may be bad, but it is no worse than central control without group autonomy.

Other critics worry that, even if all groups were equal, pluralism may, in Dahl's words, 'deform civic consciousness' by fostering particularistic or short-run demands at the expense of a broader common good or long-run social needs. Whether civic consciousness is higher or the common good is more closely attended to in those countries or policy domains where interest groups are less prevalent or less assertive is an empirical question, the answer to which is far from self-evident. A related concern is that private associations may be given essentially public functions, thereby removing or reducing democratic control over governmental decisions. One manifestation of this problem is European-style corporatism, in which bargains are worked out between peak associations representing labor, business, and farmers. In the United States, the problem of alienating democratic control may take the form of powerful groups capitalizing on the fragmentation of the system by 'capturing' governmental officials in a specialized policy domain (McConnell 1966, Lowi 1969). Dispersed power is not inconsistent with groups monopolizing power in specialized segments of the political system. Again, the extent to which this describes accurately the policy-making in any given system or policy domain is an empirical question.

Among the most common criticisms of pluralism is that it was little more than a self-satisfied apologia for American or, alternatively, Western, democracy. Whether or not those who were associated with pluralist ideas were satisfied with the state of democracy in America or Western Europe, there is no reason to think pluralism as a normative vision is inherently

conservative. Hirst (1989), for instance, uses pluralist ideas to launch a blistering critique of the British political system, particularly Parliament's claims to unlimited sovereignty. Measuring Western democracies against an exacting yardstick of pluralist principles, Hirst (1997) finds severely wanting not only the West's narrowly plebiscitarian vision of democracy, in which 'the people periodically choose who are to be their elected masters,' but also the over-centralization of the provision and delivery of public services. Taking pluralism seriously, Hirst maintains, means reshaping governance radically, not only in the nation state but at the level of international institutions like the European Community. Dahl himself draws upon pluralist ideas to argue for a dramatic devolution of power to cooperatively owned economic enterprises. And in the context of the non-Western world, even the tamest pluralist ideas can constitute a bracing challenge to entrenched state power and economic privilege. Writing in the 1968 *International Encyclopedia of the Social Sciences*, Henry Kariel wrote an obituary for political pluralism, which, he said, 'as an ideology ... only lingers quietly as a submerged, inarticulate ingredient of Western liberalism' (Kariel 1968). Over three decades later, it would appear that reports of pluralism's decline, let alone death, have been greatly exaggerated. Particularly as a normative vision, as an antidote to plebiscitary democracy or to centralized state control, political pluralism has shown renewed appeal and relevance across the globe.

See also: Civic Culture; Cleavages: Political; Democracy; Democratic Theory; Interest Groups; Interest Groups, History of; Lobbying; Participation: Political; Party Systems; Political Parties; Political Representation; Sovereignty: Political; Voluntary Associations, Sociology of; Voluntary Organizations

Bibliography

- Baumgartner F R, Leech B L 1998 *Basic Interests: The Importance of Groups in Politics and in Political Science*. Princeton University Press, Princeton, NJ
- Bentley A F 1967 [1908] *The Process of Government*. Belknap Press, Cambridge, MA
- Coker F 1934 Pluralism. In: Seligman E (ed.) *Encyclopedia of the Social Sciences*. Macmillan, New York
- Crick B 1959 *The American Science of Politics*. University of California Press, Berkeley, CA
- Dahl R, Lindblom C 1953 *Politics, Economics and Welfare*. Harper, New York
- Dahl R A 1961 *Who Governs? Democracy and Power in an American City*. Yale University Press, New Haven, CT
- Dahl R A 1982 *Dilemmas of Pluralist Democracy: Autonomy vs. Control*. Yale University Press, New Haven, CT
- Ellis R J 1992 Pluralist political science and 'the state': Distinguishing between autonomy and coherence. *Polity* **24**: 569–89
- Gunnell J G 1995 *The Declination of the 'State' and the Origins of American Pluralism*. In: Farr J, Dryzek J S, Leonard S T (eds.) *Political Science in History: Research Programs and Political Traditions*. Cambridge University Press, Cambridge, UK
- Herring P 1929 *Group Representation Before Congress*. Johns Hopkins Press, Baltimore, MD
- Hirst P Q 1987 Retrieving pluralism. In: Outhwaite M, Mulkay M (eds.) *Social Theory and Social Criticism*. Basil Blackwell, Oxford, UK
- Hirst P Q (ed.) 1989 *The Pluralist Theory of the State: Selected Writings of Cole G D H, Figgis J N, and Laski H J*. Routledge, London
- Hirst P Q 1997 *From Statism to Pluralism: Democracy, Civil Society and Global Politics*. UCL Press, London
- James W 1909 *A Pluralistic Universe*. Longmans, Green, New York
- Kariel H S 1961 *The Decline of American Pluralism*. Stanford University Press, Stanford, CA
- Kariel H S 1968 Pluralism. In: Sills D L (ed.) *International Encyclopedia of the Social Sciences*. Macmillan, New York
- Laski H J 1917 *Studies in the Problem of Sovereignty*. Yale University Press, New Haven, CT
- Laski H J 1919 *Authority is the Modern State*. Yale University Press, New Haven, CT
- Laski H J 1921 *The Foundations of Sovereignty and Other Essays*. Harcourt, Brace and Company, New York
- Lowi T J 1969 *The End of Liberalism*. W. W. Norton, New York
- McConnell G 1966 *Private Power and American Democracy*. Knopf, New York
- Nicholls D 1975 *The Pluralist State*. St. Martin's Press, New York
- Odegard P 1928 *Pressure Politics: The Story of the Anti-saloon League*. Columbia University Press, New York
- Polsby N W 1980 *Community Power and Political Theory*, 2nd edn. Yale University Press, New Haven, CT
- Runciman D 1997 *Pluralism and the Personality of the State*. Cambridge University Press, Cambridge, UK
- Schattschneider E E 1960 *The Semisovereign People*. Holt, Rinehart and Winston, New York
- Smith M J 1990 Pluralism, reformed pluralism and neo-pluralism: The role of pressure groups in policy-making. *Political Studies* **38**: 302–22
- Solomon S G 1983 'Pluralism' in political science: The odyssey of a concept. In: Solomon S G (ed.) *Pluralism in the Soviet Union*. St. Martin's Press, New York
- Truman D 1971 [1951] *The Governmental Process: Political Interests and Public Opinion*, 2nd edn. Knopf, New York

R. J. Ellis

Pluralism and Toleration

Pluralism is a characteristic of the contemporary life-world that has developed in the course of modernity into the dominant basis for social freedom and political justice. Today, the features as well as the relevance of this basis have not, gone uncontested (Schultz 1974, Kelso 1978). We generally understand pluralism not only as a diversity, but rather as a right to an equal coexistence and confrontation of natural or social elements which lack a governing principle of unity. Tolerance is closely associated with the notion of pluralism insofar as it

refers to those individual virtues which derive from the recognition of this right to an equal coexistence of a multitude of natural and social elements. Pluralism and tolerance consider the same phenomenon from different perspectives. Whereas pluralism emphasizes the diversity of equally entitled, free groups, tolerance stresses the 'live and let live' of others' liberty.

1. The Idea of Pluralism

Pluralism can be understood in a metaphysical sense as designating those philosophies which allow for a diversity of entities in their explanations of the world. The pragmatist philosopher William James maintained the existence of an infinite number of limited facts, connections, and systems which are in constant flux; which always lead to new relationships and structures; and for the cognition of which, fixed categories and criteria do not exist, only perspective orientation points (James 1909). Pluralism can also refer to the methodological neutrality of the social sciences of an empiristic orientation, to the multiplicity of cultural forms, from which 'all sociology has its starting point' (Scheler 1980, p. 25). Today, the term is primarily understood in a social and political sense in which empirical and normative elements converge.

1.1 Empirical and Normative Elements

Empirically speaking, pluralism denotes the diversity not only of confession and religion (religious pluralism), but also of social groups together with their lifestyles (social pluralism), of basic values (pluralism of value), and of the powers which determine politics (political pluralism). For this many-dimensional diversity, our life-world has primarily benefited from the religious wars of the Reformation, the European Enlightenment, and the Bourgeois and Industrial Revolutions. Through those factors, a process of differentiation was set into motion which dissolved both the relatively homogenous and stable relationships of the so-called Old-European societies and the connection between church and state. Moreover, this process brought forth open societies, the numerous and constantly expanding groups of which had their own interests, convictions, and manners of conduct. More recently, pluralism has gained a new momentum through the process of globalization (Höffe 1999), in particular also through migration movements, many refugees, and those looking for asylum. Second, pluralism purports—and here one finds its normative content—that diversity and difference is acknowledged and approved of; all functionally different groups are equally entitled to develop themselves in a free and unconstrained manner.

The notion of pluralism is of particular relevance in modern democracy theory (Fraenkel 1979, Oberreuter 1980). It is considered here as a critical notion, namely

as that central mediative category between the individual and the state, with which the liberal atomization of individuals just as the omnipotence and an all-encompassing capacity of the state are rejected. In a democracy, so runs the empirical thesis of pluralism; many-faceted powers are at work, which, on the basis of a fundamental consensus over institutionalized arbitration with and against one another, struggle for influence and importance. As far as the normative content of pluralism is concerned, various social forces, notwithstanding their differences in content and function, have the same right to engage themselves in the political process by the means of, e.g., voting, of contributions to public deliberation, or by addressing themselves to members of parliament. These forces include trade groups, unions, and employers' associations; scientific organizations, churches, and political parties; and even civil action groups. Politics is not derived from the conception of a common good which is clearly recognizable, prior to all political confrontations and which does justice to all parties concerned. Political decisions are a result of the constant dispute between competing interest groups which have to mutually monitor one another and, in constantly new coalitions, 'scuffle' together into a compromise which might be accepted by a majority.

1.2 The Critique of Pluralism

The criticism of the idea of pluralism is leveled partly at its empirical thesis and partly at its normative content; the normative criticism pertains to either pluralism of value or political pluralism.

Pluralism of value becomes visible in the form of the ideologically neutral state. It is reproached with relativism and even nihilism, because it does not even acknowledge certain very general values as binding and is held responsible, therefore, for the crises of meaning and orientation which threaten contemporary industrialized societies. Pushed to the fore, the ideologically neutral state will abolish itself, because it cannot legitimize the framework in the limits of which various powers are capable of freely expressing themselves.

The most significant criticism of *political pluralism* stems from Carl Schmitt and his followers, who fear the centrifugal, 'anarchistic' tendencies of pluralism. Through the openness and insecurity of pluralistic democracy, the state's power of decision making is left crippled; moreover, the compromises resulting from such a situation correspond to neither moral nor economical ideals of sound politics.

Supporters of the critical theory hold pluralism for desirable, but contest the existence of the empirical presuppositions for its realization. They point to the socially underprivileged, who are either effectively shut out of participating in the bargaining process of weighing competing interests, or are, at least, clearly at a disadvantage when up against more powerful

organizations. The normative idea of an equal right for all groups is, therefore, rather a postulate than social reality. Other critics maintain a 'standardization of subjects,' who are no longer capable of an autonomous definition of their interests and, hence, of a genuine pluralistic differentiation. Where a free concert of interest groups once existed, one now finds bureaucratic apparatuses in their place, the internal dynamics of which dominates the political scene.

1.3 Pluralism and Liberal Democracy

Political pluralism presupposes two conditions as given. On the one hand, all interests and the groups which represent them should have equal rights in public deliberation and in the process of political will-formation; on the other hand, the common good should realize itself through the struggle of interests between groups of equal rights. Both conditions are fulfilled, even in developed democracies, at best in approximation. For in pluralistic society, citizens with their interests and convictions have no direct access to politics; their access stands in need of mediation through clubs, organizations, and political parties. In this manner, interests and convictions do indeed gain political power, however at different levels of intensity. The mediation of interests and convictions very often leads to perceptible distortions of their influence.

Due to varying historical or social conditions, many interests and convictions are represented more powerfully than others. The explanation of this fact might be seen in the tighter organization of certain associations, to the better support by more potent money lenders, or to the fact that such groups have attractive media at their disposal, or better relations to the mass media. An equality of rights in modern times is not realized through the abolishment of legal privileges, characterizing hierarchical feudal or class societies. Pluralism does not simply describe our social and political reality. It is to a significant degree really a postulate, according to which societies need to be changed and modified.

A modification within pluralism, however, is possible only to a limited extent: first of all, because powerful groups do not freely relinquish their predominance; second, as can be observed sociologically (Olson 1965), many interests are much easier to organize than others, which may lead to the superiority of certain economic groups (employers as much as employees). Furthermore, mergings and the oligarchical tendencies within organizations endanger civic freedom, the essence of pluralism. Dependence on the government gives way to dependence on both the bureaucracy of associations and its functionaries. Finally, even a fair arrangement between a plurality of interest groups does not simply lead to the realization of the common good; for, as a rule, long-term interests get the worst of short-term interests, just as general interests get the worst of particular interests. It is for

this reason that the protection of the environment is so easy to push aside.

In taking an overview of these arguments, one must neither adhere to a conspiracy theory, nor to an extreme criticism of capitalism, in order to preserve a residue of skepticism with regard to the affirmation of an empirically existing pluralism. Hence, the plurality of intermediary groups is without a doubt a pillar of freedom; a pluralistic society is invariably better than a nonpluralistic one; and pluralism itself an important structural element of the liberal-constitutional democracy. Nevertheless, pluralism alone cannot warrant its own significance; it is only a helpful not in any way sufficient organizational form of freedom and equality. One should not easily identify pluralism and liberal democracy because the notion of pluralism brings to light only diversity and competition, and not the complementary elements of commonality and cooperation.

A liberal democracy consists not only of the coexistence and confrontation of a diversity, but also of the regulation of coexistence and confrontation. Regulation brings about community: the mutuality of problems to be solved and, in particular, that fundamental (constitutional) consensus which defines the normative content and the rules of pluralism. In this manner, pluralism shows itself to be a one-sided, nondialectical notion that stresses plurality instead of unity, competition instead of cooperation.

1.4 A Relative Pluralism

Because empirical and normative elements converge in the notion of pluralism, the normative valuation has already begun in the empirical thesis. It is continued with a *pragmatic* consideration: the decentralizing of political power, the institutional and social separation of the process of will formation, just as the diversity of convictions and manners of conduct increase the chance of an efficient coexistence, in which superfluous frustrations and conflicts decrease.

The comprehensive *moral* evaluation begins with an ethical-anthropological insight. Because of its lack of instinct and a wanting specialization of organs, mankind is not fixed to definite manners of conduct. Due to its capacity for language and reason, mankind is much more reflective in relation to its doings and omissions; this not only enables it to act freely and responsibly, but also constitutes a particular challenge. Therefore, mankind, be it as individual, a small or a large group, has the task and the right to develop its own convictions and practices, and to determine and pursue a particular conception of a good life. Different convictions and practices are to be expected, on account of the historical factors under which this process of personal as well as social self-determination takes place. The form as well as the results of self-determination are limited by different perspectives, and a single individual cannot lay claim to the

unconditional 'truth' for himself (Rawls 1993, p. 36). In the pluralism of interests, of values, and of life-plans, the wealth of human self-realization is expressed just as well as the limitation of every concrete design.

On the other hand, different groups are already related to each other through the social division of labor; but the division of labor represents only one of the many aspects which make mediation necessary. Within one's own group, the acknowledgment of interests and convictions is commonplace. However, with regard to other groups, they stand in both competition and conflict with other groups. Because the solution of conflicts is not secured by instincts, there exists the danger of collective self-destruction of mankind. In order to check this danger and, at the same time, to avoid a crude sort of social Darwinism which makes the weak even weaker and the vital and cunning even stronger, general obligations are necessary, which guarantee a humane and just management of conflicts. For that reason, not an absolute and merely a relative pluralism that develops in the context of fundamental commonalities is really desirable.

Pluralism of this sort can encounter the suspicion of being relativistic and nihilistic. Even if modern societies are skeptical of a timelessly valid code of values, even if they do not confer infallibility upon a particular personal, political, religious, or intellectual instance, they nonetheless need at least a few generally valid obligations removed from the opposition of particular designs of faith and existence. Therefore, human rights is accrued a particular meaning: as the personal right to liberty, the political right to participation, and as social and cultural rights, they are the condition and expression of the mutual acknowledgment of fundamentally equal persons. Therefore, human rights refer to those claims which every individual may legitimately raise: they belong to a sphere beyond a legitimate pluralism and which every legal order, every state, especially pluralistic democracy owes to each individual. Through human rights even those minorities which come off badly in the pluralistic struggle of competition are protected (Rawls 1971, Höffe 1995).

2. The Idea of Toleration

Tolerance is realized through the recognition of those groups which are differentiated by interests, beliefs, and lifestyles. In this manner, the idea of pluralism is, through its normative content, aligned from the outset with tolerance; and tolerance, like pluralism, has its basis of legitimation in the idea of justice and in the modern principle of liberty.

2.1 Its Origin in Religious Toleration

Like the general principle of liberty, the particular principle of tolerance found its principal acknowledged

ment only after long intellectual and sociopolitical debate. In this context, religious tolerance plays a particular role (Püttner 1977).

Because polytheistic religions make a lesser claim to truth than monotheistic ones, they do not face many difficulties as far as the toleration of diverse religious beliefs is concerned. Pre-Christian Rome allows all conquered peoples to practice and even propagate their own cults. Yet the Christians were excluded from the rule of tolerance, but not for religious reasons; from their refusal to observe the transmitted gods, one fears a backlash within the entire commonwealth.

As long as the Christians are persecuted, they appeal to the Roman Emperors, referring to the rights which all mankind are due: that freedom of religion is granted internally and externally. But after Christianity is promoted to the status of a national religion, it soon forgets this demand and its underlying principles contained in the New Testament: even the principle of reciprocity (*Matt.* 7,12; also *Tob.* 4,16); the sermon on the mountain; the parable of the weed under the wheat (*Matt.* 13,24–30, 36–43); Jesus' conduct in general of welcoming, but not of forcing his imitation; moreover his magnanimous attitude toward sinners; finally, Paul's conception of freedom of religion as free from enforcement and his admonition to mutual tolerance (*Col.* 3,12f.; *Cor.* 1, 4,12; *Cor.* 2, 11,1; *Gal.* 5,1.13).

Whereas the *Edict of Milan* (313) leaves religious belief to its practitioners and, moreover, differentiates between state and religion and does not admit mortal or corporeal punishment for religious offences, the swift rise of schisms and heresies fills the bishops with such alarm, even in the patristic times, that they begin to bind, as in the Old Covenant, religion closely with political affairs. Especially the conversion to a foreign cult is punished to the same extent as capital crimes and the disruption of civil order.

Granted, medieval Christendom should not to be considered intolerant through and through (Patschovsky 1998). Because it regards the transmittance of faith as a matter of free conscience, the conversion of pagans by force represents a clear exception. Jews are generally and, often, truly tolerated; and for a long time the Christians live in mutual tolerance with the Muslims. But, the heretic in Christendom, accused of both arrogance and impertinence, was considered, like the Islamic apostate, as it were, an outlaw in the eyes of the law. In the eleventh and twelfth centuries, capital punishment extends to the heretic and is maintained until well after the Reformation. The foundation of medieval intolerance, the all too intimate fusion of *Sacerdotium* and *Imperium*, of spiritual and secular matters are all well preserved in Protestant countries, more likely being advantageous to the state than to the church.

In the 'confessional age' (the age of the Reformation), the idea of (religious) tolerance has to be re-established; and it is effected only after hefty intellectual debate and painful political experience. In

order to encourage the victory of tolerance in the form of the religious neutrality of the state, the mutual autonomy of the state and the church, a few fundamental insights first have to be (re)attained. At first the pragmatic consideration of the 'greater benefit,' the fact that in most cases mildness is better suited than force in order to gain back 'lost souls.' Then one also has to reach the insight that the New Testament is based on tolerance and love, moreover, that not only the conversion but also the retention of faith is both an act of freedom and of mercy, and for that reason, cannot be enforced; furthermore, that the church is a free communion, whose task is to be fulfilled in the spiritual, not in the secular realm, and that for this task every possible constraint is useless, as it would at best lead to a feigned and not to a real acknowledgment of faith.

The way for this new form of tolerance is prepared by Christian humanists like Dante, Marsilius of Padua, Nicolaus von Kues, and Sebastian Castellio, who reject all acts of violence against religious dissidents. However, they are still fixed to ideas of religious unity within Christendom, and, therefore, they search for a peaceful reconciliation by way of religious forum and dogmatic compromise between rivaling confessions. On the contrary, mystic spiritualists as well as some sects engage themselves directly for the religious liberty of each individual; they do this by appealing to the Holy Ghost, who forbids every form of slavery in the realm of the religious, including that of the institutional church.

Whereas no lasting success is allotted to either of these thoughts (peaceful reconciliation or religious individualism), the second half of the sixteenth century sees the problem shift from the religious to the 'secularized' level. Theological arguments are succeeded by considerations of politics and natural right, occasionally even of economics; unsuccessful reconciliatory talks give way to efforts of toleration, in which the denominational heterodox gains ground.

Starting in France, where the preservation of the nation is considered more important than the re-establishment of religious unity, the idea of civil tolerance, which unambiguously differentiates between the aims of the state and the aims of religion, is established against the rule of 'one confession, one law, one king.' With the idea of the confessional, later the religious neutrality of the state, the bloody experiences of confessional civil wars which, at the European level, ended in a stalemate, are finally digested. Consequentially, the relative autonomy of the state and church/religion and, with that, of confessional pluralism, is acknowledged as a necessary condition for political, even physical survival. Nevertheless, the numerous edicts of tolerance, which follow on the Peace of Augsburg (1555) and the Edict of Nantes (1598), are issued only reluctantly; Brandenburg-Prussia represents an exception for a long period. In auspicious circumstances, as in France (1685), the

edicts are repealed or, as in Great Britain, where the Catholics remain second-class citizens until the nineteenth century, they are not valid for everyone from the outset.

Because intolerance endangers the free development of business and trade, Europe and North America also have economic reasons to thank for the following two phenomena: that religious liberty does indeed assert itself in the end, and that the remnants of the medieval connection between the state and the church are vanquished. The European Enlightenment, however, plays a particular role in the ultimate victory of the principle of religious tolerance. Inspired by the idea of universal human reason and liberty, such figures as Grotius and Spinoza, Milton, Locke and Voltaire, Lessing (*Nathan the Sage*), later also Schiller (*Don Carlos*) turn to the remnants of dogmatic and political intolerance within the high church and to their corresponding, amalgamated states; moreover, they lay claim to religious liberty as a fundamental personal right to liberty. But, even those propagating the Enlightenment as a rule exclude atheists and nonbelievers from the idea of tolerance, granted, as with Locke, not out of religious but rather out of political reasons; for, whoever does not believe in God cannot take a legally admissible oath.

2.2 *Passive and Active Tolerance*

The idea of tolerance kindled itself on the freedom of religion in Western Christianity. In our advanced, secularized world this idea no longer represents a special problem with regard to the social interaction of diverse confessions and religions. Tolerance is often enough primarily called for in relation to beliefs which are socially, politically, and culturally differentiated. Thereby, a weaker, more passive form of tolerance, the mere toleration of otherness, is to be differentiated from a stronger, more active, and creative form, the openness to differences, yes, even to their candid recognition.

As pluralistic democracy cannot arise nor continue to subsist without the mutual toleration of otherness, tolerance is, at least in the weaker form, a fundamental principle of modern states (for developments see Lecler 1955, Lutz 1977, Becker 1996). However, tolerance is not only necessary under the condition and in favor of a political pluralism. If not always immediately and distinctly perceptible, there are nonetheless differences among mankind: in need, talent, and interest; in taste and in sociohistorical origin; moreover, in the perception and interpretation of the personal or social situation in which one lives. Nor is one free of faults, prejudices, or mistakes. Therefore, at least passive tolerance is held to be a universally valid posture with regard to the civilized intercourse of mankind; it is through just such relations that the coexistence in marriage and family, in industry, in

society, in the government, even between states is made bearable as well as efficient. So, even an enlightened self-interest speaks in favor of tolerance.

However, the core dimension of tolerance is, with that aspect, not yet achieved. Its stronger form, the candid recognition of the other and otherness, is grounded in the dignity and the liberty of every human being. Therefore, tolerance consists not only of that forbearance of the peculiarities and weaknesses of one's fellow man which can be solicited from them and freely granted to them. As the coexistence of individually responsible persons with equal rights is not possible without reciprocal recognition, tolerance is a fundamental condition for the justice of mankind's coexistence in freedom; and, as a condition of justice, it is a claim that can be refused to no one.

For active tolerance, not merely the practical (moral-political) principle of liberty but also a theoretical interest in truth be adduced. Human perception being limited, the open debate of different opinions offers a better chance for establishing truth than the 'intolerant,' namely, dogmatic and of a conviction considered infallible (Mill 1989).

3. *Tolerance as Both a Virtue and a Principle of the State*

Religious liberty is only one example that brings to the fore the specific intellectual and political difficulties which have to be overcome in order for the principle of tolerance to become historically significant. This principle, which goes beyond religious liberty, takes primarily two different forms: tolerance as a personal attitude and as a political principle of the state and of civilization, both forms mutually supplementing and yielding one another.

Whoever affirms the self-determination (the right to life, the will to development, and the liberty) of the dissident is realizing the attitude of tolerance. Only when one has one's own convictions, can one speak of recognizing an 'other'; and only there, where one doesn't indulge in indifference, or in the type of cynical nihilism which plainly lets everything pass as valid, including even gross injustice, can one speak of this particular attitude. Tolerance does not mean the fig leaf, behind which moral indifference and intellectual weaknesses conceal themselves. It is a critical autonomy which does not even allow certain conflicts to arise in coexistence, especially not in situations of competition; and which does not go through the inevitable debates and quarrels with the inexorable severity of a religious war. Whoever is tolerant no longer pursues a life which is founded upon self-assertion through violent conversion, or through the surmounting of adversaries. He or she endeavors to establish a communality on the basis of civil equality and communication to which the following qualities belong: the ability to listen, the capacity to take others

seriously, and finally the willingness to let oneself be taught by new situations and new information. Tolerance is then accomplished by the willingness and ability to sympathize with the views and lifestyles of others.

Like pluralism, tolerance too is grounded in the following factors: the understanding that no one is entirely free from faults; the knowledge of the wealth as well as the perspectival limitations of every concrete instance of self-realization; and the insight into the limits of human knowledge. Tolerance, however, originates primarily in the recognition of the other as a free and equal person, who, on the basis of the inviolable dignity of mankind, has the right to form his or her own opinions, and to live according to them alone, or with like-minded individuals—provided that the exercise of this right does not encroach upon the same right of all others. The virtue of tolerance is one of the most important political virtues of a democratic citizen (Rawls 1993, p. 194).

As tolerance does not represent an end in itself, but rather a condition and an expression of the liberty and dignity of mankind, it terminates at the point where liberty and dignity are violated. Tolerance does not consist, as Marcuse (1965) fears, of the stabilization of existing social states, in the preservation of the status quo, and even less of the (masochistic) acceptance of injustice. Wherever the liberty and dignity of mankind needs to be protected, tolerance crosses over into accusation, criticism, and protest. The obligation to impose on no one a limitation of his or her liberty which is not imposed on all others in the same manner can be considered a criterion for the limits of tolerance. The highest principle of justice consists of this criterion of an equal liberty which is primarily spelled out in terms of human rights.

Because the justification of tolerance is found in liberty and justice, it cannot be left to the discretion of individuals and political powers. Its recognition is not only a voluntarily assumed attitude, but also a duty of justice, the implementation of which the legal order, the state, is responsible. It is therefore a political task to institutionally guarantee basic elements of tolerance, such as human rights. The first task of tolerance as a principle of the legal order and the state is therefore found in the positivation of human rights and their recognition as basic rights (personal right to liberty, political right to participation, and social and cultural rights), or, as the case may be, as fundamental aims of the state (liberty, democracy, welfare state).

As long as human rights are only recognized by coercion, tolerance remains nothing more than an acceptance accompanied by contempt toward those who think and live differently. In order to guarantee the stronger form of tolerance—coexistence based on mutual respect—tolerance as a duty of justice has to be acknowledged on its own terms; moreover, it must become through continuous practice a component of one's personal character, a genuine attitude.

Tolerance as a political principle is only accomplished when this duty of justice is freely accepted as a moral virtue of the citizen.

At the international level tolerance demands the mutual respect of different cultures and traditions, instead of valuing European culture or American lifestyle as the ideal of the development of all civilization and to devalue all other forms as primitive and barbarian. This type of respect was articulated first in Herder's (1774/1994) and later in Hegel's (1821/1970) theory of the 'spirits of nations' (*Volksgeister*), according to which different social and cultural patterns in which different peoples think and live assert themselves in the course of history. As these forms are coherent, they justify themselves to this extent.

One facilitates the recognition of other cultures' intrinsic value whenever one looks not only to that which divides but also to that which binds, that is, universal human attitudes such as readiness to aid, civil courage, and honesty. These commonalities are not only acknowledged as morally binding in the Christian West, but also in the most diverse of societies and epochs. The tolerance that began with religious liberty is accomplished only at the point, where the colorful abundance of personal and sociohistorical self-realization is freely acknowledged as both a task and a chance for creative humanity.

See also: Globalization, Subsuming Pluralism, Transnational Organizations, Diaspora, and Postmodernity; Human Rights in Intercultural Discourse: Cultural Concerns; Liberalism; Pluralism; Relativism, Pluralism, and Skepticism (Philosophical Perspective); Cultural Concerns; Tolerance; Value Pluralism

Bibliography

- Becker W 1996 *Anerkennung und Toleranz. Über die politischen Tugenden der Demokratie*. Palm & Enke, Erlangen
- Fetscher I 1990 *Toleranz. Von der Unentbehrlichkeit einer kleinen Tugend für die Demokratie*. Radius, Stuttgart
- Fraenkel E 1979 *Deutschland und die westlichen Demokratien*, 7th edn. Kohlhammer, Stuttgart
- Hegel G W F 1970 *Grundlinien der Philosophie des Rechts (1821)*. Suhrkamp, Frankfurt
- Herder J G 1994 *Auch eine Philosophie der Geschichte zur Bildung der Menschheit (1774)*. Deutscher Klassiker Verlag, Frankfurt
- Höffe O 1995 *Political Justice*. Polity Press, Cambridge, UK
- Höffe O 1999 *Demokratie in Zeiten der Globalisierung*. Beck, Munich
- James W 1909 *A Pluralistic Universe*. Longmans, New York
- Kelso W A 1978 *American Democratic Theory. Pluralism and its Critics*. Greenwood Press, Westport, CT
- Lecler J 1955 *Histoire de la tolérance au siècle de la réforme*, 2 Vols. Aubier, Paris
- Horton J, Mendus S (eds.) 1991 *John Locke: A Letter Concerning Toleration (1689)*. Routledge, London
- Lutz H (ed.) 1977 *Zur Geschichte der Toleranz und Religionsfreiheit*. Wissenschaftliche Buchgesellschaft, Darmstadt

- Marcuse H 1965 Repressive tolerance. In: Wolff R P, Marcuse H, Moore B (eds.) *A Critique of Pure Tolerance*. Beacon Press, Boston, MA
- Mendus S (ed.) 1988 *Justifying Toleration. Conceptual and Historical Perspectives*. Cambridge University Press, Cambridge, UK
- Mill J S 1989 *On Liberty (1859)*. Cambridge University Press, Cambridge, UK
- Oberreuter H (ed.) 1980 *Pluralismus. Grundlegung und Diskussion*. Leske und Budrich, Opladen
- Olson M Jr 1965 *The Logic of Collective Action*. Harvard University Press, Cambridge, MA
- Patschovsky A (ed.) 1998 *Toleranz im Mittelalter*. J. Thorbecke, Sigmaringen
- Püttner G 1977 *Toleranz als Verfassungsprinzip. Prolegomena zu einer rechtlichen Theorie des pluralistischen Staates*. Duncker & Humblot, Berlin
- Rawls J 1971 *A Theory of Justice*. Belknap Press of Harvard University Press, Cambridge, MA
- Rawls J 1993 *Political Liberalism*. Columbia University Press, New York
- Scheler M 1980 *Die Wissensformen und die Gesellschaft (1926)*, 3rd edn. Francke, Bern
- Schnapper D 1998 *La relation à l'autre. Au coeur de la pensée sociologique*. Gallimard, Paris
- Schultz U (ed.) 1974 *Toleranz. Die Krise der demokratischen Tugend*. Reinbek Rowohlt
- Thierrey P 1997 *La tolérance. Société démocratique, opinions, vertus*. Press Universitaires de France, Paris
- Walzer M 1997 *On Toleration*. Yale University Press, New Haven, CT

O. Hoffe

Polanyi, Karl (1886–1964)

1. Biographical Notes

Karl Paul Polanyi was born in Vienna on 25 October 1886 to Michael Pollacsek and Cecile Wohl. They had married in 1881 and had six children: Laura, Adolf, Karl, Michael, Zsófia, and Paul. Polanyi's father worked for the Swiss Railways and Vienna's Stadtbahn, before setting up on his own as a railway builder in Hungary. In the 1890s the family moved to Budapest where Karl Polanyi began his secondary studies in the model school of the teachers' training institute in 1898. At the turn of the century Polanyi's father's business went bankrupt and after indemnifying the creditors, the family was impoverished and he was forced to take a job abroad again. In spite of that, there was lively intellectual activity in the parental home. Polanyi's mother, Cecile Wohl, organized a literary salon attended, among others, by Oszkár Jászi, György Lukács and Karl Mannheim. After finishing secondary school in June 1904, Polanyi enrolled in the Faculty of Law in Budapest University. He lost his maternal grandfather and father in 1905, both of

whom had especially deeply inspired him. After attending courses in Vienna and Budapest he took part in organizing demonstrations for the progressive thinker Gyula Pikler, a professor of the Philosophy of Law in 1907. Polanyi obtained his Doctoral degree on 26 June 1909 from Bódog Somló at the Kolozsvár (Cluj) University. In 1908 he became president of the recently formed Galilei Circle, a student body set up to disseminate the scientific world view in readings and debates, without—as Polanyi later self-critically recalled—being politically involved. He qualified as a lawyer in 1912 and began to co-edit the review *Szabadgondolat* [*Free Thought*] in 1913. His writings argued for free education and were devoid of the influence of the church. In 1914, Polanyi was involved in establishing the Radical Party, of which he became the secretary.

However, the next year he was sent to the front and discharged two years later as disabled. During the Károlyi government of 1918, he was a member of the legal committee of the People's Council in Budapest and of the Students' Council. However, on the eve of the Hungarian Soviet Republic in 1919 he delivered a speech against the dictatorship of the proletariat, then immigrated to Vienna. He soon underwent a serious operation from which he took a long time to recover. He was working on a voluminous unpublished work, *Behemoth*, concerned with the destructive capacity of science. In this he says that the future is constantly being remade, therefore it can't give validity to our actions in the present.

In addition to Ludwig von Mises, Jakob Marschak, Felix Weil, and others, he contributed to the theoretical debate about the possibility of the planned economy in which he elaborated a proposal to solve the problem of 'Sozialistische Rechnungslegung' on the basis of the principles of guild socialism. He met a well-known figure of the revolutionary students' movement, Ilona Duczynska, in 1920. They married and had a child, Karoline Helene (Kari) in 1923. In Vienna, Polanyi first worked for Hungarian papers (Bécsi Magyar Ujság and Bécsi Kurir), and then from 1924 to 1933 he contributed to *Der Österreichische Volkswirt*. Having lost his job because of the rising danger of National Socialism, he immigrated to England where he first lived in privation, earning money by giving lectures. He was one of the masterminds behind the Christian Left movement, and concentrated on criticism of the essence of fascism.

This time he joined the Workers' Educational Association and the Extra-Mural Delegacies of Oxford and London, lecturing mainly in general economic history and social theory. His lectures focussed on the shift from *laissez-faire* to regulated capitalism, when state and industry began to interpenetrate, and interference of government with the captains of industry became regular.

In 1940 he became a British citizen, but from that year, he was resident lecturer at Bennington College in

the USA and toured America with lectures. He also worked on *The Great Transformation* first published in 1944 (Polanyi [1944] 1957). After returning to the UK during the war this book earned him the post of visiting professor at Columbia University, New York, in 1947. He lectured in economic history and co-ordinated research into the origin of economic institutions. These efforts were summed up in the posthumously published *The Livelihood of Man* (Polanyi 1977). He continued to do research with a team after his retirement in 1953, publishing their findings in *Trade and Market in the Early Empires* (Polanyi et al. 1957). The posthumously published *Dahomey and the Slave Trade* (Polanyi 1966) was another result of the cooperation.

The family settled in Canada as US authorities had denied an entrance visa to Polanyi's wife. In 1963 they visited Hungary, where Polanyi gave a lecture about the American Economic Sociology. He died of a stroke during the night of April 23, 1964 in Toronto, Canada.

2. The Substantive View

The major question of Polanyi's substantive view is how to identify the place of economies in societies (Polanyi 1947, 1977). He argues that economic activity is embedded in the social and cultural setting and that the predominance of price-making markets does not apply to most primitive, archaic, or even contemporary economies. He therefore taught that the changing place of economies could be best described by combinations of comparative concepts that shed light on the alternative logic of economic reproduction. Polanyi's substantive theory was discernible in its essentials in *The Great Transformation* (Polanyi [1944] 1957) and in the lectures he wrote in preparing it. The book, a critique of the market system, is a derivative of historical knowledge and skills accumulated in political journalism. This is borne out by a later remark of his saying he had two aims with the work he wished to: transfer the more combatant Austrian workers' culture into the English labor movement; and give an up-to-date criticism of capitalism, providing intellectual support to Roosevelt's New Deal. In Polanyi's view, the imminent problem of the market system was not the existence of markets but the turning land and labor into commodities by this system. The analysis of the so-called Speenhamland law, temporarily hindering the creation of a competitive labor market as well as of the ensuing welfare acts, revealed that welfare legislation failed to efficiently remedy the mass impoverishment paradoxically entailed by growth and could only delay the emergence of a labor market from masses of wage workers.

The book was written during World War II, parallel with Hayek's *Road to Serfdom* (Hayek [1944] 1976) and after Burnham's *The Managerial Revolution*

(Burnham [1941] 1960). While Hayek argued that deserting the road of self-regulation leads to fatal social deformation, Burnham envisaged the inevitable predominance of managerial-elitist hierarchies with the growing autonomy of control functions. Polanyi focussed on the double movement of *laissez-faire* economy and society. He derived the respective distortions from the market system and from the degeneration of society's defensive mechanisms. What constitutes the continuity between these and his earlier ideas was the criticism of the market system, the rejection of faith in advancement, the doubt about interpreting class interests as material interests, and the stress on non-anonymous historical processes resulting in the market system.

These ideas can be detected in the study 'Our obsolete market mentality' (Polanyi 1947) and in the volume *Trade and Market in the Early Empires* (Polanyi et al. 1957). However, the emphasis in the former is on the more accurate clarification of the concept of embeddedness, and in the latter, on the empirical analysis of primitive and archaic economic formations.

Embeddedness, according to the substantivist view, is related to the fact that an economic actor is a social being, with a mixture of motives, serving to attain social recognition and social goods. His economic efforts are therefore usually subordinated to these goals. Hunger and gain are not universal motives—Polanyi and his colleagues argued—thus the scarcity postulate, the cornerstone of formal economics, is inadequate. As is indicated in the title of Polanyi's key study, economy can be interpreted as an 'instituted process' in a substantive sense. The term process refers to 'location and appropriate movements,' that is production and transport on the one hand and transaction and disposition on the other.

Institution lends unity and stability to these processes, which can be approached via the forms of integration. Reciprocity presupposes the symmetry of groups, while redistribution is dependent upon collecting and distributing goods or disposal over them, hence it implies centrality. Exchange, in turn, depends on the existence of price-making markets. Not every exchange has integrative significance. Polanyi argued that the simple operational exchange and decisional exchange at a set rate do not integrate the economy, whereas bargained rate exchange does. The forms of integration are not arranged in phases, in stages of development as do the types of List or Bücher (Polanyi 1968); apart from the predominant forms, others are also present, constituting varying combinations. Trade, money, and markets are not necessarily functional upon exchange in the substantive approach, unless they are fitted in a market system. The ports of trade gift trade, and administered trade usually fall outside the market system.

Similarly, only one of the money uses coincides with exchange. The rest of the uses analyzed by Polanyi—

payment use, standard, ideal unit, operational device—do not necessarily tally with the market, being the typical features of the redistributive integrative form of the economy. The market, Polanyi sums up, is not so old a phenomenon, as are trade and money, and though it may crop up in history, it only integrates into a market system in exceptional cases.

While the former book contained an implicit debate with Hayek and Burnham, this one was challenged by *Economy and Society*, the book of Parsons and Smelser, the aim of which was to integrate economic and social theory and to find an alternative explanation on the place of economy in society (Parsons and Smelser [1956] 1957).

3. Polanyi's Influence

If it is contended with justification that Polanyi's views were deeply influenced by the writings of Bücher, Weber, Malinowski and others (Polanyi 1977), it must also be contended that Polanyi filtered these influences and thus conveyed them to late twentieth century social science. His ideas were discussed in several volumes, selections, and bibliographies (Stanfield 1986, Polanyi-Levitt 1990, Mendel and Salée 1991). Substantivism appears to be integrated in the history of sociology as a standard chapter (Hindess 1977).

The *Trade and Market in the Early Empires* volume, as well as the posthumous works edited mainly by his disciples (Polanyi 1966, Dalton 1968, Polanyi 1977) and the disputes coming to a head in economic anthropology suggest that a school was evolving around Polanyi. In economic history, Douglass C. North (North 1977) assessed Polanyi's influence in basically appreciative terms, though criticizing his market interpretation. Polanyi also influenced modern researches centering on the interrelation between state and economy as well as studies of socialist economy and its disintegration.

According to an evaluation, in spite of his marginal academic position, Polanyi was one of the most influential thinkers of the twentieth century. As far as the sociological analysis of the economy is concerned, this statement might be proved. Polanyi's concept of embeddedness became the starting point of the new economic sociology.

See also: Capitalism; Economic History; Economic Sociology; Economic Transformation: From Central Planning to Market Economy; Economics, History of; Economics: Overview; Hierarchies and Markets; Institutional Economic Thought; International Trade: Commercial Policy and Trade Negotiations; International Trade: Economic Integration; Market and Nonmarket Allocation; Market Areas; Market Structure and Performance; Markets and the Law; Marxian Economic Thought; Money, Sociology of; Policy

History: State and Economy; Regulation, Economic Theory of; Regulation Theory in Geography; Retail Trade

Bibliography

- Burnham J [1941] 1960 *The Managerial Revolution*. Indiana University Press, Bloomington
- Dalton G (ed.) 1968 *Primitive, Archaic, and Modern Economies. Essays of Karl Polanyi*. Anchor Books, Garden City, NY
- Hayek F [1944] 1976 *The Road to Serfdom*. The University of Chicago Press, Chicago
- Hindess B (ed.) 1977 *Sociological Theories of the Economy*. Holmes and Meier, New York
- Mendell M, Salée D (eds.) 1991 *The Legacy of Karl Polanyi. Market, State and Society at the End of the Twentieth Century*. Macmillan, Basingstoke, UK
- North D C 1977 Markets and Other Allocation Systems in History: The Challenge of Karl Polanyi. *Journal of European Economic History* 6(3): 703–16
- Parsons T, Smelser N [1956] 1957 *Economy and Society. A Study in the Integration of Economic and Social Theory*. Routledge and Kegan Paul, London
- Polanyi K [1944] 1957 *The Great Transformation*. Rinehart, New York
- Polanyi K 1947 Our obsolete market mentality. *Commentary* 3, February: 109–17
- Polanyi K, Arensberg C M, Pearson H W (eds.) 1957 *Trade and Market in the Early Empires. Economies in History and Theory*. The Free Press and the Falcon's Wing Press, Glencoe, IL
- Polanyi K (in collaboration with Rotstein A) 1966 *Dahomey and the Slave Trade*. University of Washington Press, Seattle, WA
- Polanyi K 1968 Bücher, Karl. In: Sills D L (ed.) *International Encyclopedia of the Social Sciences*. The Macmillan Company and the Free Press, New York, Vol. 2, pp. 163–5
- Polanyi K 1977 In: Pearson H W (ed.) *The Livelihood of Man*. Academic Press, New York
- Polanyi-Levitt K (ed.) 1990 *The Life and Work of Karl Polanyi. A Celebration*. Black Rose Books, Montréal
- Stanfield J R 1986 *The Economic Thought of Karl Polanyi. Lives and Livelihood*. Macmillan, Basingstoke, UK

G. Lengyel

Police and Politics

Sir Frederick Maitland pointed out in the late nineteenth century that the etymological relationship of the words police, policy, politic, polity, politics, and politicians is very close (Maitland 1885). All are derived from the Greek *polis*, roughly translated as city-state. The earlier English meanings of police reflected this common origin, since police could mean a constitution or a polity. In the eighteenth century, the word came to have a more specialized meaning as 'the civil force to which it is entrusted the duty of maintaining public order, enforcing regulations for the prevention and punishment of breaches of the law,

and detecting crime' (OED). This purely civil meaning was broadened (1837) to include 'any body of men, officially employed to keep order, enforce regulations, or maintain a political or ecclesiastical system.' The wider definition covers gendarmeries under military discipline and security services charged with the protection of public authorities.

The origins of the term have contemporary significance since unsuccessful attempts have been made, particularly in England, to divorce police from politics, and represent the police as nonpolitical guardians of society (Reiner 1992). The police and political authority are, however, intertwined—the police are executors of the decisions and the laws of duly constituted political authorities. In order to do so, they designated as having, in Max Weber's celebrated formulation about the state, the monopoly of the legitimate use of force. As Bittner (1979) remarks, the police 'are equipped, entitled and prepared to deal with every exigency in which force might have to be used.' The legitimate use of force by the police has been confined to the territories of sovereign states. The reality of a police monopoly of legitimate force within these territories has been contested because other agencies, such as prison officers, use physical constraints (Brodeur 1983). Other arguments contend that the actual use of force weakens the authority of the police by pre-empting negotiated solutions to conflicts and provoking sometimes serious opposition to the police (Monjardet 1996). Nonetheless, the concept of police necessarily includes the availability of legitimate force.

1. Police Systems

The long neglected comparative history of the relations between the police and politics has been reinvigorated by major studies since the 1980s (Bayley 1985, Liang 1992, Emsley 1999). In the eighteenth and early nineteenth centuries, professional police, paid by public funds, appeared. In the more developed parts of continental Europe, the purposes of the police were to secure the institutions of the state, enforce the law and the decisions of the ruler, and to protect propertied classes and property, both public and private, against criminals, calamities, and disturbances. In the English experience, except in London where the Metropolitan Police was established in 1829 by Act of Parliament, the modern police emerged from the office of constable, a medieval institution with equivalents elsewhere in Europe, who was charged with public tranquility at the local or parish level (Emsley 1996). In the United States, policing developed first at the local and state levels but, in the twentieth century, there was a proliferation of law enforcement agencies at the Federal level (Federal Bureau of Investigation, Drugs Enforcement Administration, Secret Service, etc.). However, the Federal agencies, in terms of numbers of police officers, constitute only a small proportion of the whole.

A distinction is made, sometimes too sharply, between an Anglo-American and a European 'continental' tradition of police-government relations (Mawby 1990). In general terms, hierarchically organized police forces controlled directly by a central government minister were characteristic of the most powerful continental European states. Mutual suspicion has often characterized police and politicians' attitudes on either side of the Anglo-American/continental European divide. The centralized systems have been regarded in Britain and the United States as instruments, often improperly used, in the hands of unscrupulous politicians. Misuse of police forces by the Nazis and other oppressive regimes gives this view credibility. In addition, national police forces are regarded as unwieldy and inefficient organizations. In the opposing European tradition, there has often been frank disbelief that police activities in decentralized systems can be professionally efficient, properly coordinated, and democratically controlled. They are also regarded as vulnerable to the influence of private interests, and as having uneven quality. Although much attenuated, this mutual suspicion has not entirely disappeared.

Differences between police systems have diminished because of a mutual learning process and because of internationalizing influences. Tendencies are apparent towards centralization in the British and some federal systems and to decentralization among the centralized continental systems. In the 1990s the British set up the National Criminal Intelligence Service (NCIS) to counter the increasing national and international mobility and activity of criminals. Local security plans, local coordination of policing, and more municipal police forces have modified the system of uniform central control. There have been increasing international police contacts and cooperation, and increased commonality of police methods and techniques, which, in turn, have an impact on police organization. The 'projection outwards' of police systems to the rest of the world, first through the European colonial empires and then US world influence after World War II, has resulted in hybrid systems and, as in Japan, some ingenious adaptations to local circumstances. Police systems are now distributed along a spectrum of types of organization, rather than based on contrasting models (Marenin 1996, Bayley 1996).

A greater degree of consensus than in the past exists among politicians in the advanced industrial democracies about the proper role of the police. Many of the less economically advanced countries have received assistance particularly from the United States, Britain, and France in police training and methods; they have, as a consequence, attempted to integrate, with varying degrees of success, philosophies about police-political relations. The consensus among the elites of the rich world (now including most electorally important parties of the Left) is that police forces exist to protect

the constitutionally public authorities, to uphold the law, and to protect liberties, rights, and property. Direct political interference in police operations is less and less acceptable. Left wing radicals, often influenced by Michel Foucault (Foucault 1995), dissent from this consensus and regard these beliefs as useful myths to uphold an obnoxious status quo.

2. Police Accountability

The elite consensus leaves intact the longstanding problems of how the police ought to be controlled and made accountable. These problems raise the most politically sensitive question concerning the police—'who guards the guardians?'—a question first asked in imperial Rome by Juvenal (*Satires* 6 1), to which there is no universally accepted answer. In recent years, public concern has been voiced about police malpractice, corruption, violence, and intrusion into privacy. The contemporary policy dilemma is how to preserve a degree of professional autonomy for the police, essential for operational efficiency, while making police forces accountable to elected authorities, accountable before the law, and professionally accountable for their competence (Walker 1998, Wilson 2000).

In liberal democratic systems, the principle of accountability to the law and judicial authorities is uncontroversial but the methods of implementing the principle vary between countries and within countries. The police, in general, are subject to the ordinary law but this has been found to provide inadequate redress for complaints about police malpractice. The result is special arrangements for investigating such complaints (Goldsmith 1991). These arrangements are always the focus of political controversy and allegations are often made that the police have operated above or outside the law (Skolnick and Fyfe 1993). The police usually have considerable operational discretion but there are wide variations in the extent to which they are directly controlled by the political authorities.

Systems of political accountability vary along two dimensions—the extent of decentralization and the closeness of control of the police by elected officials. In the continental European systems, direct hierarchical accountability to a minister who is, in turn, answerable to parliament seems to guarantee the clearest lines of democratic accountability. However, these systems tend to generate three kinds of police-political scandals. These are: allegations of improper instructions given to the police to further political or personal interests of a minister; the political covering up of improper actions by police officers, to avoid embarrassment to the government; and improper police actions encouraged by the government, such as a tough law and order policy.

In decentralized and federal systems, political accountability for different police forces is dispersed. In federations, federal, state, and local authorities control different police forces. Improper behavior by police officers or relations between politicians and the police does not, in these systems, tarnish the whole of the police institution. Police malpractice such as institutionalized racism, corruption, and excessive use of force can, however, become endemic locally. In a constabulary system like that in England where the Chief Constables have a high degree of professional independence, democratic accountability may be indirect (and ineffective) through nominated police committees. The greatest difficulties of accountability arise in societies with deep ethnic, social, or religious divisions such as Northern Ireland, Israel, and South Africa, as well as countries with large immigrant minorities with characteristics clearly distinguishable from the host society (Brogden and Shearing 1993, Brewer et al. 1988). In these circumstances, the police inevitably are regarded by the powerless as oppressive instruments of the dominant group.

3. *New Issues*

Certain policing issues, broadly connected with political accountability, emerged on the political agenda in the last third of the twentieth century, in the highly industrialized democracies. The women's movement has had considerable influence over the investigation and prosecution of certain forms of crime—rape, violence against women, and pedophilia. The movement has also accelerated the number of women police officers in all highly industrialized democracies. The great political importance accorded to drug abuse, expressed by Presidents Nixon and Reagan's 'war on drugs' and their European emulators, has resulted in considerable diversion of resources into the pursuit of drug traffickers and drug abusers, the development of new and controversial investigatory techniques (controlled delivery, money tracking, intrusive electronic surveillance), and international law agreements.

The introduction of 'value for money' financial accounting throughout government has raised new questions about police effectiveness. It has proven politically embarrassing for police agencies, as increased expenditures are rarely reflected in improved clear up rates for crime. The determination to control public expenditure has also created pressures to develop nonstate forms of policing, which had precedents in the period before the establishment of professional police forces. Modern forms were a prominent feature of United States policing since the nineteenth century; such was the fame of the private security firm, Pinkertons, that some Europeans thought it was an official American police department. But private security firms have now become a controversial feature of policing internationally. The

political left has been particularly critical of this development because it seems to provide an increased level of protection for the rich, but right wing supporters of national police forces are also very reserved about it (Johnston 1992).

4. *Transnationalization of Policing*

Developments in the international system must now be taken into account in the politics of policing because the police function is no longer completely constrained by the doctrine of territorial sovereignty. The establishment of international jurisdictions is changing, and has the potential to change radically the basis of police accountability. Protection of human rights has become a major priority of 'the international community.' International judicial protection for these rights is emerging and this erodes state sovereignty. France, for example, was condemned in 1999 by the European Court of Human Rights for torture following the brutal treatment in a police station of a Dutch citizen of African origin.

'Globalization' has led to increased concern about international crime in areas such as terrorism, drug trafficking, money laundering, international fraud, traffic in human beings, and disposal of high value stolen goods. International legal instruments have been negotiated, such as the 1989 Vienna Convention on drug trafficking which has direct effects on domestic law enforcement and criminal investigation techniques, such as 'controlled delivery.' Also, a certain transnationalization of police has occurred. This commenced as early as 1923 with the setting up of the International Criminal Police Commission, the predecessor of Interpol (Anderson 1989). US Federal law enforcement agencies have posted police officials abroad since the World War II—by the 1980s, the US Department of Drug Administration had over 60 offices in 40 countries (Nadelmann 1993). In Europe, Europol and the Schengen System (the compensatory measures of police and judicial cooperation for the abolition of border controls within the European Union) have gone much further in involving international regional institutions in police cooperation (Anderson et al. 1996). Transnationalization, together with the impact of the information technology revolution, undermines the traditional hierarchies of control (including political control) of the police. Policing is becoming a more knowledge-based, networking activity. In due course, this will transform all aspects of the relationship between police and politics.

See also: Criminal Justice, Sociology of; Justice and Law; Law: Economics of its Public Enforcement; Law, Sociology of; Leadership: Political; Police, Sociology of; Policing; Political Culture; Political Discourse; Prisons and Imprisonment

Bibliography

- Anderson M 1989 *Policing the World: Interpol and the Politics of International Police Co-operation*. Clarendon Press, Oxford, UK
- Anderson M et al. 1995 *Policing the European Union*. Clarendon Press, Oxford, UK
- Bittner E 1979 *The Functions of the Police in Modern Society*. Oelgeschlager, Gunn & Hain, Cambridge, MA
- Brewer D J et al (eds.) 1988 *The Police, Public Order and the State*. St. Martin's Press, New York
- Bayley D H 1985 *Patterns of Policing*. Rutgers University Press, New Brunswick, NJ
- Bayley D H 1994 *Police for the Future*. Oxford University Press, New York
- Brodeur J P 1983 High policing and low policing: Remarks about the policing of political activities. *Social Problems* 30(5): 507–20
- Brogden M, Shearing C 1993 *Policing for a New South Africa*. Routledge, London
- Emsley C 1996 *The English Police: A Political and Social History*, 2nd edn. Longman, London
- Emsley C 1999 *Gendarmes and the State in Nineteenth Century Europe*. Oxford University Press, New York
- Foucault M 1995 *Discipline and Punish*. Vintage Books, New York
- Goldsmith A (ed.) 1991 *Complaints against the Police*. Clarendon Press, Oxford, UK
- Johnston L 1992 *The Rebirth of Private Policing*. Routledge, London
- Liang H-H 1992 *The Rise of the Modern Police and the European State System from Metternich to the Second World War*. Cambridge University Press, Cambridge, UK
- Maitland F W 1885 *Justice and Police*. Macmillan, London
- Mawby R I 1990 *Comparative Policing Issues; the British and American System in International Perspective*. Unwin Hyman, London
- Marenin O (ed.) 1996 *Policing Change, Changing Police: International Perspectives*. Garland, New York
- Monjardet D 1996 *Ce que fait la Police: Sociologie de la Force Publique*. Editions de la Découverte, Paris
- Nadelmann E A 1993 *Cops across Borders*. Pennsylvania State University Press, University Park, PA
- Reiner R 1992 *Policing and Politics*. Harvester Wheatsheaf, New York
- Skolnick J H, Fyfe J J 1993 *Above the Law: Police and the Excessive Use of Force*. Free Press, New York
- Walker N 1998 Police accountability. *Encyclopedia of Applied Ethics* 3: 541–52 Academic Press, San Diego
- Wilson J Q 1989 *Bureaucracy: What Government Agencies Do and Why They Do It*. Basic Books, New York

M. Anderson

Police, Sociology of

Sociology of police as a specialized area of study can be defined as the empirical and theoretical analysis of those organizational agencies—usually but not exclusively established by state governments—that are charged with investigation and prevention of crime

and/or offences against formally adopted routines and rules as well as with maintaining public order, safety, and peace. It embraces the systematic sociological enquiry about the role of police in modern societies, the general characteristics of policing bodies and their relation to other institutions, the institutional development and internal structure of police forces, the collective and individual behavior of their members, and the interactions or relations of those agencies with individuals, different social groups, institutions, and the general public.

1. Police as a Matter of Sociological Concern

Both police and sociology are offsprings of modernity and progressively diversified societies. Although the modern contours of both became obvious only in the nineteenth century—the former as a new formal institutionalization of governance, the latter as an accomplishment of enlightenment and a new academic discipline—there is a line of discourse which can be traced back to ancient times, when Roman emperors struggled to get or keep their rule by using armed forces and when philosophers reasoned about the art of best governance. In spite of the fact, that the modern notions of ‘police’ and ‘sociology’ were not yet discovered then, the reasoning about the necessity of establishing and maintaining a public force to control crime and to ensure public order—and the critical discussion of its actual appearance—is as old as the discourse on the *res publica*.

Later on, the issue of police was embedded into the broader philosophical topic of the state and government—as it was for the early classic sociologists: police were discussed mainly as an instrument for the sovereign’s power to govern. Max Weber’s often quoted notion, that the state is that human community, that successfully claims a monopoly of legitimate use of force within a certain territory, still prevails within academic reflections on the police (Bittner 1991).

With the growing urbanization of the countries in industrial transition, there was also an increasing body of literature on police matters from a pragmatic perspective in the first half of the twentieth century. Yet it took some time for the police to become an explicit and original matter in sociological thought and empirical inquiry, primarily in English-speaking countries.

2. Determining the Subject

At first glance and seen from the surface it seems rather trivial to determine who and what constitutes ‘the police.’ One would expect that the police as a public institution is represented by (sworn) officers, i.e., representatives of the state and its government,

who are (often) specially trained professionals and are invested with certain powers (like the authority to search or arrest a person). Usually this 'apparent' kind of police will appear in public as officers on the beat, a crew in a patrol car, behind a desk in a police station, or as plain-clothes detectives, doing investigations in the field. This is the common and popular image (in the Western world) of what police actually are, transmitted by the media and supported by occasional ordinary encounters. From time to time this picture is widened by the appearance of riot police in full gear, underlining the state's authorization and capability of using legitimately physical force against disobedient citizens or rioting crowds.

When one looks again at who and what constitutes police, the semblance becomes more complicated: on the level of involved actors it becomes apparent that a good portion of the workforce employed by police consists of civilian staff, like secretaries or the clerks in the forensic laboratories; some sworn officers, such as staff who are responsible for conducting statistical analysis of criminal incidents or running a computer program for matching data, are rarely out in the field, while certain field officers, like undercover agents or specialists in charge of surveillance of telecommunication, often do not act openly as police. Furthermore there are other state agents who, having similar powers of investigation or intervention, perform certain functions of policing, but are not seen or labeled as 'police.' One might think here of custom officers, secret agents, public health inspectors, or prison guards. Still others appear in a policelike guise but are clearly not officers (at least lacking their full powers): city wardens, commercial guards and patrols, hired investigators, vigilante organizations like the 'Guardian Angels,' bodyguards, stewards in a football stadium, or bouncers.

As professionals trained to use force legitimately in the name of the state's power, police share this somewhat exclusive right with the military and other law enforcement officials (prison, customs). But on the level of routine work processes, this distinguishing feature is a rather rare event compared to the overall picture of police duties. Instead, research has shown that a considerable part of the modern police workload is not at all focused on crime control and investigation of criminal cases, but consists of responding to general emergency calls, mediating conflicts, regulating motor vehicle traffic, and communication with other institutions or agencies like social services, insurance companies, etc.

Although it might be clear from a commonsense point of view, what 'police' actually are—basically that formal institution which is vested with the powers and resources to respond to criminal acts or public order disturbances and calls itself 'police'—the subject in question gets more diverse, the more we see it from a perspective of a peculiar organized social activity rather than as a matter of institutionalism. Thus a true

'sociology of policing' would cover a wider area and embrace more organizations than the initial 'police studies.'

But even from an institutional approach one has to speak of 'the police' either from a very abstract level or from a single case point of view only. Besides the common similarities in terms of historical developments, organizational models, practical strategies and tactics, and legal accountabilities, every country's police system has its own particularities and unique arrangement of forces.

It should not be overlooked that police as a subject of sociological interest is linked in many ways with other social systems, all of them carrying their own, often overlapping, bodies of literature: as an instrument of executive governance 'the police' can be seen as a segment of the *sociology of the state*. Its quality of being an important part of the criminal justice system and its more or less explicit legal bindings does make it a component of the *sociology of law*. Last but not least, on the level of empirical studies there are clear ties to the field of *organizational sociology*.

3. Changes in Focus and Emphasis

Although some academic work on the police can be traced back to the 1950s, the first surge of sociological studies of police work occurred as an aftermath of the increasing social and political changes which affected most countries of the Western world from the early 1960s on. In a time when social and political conflicts were brought onto the streets, and rising concerns about crime and disorder became a public issue, 'the police became increasingly visible, controversial and politicized in response to these tensions and pressures' (Reiner 2000, p. 210).

According to Cain (1979), five types of studies, following a rough chronological sequence in the shift of areas of interest, can be identified for this period: (a) a law-related concern with civil rights, when confronted with police powers and discretion; (b) examinations of the police organization by participant observations (Skolnick 1966 and Manning 1977 are well-known examples); (c) studies of the role the police take in the social construction of deviants and the fabricating of public images of deviance; (d) analysis of the interactions and relationships of the police with the social environments they are working in; and (e) studies questioning the political contexts, presumptions, and consequences of police activities. It can be said that all these topics overlap in a way and have persisted during the following decades. There are two important points to note: one is that the main characteristics of this work, coming out of universities, were motivated by a concern with issues of police deviancy and generally tended to be critical of the police as a powerful institution of the state. The other

is that this research was created overwhelmingly in the USA and the UK.

In a second wave from the early 1980s on, police research was done more often outside the academic sphere. Governmental bodies like the Home Office in the UK began actively to do their own research on police issues. Furthermore think tanks and independent research organizations began to enter the field to offer their own agendas and expertise. This development enlarged and in a way altered the agenda of police research and with it the accessible knowledge on the sociology of police. Thus the momentum of studies changed—beyond the Anglo-Saxon universe—from an outside critique of police behavior and police institutions to questions of good practice and evaluation of reform measures, generated under policy and managerialist perspectives (Reiner 2000).

4. Current Theory and Research

It is still the case that the bulk of police studies which are relevant for the sociology of police in one way or the other is produced in English-speaking countries, with the USA in a clear lead. This might be seen as the reason why, on an international level, the discourse on police and policing issues mainly reflects on these countries. For example debates on 'community policing' or 'zero-tolerance strategies' originate from the development of police policies in the USA. But as matters of cross-border and international police co-operation have become more important since the early 1990s, there is a growing interest in comparative perspectives (Brodeur 1995, Findlay and Zvekic 1993).

The contemporary sociology of police and policing has to take two aspects into account. First of all, there are quite a lot of significant differences in how police institutions are organized and how policing is actually achieved between different countries. This is mainly due to the diverse national historical background. The diversity lies particularly in the arrangements of the police system influenced by legal, political, and occupational cultures (Busch et al. 1985, Monjardet 1996). The other aspect is that when it comes to the level of policing, i.e., how police work is actually achieved and what kind of resources and strategies are used, some striking similarities and convergences can be noted. There seems to be an increasing transfer of issues in the international police discourse.

The sociology of police is fed in a significant way by historical research on police institutions and forces, be it on the local or the national level. These studies show that there are certain issues in reflecting on police and policing, which have persisted for a long time, most prominent among them effectiveness, deviance, and accountability.

Contemporary research on police and policing embraces also immanent occupational and organizational aspects like recruitment and education, gender

and minority issues, the role of unions and professionalism and specialization. The question at stake here is what police are actually doing.

A good portion of police research literature is still concerned with issues of internal as well as external control mechanisms of individual and institutional police behavior. In this respect the institutional autonomy of the police in the face of the government and the judiciary is one side of the question, the control of the behavior of individual officers another (Marx 1988).

Increasingly matters of police management and evaluation are discussed, especially connected with institutional police reforms in the UK. Modernization seems to have been an enduring issue in police reform since the 1970s.

In theoretically oriented academic research the emphasis is clearly shifting from an institutional understanding of police towards policing as a socially structured, dynamic, and multilayered process. The background for this hidden paradigm shift is on the one hand the growing role nonstate policing bodies (security enterprises, insurance companies, vigilant citizens etc.) are gaining in many countries (Shearing and Stenning 1987) and on the other the increasing relevance of technology and information for the police (Ericson and Haggerty 1997).

5. Methodological Issues

The police is a specific institutional field and bears special problems for sociological research. Studies on police and policing in a sociological sense started to be done by university researchers. As long as police behavior could be observed from the outside, especially in respect to its effects on citizens, it was a matter of external description and analysis. In a way these studies could be seen as a type of 'legal journalism,' mostly case-oriented. A new field of empirical data was opened, when researchers were able to do participant observation of police work. With that, the internal perspective of the field actors were reflected in the research on police. Of course, there have always been some difficulties for police organizations granting outsiders access to internal functional mechanisms, because it carries the risk of 'occupational secrets' becoming subjects of public discussion.

Generally, from a methodological point of view three points deserve attention. First, police studies relevant for sociological interests can be roughly divided between pure empirical, pure theoretical, and those studies which combine or at least claim to combine both aspects.

Another issue is, who initiates and/or carries out the research in question. It might make a difference for methodological approaches, as well as for the broadness of the theoretical perspective, if studies are done by researchers coming from police or other govern-

ment institutions or whether they are carried out by university institutes. Furthermore there is a clear difference between research *for* and research *about* the police. Usually the former tends to be preoccupied with practical issues or management problems. Although there quite often exist interesting sociological aspects within these perspectives, a genuine sociology of police and policing is coupled with research *about* the police, i.e., its meaning in relation to the law, the state, or society in general.

Finally, the term 'police' might be a category too broad when it comes to empirical analysis and theoretical conclusions. Not only do national police systems embrace a variety of more or less specialized forces, with specific, sometimes unique ways of division of labor among units, but police organizations also refer to very different geographical ranges of jurisdiction. This situation becomes more complicated when nationally focused studies are compared on an international level: for a full comprehension of the differences and convergences national particulars have to be taken into account.

6. Future Directions of Theory and Research

There seem to be three major paths of future research which will contribute to knowledge about the sociology of police. One line of advancement will come from the shift that police institutions will be interpreted more and more in terms of the market. One issue will be what can be called the 'economy of police and policing.' In this regard commercialization of police functions will become more important, not only in terms of private police or privatization of policing but also in terms of adopting service orientation as a primary goal and private business structures within the framework of public police. Another research topic will be connected with the growing internationalization of police work. The emerging concept of liaison officers, the creation of multinational task forces—Schengen and Europol as well as Interpol arrangements on policing, etc.—are headed toward changes in particular in the relationship between policing and space. Finally, recently emerging approaches in maintaining order in transnational zones of disorder (i.e., the former Yugoslavia), established drug control policies, and antiterrorism policies, demonstrate the apparent convergence of internal and external security and safety concepts, and convergence between military and secret services on the one hand and police on the other.

See also: Control: Social; Crime: Sociological Aspects; Crime, Sociology of; Criminal Justice, Sociology of; Criminal Law and Crime Policy; Law: Economics of its Public Enforcement; Law, Sociology of; Police and Politics; Policing

Bibliography

- Bittner E [1980] 1991 The functions of police in modern society. In: Klockars C B, Mastrofski S D (eds.) *Thinking About Police*. McGraw-Hill, New York, pp. 35–51
- Brodeur J P (ed.) 1995 *Comparisons in Policing: An International Perspective*. Avebury, Aldershot, UK
- Busch H, Funk A, Kauss U, Narr W D, Werkentin F 1985 *Die Polizei in der Bundesrepublik*. Campus, Frankfurt am Main, Germany
- Cain M E 1979 Trends in the sociology of police work. *International Journal of the Sociology of Law* 7: 143–67
- Ericson R V, Haggerty K D 1997 *Policing the Risk Society*. Clarendon Press, Oxford
- Findlay M, Zvekic U (eds.) 1993 *Alternative Policing Styles: Cross-cultural Perspectives*. Law and Taxation Publishers, Deventer, The Netherlands
- Manning P K 1977 *Police Work: The Social Organization of Policing*. MIT Press, Cambridge, MA
- Marx G T 1988 *Undercover Police Surveillance in America*. University of California Press, Berkeley, CA
- Monjardet D 1996 *Ce que fait la police (Sociologie de la force publique)*. La Découverte, Paris
- Reiner R 2000 Police research. In: King R D, Wincup E (eds.) *Doing Research on Crime and Justice*. Oxford University Press, Oxford, UK, pp. 205–35
- Shearing C D, Stenning P C (eds.) 1987 *Private Policing*. Sage, Newbury Park, CA
- Skolnick J H 1966 *Justice Without Trial*. Wiley, New York

H.-J. Albrecht and D. Nogala

Policing

Public police are a bureaucracy created by political and legislative processes to enforce the statutory or common law of crimes, and to maintain public order and safety. Police in democratic states are accountable to the law and to democratically elected legislatures and executives. A 'social science' understanding of police draws upon theory and research in sociology, urban anthropology, law, criminology, history, and political science.

Social science research on policing scarcely existed before the 1960s. Moreover, observational and survey research on policing has largely been restricted to democratic regimes. Even in those, social science research may be constrained by police officials who limit access to investigators. Since the 1960s—a decade of rising crime rates, riot, and social protest—socio-legal research on police has flourished. An it has spread to democratic states around the world, including Canada, Eastern and Western Europe, Israel, Japan, Australia, India, Brazil, and South Africa. While earlier social science primarily sought to analyze police organization, role, and culture, later social science attempted to apply these insights to policing strategies and accountability.

1. Police Organization

Police structures vary. Some nations, like France, Italy, and Israel employ centralized police systems. The police in Japan, Australia, Great Britain, and Brazil are moderately decentralized. Police in the United States are decentralized, and include federal, state, county, urban, suburban, and rural forces. Nevertheless, like other bureaucracies, the police are everywhere shaped by social and economic conditions, by the interests and values of those who exercise or have access to state power, and a concern—or lack of it—for civil liberties and rights. Thus, police can differ significantly depending upon language, politics, tradition, and ideology.

1.1 Police Organization in Totalitarian States

In the first half of the twentieth century, aside from the militarized and centralized German police of the Nazi era, Soviet policing supplied a formidable model for policing a security state. From the beginning of the Bolshevik regime, the Cheka, the Soviet secret police, were given extraordinary powers to arrest, imprison, even to execute, in the interest of defending ‘the dictatorship of the proletariat.’ The detection and prevention of crime to provide public safety was secondary to the task of preserving the dominance of the political system. In contemporary China, although the police are striving for some independence, they have been regarded as a coercive instrument to serve the Communist party’s needs. Nations in transition from dictatorship to democracy have struggled to transform the police from their formerly authoritarian responsibilities to what would be acceptable in a democratic nation, where police are accountable to the political authority of elected officials.

2. Democratic Policing

Historians generally credit Sir Robert Peel’s London Metropolitan police to be the original, remarkable, and sustaining model for democratic policing. Yet the passage of *The London Metropolitan Police Act* of 1829 followed three-quarters of a century of deep suspicion and failed attempts to create a police force. Peel’s achievement was to coordinate London’s need for public order with democratic concerns and values. In 1826, before submitting the police bill, he abolished the death penalty for over 100 offenses, and removed numerous criminal penalties entirely. To win the support of the Whigs for his police bill, he temporarily excluded the City of London from the authority of the new police. To quell the fears of an emerging working class, he designed blue uniforms for the police to distinguish them from the army’s red coats; and, what

in retrospect must be seen as a brilliant and radical move, the men in blue did not carry guns while patrolling.

Furthermore, Peel and Colonel Charles Rowan, whom Peel appointed commissioner of the new police, strongly urged moderation in the actions of the police to gain public support. Although the purpose of the police was to protect private property and to maintain public order, it was understood that police power would be restrained by the rule of law, minimum use of force, and public opinion. As time advanced, and as class conflict diminished, English police were perceived to be efficient, incorruptible, accountable, and non-partisan.

All this was to change after 1960. Urban riots, rising crime, and police scandals combined to undermine the warm confidence the public had shown the British police for more than a century. ‘From almost complete invisibility as a political issue, policing has become a babble of scandalous revelation, controversy and competing agendas for reform’ (Reiner 1992, pp. 73–4).

British historians have offered revisionist, theoretically Marxist, interpretations of the origins of the English police, emphasizing their role in diminishing class conflict in the development of Great Britain’s industrial capitalism. Like their American counterparts, British sociologists have observed the power of subculture and political belief to mediate the influence of law on police behavior. Others, especially McBarnet, have implicated the legal structure—judges, senior officers, vague laws, case law, including those purporting to protect civil liberty—as the basic source of police deviance (McBarnet 1983).

2.1 The Police in the United States

No police department in America was so carefully planned and organized as the London Metropolitan Police. ‘The employment of police in municipal administration was governed not by theory but by convenience’ (Lane 1967, p. 221). Public order and rioting posed a more serious problem than crime in early nineteenth-century Boston, when the city established its first full-time police department in 1837. The precipitating event for the creation of the Boston police department was akin to a race riot, as roving bands of Protestants destroyed nearly every Irish Catholic home on Broad Street. Similarly, the creation of New York City’s police force reflected a growing intolerance for riots and disorder rather than a response to an increase in crime (Monkkonen 1981, p. 553). Nevertheless like their eighteenth-century London contemporaries, New Yorkers thought themselves engulfed in a rising tide of crime and disorder. Negro slaves made up one-fifth of New York City’s population and were ‘an unstable and unruly element often

treated with great brutality' (Richardson 1970, pp. 4–5).

Interracial and interethnic hostility have always posed a problem for policing in democratic nations as well as totalitarian ones. In the United States, the combination of immigration, expansion, and slavery stand behind contradictory images of America as a land of promise and of oppression. The themes of brutality against blacks, of discrimination against immigrants, of immigrant hostility toward blacks—all set against a turbulent background of social and class conflict—arise time and again in American history.

Over time, American cities witnessed a progressive exchange of functions between the citizenry and the police. Initially, the citizenry were responsible for guarding themselves against criminals through a watch system. As other agencies assumed responsibility for other areas of municipal administration, the police began to be increasingly concerned with crime. Not until the turn of the twentieth century did the American police fully resemble the blue coated forces of the present.

Twentieth-century changes in the American police reflected population growth and diversity in cities, changing social and political organization of communities and neighborhoods, and in technology available to police. During the first reform wave, from the late nineteenth century to the Prohibition-inspired 'roaring twenties,' progressive elites (such as Theodore Roosevelt and his political associates) attempted to wrest control of the police from corrupt and powerful political machines (such as Tammany Hall in New York City). They also instituted civil service examinations and other measures intended to insulate officers and commanders from political influence.

The second wave, in the 1920s and 1930s, reacted to police scandals associated with national Prohibition. The 1931 Wickersham Commission report (1931) documented shocking corruption and brutality in the criminal justice system and motivated police professionals, such as August Vollmer, to introduce reforms. The efficient, seemingly incorruptible, semi-military Los Angeles Police Department, of the 1950s and 1960s, headed by William H. Parker was hailed as a model of such 'legalistic' professionalism, although the police behavior in the Los Angeles (Watts) riots of 1965 was to suggest otherwise. That riot was not as hurtful as the 1992 riot following the Simi Valley acquittal, in a California court, of the police who beat Rodney King. But at the time it took the largest toll of any riot in American history and was followed by significant riots across the nation, most prominently in Cincinnati, Cleveland, Newark, and Detroit.

The Civil Disorder (Kerner) Commission (1968) found that, in the first nine months of 1967 alone, there were 164 disorders of varying intensity, of which 41, in 39 cities, were considered 'major.' The report singled out the police as the activating cause of the urban riots that had shocked and mystified the

American public. The 1960s was also a period of anti-Vietnam war protest, generating hostility between police and protesters.

A third reform wave followed, continuing into the twenty-first century. Social science theory and research was, and remains, implicated in this project. Two main subjects dominate: sociological theories of police role, culture, and discretion, based mainly on participant observation; and theories of police administration, especially community-oriented, problem-oriented, and quality-of-life policing, sometimes grounded in participant observation, but also based on theory and experimentation.

3. *Reorienting Police*

Historically, the police emerge as a significant, and sometimes controversial, agency of social control in periods of social and economic transition accompanied by crime and social unrest. In democratic societies, policing generates issues of how police should best be organized to deal with crime, and to be accountable to the communities they serve. Police researchers who had studied the 'professional' reactive model of policing in the 1960s and 1970s became increasingly skeptical of the capacity of police to satisfy community demands for public safety and dignity. Two decades of research suggested that motorized patrol scarcely deters crime, that detectives can solve only a small percentage of crime, that rapid response doesn't solve crimes, since by the time police are told of the crime, the offender has gone. Moreover, four US Presidential Commission reports (as well as Great Britain's Scarman Report) showed how supposedly neutral and detached policing styles could generate deep mistrust of police and estrangement from the communities they policed.

Three new approaches to police organization, values and purpose were articulated and eventually studied by social scientists in the attempt to reduce crime, increase public safety, and renew some degree of confidence in urban police: community oriented policing (Skolnick and Bayley 1986, Moore 1992), problem oriented policing (Goldstein 1990), and quality of life policing (Kelling and Coles 1996, Livingston 1997). Most social scientists agree that policing is only one of a number of causes affecting crime rates, and even then demands considerable organizational rethinking and practice (Blumstein and Wallman 2000).

3.1 *Community-oriented and Problem-oriented Policing*

By the 1980s, police relations with the communities they served had become a major issue for industrial democracies in Western Europe, Australia, New Zealand, Canada, Japan, and in the United States. James Q. Wilson had demonstrated that American police departments of the 1960s had adopted 'styles'

of policing—‘legalistic,’ ‘watchman,’ ‘service’—that greatly influenced their performance. ‘Community-oriented’ (COP) and ‘problem-oriented’ policing (POP) are best understood as styles of policing which, although widely discussed and praised, have not generated a consensus as to their meaning or effectiveness. Nevertheless, although there is no standardized set of precepts of COP, three elements are generally regarded as central: (a) that police and citizens are ‘co-producers’ of crime prevention; (b) that policing be decentralized, with focus on neighborhood groups cooperating with police; and (c) that patrol be reoriented, from vehicles to streets. In addition, COP seems especially sensitive to the police need for racial, linguistic, and gender diversity.

PO Policing is frequently associated, and sometimes confused, with COP and for good reason. Each stresses community involvement. Each is critical of traditional ‘incident oriented’ policing. POP suggests that police managers review ‘incidents’ in consultation with community members to focus on underlying causes. Furthermore, POP envisions the police executive as a skilled politician able to mobilize the larger community—the media, and organizations within and outside of government—in the service of addressing underlying problems.

3.2 *Quality of Life Policing*

In an influential essay, James Q Wilson and George Kelling (1982) maintained that disorder and crime are usually inextricably linked, in a kind of developmental sequence. Just as a broken window, if neglected, leads to more broken windows, minor disorders (such as littering, loitering, public drinking, panhandling, and prostitution), if ignored, invite more serious criminality. Applying the theory as chief of New York City’s Transit Police, William Bratton and his associates scrupulously removed graffiti, and cracked down on turnstile jumpers, panhandlers, and others violating subway regulations. The result was an astonishing reduction of subway crime and a comparably successful rise in ridership.

With the election of Rudolph Giuliani as New York City’s mayor, Bratton became Commissioner of the New York City Police Department (NYPD). He merged the Transit Police into the NYPD and introduced many of its precepts. In addition, he set up a sophisticated computer mapping system (COMPSTAT) to assist police commanders in allocating resources, and to hold them responsible for crime reduction. New York City crime rates subsequently declined sharply. Although proving cause and effect is difficult, since other factors might have contributed (improved economy, higher incarceration rates, changes in drug preference, demographic shifts) close observers of the police acknowledge that these police strategies reduced crime. Bratton’s successor,

Howard Safir, employed the strategies more aggressively. The aggressive policing, plus highly publicized instances of excessive force, led to hostile relations between the police and the communities they serve.

4. *Discretion, Culture, and Deviance*

Wherever social scientists have observed police in action, as patrol officers or detectives, in Amsterdam, London, Los Angeles, New York, Sydney, or Toronto, their orientation to the job appears to be influenced by a profusion of unwritten rules and understandings significantly influencing conduct and misconduct. Sociologists and criminal law trained observers of the 1960s were impressed by the vast amounts of discretion that police enjoyed. Police on patrol were seen ignoring minor thefts, drunkenness, disturbances, assaults, and malicious mischief in the interest of maintaining public order. These discretionary practices rendered a significant part of the criminal justice process invisible and unreviewable.

William Westley’s pioneering study of the Gary, Indiana police examined how the strains of the occupation—dealing with criminals, the insane, the intoxicated, often in an adversarial relationship to the public—influenced police conduct—and misconduct—on the job (Westley 1970), Skolnick (1996) built on Westley’s insights to develop a theoretical role construct, the policeman’s ‘working personality.’ Skolnick asserted that significant features of the police role—danger, authority, and accountability to superiors—were a precursor to understanding how police perceived and acted upon the world around them.

Bittner (1970) contended that the core of the police role is their capacity to employ force. Although police often aid sick and troubled people, Bittner asserted, their main involvement is to provide coercion where needed. By now social scientists have produced a substantial literature documenting the role strains, ordinary practices, and cultural understandings of urban police.

4.1 *Code of Silence*

Although subcultural understandings vary, the legendary silence code seems to be the one powerful norm uniting police across departments and nations. Maurice Punch, who studied police deviance and corruption in New York, London, and Amsterdam, writes that police operate by a code of silence which ‘dictates that you do not ‘rat on your mates,’ (Punch 1985). In Los Angeles, more than 20 police officers witnessed Rodney King’s beating. Those who administered it must have assumed that those watching would not come forward and would back up any stories they would tell; and that departmental super-

iors would believe them. Officers in the New York City Police Department's 70th Precinct did not protest when they saw Officer Justin Volpe marching around the station house, waving a broken broomstick stained with blood and feces, the residue of his sadistic anal attack on Abner Louima.

Testifying before the 1933 Mollen Commission in New York, which was investigating a pattern of corruption that had grown into vicious criminality, Officer Bernard Cawley affirmed the silence code, and the reason behind it: 'Just say if a cop decided to tell on me, his career's ruined. He's going to be labeled as a rat.' The code of silence may be that of aspect of police culture most frustrating to the general public and to officials responsible for police oversight, including prosecutors, judges, legislators, and high ranking police officials.

4.2 Excessive Force

Following his study of police violence in American, Caribbean, and South American cities, Chevigny concluded that police organized on a highly military model, in socially stratified societies which have slight regard for rights, and a tradition of private vengeance are more likely to permit police to use arbitrary violence (Chevigny 1995). 'Police brutality' is an evocative media term, but it has no legal referent. To ask someone to move on, to employ physical coercion, or to shoot to kill, are all appropriate police conduct, depending on circumstances. Nevertheless, since every arrest, every handcuffing, involves an imposition of force on an essentially unwilling person, departments develop subcultural understandings about how to deal with people who challenge police authority. The widely televised beating of Rodney King by Los Angeles Police Department officers revealed a use of force so brutal it shocked viewers around the world. Yet a suburban, predominantly white jury found that the force was not 'excessive' since King was allegedly resisting arrest. (In a second, federal trial, held in Los Angeles, the officers were found guilty of violating King's civil rights.)

A blue ribbon commission (headed by Warren Christopher to investigate the complicity of the LAPD in the beating found that the LAPD reflected a counter-productive organizational culture. Officers were rewarded for the number of calls they handled and arrests they made as well as for being 'hard nosed.' Based on its time-honored notion of 'professionalism,' the department, the Christopher Commission concluded, 'emphasized crime control over crime prevention and isolated the police, from the communities and the people they serve,' and evidenced a 'siege mentality' (Christopher 1991).

In New York City, a particularly sadistic assault on a Haitian immigrant, Abner Louima, followed by two shootings of innocent African-American men (Ama-

dou Diallo and Patrick Dorismond) led to angry protests against police use of force, especially deadly force. As in Los Angeles, critics pointed to the police culture and organizational strategies—especially 'zero tolerance' for street infractions—to explain the use of excessive force.

4.3 Culture and Corruption

Although researchers have distinguished among levels of police culture in management and street cops, usually, students of police culture have focused on street cops. Subcultures of police departments generate understandings about low-level deviance, such as when it is permissible to accept a favor, a free meal, a discount on purchased goods, or an outright money bribe. Corruption scandals have tainted the police in London, Amsterdam, and Brussels, as well as in the United States. Low level corruption can all too easily slide into a higher level, diluting the distinction sometimes made between 'clean' and 'dirty' money. That distinction was to break down in New York when the Knapp Commission found that officers had extorted money or narcotics from drug dealers, had failed to take enforcement action against narcotics criminals, and had given false testimony. Around a quarter of a century later, the Mollen Commission found that lying under oath was so routine the New York police had a word for it—'testilying.' A 2000 corruption scandal in the Los Angeles Police Department—which includes allegations of police shooting and killing drug dealers, beating them, stealing and selling drugs, evidence planting, false arrests, and perjury, has given the LAPD the unwelcome distinction of being the first major US police department to be monitored by the federal government.

5. Private Policing

With the development of mass private property in the United States—e.g., gated communities, malls, industrial and amusement parks—more than double is spent annually on commercial private security (\$90 billion) than on public policing (\$40 billion). Other countries, like South Africa, maintain an even higher ratio of private to public police. In light of this growth, private policing has come to have significant implications for theories of social control in democratic societies.

Although private policing sometimes resembles, overlaps, and is confused with public policing, private policing is distinctive in one major respect. Unlike public police, private police are primarily accountable to their employers, not to the citizenry. Thus, private policing is ultimately grounded in different interests, in turn influencing its understanding of its role and practices.

While public police may act as 'peacekeepers' they are nevertheless authorized and expected to enforce the criminal law, which, at least in theory, imposes punishments based on just desert. Private police, by contrast, primarily enforce norms instituted by their employer. Thus, private police usually address violations of institutional standards of comportment (dress codes, noise levels); try to prevent material damage to property (through theft, vandalism, or terrorism); and seek to avoid disruption of routines (disorderliness) of the sites for which they are responsible—such as universities, hospitals, offices, industrial parks, or airports. To achieve these goals, security firms typically adopt a *preventive* approach to misconduct—which community- and the problem-oriented police do as well—thus blurring the distinction. Indeed, Ericson and Haggerty (1997) have argued that, like private police, modern police have largely become risk managers rather than crime controllers.

In the United States, laws protecting ownership of property permit private clients to define the meaning of security and order within their property boundaries more narrowly than public police, modified only by constitutional and legislative limits protecting free speech, minorities, and women. For example, behavior acceptable in a public park, e.g., walking barefooted, may be unacceptable in a restaurant or private place of amusement, such as Disneyland. The typical sanction for violating private property rules in *exclusion* or *ejection*. Low-level criminal behavior, such as smoking a marijuana cigarette, may also result in ejection or the threat of ejection, rather than the punishments embedded in criminal statutes.

Private police in the United States are not constrained by constitutional protections (e.g., the Fourth Amendment), nor are they limited by the same kinds of civil liabilities (e.g., §1983 civil rights claims) to which public police are subjected. Private security typically derives its powers from, and to an extent is controlled by, the laws of property, contract, and labor. Nevertheless, though many states do regulate the licensing of guards and their ability to use firearms, private policing in the USA remains largely self-regulating.

6. Controlling Police Misconduct

Procedural law, especially under the US Constitution's Fourth (search and seizure) and Fifth (self-incrimination) Amendments, sets boundaries for police who are investigating crime by excluding evidence obtained in violation of these safeguards. Nevertheless, such protections have been narrowed by the US Supreme Court, especially in connection with drug enforcement.

Since street law enforcement involves stopping, questioning, sometimes searching and arresting, as

well as ordinary peacekeeping, complaining citizens usually report one or more of the following: excessive force (hitting, pushing, tight handcuffs); abuse of authority (ordering change of location); rudeness; and ethnic slurs. Citizen Review Boards have become increasingly relied upon to address such issues and are found in Canada, Australia, Northern Ireland, and the United States. In the USA more than 80 percent of major cities have introduced such organizations. Walker (2001) concludes that policy review and change is more important than officer punishment to raise departmental performance, particularly since it is difficult to resolve cases where an officer and a complainant offer conflicting accounts.

External monitors, Walker believes, should prove more effective than citizen complaint review in elevating police performance. Civil liability suits, and most recently, federal and state oversight, have also been initiated to control police departments which appear to have deviated from their mission, either through corruption, excessive force, racial discrimination, or all three. There are no perfect models for policing or for controlling police. There is more agreement by social scientists and the public on what should not happen—corruption, brutality, racism—than on the mechanisms for controlling police. Nevertheless, certain themes related to police accountability are evident in the social science literature. First, police organizations are influenced by ideology, the most significant being a commitment to political democracy. Yet even within democracies, policing is influenced on a daily and regional basis by politics and power relationships. Second, policing is strongly influenced by the sub-cultural norms of policing. Third, courts possess only slight capacity to constrain police discretion on streets, neighborhoods, or management offices. Fourth, victimless crime enforcement breeds corruption, brutality, and even racism, as when minorities are racially profiled to enforce drug laws. Finally, police are more effective and accepted to the degree that they involve communities in developing their policies and practices.

See also: Criminal Justice, Sociology of; Law: Economics of its Public Enforcement; Police and Politics; Police, Sociology of; Professions, Sociology of

Bibliography

- Bayley D H 1985 *Patterns of Policing*. Rutgers University Press, New Brunswick, NJ
- Bayley D H 1991 *Forces of Order: Policing Modern Japan*. University of California Press, Berkeley, CA
- Bayley D H 1994 *Police for the Future*. Oxford University Press, New York
- Bittner E 1970 *The Functions of the Police in Modern Society*. National Institute of Mental Health, Chevy Chase, MD
- Blumstein A, Wallman J (eds.) 2000 *The Crime Drop in America*. Cambridge University Press, New York

- Chevigny P 1995 *Edge of the Knife: Police Violence in the Americas*. The New Press, New York
- Christopher W 1991 *Report of the Independent Commission on the Los Angeles Police Department*. Los Angeles, CA
- Dixon D 1997 *Law in Policing: Legal Regulation and Police Practices*. Clarendon Press, Oxford, UK
- Fu H L 1994 A bird in the cage: Police and political leadership in post-Mao China. *Policing & Society* 4: 277–92
- Ericson R V, Haggerty K D 1997 *Policing the Risk Society*. University of Toronto Press, Toronto, Canada
- Fogelson R M 1997 *Big City Police*. Harvard University Press, Cambridge, MA
- Goldsmith A J (ed.) 1991 *Complaints Against the Police: The Trend to External Review*. Clarendon Press, Oxford, UK
- Goldstein H 1990 *Problem Oriented Policing*. Temple University Press, Philadelphia, PA
- Hungarian Helsinki Committee (ed.) 1988 *Police in Transition: A Survey of the Police Forces of Bulgaria, the Czech Republic, England and Wales, Germany, Hungary, Poland, Romania, Slovakia, Yugoslavia*, CD-Rom. Hungarian Helsinki Committee, Budapest
- Ianni F, Reuss-Ianni E 1983 Street cops versus management cops: The two cultures of policing. In: Punch M (ed.) *Control in the Police Organization*. MIT Press, Cambridge, MA
- Kelling G L, Coles C M 1996 *Fixing Broken Windows: Restoring Order and Reducing Crime in our Communities*. The Free Press, New York
- Kerner H 1968 *Report of the National Advisory Commission on Civil Disorder*. US Government Printing Office, Washington, DC
- Knapp W 1972 *Report on Police Corruption*. George Braziller, New York
- Lane R 1967 *Policing the City: Boston, 1822–1855*. Harvard University Press, Cambridge, MA
- Livingston D 1997 Police discretion and the quality of life in public places: Courts, communities, and the new policing. *Columbia Law Review* 97: 551–672
- Manning P 1977 *Police Work*. MIT Press, Cambridge, MA
- McBarnet D J 1983 *Conviction*, 2nd edn. Macmillan, London
- Miller W R 1977 *Cops and Bobbies*. University of Chicago Press, Chicago
- Mollen M 1994 *Anatomy of Failure: A Path for Success*. Commission Report, City of New York, New York
- Monkkennon E H 1981 *Police in Urban America: 1860–1920*. Cambridge University Press, Cambridge, UK
- Moore M H 1992 Problem solving and community policing. In: Tonry M, Morris N (eds.) *Modern Policing*. University of Chicago Press, Chicago
- Muir Jr K W 1977 *Police: Streetcorner Politicians*. University of Chicago Press, Chicago
- Reiner R 1992 *The Politics of the Police*, 2nd edn. Harvester, London
- Richardson J F 1970 *The New York Police: Colonial Times to 1901*. Oxford University Press, New York
- Rubinstein J 1973 *City Police*. Ballantine, New York
- Shearing C, Stenning P 1982 *Private Security and Private Justice: A Review of the Policy Issues*. The Institute of Policy Studies, Montreal, PQ
- Shearing C, Stenning P 1984 From Panopticon to Disney World: The development of discipline. In: Doob A, Greenspan E (eds.) *Perspectives in Criminal Law*. Canada Law Books, Aurora, ON
- Sherman L 1978 *Scandal and Reform: Controlling Police Corruption*. University of California Press, Berkeley, CA
- Skolnick J H 1966 *Justice Without Trial: Law Enforcement in Democratic Society*. Macmillan, New York
- Skolnick J H 1969 *The Politics of Protest: A Task Force Report Submitted to the National Commission on the Causes and Prevention of Violence*. Simon and Schuster, New York
- Skolnick J H, Bayley D H 1986 *The New Blue Line: Police Innovation in Six American Cities*. The Free Press, New York
- Skolnick J H, Fyfe J J 1993 *Above The Law: Police and the Excessive Use of Force*. The Free Press, New York
- Stone C E, Ward H H 2000 Democratic policing: A framework for action. *Policing and Society* 10: 11–46
- Szikinger I 2000 The Roma in Hungary. *Policing and Society* 10: 91–120
- Tonry M, Morris N (eds.) 1992 *Modern Policing*. University of Chicago Press, Chicago
- Walker S 2001 *Police Accountability: The Role of Citizen Oversight*. Wadsworth, Belmont, CA
- Westley W 1970 *Violence and the Police*. MIT Press, Cambridge, MA
- Wickershan G W 1931 *Report on Lawlessness in Law Enforcement*. National Commission on Law Observance and Enforcement. US Printing Office, Washington, DC
- Wilson J Q 1968 *Varieties of Police Behavior*. Harvard University Press, Cambridge, MA
- Wilson G Q, Kelling G L 1982 Broken windows: The police and neighborhood safety. *The Atlantic Monthly* 249 (March): 29–38

J. H. Skolnick

Policy History: Origins

1. Introduction

Public policy is as old as the history of human governance. All rulers, irrespective of the form of public rule, have sought to shape and manage the substantive order governed under their authority. Modern public policy composes a narrower domain. Co-extensive with the complex, and continuing, process of state formation (and de-formation) begun in late medieval and early modern Europe (Tilly 1975), public policy has come to refer more specifically both to what governments do as they transact with civil society, the economy, and states within a global state system and to the creation and deployment of knowledge about these sites of authoritative transactions. The semi-independent origins of policy knowledge and policy activity constitute my subject. I focus primarily on two formative moments in the West: (a) the early modern period marked by the growing concentration of state sovereignty over territory and people, and (b) the midnineteenth through to the early twentieth century when the ‘social question’ and unaccustomed levels and forms of international violence pressed to the fore. ‘Origins’ itself a word with

a composite meaning. It refers both to the beginnings of a process and to the constituent elements of what comes next. The results of decisions taken at the critical junctures I discuss, set in motion developmental pathways that further defined the contours of modern public policy.

2. *Early Modern Europe*

The very existence of public policy in the modern sense can be traced to sixteenth and seventeenth century postfeudal Europe, a period marked by the accumulation and concentration of sovereignty into national territorial states (Watkins 1934). These units claimed unique authority within their borders and recognized boundaries demarcating states from each other (in this way, international relations became distinguishable from domestic politics). Each State possessed sovereignty based on law, and with it legitimate force within a distinctive territory, an ensemble of institutions, and a vision and articulation of the common good. With these attributes, the State emerged as a calculating actor vis à vis other States, the newly separated economic sphere, and a newly distinguishable civil society (Skinner 1989).

This consolidation of sovereignty into increasingly centralized units produced the standardization of laws, coinage, taxation, language, and responsibility for security, all of which were requisites for systematic public policy. A corollary process was that of a growing set of structural separations within national boundaries distinguishing the realms of sovereignty, property, and civil society. Now, states could not simply impose their will despotically. They had to create and deploy policies to transact and coordinate with other powerholders. State–society relations now became reciprocal as policy came to be synonymous with the outputs of governments seeking to regulate an increasingly challenging environment (Mann 1988; Tilly 1990; Spruyt 1994).

The modern economy likewise was the product of this postfeudal differentiation of spheres. The key point in the development of capitalism as the dominant framework for economic development came only with the concentration of sovereignty and the establishment of an authoritative framework for property rights and economic transactions on a large scale. Capitalism prospered and urbanization (as well as rural proto-industrialization) accelerated once states emerged that were not merely extortionist but shared an interest in creating the conditions required to organize independent market transactions. Whereas early medieval face-to-face exchanges had taken place in towns whose existence depended on grants of autonomy from local lords, thus creating insecurities outside this tight embrace, by the sixteenth century, kings, in exchange for revenue, protected towns and traveling merchants

on a much larger scale. With states providing a framework for capitalist development and not acting merely as rent-seekers, the linkage between the economy and the polity was utterly transformed into a new kind of strategic game. States did not ensure just any property rights, but only those consistent with state interests; capital did not give obeisance to any state, but only to those that secured its economic activities. The terms of this relationship thus became the first of two fundamental pivots of modern domestic politics in the West (North 1985).

The second was the hinge between the State and civil society. Separated from property, and concentrated in authority, the States of postfeudal Europe had to forge new, and uncertain, ties to civil society. These exchanges were marked by contestable questions about the range of activities over which the State would have authority—that is, the extent of its *hubris* in penetrating and regulating civil society—and about the ways in which members of civil society could affect the activities of the State. With the breakup of the tightly knit juridical, economic, and social units characteristic of feudalism, states could not simply compel by imposition. Instead, it was a condition of their effectiveness that they transact with society and coordinate aims with ‘private’ powerholders.

State–society relations thus became more reciprocal and less arbitrary; over time, the very ability of the State to utilize its authority to intervene in society produced commitments that subsequently bound the State. (Pitkin 1969; Downing 1992; Ertman 1997) The terms of these exchanges became the objects of political struggle and the subjects of expert consideration by political theorists and by political economists. Hobbes, Locke, Montesquieu, Rousseau, Millar, Steuart, the Physiocrats, and Smith, amongst a cast of many, mostly lesser, others, theorized these relationships and sought to find rules of interested engagement between the relevant parties. (Hirschman 1977) At stake was both the scope of responsibility of states for economic and moral situations and the ways in which the actions of states could be shaped by people, interests, and values outside the State itself.

The last key zone of transaction was international. Based on sovereignty understood as exclusive control over a particular territory and its people, a new international system came into being that substituted modern states with claims to unique control for formerly more fluid and overlapping jurisdictions. This change did not occur all at once, and its dimensions and decisiveness were masked for quite some time by the existence of dynastic empires, but the new global order of sovereign actors transformed the meaning of international policy about war and peace. Now, offense was geared to the extension of exclusive geographical sovereignty; defense to its protection. Contests over borders were settled by force or by diplomacy, both of which required new kinds of social knowledge (Doyle 1997).

These three sets of transaction were arrayed in a hierarchy, with the international dimension of war and peace designed to secure or extend sovereignty first amongst them because the existence and protection of such exclusive claims was the premise of the very existence of domestic policy. International policy also defined ambitions for overseas empires and for the guardianship of commerce in the globe's new Europe-centered market system (Zolberg 1981). As conflict between and amongst states now took on a global cast, so, too, did specialized policy knowledge about subjects as wide-ranging as military organization and the human qualities of 'primitive' subject populations.

Producers of modern policy knowledge were those intellectuals who came to understand and manage these new sets of transaction (Bauman 1992). Such policy intellectuals, after Machiavelli, created a modern science of government with two main ends: to point rulers towards more effective transactions among the State, economy, and civil society, and to police, educate, and civilize the population in order to improve its character and render it more rational and governable. They counseled how to expand the policy role of governments to address the security and portability of property under broadly mercantile economic arrangements and doctrines and how to govern populations subject to the authority of the State. Guided by such policy knowledge, as the territorial scope of states either expanded or became more secure, as the infrastructural capacities of governmental bureaucracies grew (including their abilities to gather and deploy statistical and other forms of information), and as a communications revolution in print and in advances to transportation took hold, governments took on increasing responsibilities for the solution of collective problems. These included new roles for both local and national authorities in such areas as poor relief, sanitation, and schooling (de Swaan 1988). Each time a new zone of policy activity was established under public aegis, either complementing or supplanting the responsibilities of private religious or secular responsibility, new specialist managers and intellectuals fashioned distinctive sets of policy ideas to organize these State-based activities. This extension of responsibility transcended even such fundamental divides as the *ancien régime* and post Revolutionary France (Woloch 1994).

In the long epoch from early modern Europe to the nineteenth century, mercantile thought produced a powerful and dominant recipe for state-building and for rules to govern international economic transactions, to shape those joining the State to the domestic economy and civil society. Its main aspects included 'regulation of the economically strong, support and direction for the economically weak, and the State's own enterprise where private initiative is wanting' (Krieger 1963, p. 557; Anderson 1974). At its early moments mercantilism was put to use for

conservative purposes to buttress a social order in crisis; later, under the initiatives of Colbert, mercantilism went hand in hand with the protection of property rights and the reduction of transaction costs to provide a dynamic basis for capitalist economic development. Both facets were integral to the process of State-building in which:

regional rulers utilized provisions for welfare as well as force to extend their control over all the inhabitants of their realms. Indeed, it was precisely the need for such a politics of welfare over a larger area when the old local welfare arrangements were being broken down that went far to establish the ruler as the head of the State. (Krieger 1963, p. 557)

3. *The Late Nineteenth and Early Twentieth Centuries*

During the first half of the nineteenth century, these formulas were pressured by the emergence of independent labor markets of more than local and regional scope (Polanyi 1944; Agnew 1986). With the introduction of such mechanisms to allocate people to jobs and set wage levels, capitalism entered a new phase. Concurrent with the development of labor markets (as well as autonomous housing markets) that helped generate new working class identities, dispositions, and patterns of collective action (Katznelson and Zolberg 1986), the period was marked by the development of compelling conceptions of popular representation and citizenship defined by the American and French Revolutions. All members of civil society now became actual or potential citizens. No longer could citizenship be restricted to those who, by holding property, demonstrated a stake in the existing commonwealth. Instead, the possibility of political (and military) participation was universalized, in the expectation that in a community of citizens people would act rationally in pursuit of individual, group, and common interests. Perhaps more so at this time than at any moment since, citizenship now possessed a radical, emancipatory edge (van Gunsteren 1985).

Within this field of tension, public policy was transformed radically in range and content. In this novel circumstance the 'social problem' emerged as a key category of public policy. This emblematic term condensed the new tensions of markets and citizenship. It put at the core of public life the question of whether, and to what extent, the political relations of citizens might modify the operation of markets, as well as the extent to which the rationale of the marketplace would dictate limits to citizenship. Prerevolutionary approaches to citizenship (whether Roman, medieval, or Enlightenment) had not ignored the connection between citizenship and minimum levels of security, but they had dealt with it by excluding those without economic means from the community of citizens. No longer was this solution possible. As citizenship

became more inclusive, states extended the scope of their policies aimed at organizing markets and mitigating their distributional inequities. Further, a modernist policy intelligentsia, nourished by Enlightenment conceptions of reason, considered citizenship, markets, and their intertwining with the confidence that they could find scientific solutions to ethical problems. Their methods became more specialized, and the professionals overtook the amateurs.

In situations characterized by the crystallization of the new institutions of markets and citizenship, practitioners of social knowledge perforce could accept or repudiate these foundations of policy. Throughout the nineteenth and early twentieth centuries, deep reactionary currents of refusal and restoration, radical renunciations of the new moral economy (Hirschman 1991), and forward-looking blueprints, the most important of which was a powerful Marxist critique and teleology, assertively rejected the new institutional and normative framework of liberalism. By contrast, the great majority of social analysts accommodated with varying degrees of enthusiasm to the new order and to its immanent power relations, and they sought to solve problems within the liberal doctrinal and institutional premises.

Even within this framework, there existed a wide array of alternative conceptualizations, ways of organizing knowledge, and contestations about policies, from right to left (at both poles straddling the liberal/non-liberal divide); but these debates, no matter how hotly contested, came to be circumscribed within a distinctive family of possibilities and they focused on a shared subject matter: the field of tension created by the simultaneous development of transactions between states and markets, especially as they concerned labor, and between states and actual or prospective citizens within civil society. The harsh realities of unequal and exploitative relations between economic actors helped create contentious class representations that intersected the radical impetus of citizenship to produce incendiary political possibilities, making both the modern State and modern capitalism precarious.

In these circumstances, producers of social knowledge—as individual scholars and theorists, and in organized settings such as universities, learned societies, political parties, organizations devoted to policy goals, and governmental bureaucracies—struggled with how to interpret and manage the new transactions linking State and market and State and citizen with very mixed and often confused motives: to preserve the social order and to reform it, and frequently both. Sometimes cast in a language of socialist transformation, sometimes in a conservative discourse, and sometimes in a vocabulary of individual rights, the center of gravity of the new social knowledge in the West was located in various attempts to find reference points and tools of administration for a form of social policy which could integrate conceptions of individual

and collective goods. The policy intellectuals who undertook this quest were not utopians. Most saw no way to go back to a moment before citizenship and (labor) markets (indeed, in Britain, neither progressives nor conservatives in the late nineteenth century sought to return to the pre-1834 epoch, when social policy and economics were treated autonomously), and no way forward to a postliberal era within the ambit of normal politics (Harris 1990; Mandler 1987).

The features of policy and knowledge tended to be nationally-specific. In France, an optimistic anticlerical and republican program of the pupils of Comte and then of Durkheim articulated with the Third Republic. In Britain, New Liberalism conquered the field, displacing and reshaping virtually all its competitors; as it did in the United States, within the framework of Progressivism. In Germany, like so much else, social science and policy studies suffered the fate of the country's precarious liberalism, developing especially skittish qualities (Lepines 1988; Ringer 1969).

4. Modern Policy Knowledge and Practice

The place of pride in social analysis in the *ancien régime* had belonged to the moral sciences. Classical political economy had been elaborated within its embrace. In Britain, figures as diverse as Stewart, Malthus, Ricardo, the Philosophic Whigs, John Stuart Mill, and Bagehot, sought to maintain holistic portrayals of modern society while treating the specificity of the new transactions between its distinctive domains, and to do so by making the role of the State, in form and in policy, the centerpiece of their analyses. These pivotal thinkers were neither traditional political philosophers who deduced the character of the good society from first principles or moral ends, nor were they empirical social scientists who sought to comprehend the regularities governing the new transactions.

Over the course of the nineteenth century into the twentieth, this kind of relatively integrated social knowledge was replaced by professionalized, discipline-based attempts to grapple with the specific tensions inherent in the elaboration of markets and citizenship in liberal societies. At stake was the character of the extended responsibilities and jurisdiction of the State with respect both to markets and citizens. From the vantage point of the transactions between the states and markets, the central issue was how to shape the State's interventions to make markets function effectively and to make their distributional effects tolerable. From the perspective of transactions between states and citizens, at issue was the balance between freedom and domination. From the standpoint of international relations, the key issue was how to regulate the quest for power and plenty in a

manner that enhanced the prospects of peace and prosperity. Together, these questions elaborated the themes of moral philosophy under radically revised conditions (Thomas and Meyer 1984).

These were the concerns of a new genre of policy institution, the first example of which is Britain's Social Science Association founded in 1857. Created to consider the problems of social economics and to promote legislative reforms in the House of Commons with regard to education, law, penal policy, public health, and the regulation of markets, the Association self-consciously viewed itself as a conduit between politicians and an expanding citizenry with political rights, and it came to play an important interstitial role in the Victorian State by proposing ways to manage the 'social problem' attendant upon the simultaneous existence of labor markets and an expanded citizenry. Like Germany's *Verein für Sozialpolitik*, the Association fused reform and social science in its attempt to provide expert knowledge on behalf of a project its members conceived of as an endeavor of social integration. Working before the development of separate social science disciplines and before the creation of an autonomous and politically ambitious labor movement, the Association succeeded in joining under liberal auspices a wide array of intellectuals as well as Liberal MPs, trade union and cooperative movement leaders, and other reformers in a quest to bring expertise to bear on the project of creating a new liberalism capable of surmounting the now manifest inadequacies of traditional political economy (Goldman 1986).

This subsequent transformation in subject matter, style of work, and institutional organization was no simple linear development. There were two overlapping moments, each characterized by an important partition. In the first, spanning the 1860s through the 1910s, the knowledge community divided into specialists concerned with one or the other transactions of State, economy, and citizenship. Integrated knowledge separated into distinctive zones for the study of economics, politics, and society, ultimately devolving into wholly separate professional disciplines, each with distinctive intellectual tendencies (Oleson and Voss 1979). The second moment, from the 1890s through the 1920s, was characterized by a trend complementary to these new disciplinary divisions: a split between academic practitioners of the social sciences, located mainly in burgeoning institutions of higher education, whose policy concerns and impact were indirect, and scholars whose emphasis was placed far more directly on policy matters and public influence.

As late as 1870, political science was hardly distinguishable from political economy and history. In the next 50 years, political science secured a disciplinary status divorced in large measure from both. It came to focus in a specialized way on constitutionalism, law, political parties, interest groups,

voting and other mechanisms of representation; that is, on the rules of State-citizen transactions. During the same period, political economics shed its adjective and its claim to be an aspect, in its science, of moral philosophy. Recast as a scholarly craft in the mathematical language of post Jevons marginalism, it was not so much that economics no longer wished to concern itself with larger questions of human nature and social organization. Rather, the economists of the period sought to elaborate a science of markets on the basis of an elegant simplification about just these issues, but not for the sake of science alone. This new economics, for whom the key figure certainly was Alfred Marshall, was created in the service of the ethical goals of the period's New Liberalism. Under his remarkable influence, deductive work within economics replaced historical scholarship. This narrowing of focus was accompanied by a widening of theoretical ambition to create an economics as a timeless and institutionless body of knowledge (Soffer 1978).

Indeed, Marshall was an exceptionally important figure at two junction points: between classical political economy and disciplinary social science; and between the latter and the world of public policy. His *Principles of Economics*, published in 1890, explicitly sought to place pioneering analytical techniques at the service of human improvement. While Marshall's economics was neoclassical in the sense that all its aspects concerned the fluctuating equilibrium of supply and demand, this equilibrium depended for him on the subjective decisions of producers and consumers in pursuit of qualitative time-bound goals in pursuit of a better life. Once his new mathematical tools became central to economics, the discipline developed in ways that elaborated technique at the expense of his own value-laden preferences. The new Marshallian economics deepened its scientific capacities at the expense of its practical effects on policy issues of the day and on its ability to analyze, or even take an interest in, the rules of transaction governing the relations of State and market in particular historical settings. Just at the moment when, in the late nineteenth and early twentieth centuries, the social tensions of markets and citizenship intensified under the impact of economic depressions, the radical transformations in industrial size and class structure entailed in the second industrial revolution, international rivalries for markets, a new geopolitics characterized by imperial ventures, and, ultimately, by total war, the academic discipline of economics lost a good deal of its capacity to function as a practical policy science. In part this was a matter of a distancing from practical affairs, in part a flight to abstraction, in part a deliberate decision in favor of the ambitions of science at the expense of the ambitions of moral discourse (Church 1974; Nelson 1987).

By the end of the nineteenth century, classical political economy virtually had disappeared, as had

the gentlemen generalist historians and lawyers who had become archaic figures. These 'all-rounders' were steadily replaced by academic scholars in specialized social sciences and by a growing body of policy specialists who claimed knowledge that both drew upon and crossed over the boundaries that divided the social sciences. Labor experts, social workers, organizers of social surveys, public administration specialists, reformers concerned with social policy, and intellectuals within social movements and political parties—both inside and outside the administrative apparatuses of states—tried to come to grips with the intensification of the problems inherent in interactions between States, citizens, and markets. These policy intellectuals were a novel and distinctive breed. Their main orientation was not to the social science disciplines as such; nor were they amateur seekers after a science of legislation in the manner of their mid-nineteenth century predecessors.

In spite of its abstraction and distancing from practical affairs economics, and economists, were not left without influence. On the contrary, the new scientific economics achieved something of an encompassing contextual role with regard to the development in this epoch of a new kind of policy-relevant social knowledge that coexisted with, and drew upon, university-based social science disciplines. Economics set limits to common sense, by putting on a 'scientific' basis the assumption that market outcomes are likely to be the most optimal, at least in the sense of the most efficient. But if economics emerged secure in this role, policy analysis (unlike the policy studies of today grounded primarily in microeconomics) principally adopted the language of public administration and social engineering. This language, nestled within the assumptions of Marshallian economics, facilitated a self-image for policy intellectuals much like that of the earlier liberals of the Social Science Association which stressed the role of social knowledge in helping to transcend class conflict within a framework that accepted, at least implicitly, the divisions between State, economy, and civil society. The central concern of these policy intellectuals was how to build a State capable of dealing with the challenges of increasingly assertive working classes both as economic actors and as poorly integrated citizens. The focus of these efforts pivoted on what came, only later in the 1940s, to be called the Welfare State, to encompass the cluster of policies and institutions concerned both with the mitigation of market distributions and with the substantive content of citizenship. To one side of this new policy knowledge lay the preliberal conservatism of a Bismarck who had sought to make workers a new estate subservient to and dependent on the State; to the other, lay the strong socialist assertion that only the transcendence of markets and the integration of the domains of State, market, and citizen could resolve the contradictions of modern capitalist societies. The new liberal social knowledge decisively rejected these

alternatives. Public policy today in the West is encompassed within these boundaries.

5. Origins as Consequences: A Post World War II and Post Cold War Postscript

Totalitarianism aside, what is striking is just how encompassing this orientation to public policy proved to be, and how dominant the policies promoted by practitioners of liberal social knowledge had, by the end of World War II, become broadly based on the model established under conditions of extreme economic and geopolitical duress by the American New Deal (Ruggie 1982). With Fascism defeated and Communism the new enemy, this expanded liberalism faced no serious competition in the normal politics of the western democracies. This is not to say that consensus replaced conflict; rather, conflicts about the contours of domestic issues took the form of debates about this or that essentially social liberal alternative, even in situations where some of the most important contestants for power presented more radical alternatives for the long haul. From quite far left to quite far right, political competitors accommodated to liberal principles and institutions. Political contests became no less vigorous, but certainly less fundamental: not capitalism versus socialism or democracy versus totalitarianism (though this antinomy defined the discourse of competition between the Soviet and American blocs), but a politics about rules of transaction between spheres within capitalism and democracy. Coherent repertoires of positions about these questions pivoting on the degree to which the State should be assertive in organizing markets and in utilizing social policy for distributional goals and on how institutions of political participation and the representation of interests should be organized distinguished 'left' from 'right' formulas about State-market and State-citizen transactions. Each position was given voice by its own identifiable policy intellectuals, but together they constituted a single, inherently liberal, knowledge community.

The benefits secured for liberal societies by the robust development of social knowledge by this kind of policy intelligentsia ought not to be gainsaid, especially in light of this century's cruel alternatives. But we should also be aware that by conceptualizing their problems at the level of economic and political market exchanges, most policy intellectuals have left unexamined a host of issues about class and State, inequality and power. By working at the level of transactions between states and markets and states and citizens and by separating these from specialized inquiry by students of international relations, they tacitly have pushed the macrostructures of capitalism and states into the background of their work. In so doing, they have underscored the great normative achievements of liberalism, but also became ac-

complices in the silences deep-seated in the liberal project.

See also: Capitalism; Citizenship and Public Policy; Civil Society, Concept and History of; Democracy; Economic History; Governments; Liberalism; Marshall, Thomas Humphrey (1893–1981); Moral Sentiments in Society; Nations and Nation-states in History; Political History (History of Politics); Political Science: Overview; Property Rights; Public Policy Schools; Rule of Law; Socialism; Sovereignty; Political; Welfare State

Bibliography

- Agnew J C 1986 *World's Apart: The Market and the Theatre in Anglo-American Thought, 1550–1750*. Cambridge University Press, Cambridge, UK
- Anderson P 1974 *Lineages of the Absolutist State*. New Left Books, London
- Bauman Z 1992 'Love in adversity': on the State and intellectuals, and the state of the intellectuals. *Thesis Eleven* 31 (March)
- Church R L 1974 Economists as experts: The rise of an academic profession in the United States. In: Stone L (ed.) *The University in Society. Vol. 2: Europe, Scotland, and the United States from the 16th to the 20th Century*. Princeton University Press, Princeton, NJ
- de Swaan A 1988 *In Care of the State*. Oxford University Press, New York
- Downing B M 1992 *The Military Revolution and Political Change: Origins of Democracy and Autocracy in Early Modern Europe*. Princeton University Press, Princeton, NJ
- Doyle M 1997 *Ways of War and Peace: Realism, Liberalism, and Socialism*. W. W. Norton, New York
- Ertman T 1997 *Birth of the Leviathan: Building States and Regimes in Medieval and Early Modern Europe*. Cambridge University Press, New York
- Goldman L 1986 The Social Science Association, 1857–1886. *The English Historical Review* 110 (January)
- Harris J 1990 Economic knowledge and British social policy. In: Furner M, Supple B (eds.) *The State and Economic Knowledge*. Cambridge University Press, New York
- Hirschman A O 1977 *The Passions and the Interests: Political Arguments for Capitalism Before its Triumph*. Princeton University Press, Princeton, NJ
- Hirschman A O 1991 *The Rhetoric of Reaction: Perversity, Futility, Jeopardy*. Harvard University Press, Cambridge, MA
- Katznelson I, Zolberg A (eds.) 1986 *Working Class Formation: Nineteenth Century Patterns in Europe and North America*. Princeton University Press, Princeton, NJ
- Krieger L 1963 The idea of the welfare state in Europe and the United States. *Journal of the History of Ideas* 24 (October–December)
- Lepines W 1988 *Between Literature and Science: The Rise of Sociology*. Cambridge University Press, Cambridge, UK
- Mandler P 1987 The making of the new poor law redivivus. *Past and Present* 117 (November)
- Mann M 1988 *States, War, and Capitalism*. Blackwell, Oxford
- Nelson R 1987 The economics profession and the making of public policy. *Journal of Economic Literature* 25 (March)
- North D 1985 Transaction costs in history. *Journal of European Economic History* 14 (Winter)

- Oleson A, Voss J (eds.) 1979 *The Organization of Knowledge in Modern America, 1860–1920*. Johns Hopkins University Press, Baltimore
- Pitkin H (ed.) 1969 *Representation*. Atherton, New York
- Polanyi K 1944 *The Great Transformation: The Political and Economic Origins of Our Time*. Rinehart & Company, New York
- Ringer F 1969 *The Decline of the German Mandarins: The German Academic Community, 1890–1933*. Cambridge University Press, Cambridge, UK
- Ruggie J 1982 International regimes, transactions, and change: Embedded liberalism in the postwar economic order. *International Organization* 36 (Spring)
- Skinner Q 1989 The State. In: Ball T, Farr J, Hanson R L (eds.) *Political Innovation and Conceptual Change*. Cambridge University Press, Cambridge, UK
- Soffler R N 1978 *Ethics and Society in England: The Revolution in the Social Sciences, 1870–1914*. University of California Press, Berkeley, CA
- Spruyt H 1994 *The Sovereign State and its Competitors*. Princeton University Press, Princeton, NJ
- Thomas G M, Meyer J M 1984 The expansion of the state. *Annual Review of Sociology* 10
- Tilly C (ed.) 1975 *The Formation of National States in Western Europe*. Princeton University Press, Princeton, NJ
- Tilly C 1990 *Coercion, Capital, and European States*. Blackwell, Oxford, UK
- van Gunsteren H 1985 Notes on a theory of citizenship. In: Birnbaum P, Lively J, Parry G (eds.) 1985 *Democracy, Consensus, and Social Contract*. Sage Publications, London
- Watkins F M 1934 *The State as a Concept of Political Science*. Harper & Brothers, New York
- Woloch I 1994 *The New Regime: Transformations of the French Civic Order, 1789–1820s*. W.W. Norton, New York
- Zolberg A 1981 Origins of the modern world system: A missing link. *World Politics* 33 (January)

I. Katznelson

Policy History: Regimes

The concept of regime has multiple cognate meanings. Most broadly, it is used as a synonym for 'political system,' connoting the most general institutions and processes characterizing the exercise of political power. More narrowly, when power is so tightly centralized and controlled that the dictator seems to constitute the entire political system, regime is used to refer to the paramount political elite, as in the Franco regime of Spain or the Nasser regime of Egypt. In international relations, regime refers to the set of principles, norms, rules, and decision-making procedures that actors mutually recognize and adhere to in their interactions. The core meaning of regime shared by these usages refers to the formal and informal rules and institutions that both reflect and shape the distribution and organization of political power and that constrain to various degrees the actions of power-holders. In this article, regimes and the

major subtypes of regimes, democracies, or dictatorships, are discussed in the context of state building and with reference to the concepts of state and government. Next, research into the origins of regime types is surveyed, with special reference to the causal role of public policies. Finally, the causal role of regime types on public policy is considered.

1. Regimes and State Building

The topic of regimes assumed increasingly greater importance in the politics of the West in the seventeenth and eighteenth centuries as state building extended the infrastructural power at the disposal of central rulers. Infrastructural power (Mann 1984) is an attribute of states, and refers to the capacity of state agencies to monitor society, regulate social and economic activity, sanction certain forms of collective action while prohibiting other types, adjudicate disputes, and engage in other activities dedicated to extracting human and fiscal resources to reproduce and enhance state power. Despotic power, on the other hand, refers to varying degrees of formal and informal constraints on rulers; as an attribute of regimes, it measures the capacity of rulers to act without consulting and bargaining with other political, economic, and social groups. While in many historical empires despotic power was virtually unlimited, the practical effect of the absence of constraints was acutely mitigated because, weak in infrastructural power, the state lacked the logistic capacity to implement the ruler's whim. State formation in Western Europe threatened to combine high levels of infrastructural power with high levels of despotic power.

European state formation, however, was accompanied by the structural differentiation of the realms of politics, economics, and civil society. Over time, although operating at a variable pace in different political units, the separation between public and private became more and more evident in both theory and practice. While structural differentiation created new opportunities for rulers to extract physical and material resources, rulers increasingly had to bargain and consult not only with traditional political actors and organizations, but also with representatives of new classes and organizations. Extracting resources, moreover, soon imposed new demands for state coordination and regulation to match the increasing complexity of economy and society. Struggles over the content of public policies such as taxation and military conscription became, by the nineteenth century, intertwined with struggles over the form of the regime itself. Over the long term, state power was both massively enhanced and, via changes in the regime, the disposition of state power by rulers was made subject to new constraints and redirected to new aims.

Regimes are composed of two forms of constraints on despotic power. On the one hand, rulers may share

the political stage with other collective actors possessing political power, such as parliaments, independent judiciaries, and even legally autonomous executive agencies. On the other hand, and perhaps more importantly, rulers may be compelled to negotiate with actors in civil society, such as capitalists, large landowners, religious organizations, labor unions, and diverse interest groups, who possess resources that can be converted into political power. The distribution of power (the relative strength of actors) and the organization of power (the institutionalized forms of interaction) act as constraints on despotic power and thus constitute regimes.

The concept of regime thus falls midway between those of state and government. State refers to the relatively permanent organizations of territorial-based rule used to generate power via coercion and administration; government refers to incumbents holding offices that deploy that power. Regime refers to the organization of power generated by the state and exercised by rulers and other state officials. It thus defines the routinized relations between rulers and other agents within the political apparatus, on the one hand, and between the state and civil society, on the other hand. Critical components of the regime are the formal and informal mechanisms for selecting leaders, procedures for representation and consultation, and rules by which decisions are made.

2. Types of Regimes

The semantic field of regime types is not timeless, but has instead been composed of those types considered legitimate contenders in distinct historical epochs. Thus, in ancient Greece, political theorists judged the best regime from among democracy, oligarchy, and monarchy. In post-feudal Europe, the main alternatives were absolutist versus constitutional regimes, distinguished by whether monarchs exercised both executive and legislative powers or shared legislative powers with an assembly. During the Cold War, regimes were trichotomized into democratic, authoritarian, and totalitarian regimes. With the collapse of the Soviet Union and the transition to democracy made by both authoritarian and former communist countries, current research treats almost exclusively democratic and authoritarian regimes.

Democracy is an example of what philosophers call an 'essentially contested concept,' a concept that straddles the divide between empirical and normative criteria. These normative criteria are evident in the proliferation of subtypes of democracy characterized by rival contentions about the proper goals of democracy, ranging from the defense of individual rights and freedoms of liberal democracy to the extension of rights to the economic and social realms in social democracy (Gutman 1993).

Wishing to avoid the complex debates surrounding this array of goals to be pursued by a regime in order for it to be called a democracy, most scholars engaged in empirical work on political regimes in the post-World War Two period have subscribed to a strictly institutional definition of democracy. In an early and still influential statement, Joseph Schumpeter (1942, p. 269) identified close links between democracy and capitalism, and defined the former as an 'institutional arrangement for arriving at political decisions in which individuals acquire the power to decide by means of a competitive struggle for the people's vote.' A minimal procedural definition, however, runs the risk of committing the 'electoral fallacy:' the failure to recognize situations in which civilian political leaders gain access to office via popular mandate but do not truly represent the electorate; because they win elections via corruption and manipulation of the political process; because they do not exercise *de facto* control over non-elected military officers and state officials; or because civil society does not have the informational and organizational capacity to monitor government actions and express dissatisfaction with policies. In any of these scenarios, elections do not produce genuine accountability of elected officials to the electorate. In many circumstances, then, it may be difficult to distinguish analytically between the institutional procedures of democracy and the goals democratic practices are designed to achieve. Indeed, insofar as both politicians and the electorate understand elections as an opportunity to exchange loyalty for material benefits and readily assent to these practices, then elections may be understood as opportunities to express mutuality and reciprocal solidarity, not as a means of holding governments accountable. In this scenario, votes may truly reflect genuine preferences of the electorate yet, given the cultural construction of the meaning of elections, it is not clear whether to label this regime democratic (Schaffer 1998).

One way to bridge the institutional and the goal-oriented approaches is to return to the definition of regime and consider democracy as a constraint on power. Modern democracies emerged in national political systems that were products of centuries of state formation, producing far greater infrastructural power than any political systems in the past. Barring constraints on despotic power, rulers of modern states would be capable of achieving their preferences by directly imposing them on all other actors expressing rival preferences, by firmly controlling the processes by which decisions are made to pre-empt the expression of rival preferences, or even to shape the consciousness and preferences of less powerful actors. Consequently, unconstrained rulers with infrastructural power at their disposal would be relatively certain of prevailing in conflicts with other power-holders or actors in civil society. Moreover, insofar as the structures of modern states are relatively durable, unconstrained rulers would have relatively permanent power. Finally, in

the absence of constraints, power could be deployed in any and all realms of society, polity, and economy.

Movements for democracy were efforts to contain the arbitrary exercise of power and to redirect its use toward new goals by making governments accountable to and representative of an electorate whose membership was enlarged in the nineteenth century and swelled in the twentieth century. These goals were achieved by placing constitutional limitations on the actions of governments, establishing rights of citizenship, and licensing citizens to participate in the formation of governments and the formulation of policies. As a result, despotic power was sharply constrained: outcomes of conflicts over access to office or policies enacted while in office became increasingly uncertain, and any advantage enjoyed by political leaders became both increasingly temporary and subject to constitutional limitations. The basic principles of democratic regimes, in other words, are that power can be exercised in only a delimited realm of areas; the outcomes of all political conflicts are uncertain; and, therefore, governments are temporary (Przeworski 1991, Lawson 1993).

Making these principles of uncertain, temporary, and restricted exercise of power concrete requires a range of specific institutional arrangements that permit civil society to express preferences and monitor government adherence to them, and that permit an opposition to challenge governments in power with a genuine possibility of becoming the government. Eight institutional criteria are necessary for citizens to have the opportunity to formulate preferences, communicate them to governing authorities, and have those preferences considered in the conduct of government: freedom to join and form organizations; freedom of expression; the right to vote; the right to hold office; the right of political leaders to compete for support; alternative sources of information; free and fair elections; and institutions for making government policies depend on votes and other expressions of preference (Dahl 1971).

By focusing on the institutional arrangements that make the exercise of power relatively restricted, uncertain, and temporary, this definition treats all nondemocratic regimes as a homogenous residual category. In fact, via reference to more specific attributes of the distribution and organization of power, a number of subtypes of authoritarian regimes can be identified, including but not limited to autocratic, totalitarian, absolutist, traditional, monarchic, oligarchic, and sultanistic regimes. By the same token, democratic regimes can—and for many purposes should—be disaggregated into numerous subtypes. Examples of 'modified' democracy in which some features of full democracy are absent include 'controlled' democracy in which parties representing some segments of the political spectrum are banned from competition, and illiberal democracy in which civil liberties are incompletely recognized.

3. *Origins of Regime Types*

The major cleavage in theoretical debates over the origins of democracy and dictatorship divides structuralist accounts that assign primary causal weight to situational factors from agency-based accounts emphasizing the role of political actors who are relatively unconstrained by environmental factors. Structural explanations can in turn be subdivided into nomothetic accounts positing quasi-universal trajectories of socioeconomic change affecting regime types and historically specific explanations. Among the latter, of particular importance are theories that view regime types as functions of particular constellations of public policies.

3.1 *Universal Preconditions*

Building on ideas that originated in ancient Greek political thought and that can be traced through the Enlightenment, the most influential theories of regime change in the post-World War Two period have been those that understand economic development and cultural change to be important, if not absolutely necessary, influences on democratization. Studies relying on increasingly refined conceptualizations, reliable data bases, and sophisticated statistical techniques have largely replicated the findings of earlier, relatively crude empirical studies: a strong relationship exists between the level of socioeconomic development—essentially, increasing wealth accompanied by multiple correlates—and democracy. A number of scholars have suggested that this relationship verges on the status of a sociological law.

Several analytic defects, however, motivate scholars to propose alternative theoretical positions. First, questions remain about the causal mechanisms underpinning the statistical relationship between prosperity and democracy. One of the earliest statements of this hypothesis briefly alluded to a number of causal pathways by which increased wealth would promote the spread of values of moderation and toleration, thereby immunizing polities from extremist politics (Lipset 1959). The articulation of increasingly sophisticated statistical studies linking socioeconomic development to democracy has not been accompanied by commensurate analysis of the intervening causal mechanisms. There is yet no systematic evidence that economic development is causally related to democracy primarily via changes in values. Consequently, many studies of political culture and democracy reject the claim that pro-democratic changes in values are necessarily preceded by socioeconomic development. Instead, the cluster of attitudes seen as conducive to democracy has been attributed to political learning, political socialization, institutional change, international diffusion effects, and habits learned from the

routine practices of politics itself (Diamond 1993). Even these sources of democratic norms of moderation and tolerance, once decoupled from economic development, have been only contingently tied to democratization.

A second problem stems from the ambiguous specification of the causal relationship between wealth and democracy: rising wealth has been hypothesized both as helping to sustain democracy, but also as helping to generate democracy. Statistical tests have frequently conflated the two issues, despite warnings by prominent scholars to maintain the distinction and treat the two issues separately. When statistical tests are designed to discover the exact nature of the empirical relationship, the best evidence is that heterogeneous factors account for the origins of democracies; once established, rising levels of wealth help to stabilize democracy without causing it (Przeworski and Limongi 1997).

Finally, in response to democratic breakdowns in the 1960s and 1970s, scholars criticized the universality of the claim, restricting its validity to the earlier developers in Europe and Anglo-American countries, and contending that modernization in later developers would issue in diverse trajectories of political change. In one variant of this position, scholars argued that economic modernization in the twentieth century might actually require a set of public policies such as capital accumulation and investment in heavy industries that could only be implemented by authoritarian political systems; developing countries faced the terrible dilemma of a trade-off between prosperity and political liberty. Alternatively, socioeconomic modernization disrupted people from old patterns of culture and social interaction, and raised their expectations of increasing material gain. When these expectations could not be met, and in the absence of robust political institutions such as mass parties, political instability and military rule typically followed.

Thus, by the mid-1970s, the prospects for new democracies were considered highly unfavorable. Either economic development would be associated with authoritarian governments or, if one maintained loyalty to the modernization hypothesis, given the global recession of mid-decade and the debt crisis that spread through the developing world in the early 1980s, democracy-inducing development was not on the horizon. Political scientists were thus caught largely by surprise by the wave of democratization that began in Southern Europe in the mid-1970s and spread, over the next two decades, to Latin America, Eastern Europe, Asia, and Africa.

3.2 *Process-based Theories*

Reacting to the predictive failures of both modernization theory and its critical rivals, scholars articulated

a model of regime change that downplayed the role of sociological, cultural, and economic conditions and highlighted the importance of relatively indeterminate transition processes. These transition processes are indeterminate because imperfect information, malleable interests and identities, rapidly changing strategic calculations, and a strong role for contingent events combine in various permutations to preclude structuralist accounts of regime change. In other words, during periods of regime transitions, when previous rules and institutions have been suspended, actors typically find that their identities and interests are subject to change; consequently, it becomes extremely difficult to devise coherent strategies driving goal-oriented behavior; and events that might previously have been considered routine may take on additional significance, forcing actors to recalculate their strategies and their goals. All of these factors do not aggregate into a single model or theory of regime change; rather, they all suggest why the constraints exercised by structural conditions will be loosened, opening up considerable space for politicians to exercise initiative and craft novel outcomes that might have seemed otherwise infeasible. Just as there are no inexorable preconditions for democracy, seemingly unavoidable democratic breakdowns in the past could have been avoided by only incremental changes in actors' perceptions that would have permitted a greater degree of trust and collaboration on behalf of shared goals. Finally, to the extent that relevant actors cannot predict outcomes, normal social science methodology oriented to discovering law-like behavior will be inappropriate to the analysis of democratic transitions (O'Donnell and Schmitter 1986).

The focus on actors and their strategic interactions has undoubtedly been a useful corrective to analyses that focused exclusively on environmental factors. But actor-oriented models may have gone too far in their rejection of any meaningful role for structural conditions, prompting critics to charge this literature with excessive voluntarism. Reviewers of this literature have regularly claimed that its authors too readily dismiss the importance of a range of structural factors other than the level of economic development. Moreover, they point out, it has proven difficult for transitions scholars completely to excise structural factors that regularly appear, either in explanations of individual cases or in the acknowledgement of the 'long-run' causal impact of macro-level factors. Methodologically, these analyses have been overwhelmingly based on countries that have successfully democratized. Persistent authoritarian regimes have not been systematically incorporated into the analysis of regime change. This bias in the selection of cases and the failure to consider the full range of structural factors raises the strong possibility that those cases that democratized might all share one or more structural factors distinguishing them from those cases that have not democratized.

3.3 *Macro-historical Analyses*

It seems highly likely, therefore, that subsequent research will continue to draw on a third approach to regime change: macro-historical studies. These analyses occupy a middle position between the universalism of modernization theory and the particularity of transitions studies by wagering that a relatively delimited set of historically specific constellations of structural factors can be identified that will account for regime change in clusters of cases. In contrast to the emphasis on values and cultural change of modernization theory, the causal mechanisms adduced by these approaches are typically rooted in the conflictual relations of rational actors pursuing their interests.

A seminal work in this approach is Barrington Moore Jr.'s (1966) analysis of the divergent regime outcomes associated with European economic modernization. For Moore, various relations between the crown, landed aristocracies, and urban bourgeoisies at the time of the commercialization of agriculture determined whether modern regimes would be democratic or authoritarian. For democracy, a strong bourgeoisie had to be the dominant partner of landed elites in a coalition that sought to check royal power. When bourgeoisies were not available to play this leading role, divergent relations between landed elites and peasantries determined whether twentieth-century authoritarian regimes would be fascist or communist. Recent theorists follow Moore's suggestion that the relative strength of classes determines regime outcomes, but propose that the key to democracy lies in the weakening of antidemocratic landed classes and the strengthening of pro-democratic urban working classes (Rueschemeyer et al. 1992).

A second approach focuses less on relative class power and more on the consequences of divergent patterns of interclass coalitions. In this approach, groups embroiled in political-economic conflicts seek allies from among urban workers or rural cultivators, or both, as a means of vanquishing rivals. Regime outcomes are by-products of the political bargains sealing these class alliances (Luebbert 1991).

Another macro-historical approach links regime outcomes to patterns of state formation. As European rulers sought to increase tax burdens and to recruit manpower from their populations, they were forced to bargain with other power-holders who controlled access to those resources: landed elites and urban capitalists (Tilly 1990). Those bargains had a powerful influence on the structure of ensuing states and on the nature of the political regime. The availability of liquid capital in countries such as England and Holland, for example, tended to produce less bulky state apparatuses and more restrictions on despotic power, as holders of moveable assets otherwise beyond the reach of rulers were able to secure some control over public policy and more transparent and accountable administrative procedures in exchange for financing war-

making and state building. Indeed, in these instances, many of the limitations on despotic power established in the late medieval period such as parliaments, independent judiciaries, rudimentary rights, and the rule of law persisted into the early modern period and beyond to the twentieth century (Downing 1992). In contrast, in regions where urban capitalists were few in number and limited in wealth, rulers tended to bargain with landed elites. In these cases, medieval limitations on despotic power were abolished, absolutist regimes were established, and the prospects for stable democracy in the twentieth century were diminished.

These state-building patterns have not been reproduced outside of Europe. Instead, a repeated theme of non-European political development has been the deviation from European patterns due precisely to the existence of wealthier European states that financed non-European state formation, precluding rulers in the periphery from the need to bargain with their own populations. Assuming some conditional relationship between taxation and representation, the reliance of states on external sources of finance such as flows of foreign aid or sale of commodities by state-owned companies forms a potent obstacle to democracy. Moreover, the sale of advanced military hardware and the training of officer corps in both external and internal warfare gives third world states a powerful coercive edge over their civil societies and often gives military officers an edge over civilian officials. Finally, in times of crisis, foreign interventions have and undoubtedly will continue to shape regime outcomes in the non-European world.

At their best, macro-historical analyses are equally attentive to historical particularity, large-scale global processes, and the linkages between situational variables and individual action. But the search for clusters of variables explaining groups of cases can easily be replaced by purely idiographic explanations of individual cases, especially as the wealth of plausible hypotheses and of partially confirmatory evidence renders difficult the extension of any one model to multiple cases in different world-historical settings. Moreover the link between macro-structural variables and individual action is more often imputed than carefully theorized and researched. Finally, in light of efforts to capture at least some particularity of individual cases, there has perhaps been insufficient attention to hypotheses motivating more universalistic theories, such as the development of capitalism. It has been noted with increasing concern, for example, that recently created democracies in Latin America and elsewhere may be grafts of formal institutional structures on to deeply antidemocratic behavioral patterns. To what extent can these variants of democracy be explained by late development and global capitalism? Scholars seeking middle-range hypotheses have underemphasized this sort of research, the provenance of large-n statistical studies. Future research may have to give even more weight to both nomological theorems

and the specification and verification of macro-micro links.

4. Regime Types and Public Policy

Political scientists have frequently invoked regime types as an influential cause of public policies, particularly in the study of the political determinants of economic policies. In the 1960s, scholars openly wondered whether economic development in the post-colonial world would require authoritarian governments. By the mid-1970s, the end of the postwar growth period and manifestations of domestic political instability prompted some theorists to suggest the onset of a crisis of democratic governability in the advanced industrial nations as well.

In the short term, the hypothesized implication of democratic regimes is the 'political business cycle.' According to this theory, instrumentally rational political leaders who seek above all else to win elections will use their power to alter the economy by reducing unemployment in the period leading up to elections. After being rewarded by voters with re-election, leaders will impose austerity policies to eliminate the inflation induced by pre-election policies. Under democracy, then, economic policy would always be held hostage to the goals of egoistic politicians, resulting in regular macroeconomic instability that follows a 'stop-and-go' pattern. Over the longer term, the hypothesized consequences of democratic regimes were even more alarming. In the well-known analysis of Mancur Olson (1982), special-interest groups formed, under most circumstances, distributive coalitions that sought to increase the economic welfare of their members, typically via pushing for policies that were highly detrimental to overall growth, such as protection or halting the spread of labor-replacing technology. Under conditions of liberal democracy, there were no impediments to the growth of distributive coalitions other than collective action problems within each organization, problems that were likely to be overcome with sufficient time. Consequently, the fastest growing of the advanced economies in the postwar period were those like Germany and Japan which had experienced a fascist interlude that had eliminated most special-interest groups, granting them a period of rapid growth before the inevitable decline set in. The basic logic of these arguments was applied to the developing world as well. Given a trade-off between consumption and investment, democratic governments were seen as more likely to engage in vote-buying behavior and thus less likely to impose policies that delayed consumption on behalf of capital accumulation and long-term growth.

More recent research and theorizing have cast considerable doubt on these arguments. While data

demonstrates a political business cycle in some countries, that evidence is equivocal. When larger samples of countries are studied, the evidence for a business cycle is very weak. One reason that no such cycle can be shown to be consistently evident is that voters, while certainly influenced by economic matters, are best understood as considering both retrospective and prospective components of the economy, effectively canceling out the efforts of politicians to manage the economy for electoral gains (Lewis-Beck 1988). Arguments about the long-term effects of democratic regimes on economic performance also ignored a host of institutional factors that have considerable impact on economic policy-making and performance. States with democratic regimes, for example, display considerable diversity in their structures of interest aggregation and articulation: they thus offer considerably different prospects for particularistic interests to gain access to policy makers and correspondingly different opportunities for policy makers to act autonomously. More importantly, economic performance is closely linked to the specific features of non-regime institutions. The structure of financial system intermediation and of the trade union movement, for example, have powerful effects on growth, inflation, and unemployment rates as well as on the capacity to adjust successfully to international economic change. Broadly similar arguments about the developing world attribute level of development more to specific institutions such as bureaucratic structures and financial systems than to regime types. Indeed, numerous studies have demonstrated that even authoritarian regimes frequently sacrifice long-term investment to short-term political calculations, and that democratic regimes often have the legitimacy and popular support needed to enact painful reforms.

See also: Democracy; Democratic Transitions; Development and the State; Development, Economics of; Dictatorship; Globalization: Political Aspects; Leadership: Political; Modernization, Political: Alternate Paths; Modernization, Political: Development of the Concept; Nations and Nation-states in History; Parliamentary Government; Policy History: Origins; Policy History: State and Economy; Power: Political; Presidential Government; Sovereignty: Political; State Formation; State, History of; State, Sociology of the

Bibliography

- Cheibub J A 1998 Political regimes and the extractive capacity of governments: Taxation in democracies and dictatorships. *World Politics* **50**: 349–76
- Dahl R 1971 *Polyarchy: Participation and Opposition*. Yale University Press, New Haven, CT
- Diamond L 1992 Economic development and democracy reconsidered. *American Behavioral Scientist* **35**: 450–99
- Diamond L 1993 Political culture and democracy. In: Diamond L (ed.) *Political Culture and Democracy in Developing Countries*. Lynne Rienner Publishers, Boulder, CO
- Downing B 1992 *The Military Revolution and Political Change: Origins of Democracy and Autocracy in Early Modern Europe*. Princeton University Press, Princeton, NJ
- Gutman A 1993 Democracy. In: Goodin R, Pettit P (eds.) *A Companion to Contemporary Political Philosophy*. Basil Blackwell, Oxford, UK
- Karl T L, Schmitter P C 1993 What democracy is ... and is not. In: Diamond L, Plattner M F (eds.) *The Global Resurgence of Democracy*. Johns Hopkins University Press, Baltimore, MD
- Lawson S 1993 Conceptual issues in the comparative study of regime change and democratization. *Comparative Politics* **25**: 183–205
- Lewis-Beck M S 1988 *Economics and Elections: The Major Western Democracies*. University of Michigan, Ann Arbor, MI
- Lindenberg M, Devarajan S 1993 Prescribing strong economic medicine: Revisiting the myths about structural adjustment, democracy, and economic performance in developing countries. *Comparative Politics* **25**: 169–82
- Lipset S M 1959 Some social requisites of democracy: Economic development and political legitimacy. *American Political Science Review* **53**: 69–105
- Londregan J B, Poole K T 1996 Does high income promote democracy? *World Politics* **49**: 1–30
- Luebbert G M 1991 *Liberalism, Fascism, or Social Democracy: Social Classes and the Political Origins of Regimes in Interwar Europe*. Oxford University Press, New York
- Mann M 1984 The autonomous power of the state: Its origins, mechanisms and results. *Archives Européennes de Sociologie* **25**: 185–213
- Moore B 1966 *Social Origins of Dictatorship and Democracy: Lord and Peasant in the Making of the Modern World*. Beacon, Boston, MA
- O'Donnell G 1974 *Modernization and Bureaucratic-Authoritarianism*. University of California, Berkeley, CA
- O'Donnell G, Schmitter P C 1986 *Transitions from Authoritarian Rule: Tentative Conclusions about Uncertain Democracies*. Johns Hopkins University Press, Baltimore, MD
- Olson M 1982 *The Rise and Decline of Nations: Economic Growth, Stagflation, and Social Rigidities*. Yale University Press, New Haven, CT
- Przeworski A 1991 *Democracy and the Market: Political and Economic Reforms in Eastern Europe and Latin America*. Cambridge University Press, Cambridge, UK
- Przeworski A, Limongi F 1997 Modernization: Theories and facts. *World Politics* **49**: 155–83
- Rueschemeyer D, Stephens E H, Stephens J D 1992 *Capitalist Development and Democracy*. University of Chicago Press, Chicago
- Schaffer F C 1998 *Democracy in Translation: Understanding Politics in an Unfamiliar Culture*. Cornell University Press, Ithaca, NY
- Schumpeter J A 1942 *Capitalism, Socialism and Democracy*. Harper Torchbooks, New York
- Tilly C 1990 *Coercion, Capital, and European States, AD 990–1990*. Basil Blackwell, Oxford, UK

D. Waldner

Copyright © 2001 Elsevier Science Ltd. 11553
All rights reserved.

Policy History: State and Economy

The appropriate relationship between state and economy has been one of the central faultlines of political debate, policy conflicts, and scholarly debates for centuries. Scholars are no closer to reaching consensus on these issues at the beginning of the twenty-first century than they were when Adam Smith first published *The Wealth of Nations* (1776). Yet despite the venerable nature of this faultline, there have been remarkably few efforts to provide a history of the centuries old debate on the most desirable relationship between state and economy. This article represents a preliminary effort to sketch out a brief history of these policy debates, concentrating on the nineteenth and twentieth centuries. The critical ingredients of this history are doctrines, global constraints, and theoretical confusions.

1. Doctrines

The terms of debate on the relationship between state and economy over the last two centuries is radically different from what came before. For centuries, European governments used a variety of mercantilist practices to enhance the economic growth of their territories by setting up barriers against imports, investing in education and technology, providing concessions and other support to promising enterprises, and establishing relief policies to protect the poor from economic hardship (Heckscher 1934, Reinert 1999). There is disagreement over when and how effectively these mercantilist practices were codified into a coherent body of economic doctrine, but there is consensus that for centuries, the use of governmental power to enhance economic development was seen as legitimate and desirable.

The classical economists, particularly Ricardo and Malthus, fundamentally changed the terms of discussion by conceptualizing the economy as an autonomous entity that is governed by its own principles, such as the law of supply and demand. Both Polanyi (1957) and Rothschild (1992) stress the important discontinuities between Adam Smith and the later writers of the classical school. Writing in the shadow of Newton's scientific achievements, they imagined that the workings of the economy were similar to the harmonious movements of the planets in the solar system that could now be fully explained by the laws of physics (Redman 1997). In searching for the economic equivalents of the laws of physics, the classical economists generally viewed the state as an 'outside' force whose actions were likely to interfere with the economy's capacities for self-regulation. In fact, their arguments against various mercantilist policies often emphasized that government actions were likely to have perverse consequences—opposite to those initially intended—precisely because such actions

would disrupt the natural and self-regulating operation of economic laws (Hirschman 1991).

The classical economists recognized that there was a fundamental tension between their abstract model of an economy and the real world in which human beings were organized into separate nations governed by states. Just as their mercantilist predecessors, they were writing in order to shape the policies of their own government; they recognized that their vision of the proper organization of the economy could only be implemented through the use of the state. Hence, they distinguished between appropriate government policies—those that supported the full emergence of the economy as autonomous—and those government policies that were inappropriate—that were likely to retard or interfere with an autonomous economy. But while the classical economists agreed on this distinction, they were unable either to define precisely where this dividing line should be drawn or to develop a theoretically sound justification for its location. In fact, they had profound disagreements among themselves on where the line should be drawn. Despite the centrality of free trade to the classical tradition, one of its key thinkers—Malthus—argued consistently that it was appropriate for England to retain the Corn Laws that limited agricultural imports. John Stuart Mill, who played a critical role in codifying classical economics, also entertained a variety of heterodox policy ideas (Schumpeter 1954).

1.1 Purists and Pragmatists

This inability to derive the proper functions of the state from their abstract model of the economy has characterized the mainstream tradition of economic thought through the entire nineteenth and twentieth centuries (for accounts of this tradition that are less Anglo-American centered than most, see Schumpeter (1954) and Screpanti and Zamagni (1993)). The result has been a continuous tension within the tradition between 'purists' and 'pragmatists.' The purists argue that since we know that expanding the scope of markets improves economic efficiency, it follows that international free trade is always preferable to any form of protectionism. Furthermore, since governments can subvert free trade by providing subsidies to their own producers or by using regulations to discourage imports, it follows that a domestic regime of *laissez faire* in which government exercises the smallest possible role in the economy is to be preferred. Pragmatists, in contrast, starting with John Stuart Mill, have argued that precisely because the real world of actual economies diverges from the abstract model, there is an affirmative role of government that extends beyond establishing the minimal conditions for private contracting. Pragmatists have differed among themselves as to what government actions they would endorse, but some have favored tariff protection of

infant industries, extensive regulatory regimes to overcome market failures, substantial measures to redistribute income, and the provision of a wide variety of public goods, including government participation in planning of the economy's future growth path (Hutchison 1978 is particularly valuable in tracing out the pragmatism of Jevons, Marshall, and Pigou).

But this longstanding argument between purists and pragmatists within the central tradition of economics has usually been overshadowed by conflicts generated by two other important sets of ideas. The first is the challenge to the market model posed by socialist ideas, particularly those of the Marxist tradition. The second is the emergence of a counter tradition in economics that rejected both the methodology and the free trade preferences of the classical tradition.

1.2 Socialism

The socialist tradition predates Marx and includes a range of different theoretical currents, but the Marxist current has been the strongest for almost a century and a half. Moreover, it was Marx who effectively defined the intellectual achievement of the classical economists; he argued that they were the theorists and apologists of the emerging bourgeois social order whose triumph was signaled by the French Revolution. For Marx, this new social order was a coherent system that revolved around the imperatives of profit maximization. Precisely because the capitalist mode of production was a unified system, efforts to reform it were fundamentally futile. Marx repeatedly insisted that the only way for the working class to advance its collective interests was by a revolution that replaced capitalism with a radically different social order in which markets, private property, and the pursuit of profit would be eliminated.

By the end of the nineteenth century, the Marxist tradition developed its own division between purists and pragmatists. Purists continued to oppose any efforts to reform capitalism as inherently futile since they viewed it as impossible to change capitalism's fundamental logic. Pragmatists argued that meaningful reforms could be won that would strengthen the economic and political position of the working class and make possible further advances in the direction of socialism. Since the pragmatist position often resonated with the views of non-Marxist socialists and with the views of reformists in the trade union movement, it played an important part in expanding the state's role in the provision of social insurance, the regulation of industry, and the redistribution of income (Sassoon 1996, Hicks 1999).

While purists in the Marxist tradition and purist defenders of the free market have been bitter antagonists, it is worth stressing that they share certain common assumptions. They both insist that the only

real alternatives for organizing societies are free market capitalism or socialism. They both see any effort to mix socialist and capitalist principles within a single society to be inherently unstable; sooner or later the instability will be resolved by movement to one of the stable poles. Moreover, they have the same view of the globe—that a world divided between socialist and capitalist societies is unstable; the world ultimately will have to be organized by only one of these principles.

1.3 The Counter Tradition

Marxism, however, was not the only oppositional doctrine that was spawned by the theoretical achievements of the classical economists. Starting as early as the 1820s, a counter tradition emerged that challenged both the classical economist's emphasis on the desirability of free trade and their methodological assumptions. But while important work has been done in this tradition in virtually every decade from 1820 to 2000, the identification of this counter tradition is largely retrospective (Reinert 1999). It does not have either the organizational or the doctrinal continuity of either mainstream economics or Marxism. Its founding figure was the German, Friedrich List, who argued that the policy proposals of the classical economists were not universally applicable; if poorer nations were to embrace free trade without qualification, they would never be able to catch up with the world's leading economic power—England. List advocated the use of governmental powers, including tariffs to protect infant industries, as a means to accelerate development of the economy's 'productive forces' (Henderson 1983). The American economist, Henry Carey, also made important early contributions to this counter tradition.

The counter tradition was further elaborated in the 1870s and 1880s by the younger generation of the German historical school, especially in the writings of Gustav von Schmoller. Schmoller rejected both the 'classical tradition' view of the economy as autonomous and their *laissez faire* policy prescriptions (Schumpeter 1954). The German historical school, in turn, inspired both the first generation of academic economists in the United States (Ross 1991) and a group of English historical economists, including William Cunningham, J. W. Ashley, and W. A. S. Hewins who revisited English economic history and found powerful evidence that English economic development had depended heavily on state policies (Koot 1987). In the first part of the twentieth century, further contributions to the counter tradition were made by the American institutionalists and the English Fabians. In the 1930s, despite his orthodox intellectual lineage, John Maynard Keynes joined this counter tradition by breaking with free trade and advocating active government management of the economy. From

the 1930s through the 1970s, this counter tradition was enriched by a broad range of scholars including Japanese economists who are little known in the West (Gao 1997), the Latin American group around Raul Prebisch that developed the idea of import substituting industrialization (McMichael 1996), and others. A particularly important contribution was the study of late-industrializing countries by Gerschenkron (1962).

Political leaders in late-industrializing countries have turned repeatedly to theorists of this counter tradition to help develop policies to accelerate their nation's development. Count Witte, the Russian modernizing minister under Tsar Nicholas II had studied List closely before taking office and followed many of List's policy prescriptions (Henderson 1983, pp. 215–17). After the Meiji Restoration in Japan, the most influential economic work was that of the German historical school. As Gao writes, 'The intellectual roots of Japanese developmentalism can be traced to the German historical school' (Gao 1997, p. 60). More recently, the spectacular economic advances of the South Korean and Taiwanese economies have been attributed to the influence of developmentalist policies inspired by the counter tradition (Wade 1990, see also the important book by Evans 1995).

The boundary lines separating these various currents of economic doctrine have sometimes been vague and indistinct. In the twentieth century, the dividing line between pragmatists within the mainstream tradition and proponents of the counter tradition has been crossed by major figures. (There are also important figures whose work spans the boundary line between Marxism and other traditions; of particular interest are those who have contributed to the theory of market socialism.) Two of the greatest economists of the twentieth century—Keynes and Schumpeter—can legitimately be claimed by both the mainstream and the counter tradition. A further complication is that the mainstream tradition of Anglo-American economics whose successive paradigms were codified by J. S. Mill, Marshall, and Keynes, entered into crisis in the early 1970s. As the influence of Keynesian ideas declined, no single theory has become dominant and the tradition has splintered into a number of conflicting views (Heilbroner and Milbert 1995). This is the theoretical context in which the political influence of the purist wing of the tradition has increased dramatically in the USA and throughout the world.

2. Global Constraints

It would be a mistake to imagine that political leaders have been free to choose among these diverse doctrines in deciding what role the state should play in their particular economy. On the contrary, these diverse doctrines have existed within a global structure of power that has constrained and limited the choices of

political leaders. But this global structure of power has itself been in continuous flux, so that the severity of the constraints has varied considerably over time. Moreover, the constraints also vary within different historical periods; some countries have had more freedom to maneuver because of the specifics of their location.

The constraints are a direct result of successive national hegemonies over what Wallerstein has termed 'the modern capitalist world-system' (Wallerstein 1974, Arrighi and Silver 1999). Dutch hegemony was established initially in the seventeenth century, but it gave way, after an interregnum, to English hegemony in the early years of the nineteenth century. England's hegemony was undermined first by World War I and then by the global economic collapse of the early 1930s. After another interregnum, the US emerged as the new hegemonic power in the final years of World War II. Moreover, the geographical scope of this capitalist world-system has steadily expanded from its European core so that by the middle of the twentieth century, it encompassed the entire globe.

Hegemony represents a combination of economic, military, and cultural power. In the nineteenth century, for example, the new ideas of the classical economists were a constitutive part of the English-dominated global order. Other nations were under pressure to adopt the ideas of free trade and *laissez faire* as a way to gain access to English capital, to avoid England's military wrath, and to partake in the industrial modernity that England exemplified. The same point can be said of US hegemony in the second half of the twentieth century; it is impossible to determine whether the influence of US-sponsored economic doctrines has more to do with financial power, military power, or the powerful imagery of Hollywood and Madison Avenue.

2.1 A Rough Periodization

Even in England, there was a considerable time lag between the writings of the classical economists and the full embrace of free trade. It was only with the repeal of the Corn Laws in 1846 that England began exerting considerable pressure on other nations to embrace the doctrines of free trade and *laissez faire*. Hence, the period from 1850 to 1875 can be thought of as the first great epoch of free trade. During this period, most European nations lowered their tariff barriers and made at least some efforts to embrace *laissez faire* (Chase-Dunn et al. 2000). By 1870, many nations had embraced the gold standard by making their currencies freely convertible into gold (Eichengreen 1996). The gold standard represented the fullest expression of English hegemony, since it was designed to equilibrate automatically the global economy without governmental interference.

The dynamic shifted sharply in the next period from about 1875 to 1900. This was a period of slower

growth and declining prices that in combination with the gold standard placed severe pressures on national economies (Polanyi 1957). The result was a marked retreat from both free trade and *laissez faire*. Tariff barriers increased significantly in this era (Chase-Dunn et al. 2000), and rising protectionism was clearly a component of the rush for empire as the European powers, the US, and Japan moved to establish formal colonies (Polanyi 1957). The trade barriers meant the potential loss of access to both markets and raw materials in those regions that had been colonized by other powers. Moreover, national regulations to control market forces expanded dramatically; this was the era in which government-sponsored social insurance was first developed and began to spread.

The brief period from 1900 until the outbreak of WW I is an anomalous and contradictory era. On the one side, the pendulum swung back in the direction of lowering trade barriers (Chase-Dunn et al. 2000), and international capital movements reached unprecedented levels in this period as the global financial system became more tightly integrated. On the other side, there were few signs of a return to *laissez faire*; the state's role in managing domestic economies and providing social welfare continued to expand in many nations. Governments were under increasing pressure to make concessions to socialist movements that were growing in power and influence. Moreover, in this period, late-industrializing nations, such as Russia and Japan, did not hesitate to use governmental power to accelerate the processes of industrial development. Moreover, the intensification of the military rivalry between England and Germany in this period was very much a continuation of the imperialist conflicts that had emerged in the previous period.

Between 1914 and 1945, nineteenth-century civilization collapsed, and it appeared that the ways of thinking about state and economy developed by the classical economists had become completely obsolete. From Bolshevism to fascism to the experiments of the New Deal, it seemed that there were a multitude of divergent paths for organizing national economies (Polanyi 1957). Moreover, the imperatives of war meant that the major powers, even those that retained democratic institutions, went through extended periods of time when economic life was completely subordinated to political and military control. While much of Asia and Africa remained under colonial control, the larger Latin America nations were able to take advantage of the breakdown of global power to follow the teachings of the counter tradition (McMichael 1996). The overwhelming reality of this period was the human suffering generated by war, dislocation, depression, and unprecedented violence directed at civilians.

During the first period of US global hegemony from 1945 to 1973, the USA struggled, with the help of multilateral institutions, to limit the range of permissible variations in national economic policies. This

meant an intensification of the Cold War to restrain the growth of socialism that had spread to Eastern Europe and China. The process of decolonization meant that this battle was now fought out over a far larger group of nations. But while the USA pushed strongly in this period for trade liberalization, there was no effort to restore *laissez faire*. It was taken for granted that, as with the New Deal regime in the USA, the government would play a large role in the domestic economy in managing the business cycle, regulating private firms, and producing a broad array of public goods. In this period, the pragmatists within the mainstream tradition of economics were dominant with their vision of a mixed economy. The ability of governments to pursue macroeconomic objectives was facilitated further by global rules that severely restricted international capital movements. Even poorer countries were expected to develop ambitious development plans and an elaborate system of foreign aid emerged to finance these efforts. Yet these efforts were biased strongly towards a model of development led by the private sector and countries that were too aggressive in fostering state-led development were pressured to stop (Wood 1986). A few nations, such as South Korea and Taiwan, were able to resist these pressures because of their location close to the fault lines of the Cold War, but their experience was deeply atypical (Wade 1990).

Finally, the period from 1973 to 2000 is the second great epoch of free trade. The steady progress towards reduction of trade barriers that began in 1945 was continued, and it was now joined by the elimination of barriers to international capital movement. The integration of global financial markets reached levels that had only previously been attained in the period between 1900 and 1914. Starting in the late 1970s, with Margaret Thatcher and Ronald Reagan, and intensifying after the collapse of the USSR, the old doctrines of the classical economists were now recycled under the rubric of neoliberalism. *Laissez faire* ideas were reintroduced on the grounds that government efforts to manage the business cycle were futile and excessive levels of taxation and regulation were destroying the vitality of market economies. Trade negotiations were now expanded to dismantle 'non-tariff barriers' that included domestic legislation that had the unintended consequence of discriminating against foreign products. Developing nations were told to scrap their planning processes, to privatize state-owned enterprises, and to end subsidies that made food, transportation, and housing more available for the poor (McMichael 1996).

3. Theoretical Confusions

In reviving *laissez faire*, neoliberals assume that they can distinguish between the legitimate, albeit limited, responsibilities of government and those illegitimate

government actions that are bound to have perverse consequences. But neoliberals have been no more successful than their classical predecessors in providing a theoretically justified demarcation line. That this problem has not been solved after two centuries of work suggests that there are deep theoretical confusions operating here.

3.1 Can the Economy be Autonomous?

The fundamental problem lies with the assumption, introduced by the classical economists, that an economy can be organized around a set of self-regulating markets. Even with the added stipulation that there will be a state and a legal system that enforces property rights and contracts, this is still not enough. This point was elaborated most forcefully by Polanyi (1957), a Hungarian refugee intellectual who developed his own theoretical synthesis drawing on insights from Marxism and the German historical school. Polanyi argued that the flaw in the belief in an autonomous economy is that land, labor, and money are not true commodities; they are not things that were produced for sale on a market. Economists assume, however, that the price mechanism will equilibrate supply and demand for these fictitious commodities just as it does for true commodities. The difficulty is that when the demand for labor suddenly falls, human beings cannot be warehoused like unsold merchandise; those who are out of work must still have income. Similarly, countries that leave it to the market to determine the supply of money and credit are likely to suffer from chronic deflation when credit is too scarce or chronic inflation when it is too abundant. With labor, money, and land, a continuous governmental role is needed to make the market for these fictitious commodities work like other markets.

In short, a market economy cannot be autonomous because the state is involved intimately in three of the most important markets. What follows from this insight is that capitalism, defined as a society organized around an autonomous market in which actors seek to maximize utility, has never existed. Polanyi went on to argue that the market societies that actually exist are constituted by two opposing tendencies. The first are the ongoing efforts to expand the scope of markets—the project first announced by the classical economists. The second are parallel efforts to protect the social order from the consequences of market competition that includes regulatory efforts, social welfare policies, and other macroeconomic policies. Polanyi emphasizes that business firms participate in both of these tendencies, since they often need to protect themselves from market situations where uncertainty is uncomfortably high.

In recent years, scholars have produced a body of work that elaborates this Polanyian insight by

analyzing the existence of a number of varieties of capitalism that vary significantly in the way that key markets are organized and in the relationships between state and economy (Block 1994, Crouch and Streeck 1997, Hollingsworth and Boyer 1997). These studies represent a fundamental challenge to the assumptions of both mainstream economists and Marxists that market societies should converge towards some common pattern. At the same time, however, some commentators (Gray 1998) have expressed concern that the neoliberal rules of the global economy might have the consequence of destroying varieties of capitalism that are actually more efficient than the Anglo-American type.

3.2 Transition to a Borderless World

A second theoretical confusion has to do with the tension between a unified global market and a world of nation states. Both the classical economists and contemporary neoliberals have been motivated by a vision of a peaceful planet in which military rivalries and warfare are replaced by the competition of individuals and firms to reach higher levels of efficiency. Both have imagined a gradual withering away of the nation state as the globe becomes economically unified. While there is much that is appealing in this vision, there has been a lack of clear thinking about the actual staging of this transition to a borderless world.

Ironically, the historical periods in which this vision has been most influential were marked not by the disappearance of state power, but by the unchallenged military and political hegemony of a single power—first England and then the USA (Kindleberger 1973). The temporary dominance of a single power hardly seems a sound foundation for a transition to a world beyond nations. A closely related problem is that this transition requires that nations give up the various measures that buffer their populations from the pressures of the global economy just as those pressures are intensifying. Creating a unified global market for agricultural goods, for example, requires that governments dismantle programs designed to stabilize the income of farmers just as those farmers are becoming more vulnerable to cheaper agricultural imports. The neoliberal vision rests on the willingness of millions of ordinary people—farmers, workers, and small businesspeople to be extraordinarily flexible—to accept that years of reasonable income periodically will be interrupted by years in which their income might fall by as much as 40–60 percent. (These figures reflect income losses in Asia and Mexico in the financial crises of the 1990s.)

For the affluent, such flexibility might seem logical and reasonable; they know that years of high returns on their portfolios can alternate with periods of

negative returns. But those who lack an investment portfolio tend to resist this type of flexibility; their ability to pursue a life plan depends on some stability and predictability in their income flows. When such people are able to exercise political influence on their national leaders, they demand to be protected from extreme instability of their income. This is precisely what happened in the period from 1875 to 1900; ordinary people demanded to be protected from the costs imposed on them by global economic integration, and the result was a significant retreat from global free trade. Moreover, if governments are constrained by global power and are unable to respond to these popular demands to be protected from income instability, discontent does not disappear. It is often channeled into militant movements of the left or of the right that target both the ineffectiveness of the government in power and the foreign enemies who have imposed economic hardship on the local people. The efforts of centrist political leaders in the 1920s and 1930s to restore the gold standard and a liberal world economy were defeated by these types of movements. The deep historical irony is that the effort to create a pacified, borderless world produced exactly the opposite; extreme forms of nationalism and war on a global scale.

4. Conclusion

The period from 1973 to 2000 has seen unprecedented growth in a variety of indices of economic globalization. However, these theoretical confusions about the autonomy of the economy and the transition to a borderless world raise concerns about the sustainability of these trends. How the organization of economies and states will evolve remains an open question.

See also: Development and the State; Development, Economics of; Economics and Ethics; Economics, History of; Economics: Overview; Policy History: Regimes; Political Economy, History of

Bibliography

- Arrighi G, Silver B 1999 *Chaos and Governance in the Modern World System*. University of Minnesota Press, Minneapolis, MN
- Block F 1994 The roles of the state in the economy. In: Smelser N J, Swedberg R (eds.) *The Handbook of Economic Sociology*. Princeton University Press, Princeton, NJ
- Chase-Dunn C, Kawano Y, Brewer B D 2000 Trade globalization since 1795: Waves of integration in the world-system. *American Sociological Review* 65: 77–95
- Crouch C, Streeck W 1997 *Political Economy of Modern Capitalism: Mapping Convergence and Diversity*. Sage, Thousand Oaks, CA
- Eichengreen B 1996 *Globalizing Capital: A History of the International Monetary System*. Princeton University Press, Princeton, NJ
- Evans P 1995 *Embedded Autonomy: States and Industrial Transformation*. Princeton University Press, Princeton, NJ
- Gao B 1997 *Economic Ideology and Japanese Industrial Policy*. Cambridge University Press, Cambridge, UK
- Gerschenkron A 1962 *Economic Backwardness in Historical Perspective*. Harvard University Press, Cambridge, MA
- Gray J 1998 *False Dawn: The Delusions of Global Capitalism*. Granta, London
- Heckscher E 1934 *Mercantilism*, 2 vols. George Allen & Unwin, London
- Heilbroner R, Milbert W 1995 *The Crisis of Vision in Modern Economic Thought*. Cambridge University Press, New York
- Henderson W O 1983 *Friedrich List: Economist and Visionary 1789–1846*. Frank Cass, London
- Hicks A 1999 *Social Democracy and Welfare Capitalism*. Cornell University Press, Ithaca, NY
- Hirschman A O 1991 *The Rhetoric of Reaction: Perversity, Futility, Jeopardy*. Harvard University Press, Cambridge, MA
- Hollingsworth R, Boyer R 1997 *Contemporary Capitalism: The Embeddedness of Institutions*. Cambridge University Press, Cambridge, UK
- Hutchison T W 1978 *On Revolutions and Progress in Economic Knowledge*. Cambridge University Press, Cambridge, UK
- Kindleberger C P 1973 *The World in Depression 1929–1939*. University of California Press, Berkeley, CA
- Koot G M 1987 *English Historical Economics, 1870–1926*. Cambridge University Press, Cambridge, UK
- McMichael P 1996 *Development and Social Change: A Global Perspective*. Pine Forge Press, Thousand Oaks, CA
- Polanyi K 1944/1957 *The Great Transformation*. Beacon Press, Boston
- Redman D A 1997 *The Rise of Political Economy as a Science*. MIT Press, Cambridge, MA
- Reinert E 1999 The role of the state in economic growth. *Journal of Economic Studies* 26: 268–326
- Ross D 1991 *The Origins of American Social Science*. Cambridge University Press, Cambridge, UK
- Rothschild E 1992 Adam Smith and conservative economics. *Economic History Review* XLV: 74–96
- Sassoon D 1996 *One Hundred Years of Socialism: The West European Left in the Twentieth Century*. Tauris, London
- Schumpeter J A 1954 *History of Economic Analysis*. Oxford University Press, New York
- Screpanti E, Zamagni S 1993 *An Outline of the History of Economic Thought*. Clarendon Press, Oxford, UK
- Smith A 1776/1976 *The Wealth of Nations*. University of Chicago Press, Chicago
- Wade R 1990 *Governing the Market: Economic Theory and the Role of Government in East Asian Industrialization*. Princeton University Press, Princeton, NJ
- Wallerstein I 1974 *The Modern World-System: Capitalist Agriculture and the Origins of the European World-Economy in the Sixteenth Century*. Academic Press, New York
- Wood R 1986 *From Marshall Plan to Debt Crisis: Foreign Aid and Development Choices in the World Economy*. University of California Press, Berkeley, CA

F. Block

Copyright © 2001 Elsevier Science Ltd.

All rights reserved.

11559

Policy Information

Information is used by governments for a wide variety of policy purposes. It can be used in the formulation of public policies, by helping to identify the need for a public policy, or to select one among alternatives. It can be used in the implementation of policies, to help target or improve programs. It can be used in the evaluation of policies, such as measures of progress toward desired outcomes, or of the intended or unintended social, economic, and environmental effects. Here, however, *policy information* is taken to mean information used in the formulation, design, and selection of public policies. It may also include information from the evaluation of public policies to the extent that effects of the policies inform consideration of alternative policies.

Policy information differs from political information, such as polls of public opinion, or of support for a policy among legislators. Policy information is objective. It may be quantitative or qualitative, but it is based on evidence, rather than opinions, attitudes, or beliefs (see Citro and Hanushek 1991). Policy information comprises both data and analysis. Indeed, it is often the result of *policy analysis*, by which is meant the objective prospective, and sometimes retrospective, evaluation of the social and economic implications of changes in public policy. For example, in considering choices among social welfare policies, decision makers typically want to know what populations would be affected, in what ways, and at what cost. Policy analysis provides answers to these 'what if' questions. In order to inform choices among policy alternatives, policy analysis involves forecasts or projections into the future. As a result, they are more uncertain than evaluations of the past. Three key ingredients of policy analysis are data; analytical models, including their documentation and validation; and effective communication of the results to policy makers, in particular of the assumptions in models and uncertainty in estimates.

1. Data

Accurate and relevant data are critical for policy analysis. Government surveys are a common source (see *Government Statistics*). Since the behavior of individuals in response to decisions of employers or government agencies is often what is being modeled, data may be needed on the interactions of individuals with organizations and institutions. For example, an individual's access to health care services will depend on providers, insurers, and government health care programs and policies. Often, however, data are available only on a single unit, such as an individual, family, or firm. In this situation, analysts must meld together data from different sources.

Useful sources of data to augment survey data are administrative records, such as health care claims, tax returns, earnings reports for social security, and case files from public assistance programs. Such data may need to be linked, or matched, to the survey data. Increased concern in government statistical agencies about possible violations of the confidentiality of survey data, however, militate against such linking. Analysts in this situation must make assumptions, the validity of which the analysts may not be able to determine without the linked data.

As important as the availability of data is, it is not sufficient. Data for public policy analysis must reflect relevant and accurate measurement. For this reason, government statistical agencies need to conduct research on concepts, definitions, and measurement, and assess the quality of their data. In the highly decentralized statistical system in the United States, however, it is not uncommon for agencies to report different measures from different surveys of ostensibly the same concept, such as income, employment status, or disability. Efforts may be made to explain the differences, or a result from one survey may be adjusted, or controlled, to agree with the result from a more prominent survey. Good policy analysis, however, requires integrated databases with the best estimates that can be developed from multiple sources.

Policy analysts, as well as other sophisticated users, cope with the lack of integrated data by performing additional adjustments to make the data suitable for modeling and analysis. But not all users have the information, much less the resources, to do so. And those that do may either duplicate efforts of others or, much worse, wind up with different results.

2. Models

Models are fundamental tools of policy analysis. *Microsimulation models* are one important type of policy analysis tool. With data from individual records, a microsimulation model simulates how the records change over time and as a result of a proposed policy. To predict the effect of a change in tax law, for example, the model processes records for people as if they were filling out their tax forms. As Citro and Hanushek (1991) note:

Models based on microsimulation techniques are conceptually highly attractive because they operate at the appropriate decision level and take into account the diverse circumstances and characteristics of the relevant population, whether it be low-income families, taxpayers, or health care providers. Such models are able to respond to important needs of the policy process for information about the effects of very fine-grained as well as broader policy changes, the effects of changes that involve complicated interactions among more than one government program, and the effects of changes on specific population groups as well as total program costs and caseloads.

Microsimulation models, however, tend to be highly complex and expensive to develop and maintain. Because of the lack of integrated data, analysts using the models must often patch together a variety of data and research results of varying, if even known, quality. As a result, often many assumptions are made; and some are unsupported. Because of these problems, some analysts may rely on simpler models, akin to accounting methods, to understand the effects of changes in a program. But the value of microsimulation models in informing a policy debate is that they can model behavioral responses to changes in policies and programs.

Many other statistical models are also used in policy analysis, from linear regression models, to simultaneous equation models, to complex macroeconomic ones. Macroeconomic models are also tools for 'what if' questions. They are often used to make unconditional forecasts, such as the rate of growth of the economy, as opposed to conditional projections. In addition to these models, a cornucopia of social science methods are also employed in policy analysis, including cost-benefit studies, risk assessments, meta-analyses, and even social experiments.

While microsimulation models seek to answer 'what if' questions, other models seek to answer 'why' questions, such as why people save, or fail to save, and to provide estimates of relationships that may be used in microsimulation models. Sometimes the research model may not be explicit, as when a database is constructed by piecing together data from different sources, in particular to determine a trend for which there is no survey-based indicator, such as the retention of children in the same grade in school. Sometimes the model may be difficult to ascertain, because political factors must also be taken into consideration; for example, in formulas used in legislation to allocate funds to school districts according to the numbers of children from families with incomes below the poverty level.

3. Uncertainty

Even when the model is explicitly clear, however, its uncertainty may be unclear. The estimates from models in policy analysis are conditional forecasts of the effects of a policy change. The effects may be poorly understood. They may be unknown or depend on unknown factors. The model relating cause and effect may be incorrect; or it may be uncertain as to which of many factors contribute to the effects. Moreover, the lack of data is a ubiquitous problem.

The understanding, however, of the effects of a proposed policy is important, if decisions on it are to balance the benefits against the costs. Key to that understanding are reasonable judgments about the quality of estimates of the likely effects of a proposed policy change. Policy makers need to understand the

uncertainty or variability in the estimates and to judge the track record of a policy analysis model. Unfortunately, they are rarely given the information to enable them to do so.

3.1 Model Validation

Assessing the quality of model estimates is a common problem in statistical research, but the problem is compounded in policy analysis, because it involves conditional forecasts. Some, if not all, of the policy alternatives studied may not be adopted. And, even for alternatives that are, data may never become available on the outcome. Yet it is crucial to understand the uncertainty in model estimates, if only to evaluate the utility of different models. *Model validation* is the process for measuring the uncertainty or variability in a model's estimates and identifying the sources of that uncertainty (Citro and Hanushek 1991).

One technique of model validation is *external validation*, in which the model's estimates are compared with the outcomes the model is forecasting. Because the outcomes may never be known, however, two variants of external validation may be used. One is *ex post forecasting*, in which one takes a policy or program change adopted in the past and, applying a model to data available then, forecasts the effects in a more recent past year when outcome data are available to compare. Another variant is *backcasting*, in which the model and current data are used to simulate the effects of a reversion to program conditions in the past, and the results are compared to measures of the effects based on data in that period.

Internal validation refers to studies of the model, its components, and how it is used, as well as of the data used. Measures of internal validity derive from the data used and model estimates obtained, without any comparison to external reality. They include sampling variability and other errors in the data or other inputs, and errors of model misspecification. Estimates of variability may employ standard sample variance calculations or more modern resampling techniques. *Sensitivity analysis* looks at the effects on model estimates of changes in one or more assumptions, specifications, or components of the model. It is used to determine the factors that most affect the results, or how robust the results are to changes in assumptions.

3.2 Documentation

Documentation of data, models, and analyses is important, not only for validation studies but also for a better understanding of the model estimates and their quality. Documentation should include information on the policy alternatives under analysis, the data and other inputs, assumptions, interpretations,

and changes made. The information should be specified in sufficient detail so that another person could replicate the results. Otherwise, the utility of the model may be limited to only a few analysts who are experienced in its use. Moreover, as models invariably grow in complexity, even experienced users may not understand the implications of underlying assumptions, or how components of the model interact.

3.3 Communication of Uncertainty

Policy analysts should include information about the uncertainty of their estimates and the sources of that uncertainty in the presentations of results to policy makers. Sampling variability is often the most straight forward to calculate and report, but it neglects errors due to model misspecification, nonsampling errors in surveys, and other factors. Efforts should be made to convey a more complete description of uncertainty through a variety of techniques, including variance calculations, sensitivity analysis, and the proven record of a model.

It is important for policy makers to understand the uncertainty attendant in model estimates, for a number of reasons. First, the information on uncertainty allows them to distinguish differences from those that could occur solely by chance. Second, they can use the information to compare the quality of different estimates. Third, the information may help them decide how much weight to give to a policy analysis, as opposed to other, political considerations.

But there is an even more important reason for agencies that are responsible for policy analysis to measure and communicate the uncertainty in data and model estimates. It is only through such measures of quality that resources to improve the data and models can be justified. For example, in the United States, many national surveys are so sparse that they are unable to provide direct survey-based estimates for all but a handful of states. National policy makers have devolved some programs to the states, such as children's health insurance and welfare reforms, and expect estimates of outcomes to hold states accountable. The estimates can be modeled, but their much larger uncertainty than that for the national average must be communicated. Not only would policy makers know if a reported difference reflects a real difference; they could see the need to invest resources to improve data or models, or both. For example, the sample size of the Current Population Survey has been increased to improve estimates, in each state, of the numbers of children covered by health insurance.

4. Organization of Policy Analysis

Agencies conducting research in support of public policies in a department of government must an-

ticipate policy issues if its research is to inform the policy debate when it occurs. As a result, a policy research agency may invest resources in research on policy issues for which it may never be asked for the results. When consideration of an unanticipated policy comes to the fore, the agency must be able to draw upon a synthesis of research or a review of extant data to provide the required policy information. For that reason, the policy research agency must develop data and research continually, and be cognizant of that developed elsewhere in the important subject matters of its department.

The national government may provide for unsolicited basic and applied research, in particular in academic institutions, but policy analysis rarely, if at all, is a byproduct of this research. Moreover, policy research conducted by private organizations with a stake in the outcome may not be objective. Effective policy analysis, therefore, can neither be delegated to academia nor to interested organizations. Government policy analysis agencies must either conduct or direct the research. In doing so, they may draw upon research conducted by a wide variety of institutions, such as quasi-governmental organizations, universities and academic research institutes, private foundations and other nonprofit organizations, and for-profit consulting firms.

Policy analysis in a department of government may or may not be the responsibility of a distinct, independent unit. For example, the United States has many units in the executive branch with responsibilities for policy analysis, such as the Office of Tax Analysis in the Department of the Treasury, and the Office of the Assistant Secretary for Planning and Evaluation in the Department of Health and Human Services, but no single unit in the Department of Transportation is responsible for policy analysis. Moreover, policy analysis need not be the responsibility of just the executive branch. The US Congress has developed policy analysis capabilities, through such agencies as the Congressional Budget Office, the Joint Committee on Taxation, and the Congressional Research Service, to counterbalance the policy analysis capabilities in the executive branch.

For government policy research agencies, maintaining independence from the political process itself is essential to the agency's effectiveness. 'Research in support of public policy depends for its credibility, in great part, on its conduct by those with a greater interest in the quality of the work than in the substance of the conclusion' (Geweke et al. 1999).

5. Conclusion

Developing information to inform debate on public policies requires data, models, and an understanding of uncertainty in the resulting estimates. By far the

most common problem and the bane of policy analysts is the lack of adequate data. When a panel of the National Academy of Sciences–National Research Council was asked by the US Department of Labor to advise on the development of models to assess policies for retirement income, the panel concluded instead that priority be given to funding data collection and behavioral research in preference to significant investments in large-scale projection models:

Generally speaking, it takes more time to collect new data and analyze them than it does to build a model to use data and research results. There are more than a few instances in the history of policy analysis when complex projection models were built in a span of weeks or months. It is rare that a critical unmet data or analysis need can be supplied in that short a time. (Citro and Hanushek 1997)

Government, the panel concluded, must take the lead in data collection, so that high-quality, comprehensive databases are available for policy analysis use in all sectors—government, private, and academic.

See also: Policy Knowledge: Census; Sample Surveys, History of; Sample Surveys: Methods; Sample Surveys: The Field; Survey Research: National Centers

Bibliography

- Citro C F, Hanushek E A 1991 *Improving Information for Social Policy Decisions: The Uses of Microsimulation Modeling*, Vol. I. Panel to Evaluate Microsimulation Models for Social Welfare Programs, Committee on National Statistics, National Research Council. National Academy Press, Washington, DC
- Citro C F, Hanushek E A 1997 *Assessing Policies for Retirement Income: Needs for Data, Research, and Models*. Panel on Retirement Income Modeling, Committee on National Statistics, National Research Council. National Academy Press, Washington, DC
- Geweke J F, Bonnen J T, White A A, Koshel J J 1999 *Sowing Seeds of Change: Informing Public Policy in the Economic Research Service of USDA*. Panel to Study the Research Program of the Economic Research Service, Committee on National Statistics, National Research Council. National Academy Press, Washington, DC

M. Straf

Policy Knowledge: Advocacy Organizations

Substantive policy knowledge is a set of beliefs that encompass perceptions of (a) the seriousness of various dimensions of a policy problem (e.g., the health impacts of air pollution, the declining quality of health

care), (b) the relative importance of various causes of those problems, and (c) the expected impacts of policy options for changing current conditions. In conflicts between advocacy organizations that pursue their interests through the policy-making process—including interest groups, corporations, administrative agencies, and associations of governments—debate surrounding the various elements of policy knowledge is often technical in nature. These debates frequently employ scientific research, technical policy analyses (such as benefit/cost analyses), and policy evaluations that follow scientific norms of data acquisition and analysis.

There are four main frameworks for examining the nexus between substantive policy knowledge and the political behavior of advocacy organizations: (a) a ‘civics textbook’ approach; (b) an ‘atheoretical skeptics’ perspective; (c) the ‘multiple streams’ approach; and (d) the ‘advocacy coalition framework.’ All vary in their attention to the political process underlying public policy debates and the potential biases inherent in substantive policy information.

1. The Civics Textbook View of the Role of Science

This is basically a normative model, derived from certain fundamental principles of democratic theory. It assigns very distinct roles to three categories of actors in the policy process:

Elected officials are responsible for making basic policy decisions in a manner which reflects the distribution of values in society. They use scientific findings to help them understand trends in various problems, the factors affecting them, and the means of alleviating those problems. Their basic role is to establish clear laws and budgetary priorities for implementation by agencies.

Governmental agencies are composed largely of civil servants who should be politically neutral (i.e., faithfully implement whatever the legislature decides) and who have a special role in fostering applied research in areas of interest to the agency. Political appointees within the agency are responsible for seeing that the agency reflects the government’s priorities—to the extent permitted by law.

Scientists are supposed to be neutral seekers of the truth. Their role is to understand the world and to present this information to policy-makers.

This is a normative model of how science *should* be used in making public policy, and many scientists involved in policy disputes do, in fact, view themselves as ‘objective technicians’ (Meltzer 1976).

According to this model, the major problem impeding communication between scientists and governmental officials is that they inhabit two quite distinct communities with different value priorities, time-frames, and methods for resolving conflict (Dunn

1980). Proposals for reform thus focus on improving communication between the two communities by exchange programs, the development of facilitating or 'translating' institutions, etc.

The textbook model has substantial limitations in practice, in large part because most actors do not behave as the model indicates they should.

Elected officials seldom pass clear laws on contentious issues with substantial technical components because very few have the expertise to obtain a decent understanding of the technical issues. Instead, they tend to pass procedural or 'framework' laws which hand over the problem to an administrative agency without clear policy priorities.

Agency officials are seldom as neutral as the civil servants in the Weberian model of bureaucracy (Knott and Miller 1987). Most agencies have a fairly clear overall mission which tells them to give priority to some values over others. Most officials who join the agency come to accept those priorities, whether out of self-selection or gradual indoctrination. Agencies are often dominated by members of a particular profession or scientific discipline who share the norms of colleagues outside the agency. Thus, most agencies can be expected to sponsor research consistent with their mission and to be skeptical of findings which cast doubt on its wisdom.

Scientists are often not neutral participants. Virtually all scientists operate within a specific 'paradigm,' i.e., a set of often implicit assumptions about basic causal assumptions and proper methods of investigation which guide research (Kuhn 1970). More importantly for our purposes, almost all scientific disciplines contain important normative assumptions which members often come to accept in an uncritical fashion. The normative assumptions behind welfare economics and benefit/cost analysis have been widely discussed (Rhoads 1985). With respect to nuclear waste disposal, Barke and Jenkins-Smith (1993) provide evidence that biologists perceive significantly greater risks than physicists, chemists, and engineers. The latter think basically in terms of dose-response curves, while biologists are more wary of the effects of *any* dose on living organism.

Moreover, scientists are often drawn to applied—rather than basic—research because they want to help solve a particular problem. Having a demonstrable effect on policy normally requires the accumulation of results over an extended period of time (Weiss 1977). The more neutral and 'apolitical' scientists are unlikely to remain interested in an issue long enough to have such an impact. Thus the most active scientists in a particular dispute are likely to be those who have been involved the longest and who are most committed to defending a particular point of view.

The end result is that scientists who have something to contribute to important policy disputes are seldom neutral (Mazur 1981). The argument does not imply that they manipulate or falsify data. Instead, disci-

plinary paradigms, the values underlying their discipline, and their desire to solve particular problems affect the topics they choose to research, the variables they focus on, the methods they utilize, where they place the burden of proof in cases of uncertainty, and how quickly they present various results. For example, wildlife biologists are much more likely than engineers to look for species in trouble because their disciplinary norms define species extinction as a serious problem. They are more likely to look to human technological interventions as a likely explanation because they tend to respect the beauty of natural systems. In contrast, engineers assume they can improve on nature. Members within each discipline will readily present results which are congruent with these assumptions, while incongruent results are likely to be interpreted as tentative and in need of further verification (Brown 1977).

2. The Atheoretical Skeptics' View of the Role of Scientists in Policy Disputes

Over the years, a number of studies have seriously questioned the textbook view that scientists participating in policy are completely neutral. First, Weinberg (1972) argued that some topics—such as risk assessments of very improbable, catastrophic events—are essentially in the realm of 'trans-science.' Second, as argued above, there is increasing recognition that the topics chosen for research, the allocation of the burden of proof in areas of uncertainty, and the presentation of results are affected by scientists' disciplinary paradigms (Stewart 1986), organizational interests (Primack and von Hippel 1974, Jasanoff 1987), and/or policy concerns (Nelkin 1971).

Although suggestive, most of these studies have suffered from two serious limitations. First, they have failed to articulate a comprehensive theory (or theories) of the role of 'advocacy science' in the overall policy process. Instead, they have usually limited themselves to identifying certain institutional/organizational features that inhibit or exacerbate the extent of advocacy (Mazur 1981). But there is very little recognition of the role of broader social norms, economic conditions, and the myriad of advocacy organizations involved in most policy subsystems. Second, almost all of this research has consisted of essays or of rather traditional case studies which (a) lack explicit theoretical grounds for case and variable selection and (b) utilize quite subjective methods of data acquisition and analysis. The result has been a plethora of interesting arguments and illustrative examples, but very little systematic, intersubjectively reliable analysis of the extent of advocacy science, the factors affecting it, and its effects on the overall policy process.

3. *The Multiple Streams Framework of the Policy Process*

Drawing upon the 'garbage can model' of organizational choice (Cohen et al. 1972), Kingdon (1984) developed an approach to agenda-setting and policy formulation that deals explicitly with the role of substantive policy information. In his view, policy-making can be conceptualized as three largely unrelated 'streams':

(a) A *problem stream* consisting of information about real world problems and the effects of past governmental interventions. This would include large databases of problem severity and causes gathered by governmental agencies and research organizations, as well as policy analyses of the benefits and costs of previous attempts to address those problems. This information is treated as relatively unbiased.

(b) A *policy stream* consisting of proposals to solve one or more problems. This is the province of policy analysts in government think tanks and consulting firms, as well as many legislative staff. Kingdon notes the tendency for analysts to become infatuated with a particular solution and then to seek to apply it to a wide variety of problems. For the past 30 years, for example, many resource economists have viewed tradeable permit schemes as the solution to virtually all environmental problems (Tietenberg 1984).

(c) A *political stream* consisting of elections, legislative leadership contests, and other openly political competition.

According to Kingdon, major policy reforms result when a 'window of opportunity' joins the three streams. The window may be opened by a 'focusing event,' such as an airplane crash or oil spill, that dramatically illustrates the potential seriousness of a problem, or it can be opened by a political election that brings a new leader with a new agenda to power. In response to the window, major policy change will result if the policy community develops a proposal that is financially and technically feasible and if elected officials find it advantageous to approve it.

The multiple streams approach has a number of promising features. It gives a prominent role to substantive policy information about real world problems and the impacts of past governmental interventions. It was one of the first frameworks that treats researchers and policy analysts in consulting firms, think tanks, and universities as potentially important political actors. And it acknowledges the role of serendipity—particularly in the case of focusing events—in the policy process.

On the other hand, it has several limitations. First, we are not convinced that the problem stream is nearly as neutral and apolitical as Kingdon suggests. Bureaucracies gather information on some aspects of a situation and not on others. The US Department of Agriculture gathers much more extensive information on crop production and prices than it does on nitrogen

runoff from fields into streams and groundwater basins. Second, the conditions creating windows of opportunity need further analysis. Third, the multiple streams framework contains virtually no falsifiable hypotheses, and thus very little has been learned about its strengths and limitations over time (cf. Kingdon 1995, Chap. 10).

4. *An Advocacy Coalition Framework of the Policy Process*

The advocacy coalition framework (ACF) starts from the premise that political elites concerned with a specific problem or policy domain—such as transportation, education, health, or air pollution—will form relatively autonomous subsystems composed of interest-group leaders, administrative agency officials, researchers, and a few legislators with a specific interest in the topic. These subsystems (or policy communities) usually involve actors from multiple levels of government. Once a subsystem is formed, policy disputes become increasingly technical and specialized as actors learn more about the complexity of problems and solutions. In such a system, the role of legislatures and political parties is primarily to aggregate the decisions made by policy specialists and to provide some general guidance.

The ACF views policy change over time as primarily the result of competition among advocacy coalitions within a policy subsystem (Sabatier and Jenkins-Smith 1993, 1999). An advocacy coalition consists of interest-group leaders, legislators, agency officials, researchers, and even journalists who share a set of basic beliefs (policy goals plus critical perceptions of causal relationships) and engage in some degree of coordinated behavior in an effort to make governmental policy more consistent with those beliefs. Conflict among coalitions is mediated by 'policy brokers,' i.e., actors more concerned with fashioning an acceptable compromise than with achieving specific policy goals. While the framework focuses on competition among coalitions within the subsystem, changes external to the subsystem (such as fluctuations in socioeconomic conditions) and stable system parameters (such as constitutional rules) play an important role in major policy change.

The belief systems of advocacy coalitions are assumed to be hierarchically organized. At the highest/broadest level, the 'deep core' of a coalition's belief system consists of fundamental normative beliefs, such as the familiar Left/Right scale, that operate across virtually all policy domains. Within any given policy subsystem, however, it is the 'policy core' and the 'secondary aspects' that are most critical. The former consists of basic positions, some of them purely normative (e.g., the relative importance of different values, such as environmental protection vs. economic development), while others are a mixture of normative

and empirical (e.g., the proper scope of governmental vs. market authority for realizing those values), that operate across most or all of the policy subsystem. These policy core positions are very resistant to change, are only intermittently the subject of policy debate, and are usually changed as a result of perturbations external to the subsystem, although long-term 'enlightenment' may also play a role (Weiss 1977). Science plays a much more important role in the secondary aspects of coalitions' belief systems, as these involve disputes over the seriousness of a problem or the relative importance of various causal factors in different locales, the evaluation of various programs and institutions, and specific policy preferences.

The ACF assumes that members of a coalition will readily accept new evidence consistent with their views and seek to discount information which conflicts with their perception of the seriousness of a problem, the relative importance of various factors affecting it, or the costs and benefits of different alternatives (Lord et al. 1979). The result is 'a dialogue of the deaf?' in which members of different coalitions talk past each other. Given that policy core beliefs are resistant to change, the composition of coalitions is hypothesized to be stable over periods of a decade or more.

The ACF explicitly rejects the assumption that most bureaucrats and researchers involved in a policy area will be neutral. Some may well have no strong policy preferences, at least initially. But the framework contends that, as conflict between coalitions increases and as the interrelationships among sets of beliefs become clearer over time, initially loose groups with amorphous beliefs will coalesce into increasingly distinct coalitions with coherent belief systems; in the process, most neutral actors, particularly university scientists, will drop out. The ACF thus contends that, in well-developed subsystems, most agency officials and researchers who are active will be members of specific coalitions in terms of sharing a set of policy core beliefs and acting in concert to some degree (Sabatier and Zafonte 2000).

Policy-oriented learning involves relatively enduring alterations of thought or behavioral intentions that result from experience and/or the assessment of new information involving the precepts of belief systems (Hecklo 1974). Learning *within* coalitions is relatively easy, as members share the same value-premises and are looking for the most effective means to those ends. Learning *between* coalitions is much more difficult, as people from each coalition have quite different values and distrust each other. Nevertheless, the original versions of the ACF hypothesized three conditions facilitating such learning (Sabatier and Jenkins-Smith 1988).

(a) *Issues on which there is an intermediate level of conflict.* Issues have to be important enough to generate sufficient research—usually by members of several coalitions, as well as neutrals. On the other hand, issues involving conflict between the core beliefs

of different coalitions (e.g., the rights of non-human beings or the ability of humans to improve on nature) generate more heat than light. Learning across coalitions is thus most likely on issues involving important secondary aspects of the relevant belief systems.

(b) *Issues involving primarily the natural sciences.* Across-coalition learning is generally easier in the natural than in the social-behavioral sciences because the theories and accepted methods are better established and the objects of study are not themselves actors in the policy debate.

(c) *The existence of a forum* that is (i) prestigious enough to force professionals from different coalitions to participate and (ii) dominated by scientific norms. The latter assures a general consensus on the appropriate rules of evidence and a minimum of *ad hominem* attacks, as well as some attention to underlying assumptions. Possibilities for professional forums include 'blue ribbon' committees appointed by the National Research Council or a professional association, studies by organizations with a strong reputation for neutral competence (e.g., the Congressional Budget Office), or, in some cases, inter-agency technical advisory committees.

The following set of conditions have been developed for successful professional forums, i.e., those in which experts from the competing coalitions come to consensus on technical and/or policy issues placed before it. In general, such forums are most useful when a fair amount of scientific/technical information on various aspects of the topic exists, but the conclusions of different studies vary and the validity of much of the evidence is questioned by members of opposing coalitions. The arguments below assume that the essential task of a professional forum is to convince coalition experts that professional norms require the alteration of one or more beliefs important to a coalition.

Composition. In order for scientists with very different points of view to come to consensus and for that consensus to be accepted by the major coalitions, the technical advisory committee should include both (a) scientists clearly associated with each of the major coalitions and (b) neutral scientists. The chair should come from the latter.

The ACF assumes that the various coalitions involved in important policy disputes have scientists (or technical experts) whom they trust, presumably because they share most of the coalition's policy core beliefs and can help the coalition understand the scientific aspects of such disputes. Many interest-group leaders, legislative personnel, and agency officials who constitute the leadership core of coalitions have an intuitive awareness of the role of values in determining the agenda of technical advisory committees and how uncertainty is treated. Thus they are very unlikely to accept the recommendations of any technical advisory committee on which their point of view has not been argued by someone whom they trust.

And, if they do not trust the committee's report—particularly in a relatively decentralized political system like that of the USA—they almost always can find some decision-making venue that will enable them to circumvent or obstruct the committee's recommendations. In short, selecting a committee composed entirely of relatively 'neutral' scientists—which we perceive to have been the strategy of the National Academy of Scientists (Boffey 1975), for example—may facilitate short-term consensus but will probably fail in the long run.

On the other hand, the technical advisory committee needs to be chaired by a neutral (and probably to include a few other neutrals) whose task is to impose professional norms on the debate regarding acceptable evidence, methodologies, etc., and to indicate to coalition experts when a professional consensus is beginning to emerge. Nonprofessionals must be excluded from these deliberations in order that scientific norms can prevail in weighing the evidence for—and the assumptions behind—different points of view.

Funding and Sponsorship. Funding must come from an institution which is not perceived as being controlled by a single coalition. This will usually require funding either by (a) a neutral foundation or (b) multiple agencies representing the various coalitions. This is important because policy participants are unlikely to trust a committee funded—and therefore potentially controlled—by their opponents. Since many agencies are perceived as being part of a specific coalition (Sabatier and Zafonte 2000), single-agency technical advisory committees will lack legitimacy in the eyes of opposing coalitions.

Duration. A forum should meet at least half a dozen times over a year or so. It takes time for scientists from different coalitions to analyze their hidden assumptions, to evaluate the evidence critically, and to begin to trust each other. One-shot committees of short duration will probably not work.

Context of Policy Stalemate. A forum will be successful only in a context of policy stalemate, i.e., when each of the coalitions views a continuation of the status quo as unacceptable. In any policy dispute, arriving at a compromise acceptable to everyone requires that all the coalitions view the status quo as unacceptable. If any should view the status quo as acceptable, then it will be much less willing to compromise and may not even participate.

The assumption thus far has been that lasting policy change requires genuine learning between/across coalitions. In a relatively open and decentralized political system, such as the USA's, coalitions that do not agree with the conclusions of a professional forum have numerous points of appeal (courts, legislatures, agencies at different levels of government) which have different biases. In more centralized systems, such as UK's, where routes of appeal are severely restricted, it may be possible for a coalition to change

policy simply by convincing a policy broker of the merits of its point of view without having to change the views of the other coalition(s).

5. Summary

This article has presented four conceptual frameworks that seek to understand the role of advocacy organizations in processing substantive policy information of a relatively technical nature in policy disputes, particularly in the USA. Both the civic textbook model and the multiple streams model view substantive policy information as being relatively unbiased, but the multiple streams has a much more sophisticated understanding of the policy process. In contrast, both the atheoretical skeptics and the ACF view substantive policy information as 'filtered' through organizational, disciplinary, and coalition lenses, but the ACF has a much more sophisticated model of the policy process in intergovernmental policy subsystems. This sophistication, however, comes at the cost of parsimony. All four perspectives in combination can be useful in generating a clearer understanding of the role of policy knowledge in the behavior of advocacy organizations.

See also: Advocacy and Equity Planning; Advocacy in Anthropology; Interest Groups; Labor Unions; Lobbying; Policy Knowledge: Nonprofit Institutions; Think Tanks

Bibliography

- Barke R, Jenkins-Smith H 1993 Politics and scientific expertise: Scientists, risk perception, and nuclear waste policy. *Risk Analysis* **13**(4): 425–39
- Boffey P 1975 *The Brain Bank of America*. McGraw-Hill, New York
- Brown H 1977 *Perception, Theory, and Commitment*. University of Chicago Press, Chicago
- Cohen M, March J, Olsen J 1972 A garbage can model of organizational choice. *Administrative Science Quarterly* **17**: 1–25
- Dunn W 1980 The two-communities metaphor and models of knowledge use. *Knowledge* **1**(June): 515–36
- Hecl H 1974 *Social Policy in Britain and Sweden*. Yale University Press, New Haven, CT
- Jasanoff S 1987 Contested boundaries in policy-relevant science. *Social Studies of Science* **17**: 195–230
- Kingdon J 1984 *Agendas, Alternatives, and Public Policies*. Little, Brown, Boston
- Kingdon J 1995 *Agendas, Alternatives, and Public Policies*. Little, Brown, Boston
- Knott J, Miller G 1987 *Reforming Bureaucracy*. Prentice-Hall, Englewood Cliffs, NJ
- Kuhn T 1970 *The Structure of Scientific Revolutions*, 2nd edn. University of Chicago Press, Chicago
- Lord C, Ross L, Lepper M 1979 Biased assimilation and attitude polarization: The effects of prior theories on subsequently considered evidence. *Journal of Personality and Social Psychology* **37**: 2098–109

- Mazur A 1981 *The Dynamics of Technical Controversy*. Communications Press, Washington, DC
- Meltsner A 1976 *Policy Analysts in the Bureaucracy*. University of California Press, Berkeley, CA
- Nelkin D 1971 *Nuclear Power and Its Critics: The Cayuga Lake Controversy*. Cornell University Press, Ithaca, NY
- Primack J, von Hippel F 1974 *Advice and Dissent*. Basic Books, New York
- Rhoads S 1985 *The Economist's View of the World*. Cambridge University Press, London
- Sabatier P, Jenkins-Smith H (eds.) 1988 Symposium on 'policy change and policy-oriented learning.' *Policy Sciences* 21(2-3): 123-77
- Sabatier P, Jenkins-Smith H (eds.) 1993 *Policy Change and Learning: An Advocacy Coalition Approach*. Westview Press, Boulder, CO
- Sabatier P, Jenkins-Smith H 1999 The advocacy coalition framework: An assessment. In: Sabatier P (ed.) *Theories of the Policy Process*. Westview Press, Boulder, CO, pp. 117-68
- Sabatier P, Zafonte M 2000 Are scientists and bureaucrats members of advocacy coalitions? In: Sabatier P (ed.) *An ACF Lens on Environmental Policy*. Under review
- Stewart J A 1986 Drifting continents and colliding interests. *Social Studies of Science* 16: 261-79
- Tietenberg T 1984 *Environmental and Natural Resource Economics*. Scott Foresman, Glenview, IL
- Weinberg A 1972 Science and trans-science. *Minerva* 10: 209-22
- Weiss C 1977 Research for policy's sake: The enlightenment function of social research. *Policy Analysis* 3 (Fall): 531-45

P. A. Sabatier and M. A. Zafonte

Policy Knowledge: Census

A census is the count of the entire population of a nation, normally tabulated by jurisdictions within the nation such as provinces, districts, cities, towns, or villages. A modern census also collects social, economic, and housing characteristics as part of the population count. Census information is heavily used in state planning and in public policy. The social sciences, especially demography, sociology, history, and economics, draw from census data across a wide range of research topics: fertility, migration, race and ethnicity, economic growth, and many others. However, with a few exceptions, the census itself has not been a research subject in social science; for example, there are no cross-national comparisons analyzing differing ways in which the census is involved in social, political, and economic processes.

1. History/Definition

The archaic definition of a census is 'poll tax,' indicating its early association with the state's function of revenue extraction. Although the term census dates to ancient Rome, where it referenced the registration

of citizens and their property, the idea of a census predates Roman times. For example, *Numbers*, the fourth book of the Hebrew Bible, starts with the Lord instructing Moses: 'Take ye the sum of all the congregation of the children of Israel, after their families, with the number of their names, every male ... from twenty years old and upward.' That is, take a census of those 'able to go forth to war.'

It is probable that taxation or military conscription or both gave rise to other census operations indicated by archeological records, such as the *Nepohualco* or 'counting place' built by the Chichimecas a millennium ago. This site, near today's Mexico City, consists of stones deposited by each inhabitant of the region, an early census databank. Other archeological records from various ancient civilizations indicate similar preoccupation by the state with carrying out systematic statistical studies. In these ancient records it is difficult to detect all of the state purposes that might have been served by a census. Certainly there was interest in economic data, as is suggested by statistical records of harvest and grain storage. Ancient Chinese records, dating from the Han dynasty (206 BC to AD 220), indicate that detailed population counts were made after each harvest. Subsequent systematic census records appear throughout Chinese history, and many scholars concur that China should be credited with the first comprehensive census (Domschke and Goyer 1986).

There were attempts to establish a periodic census in early modern times, such as the ambitious seventeenth century proposal for an annual census directed to France's Louis XIV:

Would it not be a great satisfaction to the king to know at a designated moment every year the number of his subjects, in total and by region, with all the resources, wealth & poverty of each place?... [Would it not be] a useful and necessary pleasure for him to be able, in his own office, to review in an hour's time the present and past condition of a great realm of which he is the head, and be able himself to know with certitude in what consists his grandeur, his wealth, and his strengths? (cited in Scott 1998, p. 11).

In the modern period, however, it was the United States that initiated the first periodic census. The US Constitution (1787) requires that an enumeration of the population be taken every 10 years. This decennial census, first taken in 1790, served two state-building functions. One was periodically to reallocate seats in the House of Representatives, the legislative branch to which positions are allocated proportionate to size of population in each of the member states. Because the population might grow at different rates from one state to the next, provision was made to take a census every decade and then to calculate how many legislative seats would go to each state. The second nation-building task addressed the challenge of geographic expansion. The country was expected to expand into new territories, and these areas were to join the nation

on equal footing with the original states. The decennial census measured population growth and its geographic dispersion and thereby served as the mechanism regulating how territories were added to the state.

Though the US was early in establishing a periodic census, most European nations shortly followed in instituting some form of regular census to assist in nation-building and policy-making—the United Kingdom and France in 1801; The Netherlands, shortly after independence, in 1830; Sweden in 1855; Spain in 1857; Italy in 1861; the German Empire in 1871; and the Russian Empire in 1897. In nearly every instance, the postcolonial nations of Latin America, Asia, and Africa established a periodic census as an early step in nation building. By the end of the twentieth century only a handful of countries do not have a census.

2. Census and Public Policy

A wide array of public policy purposes are served by a regular official count of the population and its characteristics. Before turning to these purposes, it is necessary to distinguish between a census and a sample survey. The vast majority of state-generated statistical information comes from sample surveys, which elicit information from a part of the population in order to estimate characteristics of the entire population. A census customarily describes the effort to count everyone in the population. There is an important statistical interaction between sample surveys and the census, for the latter provides the frame necessary to draw samples for survey purposes. It also provides the basic denominator that serves as a control to determine whether a sample is demographically and geographically representative of the country. That is, does the sample have the correct proportion of age, race, ethnic, gender, and other demographic categories that have been measured in the population census, and does it reflect how the population is distributed geographically, again as measured in the census. The close relationship of a census and surveys indicates that many policy functions that rely on sample data are, in fact, drawing from a prior census.

An important distinction between the census and the survey is that the former allows estimates of population or housing characteristics to be made at comparatively small levels of geography. Policy purposes that require fine geographic detail will directly draw on census information. Policy purposes that are targeted at higher levels of aggregation tend to rely on sample survey data and thus have an indirect relation to the census.

A census supports policy-making in four broad arenas: election systems, taxation, public funding of social services, and social control.

2.1 Election Systems

When the basis of election to a legislature is size of the population being represented, census information is used to draw the boundaries of election units. As indicated above, this was the primary purpose for which a census is required by the US Constitution. It is now the practice in dozens of democratic countries, and not just for national legislatures. Election districts for city governments and other local and regional governments are frequently drawn to ensure approximate equal population size for each geographic area that constitutes an election unit.

Size is not the only population characteristic that can be taken into account in drawing election districts. Demographic traits such as ethnicity, language, and race have been used. This occurs when the state wants to ensure that the legislative body is ‘demographically representative’ of the population at large, for instance, that racial minorities have the opportunity to elect someone from their race to the legislature. Under these circumstances, census data are commonly used to monitor compliance with election laws. In the United States, for instance, census information on ancestry, gender, race, ethnicity, language spoken at home, place of birth, and year of entry into the country are critical for monitoring compliance with voting and civil rights laws.

When political party affiliations vary across different ethnic, language, or racial groups—as they often do—the use of the census to promote ethnic, linguistic, and racial balance or access will invite partisan disagreements. Census counts become the target of such disagreements, as in the events that led to the Nigerian civil war. Even census methodology can come under attack, as has happened in the United States in connection with its 1990 and 2000 census.

Basing political representation on population counts also invites jurisdictions to inflate their numbers. This incentive confronts census taking with serious methodological challenges. Unless there is a corresponding disincentive—such as basing tax rates on population size—a census can be vulnerable to strong political pressure by those with a direct stake in high counts.

2.2 Taxation

The most direct relation of a census to taxation policy occurs when taxes are collected from a jurisdiction proportionate to its population size. Thus a central government might calculate the tax obligation of its administrative units depending on the size and wealth of the unit, and use a census to make the determination.

A more indirect use of census (and survey) data occurs when the state measures the economic life of

the nation, using that information to help design a taxation system—whether to tax manufacturing activity, natural resources, individual income, property, at what rates and in what combinations. For example, population projections based on census data are used to design social security tax rates and to estimate probable social security payment levels well into the future.

2.3 Public Funding and Planning

Social welfare policies are generally targeted to specific population groups, such as families in substandard housing, children in poverty, the physically disabled, minorities historically disadvantaged or mistreated, victims of particular diseases. Assessing the size and geographic distribution of groups targeted for welfare payments or subsidies is the task of a nation's statistical system. Frequently public funding formulae are based on census counts of particular population groups (Nathan 1999). In the United States, for instance, in 1999 approximately \$US185 billion in medical, education, housing, transportation, unemployment, and other payments depend in part on census data. Similar funding formulae using census information occur in many countries with advanced social welfare systems.

When public funds are allocated to target groups based on their size, they have a strong interest in being fully counted. But these are also population groups that are difficult to reach in standard census operations, and even if reached, will sometimes resist being counted because of concerns that census information can be used against them. This situation has given rise to advocacy organizations that work with and put pressure on census bureaus to assure as full a count as possible of their particular constituencies. Where this has occurred, census taking is no longer the domain only of professional statisticians and their government minister, but becomes of much more generalized political interest.

In addition to targeted welfare spending, national governments frequently use census information for program planning across many policy domains such as public transportation, health and education services, job training, police and fire protection, sanitation and sewage disposal, and land use.

The census also plays a role in national security policy. As indicated above, since biblical times the state has relied on census information to assess the numerical strength of its fighting age population, and also at times to design and implement actual conscription. Of course, a reliable count of the fighting strength of one nation's population will be used by another nation to calculate whether that nation poses a military threat. The first US census in 1790 was declared by the nation's leaders to have missed a sizeable portion of the population, and the Secretary of State took care to announce that fact to European

powers that might otherwise have concluded that the young nation was too weak to protect its independence. Of course not just population censuses but also economic censuses help the state assess its military capacity and that of other nations.

The relative strengths and weaknesses of contending groups within a nation are also revealed by census information, and have often played a role in strategic choices that precede internal warfare. The suppression of census data in several African countries reflects a decision by the ethnic group in power to disguise the comparative numerical strength of ethnic groups challenging its rule. In the period preceding the Civil War in the United States the more rapid growth of 'free-states' compared to 'slave-states,' as revealed in the 1850 census, was a factor leading to the succession of the South from the national union.

2.4 Social Control

More generally the state uses the census as an instrument of social control, as it was routinely used during the colonial era. Great Britain, as early as the seventeenth century, introduced a census in many of its colonies in order to control more effectively its distant subjects and to assure that resource extraction proceeded smoothly. Decisions about which territories needed policing, required additional labor, could benefit from infrastructure investments, and so forth were facilitated if a reliable census could be taken. Great Britain and other colonial powers were early practitioners of what has now become commonplace. Social control policies that police borders, population movement, and even reproduction and fertility are fashioned in part on the basis of census information.

The boundary between routine social control and abusive exercise of power by the state is not easily drawn, but certainly history offers many examples of the census as an instrument of the latter. Family planning policy and fertility regulation in China and India, for instance, is often cited in this regard.

Population displacement is a further example. Census data have been used to place restrictions on freedom of movement or to force population groups to live in places they would not have chosen. An important instance of the latter is the role of US census data in the nineteenth century relocation of native American Indian tribes from their homelands to reservations. US census data were also used to facilitate the internment of Japanese-Americans at the beginning of World War Two, because they were said to pose a security risk should Japan invade America's west coast.

Census data have also been used in genocidal policies, the most tragic example being the Holocaust. Population registers or special censuses identifying Jews greatly assisted in the location of Jews and their transport to concentration and extermination camps. The 1939 German census, for example, was used to

identify specific individuals. In The Netherlands small-area tabulation of a census of religions taken in 1930 was used in dot-maps of Amsterdam to indicate the density of the Jewish population. Poland, France, and Norway offer additional evidence linking census data to Holocaust operations (Seltzer 1998).

These historical instances of gross misuse of census information have had one beneficial outcome. There is now much more professional and political sensitivity to issues of data confidentiality, the independence of statistical agencies, and related ethical considerations (Seltzer 1994).

3. Census Data and Democracy

Census information connects to politics, and especially to democratic politics, in ways other than its use by the government to design and implement public policy. Statistics and especially statistical series also advance key features of democratic politics. One purpose is helping the electorate to assess government performance and hold it accountable. Another is in providing a platform for social reform. A third is in the assertion of group interests (Prewitt 1999).

3.1 Democratic Accountability

The theory of electoral accountability holds that competitive elections offer to the electorate alternative portrayals of how well the government in power has performed or will perform against future challenges. Aspirants for political office present themselves in terms of their past accomplishments and their promises of future accomplishments in managing the economy, protecting the nation's security, and generally enhancing national well-being. Voters then elect, re-elect, or evict accordingly.

This simple model presumes that the voters have some way of assessing government performance. Statistical trends and social indicators play a major role. They indicate whether the economy is growing or stagnating, whether education or health or housing is improving, whether crime rates are down, whether the environment is being protected, and so forth. Arguments about who can take credit for improvements, or should be blamed for failure, are the common currency of competitive elections and are often advanced by citing statistical trends.

Consider, for example, the sharp political debate over immigration policy in many wealthier countries as they face heavy demands for entry from Central Asia, Northern Africa, and Latin America. These debates are framed in terms of the 'proper' balance between native population groups and newer immigrant groups, with census and related data providing the subtext. Governments are then judged in terms of how well they protect racial purity with anti-immigration policies; or, conversely, display humanitarian compassion or perhaps, as in the United States

at present, replenish the labor pool with more open immigration policy. In either instance, as in so many areas of government programming, long-term statistical trends are invoked when assessing policy.

3.2 Social Reform

Democratic theorists have long been interested in how issues get on the political agenda, especially issues that matter most to the unorganized or disenfranchised interests of society. For instance, how do issues such as child labor or male-female wage differentials become politically salient? One route to the political agenda is through social reform action that takes as its starting point statistical profiles. Since the early days of industrialization, reform groups have used statistics to document the social conditions that they were committed to eradicating (Cohen 1982). Examples are poverty laws, prison reform, child abuse, racial discrimination in education, housing, and employment. Reform activists mobilize political participation and inform public debate by transforming previously unnoticed social conditions into highly visible social injustices. In this way resource-poor groups compensate for their lack of political organization or low levels of political participation. They use state provided statistical information as a political resource (Lacey and Furner 1993).

3.3 Group Interests

Census information can be the basis on which to establish group identity. When, for instance, the census differentiates the population into its component racial, ethnic, or tribal groups, it promotes these categories in public consciousness and in public policy. To be measured is to be politically and socially visible, and this visibility can become the basis for group representation. Public policies such as affirmative action targeted to India's lower castes or racial minorities in many countries are formulated on the principle of statistical proportionality. That is, there is pressure to ensure that racial or ethnic groups historically discriminated against are represented in high status occupations or the legislative bodies proportionate to their numbers in the general population.

This process, once started, gathers momentum. More and more groups insist on being fully counted in a census. The leaders of these groups then assert political claims on the basis of the counts. A census which is judged to undercount particular groups, such as the homeless or the linguistically isolated, is attacked by advocates for the undercounted population groups. The census and even census methodology can become a political battlefield.

The increasingly central role of statistical proportionality in democratic politics and public policy is welcomed by some but resisted by others. Proponents argue that decades or even centuries of injustice can be

reversed through social policies that correct for a history of discrimination. Because this discrimination was and even remains group-based, it requires group-based measures to overcome it. Others express concern that attaching policy benefits and penalties to persons on the base of their race or ethnicity is to abandon the liberal principle of individual rights in a democracy.

The census and what it measures is of great political interest to both sides of this debate. What is unmeasured cannot easily become the target of public policy, and certainly cannot be used in policies that employ statistical proportionality to achieve social purposes.

4. The Census as a Subject of Public Policy

The census, being an instrument of nation building, policy-making, public planning, resource allocation, and democratic governance, will itself be the focus of policy. The overriding policy consideration for a government considering its census is how to strike the proper balance between accountability and autonomy. On the one hand, the census is instituted and funded primarily because it serves government purposes. If it serves no useful purpose it will be underfunded and eventually discarded. On the other hand, the census shares with other scientific enterprises the requirements of professional independence. When a government does not respect the independence of census operations, it places at risk the credibility of the counts and can easily be accused of manipulating census numbers in order to protect itself or enhance its prestige.

This tension between policy accountability and scientific autonomy plays itself out in multiple ways within the policies that govern a census. At one extreme, census results and other official statistics are only very selectively released. For example, China during its Cultural Revolution (1966–76) tended to suppress all statistics or to make them available only through highly controlled government reports. In the former USSR, the 1937 Population Census was never made public, in part to keep from public attention the magnitude of the 1932–3 famine. In less extreme instances, particular statistical series are presented in a manner designed less to inform than to persuade the public. In democracies, for example, the release of inflation rates or poverty measures can be timed to benefit the incumbent party during an election period. Practices such as suppression of data or political manipulation of its presentation and release will of course undermine public confidence in census counts and related statistics.

At the other end of the policy continuum is the assumption that the census is a public good. From this perspective, national statistics serve broad social purposes that reach well beyond the immediate needs of the government in power. This perspective, of

course, is compatible with granting considerable independence and autonomy to a census bureau staffed less with political appointees than with professional statisticians.

In most countries, census design and operations exist uneasily in a policy space that simultaneously includes close government oversight, on the one hand, and acceptance of the importance of scientific autonomy, on the other. The weight assigned to one or the other will vary depending on the ideology of the government in power. When the conservative government of Margaret Thatcher came to power in the United Kingdom, an explicit decision was taken to narrow the functions of the statistical service and make it serve specific, immediate policy needs. In the United States, a number of questions previously asked were eliminated from the 2000 census forms because they were not directly related to a federal program. This narrowing of the function of the US census dismayed demographers and sociologists who pointed out that the census was becoming less analytically useful, that is, was less of a public good.

The mission a census is expected to serve—narrow, immediate, and even secret government functions at one extreme or broad public good functions at the other—is the first-order policy issue that confronts a government as it establishes and funds the periodic census. There follow, then, a host of subsidiary policy issues. How are government statisticians expected to balance their duties as civil servants with their obligations to their professional peers? What criteria should be used to select the director of a census operation? How much autonomy is granted a census agency in its ongoing effort to improve methods and concepts, especially in redefining politically sensitive measures in areas such as race and ethnicity or income and poverty? What rules govern the timing and nature of data release? What is the appropriate tradeoff between data confidentiality and respondent privacy, on the one hand, and the society's demand for detailed information, quickly provided, on the other? Should a census agency only provide raw data, or should it have its own analytic capacity and thereby provide reports and interpretations? How can appropriate government oversight of a census operation guard against waste of public money or agency incompetency without itself being accused of partisan intrusion in the census?

All of these and other issues are the proper subject of political debate in the articulation of policies that govern the funding, organization, and implementation of a census. The issues are similar in nature to those that arise in the government sponsorship of science or the arts, each being an activity that requires a high degree of professional autonomy but each also being an arena of great political sensitivity.

With respect to a census, or statistical activities more broadly, thoughtful observers have articulated useful policy principles. In essence these principles

recommend that the statistical products should be designed to serve the needs of multiple political interests and not just those of the government currently in power, and should be sensitive to future policy issues rather than focused only on the present. Success in institutionalizing this principle will maintain both the relevance and the credibility of a census (Seltzer 1994).

5. *The Census as a Subject of Social Science*

The social sciences make heavy use of census data across a wide variety of topics. The quantitative study of society is severely hampered in the absence of census information. It is perhaps surprising, therefore, that studies of the census itself are so few. That is, though the census is everywhere the subject of public policy, it is seldom the subject of public policy analysis.

The exceptions to this general proposition are important, and should not be overlooked. There are historical treatments and case studies of particular censuses that are often rich in analytic detail (Anderson 1988, Anderson and Feinberg 1999, Arowolo and Daramola 1982, Medvedev 1989). There are studies of particular policy issues that emerge in the census context, such as analysis of confidentiality or of the census as an instrument of social control or the role of the census in revenue sharing schemes.

What is generally lacking, however, is social science theory about the emergence, role, function, and consequence of the census. It is largely assumed that the modern nation-state needs a census, else it would not have emerged in nineteenth century Europe and North America. This assumption, however, has not been subjected to investigation in political sociology. Theoretically rich treatments of nation building and citizenship (Bendix 1964) or of the state's effort to improve society (Scott 1998) provide a platform for studies of the political causes and consequences of the census, but with few exceptions (Alonso and Starr 1987) the work is yet to be undertaken.

The political economy literature is similarly deficient in its theoretical treatment of the census as an object of inquiry. It is easy to find assertions that the modern economy requires fine-grained demographic data for making decisions about labor market dynamics, location of retail outlets, types of products that will meet future consumer needs, and so forth. From this latter assertion, it is an easy step to claim that census data are a public good and to invoke microeconomic theory to explain the government's decision to fund a census. Using the United States as a case study, one sustained examination of this hypothesis reports that public good theory offers an insufficient explanation of the rise of the modern census (Kelman 1985).

Neither political sociology nor political economy has yet turned its analytic tools on the census as a feature of the modern state and modern economy.

This will likely be corrected. If metaphors about how the 'knowledge society' will replace industrial society are taken seriously, so also is the assertion that information is the emerging infrastructure for the knowledge revolution just as railroads or canal systems were the earliest infrastructure for the industrial revolution. An accurate and credible census is basic to establishing an information infrastructure and thereby requires systematic attention by the social sciences.

See also: Policy Information; Policy Knowledge; Epistemic Communities; Policy Networks; Population Cycles and Demographic Behavior; Population Forecasts; Population Size Estimation and Capture–Recapture Methods

Bibliography

- Alonso W, Starr P (eds.) 1987 *The Politics of Numbers*. Russell Sage Foundation, New York
- Anderson M J 1988 *The American Census: A Social History*. Yale University Press, New Haven, CT
- Anderson M J, Feinberg S E 1999 *Who Counts*. Russell Sage Foundation, New York
- Arowolo O, Daramola O (eds.) 1982 *Philosophy of Population Census in Nigeria*. National Population Commission, Lagos, Nigeria
- Bendix R 1964 *Nation-Building and Citizenship*. John Wiley & Sons, New York
- Cohen P C 1982 *A Calculating People: The Spread of Numeracy in Early America*. University of Chicago Press, Chicago
- Domschke E, Goyer D S 1986 *The Handbook of National Population Censuses: Africa and Asia*. Greenwood Press, Westport, CT
- Economic Commission for Europe 1992 *Handbook of Official Statistics in ECE Member Countries*. United Nations, Geneva
- Kelman S 1985 Why should government gather statistics, anyway? *Journal of Official Statistics* 1(4): 361–79
- Lacey M J, Furner M O 1993 *The State and Social Investigation in Britain and the United States*. Cambridge University Press, Cambridge, UK
- Medvedev R 1989 *Let History Judge: The Origins and Consequences of Stalinism*. Columbia University Press, New York
- Nathan R 1999 The politic of printouts: The use of official numbers to allocate Federal grants-in-aid. In: Alonso W, Starr P (eds.) *Who Counts*. Russell Sage Foundation, New York
- Prewitt K 1999 Public statistics and democratic politics. In: Alonso W, Starr P (eds.) *Who Counts*. Russell Sage Foundation, New York
- Scott J C 1998 *Seeing Like A State: How Certain Schemes to Improve the Human Condition Have Failed*. Yale University Press, New Haven, CT
- Seltzer W 1994 *Politics and Statistics: Independence, Dependence, or Interaction?* United Nations Department for Economic and Social Information and Policy Analysis, Working Paper Series No. 6. New York
- Seltzer W 1998 Population statistics, the Holocaust, and the Nuremberg Trials. *Population and Development Review* 24(3): 511–22

K. Prewitt

Copyright © 2001 Elsevier Science Ltd.

11573

All rights reserved.

Policy Knowledge: Contract Research

1. 'Contractual Research' Versus 'Academic Research'

'Contractual research' as a modality of the science-based generation and production of (policy-related) knowledge is basically characterized by the commissioner/producer or consumer/contractor principle: 'the consumer says what he wants, the contractor does it (if he can), and the consumer pays' (to quote Lord Rothschild's much-referred to formula, Wittrock et al. 1991, p. 47). Hence, the 'request for proposal' (RFP) through which the commissioning agency addresses the would-be contractors (in public bidding, selective bidding, or directly), as a rule, defines and specifies the questions to be answered and the time frame. In his project proposal the would-be contractor explains his research plan within the parameters set by the 'customer' and makes his financial offer which is usually calculated on a 'personnel costs plus overheads' formula.

By contrast, academic research has traditionally hinged on the claim to institutional and intellectual autonomy of scientific work. Ideally committed to the 'search for truth,' academic (social) science research is traditionally understood as selecting its subject matter, research questions, and methodology solely by the 'free' decision of the individual researcher on the basis of 'intra-scientific' criteria. The essential frame of reference is the scientific community and its reputation system centering on peer review and peer recognition. Hence, the publication of results is an indispensable vehicle of allowing peer review as well as scientific knowledge accumulation and theory building. The classical locus of academic research is the 'research university,' based on the Humboldtian idea of the 'unity of research and teaching' as well as the (public/quasi-public non-profit) research institutes operating on a 'university without students' logic (Weaver 1989). While academic research is essentially geared to 'basic' research, it may well, through an intra-scientific search for 'relevance,' take some 'applied,' policy-related orientation. The cognitive autonomy of academic research is essentially secured by its 'independent' modalities of research funding, be it through the host institution's own resources ('untied' public money, private donations, endowments, etc.), be it through funding agencies (public or philanthropic foundations) usually operating on a competitive, peer-reviewed application formula.

At this point, it might be useful, in a side remark, to recall that recently the term and concept of 'think tanks' has been occupying the discussion on social science-provided policy-advice and its institutionalization (see Stone et al. 1998). In order to avoid the present use of the inflationary use of the term, 'think tanks' might be understood and defined as research

institutes which are institutionally and financially independent and see their task in generating and (most importantly) diffusing policy-relevant knowledge and therefore in exerting an influence on the political debate and agenda setting. 'Think tanks' that assume an explicitly ideological stance have been labeled 'advocacy think tanks' of which the Heritage Foundation in the USA and the Adam Smith Institute in the UK (both acting as neo-conservative/neo-liberal intellectual spearheads) are perhaps the most prominent examples.

This article is meant, within the limits of its space, to do two things. First, in its main body, it will take up a historical-institutionalist approach in 'mapping' the institutional landscape of contractual research (Sect. 2). Second, some conceptual (and ethical) key issues of contract research will be addressed (Sect. 3).

2. Across-countries Commonalities and Variance in Institutionalization of Contractual Research

2.1 Institutional and Cultural Factors Influencing the Institutional Profile ('Who is Who') of Contractual Research

The size and profile of contractual research in a country plausibly depends on institutional and cultural factors both of the political system ('on the demand side') and of the (social) science system ('on the supply side').

The political, governmental, and administrative demand for (external) science-generated policy knowledge may be contingent on the *institutional structure of government* (whether presidential or parliamentary), on the *political culture* (whether government seeks policy knowledge and advice primarily 'in house' in its own governmental apparatus, or at best 'quasi in-house' in governmental research institutes, or whether it is disposed to turn to external research capacities), and on the *administrative culture* (whether the administrative elites see themselves as professionally self-sufficient and secluded or whether they are perceptive for, and receptive to, external science-generated knowledge).

The supply of science-generated policy knowledge and advice may hinge on the *institutional setting of the social science system* (whether it traditionally revolves around 'research universities' (in the Humboldtian sense) or on non-university research institutes) as well as on *research attitudes and cultures* (whether university-based social scientists have the propensity to engage in policy-related knowledge production).

As some of these factors appear deeply, even in a 'path-dependency' manner, rooted in the countries' institutional and cultural fabric, they are likely to bring forth country-specific trajectories in the way contractual research has been institutionalized.

In Sects. 2.2 and 2.3 the attempt is made to elaborate (and explain) some distinct patterns and types of that institutionalization by dwelling on a number of countries as cases in point.

2.2 *Emergence Of Contractual Research up to World War II: America's Exceptionalism*

The development of the relation between university-based social science and government up to 1945 is a story of 'American exceptionalism' for a number of reasons. When the social sciences emerged at American universities, most of which were privately funded, the explicit attempt was made in the 1920s, spearheaded by the political scientist Charles Merriam at the University of Chicago, to combine the ambition to make them 'truly scientific' with an 'applied,' policy-related orientation. This applied stance of university research was fostered and supported by the early appearance of philanthropic foundations (Rockefeller, Carnegie) which made for another US American particularity and innovation. In this context the Brookings Institution was established in 1927 as a private, independently funded, policy-related institute—another American 'first'. While the federal government abstained from setting up governmental research institutes of its own, it started to draw on external, mainly university-based, social science expertise since the 1920s, opened up by President Hoover and stepped up by President Roosevelt during the New Deal (Wittrock et al. 1991, p. 39). When, during World War II, the federal government made an all-out effort to mobilize the country's entire scientific and technological potential, including the social sciences (systems analysis, psychology, etc.), 'contract research' became the decisive tool by which the federal government succeeded in catalyzing the huge technological, industrial, and intellectual resources of America's private sector which, to a large extent, include the universities (Lieske 2000, pp. 51ff.). In view of the salience of the private sector in the American institutional world it has been aptly remarked that 'contract is symbolic of the extent that it embodies and represents the values of private enterprise and the "free" market economy that have distinguished much of the nation's history' (Owens 1996, p. 249).

In contemporary European countries the development under consideration was distinctly different. On the Continent governments showed little interest in drawing on external science-generated knowledge. Instead, for instance in Germany, governmental research institutes were put in place for the (quasi in-house) provision of policy-relevant knowledge. While the university-based social sciences, in their struggle for scientific and academic recognition, accentuated their scientific profile and shunned an applied one. Sweden was an exception to this, where since the 1930s a policy-related interaction between government and

university-based social sciences has caught root. To some degree this also held true in the UK where reformist Fabians established the London School of Economics as an explicitly policy-related academic institution. It should be added, though, that in the economic field most European governments showed an early interest in eliciting and using external expertise from newly established public and private economic research institutes that were founded in Germany, France, Austria, and Sweden during the 1920s and 1930s (Gellner 1998, p. 90, Wittrock et al. 1991, p. 48).

2.3 *The Upsurge of Contractual Research After 1945: Commonalities and Variance*

In the following, the development in a number of countries will be highlighted as cases and types in point instead of venturing in a systematic overview.

2.3.1 *USA: private for-profit institutes/firms prevail in contractual research.*

Responding to the mounting policy knowledge demand which came from the post-war role of the US as the hegemonic Western power and from the ever more intensive Cold War confrontation, the US government continued to heavily draw on contractual research, at first particularly in foreign and military policies. The RAND Corporation, founded in 1948, was an institutional innovation in that it was the first private, for-profit research institute which was basically financed through governmental research contracts, in that case particularly from the US Department of Defense (Abelson 1998, pp. 112ff.). During the 1960s, contractual research experienced a tremendous upsurge. The Social Action programs under the Johnson administration and the legislative mandate to have them evaluated unleashed a contractual research money flow from the federal ministries and agencies on an unprecedented scale. At the same time, typical of the US government system, Congress built up its own evaluation capacities (GAO, CBO counting some 6000 staff members) and the federal departments too set up extended 'in-house' evaluation units and personnel (at the cost of some 1.2 billion US\$ in 1984). A huge amount of contractual money was put into external evaluation (amounting to some 460 million US\$ in 1984, Rist 1990, p. 80); thus, evaluation research became, 'in the words of Wall Street, a growth industry' (Rossi et al. 1999, p. 11). The institutional response to the contractual money market was threefold. First, linking up with their traditional applied and policy-related orientation and experience, university-based social scientists successfully sought access to the contractual research money market, sometimes operating at self-standing (non-profit) research institutes that, while remaining affiliated with their 'mother' university, were founded for the

very purpose of doing contractual research. Second, the (quasi-public or quasi-private) non-profit institutes (such as The Urban Institute, but also the time-honored Brookings Institution) got engaged in contractual research. Third and most conspicuously, a growing number of private for-profit research institutes and consulting firms, doing 'research for business,' were founded and entered the race for research contracts (such as Abt Associates, established in 1966). Sometimes ironically dubbed 'Belt Way Bandits' (after the high way around Washington, DC, where many of them are located) they have succeeded in taking the lion's share of the contractual research money market.

2.3.2 Sweden: university-based research prevails in contractual research. Among the European countries, Sweden also makes for an exceptionalist story of contractual research. Sweden's policy-making is traditionally characterized by the practice of creating, for each major piece of legislation or policy-making, *ad hoc* commissions made up of members of parliament, administrators, representatives of interest groups, and academics; in their mandate to seek agreed upon policy solutions they epitomize Sweden's traditional consensus-seeking political style (Vedung 1992, pp. 74ff.). For preparing their recommendations to government and parliament the *ad hoc* committees dispose of contractual research money to commission policy-related information and expertise. During the 1960s when Sweden experienced a (short-lived) PPBS boom and a 'scientific revolution' in policy-making (Wittrock et al. 1991), the country became one of the European frontrunners in policy evaluation, making evaluation look 'endemic throughout the Swedish system' (Levine 1981, p. 50). Thanks to the traditional policy-related interaction between government and university-based social scientists, the latter have prevailed in the conduct of contractual research.

2.3.3 Germany: public/non-profit and private/for-profit mix in contractual research. Germany presents the case of an (at first) retarded and (then) accelerated development of contractual research. Well into the 1960s, Germany's administrative (law-trained) elite adhered to the traditional belief in its self-sufficiency and, for the rest, relied on the chain of governmental research institutes (Hohn and Schimank 1990, pp. 303ff.) for informational support. At the same time, university-based research, while still absorbed by its post-war reconstruction and expansion, largely cherished its traditional dedication to basic research and continued to be wary of applied research. In the late 1960s, this traditional distance between, if not mutual neglect of, the government and the university-based social science experienced a

conspicuous rupture and reversal which was significantly the work of 'reform coalitions' formed by reformist politicians, civil servants, and university-based (social) scientists (Wagner and Wollmann 1991, pp. 74ff.). In conspicuously revamping its governmental machinery under the then dominant planning imperative and drawing, in an attempt at the 'scientification of policy making,' on social science advice on an unprecedented scale (Wollmann 1989), Germany became a European frontrunner, besides Sweden, in evaluation research (Derlien 1990, p. 148, Wagner and Wollmann 1986). The dramatic shift was evidenced by a rapid expansion of the contractual research money market with three groups of researchers competing for the research contracts. First, pursuing their new reorientation to applied and policy-related research, university-based researchers have sought and attained some access to contractual research, partly in institutional 'spin-offs' affiliated with the 'mother' university. Second, quasi-public non-profit research institutes have held a significant share in contractual research. Third, the new sector of private for-profit research institutes and consulting firms has been able to take lion's share of the contractual money market.

2.3.4 France: production of policy-related knowledge still largely in in-house or quasi-in-house manner. France can be seen as an example of the Continental European (State) tradition in which the production of policy-related knowledge has still remained largely within the government's domain in an in-house or quasi-in-house manner. France's administrative (*grands corps*) elites have retained their traditional belief in their professional self-sufficiency (Nioche 1992, p. 31). The Court of Accounts and the General Inspection of Finances constitute another barrier of traditional policy monitoring and auditing agencies. In addition, a chain of quasi-governmental research institutes, subordinated to the central ministries, has been established to provide ancillary policy information and advice. On the science system side, France has traditionally known the functional division between the teaching role of the universities and the research-focused CNRS whose close ties with the State have been seen as 'making them an integral part of the state apparatus' (Fieschi and Gaffney 1998, p. 54). Hence, France has been a latecomer to policy evaluation (Nioche 1992, p. 24); as late as in 1990 a typically centralized evaluation procedure was institutionalized, comprising, *inter alia*, a Scientific Council of Evaluation made up of academics (Duran et al. 1995, pp. 55ff.). While the contractual research money market is comparatively small, it has been dominated by CNRS institutes which quite often have teamed with university researchers. The share of private for-profit institutes and consulting firms has so far been quite modest.

2.3.5 Japan: private for-profit research institutes prevail in contractual research. While Japan's governmental bureaucracy has traditionally cherished a culture of self-sufficiency and elitist seclusion, it was also seized, during the 1960s, by a (passing) PPBS boom which resulted in the establishment of planning and analytical units in central ministries (Ide 1969) and in growing demand and contractual funding for external analyses. In 1974 the quasi-governmental National Institute for Research Advancement (NIRA) was established to initiate, coordinate, and monitor the contractual money research. On the 'supply side,' the university-based social science research has shown little propensity for applied and policy-related work (Watanuki 1991), not least because the universities have traditionally been under-equipped for research purposes. Thus, contractual research has been procured and conducted almost exclusively by private research institutes. Out of the some 225 private research institutes (as of 1992) some 40 percent are for-profit (Ueno 1998, p. 195). While the largest ones (such as Mitsubishi Research Institute and Nomura Research Institute, founded in 1970 and 1975, respectively) have been established by their 'parent' corporations, most of the smaller ones have typically been established by former (well-connected) high-ranking civil servants after their early retirement as well as by university professors who combine teaching at the university with doing paid contractual research outside.

2.3.6 European Union (EU): dynamic contractual research money market dominated by the private for-profit research sector. While EU policy-making lies largely with the European Council and the Commission, the implementation of EU policies (that is, EU regulation and EU funding) is carried out by the member states and their (for the most part regional and local) bureaucracies. Thus the EU central actors, particularly the Commission, have a 'natural' need for external policy information. With regard to EU structural funding, for instance, the procurement of such information has been secured by a technically sophisticated evaluation system consisting of *ex ante*, interim, and *ex post* evaluation and based almost entirely on external evaluation. Consequently the EU has, since the late 1980s, opened up and has massively funded a contractual research money market. The evaluation of EU policies and programs is conducted on two levels. On the EU central level the evaluation contracts (amounting to 13.5 million Euro in 1999) relate to the entire EU programs and have been won and carried out exclusively by for-profit institutes and consulting firms among which big international firms (like Anderson, Ernst and Young and the like) loom large. Another EU contractual money flow is directed at evaluating the implemen-

tation of the EU programs in the individual member countries. As these projects are smaller and require some familiarity with the national, regional, if not local, contexts, 'national' research institutes have prevailed in gaining the research contracts. The profile of the contractors differs from country to country. While, for instance, in Spain university-based research groups have the Spain-related evaluation contracts, in most other countries the majority of the contractors are private for-profit. The contractual research money flow of the EU is bound to have significant repercussions not only on the contract research scene, but also on the social science community at large. This holds true particularly in South European member countries (Greece, Spain, Portugal, and also Italy) which receive the bulk of EU funding and which, consequently, are the target area of correspondingly massive evaluation research. There is evidence that, in fact, EU contractual research funding has given strong incentives and impulses to the national research scene in these countries.

3. Contractual Research Constituting a Threat to the (Intellectual Etc.) Integrity of the Individual Researcher or Even of the Social Science Research Community at Large?

The expansion of contractual research has been accompanied, both within the contractual research camp as well as on the academic scene, by a debate about the threats which the crucial consumer/contractor constellation may pose to the intellectual (and moral) integrity of the researchers involved and beyond. It is particularly the basic commissioning and funding logic of contractual research, described in the 'Rothschild principle' as 'the consumer says what he wants, the contractor does it (if he can), and the consumer pays,' that is being suspected of jeopardizing and undermining its intellectual integrity and of translating into reality the adage 'who pays the piper calls the tune'. First of all private for-profit institutes and firms that seek and conduct research contracts as 'research for business' are seen exposed to an erosive, if not corruptive mechanism (Ham 1999, p. 276). As a spill-over from the realm of contractual research onto social science research at large, including academic research, even a 'colonization process' has been foreboded in the wake of the external setting of the research agenda, 'whereby the bureaucracies' perspective and conceptual framework (is) established as the valid epistemological framework also for scientific research' (Elzinga 1985, p. 211).

In the face of such challenges to the intellectual integrity and honesty of contractual research, initiatives have been taken, particularly among evaluators as probably the most numerous and most

professionalized group among those engaged in contractual research, to formulate and lay down standards that could guide them in their contractual work and in negotiations with their 'clients' (Rossi et al. 1999, pp. 425ff.). Exemplary are the *Guiding Principles of Evaluation* that were adopted by the American Evaluation Association in 1995. Among its five principles the maxims of integrity and honesty of research ('evaluators must ensure the honesty and integrity of their entire evaluation process') are writ large (full text in: Rossi et al. 1999, pp. 427ff.).

Bibliography

- Abelson D E 1998 Think tanks in the United States. In: Stone D, Denham A, Garnett M (eds.) *Think Tanks across Nations*. Manchester University Press, Manchester/New York, pp. 107–26
- Derlien H-U 1990 Genesis and structure of evaluation efforts in comparative perspective. In: Rist R (ed.) *Program Evaluation and the Management of Government*. Transaction, New Brunswick/London, pp. 147–77
- Duran P, Monnier E, Smith A 1995 Evaluation à la Française. *Evaluation* 1(1): 45–64
- Elzinga A 1985 Research, bureaucracy and the drift of epistemic criteria. In: Wittrock B, Elzinga A (eds.) *The University Research System*. Almqvist and Wiksell, Stockholm, Sweden, pp. 191–220
- Fieschi C, Gaffney J 1998 French think tanks in comparative perspective. In: Stone D, Denham A, Garnett M (eds.) *Think Tanks across Nations*. Manchester University Press, Manchester/New York, pp. 42–58
- Gellner W 1995 *Ideenagenturen für Politik und Öffentlichkeit: Think Tanks in den USA und in Deutschland*. Westdeutscher Verlag, Opladen, Germany
- Gellner W 1998 Think tanks in Germany. In: Stone D, Denham A, Garnett M (eds.) *Think Tanks across Nations*. Manchester University Press, Manchester/New York, pp. 82–106
- Ham V 1999 Tracking the truth or selling one's soul? Reflections on the ethics of a piece of commissioned research. *British Journal of Educational Studies* 47(3): 275–82
- Hohn H-W, Schimank U 1990 *Konflikte und Gleichgewicht im Forschungssystem*. Campus, Frankfurt/New York
- Ide Y 1969 Administrative reform and innovation: The Japanese case. *International Social Science Journal* 1: 56–76
- Leeuw F L, Toulemonde J, Brouwers A 1999 Evaluation activities in Europe: A quick scan of the market in 1998. *Evaluation* 5(4): 487–96
- Levine R A 1981 Program evaluation and policy analysis in western nations: An overview. In: Levine R A, Solomon M A, Hellstern G-M, Wollmann H (eds.) *Evaluation Research and Practice*. Sage, Beverly Hills, CA, pp. 27–60
- Lieske J 2000 *Forschung als Geschäft*. Campus, Frankfurt/New York
- Nioche J-P 1992 Institutionalising evaluation in France, skating on this ice? In: *OECD 1977, Review of Social Science*. OECD, Japan, Paris
- Owens L 1996 MIT and the federal 'Angel': Academic R&D and federal-private co-operation before World War II. In: Numbers L, Rosenberg C E (eds.) *The Scientific Enterprise in America*. Chicago, pp. 247–72
- Rist R (ed.) 1990 *Program Evaluation and the Management of Government*. Transaction, New Brunswick/London
- Rossi P H, Freeman H E, Lipsey M W 1999 *Evaluation. A Systematic Approach*, 6th edn. Sage, Thousand Oaks
- Stone D, Denham A, Garnett Mark (eds.) 1998 *Think Tanks across Nations. A Comparative Approach*. Manchester University Press, Manchester, UK
- Ueno M 1998 Think tanks in Japan: Towards a more democratic society. In: Stone D, Denham A, Garnett M (eds.) *Think Tanks across Nations*. Manchester University Press, Manchester/ New York, pp. 188–201
- Vedung E 1992 Five observations on evaluation in Sweden. In: Mayne J, Bemelmaus-Videc M L, Hudson J, Conner R (eds.) *Advancing Public Policy Evaluation*. Elsevier, Amsterdam, pp. 23–36
- Wagner P, Wollmann H 1986 Fluctuations in the development of evaluation research: Do 'regime shifts' matter? *International Social Science Journal* 38(2): 205–18
- Wagner P, Wollmann H 1991 Beyond serving state and bureaucracy: Problem-oriented social science in (West Germany). *Knowledge and Policy* 4(1,2): 56–88
- Watanuki J 1991 The impact of social sciences on the process of development in Japan. In: Wagner P, Weiss C H, Wittrock B, Wollmann H (eds.) 1991 *Social Sciences and Modern States*. Cambridge University Press, Cambridge, UK, pp. 221–9
- Weaver K 1989 The changing world of think tanks. *Political Science and Politics* 22(3): 563–78
- Wittrock B, Wagner P, Wollmann H 1991 Social science and the modern state. In: Wagner P, Weiss C H, Wittrock B, Wollmann H (eds.) *Social Sciences and Modern State*. Cambridge University Press, Cambridge, UK, pp. 28–85
- Wollmann H 1983 *Development and State of Policy Research in Japan*. National Institute for the Advancement of Research, Tokyo
- Wollmann H 1989 Policy analysis in West Germany's Federal Government Governance. *An International Journal of Policy and Administration* 2(3): 233–66

H. Wollmann

Policy Knowledge: Epistemic Communities

Political science and policy studies have increasingly developed an interest in the role of ideas, values, and technical understanding in shaping political outcomes, particularly under conditions of perceived complexity. 'Epistemic communities' is a concept developed by 'soft' constructivist scholars of international relations concerned with agency to understand the actors associated with the formulation of ideas, and the circumstances, resources and mechanisms by which new ideas or policy doctrines get developed and are introduced to the political process.

1. Constructivism

Constructivism analyzes the social process by which actors construct meaning of the world through the application of broad ideas and reasoning patterns. Thus, constructivists specify how actors come to

recognize interests and make choices (Ruggie 1998, Haas 1992/1997). While there has been a long-standing appreciation of the importance of ideas—the fabric with which Weberian webs of significance are spun—on politics and on choice, the mechanism by which such constructions arise and persist has long remained elusive. Epistemic communities are important actors responsible for developing and circulating causal ideas and some associated normative beliefs, and thus helping to create state interests and preferences, as well as helping to identify legitimate participants in the policy process and influencing the form of negotiated outcomes by shaping how conflicts of interest will be resolved. Attention to epistemic communities provides a way to understand agency in politics and policy formation.

Assigning context and determining appropriate behavior—and thus policy—are fundamentally political processes, both in terms of actors' efforts to directly influence the process and in terms of their distributional consequences. Most IR theorists now agree that contemporary politics and international relations are characterized by conditions of complexity, as a consequence of growing interdependence and globalization. Complexity complicates decision-making. Substantive and procedural rationality are impossible under high degrees of systemic complexity (Simon 1983).

Constructivists rely on discussions of complexity to conclude that actors are unable to easily deduce preferences and understanding of options from contemporary circumstances. States and other actors are uncertain about their ultimate goals, and how to achieve them. Even whether actors choose to apply a logic of consequences or of appropriateness is itself conditioned by their prior understanding of the nature of the political circumstances in which they find themselves (March and Olson 1998). The very nature of political strategies actors deem appropriate is a consequence of the advice and framing that actors receive about the uncertain policy environment in which they find themselves, or what John Searle calls establishing that 'x counts as y under conditions c' (Searle 1995).

For international relations, more specifically, actors are uncertain about how to achieve particular interests in given circumstances (their preferences); with whom they have to interact to achieve their goals, and by what means to do so. For instance, when first confronted with new issues political actors are typically uncertain about whether to pursue their ends unilaterally or multilaterally, through which institutions, which nests to apply, which linkages are relevant, and what policies should be collectively embraced. For the policy studies literature (Ascher 1986, Thomas 1997, Social Learning Group 2001) the principal analytic focus that results from an appreciation of complexity and the role of ideas and epistemic communities is on the stages of policy making:

framing, agenda setting, social learning, participation, policy choice, and, to a lesser extent, compliance or enforcement.

National and international politics and policy choice then becomes a matter of applying embedded and institutionalized beliefs about the nature of problems and the appropriate means of collective response as argued by many current theorists of IR (Ruggie 1998, Adler 2001, Haas 1990, 1992/1997), rather than the process of resolving rationally formulated state preferences. In this conception politics becomes a process of learning about the world.

Thus, ideas may play important roles in framing the policy debate, and in predisposing outcomes highlighted by each perspective. Epistemic communities are the cognitive baggage handlers of constructivist analyses of politics and ideas. They are invoked by Constructivist scholars to focus analytic attention on the process by which states formulate interests and reconcile differences of interest. Epistemic communities are a principal channel through which consensual knowledge about causal connections is applied to policy formation and policy coordination. States may learn through processes of international cooperation mediated by epistemic communities. As a consequence collective patterns of behavior reflect the dominant ideas initially brought to bear to understand the area of activity which is being governed.

Ruggie borrowed *episteme* from Foucault, to describe the overarching perspective through which political relationships are visualized and understood during historical eras (Ruggie 1975). The winter 1992 issue of *International Organization* on Knowledge, Power and International Policy Coordination refined the concept by restoring agency to the study of politics and knowledge, by focusing on the actors responsible for articulating and aggregating knowledge based understanding in areas of security, environment and international political economy (Haas 1992/1997). Epistemic communities are networks—often transnational—of knowledge based experts with an authoritative claim to policy relevant knowledge within their domain of expertise. Their members share knowledge about the causation of social or physical phenomena in an area for which they have a reputation for competence, and a common set of normative beliefs about what actions will benefit human welfare in such a domain. In particular, they are a group of professionals, often from a number of different disciplines, who share all of the following characteristics:

(a) Shared consummatory values or principled beliefs. Such beliefs provide a value-based rationale for social action by the members of the community.

(b) Shared causal beliefs or professional judgment. Such beliefs provide analytic reasons and explanations of behavior, offering causal explanations for the multiple linkages between possible policy actions and desired outcomes. While they need not agree on every element—in fact they are likely to disagree vehemently

about some elements—they do agree about the core assumptions and causal forces from which their models of the world derive and agree on the means by which such differences can potentially be reconciled.

(c) Common notions of validity: intersubjective, internally defined criteria for validating knowledge.

(d) A common policy enterprise: a set of practices associated with a central set of problems which have to be tackled, presumably out of a conviction that human welfare will be enhanced as a consequence.

This combination of factors—especially the socialized truth tests and common causal beliefs—distinguish epistemic communities from other types of policy networks and groups active in politics and policy-making. Unlike other organized interest groups active in politics and policy making, epistemic communities are bound by the truth tests to which they were socialized, and thus are more likely to provide information that is politically untainted, and thus more likely to ‘work,’ in the political sense that it will be embraced and followed by political authorities concerned about the need for appearing impartial, and also technically.

2. *Cognate Literatures*

The policy sciences, since Lasswell, have stressed the difficulty of applying strong rationality assumptions when conducting policy analysis, as well as advocating a neo-pluralistic approach to understanding the politics of policy making (Ascher 1986, Braybrooke and Lindblom 1963, Healy and Ascher 1995). Discursive approaches to understanding the policy sciences reinforce constructivist studies of politics and policy (Linder and Peters 1995, Stone 1989).

Similar concepts to epistemic communities were developed in other social sciences in the 1980s and 1990s to better understand the role of expertise in constructing meaning for policy makers and for developing policy usable knowledge. Knorr-Cetina, in the sociology of knowledge, coined the term ‘epistemic communities’ (Knorr-Cetina 1981, 1999), although this neologism was not read by the international relations research program until after the 1992 publication of the epistemic communities volume.

In comparative politics and policy studies complementary notions have been put forward to study networks of experts involved in policy making. Richardson coined the term policy communities (Richardson 1995). Reinicke has extended policy communities to include global policy networks (Reinicke and Deng 2000). Sabatier developed the notion of policy advocacy coalitions, which were an extension of earlier US work on iron triangles (Sabatier and Jenkins-Smith 1993). The concept of policy communities stresses the causal beliefs shared by policy analysts involved in policy development, but do not

fully develop a causal model of their development and spread. Analysts of advocacy coalitions note the presence at different points in time of people from multiple interest groups and organizations, but do not address their beliefs systematically, nor do they look at the causal role of a kernel of individuals sharing beliefs and involved in the same policy enterprises over time. Analysts of epistemic communities, with their focus on agency and the shared beliefs of a community that shares a common policy activity over time, provide the leverage for generating social science knowledge about the development and diffusion of knowledge for policy and politics.

3. *A Constructivist Model of Institutionalized Ideas*

The development and application of an epistemic community’s consensual knowledge can be best understood within a stepwise framework. Such a framework provides independent mechanisms by which new ideas are generated, selected by authoritative actors, and persist over time. It is a path-dependent model of punctuated equilibrium because many choices or decisions are largely irreversible and the array of possible choices at time (t) is causally related to decisions taken at time ($t-1$) in ways that may not be perceptible to actors at time $t+1$ (Adler and Haas 1992/1997). Ideas may exercise a causal impact later in time regardless of the reasons for their initial selection, as they continue to shape actors’ understanding and guide decisions even in the absence of original conditions that contributed to their widespread acceptance.

While actors are always constructing the world, to some extent, the challenge to analysts is to identify what constructions are likely to prevail, and under what circumstances. Since current constructions are made in the shadow of past constructions, this framework provides the means to specify which ideas will be influential and the factors influencing the likelihood that they will exercise enduring influence. Unique collective outcomes involving learning are possible with the existence and involvement of an epistemic community to focus beliefs. Otherwise, political behavior remains a matter of applying prior frames to understanding political choices and options, with conflicts of interest resolved through the application of existing institutions and power resources.

4. *Constructivist Assumptions*

A number of key assumptions underlie this constructivist model. States are key juridical actors, and act willfully. States are functionally differentiated, and vary widely according to their state/society relations

and the technical capacity of the state to formulate and enforce public policies in technical domains (Risse-Kappen 1995, Katzenstein 1977, Gourevitch 1996, Evans et al. 1993). States operate in a complex policy environment, so that information about the environment is highly valued. Yet knowledge is asymmetrically distributed within and between countries. Finally, while perceptions are socially constructed, political actors and social scientists are not subject to the same set of influences. Thus social scientists may identify social forces influencing actors' choices without being equally prone to their influence.

A number of propositions about state behavior and its explanation immediately follow from these assumptions. States are uncertain about their policy environment and how their interests are to be recognized in particular instances, and thus they exercise bounded rationality. The recognition of threats and the identification of responses are subject to socialized patterns of reasoning, rather than through formal rationality in which actors have complete information and choose the most efficient alternatives. Decision-making operates according to a satisfying model in which actors express procedural rationality, at best, subject to the time constraints and structural blinders imposed by bounded rationality (Simon 1983, 1985, March 1994). Consequently, decisions and institutions are sticky, since decision-makers do not return to issues until there is a crisis.

States make choices subject to multiple sources of influence, whose organization varies by issue area (Keohane and Nye 2001). Thus, governance varies by issue area. Consequently, policy networks, organized around specific issues become the appropriate level and unit of analysis, because the array of actors, interests, institutions, and capabilities varies by issue. Even within a particular state, the design of policy-making institutions and procedures vary by issue area. Consequently, it is not surprising that international regimes in related areas can differ widely, because the articulation of interests for each regime takes a different form.

5. Generation of New Ideas

Epistemic communities are the agents by which policy relevant information is developed and transmitted. They are likely to be found in substantive issues where scientific disciplines have been applied to policy oriented work. Thus, epistemic communities arise in disciplines associated with natural science, engineering, ecology, and even economics. They will also be most likely to emerge in countries with well-established institutional capacities for public administration, and science and technology. They are also most likely to be present in rich (measured as a high ratio as share of GDP, and with a significant number of scientists and

engineers in the country) democracies. Democracies are most likely to encourage open-ended, spirited discussions because the knowledge enterprise is independent of the state, and yet the state has the resources to access and support science.

For governments to defer to new knowledge, the presumptive knowledge claim must be regarded as authoritative. Authoritative knowledge comes from groups that command legitimacy and authority in society. Epistemic communities' reputation for authority rests on their expertise as well as their reputation for impartiality. If not seen as being impartial their advice will be widely contested and lack any independent political influence.

Such pockets of independent research are most likely to emerge when the conduct of research is seen to be far from the direct interest and influence of the sponsoring sources of the research (Dasgupta and David 1994, David 1995). Its impartiality and accuracy depends heavily on the ability to publish in peer-reviewed journals, as well as the internal commitments to truth tests by epistemic community members. A distance from research sponsors improves the actual and perceived impartiality of epistemic community members as it is less likely to be shaped by needs of the state or sponsoring agency, and will be less politically tainted. Activities most likely to be regarded as impartial are those with relatively low costs that are not directly tied to the mission of the sponsoring agency. Epistemic communities are likely to offer policy advice that will be effective, unlike the more common process of policy advice which corresponds to specific interests, because the policy proposals are developed far from the political influence of sponsoring agencies and will be more acceptable to the public when the ideas do not appear politically compromised.

6. Selection

New ideas will be solicited and selected only after crises, for crises will alert politicians to the need for action and will seek to gather information about their interests and options. Crises will trigger agenda setting and the search for new information.

Ideas will be selected by states and international institutions. However, not all states or international institutions are equally likely to consult members of an epistemic community or defer to their advice. Institutions with strong science and technology capabilities, and those representing pluralistic societies will be most likely to be in the first wave of those embracing new ideas. It is there that the scientific culture of epistemic communities will have a close affinity, and new groups will be able to quickly articulate new policy initiatives. Selection will probably take longer in non-democratic societies with technically strong

states that have a strong domestic scientific capacity because it takes longer for ideas to percolate up from society to the government outside of routinized channels.

Different styles of policy selection occur in different political systems, due to organizational and cultural differences between types of democracies (Weaver and Rockman 1993, Katzenstein 1977). In presidential systems, with multiple checks and balances, the flow of new ideas to the state will emerge from think tanks, universities, and governmental labs; and will be solicited directly by Congressional Committee staff, the judiciary and executive branch bureaucrats. In Parliamentary systems the sources of new ideas are more limited, and much more likely to emerge from party think tanks and be deployed through political parties to ministers and suitably socialized civil servants. At times ideas may go from universities and independent think tanks to mid level civil servants, but there are fewer universities and independent think tanks in parliamentary systems because of the predominant influence of political parties.

International institutions will also vary in their propensity to take on board new ideas. In a study of international institutions that selected new programmatic ideas from epistemic communities leading to a redirection of the core mission of the institution, Haas and Haas (1995) found a number of common institutional properties among them. Their governing bodies were not split by major political disagreements, they were led by skilled institutional heads, they had skilled professional staffs recruited on merit, and the institution maintained a porous connection with external sources of information. Few institutions have these properties. Those that do were created by their initial funders and sponsoring states to have such resources. The postwar settlement institutions had to reflect the geopolitical goals of the superpowers, and be consistent with overarching values and widely shared norms about economic management because decisions were made at the top levels of government in the wake of such system wide crises, and these policy makers were alert to consistencies between perceived state imperatives and institutional design. Institutions that were established in the interstices of the Cold War, when state leaders were paying less attention to perceptions of systemic imperatives, have greater prospects for flexibility when they reflect domestic sentiment in the superpowers, and they were created during an absence of widespread geopolitical schisms.

There are several alternatives regarding the likelihood of which ideas will be selected at a moment of crisis. If there is only one new candidate, there is little problem, and the new aspirant will be selected. If there are many, choices may be conditioned by calculations of relative political gain from each idea. Studies of trade policy and the use of ideas in formulating international trade regimes have offered a number of

propositions about conditioning factors influencing the selection of competing policy ideas (Odell 1990, Goldstein 1993, Garrett and Weingast 1993). Ideas that provide a potential for building coalitions will be favored over those that do not. Simple ideas that have perceived practical utility will be chosen over more complex approaches. In either case, new ideas will only endure if they are loosely commensurate with deeper-seated beliefs and do not endanger strategic political alliances.

7. Diffusion

Over time ideas will be diffused. Whereas they were selected initially by deferring to epistemic communities and recruiting their members by states and international institutions that were seeking new information, the diffusion of ideas occurs by many channels that include direct persuasion, demonstration effects, and inducements and incentives.

Diffusion may occur by institutional learning, as groups apply evolving consensual knowledge to manage institutional practices. Learning occurs directly, through interpersonal persuasion, communication, exchange and reflection, that leads to the recognition or appreciation of new causal models and shared values. Leaders are socialized to accept new views and to empower their expositors.

One other diffusion mechanism is administrative recruitment, as epistemic community members or their confidantes replace officials informed with alternative perspectives. Diffusion may also occur by other mechanisms, as actors alter their behavior subject to the influence wielded by institutions that embody the ideas (those that selected an epistemic community and its ideas earlier).

International regimes informed by institutionalized ideas promulgated by strong states and by international institutions will have the effect of disseminating the ideas to other parties in which epistemic communities were not present, and through their socialization and exposure to new ideas presented during regime negotiations.

Diffusion may also occur directly as a consequence of incentives and inducements to adopt those ideas coming from states, international institutions, and international regimes. These incentives and inducements typically take the form of technology assistance and financial transfers. These will only promote diffusion to states with weak capacity, for whom the provision of such resources would be a powerful incentive to change policy. Demonstration effects by multilateral financial institutions will induce borrowing states to adopt the policy being demonstrated.

States may draw lessons from one another's experiences, and thus borrow policies and ideas from countries that they are trying to emulate (Rose 1993). This process of diffusion is not one of learning,

though, it is one of mimesis or borrowing because the lesson drawers are not adopting the policies based on the persuasive power of reason by an advocate of the policy. Rather they are imitating the actions of someone who commands respect and appears successful.

These diffusion mechanisms will operate along different time frames. While some effects may be felt immediately through new recruitment decisions, and persuasion, broader shifts in public opinion and societal effects may occur on the order of decades. As these processes overlap, the general diffusion of an epistemic community's ideas will accelerate.

8. Persistence

Once in place, ideas are likely to persist once they acquire a taken-for-granted element and as states and their diplomats become socialized to institutionalized regimes and practices generated by the application of new ideas. Eventually ideas get converted to domestic laws, and thus become enforced out of habit and become the basis for policy enforcement by the state. It also becomes politically costly to reverse such practices as new interest groups and policy communities mobilize around them after recognizing that material gains are possible from the application of the new ideas. Ultimately new markets may form through the application of these ideas, and thus generate economic incentives for their continued adoption once economic actors associate their profits with the application of such ideas.

9. Empirical Examples

9.1 International Political Economy (IPE)

Analysts of IPE have identified ideational dynamics in the popularization and dissemination of major economic doctrines. For instance, the repeal of the corn laws occurred after the conversion of Prime Minister Peel, in the face of likely political defeat. The adoption of liberal economic policy occurred following the creation of an influential epistemic community led by Richard Cobden able to persuade top levels of government that the national interest could be advanced by low tariffs, even in the face of economic losses to a powerful political constituency expressing particularistic protectionist interests. England's hegemonic power was then deployed to convert Europe to free trade through the 1860 Cobden Chevalier Treaty, and the US through bilateral treaties, and most European governments kept tariffs relatively low in accordance with the institutionalized ideas until

the 1890s, despite a lengthy and deep economic depression.

Keynesianism spread internationally through a similar process of epistemic persuasion in a core set of countries, followed by the creation of international institutions to legitimate and create international structures based on Keynesian dictates. Keynesianism first caught on in England, then traveled to the US through academic networks of economists. After its adoption at the highest levels in the two countries, it provided the intellectual foundations for the Bretton Woods System, through which it was institutionally embedded for decades.

Structuralism and Development Economics was developed in ECLA in the 1940s, and then embedded in ECLA and UNCTAD, from where it continued to serve as development orthodoxy to a number of developing countries throughout the 1970s, as well as providing the economic justification behind the New International Economic Order reforms of the 1970s. Odell tells how a group of 'floaters' in the US treasury department influenced ending the Bretton Woods fixed exchange rates by designing its replacement (Odell 1982). The Washington Consensus on macroeconomic management has guided international public debt rescheduling during the 1990s.

9.2 International Environmental Protection

Although multilateral economic activities have been governed, since World War II, by the application of consensual liberal ideas about free trade, efforts to control the impact of economic activity on transboundary and environmental pollution has been based on non market and regulatory principles. Since 1972 an ecological epistemic community, articulating a new ecological management doctrine, has institutionalized its ideas in state policies and practices, in the programmatic activities of international institutions, and in international regimes. As a consequence of the international institutionalization of its ideas, states were caused to undertake more comprehensive styles of environmental management for transboundary and global environmental threats, leading to selective improvements in environmental quality (Haas 2001).

A new environmental management doctrine based on ecological principles emerged in the 1960s. The ideas were developed outside the scope of most state interests and are generally regarded as being relatively uncompromised by political and institutional influence, as ecology was a relatively cheap research enterprise and its subject of study was far removed from most state interests. Members of the ecological epistemic community subscribed to holistic ecological beliefs about the need for policy coordination subject to ecosystemic laws (Haas 1990). Their ideas about

ecological management were based on a systems perspective of environmental and social systems. Ecological management proposals were based on setting comprehensive environmental standards based on conservative estimates of the ability of ecosystems to sustain stress, subject to their technical understanding of the behavior of particular ecosystems. They promoted international environmental regimes that are grounded on policies which offered coherent plans for the management of entire ecosystems, sensitive to interactions between environmental media (such as air and water), sources of pollution, and contending uses of the common property resource, rather than being limited to more traditional policies for managing discrete activities or physical resource spaces within fairly short term time horizons. They proposed treaties in which bans and emission limits were set for multiple contaminants, with environmental standards for each contaminant set according to scientific understanding about its environmental impact and its interactive effects with other contaminants.

There were few ideational competitors. Resource management bodies had traditionally been staffed by neoclassical economists and resource managers, who had been discredited by broadly publicized environmental disasters and the energy crisis of the 1970s as well as the limits to growth debate, which they had been unable to predict, and attendant popular fears of widespread resource depletion.

Following the politicization of international environmental issues in the late 1960s galvanized by widely publicized environmental disasters occurring in global commons, the United Nations convened the United Nations Conference on the Human Environment (UNCHE) in 1972. As a consequence of these new domestic demands, and in order to prepare national papers for UNCHE, governments created new environmental agencies. Most state officials were unfamiliar with environmental threats and were unable to rank order environmental threats, amounts of national emissions, what would constitute safe or dangerous concentrations of possible contaminants, and what were appropriate policies to reduce emissions. Many of these national agencies recruited epistemic community members to serve as officials or as consultants.

International organizations were created as well. The United Nations Environment Program (UNEP) was established in 1973 and was staffed principally by young epistemic community members eager to put their professional knowledge to work. The UN Economic Commission for Europe (UNECE) played a strong role in managing European air pollution. The UNECE's environmental unit was led by a former UNEP official who carried the ecological management ideas from UNEP to UNECE. After 1987 the World Bank also became active in international environmental matters. Part of the environmental reforms

introduced at the World Bank was the recruitment of ecological epistemic community members and their assignment to key posts to evaluate the environmental consequences of development projects.

Diffusion occurred principally through efforts of these major environmental IOs and through environmental regimes. The IOs encouraged other IOs to internalize environmental concerns into their missions through joint projects, and encouraged governments to pursue more comprehensive environmental policies through public education campaigns, publicizing environmental monitoring findings, resource transfers, and demonstration effects.

International regimes have been even more important for the diffusion of ecological management techniques. This spread is due to the growing influence of ecological epistemic communities and the growing confidence and influence of strong international institutions informed with their ecological beliefs.

Ecological epistemic communities, often working with UNEP, helped draft international environmental regimes governing marine pollution, acid rain, stratospheric ozone protection, wetlands protection, protecting migratory species, polar bears protection, and the preservation of Antarctica. An increasing proportion of them are now based on the comprehensive ecological approach promoted by the ecological epistemic community. In 1973, three of 11 international environmental regimes were based on ecological management styles. In 1985, seven of 22 were comprehensive. In 1995, 10 of 25 were comprehensive (Haas 2001). The application of ecological management ideas to environmental regimes spanning a number of geographic areas and functional activities means that most states have accepted ecological obligations for governing a wide variety of human activities.

Ecological practices based on ecological management ideas have become locked in through a variety of mechanisms. Following ratification of international regimes, governments enforce these obligations domestically. Ecological practices get institutionalized domestically through legal precedents, bureaucratic standard operating procedures, and policy enforcement. In many countries they have acquired domestic constituencies—composed of lawyers and civil engineers who subcontract services, firms selling pollution control technologies, and environmental NGOs—that contribute political pressure for continued state enforcement of policies grounded on these ideas.

Exogenous forces have also reinforced these state commitments. A Gallup poll conducted in 1992 demonstrated new worldwide concern about the need to respond to global and transboundary environmental threats (Dunlap et al. 1992). In democratic societies public demands for environmental protection reinforce the influence of the epistemic community's ideas. International markets for pollution control technology also came into existence in the 1990s.

10. Further Research

Epistemic communities is a core concept in a progressive constructivist research program. Its application in international relations has restored agency to the study of ideas, improved the understanding of the sources of state interests under conditions of uncertainty and complexity, and expanded an understanding of regime dynamics. In policy studies it fleshes out with accounts of agency the insight that policy debates are discursive rather than rational, and thus 'best' policy analytic suggestions do not always prevail. It provides the potential for understanding the mechanisms by which knowledge change contributes to policy change.

Further work needs to be done to integrate this work into institutional studies, to better understand the character of crises that are capable of generating political upheavals and discrediting institutionalized ideas as well as how ideas become institutionalized at the national and international levels. Research opportunities beckon on the creation, mobilization and maintenance of epistemic networks. Further prospects lie in studies of variation in generation and selection of epistemic communities, and how they communicate with the public. Applications of epistemic communities' research would contribute to the development and organization of usable technical knowledge for improved policy debates on contemporary issues.

See also: Constructivism/Constructionism: Methodology; Decision Making, Psychology of; Environmental Planning; Environmentalism: Preservation and Conservation; International Organization; Organizational Behavior, Psychology of; Organizational Climate; Organizational Decision Making

Bibliography

- Adler E 2001 Constructivism in international relations. In: Carlsnaess W, Risse T, Simmons B A (eds.) *Handbook of International Relations*. Sage, Beverly Hills, CA
- Adler E, Haas P M 1992/1997 Conclusion. In: Haas P M (ed.) *Knowledge, Power, and International Policy Coordination*. University of South Carolina Press, Columbia, SC
- Ascher W 1986 The evolution of the policy sciences. *Journal of Policy Analysis and Management* 5(2): 365–89
- Braybrooke D, Lindblom C E 1963 *A Strategy of Decision; Policy Evaluation as a Social Process*. Free Press, New York
- Dasgupta P, David P A 1994 Toward a new economics of science. *Research Policy* 23: 487–521
- David P A 1995 Reputation and agency in the historical emergence of the institutions of 'open science.' Mimeograph, Stanford University, Stanford, CA
- Dunlap R E, Gallup G H Jr, Gallup A M 1992 *Health of the Planet Survey*. Gallup International Institute, Princeton, NJ
- Evans P B, Jacobson H K, Putnam R D (eds.) 1993 *Doubledged Diplomacy*. University of California Press, Berkeley, CA
- Garrett G, Weingast B 1993 Ideas, interests and institutions. In: Goldstein J, Keohane R O (eds.) *Ideas and Foreign Policy*. Cornell University Press, Ithaca, NY
- Goldstein J 1993 *Ideas, Interests, and American Trade Policy*. Cornell University Press, Ithaca, NY
- Gourevitch P A 1996 Domestic sources of international cooperation. *International Organization* 50(2): 349–73
- Haas P M 1990 *Saving the Mediterranean*. Columbia University Press, New York
- Haas P M (ed.) 1992/1997 *Knowledge, Power, and International Policy Coordination*. University of South Carolina Press, Columbia, SC (originally published in *International Organization* 46(1) (Winter 1992))
- Haas P M 2001 Institutionalized knowledge and international environmental politics. In: Ikenberry G J, Parsi V (eds.) *Manual di Relazione Internazionale*. Gins, Laterza & Figlie, Rome
- Haas P M, Haas E B 1995 Learning to learn. *Global Governance* 1(3): 255–85
- Healy R G, Ascher W 1995 Knowledge in the policy process. *Policy Sciences* 28: 1–19
- Katzenstein P J (ed.) 1977 *Between Power and Plenty*. University of Wisconsin Press, Madison, WI
- Keohane R O, Nye J P 2001 *Complex Interdependence*, 3rd edn. Longman, New York
- Knorr-Cetina K 1981 *The Manufacture of Knowledge*. Pergamon, New York
- Knorr-Cetina K 1999 *Epistemic Cultures: How the Sciences Make Knowledge*. Harvard University Press, Cambridge, MA
- Linder S H, Peters B G 1995 The two traditions of institutional designing: Dialogue versus decision. In: Weimer D L (ed.) *Institutional Design*. Kluwer, Boston, pp. 133–60
- March J G 1994 *A Primer on Decision Making*. The Free Press, New York
- March J G, Olson J P 1988 The institutional dynamics of international political orders. *International Organization* 52(4): 943–70
- Odell J S 1982 *U.S. International Monetary Policy*. Princeton University Press, Princeton, NJ
- Odell J S 1990 Understanding international trade policies. *World Politics* 43: 139–67
- Reinicke W H, Deng F 2000 *Critical Choices*. IDRC, Ottawa
- Richardson J 1995 Actor based models of national and EU policy-making: Policy communities, issue networks and epistemic communities. In: Kassin H, Menon A (eds.) *State Autonomy in the European Community*. Routledge, London
- Risse-Kappen T (ed.) 1995 *Bringing Transnational Relations Back In*. Cambridge University Press, Cambridge, UK
- Rose R 1993 *Lesson Drawing in Public Policy*. Chatham House, Chatham, NJ
- Ruggie J G 1975 International responses to technology. *International Organization* 29(3): 557–84
- Ruggie J G 1998 What makes the world hang together? *International Organization* 52(4): 855–86
- Sabatier P A, Jenkins-Smith H C (eds.) 1993 *Policy Change and Learning: An Advocacy Coalition Approach*. Westview, Boulder, CO
- Searle J 1995 *The Construction of Social Reality*. The Free Press, New York
- Simon H A 1983 *Reason in Human Affairs*. Stanford University Press, Stanford, CA
- Simon H A 1985 Human nature in politics: The dialogue of psychology with political science. *The American Political Science Review* 79: 293–304

- Social Learning Group 2001 *Social Learning and the Management of Global and Transboundary Risks*, 2 Vols. MIT Press, Cambridge, MA
- Stone D A 1989 Causal stories and the formation of policy agendas. *Political Science Quarterly* 104(2): 281–300
- Thomas C W 1997 Public management as interagency cooperation: Testing epistemic community theory at the domestic level. *Journal of Public Administration Research and Theory* 7(2): 221–46
- Weaver R K, Rockman B A (eds.) 1993 *Do Institutions Matter?* Brookings Press, Washington, DC

P. M. Haas

Policy Knowledge: Foundations

Private foundations, typically found in countries with strong market economies and comparatively weak states, directly fund public policy analysis by the support they provide to research projects, policy institutes, and individual scholars who investigate policy questions. Foundations also advance public policy agendas by supporting advocacy groups and social movements that in turn have policy agendas. In most countries foundations are prohibited from direct lobbying of legislatures, but they are generally free to support public education campaigns that can contribute to an environment within which different public policies will be favored or disfavored.

1. Definition/History

The private foundation is an institution that privately manages private funds in pursuit of a public good. The source of these funds is normally a wealthy individual, almost always a person whose wealth was acquired in commercial activities. Foundation funds are privately managed by self-perpetuating trustees. The requirement that foundations advance a public good derives from the law under which they are chartered. That law stipulates that the funds made available for the endowment of a foundation, being ‘tax-free,’ must be used for charitable or philanthropic purposes. Foundations operate primarily by giving grants from earnings on their endowments, though in rare instances will also use their endowed wealth for grant making.

The wealthy donor inclined to the philanthropic act as a way to perpetuate his or her name and social influence is not a recent practice. Plato bequeathed property so that his Academy would continue after his death, and in fact the Academy lasted nearly a millennium, from 347 BC to 529 AD, until it was deemed unchristian and dissolved by the Emperor

Justinian. Popes and princes became patrons of the arts to enhance their esteem in the public eye. The royal gift, especially to provide a health or educational service, is known across history and across royal regimes from China to India to Egypt to England. More than a millennium of Islamic philanthropy built libraries, universities, and mosques. Commercial wealth has generated philanthropic endeavors across the centuries, and not just since industrial capitalism. The famous Hotel Dieu in Burgundy was established in 1443 when the benefactor donated a highly valued vineyard. The University of Uppsala was similarly endowed with a gift of productive land. Philanthropy as reflected in the royal gift, the papal patronage, and even the private endowment is the backdrop to what emerged in the advanced industrial democracies, particularly the United States, toward the end of the nineteenth century.

2. Foundations and Industrial Wealth

Late nineteenth century industrial wealth became the basis for a new type of philanthropy—the private foundation. Modern foundations were a self-conscious break from Victorian era charity focused on the immediate relief of suffering and poverty. The distinction between foundation grant-making and charitable giving is often captured in the ‘root cause’ metaphor. Rather than relieve suffering, foundations see their purpose as doing something about the deeper causes that lead to suffering in the first place. The classic example is public health such as a vaccine regimen that is designed to prevent the suffering that accompanies epidemics.

Foundations also differed from the one-time gift to endow a hospital or center of learning. The private foundation, a legally immortal institution, extends the bequest and its purposes into the distant future. That is, rather than endow an institution, the philanthropic act endowed the foundation itself. There are a number of reasons for this institutional innovation, one being the sheer magnitude of the private fortunes gained by controlling a key industry (oil, steel, banking, transport) during the unregulated and largely untaxed decades of the late nineteenth and early twentieth centuries. The private foundation, then, is a place to put wealth too substantial to give away in one’s lifetime.

The twentieth century saw a vast growth in the number of private foundations, and in their aggregate asset base. There are more than 40,000 in the United States alone, and many fewer though still substantial numbers in Western Europe, Japan, Brazil, India, Mexico, and other of the wealthier nations of the world. Many foundations are comparatively small, with annual grant budgets of \$1 million or less. At the other extreme are a few substantially endowed

foundations—such as the Gates and Ford Foundations in the United States and the Wellcome Trust in Great Britain, whose annual grant-making approaches \$1 billion. There are also corporate foundations and community foundations, which together comprise about one-fifth of all foundations in the United States and a smaller percent in other countries.

There was a surge in the number of independent private foundations in the latter decades of the twentieth century, for reasons similar to the first surge that occurred a century earlier. The late twentieth century wealth gained through such new industries as electronics, information, media, and entertainment was at a scale comparable to the nineteenth century fortunes in oil, steel, and railroads. Such growth occurs in a period when the rules of the liberal economy allow, and even encourage, the accumulation of inordinate wealth in private control and which tolerate wealth inequalities at a scale thought improper if not immoral only a few decades earlier. If wealth redistribution through public policy was not in favor in the early twentieth century, neither is it in favor in the early twenty-first. The private foundation becomes, then, a means to recycle some of this wealth, though on terms largely set by the individuals who control it.

3. Foundations and the Political Culture

Social Darwinian theories favored by late nineteenth century industrialists and their apologists was particularly favorable to the emergence of the private foundation (see *Darwinism: Social*). An influential statement was Andrew Carnegie's widely circulated essay, *The Gospel of Wealth* (1889a). Carnegie acquired his vast wealth in the steel industry. He argued that industrial capitalism was providing more and more benefits to ever greater proportions of the population, but acknowledged that its workings inevitably generated substantial surplus wealth under control of the few. 'We accept and welcome, therefore, as conditions to which we must accommodate ourselves, great inequality of environment; the concentration of business, industrial and commercial, in the hands of the few; and the law of competition between these, as being not only beneficial, but essential to the future progress of the race.' (p. 656). Having acquired this wealth under the benign influence of social Darwinianism, Carnegie then concluded that it was his Christian obligation to give it away.

It was Max Weber, in *The Protestant Ethic and the Spirit of Capitalism*, who gave sociological depth to the association between Protestant asceticism and responsibility—with its emphasis on discipline, secular engagement, and deferred gratification—and the thrifty, hardworking business ethic that shaped modern capitalism (see *Weber, Max (1864–1920)*). The Protestant ethic stressed the greatest possible pro-

ductivity, but frowned on the luxurious enjoyment of the wealth so earned. If work is to manifest God's glory, the profits of work are to be reinvested in that which is productive and socially beneficial. To do good works in this life is a sign of grace. Not that salvation could be earned by good works, but doing good as well as being good was to be expected of wealth. This was a convenient doctrine for men, Carnegie notably among them, who were both devoutly religious and wealthy beyond easy measure. It is comforting to know that one can dedicate oneself to acquisition and yet be virtuous.

Although explicitly religious motives are not often cited in today's more secular culture, the Protestant ethic retains its hold. The notion that one is a 'steward' of surplus wealth still appears in explanations of the philanthropic act, as do echoes of the belief that good works in this life signal moral worth. Late twentieth century billionaires speak of an obligation to repay society for having been so good to them. They are in this respect reflecting the notion that the wealthy *because* they are wealthy are obliged to be philanthropic. Inherent in wealth is the responsibility to be responsible for the public good that one's wealth can promote.

The social Darwinian environment within which foundations emerge was compatible with another feature of, especially, American political culture—the weak state tradition. Understood as private funds linked to public purpose, the private foundation is a social invention consistent with this tradition. If wealth is accumulated in an amount that is too substantial to be consumed or given away in one's lifetime, and thought to be too great to bequeath to family members, there are a limited number of ways to dispose of it. Furthermore, if the political culture as well as law and taxation policy invite the wealthy to create private institutions that can function as an alternative to the state, there is much satisfaction to be gained by creating a foundation. In the twentieth century, then, the private foundation has become a major mechanism by which large-scale, privately held capital memorializes the donor, endows his or her philanthropic inclinations, and not incidentally substantially reduces tax obligations. In some instances it also serves to keep control of a business enterprise in the hands of a founder's family and descendants.

There is a long tradition of inquiry that sees in the establishment of foundations the intent by the donor to exercise lasting political-economic-social control. The very wealthy do not earn their fortunes without exercising economic control—of an industry, a marketing strategy, a scientific innovation, a new product. Nor are these individuals indifferent to how policies, regulations, tariffs, and taxes can help or hinder the acquisition of wealth. It is not much of a leap from these facts to presume that transferring private wealth to a private foundation is a way to extend political-economic influence.

For example, a Commission of the United States Congress (1915–16), observed that industry and wealth in the United States had come under the control of a few very wealthy persons. The Commission alleged that this small group intended to continue and extend this control by establishing privately managed funds for indefinite purposes. Here, then, is coupled the late nineteenth century anxiety about the concentration of economic power with the fear that private foundations would perpetuate the dominance of business interests in American political and social life.

Careful historical analysis needed to test this hypothesis is in short supply. There is, however, ample anecdotal evidence that donors who establish foundations have a strong inclination to see that their funds enhance particular political and economic doctrines, especially private enterprise and market capitalism. Perhaps there is an irony in the fact that the foundation results from the accumulation of substantial private wealth and yet in instance after instance declares that its mission is to improve the lot of the poor and powerless. It goes about that mission by helping to lower the barriers to upward mobility, by working to insure basic civil and political rights or by contributing to education and health for the poor. It does not go about that mission by calling into question the political-economic arrangements that allow for unequal wealth acquisition in the first place. This despite the oft repeated claim that foundations will find and eliminate the ‘root causes’ of poverty, discrimination, and illness. These root causes, however, seem not to include the economic doctrine and public policies that govern wealth acquisition.

4. The Public Policy Agenda

Based on statistics it collects from thousands of private foundations, the US based Foundation Center uses a ten-part subject-matter classification to report grant making: arts and culture, education, environment, health, human services, international affairs and development, public/society benefit, science and technology, social science, and religion. Although data collection is less systematic in other nations, these categories are generally serviceable across the private foundation world. Public policy is not its own discrete category, because policy analysis and policy advocacy can appear in each of the ten categories. Foundations have taken up arts policy, education policy, health policy, science policy, security policy, and much more.

Of course the policy interests of foundations alter with the times. To note obvious examples: the sizeable investment by American foundations in area and international studies, language training, and student exchanges is a product of post World War II conditions; the rise of environmental consciousness in the

1970s led to a marked increase in funding for environmental policy in Japan, Europe, and the US; the collapse of communism suddenly opened up opportunities in Eastern Europe and the former Soviet Union to which both European and American foundations were quickly responsive. Such examples could be multiplied over and over, indicating that the foundation mission and especially how it expresses itself in grant making is not frozen but rather adapts and evolves as the policy conditions change.

The foundation has a number of strategic choices available as it considers whether and how to confront the public policy agenda. Considered from the broadest possible perspective, foundations adopt strategies based on their notions about how social change can best be brought about. Does the foundation operate on the assumption that ideas drive history; or that technologies do; or social movements, market incentives, government interventions, moral exhortation? If ideas drive history, invest in intellectual efforts; if government interventions drive history, invest in specific policy analysis and advocacy; if exhortation drives history, invest in public education. Or, to use a metaphor much favored by foundations, grant making should seek out the point of ‘leverage’ judged to be most productive of desired social change. The point of intervention is sometimes indirectly and at other times directly linked to public policy knowledge or policy advocacy.

4.1 New Knowledge

Foundations have made major investments in universities and in research institutions, equipment, and personnel, as part of a broad strategy to advance basic human understanding in nearly every field of study imaginable: anthropology, astronomy, biology, chemistry, economics, history, medicine, music, physics, political science, and psychology, to name only the most obvious. These investments have, at times, been sustained over long periods, illustrated by the Rockefeller Foundation’s century-long involvement with medical science. And for some foundations for certain periods of their history, the search for new, fundamental knowledge defined what they were about, what they took to be their vocation. Seldom, however, was new knowledge sought for its own sake. The foundation is focused on *public* benefits and thus even an investment in basic science will always be linked to how the new knowledge is to be applied.

4.2 New Knowledge Applied

Foundations have always been focused on how new knowledge can be applied in socially beneficial ways. An example is the investment in the public health profession, viewed as a means to ensure that dis-

coveries from biomedical science will reach the broadest population possible, and especially those population groups that might otherwise be ignored. There are many other examples: basic knowledge in learning theory being coupled with school reform efforts, basic knowledge in plant physiology leading to higher yielding food crops and their dissemination in poorer countries. In this application effort, the foundations make heavy use of the social sciences more generally and policy studies more specifically. (See *Policy Knowledge: Universities*.)

4.3 Direct Policy Analysis

Foundations have had a long, close relation with studies and publications that would inform public policy. This can be seen in the heavy investment in policy research institutions in the United States, and in policy reports and white papers in European countries. In the United States the independent policy institute and the independent foundation are linked because they are responsive to the same deep strand in the political culture that prefers the private to the public sector (Prewitt 1995). Foundation funding of the social sciences, which starts before the turn of the century, has generally been motivated by the quest for improved public policy. The close relationship between foundation funds and public policy analysis underscores how privately held philanthropic resources attempt to advance public purposes. This relationship has been a complicated one, attracting its share of sharp criticism from both the political left and political right. Notwithstanding this criticism, foundations have forthrightly declared their intent to fund analysis that will lead to what they view as public policy improvements. For example, the German-based Bertelsmann Foundation articulates as one of its primary goals to provide new concepts and models for the social market economy and for public administration.

4.4 Policy Advocacy

Often in combination with policy analysis, foundations have funded a seemingly endless number of organizations dedicated to fighting for or against particular public policies: in human rights, equal opportunity, school reform, nuclear disarmament, environmental protection, free speech, children's welfare, and on and on. Although a dimension of American philanthropy early in the century (Sealand 1997), such funding gained momentum in the post World War II period as the federal government became increasingly active in areas long of interest to the philanthropic sector. If foundation funds could stimulate the government to adopt the 'right' policy in

civil rights or social welfare, the impact would be far greater than what could be accomplished with foundation funds alone. From this emerged the notion that philanthropic dollars could be taken to scale by leveraging government funds. That is, if government through its regulatory and taxation powers influences private expenditure patterns, philanthropy reverses the direction of influence by spending private funds on policy analysis and advocacy that will change government spending priorities. In earlier years, the policy advocacy groups supported by foundations largely focused on domestic policy. In recent years, however, attention has broadened to include the international arena and international actors; policies of the World Bank or the World Health Organization or multinational corporations are targets for internationally organized advocacy groups, especially in the fields of environment and human rights. The post-1989 preoccupation with what is often if not very precisely described as civil society funding is a combination of policy advocacy and social movement philanthropy. (See *Policy Knowledge: Advocacy Organizations*.)

Though overlapping with policy advocacy, social movement philanthropy is sufficiently different to merit independent mention. In recent decades, the major beneficiaries have been social movements organized around equality for women, environmental protection, and civil rights for racial and ethnic minorities. Although each of these social movements originated outside philanthropy, foundation dollars rather quickly led to their stabilization and professionalization. Support for social movements is largely a post World War II phenomena, getting its start with the activism that characterized the 1960s. This type of funding has also been internationalized, as worldwide organizations have formed to advance environmental protection or women's rights or children's welfare. What foundations refer to as support for grass-roots organizations, or community groups, is often an element of social movement philanthropy.

Closely associated with such funding is an emphasis on *social empowerment*. This term refers to the direct empowerment of individuals, especially the traditionally powerless or disenfranchised social groups. An example is micro-credit lending, which is designed to put resources directly in the hands of the powerless poor. Social movement and social empowerment philanthropy are linked in that the former has the dual purpose of empowering its adherents and changing the social landscape. But direct empowerment need not be part of an effort to create a social movement. The beneficiaries of a micro-credit lending program or a female education strategy are not viewed as members of a movement in the same way that feminists or ecological activist are. Social empowerment, of course, is a close cousin to charity—the direct relief of suffering. But philanthropy has defined itself as not about charity so much as by its intent to remove the 'root causes' that lead in the first place to the suffering

for which relief is sought. In its effort to permanently change the condition of the disenfranchised, empowerment is distinct from direct charity.

4.5 Public Education

Foundations have often tried to influence the policy knowledge of the general public. There are numerous examples where foundation funds have been used to improve public understanding as a way to correct for what are considered 'misguided' public practices. A public inclined to permit inefficient and corrupt partisan politics, for example, might be taught to appreciate the virtues of independent expertise in government and a professional civil service—and turn of the century philanthropy took up this progressive cause. Or, a public too given to racial prejudice should be enlightened to the moral worth of racial tolerance—a major philanthropic project across the twentieth century. The examples multiply endlessly—the public is inefficient in its use of energy, unclear about the dangers of population growth, illiterate about its national history, prone to unhealthy eating habits, smokes too much, and, repeatedly, is insensitive to the plight of, depending on the times, women, workers, children, the poor, immigrants, minorities, the handicapped or others from a long list that changes its emphasis from decade to decade, but is never absent from the philanthropic portfolio.

These deficiencies in public knowledge, understanding or attitude can be corrected by the selective use of foundation resources, such as introducing new elements into the school curriculum or through targeted media campaigns. More than public education is at stake, for these efforts often have a change in public policy as the goal. The anti-smoking campaign illustrates this. Foundation supported public education intends not just to discourage smoking, but also to modify public policies on, for example, where tobacco can be sold or how it can be advertised.

What are generally described as 'liberal foundations' have been especially adept at using public education campaigns to shape public opinion, though in recent decades foundations with a more conservative inclination have supported the writings of public intellectuals on what they see as failures in public education, social welfare, criminal justice, and other projects long associated with liberal foundations.

5. Entrepreneurial Philanthropy

Through the twentieth century the attention of foundations to policy knowledge and analysis focused almost exclusively on government policies. Insofar as the behaviors of business sector firms were of interest,

the strategy of choice was to affect regulation, taxation and other public policies to influence market behavior. In the latter decades of the twentieth century this exclusionary focus on government policy began to alter. Some foundations directly addressed the policies and practices of market actors. For example, in the field of public health, foundation funds were used as an incentive to induce pharmaceutical firms to invest in AIDS vaccine research and with respect to energy efficiency, foundation funds attempted to influence utility companies to give more attention to renewable energy sources such as wind and solar power. Micro-credit lending (below market loans in impoverished areas for small-scale entrepreneurs) also became a favored target of foundation funds.

With the arrival on the foundation scene of young, recently wealthy entrepreneurs from the high technology industries this form of philanthropy is likely to increase. The degree to which it will supplant rather than simply add to traditional patterns of public policy funding is unknown, but it is not likely that foundations will soon if ever exit the long-established field of public policy analysis and advocacy.

6. Conclusions

Foundations are well established in advanced market economies, especially in those countries with a comparatively weak state tradition. Over the decades a receptive network of regulatory and tax policies have emerged to make it easy to establish foundations, and thus to withdraw funds from the tax stream on the assumption that these funds will serve the public good. Foundations have used these tax-free funds to support studies of alternative public policies in areas of interest such as health, environment, education, the arts and so forth. They have contributed significantly to the growth of policy knowledge. As more and more nations of the world adopt liberal market economies, it is likely that foundations will emerge in ever greater numbers. It is also likely that public policy analysis and advocacy will remain a foundation funding priority.

Bibliography

- Arnove R (ed.) 1980 *Philanthropy and Cultural Imperialism: The Foundations at Home and Abroad*. G. K. Hall & Co., Boston
- Carnegie A 1889a Wealth. *North American Review* **June**: 653–64
- Carnegie A 1889b Wealth. *North American Review* **June**: 682–98
- Cunningham M 1972 *Private Money and Public Service: The Role of Foundations in American Society*. McGraw-Hill Book Company, New York
- Foundation Giving: Yearbook of Facts and Figures on Private, Corporate and Community Foundations*. 1999 The Foundation Center, New York

- Fosdick R 1952 *The Story of the Rockefeller Foundation: Nineteen Thirteen to Nineteen Fifty*. Harper and Brothers, New York
- Keppel F 1930 *The Foundation: Its Place in American Life*. Macmillan Company, New York
- Mohn R 1997 Objectives of an operating foundation. In: (ed.) *The Work of Operating Foundations: Strategies—Instruments—Perspectives*. Bertelsmann Foundation Publishers, Gutersloh, Germany
- Prewitt K 1995 *Social Sciences and Private Philanthropy: The Quest for Social Relevance*. Indiana University Center on Philanthropy
- Sealand J 1997 *Private Wealth & Public Life: Foundation Philanthropy and the Reshaping of American Social Policy from the Progressive Era to the New Deal*. Johns Hopkins Press, Baltimore

J. Goldenberg

Policy Knowledge: New Liberalism

New Liberalism is at its core a contrast term, designed to distinguish it from and to set it against classical liberalism, as the latter developed in the West between the seventeenth and mid-nineteenth centuries. In use internationally between the 1880s and the 1940s, the term referred to a broad-gauged movement in social knowledge, public philosophy, state policy, and social practice intended to address the 'critique of political economy' and the 'social question' that the pernicious instabilities and distributive inequities of modern capitalism had engendered, and to provide new critical thinking about the crucial relationships between *state*, *economy*, and *society*. The movement should not be taken as a total repudiation of the older tradition, with its characteristic stress on individual autonomy, private property, and limited government. Rather, new liberalism was undertaken initially as an exploration of how and where to amend the old ideals in light of what had come to be seen as negative consequences of capitalist industrialization, and as a method for helping to shape and choose among new civic and social moralities under construction in the industrializing nations of the West.

New liberals were ultimately unable to sustain their critical assessment of the classical tradition without changing quite fundamentally their understanding of the relationship between society, economy, and state. In an intricate process of reaction and inversion, these changes shifted the meaning of political conservatism in the West, gradually severing it from its basis in quasi-feudal and aristocratic precedents and reorienting it toward a defense of personal, civil, and commercial liberty as it had earlier been defined in the 'old,' classical liberalism. The account of new liberalism provided here refers mainly to the American

case, with occasional references to other (mainly the British) experiences. It identifies four phases in the development of new liberalism that aid in heightening the visibility of critical moments in its construction. These phases are: (a) market failure and the drive for new social knowledge, (b) conflicting approaches to framing the social question, (c) the 'turn' to social insurance, and (d) the social sciences, planning, and the limits of liberal statism.

1. Market Failure and the Drive for New Social Knowledge

The devastating market failures of the late nineteenth century, destabilizing waves of bankruptcies, mass unemployment, and chronic, desperate poverty produced an enormous disappointment of traditional liberal expectations regarding the efficacy of the market. New liberals recognized market failure and attendant social conflicts as evidence of a wider collapse, more disruptive to intellectual, social, and political order and more menacing to communities of a social contract based on unfettered competition between white males that had been implemented during the market revolution of the Jacksonian era. Rejecting the excessive individualism, localism, and states-rights obstructionism of that period as morally and intellectually bankrupt, they put themselves quite deliberately in the business of conceptual innovation, seeking a higher rationality that would place knowledge in the service of a more democratically constituted society, which would maintain itself in part by allotting intellectuals a special role in shaping a more vigorous, capable, developed, and purposeful national government.

Across the range of nations involved, the conceptual revolution that new liberals sought to promote was grounded institutionally in an international social science association movement that fostered transatlantic borrowing, and in critical journalism, settlement houses, and social surveys, all of them cranked up to record, analyse, and disseminate empirical data, propound theory, encourage the development of political leadership responsive to the 'teachings' of liberal social knowledge, improve the execution of policy, and advocate reforms intended to renovate and enable Enlightenment ideals of political equality and social justice. Within government, in investigative agencies such as improved censuses, labor, women's, and children's bureaus, trade boards, and bureaus of corporations, high-level civil servants such as Llewellyn Smith and Lord Beveridge in England, and Carroll Wright, Francis Amasa Walker, and Julia Lathrop, in the United States provided conduits for policy ideas from those burgeoning 'parastates' into the state. Within this 'great transformation' in social thought and political practice, crucial early players were the

academic social science disciplines, whose missions and boundaries were historically determined, differing from country to country. Academic professionalization proceeded rapidly in the American research universities, fueled by rapid economic growth, academic entrepreneurship, and the relative shortage of other types of creditable specialized intelligence.

But from the beginning, progressive academic economists, political scientists, and sociologists were interested in more than the provision of expertise; even as conservative university trustees (and at times conservative colleagues) disciplined their advocacy, social scientists saw themselves as crucial links in the interpretive processes needed to make the new political and social economy intelligible. Thus their specialized investigations and discourses were a critical factor in reworking the political agenda so that it placed prominently before policy makers and emerging, activist publics as problems urgently needing comprehension the range of issues that came to define the politics and policymaking of the new liberal era. Most prominent among these issues were business cycles and unemployment, rights and conditions of labor, wages and distribution, industrial organization and corporate reorganization, family welfare and gender roles, problems of the low-paid, exploited immigrant work force and—often with disabling assumptions rolled in—issues of race. They also addressed related issues of governance: the detection of wholesale corruption of law and government that inspired the clean government crusade, and the need for methods of building up the cognitive abilities and administrative capacities of public officials.

Strategic rhetorical steps were taken early that shaped the ties between American governance and intellectual life in the period. In the United States, Woodrow Wilson's early and generative 'Science of Administration' (1887) addressed republican fears of bureaucracy by placing a firewall between politics and administration. Reflecting distaste for the intense, tribal partisanship of Gilded Age politics, academics and bureaucrats (unlike many female social scientists and social reformers) adopted nonpartisanship as a standard description of their professional ethic. This pristine canon elevated everyday standards of disciplinary and agency practice, but it often hampered the effectiveness of social scientists as advocates for change, routing their activism through ostensibly apolitical venues such as foundations and think-tanks, which ceded partial control of research agendas to corporate elites.

2. Conflicting Approaches to Framing the Social Question

Yet despite frequent and often calculated professions of scholarly and professional neutrality, the building

of new liberalism was neither a consensual nor a non-ideological process, but rather one in which newly experienced social realities disrupted expectations and categories in different and conflicting, contested ways. Deep divisions over the substance and implications of new empirical and theoretical knowledge led to a branching of new liberalism, early in its American career, into what participants saw as two distinct new liberal movements—a *statist-democratic* strand of reform thought born in the 1880s, and an opposing, *corporate-voluntarist* analysis that came together around the turn of the century. These conflicting reform discourses, worked out in dialogue with each other between the 1880s and the 1940s, opened up greater possibilities for genuine policy innovation than existed either earlier or since.

Assessing deficiencies in the mechanisms for translating an aroused social conscience into effective social action, the statist strand of new liberalism looked in new ways and on a number of fronts to a redefinition and a principled restriction of the meaning of 'private' in many of its applications, and to a corresponding expansion of what were to be taken as proper, morally necessary functions of government. Liberal statistes sought more effective methods for mobilizing state agencies in behalf of a reinvigorated national interest, taken as something higher and more ethically compelling than the sum of the separate private, local, class, or sectoral interests.

The competing voluntarist vision rejected with equal conviction what—telegraphing their disapproval—its founders referred to as 'state interference.' Seeking ways to insulate and energize civil society, the anti-statists preferred to build new ordering strategies upon what they took to be the naturally evolving social institutions of a modern pluralist society, premier among them the corporation, and what they saw as organic, corporate bodies such as trade associations, trade unions, professions, and charitable associations, rather than around invigorated democracy, political mobilization, and a bureaucratic, regulatory state, which they represented as intrusive upon 'natural' evolutionary processes and prone to manipulation, most dangerously by social levelers who failed to appreciate the vulnerability of the investment system. More than merely Tocqueville revivalists, they were particularly interested in what could be achieved by way of ordering a chaotic, excessively competitive economy through innovative uses of the business contract, within a rich climate of information about the market, and through enlightened management. This effort culminated in the liberal-corporatist design fleshed out during the 1920s for peak-level business-government cooperation in achieving economic and social stabilization through energized philanthropy and local relief in hard times, but mainly preventively, through the development and legitimation of industrial self-government oriented toward controlling chronic instabilities such as overproduction.

The liberal statist placed greater confidence in deliberative democracy than this, and less in capitalist rationality or the unity of the corporate and the public interest. Drawing energy from heartland republicanism, evangelical Christianity, and European socialism, a leading theorist of the new liberal state, political economist Henry Carter Adams, marked out a path that two generations of unabashedly statist progressives would define as a democratic, progressive, collectivist or communitarian public philosophy—a distinctively American middle way between what they framed as the excessive individualism of the British model and the excessive statism of the German. The founding text for the liberal statist approach was Adams's classic paper, 'Relation of the State to Industrial Action,' which interrogated standing assumptions about the beneficence of competition and called into question the fundamental morality and the long-term serviceability to the public interest of the way the modern capitalist market was being constituted in its corporate phase. For those engaged in reflecting on issues of public philosophy, Adams, along with Richard Ely, Florence Kelley, Edmund James, Simon Patten, John R. Commons, and Jane Addams among numerous others, did more than establish a 'context of refutation' for socialists and die-hard believers in *laissez faire*. More important, in relation to later developments in theoretical and policy knowledge, they directed their arguments quite explicitly to new-liberal antistatists, who were forming their own intellectual identity in the 1890s as part of a corporate-voluntarist counter-attack on liberal statism.

Reflecting back on the historical construction of the ante-bellum US liberal order, Adams pointed to three deficiencies of a regime of market competition that set the initial terms on which this new, positive vision of liberalism would be built: (a) its failure to allow society to enjoy the benefits of monopolies occurring naturally in increasing-returns industries, and thus to reap the benefits of human genius enacted in advancing technology, (b) the inevitable decline of business ethics to the lowest level, absent a generally credited method for communal determination and legal establishment of labor standards, among them a right to employment, and (c) a corruption and general hollowing out of the substance and potential of representative government in a deliberate, misguided weakening of state capacities, which Adams traced to anti-statism in the Jackson period, the lack of a respectable science of public finance in the United States, the rise of powerful corporations, and a compliant legal profession bent on shaping the protections of the law to private ends.

Naturalizing certain instances of monopoly was a critical step in the modernization of civic republicanism, designed to pull that venerable tradition out of the Jeffersonian, anti-monopoly, anti-statist discourse that had assimilated it to market liberalism. The reshaping of this core category laid an authoritative

basis for state supervision of entire segments of the transportation, communication, and production systems, designed to control monopoly power. This strategic conceptual move gathered force in the regulation movements of the early progressive era, and in Theodore Roosevelt's vision of a national law of corporations, a robust national citizenship, and a muscular 'new nationalist' bureaucracy capable of disciplining the modern corporation. Classic texts such as Herbert Croly's *Promise of American Life* (1909) and the Progressive Party Platform of 1912 represented the maturing new liberal orientation to planning, in which capable, informed governance could instruct and protect a collective interest in conservation, labor standards, and consumer protection. More than merely an affirmation of the need for aggressive regulation in these regards, Rooseveltian progressivism argued for a public presence in the civic discourses and decisions regarding production, investment, and pricing. This broader vision of a deliberative process that could detect and shape the impact of American capitalism on society was later the unifying insight at work in the planning phases of the New Deal, most evident in the activities and discourses of the Public Works Administration (PWA), the National Resources Planning Board (NRPB), and the Temporary National Economic Committee (TNEC).

The Adams-Ely view was challenged immediately by procorporate academic economists Arthur Hadley and Jeremiah Jenks, close students of the sources of market failure in specific industries, and political advisers to Republican governors and presidents. Though these two extended what they saw as a natural tendency to monopoly even further than Adams did—into the whole broad range of capital-intensive industries, they considered even indirect state control of production and profit a threat to the price and investment systems that neither liberalism, capitalism, nor private morality could long endure. Conservative institutionalists such as these began constructing an alternative, more hierarchical, less democratic discourse of elite responsibility and social control, forming around the institutions of high finance and the corporation. They hoped to make the new corporate elites responsible for transclass and intercorporate cooperation, for personal growth and individual mobility along career ladders provided by corporate business, and even for community economic and moral development—to the extent that experience within a corporate culture would inculcate values of efficiency, cooperation, interdependence, and service.

This rudimentary modernization theory was the founding discourse of corporate liberalism, which down through Herbert Hoover's associative state to its full flowering in pluralist theory and liberal-corporatist public policy during the Eisenhower years equated the (properly enlightened) corporate interest with the (properly conceived and limited) public interest in sustained growth of the economy. Cor-

porate liberalism provided a systematic alternative to liberal statism, supplying a constraint on statist progressive politics in the US that was available, in anything like comparable strength, nowhere else.

More than other distinctive qualities, such as early (white manhood) suffrage, weak parties, weak labor, weak socialism, or state structure and capacities, this continuing tension between statist and corporatist discourses of the public interest, increasingly fleshed out in programmatic terms, explains—along with issues of racial exclusion that new liberals for the most part only minimally addressed—the special character of modern United States liberalism. It also helps to account for the distinctive pattern of US state formation, and particularly welfare-state development, that students of social policy, often taking the more comprehensive European welfare states as the normal path of history, have seen as tardy, insufficient, and incomplete—kept from developing by traditions of individualism and *laissez faire*.

In reality, such commitments were highly attenuated where they survived at all among the knowledge-producing classes most involved in shaping new liberalism. For democratic theorists from Adams to John Dewey, the public interest was a potent category—reclaimed from the republican past, but extensively recast to take account of concentrations of social and economic power that threatened family subsistence, community morals, responsive and responsible governance, and political, social, and economic democracy. In his seminal *The Public and Its Problems* (1927), another powerful exemplar in the new liberal ‘role of the state’ genre that H. C. Adams inaugurated for Americans, Dewey located the public interest as the place in social relations where effects spilled over from the parties to a contract to third parties, who bore the often devastating ancillary consequences of others’ private bargains.

In the construction of democratic statism, problems raised by shoddy business ethics and new, manipulative forms of ‘scientific’ management rivaled concerns about monopolistic corruptions of the corporate form. Instituting another enduring theme in new liberal reflection about the role of the state, the second ‘evil of competition’ in Adams’s treatise—pernicious insecurity, exploitative working conditions, and relentless time pressure—he proposed to remedy by labor-standards legislation based on a community moral consensus derived discursively, from detailed knowledge of the impact of industrial labor as then constituted, on the moral, mental, and civic capacities of the working class. The debilitating effect of casual labor, low wages, and chronic unemployment on working class men and women was a constant theme that women’s social investigation and reform activism also did much to develop. But this emphasis on legislated labor standards should not be mistaken for an early blueprint for the welfare state. In the formative moment, the 1880s to 1910s, American

statist progressives were looking not to compensatory welfare measures but to independent unions, fostered by public intellectuals and protected by the state, and to community moral pressure and democratic politics, as the means for achieving an ‘economic republic’ and ‘industrial democracy.’ Their vision included recognition of two inalienable rights of worker citizens—a ‘living’ wage, sufficient to support a family in reasonable comfort, and a voice in management.

3. The ‘Turn’ to Social Insurance

Underway in Europe much earlier, the new liberal ‘turn’ to social insurance came in the American case only in the second decade of the twentieth century. Growing up in social work, settlement, bureaucratic, and academic circles, it signaled the onset of a second phase in new liberal thinking, running parallel, but in important respects contrary, to the older movement for industrial democracy, which also continued to gather force through the New Deal and into the late 1940s. Several factors can be seen to have promoted the adoption of a social insurance perspective, among them the arrival of a broad environmentalist consensus on the causes of poverty, accumulating realism about distinctive constraints on labor’s capacity to organize a positive political program, and the social tensions that compelled American cities to provide relief during cyclical downturns from the 1870s onward. Once again, academic and expert knowledge played a major role as well. European reform discourses and precedents, culminating with the British adoption in 1911 of broadly accessible unemployment insurance, were faithfully catalogued in works such as Isaac Rubinow’s *Social Insurance* (1913). As class conflict accelerated, inquiries staffed by social scientists such as the famous Pittsburgh Survey and the United States Commission on Industrial Relations documented economic insecurity as the central cause, along with corporate arrogance and powerlessness to combat it in their communities, of worker unrest, suggesting an attack on inequality through aggressive labor standards reforms and state policies promoting redistribution.

Contrary to the strong government role envisaged by the USCIR and in wartime programs for improving labor standards, in the post-World War I period, with Herbert Hoover in a prominent organizing role, the Wilson and Harding administrations arranged a series of high-level conferences for planning postwar stabilization. Allied with the President’s Conference on Unemployment, Wesley Mitchell’s National Bureau of Economic Research drew in economists such as Leo Wolman and Wilford I. King for the vital statistical work and propagated new methods for understanding the pesky business cycle. This chain of policy thinking suggested alternative policy options, among them more responsible administration of credit by banks and government, enrichment of the informational climate for business decisions to facilitate a

microeconomic approach to steadying output and employment, the accumulation of reserve funding for countercyclical public works by (from Hoover's perspective, preferably local) government, and an actuarial approach to relieving unemployment through funds amassed by employers, unions, or government. In what must be seen as one of the most significant conceptual innovations in modern governance, these more conservative approaches transformed unemployment from a problem of the political order rooted in inequality into a 'risk,' the key policy decision becoming therefore how, and by whom, that risk—understood as a problem in the smooth management of the elastic supply of reserve labor accepted as an indispensable part of the capitalist process—could be insured. In the discourse manufactured in Hooverian circles, most of these remedies could be implemented by informed business decisions and enlightened labor-management cooperation.

The impulse for what from the perspective of liberal-statist, laborist, and social-feminist proposals for extensive redistribution of income and political voice became a critical retrenchment in social policy thinking arose not only from these liberal corporatist initiatives, but from at least three other, at first glance unlikely sources as well: the maternalist turn in the women's movement of the progressive era; a vision of the immigrant and non-white poor propagated in the social sciences, by the Chicago school of sociology most particularly; and the appropriation and aggressive expansion of carefully targeted forms of social provision in the private-sector, imagined by corporate liberals in business management, government, and the academy as a strategy for making the 'modern manors' of welfare capitalism reliable instruments for securing the loyalty and uninterrupted productive service of industrial employees.

As the central sites for the development of reform discourses that transported women's issues from the private to the public spheres, the women's movement and female social science were neither immediately nor exclusively committed to a social insurance perspective. Female social scientists such as Jane Addams, Florence Kelley, Alice Hamilton, Crystal Eastman, and Mary van Kleeck initially broadened the social question to include the appalling conditions faced by working class families in immigrant enclaves, conditions they traced as much to deliberate capitalist exploitation as to the relentless, patterned instability of the capitalist economy. Constitutive new liberal texts such as the pioneering social survey *Hull House Maps and Papers* (1895) and Margaret Byington's 1907 study for the Pittsburgh Survey of steel workers' families traced their poverty to the most discretionary of environmental causes—a calculated policy of exceedingly low wages and insufficient employment, yielding family incomes well below what was needed for subsistence, chosen by leading industrialists as a way to socialize much of the cost of maintaining a low-

skilled industrial work force, including a good-sized reserve labor supply.

Female reformers such as these translated their expert knowledge of the working conditions of women and children into an ethic of public responsibility that matched and complemented the initiative of academic new liberals such as H. C. Adams for finding ways of translating communal judgments regarding the 'moral plane of competition' into policy. Often promoting and then working in state and national investigative and protective agencies, rights-thinking feminists were thus among the earliest new liberal state-builders. Yet, as conservative opposition accumulated and the suffrage crusade shifted to political mobilization, important segments of mainstream feminism took on a more elitist, exclusionist cast, and the 'female dominion' of women's institutions increasingly followed a separatist course in identifying its reform goals. Leading female reformers, sociologists, and legal realists built a maternalist case for special state-based income assistance to mothers and children, special provisions for safeguarding maternal and infant health, and special protections for republican mothers in the workplace.

Maternalists helped to shape as the American pattern what women's historians have identified as a two-channel welfare state, in which one channel provided universal coverage for male workers under reasonably generous programs such as workers' compensation, unemployment insurance, and Social Security, whereas the other offered only meager, means-tested pensions for virtuous poor widows, and little assistance of any kind for most female workers and people of color. Worthwhile in themselves as strategies for meeting real needs among the 'deserving' poor, these approaches sidestepped, or confronted only indirectly, and without formulating them as such, issues of monopoly power, proletarianized, low-wage work, and obstacles to unionization for workers of both sexes that had initially been catalysts for the development of democratic-statist new liberalism.

The social sciences provided powerful legitimization for the social insurance track as an alternative to empowering the bottom ranks of the working class. Between the 1890s and the mid-twentieth century, even as institutionalists continued to fashion economic knowledge in forms that emphasized law and capitalist agency as contexts for the political and cultural creation of markets, rival neoclassical economists cultivated statistical and mathematical models of analysis that extracted economic behavior from history and culture, naturalized the workings of the market, and turned the discipline into a potent, technocratically organized form of social engineering. As Dorothy Ross's account of the rise of scientism also documents, quantitative, behavioristic and functionalist models vied with historicism and institutionalism in other social sciences as well, perhaps most visibly in two massive collaborative efforts in Hoover-

era policy knowledge, *Recent Economic Changes* (1929) and *Recent Social Trends* (1933). The dominant sociological discourse of the 1920s, the 'Chicago School' established by Robert Park, W. I. Thomas, and their students at the University of Chicago offered an alternative to the liberal-statist explanation for poverty and urban ills that took the form of a cultural theory of urban family and community 'disorganization.' Describing a succession of discrete stages—contact, disorganization, acculturation, assimilation—that accounted for the rise of problems such as dysfunctional families, juvenile delinquency, and crime, these scholars offered a vision of the immigrant, and especially the African-American migrant poor as culturally deficient—whether from environmental, or more enduring cultural causes. They uncovered 'pathologies' in Negro personality and family patterns, an alleged inability to defer gratification rooted in the experience of slavery, and a tendency toward male improvidence and female-headed families, for example, that predicted a longer than normal acculturation process for African Americans. Culminating with Gunnar Myrdall's *American Dilemma* (1944), this new cultural analysis of poverty separated the poor from the capitalist strategies that shaped their lives, and rendered those historically constituted structures less visible.

Maternalist social knowledge and the sociology of ethnic difference merged with developments in the new sciences of management to promote the development of corporation-based programs for linking carefully constructed forms of social provision to employment with specific employers. When wartime controls on the economy were quickly abandoned, corporatism flourished in the United States (as it did in different forms during the interwar period in Europe as well). Doctrines of imperfect competition, an institutionalist analysis of the modern corporation's separation of ownership from control, arts of advertising based on a new consumer psychology that legitimized desire, and theories of an administered market allowed liberal corporatists to proffer a new economy of mass consumption as a palatable alternative to nineteenth-century producerist dreams of worker control.

Resisting this trend only formally, the American Federation of Labor, along with industrial relations experts constructing a new profession in the 1920s, recognized productivity increases as the indispensable basis for increases in wages. Where company welfare programs existed, worker loyalties based on them ran quite deep. Thus the shock of the sudden collapse of Hooverian strategies for industrial self-government in the Great Depression crisis, the failure of voluntarism more generally, the welcome proliferation of New Deal programs targeting unemployment and poverty, and the aggressive organizing efforts finally unleashed by national labor-rights legislation helped to ratchet up workers' expectations of the state, and to reinvigorate impulses, largely quiescent since the im-

mediate postwar era, toward the development within government of planning capacities capable of heading off such emergencies in the future, and of democratizing both industrial governance and the political state.

4. *The Social Sciences, Planning, and the Limits of Liberal Statism*

From their beginnings in the Enlightenment and especially since the late nineteenth century, the social sciences were oriented toward planning, defined imprecisely enough to include methods all the way from enriching the informational climate of decision making to managing economic performance and shaping cultural expression. Most centrally a conservationists' discourse of efficient natural resource management in the 1910s and a welfare corporatists' discourse about the benefits of social cooperation in the 1920s, the American conversation about planning matured rapidly in the depression emergency. As critics correctly charged, stressing continuities in the development of a national planning capacity that bridged private and public, the National Recovery and Agricultural Adjustment Acts could seem like little more than surreptitious continuations of corporatist strategies for contriving scarcity under the cover of state compulsion. Yet a parallel, consciously statist movement also reappeared in the 1930s, fueled by FDR's intense commitment to planning, challenges on the left that promised full utilization of idle resources and full employment, and the recruitment to government planning agencies of a social scientists, including a large number of institutionally oriented economists, who energized and updated the democratic statist strand of new liberal thinking. When crisis suddenly opened the way for bureaucratic efforts to comprehend and combat depression, many of the plans necessarily came 'off the shelf,' as did the comprehensive programs for national old age, survivors, unemployment, and disability insurance. Yet by the mid-1930s institutional and policy innovations shifted the planning discourse toward methods of using resources efficiently over the long term, expanding public investment, crushing debilitating tendencies toward monopoly, and rekindling demand—policy goals institutionalists insisted were related causally. Social scientists in key nodes of conceptual innovation redefined planning and, within the constraints of opposition that also bedeviled Franklin Roosevelt's own narrower planning ambitions, attempted to give it an overt civic dimension. Thus, in the early leadership of the National Labor Relations Board, William Leiserson and others experimented with methods for linking the civic and stabilization purposes of the Wagner Act, which its framers intended to ensure the worker's right to organize independently, but also to equip trade unions

to play a role in redistributing income, supporting the workers' ability to consume, and democratizing industry.

Economists' knowledge and planners' leanings toward infusing market structures with civic purposes also blended in agencies such as the venerable Bureau of Agricultural Economics, the path-breaking, community-reconstruction oriented, ultimately disappointing Tennessee Valley Authority, and the Farm Security Administration, which addressed the issue of chronic farm surpluses, but also the economic and social marginality experienced by people at the bottom of rural hierarchies. A major study by TNEC of the concentration of economic power steered by Thurman Arnold and Leon Henderson referred the stickiness of prices, excess saving, and the slow pace of investment back to the monopoly power that allowed large corporations to administer markets and perpetuate extreme maldistribution of income and wealth.

The meaning of planning was shifted most decisively by theoretical, institutional, and practical innovations generated by the NRPB, as that New Deal agency's role was enlarged by Roosevelt from resource management to guiding government efforts to achieve recovery, and later to maintain prosperity and promote economic democracy after World War II. A body of economic knowledge produced in the early 1930s by economists such as Edward Chamberlain, Joan Robinson, Alvin Hansen, and John Maynard Keynes explained how a stubborn equilibrium, well below the level of full employment of the factors of production, could have been reached and sustained over nearly a decade. Building upon this, economists at NRPB and the Federal Reserve zeroed in on underconsumption as a factor in the slow and uneven pace of recovery and pointed to gains to be achieved by more aggressive deficit financing of public expenditures in the form of worthwhile civil works. By the definition of planning that flowed from this new thinking, the state became responsible for continuously monitoring and occasionally dramatically augmenting or shifting investment flows in socially constructive directions not attractive to the private capital markets, and for limiting entrenched corporate power over prices and redistributing income to support consumption. Among the most advanced elaborations of New Deal statism, *Security, Work, and Relief Policies* (1942) conveyed NRPB's synthesis of three related but discrete government roles: guaranteeing work for the segment of the population securely attached to the labor market, financing and extending social insurance to care for those unable to participate in the work economy, and underwriting a still highly contested version of industrial democracy through the NLRB and the self-organization of labor. Through its promotion of this triad, the NRPB enunciated a liberal statist's vision of a new social contract.

New Deal planners managed only a partial and temporary fulfillment of the moralizing role Henry

Carter Adams had charted for knowledge-based civic discourse back in the 1880s, implemented in limited ways through the Fair Labor Standards Act and the Fair Employment Practices Commission. The NRPB's interest in postwar planning was one of the factors that led to the unceremonious termination of the agency in 1943, a milestone in the history of new liberalism and in the history of the social sciences as well. The commitment of New Deal planners to full employment proved untenable in the face of rising opposition to the New Deal. Though the Council of Economic Advisers authorized by the watered-down Employment Act of 1946 lacked the authority to trigger government spending at a level needed to achieve full employment, the agency continued at least through the Truman era to press for a major government role in the postwar expansion of the economy, and to jawbone against monopoly as an obstacle to growth.

Meanwhile, opponents of planning clustered around nodes represented by the resurgent free market liberalism of Frederick Hayek and by concerns among business conservatives and Southern industrialists, fanned by legal activism and the uproar created by *American Dilemma*, that federal government controls might invade local prerogatives regarding labor and race. The most influential opponents of government authority in these regards were undoubtedly corporate elites and corporate liberal thinkers who considered the associative vision of business-government cooperation a better method longterm. Herbert Stein, Allan Brinkley, and others have located a 'fiscal revolution' in economic policy thinking, underway by 1937, in which a competing body of policy knowledge suggested that the measured use of fiscal and monetary levers would enable macroeconomic stabilization without intrusion into capitalist decision-making processes or extensive interference with labor markets, which in this view could safely be left to private management in the context of 'commercial Keynesianism' and a compensatory state.

Whereas the first generation of American liberal-statist social scientists had reconstituted government as a knowledgeable, morally responsible regulator, and the second generation of planners had modeled the state as social insurer and mediator of class antagonisms, the third, New Deal generation of planners conceived a state role not well captured as state capitalism. It was closer to older Hamiltonian and Whig models of a development state responsible for liberating and energizing community efforts and for facilitating the balanced economic and civic development of the commonwealth. Policy discourses of the 1950s, developed around Cold War internal security issues, an increasingly consensual growth-oriented commercial Keynesianism, modernization theories exported from the United States as technologies for liberal internationalism, and domestic discourses of racial and gender identity, challenged the

assumptions of new liberal social science, fostered new methods for interpreting modern social relations, and inaugurated a new ideological era. But embodied in resilient structures of disciplinary knowledge formation, in durable habits that demand grounding of political argument in social empiricism, and in a chastened but resilient Enlightenment tradition of political obligation, the legacy remains—indeed, in important though little recognized ways, recurs in the dialogue between contemporary liberals and communitarians. Absent the polarizing tendencies of a world divided for half a century between capitalism and state communism, and given now the resurgence over recent decades of libertarianism and free market ideology, it becomes ever more obvious that the critical decisions in democratic governance in the West, and beyond, will henceforth be—as they were in the new liberal era, 1880s to 1940s—hard choices between variants of liberalism.

See also: Capitalism; Feminist Theory; Liberal; Liberalism; Liberalism: Historical Aspects; Liberalism: Impact on Social Science; Social Justice; Social Question: Impact on Social Thought

Bibliography

- Alchon G 1985 *The Invisible Hand of Planning: Capitalism, Social Science, and The State in The 1920s*. Princeton University Press, Princeton, NJ
- Barber W 1996 *Design Within Disorder: Franklin D. Roosevelt, the Economists, and the Shaping of American Economic Policy, 1933–1945*. Cambridge University Press, New York
- Berkowitz E, McQuaid K 1992 *Creating The Welfare State: The Political Economy of Twentieth-century Reform*, rev. edn. University Press of Kansas, Lawrence, KN
- Brinkley A 1995 *The End of Reform: New Deal in Recession and War*. A. A. Knopf, New York
- Clawson M 1981 *New Deal Planning: The National Resources Planning Board*. Johns Hopkins University Press, Baltimore, MD
- Collini S 1979 *Liberalism and Sociology: L. T. Hobhouse and Political Argument in England, 1880–1914*. Cambridge University Press, New York
- Croly H 1909 *The Promise of American Life*. Macmillan Co., New York
- Eisenach E 1994 *The Lost Promise of Progressivism*. University Press of Kansas, Lawrence, KN
- Ethington P 1992 Hypotheses from Habermas: Notes on reconstructing American political and social history, 1890–1920. *Intellectual History Newsletter* 14: 21–40
- Farr J 1989 Understanding conceptual change politically. In: Ball T, Farr J, Hanson R (eds.) *Political Innovation and Conceptual Change*. Cambridge University Press, New York pp. 24–49
- Fraser S, Gerstle G 1989 *The Rise and Fall of the New Deal Order, 1930–1980*. Princeton University Press, Princeton, NJ
- Freeden M 1978 *The New Liberalism: An Ideology of Social Reform*. Clarendon Press, Oxford, UK
- Furner M O 1975 *Advocacy and Objectivity: A Crisis in the Professionalization of American Social Science, 1865–1905*. The University Press of Kentucky, Lexington, KY
- Furner M O, Supple B (eds.) 1990 *The State and Economic Knowledge: The American and British Experiences*. Cambridge University Press, Cambridge, UK
- Gerstle G 2001 *American Crucible: Race and Nation in the Twentieth Century*. Princeton University Press, Princeton, NJ
- Hawley E 1966 *The New Deal and the Problem of Monopoly*. Princeton University Press, Princeton, NJ
- Hecl H 1974 *Modern Social Politics in Britain and Sweden*. Yale University Press, New Haven, CT
- Jackson W 1990 *Gunnar Myrdal and America's Conscience: Social Engineering and Racial Liberalism, 1938–1987*. University of North Carolina Press, Chapel Hill, NC
- Jacoby S 1997 *Modern Manors: Welfare Capitalism Since the New Deal*. Princeton University Press, Princeton, NJ
- Joas H 1993 *Pragmatism and Social Theory*. University of Chicago Press, Chicago, IL
- Judis J B 2001 *The Paradox of American Democracy: Elites, Special Interests, and the Betrayal of the Public Trust*. Pantheon Books, New York
- Kloppenborg J 1986 *Uncertain Victory: Social Democracy and Progressivism in European and American Thought, 1870–1920*. Oxford University Press, New York
- Koven S, Michel S 1990 Womanly duties: Maternalist politics and the origins of welfare states in France, Germany, Great Britain, and the United States, 1880–1920. *American Historical Review* 95: 1076–108
- Kraditor A 1971 *Ideas of the Woman Suffrage Movement, 1890–1920*. Columbia University Press, New York
- Lacey M J, Furner M O 1993 *The State and Social Investigation in Britain and the United States*. Cambridge University Press, Cambridge, UK
- Lacey M (ed.) 1989 *The Truman Presidency*. Cambridge University Press, Cambridge, UK
- Milkis S M, Mileur J M (eds.) 1999 *Progressivism and the New Democracy*. University of Massachusetts Press, Amherst, MA
- O'Connor A 2001 *Poverty Knowledge: Social Science, Social Policy, and the Poor in Twentieth Century US History*. Princeton University Press, Princeton, NJ
- Patterson J T 2000 *America's Struggle Against Poverty in the Twentieth Century*. Harvard University Press, Cambridge, MA
- Persons S 1987 *Ethnic Studies at Chicago, 1900–45*. University of Illinois Press, Urbana, IL
- Reagan P 1999 *Designing a New America: The Origins of New Deal Planning, 1890–1943*. University of Massachusetts Press, Amherst, MA
- Rodgers D 1998 *Atlantic Crossings: Social Politics in a Progressive Age*. Belknap Press of Harvard University Press, Cambridge, MA
- Ross D 1991 *Origins of American Social Science*. Cambridge University Press, Cambridge, UK
- Rueschmeyer D, Skocpol T (eds.) 1996 *States, Social Knowledge, and the Origins of Modern Social Policies*. Princeton University Press, Princeton, NJ
- Silverberg H (ed.) 1998 *Gender and American Social Science: The Formative Years*. Princeton University Press, Princeton, NJ
- Sklar K K 1993 The historical foundations of women's power in the creation of the American welfare state, 1830–1930. In:

- Koven S, Michel S (eds.) *Mothers Of A New World: Maternalist Politics and the Origins of Welfare States*. Routledge, New York pp. 43–93
- Sklar M 1988 *The Corporate Reconstruction of American Capitalism, 1890–1916: The Market, Law, and Politics*. Cambridge University Press, New York
- Skocpol T 1992 *Protecting Soldiers and Mothers: The Political and Social Origins of Social Policy in the United States*. Belknap Press of Harvard University Press, Cambridge, MA
- Stein H 1969 *The Fiscal Revolution in America*. University of Chicago Press, Chicago, IL
- Tomlins C 1985 *The State and the Unions: Labor Relations, Law, and the Organized Labor Movement in America, 1880–1960*. Cambridge University Press, Cambridge, UK

M. O. Furner

Policy Knowledge: Nonprofit Institutions

The complexity of advanced industrial systems requires governments to acquire an enormous range of specialized knowledge and administrative expertise with which to order their affairs. Uniformity in demand, however, has not produced uniformity in the character and supply of knowledge for government purposes. There are significant variations, related to diverse cultural traditions in the means of acquiring knowledge and applying methods of expertise. For example, central government bureaucracies in Western Europe and Japan directly provide many more sources and opportunities for the development of policy expertise than they do in the USA, which in modern times has relied heavily on private research organizations operating outside government bureaucracies for functional expertise. As the federal role in economic and social policy expanded in the post-World War II era, increased federal funding did not produce a commensurate increase in the size of government. Instead, it strengthened the complicated network of relationships among the federal government, state and local governments, and private groups, including nonpartisan research institutes (Alchon 1985, Critchlow 1985, Ricci 1993, Smith 1991).

In this policy system, the federal government operates indirectly, through intermediary organizations in state and local government, and in civil society (Heclo 1980). This reliance on private, nonpartisan research organizations imparts a unique quality to the acquisition and use of knowledge by the administrative state in modern America.

The vast size, complexity, and expense of this system of acquiring knowledge by government in the USA suggests the pluralistic nature of the American polity, deeply rooted in its republican distrust of centralized government. This distrust of centralized government found expression in the nation's constitution, drafted

in Philadelphia in 1787. Its famous beginning 'We the people' asserted that the source of its authority rested in the people rather than state or centralized government. This distrust of centralized government found expression in the antistatist Jeffersonian–Jacksonian tradition that emerged early in the nation's history.

1. Social Investigation in the Nineteenth Century

In the environment of relatively weak centralized government prior to the US Civil War, the craft of social investigation developed within a middle-class subculture of reformers, many of them women, who documented the plight of slaves, the poor, prisoners, and the insane. The affinity of antebellum social reform and feminism became evident in reports issued by American abolitionists who documented the conditions of slave life, as well in reports on the care of prisoners, the indigent, and the insane. In the 1840s, social reformer Dorothea Dix issued a series of groundbreaking reports to state legislatures calling for the humane care of the mentally ill. This tradition of empirical research into social problems conducted by private individuals associated with reform movements was extended late into the nineteenth century (Marshal 1937, Nelson 1992, Leach 1989, Daniels 1968, Dupree 1957).

By the outbreak of the Civil War, state legislatures in the most populous states regularly received detailed narrative and statistical reports on dependency, crime, public health, labor conditions, and economic development. These reports relied more on democratic and humanitarian sentiments associated with moral reform than they did on claims of science and disinterested expertise. Female reformers, as suggested by the role of Dorothea Dix, played an especially important role in this early endeavor to bring social knowledge to state legislatures and the public (Furner and Lacey 1993, Ross 1991).

While the capacity for social knowledge grew on the state level, the role of the federal government's activities in the pursuit of knowledge remained limited generally to surveying, land exploration, and modest technological development through private contracts. This relative institutional impoverishment reflected a unique antistatist tradition in America, reinforced by sectional politics of the period. The major institution in Washington for the gathering and accumulation of scientific and social knowledge remained the Smithsonian Institution, which opened in 1846 at the instigation of the individual for whom it was named. In the decades following the Civil War, the Smithsonian Institution became a center for people such as Francis Walker, John Eaton, John Wesley Powell, and Carroll D. Wright, who took leading roles in establishing a rudimentary infrastructure of social investigation in the nation's capital. These individuals served

as conduits for a transition from moral reform to scientific reform.

2. The Twentieth Century

By the beginning of the twentieth century a large resident class of scientific and professional civil servants had emerged in Washington. In 1904, the US Bureau of Census reported that 3,919 civil servants fell into the classification of 'professional, technical and scientific' workers. Within this group, nearly 1,100 people were described as 'special agents, experts, and commissioners,' while another 300 or so were designated simply as 'scientific experts and investigators' (Lacey 1993, pp. 127–70, Ross 1991, Blumer et al. 1991, Keller 1977, Skowronek 1982, Wilson 1909).

Many of the early organizers in the Smithsonian Institution played a critical role in the establishment of the American Social Science Association (ASSA) in 1865. The ASSA sought to serve as a clearinghouse of social knowledge in order to promote reform in education, public health jurisprudence, finance, and social economy. Directly influenced by the achievements of the Sanitary Commission during the Civil War and modeled on the British National Association for the Promotion of Social Science, the ASSA attracted commercial elite groups interested in promoting sound fiscal and monetary policy; lawyers, doctors, ministers, and teachers who sought professional affirmation; and middle-class reformers who combined humanitarian moral reform with the new scientific spirit. In this way, the ASSA, and other professional organizations that arose during this period, paved the way for a new liberalism, grounded in social empiricism (Haskell 1977, Furner 1975, 1993).

These reformers called for new measures of efficiency and economy in voluntary charity organizations and in public administration. Behind this concern with efficiency and economy lay a deep conviction that the polity was threatened by corruption of partisan political machines serving special economic interests. These political machines, run by local and state political bosses, consistently manipulated the mass of voters to maintain political power in flagrant disregard of any notion of public morality or efficiency and economy in government. Thus progressive reformers called for new social knowledge to help restore political order and representative government to American society. Beneath the catchphrase 'efficiency and economy'—the rallying cry of reformers—lay strong antimajoritarian values fused with antiparty perspectives (Critchlow 1985, Alchon 1985, Karl 1974).

This movement for efficiency found early expression in late nineteenth-century charity endeavors. Amid concern that relief was being wasted on the undeserving poor, new professionals were brought into charity organization to centralize recipient roles,

provide vouchers for relief funds to be distributed to recipients, institute periodic checks on recipients, and implement budget systems within agencies to keep a close watch on funds. Essential to this reform was the use of data gathering through statistical measures and new accounting procedures. Behind this movement lay a sentiment expressed by industrialist Andrew Carnegie, who declared, 'One of the most serious obstacles to the improvement of our race is indiscriminate charity.' Carnegie's infatuation with the new social sciences would lead him to create a number of new nonprofit philanthropic institutions concerned with the gathering of public knowledge (Lagemann 1989).

At the vanguard of this movement for efficiency and economy in government were university trained and highly specialized social scientists. In the latter half of the nineteenth century, the basis of social science research was transformed from deductive reasoning derived from natural law into inductive procedure founded on historical analysis and empirical research. This new methodology allowed social scientists to mask any preoccupation with political matters in the terminology of scholarly objectivity and empirical research. The transformation of political impulses into the quest for public knowledge and administrative solutions reflected the aspirations of a new professional class anxious to gain influence in a highly industrialized society through expertise. They sought reform by placing 'the best people'—objective-minded experts—in government (Ross 1991, Haskell 1977).

The call for efficiency and economy in government, and the need to develop new public agencies for the gathering of public knowledge essential to restoration of representative government, began on the municipal level first in New York with the creation of a Bureau of Municipal Research in 1905, organized by social scientists and civic-minded philanthropists and businessmen. Municipal research bureaus would be established in several other US cities. Central to these municipal research bureaus was the creation of executive budget systems that allowed a strong executive office, public officials, and the general electorate to track how public funds were spent. Essential to the functioning of an executive budget was the creation of an agency staffed by nonpartisan experts who would prepare and analyze budgetary finances and expenditures (Critchlow 1985, Karl 1974, Smith 1991).

This demand for an executive budget system reached the national level under the administration of William H. Taft in 1910. When Taft assumed office in 1909, the federal budget stood at over a billion dollars and the federal debt had reached disturbing proportions, largely as an aftermath of the 'Panic of 1907.' In these circumstances, Taft was able to persuade Congress to appropriate \$100,000 for a commission on efficiency and economy in government. To staff the commission, Taft called on the best social scientists in the country. After two years of painstaking work, the Taft Com-

mission came out with its report, *The Need for a National Budget*. Yet, because the proposal was made in an election year, a recalcitrant Congress refused to enact an executive budget system. The failure of the Taft Commission did not dismay reformers, however. In 1915 they created the Institute for Government Research—the predecessor to the Brookings Institution. Only in 1921 during the Harding administration would budget reform be enacted in the Budget and Accounting Act of 1921.

Taft's successor, Woodrow Wilson reflected in many respects the deep ambivalence that middle-class Americans felt about the state during the early twentieth century. An array of new government agencies was now involved in conservation, consumer protection, corporate regulation, trade practices, and monetary and financial policies. Yet Wilson, in his pioneering essay on public administration warned that the model of Prussian administration should be avoided in America. He declared that European bureaucratic practices would have to be adapted to fit the highly decentralized form of government in the USA. If the European model is to be employed in the USA, he told his fellow reformers, 'We must Americanize it ... We must learn our constitutions by heart and we must get the bureaucratic fever out of its veins' (quoted in Cuff 1993, p. 390).

Until America's entry into the war in April 1917, neither the federal government's defense department nor civilian departments had considered mobilization tasks. As a consequence, private planning proceeded public planning. Financial and manufacturing firms such as J. P. Morgan and Du Pont Corporation, by receiving Allied war orders, were among the first to assess the financial, technical, and labor requirements of munitions production (Brownlee 1993).

War, however, forced the enormous administrative expansion of the federal government in the short period between April 1917 and November 1919. New government agencies appeared in haphazard fashion, drawing corporate and academic personnel into government management. This involvement led many to the belief that the capacity for public knowledge and for effective social action rested in the mobilization of private, public-spirited professional and business groups willing to work in cooperation with a sympathetic state. The war demonstrated for these individuals how to forge a uniquely American combination of state-society relations.

Herbert Hoover sought to put this insight into practice as Secretary of Commerce in the Harding and Coolidge administrations, and later as president, by using the techniques of private-public cooperation he had learned as a wartime administrator. Hoover sought to develop state-voluntary cooperation in the collection of specialized data required for rational decision-making in both the private and public sectors. For Hoover and others concerned with public knowledge, the war revealed just how essential government

knowledge was for economic and social planning. Through the development of an administrative state he sought to develop new mechanisms for collecting public knowledge without the coercive element of state power (Hawley 1974).

In keeping with his views about the credibility of privately financed researchers, Hoover arranged for the Rockefeller Foundation, founded in 1913, to fund a comprehensive survey of social changes in American life between 1900 and 1930. The three-year study resulted in the massive 1,600-page *Recent Social Trends in the United States: Report of the President's Research Committee of Social Trends* (1933). This study presented the full panoply of available knowledge for public purposes (Karl 1974, Hawley 1974, Alchon 1985).

Although state capacity for gathering public knowledge expanded dramatically during the Great Depression era, World War II, and the postwar period, the federal government continued to rely indirectly on intermediary organizations in state and local government, and in civil society, for the accumulation of knowledge. The commensurate increase in the size of the federal government led to the creation of fluid networks of policy specialists, who operated in and between government and the private sector, identifying policy issues, generating 'solutions' to policy problems, and overseeing and evaluating the implementation and administration of policies within their specific areas of competence (Heclo 1980).

3. Role of the Think Tanks

Essential to these policy networks are private policy research institutes or 'think tanks,' a peculiarly American social invention. By the 1990s there were an estimated 1,200 non-government think tanks of various descriptions. World War II and the postwar continuation of a large defense program, as well as expansion of federal social programs, was a watershed in the evolution of the American think tank.

At the end of World War II, Americans made a political decision to have government train, subsidize, and regularly consult a sizable population of civilian experts on such issues as defense, welfare, health, and economic stability. Of the major think tanks, only the Brookings Institution (1916) and the Carnegie Endowment for International Peace (1910) were founded before World War II. The American Enterprise Institute was founded during the war in 1943 (Critchlow 1985, Smith 1991, Lagemann 1989, Dickson 1971).

Although think tanks are ostensibly nonpartisan, in many instances they function as extensions of state power, their influence rising and falling with changes in governments and shifts in the ideological climate of the country. In other cases, think tanks function more independently, questioning and monitoring state strat-

egies and structures. For example, the Rand Corporation of the 1940s and 1950s was designed to strengthen the programs of the US Air Force, whereas the Institute for Policy Studies (IPS) has been a major center of criticism of US foreign policy, domestic social programs, and policies in the areas of civil rights and poverty.

Often the ideological orientation of these think tanks reflected changes in the historical context and political climate, as indicated by the dominant trends in high and popular culture, changes in voting behavior, and shifts in political leadership (Ricci 1993).

The course of the Brookings Institution from the New Deal through the end of the twentieth century reflected these ideological currents. While Brookings studies in the early 1930s criticized New Deal programs, the Brookings Institution under its president Harold Moulton tried to avoid becoming directly involved in partisan politics. By 1937, however, the institution had emerged as a vociferous critic of New Deal fiscal policy.

In response to the new postwar environment and the reluctance of foundations to fund an institution they perceived as ineffective or out of touch, Robert Calkins, former head of the General Education Fund at the Rockefeller Foundation, agreed to head the institution. Calkins reorganized Brookings and recruited social scientists with liberal credentials and government experience. This new group had close ties with government, and unlike the devotees of the earlier nonpartisan ideal, aligned themselves closely with presidential administrations. In 1965, when Calkins retired, Brookings was representative of mainstream Keynesian economic thinking, and its growing influence was reflected in renewed foundation support, especially from the Ford Foundation. Under Calkins' successor, Kermit Gordon, Brookings' reputation as a liberal Democratic think tank was well entrenched. As a result, the institution became a major research center and a holding area for policy experts who were temporarily out of office. In these years, the Brookings Institution became a major center for policy innovation in welfare, healthcare, education, housing and taxation policy (Critchlow 1985, Smith 1991).

Divisions over foreign and domestic policy opened cracks in the liberal consensus in the early 1970s, however, leading many observers to see the institution as isolated from large segments of business. In 1976, the board of trustees appointed Bruce MacLaury to head the institution. A former regional Federal Reserve banker and Treasury official, MacLaury successfully courted business support, increased corporate representation on the board of trustees, and moved the institution toward a more moderate ideological stance.

By the 1970s, the Brookings Institution confronted competition from other major policy research institutions, especially the American Enterprise Institute and the Heritage Foundation—both viewed as conserva-

tive research institutions close to the Republican Party.

The American Enterprise Institute (AEI), which was founded in 1943 as the American Enterprise Association (AEA), illustrates the experience of a conservatively oriented research institution that expressed deep ambivalence about the post-World War II policy consensus (Critchlow 1985, Smith 1991). The key figure behind the establishment of the AEA was Lewis Brown, chairman of Johns-Manville Corporation. Brown realized that the New Deal had changed the political environment for business and, although he expressed anxiety about the level of government intervention, he accepted much of the New Deal program. From the start, the AEA reflected a conservative bias.

In 1954, A. D. Marshall, head of General Electric, assumed the institution's presidency and immediately hired William Baroody Sr. and W. Glenn Campbell, both staff economists at the US Chamber of Commerce, to head the research program. Under their guidance, AEA was gradually built into a modern research institute under its new name, the American Enterprise Institute. Principal support came from private foundations and major corporate sponsors. The institution's reputation was enhanced when the Nixon administration called upon a number of AEI associates to fill government positions. The AEI also emerged as a successful proponent of economic deregulation.

In 1977 Baroody retired and his son, William Baroody Jr., took over the presidency of the institution. To improve its standing in the academic community, the AEI assembled an impressive staff including Melvin Laird, William Simon, Robert Bork, Michael Novak, and Herbert Stein. The tenure of William Baroody Jr., however, ended abruptly in the summer of 1987, when an increasingly restive board of trustees forced his resignation because of cost overruns and declining revenues. Moreover, despite the large number of AEI staffers who held appointments in the Reagan administration (30 left AEI to join the administration in its first two years alone), members of the New Right, suspicious of what they considered the centrist orientation of the institution, voiced bitter criticism of the AEI.

Baroody's successor, Christopher DeMuth, bolstered the conservative orientation of the institute by bringing on board several former Reagan administration officials with strong rightist reputations.

The founding of the Heritage Foundation in 1973 revealed a new ideological climate in the analysis of public knowledge (Edwards 1997, Peschek 1987). Founded by Edwin Feulner and Paul Weyrich to provide rapid and succinct legislative analysis on issues pending before Congress, the Heritage Foundation sought to promote conservative values and demonstrate the need for a free market and a strong defense. The Heritage Foundation's articulation of conserva-

tive values in social policy, education, and government activities placed it at the forefront of New Right activity. Although sympathetic to the Reagan wing of the Republican Party, Heritage's commitment to conservative economic theories and social values allowed the institution to be critical of both Republicans and Democrats. The Heritage Foundation remained relatively small in its early years, but the election of Ronald Reagan to the presidency in 1980 enhanced the institution's prestige. By the mid-1980s the Heritage Foundation had established a solid place in the Washington world of think tanks as a well-organized, efficient, and well-financed research organization that called for the turning over of many government tasks to private enterprise, a strong defense, and a cautious approach to Russia and China.

During these years, myriad other think tanks emerged in Washington representing a range of ideological positions and specialized policy interests, including the left-oriented Institute for Policy Studies (1963) and the libertarian-oriented CATO Institute (1977). Think tanks concerned with national security included the Center for Strategic and International Studies (1962) and the Center for National Security Studies (1962) affiliated with the American Civil Liberties Union (Friedman 1997).

These private institutions served to supplement social research conducted by federal agencies and institutions. By the 1990s the number of personal and committee staff of Congress exceeded 11,000, with a further 20,000 people responding to the information needs of the legislature from the Congressional Budget Office, the Office of Technological Assessment, and the General Accounting Office (Lacey 1993). The executive branch commands far greater numbers. By the 1990s the scale of spending by the federal government for the acquisition and application of public knowledge was massive. Funding is currently routed through an estimated 180 bureaus and agencies. This research includes a mixture of 'in-house' research conducted by career civil servants, and a larger body of research conducted by outside parties via grants and contracts on a 'for profit' basis and some of it by university-based scholars on a 'not for profit' basis. An estimated third of this research activity is supported by the federal government to meet its needs, while the remaining two-thirds is authorized for state and local government, school systems, policy forces, welfare agencies, and industrial and agricultural groups. In addition, a vast array of trade and professional organization and 'public interest' groups conduct research in the areas of environmental policy, consumer affairs, civil rights, and foreign policy issues (Critchlow 1985). While the government relied on think tanks for policy innovation, program review, and policy assessment, the federal government relied on nonprofit service providers to implement social programs. In this way, the expansion of the welfare

state in the 1960s and 1970s created a seeming paradox that remains apparent in the contemporary regime: the rapid growth of the nonprofit, voluntaristic sector occurred at the same time that federal social policy became increasingly centralized and active. The parallel development of the centralized welfare state and the voluntary sector was no coincidence. The growth of federal social programs mobilized the nonprofit sector, while in turn the nonprofit sector, through policy innovation and program implementation, helped legitimize the modern welfare state.

Thus, as the traditional federal system that had existed since the founding of the nation eroded in the 1960s, when the federal government extended its powers at the expense of state power, the nonprofit sector was called upon to help design, implement, and administer social programs that had once been within the jurisdiction of the public sector (Hall 1992, Kameran and Kahn 1989, Salamon 1992). As a consequence, the nonprofit sector expanded in the 1960s at the very time that the federal government enlarged its powers over the states. The result produced an unusual phenomenon—government centralization produced greater nonprofit sector voluntarism (Critchlow 1999).

4. The 'Mixed' Welfare State

Between 1965 and 1980 the American social welfare state underwent a significant transformation, with government spending on social welfare increasing by 263 percent in inflation-adjusted dollars (Gilbert 1983, Critchlow 1999). The expansion of the welfare state promoted the emergence of a government–nonprofit partnership. Increasingly in the 1960s, the federal government generated the funds but then turned over the delivery of services to other public state and local organizations, and private nonprofit organizations. The private, nonprofit sector became a major beneficiary of this system of services. By the 1980s, the federal government alone provided \$40 billion in support to the private nonprofit sector, while state and local governments provided additional amounts from their own resources (Salamon 1992).

Federal demand enabled the nonprofit sector to grow by leaps and bounds. In 1940 there were only 12,500 charitable, tax exempt organizations. By 1990, there were over 700,000, with most of this growth taking place after the 1960s (Salamon 1992, Hall 1992, Critchlow 1999). While the federal government centralized social policy through congressional mandates, administrative and legislative regulations, and new funding arrangements, the delivery and implementation of these programs were placed in the hands of state and local governments and nonprofit organizations.

The emergence of the 'second welfare state' in the last three decades of the twentieth century—contrary

to the concerns of those who feared that government would replace the nonprofit sector altogether—created a vibrant nonprofit sector. Government, in turn, came to rely on this sector to design and implement programs, undertake demonstration projects, contract with other private agencies, and provide outside consultation and independent review agencies and activities.

As this second welfare state emerged, the nonprofit sector was called upon to fulfill a number of distinct functions. Nonprofit organizations served as *funding intermediaries* distributing funds to other parts of the nonprofit sector, *service-providing organizations* involved in healthcare, education, counseling, adoption assistance, and *political action agencies* supporting particular pieces of legislation or candidates for political office (Salamon 1992). By providing public services, these nonprofit organizations developed a symbiotic relationship with the federal government. By 1990, government grants, contracts, and reimbursements accounted for 31 percent of nonprofit service organizations' income. Just 18 percent of the total income that nonprofit organizations receive comes from private giving (Salamon 1992, Hall 1992).

The proliferation of government and nonprofit institutions involved in public policy marshals expert against expert, testing the credibility of each and enriching the information environment in which policy discourse occurs and decisions are taken. Indeed, on highly technical issues, opposing experts often directly contradict one another. Yet the growth of public and private agencies involved in collecting public knowledge allows for a more democratized policy process—much closer, in a sense, to the Jeffersonian–Jackson vision of an open, competitive democracy.

Still, two qualifications are in order: neither Jefferson nor Jackson envisioned a world of competing experts representing various interests and ideological positions; and secondly, this contemporary world of competing institutional research is not equally open to all comers. Access to financing is one limitation, and access to those in power is another.

The proliferation of expertise and the expansion of government knowledge that has arisen in response to the demand for more policy inquiry have not necessarily improved the information base on which political decisions are made. The emergence of more complicated patterns of political mobilization by a larger and more diverse array of interests and groups has necessarily politicized the nature of policy knowledge and thereby redefined the nonpartisan ideal espoused at the beginning of the twentieth century. Competition for attention to research places heavy demands on the time and attention of legislators and policy-makers who, in a value-laden universe, are left, ultimately, to decide what information counts for them on the basis of their own values. A widespread acceptance of relativism, embodied institutionally in public and

private research organizations, has irrevocably altered the process of social investigation and the uses of social knowledge.

See also: Centers for Advanced Study: International/ Interdisciplinary; Intellectuals, Sociology of; Non-profit Sector, Geography of; Policy Information; Policy Knowledge: Advocacy Organizations; Policy Knowledge: Universities; Strategic Planning; Think Tanks

Bibliography

- Alchon G 1985 *The Invisible Hand of Planning: Capitalism, Social Science, and the State in the 1920s*. Princeton University Press, Princeton, NJ
- Bensel R F 1990 *Yankee Leviathan: The Origins of Central State Authority in America, 1859–1877*. Cambridge University Press, Cambridge, UK
- Blumer M, Bales K, Kish Sklar K (eds.) 1991 *The Social Survey in Historical Perspective, 1889–1940*. Cambridge University Press, Cambridge
- Brewer J 1990 *The Sinews of Power: War, Money and the English State, 1688–1783*. KNOF, Cambridge, MA
- Brock W R 1984 *Investigation and Responsibility: Public Responsibility in the United States, 1865–1900*. Cambridge University Press, Cambridge, MA
- Brownlee W E 1993 Social investigations and political learning in the financing of World War II. In: Lacey M J, Furner M O (eds.) *The State and Social Investigation in Britain and the United States*. Cambridge University Press, Cambridge, pp. 323–61
- Cassedy J H 1969 *Demography in Early America: Demography in Early America: Beginnings of the Statistical Mind*. Harvard University Press, Cambridge, MA
- Cohen P C 1982 *A Calculating People: The Spread of Numeracy in Early America*. Chicago University Press, Chicago
- Critchlow D T 1985 *The Brookings Institution, 1916–1952: Expertise and the Public Interest in a Democratic Society*. Northern Illinois University Press, DeKalb, IL
- Critchlow D T 1999 *Intended Consequences: Birth Control, Abortion and the Federal Government*. Oxford University Press, New York
- Cuff R D 1993 War mobilization, institutional learning, and state building in the United States, 1917–1941. In: Lacey M J, Furner M O (eds.) *The State and Social Investigation in Britain and the United States*. Cambridge University Press, Cambridge, pp. 388–427
- Daniels G H 1968 *American Science in the Age of Jackson*. Columbia University Press, New York
- Dixon P 1971 *Think Tanks*. New York
- Dupree, Hunter A 1957 *Science in the Federal Government*. Harvard University Press, Cambridge, MA
- Edwards L 1997 *The Power of Idea: The Heritage Foundation at 25 Years*. Jameson Books, Ottawa, IL
- Friedman J (ed.) 1983 *The First Harvest: An Institute for Policy Studies Reader, 1963–1983*. Grove Press, New York
- Furner M O 1975 *Advocacy and Objectivity: A Crisis in the Professionalization of American Social Science, 1865–1905*. University of Kentucky Press, Lexington, KY
- Furner M O 1993 The Republican tradition and the new liberalism: Social investigation, state building, and social learning in the gilded age. In: Lacey M J, Furner M O (eds.)

- The State and Social Investigation in Britain and the United States*. Cambridge University Press, Cambridge, pp. 171–242
- Furner M O, Lacey M J 1993 Social investigation, social knowledge and the state: An introduction. In: Lacey M J, Furner M O (eds.) *The State and Social Investigation in Britain and the United States*. Cambridge University Press, Cambridge
- Furner M O, Supple B (eds.) 1990 *The State and Economic Knowledge: The American and British Experiences*. Cambridge University Press, Cambridge
- Gilbert N 1983 *Capitalism and the Welfare State: Dilemmas of Social Benevolence*. Yale University Press, New Haven, CT
- Hall P D 1992 *Inventing the Nonprofit Sector and Other Essays on Philanthropy, Voluntarism, and Nonprofit Organizations*. Johns Hopkins University Press, Baltimore
- Haskell T 1977 *The Emergence of Professional Social Science: The American Social Science Association and the Nineteenth Century Crisis of Authority*. Illinois University Press, Urbana
- Hawley E 1974 Herbert Hoover, the commerce secretariat, and the vision of an associative state, 1921–1929. *Journal of American History* 61(June)
- Heclo H 1980 Issue networks and the executive establishment. In: King A (ed.) *The New American Political System*. American Enterprise Institute Press, Washington, DC
- Kameran S B, Kahn A J 1989 *Privatization and the Welfare State*. Princeton University Press, Princeton, NJ
- Karl B 1974 *Charles Merriam and the Study of Politics*. University of Chicago Press, Chicago
- Karl B 1983 *The Uneasy State: The United States from 1915 to 1945*. University of Chicago Press, Chicago
- Keller M 1977 *Affairs of State: Public Life in Late Nineteenth Century America*. Harvard University Press, Cambridge, MA
- Lacey M J 1993 The world of the bureaus: Government and the positivist project in the late nineteenth century. In: Lacey M J, Furner M O (eds.) *The State and Social Investigation in Britain and the United States*. Cambridge University Press, Cambridge
- Lagemann E C 1989 *The Politics of Knowledge: The Carnegie Corporation, Philanthropy, and Public Policy*. University of Chicago Press
- Leach W 1989 *True Love and Perfect Union: The Feminist Reform of Sex and Society*. Basic Books, New York
- Marshall H E 1937 *Dorothea Dix: Forgotten Samaritan*. University of North Carolina Press, Chapel Hill, NC
- National Academy of Sciences 1978 *The Federal Investment in Knowledge of Social Problems*. Washington, DC
- Nelson W 1992 *The Roots of American Bureaucracy, 1830–1900*. Cambridge
- Peschek J G 1987 *Policy-planning Organizations: Elite Agendas and America's Rightward Turn*. Temple University Press, Philadelphia
- Ricci D 1993 *The Transformation of American Politics: The New Washington and the Rise of Think Tanks*. New Haven, CT
- Ross D 1991 *The Origins of American Social Science*. Cambridge University Press, Cambridge
- Salamon L M 1992 *America's Nonprofit Sector: A Primer*. Foundation Center, New York
- Skowronek S 1982 *Building a New American State: The Expansion of National Administrative Capacities, 1877–1920*. Cambridge University Press, Cambridge
- Smith J A 1991 *The Idea Brokers: Think Tanks and the New Policy Elite*. The Free Press, New York
- Wilson W 1909 *The State: Elements of Historical and Practical Politics*, rev. edn. D C Heath and Company, Boston

D. T. Critchlow

Policy Knowledge: Universities

All civilizations have needed higher education to train their ruling, priestly, military, bureaucratic, and other service elites. However it was in twelfth-century Europe that an institution recognizable as a university came into being, as a school of higher learning combining teaching and scholarship and characterized by a corporate autonomy and academic freedom. This institution of higher learning has survived in varying forms down to the present day, while evolving a flexibility which has permitted it to serve the needs of enormously different societies in every part of the world. One of those common societal needs which universities in recent years have striven to serve is that of public policy, that can be defined as national policy which is at some level an expression of the public will. Universities have come to interact with the policy process in multiple ways.

The idea of the practical application of thought to the solution of technical and social problems is virtually timeless and universal. However the direct and total involvement of the corporate universities in social and economic activity is of relatively recent origin. It is identifiable in the later stages of the Industrial Revolution and in the Europe of the nineteenth century, when the direct relevance of scientific and intellectual thought was generally discussed and debated in public. However, it was only after World War Two had dramatized the life-saving significance of knowledge production and application that in the second half of the twentieth century a clear relationship to public policy can be discerned: 'The systematic attention of universities to social reconstruction and economic innovation to all aspects of economic development—financial, industrial, entrepreneurial, organizational, managerial—its concerns for the urban infrastructure, epidemiology, family welfare and schooling, the enhancement of life chances and the encouragement of vertical and horizontal mobility, are changes occurring within living memory' (Rothblatt 1995).

1. The Foundations of the Relationship Between Universities and Public Policy

That issues of social and economic development and policy claimed the attention of the university was a combined consequence of developments within the institution itself and in its external political context. The great nineteenth-century development within the university was the emergence of a new paradigm of the institution, in which research was added to teaching as an equal part of its mission.

Twentieth-century universities have systematically developed structures that allowed them to add the function of generating knowledge to the previous ones of preserving and transmitting it. This was followed by

the rapid development of the empirical social sciences as an expanding part of university curricula. Thus research and the production of knowledge emerged as a service to society and the social sciences arrived as a science of society. The potential of both brought the university into a systematic engagement with public policy.

Outside the university in the second half of the twentieth century came the solidification of political systems that derived more fully than before from the wishes of the national public, and required technical and research-based knowledge to satisfy those wishes and to inform policies. National competition to produce new knowledge made universities the key institutions of postindustrial society.

Historically there has always been interplay between what happens inside universities and what goes on in society outside. In the twentieth century this bond became close. The kind of people who work in universities, the type of topics on which they work and the theories, methodologies, and knowledge which they produce all have relevance to the solution of policy problems and improved social wellbeing. As problems became increasingly complex and multidimensional, cross-disciplinary and multidisciplinary concentrations began to develop within universities as a way of responding to the different problems. Specific techniques, such as survey research, manpower planning, opinion polling, rate of return analysis, and systems theory are a few among the techniques and vocabularies of social analysis which universities developed and exported, to societies wishing to digest and utilize knowledge, and to policy-makers anxious for a scientific base to their choices. There was, in the second half of the twentieth century in Europe and America, a social demand from the policy arena for the kind of knowledge, ideas and techniques which trained people within universities tend to produce.

2. Autonomy and Relevance

The extent of this demand created a dual temptation: within the academy for faculty members to direct their work towards perceived policy issues and, outside it, for governments, with the power of the purse, to put pressure on universities to create specific policies backed by suitable academic rationales. The result was a tension between the traditional autonomous role of the university and the demands of partisan government for knowledge that would make policies competitive.

In this situation the critical issue becomes the institutional relationship between the university and the government, which has taken over the formulation of public policy, and the nature of government's claim on the ideas and people producing them, in a context where the essence of the university ethos is autonomy. In the USA and Europe, where concepts of university autonomy are relatively well established, this auton-

omy was protected in large measure by a *de facto* hiving off of the policy function to independent policy research institutions. Often physically close to universities, they are staffed mainly by former university academics who focus on specific policy issues in the company of like-minded scholars.

A dense network of such institutions has grown up in the USA addressing a wide range of policy issues and there is a constant two-way traffic between them and the universities, and between both these and the government and the private sector. With regard to public policy, individuals from universities who had developed strong knowledge and skills in particular policy areas would be recruited for policy positions and consultancies in government. In the USA, Henry Kissinger, Daniel Patrick Moynihan, and Joseph Stiglitz are no more than the better known among the many who took that route in the current era. The critical point here is that they were recruited because of their individual competence and interest, not because of a particular position reflected by their university. Universities as institutions in general try to avoid espousing specific public policy positions and make a virtue of eclecticism. The guiding, albeit unstated, principle by which universities diffuse the tension is that they do not make policies but rather see themselves as enlarging the knowledge base on which policies can be made. The balance is sometimes hard to hold as when, for example, a policy philosophy becomes so strong that it is associated with the institution itself, and, in this connection, the reputation and influence of the so-called 'Chicago School' in the 1960s comes to mind. Other forms of encroachment of the policy-making function on university autonomy become controversial when the policy institute and the university appear synonymous at least to the outside eye, or when the university takes government money for training for research related to debatable policy purposes.

3. The Internationalization of the Social Sciences

Outside the USA and Europe the postwar era of independence in Asia and Africa, and before that in Latin America, led ultimately to the internationalization of social science and hence the university's role in relation to public policy. First, it became clear that theories derived from industrialization and market behavior did not automatically apply in rural settings, imperfect markets, or centralized and state-controlled economies. Policy-making required more contextual understanding derived from history, anthropology and cultural studies and to develop a theory of problem-solving. The content, theories, and techniques available for policy consideration were enriched. Second, the issue of autonomy and relevance, as encapsulated in the relationship between government and universities, took very different expression in developing countries than in the industrialized west.

Instead of a dense array of public policy institutes, policy and research units existed as full and functioning parts of what were to begin with often single national universities. In this period the relationship between university and government was so close that either through a honeymoon sense of postindependence common purpose, i.e., to take control of the foreign-dominated state policy-making apparatus, or because a single party regime did not see much room for university autonomy in the sense of expressing views contrary to national policy, there was little call for independent policy institutes. In many such situations in Africa the universities came close to becoming part of the civil service with a responsibility to develop and present public policy, solve problems, and in so doing be relevant.

The tension between autonomy and relevance comes to the fore and becomes more acute to the extent that the university's role in producing new knowledge and a workable science of society is supplemented by additional social demand for relevance from universities, where relevance is defined not simply in terms of the institution's ability to train its students for jobs but also to define and support specific public policies. This tension was contained by a sense of common purpose between government and universities in Africa for a period but, in Latin America, autonomy was an early casualty to conformity and the independent and knowledge-producing function moved out into independent centers which in essence became the source of alternative policies that in Africa during the same period emanated from the universities.

However, with the end of the Cold War, sweeping international forces transformed the situation and created conditions permitting a new kind of relationship between universities and the policy-making process. Economic liberalization, political democratization, and administrative decentralization all created a favorable environment for genuinely autonomous universities. Universities could now redefine their relationship with the policy-making process in a context of autonomy and chose to favor coexistence with separate policy institutes in line with the Western pattern. This move was strongly promoted by the World Bank, anxious as it was to have recognizable partners for policy dialogue in an economic context increasingly dominated by debt ratios. At the end of the twentieth century the pattern worldwide is for research centers, institutes, and think tanks to multiply on the periphery of universities while university faculties and departments are becoming the place for the provision of teaching.

4. Universities and Public Policy in the Twenty-first Century

Universities were never the sole producers of knowledge for policy consideration but they were para-

mount, and they are perhaps the only institutions with the objectivity and neutrality to train and make provision for the public good dimension of public policy. However, as they enter the twenty-first century they face global forces and technological change which promise a massive transformation and significant reduction in their paramount role. They are being required to function in a larger, more complex environment than existed when they became basic knowledge producers after World War Two. Knowledge production and dissemination, research and teaching will no longer be self-contained activities but will involve integration with a far wider variety of knowledge producers than before, including industry. As always, the universities are being required to respond to external change.

The biggest single source of change in the relationship between universities and the policy process is the emergence of the Internet and the pace of change in information technology. It is influencing enormous changes in the nature of knowledge production, the corresponding role of the university and its relationship to other actors in the informing of public policy. At the same time, globalization and economic pressures of international competition are dissolving the boundaries between national institutions and disciplines and creating widely dispersed knowledge production systems in which universities are a part but have many partners.

Universities have traditionally organized their work into disciplinary science structures. Such structures will retain a privileged place in knowledge production but they will no longer be alone. The disciplinary structure of the production of knowledge is itself changing, as we witness the emergence of a new model in which knowledge is produced in the context of its application rather than in a single theoretical disciplinary structure and is becoming heterogeneous and more socially accountable. Research done in the context of application and global distribution means that 'science' cannot remain easily within the confines of a university department or academic center. The emergence of a host of new institutional arrangements linking government, industry, NGOs, and private consultancy groups will arrange knowledge production in new forms of organization that erode the traditional place of the university.

To the extent that research is cutting loose from disciplinary structures and generating knowledge which does not seem to wish to institutionalize itself in university departments and faculties, universities will be forced to pay more attention than before to links with the new institutions and centers where the research is being generated. Coalitions, partnerships, and networks are becoming more important than institutions. In terms of their relation to policy, the old division of labor in which universities focused their knowledge production on government, and industry on private application, is less relevant and the re-

lationshps which count are those with the host of diverse and globally scattered institutions which now have a voice and a hand in the production of policy-related knowledge.

See also: Centers for Advanced Study; International/ Interdisciplinary; Policy Knowledge: Nonprofit Institutions; Think Tanks; Universities and Science and Technology: Europe; Universities and Science and Technology: United States; Universities, in the History of the Social Sciences

Bibliography

- Altbach P, McGill Peterson P *Higher Education in the 21st Century: Global Challenge and National Response*. Institute of International Education, Annapolis
- Caraca J, Conceicao P, Heitor M V 2000 Towards a public policy for the research university in Portugal. *Higher Education Policy* **13**: 181–201
- Court D 1979 The idea of social science in East Africa: An aspect of the development of higher education. *Minerva* **17**(2): 244–82
- Dill D B, Sporn B 1995 *Emerging Patterns of Social Demand and University Reform: Through a Glass Darkly*. Pergamon, Oxford
- Gibbons M 1998 *Higher Education Relevance in the 21st Century*. The World Bank, Washington
- Perkin H 2000 History of universities. *International Higher Education International* **13**: 169–204
- Neave, G, van Vught F A 1994 *Government and Higher Education Relationships Across Three Continents: The Winds of Change*. Pergamon, Oxford
- Rothblatt S 1995 An historical perspective on the university's role in social development. In: Dill D, Sporn B (eds.) *Emerging Patterns of Social Demand and University Reform: Through a Glass Darkly*. Pergamon, Oxford
- Scot P 1999 Globalization and the University. Unpublished manuscript
- Stifel L D, Davidson R K, Coleman J S 1982 *Social Sciences and Public Policy in the Developing World*. Lexington Books, Lexington, MA
- Task Force on Higher Education and Society 2000 *Higher Education in Developing Countries: Peril and Promise*. World Bank, Washington

D. Court

Policy Networks

Although the concept of policy network has taken on a variety of different definitions since its first appearance as an analytical term in the late 1970s, gradually a consensus has appeared on its meaning. A policy network refers to a set of informal and formal interactions between a variety of usually collective

public (state) and private actors, who have different, but interdependent interests. Operating in a more or less institutionalized setting, these actors are engaged in horizontal, relatively nonhierarchical discussions and negotiations to define policy alternatives, or formulate policies, or implement them. This definition suggests that policy networks vary along two core dimensions: the structural pattern according to which public power is shared among the members and the degree and patterns of integration among the members. These two dimensions developed along distinct conceptual routes, only merging into a unified concept in the mid-1990s.

1. The Changing Context of Policy-making

This definition of policy networks signals a change in the way political decision-making has been conventionally understood. It contradicts the standard notion of a clear separation between state and society, with the state being seen as the highest center for social and political control. As Mayntz (1993) observes, social modernization has brought more importance to formal organizations in all areas of society and with their strengthened position comes an increasing fragmentation of power. Policy networks represent one response to the escalating functional differentiation of social subsystems. With the state no longer being the sole entity capable of organizing society, there is a dispersion of expertise and competence, a multiplication of channels for mediation and agreement, and the involvement of different levels of decision-making from the local to the supranational (Jouve 1995). Consequently, public policy is much less likely to be produced by a unitary state with adequate control over all the resources necessary for implementing decisions in the public interest.

Rather, public policy is more likely to result from strategic interaction among several or many policy actors, each with a particular understanding of the problem and its own individual or institutional self-interest (Scharpf 1997).

1.1 Historical Development

The sharing of political power between public and private actors became the first avenue of enquiry that led to the development of the policy network concept. In the mid-1970s, Schmitter and Lehmbruch (1977) suggested the concept of corporatism be reconsidered as a conceptual tool for understanding the particular relationships between public and private actors that had emerged for the management of macroeconomic policy in several European democracies. The revitalization of the corporatism concept coincided with

other suggested new concepts for capturing the changing patterns of power sharing between private and public actors: 'corporate pluralism' in Scandinavia, 'issue networks' in the US (Hecklo 1978), 'policy communities' in the UK (Richardson and Jordan 1979), and 'policy networks' (Katzstein 1978).

The early focus in each of these concepts was on broad, macro characteristics of the political system. As these concepts became more familiar, however, a number of scholars noted their potential relevance for understanding private-public relations in more restricted sectoral policy arenas. Drawing on a growing number of empirical studies, these scholars began to develop typologies of different structural patterns of public-private interaction at the sectoral or meso level (Atkinson and Coleman 1989, van Waarden 1992), prompting a certain spate of definitional disputes. By the mid-1990s, however, sectoral and state corporatism, pressure pluralism, clientelism, issue networks, and state-directed networks generally were accepted as core structural types.

In addition to patterns of power sharing between public and private actors, a second line of inquiry developed focusing on the degree of integration among the various actors involved in these horizontal policy-making settings. Rhodes (1990) and Rhodes and Marsh (1995) targeted the degree of integration as their key variable and outlined a typology of different types of 'policy networks' that ranged from fairly tightly integrated 'policy communities' to very loose and open 'issue networks.' The prevalence of large organizations in horizontal arrangements in policy-making also triggered the interest of political sociologists interested in patterns of communication between large organizations. Laumann and Knoke (1987) sought to identify network structures by examining how information was transmitted between policy actors, the nature of the resource transactions between actors, and the relative openness of boundaries in various policy domains. These studies employed quantitative social network analysis developed in sociology. By the mid-1990s, this quantitative approach had incorporated insights from other policy network writings (Pappi and Henning 1998).

1.2 Current Views on the Concept

When speaking of the degree of integration of the policy network, it is usual to employ a concept that refers to the aggregate of actors involved. The term 'policy community' is utilized most commonly for these purposes, but terms like 'policy domain' or 'policy subsystem' are also used. Research has established a number of factors that are related to integration:

(a) Openness of the policy community. The more open deliberations in the policy community tend to be, the less integrated it is likely to become.

(b) Size of the policy community. The larger the number of members of the policy community, the greater the obstacles to high integration.

(c) Frequency of interactions among community members and the degree to which these interactions become institutionalized as opposed to remaining *ad hoc* and fluid. Institutionalization normally enhances the degree of integration.

(d) Membership stability. The more changeable the membership of a policy community, the less integrated it will be.

(e) Density and structure of information exchange. The denser the ties for information exchange and the more direct relationships are, the more integrated the policy community.

Another, particularly important, factor affecting the level of integration of a policy community is the degree to which its members share basic beliefs, values, and norms. Some actors within a given policy community may come to share a set of deep normative core beliefs and from these develop basic strategies and policy positions (Sabatier and Jenkins-Smith 1993). These 'advocacy coalitions' create 'spaces of meaning' within the policy community that provide a framework for political action. The degree of integration of the policy community will be affected by how many distinct spaces of meaning it holds. Sabatier and Jenkins-Smith suggest that policy communities will normally contain two to four advocacy coalitions. They add that the number will likely be higher when the policy community is newly formed, and may dwindle to one dominant coalition as policy-making becomes routinized. The concept of an epistemic community developed in international relations has certain properties in common with Sabatier and Jenkins-Smith's advocacy coalition (see *Policy Knowledge: Epistemic Communities*). All or part of an epistemic community may act as an advocacy coalition. When this development occurs, these epistemic communities often provide important links between policy communities, whether in other nation-states or at the transnational level.

Different types of policy networks may be distinguished depending on whether state actors share political power and on the relative resource balance between private and public actors. State actors may share political power by inviting private actors to participate directly in the formulation or implementation of policy (see *Authority: Delegation*). Here private actors become policy participants. Corporatist and clientelist networks have this feature. Or, state actors may monopolize policy formulation and policy implementation, while listening to, or receiving submissions from, private actors. In this relationship, private actors perform as policy advocates. Pressure pluralism is the most common network with this feature, but it also occurs in issue networks and state-directed networks. The decision by state actors to share political power affects the overall balance of

resources between public and private actors in the policy network. Resources may be relatively balanced, they may favor the state actors, or they may advantage civil society actors. Both pressure pluralist and sectoral-corporatist networks feature a balanced relationship. Private actors have the stronger position in clientelist and issue networks while public actors dominate in the state-directed and state-corporatist types.

2. Theoretical Status of the Concept

Although definitional disputes surrounding the concept have gradually been resolved, the overall theoretical status of policy networks remains subject to debate. Two fairly fundamental criticisms have been directed at the concept. First, Dowding (1995) has charged that the concept is nothing more than a metaphor and will thus fail to produce theories of the policy process. Second, the concept is said to have limited relevance because of an apparent lack of suitability for explaining policy innovations and policy change.

2.1 Policy Networks as Metaphors Only

Dowding has argued that the attempts to provide a middle range theory of policy-making utilizing the policy network concept are bound to fail. He suggests that the driving force of the explanation, the independent variables, are not characteristics of the network as such, but of the components within the networks. Variables such as the relative openness of a policy network are cast as a mere redescription of policy phenomena. Related criticisms argue that policy network studies have failed to explain both the mutual dependence of public and private actors and the conditions under which particular policy networks arise (Thatcher 1998). Critics also lament the lack of specificity of the hypotheses that do exist, a failing bound to render any empirical testing rather unconvincing.

Although useful for spurring discussion and reflection, there are several reasons why these criticisms may be premature. First, the criticisms undervalue the descriptive utility of the policy network concept. One of the most difficult empirical problems faced in the analysis of public policies is determining specifically who the consequential actors might be in any policy domain. Although one might be able to begin to specify these actors by utilizing a good understanding of formal institutions and of the persons occupying key positions, such a specification is almost inevitably bound to be incomplete. Policy-making is replete with informal relationships and the operation of subtle, somewhat opaque norms that empower certain actors in unexpected ways. Accordingly, the methodologies developed for delimiting a policy network have a

broader use in policy analysis. By identifying the central actors in policy-making, they assist the analysis of the processes by which policy ideas are translated into particular policy proposals through interchanges between an identified set of actors. This identification process becomes particularly helpful in policy areas featuring multilevel governance as Peterson (1995) has argued in referring to the European Union. This argument also has relevance to such patterns of governance as they emerge in policy areas where globalization has a decided impact.

Second, numerous empirical studies subsequent to the publication of Dowding's critique have provided convincing evidence of relationships between variations in policy networks and policy outcomes. Josselin's (1997) study of the negotiation of core directives for a single market in financial services in the EU evaluates the extent to which different network types facilitate or hinder the promotion of member-states' interests in EU-level negotiation settings. She shows that a relatively closed corporatist network at the national level is a decided asset in promoting a unified position at the EU level. Conversely, however, a more fluid issue network makes such unity more difficult, while providing additional degrees of freedom in the fluid decision-making arenas that exist at the supranational level. Similarly, Daugbjerg (1998) relates the relatively closed character of corporatist networks to patterns of political change. The more closed and cohesive the network, the more incremental the pattern of change and the more stable the utilization of policy instruments. Coleman's (1996) study of the impact of globalization on the regulation of financial services at the domestic level traces relationships between particular types of policy networks and the varying adaptation of countries to these globalizing pressures.

Finally, Dowding's critique is addressed by some scholars who are using more formal models. Pappi and Henning (1998) elaborate a complex model that suggests direct relationships between network types (pluralism, corporatism, clientelism, state-direction) and policy outcomes. Scharpf's 'actor-centered institutionalism' introduces network-based negotiation settings as one of several core types of forums for policy-making. Drawing on game theory, he sets out a series of propositions that provide a basis for modeling negotiations between actors in such networks and deriving likely differences in policy outputs based on general welfare and distributive justice criteria. Both of these theoretical approaches would appear to hold considerable promise for the development of middle level theory where policy networks will be an important independent variable.

2.2 Policy Networks and Policy Change

Policy networks have been conceptualized primarily using structural variables. Such structural character-

istics are helpful for building descriptions of the policy process at a particular point in time and, more recently, for positing testable explanations of variations in policy outcomes. Dowding (1995) questions, however, whether these same variables are sufficient for explaining how and why policies change and why networks themselves change. Two avenues of investigation have been pursued for addressing these criticisms. The first focuses on the relationship between policy ideas and policy networks and the second draws on an analysis of boundary rules.

Sabatier and Jenkins-Smith (1993) argue that policy networks include one or more advocacy coalitions that are composed of actors who share deep core beliefs, near core policy beliefs, and secondary beliefs about the use of policy instruments. Changes in the relative dominance of advocacy coalitions in a policy network may prompt policy change. Or members of coalitions may negotiate adjustment of secondary beliefs that, in turn, open the door to policy change. Paralleling Sabatier and Jenkins-Smith's work, Jobert (1995) argues that a 'political forum' and a 'scientific or expert' forum exist alongside a policy network. When politicians begin to experience problems that cannot be addressed well within a policy network, they begin to reach to these other fora for advice. To the extent that this advice is acted upon, the policy network may be dissolved, with new arrangements taking its place.

What both these theoretical arguments stress is the importance of looking at the interaction between policy ideas and who shares which ideas within a given policy network. Coleman et al. (1996) follow up on these insights by arguing that the degree of sharing of policy ideas in a given policy network will determine the approach to policy negotiations among actors in the network. They hypothesize that the more ideas are shared, the more policy discussions will take on aspects of what Scharpf (1997) has called 'positive coordination' or a 'problem solving' decision-making style. They suggest that positive coordination is more likely in closed networks like corporatism than in more open networks such as pressure pluralism. By coupling policy network structures to policy ideas, they then trace out different patterns for how policy change occurs. Scharpf's work stresses the importance of looking at actor preferences and cognitive orientations in studying policy processes. The richness of his theoretical work offers further opportunities for improving theories of policy change.

The second avenue for improving explanations of policy change focuses on different types of boundaries within policy communities. In their retrospective on policy network analysis in the UK, Rhodes and Marsh (1995) note that a policy community can have several policy networks. These, in turn, may operate at two levels; a central level whose members dispose of considerable resources, and a peripheral one composed of those with few or no resources. Changes in

the actors at the center of a policy community and thus in the dominant policy network, or movements from the peripheral networks to the dominant one, will be crucial intervening variables in explanations of policy change. Similarly, new actors may build peripheral networks over time and then come to contest the exclusivity of the existing dominant policy network. To the extent that new actors enter an existing policy network, they may bring about a change in its structure. For example, entry of new actors into a previously closed corporatist network may foster sufficient dissent that the state refuses to involve private actors any longer in policy making. Once such actors are excluded, the network takes on pressure pluralist characteristics, altering, in turn, the policy process and perhaps policy outcomes (Coleman et al. 1996).

3. Methodological Issues

The empirical analysis of policy networks has centered on two variables: the degree of integration of policy communities and the division of power between state and civil society actors. Scholars have approached these empirical issues from a variety of methodological standpoints. Among these, the use of quantitative network analysis, elite interviews, and close examination of documents and archival materials are the most commonly used. The use of these methods varies depending on the particular empirical task in question.

3.1 What are the Boundary Rules of the Policy Community? Who Belongs?

Determining the boundaries of a policy network is one of the more difficult, yet more important empirical tasks in this approach. Understanding who belongs and who does not provides key information about likely political demands and patterns of political influence, factors, in turn, important to explaining policy outcomes. In addition, as noted above, observations of changes in policy network membership may be crucial to approaching the explanation of political change. Accordingly, boundary questions must be approached diachronically, with Sabatier and Jenkins-Smith (1993) suggesting a time perspective 'of a decade or more.'

Those working in the public policy field have tended to approach boundary specification by looking at relationships and then at activities and events. In the relational approach to boundary specification, they seek to identify the actors involved in the formulation and implementation of policies in selected policy domains. With these relationships in focus, they then examine variations in the organizational attributes, political ideas, and degree of participation of these actors in selected activities and events. A mix of

quantitative network analysis and qualitative techniques are used in these studies. The relationships of concern are usually determined using 'snowball-sampling' techniques to set up the right set of elite interviews. This methodology does have its weaknesses in that reliance on elite interviews is costly both in time and in money and it is difficult to replicate these interviews over time to get a dynamic perspective on boundary change.

3.2 Does the State Share Power Directly With Policy Community Members? What is the Balance in the Distribution of Resources in the Policy Community?

The answers to these questions provide the empirical base for determining the type of policy network in place in a given policy community. In beginning, to assess the nature of the policy network, it is essential to understand well the organizational attributes of the actors in the policy community. One of the crucial empirical legacies of the corporatist literature were some empirical guidelines for assessing such policy capacity.

The policy capacity of interest associations and of associational systems may be evaluated by looking successively at three types of phenomena. First, it is useful to assess whether associations in a policy network compete with one another for members and are part of a more comprehensive associative order. The development of an associative order is evaluated by examining the degree of horizontal and vertical integration among associations. Second, some evaluation of the internal resources of the associations is useful for assessing likely policy capacity. Frequently, however, these kinds of analysis are still insufficient, particularly for determining the degree to which associations participate in policy-making. Normally, before final determination of the structure of a policy network, it is useful again to turn to elite interviews for additional information on the extent to which civil society actors participate directly in policy formulation or implementation.

4. Future Challenges

The concept of a policy network was designed in the first instance for the analysis of policy making at the national and more local levels. As globalization has increased the internationalization of many policy areas, however, the likelihood rises that several policy networks will be interacting with one another across different levels of decision-making. Future research, therefore, will need to focus on improving understanding of the relationships between policy networks.

Some of these networks will have the additional property of being transnational in character, that is, they will include actors from both the nation-state and the international levels.

The concept of policy community mediators offers one avenue for research on the relationships between policy networks (Coleman and Perl 1999). In many internationalized policy domains, boundary rules will create overlapping memberships between policy networks, with mediators emerging from among those actors with joint memberships. Mediation can take place at two levels. First, a mediator may operate as a pragmatic policy broker who looks for opportunities to lessen conflict within and between policy networks in order to move the policy process forward. Second, perhaps even more important, mediation takes place on an intellectual plane, involving the 'translating' of policy ideas and policy paradigms between policy networks. Drawing on Kuhn's work on scientific paradigms, Schön and Rein (1994) suggest that translation is crucial to resolving conflicts between different policy paradigms or frames. Mediators construct a new view of the world that has a place for the policy concerns of both networks and that provides a language and set of norms that enable them to speak to one another.

Success in this second role requires that the mediator be accepted as legitimate in both policy communities. Legitimacy will derive from having both adequate expertise and a socialization experience that permits straddling the boundaries of the communities concerned (Jouve 1995). Research on linkages between policy networks and on the role of policy mediators will be crucial to further theorizing on policy networks in this globalizing era.

See also: Network Analysis; Networks: Social

Bibliography

- Atkinson M M, Coleman W D 1989 Strong states and weak states: sectoral policy networks in advanced capitalist economies. *British Journal of Political Science* **19**: 47–67
- Coleman W D 1996 *Financial Services, Globalization and Domestic Policy Change*. St. Martin's Press, New York
- Coleman W D, Perl A 1999 Internationalized policy environments and policy network analysis. *Political Studies–London* **47**: 691–709
- Coleman W D, Skogstad G, Atkinson M 1996 Paradigm shifts and policy networks: cumulative change in agriculture. *Journal of Public Policy* **16**: 273–301
- Daugbjerg C 1998 Linking policy networks and environmental policies: nitrate policy making in Denmark and Sweden 1970–1995. *Public Administration* **76**: 275–94
- Dowding K 1995 Model or metaphor? A critical review of the policy network approach. *Political Studies–London* **43**: 136–58
- Hecl H 1978 Issue networks and the executive establishment. In: Beer S H, King A (eds.) *The New American Political*

- System. American Enterprise Institute for Public Policy Research, Washington, DC
- Jobert B 1995 Rhétorique politique, controverses scientifiques et construction des normes institutionnelles: esquisse d'un parcours de recherche. In: Faure A, Pollet G, Warin P (eds.) *La Construction du Sens dans Les Politiques Publiques: Débats Autour de la Notion de Référentiel* [The Construction of Meaning in Public Policies: Debates surrounding the concept of the *Référentiel*]. L'Harmattan, Paris
- Josselin D 1997 *Money Politics in the New Europe: Britain, France and the Single Financial Market*. Macmillan, Basingstoke, UK
- Jouve B 1995 Réseaux et communautés de politique publique en action. In: Le Galès P, Thatcher M (eds.) *Les Réseaux de Politique Publique. Débat autour des Policy Networks* [Public Policy Networks: Debates on Policy Networks]. L'Harmattan, Paris
- Katzenstein P 1978 Conclusion: domestic structures and strategies of foreign economic policy. In: Katzenstein P J (ed.) *Between Power and Plenty: Foreign Economic Policies of Advanced Industrial States*. University of Wisconsin Press, Madison, WI
- Laumann E O, Knoke D 1987 *The Organizational State*. University of Wisconsin Press, Madison, WI
- Mayntz R 1993 Policy-netzwerke und die Logik von Verhandlungssystemen [Policy networks and the logic of negotiation systems]. In: Héritier A (ed.) *Policy-Analyse: Politische Vierteljahresschrift Sonderheft 24*. Westdeutscher Verlag, Opladen, Germany
- Pappi F U, Henning C H 1998 Policy networks: more than a metaphor? *Journal of Theoretical Politics* 10: 553–75
- Peterson J 1995 Decision-making in the European Union: towards a framework for analysis. *Journal of European Public Policy* 2: 69–93
- Rhodes R 1990 Policy networks: a British perspective. *Journal of Theoretical Politics* 2: 293–317
- Rhodes R, Marsh D 1995 Les réseaux d'action publique en Grande Bretagne. In: Le Galès P, Thatcher M (eds.) *Les Réseaux de Politique Publique. Débat autour des Policy Networks* [Public Policy Networks: Debates on Policy Networks]. L'Harmattan, Paris
- Richardson J J, Jordan A G 1979 *Governing Under Pressure: The Policy Process in a Post-Parliamentary Democracy*. Robertson, Oxford, UK
- Sabatier P A, Jenkins-Smith H C 1993 Policy change over a decade or more. In: Sabatier P A, Jenkins-Smith H C (eds.) *Policy Change and Learning: An Advocacy Coalition Approach*. Westview Press, Boulder, CO
- Scharpf F W 1997 *Games Real Actors Play: Actor-centered Institutionalism in Policy Research*. Westview Press, Boulder, CO
- Schmitter P C, Lehmbruch G 1977 *Trends Toward Corporatist Intermediation*. Sage Publications, London
- Schön D A, Rein M 1994 *Frame Reflection: Toward the Resolution of Intractable Policy Controversies*. Basic Books, New York
- Thatcher M 1998 The development of policy network analyses: from modest origins to overarching frameworks. *Journal of Theoretical Politics* 10: 389–416
- van Waarden F 1992 Dimensions and types of policy networks. *European Journal of Political Research* 21: 29–52

W. D. Coleman

Policy Process: Business Participation

While business firms play an important role in policy process in all capitalist societies the extent, nature, and impact of this participation varies both over time and among countries. Business political participation has changed significantly in the USA since the 1970s, and to a lesser extent in Europe. However the degree of political power exercised by business remains an important point of contention among scholars.

1. Business Political Participation in Comparative Perspective

A critical factor underlying differences in national patterns of business–government relations is the extent and scope of government intervention in the economy (Shonfield 1965). The greater a government's impact on the direction of national economic development—measured by spending, government ownership of industry, and control over capital allocation—the closer are the political ties between business and government. In capitalist nations characterized by extensive government intervention, the government will often promote the formation of business associations in order to facilitate policy coordination both with and within the private sector. The way business is organized politically also tends to mirror the structure of the public sector: the more governmental decision-making is centralized, the greater the breadth of business political associations.

On both dimensions, the USA is distinctive: government economic intervention has historically been relatively limited and public policy making is relatively fragmented. Accordingly, there is no American counterpart to 'peak' associations such as the German Federation of Industry, Confindustria in Italy, the Comité National du Patronat in France, the Keidanren in Japan, or the Confederation of British Industry—organizations which both represent and are able to speak for business as a whole (G. Wilson 1981, 1985).

In many capitalist nations, trade associations negotiate with the government on behalf of all firms in a particular sector. Such arrangements are often characterized as corporatist (Schmitter and Lehmbruch 1979). By contrast, trade associations in the USA rarely enjoy official or quasiofficial status and they tend to be much weaker and less inclusive than their counterparts in other capitalist nations. More generally, while business firms have become much more politically active in the USA since the 1970s, business is still less politically organized than in other capitalist nations.

The nature of the interaction between business organizations and government officials also differs in

the USA. Firms in the USA typically influence public policy through 'lobbying,' a term which connotes an arms-length relationship between business and government. Corporations typically need to engage in pressure group politics to gain access to policy makers and shape the policy process. But outside the USA, business and political elites are more likely to come from similar social backgrounds and there is considerable movement of personnel between the two institutions. There are numerous opportunities for both formal and informal interaction between business and government officials and firms can expect to be closely consulted about policies which affect their interests. There is less need to 'lobby' because business organizations are themselves either a quasiofficial or official part of the public policy process.

2. Changes in Business Political Participation in Europe

However, business political participation has changed in Europe. In many capitalist societies a growing portion of firms are subsidiaries of multinational corporations whose headquarters are outside national borders. Such firms are less likely both to identify their interests with those of the nation in which their business is located and to lack the kind of informal access to policy makers enjoyed by domestic firms.

In addition, the role of government in managing the economy has declined in almost all capitalist polities: the importance of national industrial policies which required close business-government cooperation has diminished while a substantial number of government-owned firms have been privatized, thus creating a more independent and pluralist business community. Both of these developments have diminished the historically close ties of business and government in a number of European countries and made the structure of business political representation more fragmented.

Within Western Europe, the increasing importance of the European Union has had a significant impact on business-government relations (Mazey and Richardson 1993, Wallace and Young 1997). The Single European Act (1987) not only significantly expanded the regulatory role of the EU, but ended the unanimity rule of the Council of Ministers. The latter constitutional change meant that each member state was no longer able to exercise a veto power over EU legislation. Accordingly, firms and trade associations could no longer rely on their influence with their national governments to protect their interests in Brussels.

The result has been a substantial increase in business political representation in Brussels. Many European-wide trade associations have been established to represent business interests in Brussels. At the same

time, many large multinational firms have established their own independent lobbying operations in Brussels. The interests of 80 American firms with subsidiaries in Europe are represented by the EC Committee of the American Chamber of Commerce. In addition, many national trade associations have shifted part of their political efforts to Brussels.

In some cases, the more stable corporatist arrangements which characterized business-government at the national level have been undermined and firms and industries no longer enjoy the privileged access they had when policy making took place at the national level. The result has been an EU policy process which often resembles the American pattern of 'disjointed pluralism' and fluid issue networks (Streeck and Schmitter 1991). This, however, is not true of all policy areas, many of which continue to be characterized by stable policy communities which usually privilege producer interests.

2.1 Contemporary Changes in Business Political Participation in the USA

The degree and forms of business participation have significantly changed in the USA since the 1970s. During the 1950s and 1960s, business devoted relatively few resources to influencing public policy in Washington. Few large firms had Washington offices and these were typically thinly staffed. An exhaustive study of business lobbying on foreign trade policies during the 1950s found most business lobbies to be poorly financed, ill-managed, and at best only marginally effective. Indeed, only a minority of either large- or medium-sized firms regularly communicated with Congress (Bauer et al. 1993).

Certainly, many firms and industries received considerable benefits from government. For example, defense contractors benefited from procurement policies, agricultural interests received substantial price support, and the oil and textile industries, and also much of American agriculture, were protected from import competition. Many industries, such as housing and energy, received important tax preferences and a substantial portion of the American economy was regulated in ways that limited price competition and stabilized markets. However, because business enjoyed considerable public support, and nonbusiness constituencies were poorly organized, business did not need to engage in extensive political activity to enjoy a wide array of economic privileges.

2.2 The Changing Political Environment of Business

The political environment of business began to change in the late 1960s. The social activism of the latter part of that decade reflected and contributed to a decline of

public confidence in all institutions, including the large corporation. It also spawned a new political force: a loosely allied network of consumer and environmental organizations who began to pressure for expanding government regulation of business. Led by skillful political entrepreneurs such as Ralph Nader, and enjoying extensive and for the most part favorable press coverage, the public interest movement had a major policy impact (Berry 1997). From the late-1960s through the mid-1970s business experienced a series of major political defeats in Congress. New laws were passed which expanded the regulatory role of the federal government in environmental protection, consumer protection, and occupational health and safety.

Unlike previous patterns of government intervention, which had primarily affected only specific industries, the impact of the new social regulation was much broader: now, for the first time, virtually every firm faced a large and complex array of federal rules that mandated or prohibited a wide range of business decisions. The additional costs of complying with these regulations also became more burdensome following the economic downturn of the mid-1970s, which also led to a substantial increase in public hostility toward particular industries, most notably the oil industry.

2.3 Implications for Business Political Strategy

Business-government relations had previously been dominated by 'insider politics': corporations had secured benefits primarily by making low-visibility appeals to key legislators and administrators. But social regulation was dominated by 'outsider politics': it was often driven by highly visible, often emotional appeals to both public and elite opinion. To counter effectively the appeals of activists and their supporters, companies had to adopt to the 'outsider' strategies of their political opponents. Moreover, within Congress power became more decentralized and party discipline weakened. This meant that business could no longer rely on a handful of influential politicians or party leaders: they had to learn how to influence the votes of 535 individual legislators.

This decade also witnessed a substantial change in the regulations governing campaign contributions by business. Since the beginning of the twentieth century, corporations had been forbidden to contribute funds to candidates for federal office. However, the investigations surrounding the Watergate break-in revealed that a large number of firms had responded to pressures from the Finance Committee to re-elect the President by making illegal contributions to President Nixon's re-election campaign. Ironically, Congress 'cleaned up' federal campaign finance by providing a legal vehicle for corporations to solicit funds from their executive employees and shareholders. The 1974

Federal Campaign Act Finance Amendments permitted corporations, and also other organizations, to establish political action committees (PACs). Corporations were allowed to pay the costs of establishing and administering PACs, although they still could not contribute to them

2.4 New Forms of Business Political Participation

As a response to these developments, business political activity both increased and became more visible (G. Wilson 1981). Between 1968 and 1978, the number of corporations with public affairs offices in Washington grew from 100 to more than 500, while many other firms employed full- or part-time lobbyists. All told, by the early 1980s nearly 2,500 business firms had some form of political representation in Washington (Vogel 1989, J. Q. Wilson 1996).

The number of trade associations also grew substantially, and many moved their headquarters to Washington in order to more effectively monitor and influence national policy. Some, such as the Chemical Manufacturers Association, the Semiconductor Industry Association, and the Health Insurance Association of America, proved highly effective. But the typical industry trade association declined in importance. This was largely the result of structural changes in the American economy which made firms more diversified. With investments in a wide variety of different industries, their interests could no longer be effectively represented by an industry-based interest group. Accordingly, many large corporations developed an independent capacity to monitor political developments and to advance their political interests (Ryan et al. 1987, Harris 1989).

In order to advance its collective interests, the business community engaged in coalition building—the identification and mobilization of large numbers of companies and trade associations who shared a common interest in a particular issue. This became a highly effective strategy to maximize the political influence of substantial segments of the business community, since it enabled many firms and trade associations to pool and coordinate their political resources. Such coalitions were often difficult to establish since many firms were reluctant to commit scarce political resources to lobby on issues in which their stake was either modest or not readily apparent. But those that were establishing were often politically effective.

Many coalitions were *ad hoc* in nature: they were formed to lobby on a particular issue—and then disbanded when the issue was resolved or disappeared from the political agenda. However, some business coalitions became institutionalized (Levitan and Cooper 1984). One of the most important new 'peak' business organizations was the Business Roundtable.

It was established during the early 1970s from a merger of three *ad hoc* coalitions which had been primarily formed to lobby on issues affecting labor. Exclusively composed of the chief executive officers of approximately 200 of the nation's largest corporations, it relied on the 'unique mobilization of the talent and prestige of the nation's top corporate heads' to lobby government officials (Vogel 1989, p. 198).

Other more established business organizations representing large numbers of firms and industries also grew in importance. Of particular significance was an increase in coordinated political activity on the part of smaller firms. Between 1970 and 1979, the membership of the National Federation of Independent Business (NFIB), which represented smaller firms, grew from 300 to 600,000. The NFIB frequently took positions opposed to those of the Business Roundtable, which many small firms viewed as too willing to accommodate itself to liberal political pressures for government expansion.

Another dimension of business political participation focused on mobilizing the 'grass-roots.' By the late 1970s, corporations and trade associations were spending an estimated \$850 million to help their supporters throughout the USA to communicate with, and thus influence the votes of, their representatives and Senators. Much of what Washington lobbyists now did was to identify strategies for influencing particular legislators, rather than lobby them directly. Business associations representing large numbers of smaller firms such as the National Association of Independent Businessmen and the Chamber of Commerce made particularly extensive use of grass-roots lobbying.

2.5 The Politics of Ideas

An important component of business participation in the policy process focused on changing public attitudes toward business in general and specific policy issues in particular (Blumenthal 1986). Corporations and trade associations made considerable use of 'advocacy advertising,' a specialized form of institutional or corporate advertising whose purpose is to influence the public's attitudes on issues of public policy. During the 1970s, more than 35 corporations and trade associations launched public advocacy campaigns and by the end of that decade, major corporate advertisers were spending about one-third of their advertising dollars on campaigns aimed at the public in their role as citizens rather than consumers.

The oil industry made extensive use of this advertising strategy during the 1970s to explain its position on antitrust policy, government regulation and industry profitability—all issues then much in the public spotlight as a result of the energy crisis. Steel firms purchased advertising space to explain their

positions on import restrictions and pollution controls, the American Electric Power Company ran a multimillion dollar campaign opposing the use of scrubbers and the Caloric Control Council employed advocacy advertising to build a successful grass-roots campaign to enact legislation overruling the Food and Drug Administration's ban on saccharin. During the 1990s, the Health Insurance Association of America waged a highly effective television campaign against the Clinton Administration's proposed health care plan.

The business community and its supporters also sought to influence the terms of public debate by funding research on public policy. Financial support from corporations and conservative foundations played a critical role in creating and expanding private research centers that issued and sponsored books, studies and reports that challenged liberal and left of center views on public policy. In contrast to advocacy advertising, these efforts were aimed at influencing the views of elites such as politicians, government regulators, academics, and journalists.

The Heritage Foundation, a conservative policy research institute, employed several dozen policy analysts who worked closely with the staffs of sympathetic members of Congress, tracking the progress of particular policy proposals and preparing detailed positions papers on them. The American Enterprise Institute, which became a kind of right-of-center counterpart to the more liberal and older Brookings Institution, released scores of studies, held numerous conferences, and prepared several analyses of legislative proposals—in addition to publishing seven journals and newsletters—primarily focusing on government regulation and tax policy. This research helped contribute to a more skeptical attitude about the role of government and a more positive attitude about the role of markets—first among elites and then among the public as a whole. More liberal research institutes also became more sympathetic to business perspectives.

3. The Debate Over Business Power

During the 1950s and 1960s, the scholarly consensus within both political science and sociology regarded business as an interest group which did not pose any distinctive problems for democratic governance (Truman 1951, Epstein 1969). There were dissenters, but they were a minority voice (Bachrach 1967). During the 1970s and 1980s the intellectual center of gravity of both disciplines began to shift: a considerable body of scholarship now judged 'the large corporation's concentration of resources and wealth (to be) an anomaly that upsets the balance between democracy and capitalism' (Menninger 1985, p. 210).

According to the critics of pluralism, business political influence has two dimensions, one structural

and the other instrumental. Structural power refers to the ability of business to shape both the terms of political debate and political outcomes by virtue of its privileged position within a capitalist system. This privileged position in turn has two roots. One is the central role of business in shaping the rate and direction of economic development. According to this analysis, if public policy does not furnish firms with sufficient inducements to invest, or alternatively if it imposes too many restrictions on their ability to accumulate capital, firms will reduce their rate of investment, thus resulting in an economic downturn. Consequently, politicians, knowing that their reelection will be adversely affected by a slowdown in economic growth, will be reluctant to support policies that might antagonize corporate investors.

Since a critical dimension of political power involves control over the political agenda and the ability to define the terms of public debate, the privileged political position of business in capitalist societies also has an ideological dimension. According to this analysis, a capitalist society is characterized by a prevailing ideological consensus that places a wide array of issues and policy alternatives off the political agenda, such as the autonomy of corporate managers, the prevailing distribution of wealth, and private ownership of the means of production—all of which protect business privileges.

While the structural dimension of business power is often exercised in subtle ways, the second, instrumental dimension of business power is more visible. It is reflected in the disproportionate resources that business has available to shape the resolution of issues that actually do come before the political process—resources which give business political advantages enjoyed by no other interest group, or the public as a whole. This dimension of business power became substantially more visible in the 1970s and 1980s as firms devoted increased resources in order to both shift the terms of public debate and affect political outcomes.

3.1 Is Business Privileged?

The claim that business enjoys a privileged political position is closely associated with Charles Lindblom, whose book *Politics and Markets* represents one of the most important and influential analyses of business political power published since the 1970s (Lindblom 1997). As he argues, capitalism in the USA enjoys considerable legitimacy: no political movement in the USA has seriously questioned private ownership or challenged the central role of markets in distributing wealth and privilege. However, while business certainly benefits from this characteristic of American society, it is not responsible for it. The hegemony of the 'liberal tradition in America' is deeply rooted in popular American culture—a culture that has long

valued individual self-reliance, limited government, and private property—and which predates the rise of the large corporation (Hartz 1955, McClosky and Zaller 1984).

At the same time, American culture also contains a strong populist dimension, which, although not anti-capitalist, has periodically been expressed in considerable public hostility to the large corporation. During the twentieth century, the USA experienced four reform movements—Populism, Progressivism, The New Deal, and the public interest movement, each of which has effectively challenged a wide variety of corporate prerogatives. Moreover, in other capitalist nations, anticapitalist movements and political parties on both the left and the right have received considerable public support, which suggests that capitalism does not enjoy equal legitimacy or hegemony in all capitalist societies.

Scholars have challenged Lindblom's thesis with reference to the USA by citing the ability of non-business constituencies to challenge effectively a wide array of corporate prerogatives during the 1960s and 1970s (Vogel 1987, G. Wilson 1981, J. Q. Wilson 1996). The effectiveness of the challenges to corporate privileges that have surfaced periodically in the USA implies that corporations, like any group of interests, must engage in interest-group politics in order to defend and advance their interests. Sometimes their efforts will be successful and other times they will not. But this does not distinguish them from any other group of interests.

Vogel claims that the power of business has varied considerably over time: it was relatively substantial in the 1950s through the late 1960s, declined through the late 1970s, and then subsequently experienced a resurgence (Vogel 1989). These fluctuations in business political influence are primarily attributed to shifting public perceptions of the performance of the economy: business tends to be more influential when the economy appears to be performing relatively poorly and is less influential when it is relatively prosperous. While this analysis does help account for many of the changes in political influence from the early 1960s through the late 1980s, it cannot account for the 1990s—when economic performance was extremely strong and yet business was able to successfully resist the efforts of the Clinton Administration to increase public social spending and expand medical coverage.

Plotke (1992) argues that the slowdown in growth and the decline in profits during the 1970s was not by itself sufficient to mobilize business politically. What was also required was an alternative vision of political economic relations which both attributed the nation's economic difficulties to the expansion of government intervention and proposed an alternative model of economic growth and development. This ability to redefine the political agenda provided a referent for business interest formation that in turn made possible the mobilization of large numbers of disparate firms

and industries. An important contribution of this perspective is to emphasize the contingent nature of business political preferences, as well as the extent to which business influence both affects and is affected by the terms of public debate.

Mucciaroni (1995) analyzed the factors that affect the ability of particular sectors or firms to maintain the benefits they receive from government. Focusing on developments in four policy areas, namely trade policy, agricultural subsidies, tax expenditures, and economic regulation, over two decades, he concluded that producer-group fortunes have varied considerably both over time and among issue areas. In three of the four policy areas, producers experienced reduced benefits due to changes in both public policy norms and the structure of political decision making.

3.2 Corporate Campaign Contributions

The number of business-sponsored political action committees grew substantially in the 1970s and 1980s. While there were no corporate PACs in 1971, there were more than 1,700 in 1992 and they gave a total of \$64 million to congressional candidates. Thanks to the wealth of data available due to federal reporting requirements, few dimensions of business political participation—and power—have been studied as extensively.

One focus of this research has examined the effect of campaign contributions on the voting of legislators, from which different conclusions have been drawn. A series of studies have found that business contributions affected roll-call votes on minimum wage legislation, the debt limit, the windfall profits tax, wage and price controls, trucking deregulation, and legislation affecting auto dealers (Conway 1991). However, a comprehensive study of a number of PACs and congressional voting behavior over an 8-year period found contributions to be rarely related to congressional voting patterns.

According to Wilson, ‘a fair summary of the existing studies (reveals) there is no consistent and systematic evidence showing that PAC contributions, independent of other factors (such as constituency pressures, legislator ideology, and party affiliation), have a significant impact on voting in Congress. Some studies find some effects in some areas, other find no effect in other areas’ (J. Q. Wilson 1996, p. 430). These findings suggest that corporations often make contributions as much to gain access as to influence votes. However, other studies reach a very different conclusion, arguing that campaign contributions do represent a critical source of business power and privilege (Clawson et al. 1992, Ferguson 1995).

The extent to which firms share similar policy preferences, and are able to cooperate to achieve them, has been an important focus on scholarly debate over the political significance of business campaign spend-

ing as well as over business political influence in general. Business campaign contributions are often coordinated. For example, in the 1984 election, business contributions went overwhelming to only one candidate in almost every race, while in three out of four congressional races, large corporate PACs have given at least nine times as much to one candidate as the other (Clawson et al. 1992, p. 160).

3.3 The Extent and Significance of Business Political Unity

This coherence, in turn, reflects not only their common interests, but also a common culture and the extent of structural ties among firms, such as loans from the same banks, sales and purchases from each other, and interlocking boards of directors. ‘Business power derives in part from this ability to achieve political unity on most issues’ (Clawson et al. 1992, p. 160). Another study found economic leverage between industries to be associated with similar patterns of donations (Mizuichi and Koenig 1986). Moreover, the American business community has numerous mechanisms that facilitate cooperation and coordination, especially under the leadership of large firms (Useem 1984).

Other scholars, by contrast, emphasize that while ‘corporate class consciousness’ does exist in the USA, it only emerges sporadically under unusual circumstances: the American business ‘community’ is more often a community in name only (Vogel 1978). The structure of the American economy is highly fragmented with firms having multiple sources of capital. Moreover, the American system of corporate governance encourages managers to support public policies that directly benefit their firms—even if such policies may not be in the interest of business as a whole. While companies may not typically give to opposing candidates, they nonetheless frequently do have opposing interests.

Many important public policies, such as trade policy, tax policy, health policy, environmental policy, and economic deregulation, divide the business community. A particularly important and long-standing division is between large firms and small businesses, with the latter often frequently opposing the more liberal policy preferences of ‘big business.’ Martin’s study of the politics of human capital investment, including the bitter debate over the expansion of medical coverage in the early 1990s, reveals sharp divisions between large and small firms which precluded the enactment of reforms that arguably would have benefited business as a whole (Martin 2000).

Ironically, the politicization of business that took place during the 1970s has exacerbated the centrifugal tendencies within American business. Because firms have become more politically sophisticated and both able and willing to devote substantial resources to

influencing public policy, corporate political strategy has become more closely linked to business strategy. Firms now frequently seek to gain competitive advantages through the use of non-market strategies, which in turn increases the likelihood of policy differences among them (Baron 1995).

4. Conclusion

The latter part of the twentieth century has witnessed a substantial increase in both scholarly and popular studies of business political activity in the USA and the European Union, in large measure because this activity has become more visible. There have also been a few comparative studies of business political participation (G. Wilson 1985, 1990, Bottomore and Brym 1989). However, there remains a dearth of comparative studies of business political power and influence. Scholars who assert that business enjoys a privileged position usually assume that business is equally influential in all capitalist societies, while most empirical studies of business political influence tend to focus on only one country, or even more narrowly on specific policy areas. An important challenge of future scholarship is to compare systematically patterns of business political influence in different capitalist societies.

See also: Business History; Business Law; Community Power Structure; Corporate Law; Ethical Codes, Professional; Business Codes; Lobbying; Participation: Political; Policy History: State and Economy; Political Money and Party Finance; Regulation, Economic Theory of; Regulation: Empirical Analysis; Regulation Theory in Geography

Bibliography

- Bachrach P 1967 *The Theory of Democratic Elitism: A Critique*. Little, Brown, Boston
- Baron D P 1995 Integrating strategy—market and nonmarket components. *California Management Review* 37(2): 47–65
- Bauer R A, de Sola Pool I, Dexter L A 1993 *American Business and Public Policy: The Politics of Foreign Trade*. Atherton Press, New York
- Berry J 1997 *Lobbying for the People: The Political Behaviour of Public Interest Groups*. Princeton University Press, Princeton, NJ
- Blumenthal S 1986 *The Rise of the Counter-establishment: From Conservative Ideology to Political Power*. Times Books, New York
- Bottomore T, Bryn R J (eds.) 1989 *The Capitalist Class: An International Study*. New York University Press, New York
- Clawson D, Neustadt A, Scott D 1992 *Money Talks: Corporate PACS and Political Influence*. Basic Books, New York
- Conway M 1991 PACs in the political process. In: Cigler A J, Loomis B A (eds.) *Interest Group Politics*, 3rd edn. Congressional Quarterly Press, Washington, DC, pp. 199–216
- Epstein E M 1969 *The Corporation in American Politics*. Prentice-Hall, Englewood Cliffs, NJ
- Ferguson T 1995 *Golden Rule: The Investment Theory of Party Competition and the Logic of Money-driven Political Systems*. University of Chicago Press, Chicago
- Harris R A 1989 Politicized management: the changing face of business in American politics. In: Harris R A, Milkis S M (eds.) *Remaking American Politics*. Westview Press, Boulder, CO, pp. 261–86
- Hartz L 1955 *The Liberal Tradition in America*. Harcourt Brace, New York
- Levitan S A, Cooper M R 1984 *Business Lobbies: The Public Good & the Bottom Line*. Johns Hopkins University Press, Baltimore
- Lindblom C 1997 *Politics and Markets*. Basic Books, New York
- Martin C J 2000 *Stuck in Neutral: Business and the Politics of Human Capital Investment Policy*. Princeton University Press, Princeton, NJ
- Mazey S, Richardson J 1993 *Lobbying in the European Community*. Oxford University Press, Oxford
- McClosky H, Zaller J 1984 *The American Ethos: Public Attitudes Toward Capitalism and Democracy*. Harvard University Press, Cambridge, MA
- Menninger D 1985 Business and Politics. *Politics & Society* xviii, pp. 207–11
- Mizruchi M S, Keonig T 1986 Economy sources of corporate political consensus. *American Sociological Review* 51: 482–91
- Mucciaroni G 1995 *Reversals of Fortune: Public Policy and Private Interests*. Brookings Institution, Washington, DC
- Plotke D 1992 The political mobilization of business. In: Petraccia M R (ed.) *The Politics of Interests: Interest Groups Transformed*. Westview Press, Boulder, CO, pp. 175–98
- Ryan M H, Swanson C L, Buchholz R A 1987 *Corporate Strategy, Public Policy and the Fortune 500: How American's Major Corporations Influence Government*. Blackwell, Oxford, UK
- Schmitter P C, Lehbruch G 1979 *Trends Toward Corporatist Intermediation*. Sage, London
- Shonfield A 1965 *Modern Capitalism*. Oxford University Press, New York
- Streeck W, Schmitter P C 1991 From national corporatism to transnational pluralism: Organized interests in the single European market. *Politics and Society* 19: 133–64
- Truman D B 1951 *The Governmental Process: Political Interests and Public Opinion*. Alfred Knopf, New York
- Useem M 1984 *The Inner Circle: Large Corporations and the Rise of Business Political Activity in the U.S. and U.K.* Oxford University Press, Oxford, UK
- Vogel D 1978 Why businessmen distrust their state: The political consciousness of American corporate executives. *British Journal of Political Science* 8(1): 45–78
- Vogel D 1987 Political science and the study of corporate power: A dissent from the new conventional wisdom. *British Journal of Political Science* 17(4): 385–408
- Vogel D 1989 *Fluctuating Fortunes: The Political Power of Business in America*. Basic Books, New York
- Wallace H, Young A R (eds.) 1997 *Participation and Policy-making in the European Union*. Clarendon Press, Oxford, UK
- Wilson G K 1981 *Interest Groups in the USA*. Clarendon Press, Oxford, UK
- Wilson G K 1985, 1990 *Business and Politics: A Comparative Introduction*. Chatham House, Chatham, NJ

Wilson J Q 1996 The corporation as a political actor. In: Kaysen C (ed.) *The American Corporation Today*. Oxford University Press, Oxford, UK

D. Vogel

Politeness and Language

1. What is Politeness?

If, as many have claimed, language is the trait that most radically distinguishes *homo sapiens* from other species, politeness is the feature of language use that most clearly reveals the nature of human sociality as expressed in speech. Politeness is essentially a matter of taking into account the feelings of others as to how they should be interactionally treated, including behaving in a manner that demonstrates appropriate concern for interactors' social status and their social relationship. Politeness—in this broad sense of speech oriented to an interactor's public persona or 'face'—is ubiquitous in language use. Since, on the whole, taking account of people's feelings means saying and doing things in a less straightforward or more elaborate manner than when one is not taking such feelings into consideration, ways of being polite provide probably the most pervasive source of indirectness, reasons for not saying exactly what one means, in how people frame their utterances.

There are two quite different kinds of feelings to be attended to, and therefore there are two distinct kinds of politeness. One kind arises whenever what is about to be said may be unwelcome, prompting expressions of respect, restraint, avoidance ('negative politeness'). Another arises from the fact that long-term relationships with people can be important in taking their feelings into account, prompting expressions of social closeness, caring, approval ('positive politeness'). There are many folk notions for these kinds of attention to feelings—including courtesy, tact, deference, demeanor, sensibility, poise, discernment, rapport, mannerliness, urbanity, as well as for the contrasting behavior—rudeness, gaucheness, the making of social gaffes, and for its consequences—embarrassment, humiliation. Such terms label culture-specific notions invested with social importance, and attest both to the pervasiveness of notions of politeness and to their cultural framing.

2. How Can We Account for Politeness?

Since politeness is crucial to the construction and maintenance of social relationships, politeness in communication goes to the very heart of social life and

interaction; indeed it is probably a precondition for human cooperation in general. Politeness phenomena have therefore commanded interest from theorists in a wide range of social sciences. Three main classes of theoretical approach to the analysis of politeness in language can be distinguished.

(a) *Politeness as social rules*. To the layman, politeness is a concept designating 'proper' social conduct, rules for speech and behavior stemming generally from high-status individuals or groups. In literate societies such rules are often formulated in etiquette books. These 'emic' notions range from polite formulae like *please* and *thank you*, the forms of greetings and farewells, etc., to more elaborate routines for table manners, or the protocol for formal events. Politeness is conventionally attached to certain linguistic forms and formulaic expressions, which may be very different in different languages and cultures. This is how the 'person on the street' tends to think about politeness, as inhering in particular forms of words.

Some analytical approaches to politeness are formulated in terms of the same sorts of culture-specific rules for doing what is socially acceptable, for example, the work by Ide and others on Japanese politeness as social indexing or 'discernment' (see Watts et al. 1992). In these approaches, politeness is a matter of social norms, and inheres in particular linguistic forms when used appropriately as markers of pre-given social categories. This approach is most appropriate for fixed aspects of language use—the more or less obligatory social marking of relatively unalterable social categories.

(b) *Politeness as adherence to Politeness Maxims*. Another rule-based approach derives politeness as a set of social conventions coordinate with Grice's Cooperative Principle for maximally efficient information transmission ('Make your contribution such as required by the purposes of the conversation at the moment'), with its four 'Maxims' of Quality, Quantity, Relevance, and Manner (See *Pragmatics: Linguistic*). Lakoff (1973) suggested that three 'rules of rapport' underlie choice of linguistic expression, rules which can account for how speakers deviate from directly expressing meanings. Choice among these three pragmatic rules ('Don't impose,' 'Give options,' 'Be friendly') gives rise to distinct communicative styles. Leech's more detailed proposal (1983) is in the same vein. Complementary to Grice's Cooperative Principle, Leech postulates a Politeness Principle—'Minimize the expression of impolite beliefs,' with the six Maxims of Tact, Generosity, Approbation, Modesty, Agreement, Sympathy. As with Grice's Maxims, deviations from what is expected give rise to inferences. Cross-cultural differences derive from the different importance attached to particular maxims.

The conversational maxim approach shares with the social norm approach the emphasis on codified social rules for minimizing friction between inter-

actors, and the view that deviations from expected levels or forms of politeness carry a message.

(c) *Politeness as face management.* A more sociological perspective places ‘face work’ at the core of politeness. Goffman (1967) considered politeness as an aspect of interpersonal ritual, central to public order. He defined face as an individual’s publicly manifest self-esteem, and proposed that social members have two kinds of face requirements: positive face, or the want for approval from others, and negative face, or the want not to offend others. Attention to these face requirements is a matter of orientation to Goffman’s ‘diplomatic fiction of the virtual offense, or worst possible reading’ (Goffman 1971 pp.138ff.), the working assumption that face is always potentially at risk, so that any interactional act with a social-relational dimension is inherently face-threatening and needs to be modified by appropriate forms of politeness.

2.1 Universal Principles of Politeness

Brown and Levinson (1987, henceforth B&L) introduced a new perspective by drawing attention to the detailed parallels in the construction of polite utterances across widely differing languages and cultures, and argued that universal principles underlie the construction of polite utterances. The parallels they noted are of two sorts: how the polite expression of utterances is modified in relation to social characteristics of the interlocutors and the situation, and how polite utterances are linguistically constructed. At least three social factors are involved in deciding how to be polite: (a) one tends to be more polite to social superiors; (b) one tends to be more polite to people one doesn’t know. In the first case, politeness tends to go one way upwards (the superior is less polite to an inferior); in the second, politeness tends to be symmetrically exchanged. In addition, (c) in any culture there are norms and values affecting the degree of imposition or unwelcomeness of an utterance, and one tends to be more polite for more serious impositions. In language there are also parallels, the linguistic structures for realizing particular kinds of politeness show remarkable similarities across languages. The politeness of solidarity is characterized, for example, by use of intensifiers, in-group identity markers and address forms, exaggerated intonation patterns, forms for seeking agreement and avoiding disagreement. Avoidance-based politeness is characterized by self-effacement, formality, restraint, deference, with the use of honorifics, hedges, indirect speech acts, and impersonalizing mechanisms like pluralization of pronouns, nominalization, and passive.

Why then are these kinds of detailed parallels across languages and cultures to be found in the minutiae of linguistic expression in socially analogous contexts? Explanations in terms of social norms or rules can

account for politeness in a particular social group, but not the cross-cultural patterns, which seem to require a strategic account in terms of what people generally are trying to do when they are being polite. B&L (1987) proposed an abstract model of politeness wherein human actors are endowed with two essential attributes: *face* and *rationality*. Face consists of two specific kinds of wants: positive face (i.e., the desire to be approved of, admired, liked, validated), and negative face (the desire to be unimposed upon, unimpeded in one’s actions). The second ingredient in the model—rationality—provides for the ability to reason from communicative goals to linguistic means that would achieve these goals. From these two assumptions—face and rationality—and the assumption that speakers mutually know that all speakers have these attributes, B&L developed a model of how speakers construct polite utterances in different contexts on the basis of assessments of three social factors: the relative power (P) of speaker and addressee, their social distance (D), and the intrinsic ranking (R) of the face-threateningness of an imposition. P, D, and R are seen as abstract social dimensions indexing kinds of social relationship (P and D) and cultural values and definitions of impositions or threats to face (R).

B&L (1987) distinguished five general types of strategies of politeness, ranging from avoiding a face-threatening act (FTA) altogether, to carrying it out but ‘off record’ (indirectly). On-record realization of an FTA can be done without any redressive action at all (‘baldly’). It may be carried out with positive redress, which is essentially approach based, addressing the hearer’s positive face wants by emphasizing closeness and solidarity. Politeness may also be carried out with negative redress, which is essentially avoidance based, addressing negative face wants for distance, deference, freedom from unexpected impositions. Speakers are assumed to choose the linguistic framing of their utterance from this set of strategic possibilities according to the weightiness of the FTA, which is assessed with reference to the three contextually dependent social factors P, D, and R. For low levels of FTA-threat, bald-on-record, or positive politeness is most appropriate and cost effective; for higher levels, negative politeness is required; for the highest threats, indirectness is the safe option.

The claim, then, is that, however culturally variable the kinds of social relationship and kinds of face threat might be, underlying them are pan-cultural social dimensions (relative power, social distance, ranking of face threateningness) which universally go into the reckoning—and the interpretation—of strategic language choice, and hence one can derive the cross-cultural similarities in choice of linguistic realizations of politeness strategies that empirically seem to be in evidence. B&L (1987) claimed further that this model of politeness universals could be applied in particular cultural settings as an ethnographic tool for analyzing the quality of social relationships. Stable social

relationships are characterized in part by stable patterns of language use, which may distinguish particular societies or social groups.

2.2 Criticisms of the B & L Model

The goal of B&L to formulate an 'etic' set of concepts in terms of which politeness can be analyzed in 'emic' terms for any particular society was ambitious. Critics of the B&L model reveal several major points of contention about how a theory of politeness should be formulated.

2.2.1 The universality of face. Many critics have challenged. B&L's formulation (via Goffman and Radcliffe-Brown) of positive and negative face wants, as a valid way of conceptualizing the universal underpinnings of politeness. Negative face, in particular, considered as wants for freedom from imposition, appears entirely too embedded in Western individualism to sit well with conceptions of face in some other (e.g., Asian) cultures. In part, this is due to a misconstrual: the B&L face wants are abstract, they do not necessarily correspond clearly to conscious emic notions. What B&L (1987) claimed that *underlying* very diverse folk notions is a core of two interactionally relevant wants (for ratification, and freedom from imposition) which seem to be cross-culturally applicable, as desires concerning one's public self-image in the context of the moment which are *assumptions oriented to in interaction*. Other theorists (e.g., O'Driscoll 1996, Arundale 1999) have argued for notions of positive/negative face that are even more abstract, in terms of merging/individuation or closeness/separation, as the universal heart of politeness.

Challenges to the universality of the model also extend to the proposed hierarchy of increasing politeness (from bald-on-record to positive to negative to indirectness). Assessments of the P, D, R factors are situationally and culturally very variable, it is possible to accumulate different strategies in one utterance and to balance elements of negative politeness with positive politeness in one act, and indirectness is not always seen as the most polite option. These observations have led some researchers to argue against the possibility of identifying any kind of universal basis for polite behavior; politeness is simply incommensurate across societies. Those who take this extreme relativistic line can have no explanation for the observable cross-cultural parallels in patterns of language use, for how people manage (sometimes) to understand others from culturally different backgrounds, or for cross-linguistic parallels in the diachronic sources of honorifics from politeness strategies.

2.2.2 Politeness as communicated or taken for granted. In contrast with rule-based approaches, B&L

(1987) insist that politeness inheres not in words or in sentences *per se*, but in utterances uttered in a context, by virtue of the successful communication of a polite attitude or intention. Polite utterances are not necessarily communicating 'real' feelings about another's social persona, but expressing contextually-expected concern for face. This concern is an 'implicature,' an inference of polite intentions, not a feature inextricably attached to particular linguistic forms. Politeness is ascribed to a speech act, or to an interactional move (if you prefer), not to a strategy or its linguistic realization *per se*.

In other approaches (for example, Fraser's (1990) 'conversational contract,' Watts et al.'s (1992) 'politic behavior'), politeness is taken to be the expected *background* to interaction; it is normally not communicated but consists in following expectations as to appropriate behavior.

2.2.3 Broad vs. narrow scopes for politeness theory. A narrower view takes politeness to be strategic orientation to potential face threats; there are some situations (e.g., task-oriented ones) where politeness may be subsumed to other goals, and there are many reasons for being indirect in speech other than politeness (e.g., humor, irony, rhetorical farce). Many motivations other than politeness guide human behavior. A more inclusive view sees politeness as orientation to the social-relationship dimension of every interaction, with attention to face taken to be an omnipresent necessity. The whole continuum from extreme politeness through a quite neutral level of politeness (maintaining the status quo, 'discernment') to rudeness (outright intentional threat) then needs to be brought into the theory.

2.2.4 Politeness from the point of view of the individual, the dyad, or the social group. The B&L model takes the interacting dyad as its unit; it is about how interlocutors make inferences of politeness from one another's deviations from Gricean efficient communication, and how stable patterns of strategies characterize dyadic interaction, providing an index to the quality of the social relationship. Many politeness strategies are quintessential examples of 'inter-subjective perspective-taking'—putting yourself in the others' position. A major goal of B&L (1987) was to insist on the centrality of social interaction as a significant level of social life, intermediate between the individual and society, where social/cultural facts (status, role, values, norms, rights, and obligations) are integrated with individual ones (goals, plans, strategies).

Yet the Gricean foundation of the theory and the speech-act based formulation of the B&L strategies have made many see the model as purely psychological (how a speaker calculates how to frame an utterance).

Arundale (1999), for example, argues for a theory of how face is *jointly constituted* in ongoing interaction. In fact, we need both perspectives: face is indisputably interactionally created and manipulated. Nevertheless, it can be considered from the point of view of the individual speaker or hearer (as B&L do in their production/comprehension model), or of the society or social group (as 'face constituting theory' and most sociolinguists do).

2.2.5 The indeterminacy problem. The usefulness of the B&L (1987) model as an ethnographic tool for analyzing the quality of social relationships in any society is undermined by the contextual dependency of ratings of the social factors. The problems of the mixing and ordering of strategies are compounded by the indeterminacy of context-dependent P, D, R assessments, which make it hard to code levels of politeness in any concrete situation. This is a problem with any theory in terms of actors' intentions when applied to empirical data; as both interactors and conversation analysts know, it is not always possible to be certain what interlocutors' intentions are at a particular point in natural interaction.

The B&L model of politeness as originally formulated clearly needs elaboration and revision. Nonetheless, it retains its centrality in research on politeness, largely because it provides a coherent set of concepts for analytically dissecting polite speech in different societies and contexts. It is also the only theory that begins to provide an account for the detailed cross-cultural parallels in polite speech.

3. Conclusions

Politeness has attracted an enormous amount of research attention since the 1970s, and continues to be a major focus for research in disciplines concerned with social interaction. The study of politeness phenomena can provide insight into widely differing issues; foci of interest correspondingly differ widely. They include, for example, analyses of the sequential development of politeness in natural interaction, the study of cross-cultural pragmatics and misunderstanding, the ethnography of speaking of face and politeness in different cultures and contexts, politeness strategies as providing the stylistic coherence of particular types of interaction (e.g., gender differences in speech style), politeness as a functional motivation for linguistic structure, the social psychology of face management and interpersonal perception, politeness theory applied to the analysis of formal ritual and to a view of culture as 'rhetoric,' or forms of effective expression.

Research on politeness has, however, been much weaker on the theoretical front. A major limitation is the kinds of data used in analyses. A large proportion of studies take as their data people's conscious

evaluations of politeness expressed in sentences, judgments which tend to be both prescriptive and stereotypical. Far fewer studies use as data recordings of situated conversational exchanges to explore how politeness is achieved sequentially in naturally-occurring discourse, and only a handful provide the crucial kind of evidence necessary to test the universality of any theory of politeness: for a particular society, an 'ethnography of speaking' providing evidence across different contexts to show how politeness is modulated in relation to social factors (P, D, R and others) in that society.

The emphasis in research has been largely on cross-cultural *differences*, with insufficient attention addressed to the cross-linguistic/cross-cultural parallels which tend to be taken for granted when they are not disputed. But politeness has a significance far beyond the P's and Q's of appropriate behavior and speech. The wider significance of politeness is in the interactional, communicative, day-to-day basis of social life, and the conduct of social relationships. Any work in this area needs to be based in a theory of social interaction that takes account *both* of our common human nature (and ability to communicate cross-culturally) and of our cultural differences (which makes us sometimes misunderstand one another), to anchor research on politeness—this is the great weakness in butterfly-collecting approaches.

See also: Communication and Social Psychology; Communication: Philosophical Aspects; Communicative Competence: Linguistic Aspects; Expression and Interaction: Microfoundations of Expressive Forms; Interactionism and Personality; Intergroup Relations, Social Psychology of; Pragmatics: Linguistic; Social Identity, Psychology of; Social Psychology; Status and Role, Social Psychology of

Bibliography

- Arundale R 1999 An alternative model and ideology of communication for an alternative to politeness theory. *Pragmatics* 9: 119–53
- Brown P 1995 Politeness strategies and the attribution of intentions: The case of Tzeltal irony. In: Goody E (ed.) *Social Intelligence and Interaction*. Cambridge University Press, Cambridge, UK, pp. 153–74
- Brown P, Levinson S C 1987 [1978] *Politeness: Some Universals in Language Use*. Cambridge University Press, Cambridge, UK
- Brown R, Gilman A 1968 [1960] Pronouns of power and solidarity. In: Fishman J (ed.) *Readings in the Sociology of Language*. Mouton, The Hague, Netherlands, pp. 252–76
- Coulmas F (ed.) 1991 New perspectives on linguistic etiquette. *International Journal of the Sociology of Language* 92
- Dufon M A, Kasper G, Takahashi S, Yoshinaga N 1994 Bibliography on linguistic politeness. *Journal of Pragmatics* 21: 527–8
- Fraser B 1990 Perspectives on politeness *Journal of Pragmatics* 14: 219–36

- Goffman E 1967 The nature of deference and demeanor. In: Goffman E *Relations in Public: Microstudies of the Public Order*. Harper and Row, New York
- Kasper G 1990 Linguistic politeness: Current research issues. *Journal of Pragmatics* **14**: 193–218
- Lakoff R 1973 The logic of politeness or minding your p's and q's. In: *Papers from the Ninth Regional Meeting of the Chicago Linguistic Society*. Chicago, pp. 292–305
- Leech G 1983 *Principles of Pragmatics*. Longman, London
- O'Driscoll J 1996 About face: A defence and elaboration of universal dualism. *Journal of Pragmatics* **25**: 1–32
- Streckler I 1998 *The Social Practice of Symbolization*. Athlone, London
- Tracy K 1990 The many faces of facework. In: Giles H, Robinson W P (eds.) *Handbook of Language and Social Psychology*. Wiley, Chichester, UK, pp. 209–26
- Watts R J, Ide S, Ehlich K (eds.) 1992 *Politeness in Language*. Mouton, Berlin

P. Brown

Political Advertising

Political advertising has been the subject of intense scholarly investigation for nearly a century. Two agendas underlie this strain of social science inquiry.

The first agenda seeks to measure how political advertising, and campaigns generally, produce votes. Candidates' campaign advertisements are a major source of information about electoral choices that voters face. The intellectual origins date to Gosnell's studies of voting in Chicago in the 1920s (Gosnell 1927, 1935). How much and how do advertisements influence election outcomes? The answers to these questions have very practical uses for political practitioners, who seek information that can guide campaign strategies, and for social scientists, who seek to explain how elections are won. Many social scientists question whether political advertising has any influence on elections, because elections are readily forecasted using indicators of peace and prosperity. Key writes in *The Responsible Electorate* that 'The hullabaloo of a presidential campaign so commands our attention that we ascribe to campaigns great power to sway the multitude. Campaigning does change votes and it does bestir people to vote. Yet other influences doubtless outweigh the campaign in the determination of the vote' (1966, p. 9). On election forecasts, see Rosenstone (1983).

The second agenda concerns the extent of voter autonomy. Are votes hollow echoes of campaign messages, or does political discourse enable voters to reach reasonable decisions that reflect their own preferences and values? Electoral reform motivates much research on advertising and voter autonomy. If political advertising easily misleads voters, then restrictions on campaign advertising or some sort of

'political consumer protection' may be necessary in order to make democracy work. To the extent that there is a scholarly consensus on this matter it is that concerns about voter autonomy are overdrawn. Behavioral research in a wide variety of settings indicates that advertising has modest but important effects on election outcomes: advertising can make a difference in a close election, but it does not appear to persuade large numbers of voters to turn against a candidate or party that is well-known, well-liked, and represents the voters' interests.

1. Advertising and Vote Production

Political advertising covers a wide range of media that candidates use to mobilize supporters to turn out or to persuade people to vote for them. Commentary on political advertising often equates it with television advertising, as, for instance, when European commentators decry the 'Americanization' of their politics. But even in the US there is, in fact, much less television in politics than is often supposed. Campaigns for the US House of Representatives, for instance, use as much direct mail as television advertising. And in local campaigns candidates use little television but a vast range of other media, including mail, billboards, lawn signs, telephones, and door-to-door canvassing (Key 1966).

Campaigns' usage of different media appears to follow the same principles that apply to firms. Candidates choose what media to use based on the productivity of the media and their relative prices. Variation in the mix of media used at different levels of office suggests that there are economies of scale to different sorts of advertising. Personal contact is reputed to be the most effective mode of political communication and it is a staple of local elections everywhere. But it is impossible to spend 15 minutes with each of the millions voters in a national election.

Mass media are most cost-effective at reaching broad constituencies. Evidence of the rational use of media comes from US House races. The average household in the typical House district receives about two pieces of mail and sees at least two television advertisements. However, candidates choose a mix of advertising media based on the prices and productivity of those means of reaching potential voters (or consumers). Campaigns for the US House in districts with expensive broadcasting media (such as Los Angeles and New York) spend \$2.50 on mail for every \$1 broadcasting, or three pieces of mail and 15 seconds of advertising per person. House campaigns in districts with inexpensive broadcasting (such as Dayton, Ohio and El Paso, Texas) spend \$10 on broadcasting for every \$1 on mail, or one piece of mail and 8 minutes of advertising per person. (For details see Ansolabehere et al. 1999a.) To my knowledge, similar sorts of

calculations have not been conducted for other countries, though it is evident from patterns of media usage that parties and candidates in Europe, Australia, and elsewhere are similarly responsive to media prices. (See, for example, Blumler 1983 and Australian Election Commission 1997.)

Social science research has focused on how voters respond to the overall volume and different sorts of campaign advertising that they receive. The estimated effects depend in part on the methods used.

Three different approaches have been taken to the study of advertising. First, laboratory and field experiments have been used to gain precise measures of the effects of advertising exposure in specific campaign settings. Such studies control what messages individuals receive and then measure differences in opinions or behaviors in response to different sorts of messages (e.g., Gosnell 1927, 1935, Ansolabehere and Iyengar 1996, Gerber and Green 1999, Norris 1999).

Second, studies of aggregate data correlate the observed vote to levels of campaign spending. Some of these studies design their inquiries around 'quasi-experiments' which exploit factors that change levels of spending (such as personal wealth) in order to measure the effect of spending on the vote. The seminal work in this vein, for both American and European researchers, is Jacobson (1978). Subsequent research has improved on this model; see, for example, Gerber (1998). Races involving three or more candidates pose statistical problems in estimation of vote production functions, for which widely accepted statistical models have not yet been developed. Third, surveys associate reported vote or political attitudes with reported or recalled exposure to messages. Almost all research outside of the US relies on survey methods. For a discussion of the problems with survey based measures of campaign effects see Ansolabehere et al. (1999b). Analogous with the aggregate studies, instrumental variables can solve such problems.

Experiments, which offer high internal validity, show a remarkably consistent set of results about the capacity of advertising to produce votes. Researchers observe exactly what messages citizens see and can design the study so that other factors do not confound the effect of the experimental manipulation. Perhaps the earliest experimental study of political advertising and campaigning took place in the 1920s in the US. To measure the mobilizing effects of direct campaigning, Gosnell (1927) sent a cadre of graduate students out to canvass a set of blocs of the city of Chicago in an effort to manipulate the levels of voter contact. They found higher in turnout ranging from 4 to 7 percent in the areas that the students canvassed over the areas that they did not canvass. Fast forward to the 1990s. Gerber and Green (1999) replicated Gosnell's study in several towns in southern Connecticut. They found increases in participation similar to those reported in the Chicago studies 70 years earlier. They also found increases in vote shares of candidates in the neighbor-

hood of 4 percentage points for two pieces of mail. In a series of laboratory experiments conducted in Los Angeles from 1990 to 1993, Ansolabehere and Iyengar (1996) found strikingly similar results for television advertising. Exposure to a 30-second television advertisement raised the percentage of viewers who intended to vote for a candidate by 4 to 10 points. These and many other experiments show substantial effects of political advertising on candidates vote shares. Those who see a broadcast message, receive a mailing, or talk to a campaign worker are between 4 and 10 percent more likely to support the candidate being promoted.

Aggregate studies of campaign spending yield similar, albeit lower estimates of the effectiveness of campaign advertising. Using fairly complicated, and controversial, statistical models, researchers have attempted to measure the marginal effect on the vote of spending an additional campaign dollar. Such techniques yield estimates ranging from miniscule effects of a dollar spend on the vote to estimates of approximately \$10 per vote. The key papers in this subfield are Jacobson (1978), Green and Krasno (1988), and Gerber (1998).

Such estimates are in the right ballpark, but the wide range of estimates in the literature on campaign spending suggests that additional work finding valid statistical models is in order before we can put much confidence in the estimates emerging from this literature.

The problem is that it is difficult to sort out cause and effect in aggregate electoral data. Consider, for example, two similar legislators: one holds a highly competitive seat and the other holds a very safe seat. The legislator in the competitive seat must spend large amounts of money to win re-election; the legislator in the safe seat is unlikely to be challenged and if they are they do not need to spend much to retain the seat. In this example the correlation between spending and vote shares is negative! To fix this problem researchers have taken a 'quasi-experimental' approach, using factors such as candidate wealth, which affect directly how much candidates spend and in turn affect their vote shares. Improving such statistical models remains an important research challenge.

Survey research has presented the most modest estimates of the effectiveness of advertising. A long line of studies report minimal effects of campaigns in general and of advertising in particular on the vote (in the US, see Campbell et al. 1960; in the UK, see Norris 1999). Serious problems of question validity lead to serious doubts about these research findings. In the laboratory setting, Ansolabehere and Iyengar (1996) document that survey-based measures of advertising exposure miss one exposure for every one that they detect. This degree of misreporting is sufficient to explain why surveys often produce insignificant findings or even estimates in the 'wrong' direction. Quasi-experimental techniques analogous to those used in

aggregate studies can correct measurement problems (Bartels 1993).

Beyond differences in media and methodologies, a further theme in the study of political advertising is the mechanism through which messages become votes. There are three. Advertising may solidify an uncertain vote (reinforcement), it may stimulate a supporter to vote (mobilization), and it may lead a voter to change their mind (conversion). This typology is due to Berelson et al. (1954). Although political observers often talk of 'swing' voters and movement in the polls, there appears to be relatively little evidence in experiments or surveys of conversion. Few Republicans decide to vote Democrat and few Democrats decide to vote Republican in response to political advertisements. Perhaps the strongest evidence against conversion, though, comes from nonpartisans. Nonpartisans, especially low information independents, are the least responsive to political advertising.

Reinforcement appears to be the dominant mechanism in political communication. For instance, in the experimental research reinforcement produces the clearest effect on the vote. Nonpartisans and viewers of the opposite party as the sponsor of the ad showed no statistically significant response to a candidate's ads. Low information partisans of the same party as the sponsor showed by far the strongest favorable response to political advertising. That effect shrank, though it remained statistically and substantively large, with the partisan's level of information and interest in politics (Ansolabehere and Iyengar 1996). This pattern suggests that political advertising fits neatly into the more general model of information acquisition developed by Zaller, in which both predisposition to believe a piece of information and prior levels of knowledge determine responsiveness to new information (Zaller 1992).

Mobilization has proved the most controversial of the three mechanisms. Experimental research on political advertising has called into question the long-standing belief among political scientists that campaigns are inherently stimulating affairs. Berelson et al. (1954) describe the process in detail. This is essentially the model behind retrospective voting developed by Key (1966) and widely used to understand elections in the US, Europe, and elsewhere. A large number of experimental studies have demonstrated that not all viewers are stimulated to vote by political advertisements. For example, nonpartisan and low information viewers exposed to negative advertising express significantly lower intentions to vote than others. As a result, negative campaigning can produce lower turnout among the key group of 'swing' voters and an electorate that reflects two ideologically opposed groups.

The demobilizing effects of negative advertisements on non-partisans underscores the difference between reinforcement and mobilization. Negative advertising might win the highest vote share, mainly by reinforcing

partisans, but it can lead to lower overall participation.

One important caveat about the state of social science research on political advertising is in order. The bulk of the research on the effectiveness of political advertising has been conducted in the United States. Many of the findings about mobilization and reinforcement likely depend on the electoral system, which influences the number of parties or candidates. The dynamics of campaigns and the consequences of advertising appear to be quite different in races involving more than two candidates. Johnston et al. (1992) provide excellent examples of political communications having complicated effects on competition in Canada's multi-party system. Careful experimental research on multicandidate and multi-party races, though, is in its nascency.

2. Voter Autonomy

The central problem for democratic theory in modern societies is that people apparently know little about the choices they face and they have little incentive to seek new information. Candidates and parties, then, have very strong incentives to provide information to voters at little or no search costs: candidates try to reach us, rather than us, them. As a result we are highly reliant on what politicians choose to say in their advertisements.

This is the problem of voter autonomy. By one telling voters have abdicated to political elites the important jobs of choosing what issues will be discussed, what elections and therefore government will be about, and even what we will think as people walk into the voting booth. From this view, people are highly susceptible to false or deceptive advertising and will vote for the candidate whom we see the most rather than the one we think is best representative. Regulation of the volume and content of political advertising may, therefore, be necessary.

The counter argument is that voters, even though they do not know much by way of details, have a clear sense of what they themselves value or prefer, and, in order to win elections, politicians must adjust their messages to what the voters want. In crafting their advertisements, a candidate selects issues on which he or she can make electoral gains, because the voters care about it, because the candidate has a strong record, or because the candidate has have taken the more popular position. This leads to majoritarian biases in what is discussed, but that is the essence of elections.

Voter autonomy presents a much more difficult set of research questions. In order to say that people are or are not misled by political communications, we must establish what it is that they prefer or value *a priori*. The effects of advertisements must then be judged against that baseline. Two questions are central.

First, do advertisements lead people to choose a candidate who is less good for their interests or values? Research showing strong reinforcement and few conversion effects suggests that the answer to this question is probably no. In both survey and experimental data it appears that people who lean toward a party become more solidly in favor of that party as the campaigns progress (Ansolabehere and Iyengar 1996, Chap. 4, Gelman and King 1990). An important variation on this arises from tactical voting in Britain, Canada, and other countries. As campaigns progress, voters supporting a candidate or party in third place or lower might vote for the second place candidate if they prefer that candidate over the front-runner. This seems to have happened in the 2000 Mexican election as support for Cardenas fell over the course of the campaign and support for Fox, who was in second place most of the race, rose.

Nonpartisans tend to move very little, and tend not to vote. The immediate political implication is that campaign discourse is polarizing: it ultimately pits Democrats against Republicans, Conservatives against Labourites. The low amount of conversion suggests, however, that people fundamentally are not misled.

Second, can advertising set the public agenda, or are people mainly responsive to what they themselves care about? While agenda setting effects have been widely documented in other areas of study, they have received surprisingly little careful study in the field of political advertising. Early evidence of agenda setting is provided by Berelson et al. (1954, Chap. 12). Two careful studies stand out. Jamieson's (1992) *Dirty Politics* documents how the Republican and Democratic strategists use their advertising and the news media to magnify their own agendas, especially Republican's usage of crime and race. Johnston et al. (1992) demonstrate that the debate over free trade completely transformed the Canadian voters' priorities, without changing their preferences on free trade itself.

Reinforcement and agenda setting can hold conflicting implications for voter autonomy. Partisan reinforcement suggests that there is little evidence that people vote against their expressed preferences; agenda setting suggests that people are vulnerable to what the campaigns choose to discuss. There may, however, be no contradiction here. It may indeed be the case in each of the examples discussed that voters were simply very receptive to new agendas, and that elections gave voters an opportunity to change public policies on trade, affirmative action, and other issues from those set decades earlier.

3. New Directions

Political advertising poses both methodological and normative challenges for social scientists, and our understanding on both levels has advanced consider-

ably over the last two decades. Advertising has important effects on election outcomes, but it does not appear to lead people to vote for candidates with whom they fundamentally disagree. It, therefore, does not on the whole undercut the normal workings of majority rule in democratic elections.

In terms of methods, a high priority for further inquiry is bridging the gaps between experimental, aggregate, and survey methods, and developing more powerful quasi-experimental approaches to correct measurement and simultaneity problems. To this end, research on vote production, which has relied mainly on quasi-experiments in aggregate data, should make greater use of experiments to map out features of the vote production function more fully, especially the effects of repeated exposures to messages and interactions between the messages and the viewers preferences.

In addition, there is considerable need for research on political advertising outside the US. The American focus of campaign research stems in part from existing regulations on elections and media. The US imposes few limits on how candidates and parties raise and spend campaign funds, on their access to media, and on what can be said on the various media. As a result, the US has proved a rich field to study advertising effects. Research on campaigning in European countries has placed a much greater emphasis on electoral laws, political parties, and tactical voting. There is, however, growing interest in campaign effects in European political science.

Beyond methodological differences, there is the matter of why. Reinforcement and mobilization are a sufficiently well established phenomenon that it is worth opening up the black box. Why are these the routes through which political advertising works, and why not conversion? The highly inductive approach of empirical researchers is likely to pay the biggest dividends, as political theory has yet to produce a powerful model of why people vote.

See also: Mass Communication: Normative Frameworks; Mass Media, Representations in; Political Communication; Political Discourse; Political Money and Party Finance; Political Parties, History of; Political Science: Overview; Polling; Television: Industry; Voting, Sociology of

Bibliography

- Ansolabehere S, Gerber A, Snyder J M Jr 1999a *How Campaigns Respond to Media Prices*. MIT, Cambridge, MA
- Ansolabehere S, Iyengar S 1996 *Going Negative: How Political Advertising Shrinks and Polarizes the Electorate*. Free Press, New York
- Ansolabehere S, Iyengar S, Simon A 1999 Replicating experiments using surveys and aggregate data: The case of negative advertising. *American Political Science Review* 93: 901-10
- Australian Election Commission 1997 *Funding and Disclosure Report, 1996*. AEC, Sydney, Australia

- Bartels L 1993 Message received. *American Political Science Review* 87: 267–85
- Berelson B, Lazarsfeld P, McPhee W 1954 *Voting*. University of Chicago Press, Chicago
- Blumler J 1983 *Communicating to Voters: Television in the First European Parliament Elections*. Sage, London
- Campbell A, Converse P, Miller W, Stokes D 1960 *The American Voter*. University of Chicago Press, Chicago
- Fritz S, Morris D 1992 *The Handbook of Campaign Spending, 1990*. CQ Press, Washington, DC
- Gelman A, King G 1990 Why are presidential election campaign polls so variable when votes are so predictable? *British Journal of Political Science* 23: 409–52
- Gosnell H F 1927 *Getting-out-the-vote: An Experiment in the Stimulation of Voting*. University of Chicago Press, Chicago
- Gosnell H F 1935 *Machine Politics: Chicago Model*. University of Chicago Press, Chicago
- Green D, Krasno J 1988 Salvation for the spendthrift incumbent. *American Journal of Political Science* 32: 884–907
- Gerber A 1998 Estimating the effect of campaign spending on senate election outcomes using instrumental variables. *American Political Science Review* 92: 401–11
- Gerber A, Green D 1999 Does canvassing increase voter turnout? A field experiment. *Proceedings of the National Academy of Sciences* 96: 10939–42
- Jacobson G 1978 The effects of campaign spending in congressional elections. *American Political Science Review* 72: 469–91
- Jamieson K H 1992 *Dirty Politics*. Oxford University Press, New York
- Johnston R, Blais A, Brady H E, Crete J 1992 *Letting the People Decide: Dynamics of a Canadian Election*. Stanford University Press, Palo Alto, CA
- Key V O 1966 *The Responsible Electorate*. Vintage, New York
- Morris D, Gamache M 1994 *The Handbook of Campaign Spending, 1992*. CQ Press, Washington, DC
- Norris P 1999 Political communications in British election campaigns: Reconsidering media effects. *British Elections and Parties Review* 5: 190–211
- Rosenstone S 1983 *Forecasting Presidential Elections*. Yale University Press, New Haven, CT
- Zaller J 1992 *The Nature and Origins of Mass Opinion*. Cambridge University Press, New York

S. Ansolabehere

Political Anthropology

1. Definitions

Political anthropology is a subfield of sociocultural anthropology and as such is concerned with the comparative, fieldwork-based study of politics in a broad range of social and cultural contexts. If sociocultural anthropology is above all a comparative discipline, concerned with the full variety of human social and cultural life, the existence of a coherent subfield of political anthropology presupposes a relatively clear definition of what constitutes ‘the political.’

Classic studies in political anthropology in the 1940s and 1950s concerned themselves with the maintenance of order and the regulation of force in societies apparently lacking the institutional forms of the modern state. In contrast, more recent studies in the anthropology of politics are as likely to be concerned with issues of citizenship and exclusion, democracy and nationalism, law and violence, but they coexist with a whole corpus of work which identifies and analyzes the workings of power in all social contexts.

There are three overlapping ways in which anthropologists might tackle the definitional issue. One would be formal: political anthropology is the study of ‘political behavior,’ wherever and however encountered, which in turn may be defined in various ways—for example, in terms of the pursuit of power, of strategizing, or of instrumentality. A second way would be to concentrate instead on function: the political refers to institutions and practices which perform political functions—the maintenance of social order, the succession to positions of power and authority, the settlement of disputes. The third would be substantive and based on a recognition that ‘the political’ is not a universal category of experience, but is a product of a certain history: it emerged in Western Europe after the Reformation, alongside the modern idea of the state, and as part of a move towards recognizably modern forms of mass politics based on the institutions of representative democracy. Insofar as such a category now also exists across the spectrum of postcolonial and postsocialist states, political anthropology is the study of whatever it is people in those places believe politics to be.

These three approaches to the anthropology of politics are further complicated by the relationship between political anthropology and other disciplines concerned with the same topics. Both the formal and the functional versions of political anthropology have obvious affinities with political science and political sociology, but less apparent overlap with political philosophy and political theory. In the early days of political anthropology, there was a reaction against philosophical accounts of human political capacities, such as social contract theories of the state, which relied on evolutionary speculation about the primitive origins of contemporary political institutions. Political anthropology’s affinity for political science and hostility to political theory also stemmed from political theory’s tendency to combine and confuse the ideal and the actual, accounts of how things *are* with speculations on how they *ought* to be (Fortes and Evans-Pritchard 1940, p. 4). However, in the latter half of the twentieth century, anthropological accounts of politics have become less confident about the reliance on observed behavior and the appeal to dispassionate science. In particular, since the 1970s, the influence of the work of Michel Foucault has focused attention on the alleged ubiquity of power, while opening the way for an empirical reexamination of the language and

claims of liberal political philosophy. At the same time, the recognition that the anthropological science of 'observed behavior' was almost always a colonial science has shifted attention from the comparative study of politics to a reflexive concern with the political constraints that have shaped the emerging discipline of anthropology. This double concern—with the anthropology of politics *and* with the politics of anthropology—animates Vincent's recent authoritative overview of the subject (Vincent 1990).

Translated into more conventional historical terms, there is a tale to be told of the rise, fall, and partial resurrection of political anthropology. Between the 1940s and 1960s the comparative study of political institutions became a central anthropological interest, especially for social anthropologists in the European tradition. From the early 1970s political anthropology swiftly fell out of fashion, before a number of political themes—power, nationalism, resistance, the state—returned to the anthropological agenda in the mid 1980s. Even so, many anthropologists who studied obviously political topics, but not in the style of 1960s political anthropology, were reluctant to identify themselves with the label of 'political anthropology.'

2. From Function to Process to Form

A comparison of three key works from the early years of political anthropology provides a better sense of the strengths and weaknesses of canonical political anthropology. In 1940 Edward Evans-Pritchard published his monograph on *The Nuer*, subtitled 'A description of the modes of livelihood and political institutions of a Nilotic people' (Evans-Pritchard 1940). Evans-Pritchard had conducted his research among this group of people in the 1930s at the request of the colonial government. His book found the key to Nuer political order in the field of kinship, specifically in the structure of dominant lineages, and the systematic way in which segments of the population would unite in antagonism to structurally equivalent segments. The result was an extraordinarily neat and coherent account of the emergence of political order in the absence of a unitary ruler. Key issues in Evans-Pritchard's account—such as the ordered exchange of violence in the institution of the feud, or the structure of so-called 'segmentary lineage systems'—generated their own specialist literatures, and inevitable critiques, in the years that followed.

Fourteen years after Evans-Pritchard's Nuer study, Edmund Leach published an analysis of highland Burma of comparable intellectual ambition (Leach 1954). Like Evans-Pritchard, Leach concentrated on a group of people on the further fringes of colonial power. As his title suggested, Leach was also concerned with the idea of the political system, or in this case *systems*, and, like Evans-Pritchard, he rooted parts of his analysis in the everyday structures of

kinship and marriage. But Leach was also concerned to historicize his account, and to insert a dynamic element. Dissatisfied with the conservative assumptions of functionalist anthropology, Leach attempted to problematize the idea of a system in equilibrium, by showing how, over a relatively wide area and a relatively long period of time, the same people could oscillate between an egalitarian and a hierarchical political structure, each of which was inherently unstable and contained within itself the germ of its opposite.

The same concern with the dynamic working through of internal contradictions could be found in Victor Turner's (1957) analysis of village micropolitics in what was then Northern Rhodesia (now Zambia). Eschewing the more geographically and historically ambitious vision of Evans-Pritchard and Leach, Turner concentrated on the meticulous analysis of specific cases of conflict within a specific village. Pattern and form were to be found not so much in a social structure, apparently frozen in time, but in the orderly sequence of events within what Turner described as his 'social dramas,' while the shared symbols in village ritual bound people together, even as political conflict drove them apart.

The studies of Evans-Pritchard, Leach, and Turner assumed canonical status in the emerging subdiscipline of political anthropology, and together they can give a fairly good idea of what kind of work political anthropologists of the 1950s and 1960s might be expected to do. First, and most obviously, political anthropologists were relatively unconcerned with the political institutions of the modern nation state, and, although all three authors were working within the political structures of late colonialism, European rule was not a central concern of their work. Their empirical focus was on relationships of kinship, marriage, and residence, and their intellectual goal was to locate the principles of local order in the details of everyday conflict and show the structural consequences of those principles as they applied across a wider social field. Yet, as we move from one author to the next, two related changes emerge, and one unresolved dilemma is left hanging. The two changes were the shift—evident in the contrast between Evans-Pritchard and Leach, and further developed in Turner's work—from a concern with *structure* to a concern with *process*; and the parallel move from relatively large-scale models of social structure to what Turner described as 'diachronic micro-sociology' (1957, p. 328). The unresolved dilemma concerned the cultural content of politics.

The 1960s were the subdisciplinary heyday of political anthropology, and many of the edited collections and monographs of this period celebrate this shift towards a processual view of politics, and evince a growing concern with the microstudy of political behavior (cf. Swartz et al. 1966, Bailey 1969). In the terms set out at the start of this article, if political

anthropology had started out as the study of political *functions* in societies apparently lacking modern political institutions, it gradually became the *formal* study of political behavior in all kinds of society. But the formal definition often rested on a hostility to the cultural meaning of politics. In *African Political Systems*, Fortes and Evans-Pritchard had insisted on the need to ignore culture for comparative purposes: 'A comparative study of political systems has to be on an abstract plane where social processes are stripped of their cultural idiom and are reduced to functional terms' (1940, p. 3). In *Political Systems of Highland Burma*, cultural differences became mere tokens in the political games analyzed by Leach (1954, pp. 16–17). Yet this antagonism to the cultural meaning of politics was never simple and complete: Turner concludes *Schism and Continuity* with a recognition of the binding force of ritual and symbols—paradoxically, a point also acknowledged by Fortes and Evans-Pritchard (1940, pp. 16–22)—and the promise, richly fulfilled in his subsequent work, of more analysis in the future of 'the cultural aspects of Ndembu ritual' (1957, p. 331).

3. After Formalism: Domination, Resistance, and the Nation

The emphasis on formal models of political behavior owed much to the prevailing positivism of political science, especially North American political science, but was increasingly out of joint with the anthropological *zeitgeist*. The 1960s and 1970s saw a major shift away from the formal, typologizing type of work found in political anthropology, and towards the more humanistic style once characterized by Evans-Pritchard as the 'translation of culture.' In Europe, this shift was given theoretical impetus by Lévi-Strauss' structuralism, which directed attention to the study of myth and ritual; in North America, the same role was played by the cultural analyses of Clifford Geertz and David Schneider, with their roots in Parsonian theory, and their concern with culture as a system of symbols and meanings. By the early 1970s, a leading political anthropologist described an intellectual landscape unhelpfully divided between those he characterized as the 'thought structuralists' and those interested in power relations (Cohen 1974). The task, ever since, has been to attempt some sort of *rapprochement* between anthropology's symbolic wing, concerned with the explication of cultural meaning, but often at the expense of any concern with social context and social process, and its political specialists, who had been increasingly identified with a rather narrow and culturally arid formalism.

One area of convergence that emerged in the 1980s was in the anthropology of power and resistance. This represented a development of themes that had first emerged in the feminist and Marxist work of the

1970s. To what extent are marginal or oppressed people—women, the poor—mystified by the ideological representation of their own position? Is culture a source of such mystification, or is it a resource for resistance? Is culture internally homogeneous or, as the slogan of the times put it, is it 'always contested'? Theoretical answers to these questions were found in a range of new sources—Marxist and post-Marxist theorists of culture like Raymond Williams, Antonio Gramsci, Pierre Bourdieu, and Stuart Hall; feminists like Judith Butler and Gayatri Chakravorty Spivak; the radical South Asian historians of the Subaltern Studies group; and, perhaps above all, Michel Foucault. Empirically, instances of resistance might be found in everything from changing domestic consumption patterns, to agricultural workers' self-conscious footdragging, to participation in new religious movements. From a considerable distance, it might be possible to discern a strong continuity with the 1960s analysis of local micropolitics, but the antipositivism and self-conscious poststructuralism of many writers on resistance left them reluctant to acknowledge too strong a kinship with the political anthropology of a previous generation.

A second area of convergence between anthropology's symbolic and political tendencies was in the study of nationalism. The wave of decolonization in the decades after World War II provided anthropology with an obvious area of research. But despite interesting attempts at interdisciplinary enquiry, such as Chicago's Committee for the Study of New Nations, the anthropology of nationalism failed to take off in the 1950s and 1960s. The publication of a provocative essay by a highly 'anthropological' political scientist, Benedict Anderson's (1983) *Imagined Communities*, put nationalism back on the anthropological agenda, and the emergence of ethnic and nationalist conflict in the years after the collapse of Soviet hegemony in Europe provided further impetus. However, while the new anthropology of nationalism has produced a great deal of important and imaginative empirical research, it has been somewhat less theoretically impressive. Some of the most impressive anthropological work has been on the experience of ethnic and nationalist violence, but this has been overwhelmingly studied from the point of view of the victims, rather than the perpetrators, of violence (e.g., Das 1990). The big political mystery of the twentieth century, mass mobilization in the name of the nation, remains as problematic for anthropology as for the other social sciences. In the face of the breakup of the Soviet empire and the war in former Yugoslavia, political anthropology's key contribution in the 1960s—instrumental analyses of the politics of ethnicity—now looks dangerously reductionist. The major theoretical gain from the anthropology of nationalism has been a new attentiveness to the political history of anthropology's central concept of 'culture,' and a reflexive awareness of anthropology's potential complicity in

nationalist interpretations of culture and cultural difference (Handler 1988).

4. Towards Complexity

Although the label 'political anthropology' still retains some of the stigma of 1960s formalism, the study of politics has returned to the center of sociocultural anthropology since the 1980s. There are two obvious and related reasons for this. One is an increased openness to interdisciplinary collaboration with historians, political scientists, and political theorists as anthropologists grapple with the spatial and temporal limitations of the ethnographic method. The other is a renewed critical interest in the received vocabulary of modern politics, obviously dense and complex ideas like 'the nation' and 'the state,' but also concepts of 'citizenship,' 'civil society,' and 'democracy.'

Ethnographic enquiry itself has changed, opening up new potential issues for political anthropology. A study like James Ferguson's *The Anti-Politics Machine* (1990) exemplifies what is now possible. Ferguson's book is based on an ethnographic investigation of a large and unsuccessful development project in Lesotho. It involves at once a critical examination of the 'development discourse' that informs projects like these, and a description of the systematic ways in which potential political issues, like the sources and consequences of social inequality, are swept aside in the path of the project, while all the while the reach of state power is extended into the countryside. Theoretically, Ferguson's work owes a great deal to Foucault's critique of liberal complacency, but empirically it combines local ethnographic detail with broader understandings of the translocal processes and institutions that are so characteristic of the age of development.

In some ways, the new anthropology of the state represents a return to the founding suspicion of political philosophy in *African Political Systems*. In the Preface to that volume, Radcliffe-Brown had dismissed the idea of the state as 'a fiction of the philosophers.' In a series of provocative essays, Michael Taussig has raised the issue of the productive potential of this 'fiction': the—mostly symbolic—ways in which people are subjected to the 'state effect' (Taussig 1997). In a collection of critical studies of African politics, John and Jean Comaroff (1999) subject the notion of 'civil society' to a similar process of empirical scrutiny. Thomas Hansen's (1999) study of Hindu nationalism, combines history, ethnography, and Lacanian theory, and locates the rise of politicized Hindu identity in the peculiar postcolonial trajectory of Indian democracy. These studies represent an especially interesting departure in the relationship between political anthropology and political theory. They focus on central concepts in liberal political philosophy but, rather than viewing that philosophy

as a potentially universal and neutral idiom, treat it as one more folk theory of government, better viewed as an object of ethnographic enquiry than a source of wholesale explanation.

See also: Evans-Pritchard, Sir Edward E (1902–73); Nationalism: General; Political Economy in Anthropology; Political Sociology; Power in Society; Power: Political.

Bibliography

- Anderson B 1983 *Imagined Communities: Reflections on the Origin and Spread of Nationalism*. Verso, London
- Bailey F G 1969 *Stratagems and Spoils: A Social Anthropology of Politics*. Blackwell, Oxford, UK
- Cohen A 1974 *Two-Dimensional Man: An Essay on Power and Symbolism in Complex Society*. Routledge and Kegan Paul, London
- Comaroff J L, Comaroff J (eds.) 1999 *Civil Society and the Political Imagination in Africa: Critical Perspectives*. University of Chicago Press, Chicago
- Das V (ed.) 1990 *Mirrors of Violence: Communities, Riots and Survivors in South Asia*. Oxford University Press, Delhi, India
- Evans-Pritchard E E 1940 *The Nuer: A Description of the Modes of Livelihood and Political Institutions of a Nilotic People*. Oxford University Press, Oxford, UK
- Ferguson J 1990 *The Anti-Politics Machine: 'Development,' Depoliticization, and Bureaucratic Power in Lesotho*. Cambridge University Press, Cambridge, UK
- Fortes M, Evans-Pritchard E E (eds.) 1940 *African Political Systems*. Oxford University Press, London
- Handler R 1988 *Nationalism and the Politics of Culture in Quebec*. University of Wisconsin Press, Madison, WI
- Hansen T B 1999 *The Saffron Wave: Democracy and Hindu Nationalism in India*. Princeton University Press, Princeton, NJ
- Leach E R 1954 *Political Systems of Highland Burma: A Study of Kachin Social Structure*. Athlone, London
- Spencer J 1997 Post-colonialism and the political imagination. *Journal of the Royal Anthropological Institute* 3: 1–19
- Swartz M, Turner V W, Tuden A (eds.) 1966 *Political Anthropology*. Aldine, Chicago
- Taussig M 1997 *The Magic of the State*. Routledge, London
- Turner V W 1957 *Schism and Continuity in an African Society: A Study of Ndembu Village Life*. Manchester University Press, Manchester, UK
- Vincent J 1990 *Anthropology and Politics: Visions, Traditions, and Trends*. University of Arizona Press, Tucson, AZ

J. Spencer

Political Communication

Political communication is an interactive process concerning the transmission of information among politicians, the news media, and the public. The process operates downward from governing institutions toward citizens, horizontally in linkages among political actors, and also upward from public opinion toward authorities. The literature in political

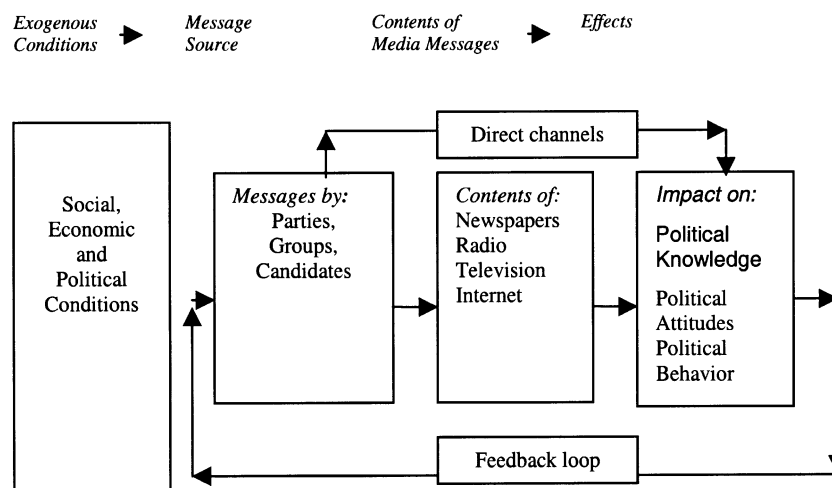


Figure 1
Systems model of political communication process

communication can be subdivided into three major categories, using a simple systems model of the process illustrated in Fig. 1 to distinguish between production, contents, and effects.

1. Production Processes

Work on the production process focuses on how messages are generated by political actors like parties and interest groups, and then transmitted via both direct channels like political advertisements and indirect channels including newspapers, radio, and television. Many studies have focused on the increased professionalization of political marketing campaigns in the postwar era, including the rise of the class of political consultants, pollsters, advertising executives, and their coterie, and the consequence of this process for strategic communications by political parties and interest groups. A large literature, particularly within Europe, has also studied the changing structure of the news industry, notably the economic basis of the newspaper industry and the legal structure regulating broadcasting and the press. Comparative studies have also commonly analyzed the news culture, especially the values that journalists, broadcasters, and editors employ as 'gatekeepers' in deciding 'what's news,' as well as the organizational structure of newsrooms. Recent work within this area has generated a growing body of research on the rise of new communication and information technologies, and use of the Internet by parties, new social movements, and the news media.

One of the major challenges in this subfield is to widen the scope of comparative research on a sys-

tematic basis so that we move from studies of the structure of the political communication process within particular nations to conceptual typologies, broader theories, and empirical generalizations that can be tested across different types of societies. Much of this work has traditionally focused upon postindustrial nations, particularly the USA and Western Europe, although in the 1990s increased attention has been paid to new democracies in Central and Eastern Europe, Latin America, and Asia, as well as, to a lesser extent, to the role of the news media in authoritarian regimes like Burma, China, and Cuba where the role of the free press and opposition movements remain severely restricted.

2. Contents

Another related mainstream research tradition has examined the content of the messages that are produced by this process, such as the amount and tone of political reporting presented in television news, the partisan balance in the press, the coverage of election campaigns and particular events, the agenda-setting reporting of policy issues, and the representation of social minorities in the news media. The most common forms of comparison are between different media, for example, differences in the messages conveyed during an election campaign in paid political advertisements, party press releases, newspaper columns, and TV news stories.

Other forms of comparison examine trends over time, such as the coverage of political scandals or social minorities in recent decades. More rarely, collaborative teams have attempted comparison

across nations, for example, concerning a specific period or event in selected major newspapers. Work in this tradition has drawn largely upon systematic forms of content analysis of a representative random sample of stories among different media, although alternative qualitative techniques for deconstructing textual and visual messages are also common. Moving beyond systematic description, which is valuable but limited, the main challenges in this area are to relate the content of the messages to either the production process (to examine their possible causes) or to their potential impact (to understand their effects).

3. *Effects*

Lastly perhaps the largest body of research, certainly in the USA, has focused at individual level on the study of the potential effects of exposure to different types of political communication messages. The most common method has been to draw upon cross-sectional or panel representative surveys or, more rarely, experimental methods. The key issues have focused upon analyzing the potential impact of exposure to different type of mediated messages (such as watching an advertisement or news story) upon either political knowledge and opinions (such as awareness about an issue, civic knowledge, or recognition of political leaders), political attitudes and values (such as support for a particular party or issue), and political behavior (such as voting turnout). Most work has focused upon the impact of the messages on the mass public or particular subgroups, like women or undecided voters, although some studies have also analyzed the effects upon middle-level elites involved in the policymaking process.

The primary challenges are threefold: (a) to expand generalizations beyond the USA, to see how far they continue to hold up within very different contexts; (b) to move beyond cross-sectional surveys, which cannot determine issues of causality, toward more dynamic designs such as panel surveys and pre-post experimental designs; and (c) to link studies of the individual-level analysis of effects to both what we know about the structure of the news industry and the contents of the messages, generating multilevel and multinational research designs. This work has made considerable progress as the study of political communications has moved increasingly into the mainstream within political science since the 1980s, but nevertheless the subfield remains predominantly US and European, and the process of internationalization in what is increasingly a global society is only starting to produce more systematic cross-national and multilevel research.

Political communication has, therefore, always been central to the electoral and policymaking process but since the 1990s certain important structural develop-

ments have fundamentally altered this process, particularly postwar trends in the mass media moving from the traditional world of newspapers, radio, and television broadcasting toward the Internet. The rest of this article outlines alternative interpretations of the nature of these trends and reflects on their consequences for the process of socioeconomic and political development around the globe.

4. *Postwar Trends in Newspapers*

Concern about traditional standards of journalism has been fuelled by major changes in the newspaper industry during the postwar era. In the USA the daily press has experienced dwindling readership and sales, especially among the younger generation, a loss of advertising market share to the electronic media, and growing concentration of ownership in larger multiple newspaper chains or a few multimedia conglomerates (McManus 1994, Baker and Dessart 1998). All these developments have had a major impact upon the profitability and economic viability of the print sector in the USA, particularly for smaller outlets (De Bens and Ostbye 1998).

Yet although the demise of newspapers has been widely predicted for decades, we should not underestimate their continued popularity and technological adaptation to new forms of production and distribution. If we compare postwar trends in circulation, controlling for population growth, the evidence shows that sales of the daily press in most postindustrial societies has not been affected by the growing availability of electronic media. The long-term trend in newspaper sales in OECD countries, per 1,000 population, has proved fairly stable. Average circulation in all OECD states was 271 newspapers per 1,000 population in 1950, rising modestly in 1980, before subsiding slightly to 263 per 1,000 in 1996 (UNESCO). In postindustrial societies, despite the massive surge in the availability of television since the 1950s, about one quarter of the population continues to buy a daily newspaper, and readership figures are even higher. The electronic media have therefore increased the choice and diversity of news outlets and formats, but at the same time they have not killed sales of the printed press. Given growing educational levels and affluence characteristic of postindustrial societies, consumption of news has not proved a zero sum game.

Moreover there are considerable variations among societies in use of the traditional mass media. Figure 2 illustrates the penetration of newspapers and TV sets per 1,000 population worldwide. The pattern confirms how far newspaper and TV penetration levels are associated with basic patterns of socioeconomic development; use of the mass media is related to a more affluent and educated population with greater leisure

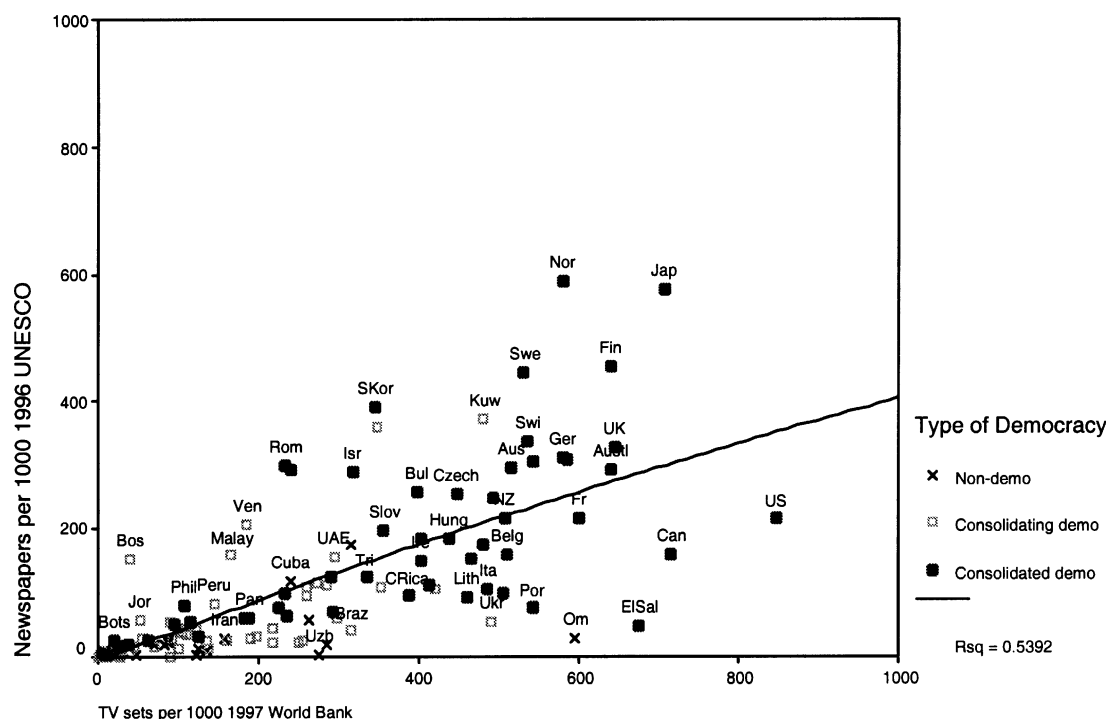


Figure 2
Worldwide distribution of newspapers and TV sets, mid-1990s. (Sources: Newspapers: UNESCO Statistical Yearbook 1997; TV sets: World Bank 2000; Type of democracy: Based on the Freedom House classification of political rights and civil liberties.)

time. Postindustrial societies tend to be the heaviest consumers of the mass media; nevertheless the pattern also shows considerable variations among these nations. Some outliers like the USA and Canada prove far more television-oriented than average (along with Southern Mediterranean Europe and much of Latin America), while others nations like Norway and Sweden (along with much of Northern, Central, and Eastern Europe) remain more newspaper-oriented.

It remains unclear whether systematic trends in the newspaper industry have changed traditional journalism, producing an increased focus on crime, sex, and entertainment, as is assumed by some critics. What seems equally plausible across OECD countries is an expansion of both lowbrow and highbrow news media in recent decades, representing a diversification of the market. A recent review of the comparative literature by Kees Brants concluded that the few available content analysis studies provide an ambiguous and sometimes contradictory picture of the growth of 'infotainment' news in different countries, rather than showing a uniform pattern: 'Where for the European countries as a whole we might see a slight tendency towards the popularization of news, there is little evidence that politicians and politics are more

dramatically personalized and sensationalized than before' (Brants 1998).

Brants found that the available content analysis shows a mixed picture of the growth of 'infotainment' news in different European countries, rather than a uniform trend. Frank Esser concluded that there were marked contrasts between Germany, the UK, and the USA in the popularity of tabloid news, and that the nature and degree of competition in a particular media market is the decisive factor explaining the degree of tabloidization (Esser 1999). Moreover systematic research on long-term trends in British newspapers from 1952–97 found that the amount of political coverage in the tabloid sector had not declined over time, as many critics assume. Instead, the tabloid press in the UK has expanded its coverage of entertainment but also maintained its political news over this period (McLachlan and Golding 1999, McLachlan 1999).

Therefore across all postindustrial societies newspaper circulation levels have remained largely stable during the postwar era, yet at the same time the range of papers published in these states has contracted. The number of daily newspapers published in OECD nations fell on average by 15 percent during the postwar era, from 160 per country in 1960 to 130 in

1996, producing greater concentration of readership in fewer outlets. Many countries have introduced measures to maintain press diversity, on the assumption that we need diverse outlets for an effective civic forum. Antitrust regulations have attempted to insure competition in the ownership of the press, for example, limiting the proportion of cross-media ownership by a single company, administered by fair trade bodies like the British Monopolies and Merger Commission or the German Federal Cartel Office. Other societies like Sweden and Norway have used press subsidies as a policy instrument to protect the financial viability of the more vulnerable sectors of the press (McQuail and Siune 1998). Countries with provincial and localized newspaper markets like the USA and Germany proved particularly prone to media mergers and acquisitions, reducing pluralism and competition in many cities. Papers in smaller countries like Austria and Belgium also often experienced takeovers or closure because of a limited domestic market and imports from neighboring states with a shared common language.

Concentration of ownership in the hands of a few multinational corporations with multimedia empires has become increasingly common, notably Rupert Murdoch's News International, and the vast holdings of Bertelsmann in Germany, or Fininvest in Italy (Tunstall and Palmer 1991, Smith 1991, Sanchez-Tabernero 1993). Hence Rupert Murdoch, who started with two small Australian newspapers, built an empire in News Corp. that includes 20th Century Fox films, the Fox TV network, a number of US stations, 50 percent ownership of Sky TV, a majority interest in the STAR Asian satellite, ownership of *The Sun* and *The Times* in the UK, additional television stations in Latin America, and the book publisher HarperCollins, as well as investments in Internet companies. In the USA, Time Warner's purchase of Turner Broadcasting Systems Inc (including CNN) in 1996 created the largest media firm in the world with strong print, cable, and programming divisions. Walt Disney Company's acquisition of Capital Cities/ABC Inc for \$19 billion in 1995 created the second largest media conglomerate with movie, television, cable, and Internet interests, although the purchase proved costly since ABC's profitability moved sharply into the red four years after acquisition. Conrad Black's acquisition of Southam in Canada in 1996 gave his company, Hollinger Inc, control of two-thirds of the newspapers in that country.

What will be the consequences of these developments? Many commentators like Ben Bagdikian (1997, see also Bogart 1995) fear that media mergers have concentrated excessive control in the hands of a few multinational corporations, who remain unaccountable to the public, and that only greater economic competition can change this situation. Recognition of this problem has arisen from the understanding that economic controls can constrain the media just as significantly as political controls. There is nothing new

about this concern, which was often expressed during the interwar era of the press barons when proprietors like Beaverbrook and Rothermere actively intervened to further their political ambitions. Yet others like Robert Picard (1988, 1998) remain more sanguine about recent developments, arguing that we need to distinguish between concentration defined by considering the number of media outlets held by dominant firms and concentration defined by dominance in a clear geographical market. It is the latter—which can harm consumers by producing fewer choices, poorer services, and higher prices—which is important for the availability of alternative sources of political information in a democracy.

Monopolies in the local market for ideas can be harmful for pluralism. Nevertheless it must be recognized that we need to look beyond any single media sector to establish the harmful political effects of concentration, since consumers use and have access to multiple sources of news and information, from newspapers to radio, television, and the Internet. Moreover the trends toward greater concentration are not universal, as some OECD countries have seen a significant expansion in the circulation and range of daily newspapers being published in the postwar era, particularly states like Mexico and Greece where educational and literacy rates have been rising sharply, as well as more modest growth evident in newer democracies like Hungary, the Czech Republic, and South Korea.

5. Trends in Broadcasting

Just as there are serious concerns about the future of newspapers, so many believe that in recent decades traditional standards of television news and public affairs have come under threat from technological and economic developments. The critical factors transforming broadcasting include the proliferation of channels on terrestrial, cable, satellite, digital, and now broadband services, fragmenting the mass audience; the crisis of identity and funding facing public service television, which once enjoyed a monopoly throughout most of Europe, following the rise of myriad commercial competitors; and lastly the more recent technological convergence with the digitization and compression of images, sounds, and data which has produced a new multimedia environment breaking down the traditional boundaries between telecommunications, the audiovisual industries, and computers.

These trends have affected all OECD countries to different degrees although their impact and timing has been strongly mediated by the existing communications landscape. In the Thatcherite 1980s, deregulation and privatization had the most profound influence on public service broadcasters throughout Western

Europe (Achille and Ibanez Bueno 1994). Following the fall of the Berlin Wall, the transition to democracy in the early 1990s produced an even more radical jolt to public television in Central and Eastern Europe (O'Neill 1998). Meanwhile in the USA the long-standing rule of the three major networks experienced an equivalent coup d'état, cut down in the 1980s by myriad competitors on cable and satellite (Kimball 1994). Despite the long-standing contrasts between the commercially dominant major networks in the USA and public television in Europe, in recent years both have faced a strikingly similar multiplication of outlets and fragmentation of television audiences, raising new concerns about the standards of programming.

Despite the deluge of commercial alternatives, the public channels remain popular; on average across all OECD states public channels maintain a 42 percent share of the television audience. This varies substantially, however, by country. Public television RAI1 and TVE1 remain market leaders in Italy and Spain, while NRK in Norway and SVT in Sweden had most of the best-rated shows in their countries. In contrast public TV has a far smaller share in some other societies such as PBS (3 percent) in the United States and NHK (18 percent) in Japan (Mediametrie 1999). OECD states can be classified into three major types: those that remain predominately *public* systems (based on an audience share of public channels of 60 percent and above), *mixed* systems (with a public share of 40–59 percent) and predominately *private* systems (with a public share of less than 40 percent). Where there is comparable data, only three OECD nations can be categorized as predominately public (Austria, Denmark and Hungary), 11 represent mixed systems, while 10 can be classified as predominately private systems. Therefore although there has been growing transatlantic convergence over the years, nevertheless broadcasting systems continue to bear the distinct imprint of their origins with radio in the 1920s, and public television continues to remain popular in many countries, rather than being swamped by commercial services.

6. The Rise of the Internet

Since the early 1990s the most important change to the political communication process has occurred through the rise of the Internet, particularly in postindustrial societies that are at the forefront of the information society such as the USA, Australia, and Sweden. Networked computing and computer-mediated email have existed for the scientific elite since the early 1960s but the number of users was too small to monitor through mass surveys. The key historic development transforming the Internet into the world's favorite virtual reference library, post office, and shopping

mall was a series of rapid innovations: the birth of the World Wide Web (in 1990) and the launch of popular browsers to access materials including those by Mosaic (1993), Netscape Navigator (1994), and Microsoft Internet Explorer (1995) (see Berners-Lee 2000). Subsequent technological applications, like the easy transfer of .mp3 music files and video formats, and WAP-enabled digital telephony, while representing important innovations, cannot yet claim to have had an impact equal to the basic invention of point-and-click browsers.

As yet no official government statistics on the online population are collected by international agencies like UNESCO and the International Telecommunications Union (ITU), although many indirect measures of technological diffusion are available, including investment in scientific Research and Development, the spread of computer hardware, and the rate of telephone density. The most comprehensive worldwide guide estimating the size of the online population is provided by NUA. This organization regularly monitors and collates survey results conducted by different market research companies in each nation. The surveys ask a representative sample of the public in each country about use of the Internet from home, work, or elsewhere during the previous three months. NUA's database 'How Many Online' (<http://www.nua.ie>) currently collects data from 179 countries, covering 5.7 billion people.

The NUA evidence highlights the dramatic rise in popularity of the Internet in recent years: between 1995 and 2000 the total number of Internet users surged from about 26 to 377 million worldwide, an explosive jump within the space of a few years. The Internet became a truly global phenomenon as more and more users came online from around the world and the proportion of Americans in the online community dropped from 70 to 40 percent in 1995–2000. Despite this remarkable expansion, about one in 20 of the world's population was online in 2000, with highly uneven diffusion globally.

According to NUA estimates, in Spring 2000 Scandinavia and North America led the world in rates of Internet penetration, with one-third or more of the population online, followed by Western Europe, with about one in 10 online. Central and Eastern Europe, Asia, the Middle East, and South America fall below the world average, all with less than one in 20 online, while minimal diffusion is evident in Sub-Saharan Africa, with only 36 users per 1,000 people. In terms of levels of human development, there are stark contrasts between rich and poor nations. Most of the world's online community (87 percent) lives in highly developed nations, as measured by the UNDP index (see UNDP 1999). In comparison, the 35 societies classified by the UNDP with low levels of human development, like Nigeria, Bangladesh, and Uganda, contained only 5 percent of the online population, although home to half a billion people.

A finer-grained comparison of countries ranked by the online population reveals a pattern of widespread adoption in four clusters of societies:

(a) Throughout the smaller Nordic social democratic welfare states, especially Sweden, Norway, Iceland, and Finland;

(b) In larger Anglo-American and English-speaking nations including the USA, Canada, Australia, and the UK;

(c) In the Asian 'tiger' economies of Singapore, South Korea, and Taiwan, as well as Japan; and lastly,

(d) In a few smaller European nations with above-average Internet use such as the Netherlands, Belgium, Switzerland, Slovenia, and Estonia.

At the bottom of the national rankings, with less than 0.5 percent of the population online, few Internet users are found throughout most of the poorer countries of sub-Saharan Africa (with the exception of South Africa), as well as in many states in central Europe, the Middle East, Asia, and Latin America.

Yet at the same time if technological diffusion can be achieved in poorer societies, and it is a big 'if,' then many observers hope that the Internet will provide multiple opportunities for socioeconomic and democratic development. Digital networks have the potential to broaden and enhance access to information and communications for remote rural areas and poorer neighborhoods, to strengthen the process of democratization under transitional regimes, and to ameliorate the endemic problems of poverty in the developing world. With connectivity as the umbilical cord, enthusiasts hope that the Internet will eventually serve multiple functions as the world's favorite public library, school classroom, and medical database; post office and telephone; marketplace and shopping mall; channel for entertainment, culture, and music; daily news resource for headlines, stocks, and weather; and heterogeneous global public sphere. In the heady words of the G-8 Okinawa Charter:

Our vision of an information society is one that better enables people to fulfill their potential and realize their aspirations. To this end we must ensure that IT serves the mutually supportive goals of creating sustainable economic growth, enhancing the public welfare, and fostering social cohesion, and work to fully realize its potential to strengthen democracy, increase transparency and accountability in governance, promote human rights, enhance cultural diversity, and to foster international peace and stability." (G-8 Okinawa Charter on Global Information Society 23 July 2000. <http://www.g8kyushu-okinawa.go.jp/w/documents/it1.html>)

The Internet may allow societies to leapfrog stages of technological and industrial development. On the production side, if Bangalore companies can write software code for IBM or Microsoft, and if Costa Rica can manufacture chips for Intel, then potentially entrepreneurs can offer similar services from Malaysia, Brazil, and South Africa. The Internet encourages market globalization: small craft industries and

the tourism industry in Bali or the Maldives can deal directly with customers and holidaymakers in New York and London, irrespective of distance, the costs of advertising, and the intermediate distribution chains of travel agents and retail businesses (International Telecommunications Union 1999, p. 7, Dugger 2000).

The Internet also offers promise for the delivery of basic social services like education and health information across the globe, a function that may be particularly important for middle-level professionals serving their broader community (Hayward 1995, Wresch 1996). Potentially, local teachers or community officials connected to the digital world in Lagos, Beijing, or Calcutta can access the same electronic journals, books, and databases as students at the Sorbonne, Oxford, or Harvard. Distance learning can widen access to training and education, via open universities in India, Africa, and Thailand, and language websites for schools (Arunachalam 1999). Net works of hospitals and healthcare professionals in the Ukraine, Mozambique, and Stockholm can pool expertise and knowledge about the latest research on AIDS. Peasant farmers using village community centers can learn about storm warnings and market prices for their crops, along with employment opportunities in local towns. Where peripheral regions lack access to the traditional media, the convergence of communication technologies mean that potentially the Internet can deliver virtual local newspapers, streaming radio and television video, as well as other services.

Numerous examples can be cited to show the potential of digital technologies for fostering new opportunities for development in societies around the world (World Economic Forum 2000). Many South East Asian nations seek to emulate the Japanese model of development in the postwar era of reconstruction, and the knowledge-based economy in Singapore, South Korea, and Taiwan. In Malaysia, for example, the Multimedia Super Corridor has been developed to bring investment from telecommunications, multimedia, and electronics companies, and the production of silicon wafers and software. The corridor has attracted major players such as Microsoft, Sun Systems, and NTT (Japanese telecom). Under the 'Vision 2020' plan, Malaysia now boasts cellular telephone penetration rates of one in every 10 people, more and more wired schools, and 21 Internet hosts per 1,000 people. Revenue generated by the production of information and communication technology goods, like office equipment, telecommunications, and consumer audiovisuals, shows that the USA leads the world but many Asian countries are close rivals, including Japan (second), Korea (third), Singapore (fourth), Taiwan (seventh) and Malaysia (eighth) (OECD 2000, p. 24, Table 2).

Southern India is most often cited as an important area of software development, producing an estimated \$3.8 billion in revenues, with this figure doubling in the

last years of the 1990s. Over one-half of India's software services are exported to the USA (OECD 2000). The Bangalore area has attracted inward investment from many major corporations, not least from the diaspora of the Asian dot.com entrepreneurs thriving in California's Silicon Valley and Cambridge's Technology Park (for a discussion see Yourdon 1996). In rural Bangladesh many isolated communities lack landline telephones. An innovative program by Grameen Telecom supplies cellular mobile phones to village women, who rent calls in their community to repay the loan and sustain thriving micro enterprises. (By Spring 2000, over 1,000 phones had been provided, serving 65,000 people, and the eventual target is 40,000 phones. See Richardson 2000, see also <http://www.grameenphone.com>.) With this service, local communities benefit by direct links to job, weather, and health information, as well as more efficient markets for their produce. Village telecom centers are being developed with email and fax services, along with computer literacy projects in selected schools.

In Central and Eastern Europe, Slovenia, Estonia, and Slovakia have made great strides in moving their populations online, moving well ahead of Portugal, Greece, and Austria in levels of connectivity. Hungary's ambitious Schoolnet program has allowed students in two-thirds of all secondary schools to browse the Web from their classrooms, with extensive teaching resources, interactive discussion forums, events, and competitions (see <http://www.SuliNet.hu>). In the Baltic, the Estonian government has provided public access points for the Internet throughout the country, using schools, post offices, community centers, libraries, police stations, and health clinics. The program has been highly successful; today more than one in 10 Estonians is online, with personal computer ownership well above average for Central and Eastern Europe (UNDP 1999, p. 64).

Progress has been slower in Africa, but nevertheless plans have been announced by Global Crossing, Lucent Technologies, and Africa One for an ambitious \$1.9 billion project to link up the whole continent by 2002 through a high-speed underwater fiber optic cable, with interior countries connected through terrestrial cables, microwave, or satellite facilities, overcoming many of the current problems of the inadequate telephony infrastructure (Wired News 2000. For further details see www.AfricaOne.com). Given a high-speed backbone, and market liberalization of telecommunication services, African nations may also be able to 'leapfrog' stages of industrialization through new technology by investing in fully digitized telecommunications networks rather than outdated analog-based systems. Cellular telephony is rapidly expanding as an alternative to conventional network services; the number of subscribers in the OECD region reached almost one-quarter of the population in 1998 (OECD 2000, p. 81). This growth has had even greater impact

in the developing world. In postindustrial economies there were 20 times as many mobile phones in 1998 as there were in 1990, and in developing economies there were 160 times as many, an astonishing rise (World Bank 2000, p. 299). Over a third of all telephone subscribers in Cote d'Ivoire, Cambodia, and Paraguay, for instance, are now connected via mobiles, a far higher proportion than in the USA (International Telecommunications Union 1999).

7. The Consequences for Democracy and Democratization

What will be the political consequences of changes in the traditional news media and the subsequent rise of the Internet? In many ways it remains far too early to say (for a detailed discussion, see Norris 2001). The Internet has generated deeply contested alternative visions about the future. Luddites fear for the worse, but technophiles hope for the better.

The most positive perspective is held by cyber-optimists who emphasize the Panglossian possibilities of the Internet for the involvement of ordinary citizens in direct democracy. In its more utopian manifestations, this view has been dubbed 'technoromanticism' (Coyne 1999), expressed in earlier eras in response to Samuel Morse's electric telegraph, Alexander Graham Bell's telephone, and Guglielmo Marconi's wireless radio (see, e.g., Rhodes 1999, Pursell 1995, Lubar 1993). In this account, digital technologies hold promise as a mechanism facilitating alternative channels of civic engagement such as political chatrooms, electronic voting in general elections and for referenda issues, and the mobilization of virtual communities, revitalizing levels of mass participation in public affairs (see Barber 1998). The use of the Internet by groups and social movements is often believed to exemplify digital politics.

The more utopian visions of the Internet suggest a future society in which virtually unlimited qualities of information become available, civic society flourishes, government decision making becomes more open and transparent, and nation-state borders are eroded as people build virtual communities for work, learning, and leisure spanning traditional boundaries of time and place. Although still in its adolescence, the core transformative capacities of the Internet include its potential for radically shrinking communications and information costs, maximizing speed, broadening reach, and eradicating distance. Compared with radio, television, and newspapers, controlled by editors and broadcasters, the World Wide Web facilitates a virtually unlimited choice of information and communication one-to-one (e.g., via email), one-to-many (e.g., via a personal home page or electronic conference), many-to-one (e.g., via an electronic poll) and, perhaps most importantly, many-to-many (e.g., an online chatroom), with a minimal role for gatekeepers or

government censors (see, e.g., Shapiro 1999, Weare 2000). Internet messages have the capacity to flow further, faster, and with fewer intermediaries.

As a result many hope that recent developments, especially the spread of new information and communication technologies, will serve to undermine authoritarian regimes, creating a 'dictators' dilemma' in countries like Burma, China, and Cuba (Drake et al.). Leaders in these nations want to facilitate economic development through the Internet and yet at the same time they seek to restrict political access. It is also believed that the Internet will have major consequences for electoral democracies, countries like Russia, Taiwan, and Mexico that are seeking to consolidate democratic political institutions.

Yet as the Internet evolves, a darker vision has been articulated among cyber-pessimists, who regard digital technology as a Pandora's box unleashing new inequalities of power and wealth, reinforcing deeper divisions between the information rich and poor, the tuned-in and the tuned-out, the activists and the disengaged. This account stresses that the global and social divides already discussed mean that Internet politics will disproportionately benefit the elite (see, e.g., Golding 1996, 1998, 2000). In this perspective, despite the potential for technological innovations, traditional interests and established authorities have the capacity to reassert their control in the virtual political sphere, just as traditional multinational corporations have the ability to reestablish their predominance in the world of e-commerce (see, e.g., McChesney 1999, pp. 182–5). In this view, so far the potential of the Internet has failed to have a dramatic impact on the practical reality of 'politics as usual,' for good or ill, even in countries at the forefront of digital technologies (Margolis and Resnick 2000). There are fears that continued inequalities in the spread of new technologies will exacerbate the traditional North–South divide evident in access to the traditional news media like newspapers, radio, and television. As such, while some developing countries may manage to leapfrog earlier technologies in the race towards the information society, other poorer societies may drop even further behind.

The debate between the cyber-optimists and pessimists continues and as the Internet continues to evolve over the first decade of the twenty-first century, then the impact of new technologies will become evident, whether for good (as some hope) or ill (as others fear). What is certainly clear is that political communications via the old world of newspapers and television is in the process of fundamental change and this process holds both threats and promises for the future of socioeconomic and political development.

See also: Advertising: Effects; Advertising: General; Broadcasting: General; Broadcasting: Regulation; Campaigning: Political; Electronic Democracy; Journalism and Journalists; Media Effects; Media, Uses of;

News: General; Political Advertising; Political Culture; Political Discourse; Political Money and Party Finance

Bibliography

- Achille Y, Ibanez Bueno J 1994 *Les télévisions publiques en quête d'avenir*. Presses Universitaires de Grenoble, Grenoble, France
- Arunachalam S 1999 Information and knowledge in the age of electronic communication: A developing country perspective. *Journal of Information Science* 25(6): 465–76
- Bagdikian B 1997 *The Media Monopoly*, 5th edn. Beacon Press, Boston
- Baker W F, Dessart G 1998 *Down the Tube*. Basic Books, New York
- Barber B R 1998 Three scenarios for the future of technology and strong democracy. *Political Science Quarterly* 113(4): 573–89
- Berners-Lee T 2000 *Weaving the Web*. Harper Business, New York
- Bogart L 1995 *Commercial Culture*
- Brants K 1998 Who's afraid of infotainment? *European Journal of Communication* 13(3): 315–35
- Coyne R 1999 *Technoromanticism: Digital Narrative, Holism, and the Romance of the Real*. MIT Press, Cambridge, MA
- De Bens E, Ostbye H 1998 The European newspaper market. In: McQuail D, Siune K (eds.) *Media Policy: Convergence, Concentration and Commerce*. Sage, London
- Drake W, Kalathil S, Boas T C Dictatorships in the digital age: Some considerations on the Internet in China and Cuba. *iMP: The Magazine on Information Impacts*
- Dugger C W 2000 Connecting rural India to the world. *New York Times* 28 May, <http://www.nytimes.com/library/tech/yr/mo/biztech/articles/28india.html>
- Esser F 1999 Tabloidization of news: A comparative analysis of Anglo-American and German press journalism. *European Journal of Communication* 14(3): 291–324
- Golding P 1996 World wide wedge: Division and contradiction in the global information infrastructure. *Monthly Review* 48(3): 70–85
- Golding P 1998 Global village or cultural pillage? The unequal inheritance of the communication revolution. In: McChesney R W, Meiksins Wood E, Foster J B (eds.) *Capitalism and the Information Age: The Political Economy of the Global Communication Revolution*. Monthly Review Press, New York
- Golding P 2000 Information and communications technologies and the sociology of the future. *Sociology* 34(1): 165–84
- Hayward T 1995 *Info-rich, Info-poor: Access and Exchange in the Global Information Society*. Saur, London
- International Telecommunications Union (ITU) 1999 *Challenges to the Network: Internet for Development*. ITU, Geneva, Switzerland
- Kimball P 1994 *Downsizing The News: Network Cutbacks in the Nation's Capital*. Woodrow Wilson Center Press, Washington, DC
- Lubar S 1993 *Infoculture: The Smithsonian Book of Information Age Inventions*. Houghton Mifflin, Boston
- Margolis M, Resnick D 2000 *Politics as Usual: The Cyberspace 'Revolution'*. Sage, Thousand Oaks, CA
- McChesney R W 1999 *Rich Media, Poor Democracy*. University of Illinois Press, Urbana, IL
- McLachlan S 1999 Who's afraid of the news bunny? The changing face of the television evening news broadcast.

- Information and Democracy Research Project: Working Paper No. 3. Loughborough University, Loughborough, UK
- McLachlan S, Golding P 1999 Tabloidization in the British press: A quantitative investigation into changes within British newspapers from 1952–1997. *Communication Research Centre Working Paper* #9. Loughborough University, Loughborough, UK
- McManus J H 1994 *Market Driven Journalism*. Sage, London
- McQuail D, Siune K 1998 *Media Policy: Convergence, Concentration and Commerce*. Sage, London
- Mediametrie 1999 *One Television Year in the World*. 1998. Mediametrie, Paris
- Norris P 2001 *Digital Divide: Civic Engagement, Information Poverty and the Internet Worldwide*. Cambridge University Press, Cambridge, UK
- O'Neill P 1998 *Post-communism and the Media in Eastern Europe*. Frank Cass, London
- Organization for Economic Cooperation and Development (OECD) 2000 *OECD Information and Technology Outlook*. OECD, Paris
- Picard R G 1988 *Press Concentration and Monopoly: New Perspectives on Newspaper Ownership and Operation*. Ablex, Norwood, NJ
- Picard R G 1998 Media concentration, economics, and regulation. In: Graber D, McQuail D, Norris P (eds.) *The Politics of News: The News of Politics*. CQ Press, Washington, DC
- Pursell C W 1995 *The Machine in America: A Social History of Technology*. The Johns Hopkins University Press, Baltimore, MD
- Rhodes R (ed.) 1999 *Visions of Technology*. Simon & Schuster, New York
- Richardson D 2000 *Grameen's Telecom's Village Phone Programme in Rural Bangladesh*. Telecommons Development Group, ON, <http://www.telecommons.com>
- Sanchez-Tabernero A 1993 *Media Concentration in Europe: Commercial Enterprises and the Public Interest*. John Libbey, London
- Shapiro A L 1999 *The Control Revolution: How the Internet is Putting Individuals in Charge and Changing the World we Know*. Public Affairs, New York
- Smith A 1991 *The Age of Behemoths: The Globalization of Mass Media Firms*. Priority Press, New York
- Tunstall J, Palmer M 1991 *Media Moguls*. Routledge, London
- United Nations Development Programme (UNDP) 1999 *Human Development Report 1999*. UNDP/Oxford, New York
- UNESCO 1964 *UNESCO Statistical Yearbook*. UNESCO, Paris
- Weare C 2000 Technology and politics: Linking the Internet to changes in democratic governance. Paper presented at the International Political Science Association World Congress, Quebec, August
- Wired News 2000 Africa One project targets 2002. 5 June
- World Bank 2000 *World Development Indicators 2000*. www.worldbank.org/data
- World Economic Forum 2000 *From the Global Digital Divide to the Global Digital Opportunity: Proposals submitted to the G-8 Kyushu-Okinawa Summit 2000*. www.ceip.org
- Wresch W 1996 *Disconnected: Haves and Have-nots in the Information Age*. Rutgers University Press, New Brunswick, NJ
- Yourdon E 1996 *The Rise and Resurrection of the American Programmer*. Prentice-Hall, New York

P. Norris

Political Culture

At the broadest level, political culture refers to the meanings attributed to politics. All aspects of meaning, subjective and intersubjective, related to politics are included in this broad perspective—including the definition of what constitutes or determines the boundaries of the political. A central concern is the role of political culture in the transformation of power into authority, and in challenges to authority. Political culture constitutes the socially constructed and tenuously shared meanings which endow or challenge legitimacy in the political institutions, offices, and procedures of a polity. At the cognitive level, political culture defines and explains or questions the rules of the political game, and at the normative level it justifies or denies the authority of those who play it and of the institutions within which the players interact. Political culture is employed to establish or contest the legitimate boundaries of collective identity. It defines a 'we' against a 'them' and marks the centrality and marginality of various groups within the social unit. At any given historical period, a political culture will likely be characterized by a fairly limited repertoire of particularly salient themes that express the values and goals most sacred to the collectivity through key symbols, myths, rhetorical strategies, and rituals accessible through the analysis of discourse. Most contemporary approaches to political culture consider a consensus on values unlikely to occur and unnecessary for the viability of polities. The public need only resonate to common symbols to which they may ascribe radically different meanings. Even in the best of circumstances the shared meanings of political culture are highly tenuous. Like a fragile spider's web, its symbolic themes are stretched in multiple directions through contested interpretations. Political culture, for all its fragility, is essential for the survival of society and civilization.

1. The Politics of Culture and the Culture of Politics

Contemporary approaches to political culture derive from two Western intellectual traditions. One emphasizes politics—viewing it as a neutral adjudicator of conflicts including those in the realm of culture. The second tradition, descending from Montesquieu and de Toqueville, views politics and governance as cultural constructs. The first tradition is expressed by Harold Lasswell's (1960) inquiry into who gets what, when, and how. The second is illustrated by Michel Foucault's (1966) concern with multiple ways in which power is implicated in the constitution of all areas of social life. In this second tradition the definition of the political and the location of its boundaries is framed as a fundamental cultural question. Political culture

encompasses both approaches—the politics of culture, and the culture of politics. Although the term ‘political culture’ is used primarily by political scientists and historians (Gendzel 1997), interest in the relationship between politics and culture—including the investigation of the extent to which each constitutes an autonomous realm—is of interest in fields in which the practitioners use the term less frequently; for example, anthropology, sociology (Swidler 1984), and cultural studies.

2. *Components of Political Culture*

Symbols are the threads from which the web of culture is spun. Sherry Ortner (1973) distinguishes between elaborating and summarizing symbols. The latter, deriving their power primarily from their emotional resonance, tend to be sacred (in the broad sense)—acting as objects of reverence. The national flag is treated with such reverence as to constitute a fetish in American civil religion. The Western Wall in Jerusalem, a place of sacred centrality in Israeli political culture, and the Black Madonna of Czestochowa in Poland constitute powerful summarizing symbols.

Elaborating symbols work in the opposite direction. They provide vehicles for sorting out complex and undifferentiated feelings and ideas, thereby making them comprehensible, communicable, and translatable into orderly action. Given their cognitive nature, they play a more dominant role in ideology. Root metaphors have conceptual elaborating power. Their value derives from providing categories for comprehending order in the world. ‘Exile’ from and ‘return’ to the ‘Holy Land’ provide the root metaphor of Zionism—the dominant discourse (master narrative) of Israeli political culture (Aronoff 1989). Key scenarios are specific guides that provide maps for socially valued action. The Horatio Alger myth is a such a guide to achieve the idealized American success story.

The myths that endow power with legitimacy are challenged constantly by iconoclasts who call them into question and undermine them, thereby tearing the web of cultural meanings. Old truths are constantly reinterpreted as political realities change, and new cultural traditions are invented by myth-makers who repair the web and drape emergent groups with the mantle of authority. Political myths are narratives which rely on the nonrational for their emotive power. Myths present more lucid and compelling images than do more abstract principles. The more compelling the image, the more powerful the myth’s appeal. Because a myth makes hidden connections between disparate ideas based on unconscious premises, public discourse in mythical form is relatively immune from rational criticism that points out its logical contradictions, or evidence which disconfirms it. When such a pattern of discourse dominates a polity, it is usually a powerful force constraining political change. Yet, political myths also play crucial roles in times of crisis and can

be used—especially in rituals through which myths are enacted—to achieve even revolutionary change (Kertzer 1988). Solidarity and the Catholic Church utilized the Pope’s visit to Poland effectively to orchestrate public rites and political rituals which undermined the hegemony of the Polish People’s Republic (Kubik 1994). Political rhetoric, as in the case of the Pope’s sermons, also constitutes an important component of political culture.

Collective memory and amnesia are molded and expressed through invented traditions which are revitalized cultural forms, such as nationalism, portrayed as continuations of more ancient customs. These elements of political culture articulate a society’s self-definition and destiny. Political cultures are effectively reified when the ‘reality’ they define becomes taken for granted by their publics as natural or divinely ordained. This ‘reality’ explains and justifies (or challenges) ideologically the allocation of resources, prestige, and power in the polity. Ideologies are more concrete and rationalized versions of the general political culture articulated by groups to give legitimacy to their identities and to justify their goals. The endurance of dominant (hegemonic) political cultures at any point in time varies depending on their effectiveness in providing models for dealing with societal problems. Political cultures are dynamic because they are embedded in changing historical, social, economic, cultural and political environments that shape them, and which are in turn shaped by them. The relationship between these variables is dialectical, making the determination of causality highly problematic.

Political cultural divisions can be understood as competing temporal and spatial perceptions. Contested interpretations of myths of origin and aspects of history forgotten and remembered selectively reveal contrasting imaginings of political community. Analysis of the relationship between the cultural construction of past, present, and future, and political ideology reveals the link between contested collective memories and competing visions of nationhood. Temporal and spatial perceptions vary within and across political cultures. Social and political boundaries are related to, but not always identical to, actual geographic frontiers of the state (Lustick 1993). Alterations of the physical borders of a state, filtered through temporal cultural frames, shape citizens’ senses of personal and collective security, which in turn influence their political sentiments and positions.

3. *Disciplinary and Interdisciplinary Approaches*

At the apex of the post-World War II behavioral revolution, the first systematic multinational comparative study of political culture based on the new research method of sample surveys was published. Gabriel Almond and Sidney Verba’s *The Civic Culture* (1963) had a major impact on the comparative study

of political culture, leading to replications, adaptations, extensions, and criticisms of both theory and methodology. By the 1970s and early 1980s, the role of the political culture approach had diminished considerably in political science. Since the mid-1980s the concept has regained stature, during a time when the role of culture in the social sciences and humanities has undergone far-reaching reinterpretation.

At the analytical level, rational modes of thought can be attributed to the cognitive dimension of political culture, and nonrational modes relate to the normative. Most political scientists, including some students of political culture, assume the rationality of political actors and essentially ignore the nonrational dimension—with notable and important exceptions. Rational choice and formal modeling approaches have strengthened this bias, making many political scientists unreceptive to cultural approaches. Reactions against attempts to make the rational actor paradigm hegemonic in the discipline, however, and the resurgence of identity politics—most dramatically in the manifestations of xenophobic nationalism and ethnic cleansing—have contributed to the renewed interest in the role of political culture. David Laitin (1998), who shows how political culture is part of an equilibrium in a coordination game, is one of several scholars who combine rational choice and political cultural approaches. The revival of interest in political culture is most apparent in the field of comparative politics (Ross 1997), but the concept has also gained importance in international relations, particularly in the areas of conflict resolution and strategic culture (Katzenstein 1996).

The two major approaches to political culture are semiotic (via anthropology) and cognitive (via social psychology). Richard Wilson (2000) subdivides them into five approaches. Hermeneutics, influenced by the work of Clifford Geertz, is an important bridge between traditional and postmodern approaches. The interpretative semiotic approach (characterized in the introduction to this article) has been very influential in the social sciences and humanities—the boundary between these two having become increasingly blurred in recent decades. James C. Scott (1990) utilizes this approach to investigate the boundaries of the political. The culture theory approach builds upon the work of Mary Douglas. Developed by Michael Thompson, Richard Ellis and Aaron Wildavsky (1990), it postulates a universal division of politics into four ideal-type subcultures based on two criteria. Analysis focuses on the contestation within politics between competing subcultures, and explains change in terms of their different competencies and biases.

Social character is the third approach influenced by anthropology, but it also draws heavily from social psychology. It emerged from studies of culture and personality, and national character. Personality variables are analyzed in terms of types of significant value orientations, including attitudes toward and depen-

dence upon authority, fear of disorder, and various religious and ideological positions. These studies generally are based on the analysis of the relationship between norms and behavior in specific cultural contexts. Although a few scholars define types of broad civilizations, most, like Robert Putnam's (1993) influential study of civic traditions in Italy, focus on selected case studies. The social learning approach, based more directly on psychological theories, utilizes social surveys to analyze values and beliefs to explain the role of socialization in determining values and behavior. Ronald Inglehart's (1997) work on the relationship between generational change and post-materialist values illustrates this approach well, following the tradition of *The Civic Culture*. Other cognitive approaches incorporate Piaget's theories and conceptualize value orientations very differently from social character studies. Analysis focuses on the reciprocal relationship between individual development and social life. In some of this work moral reasoning and development is at the center of analysis.

Ironically, although political science borrowed the concept of culture from anthropology, contemporary perspectives in anthropology challenge the utility and integrity of the concept, and either avoid the term culture or bracket it with quotation marks to signal its problematic nature (Brightman 1995). Postmodern approaches deconstruct political culture by defining all cultural phenomena in terms of domination and contestation for power. Many practitioners of cultural studies use euphemisms for culture to explore the relationship between power and the imaginative expression of culture. Much innovative work on political culture crosses disciplinary boundaries.

4. *Changes in Focus and Recent Developments*

Contemporary understanding of political culture builds on the foundations of Karl Marx's notion of reification, Max Weber's analysis of legitimation, and Michel Foucault's insights into the cultural definitions of the political and his analysis of power in institutional discourses. Many scholars utilize Antonio Gramsci's conceptualization of hegemony, Eric Hobsbawm and Terence Ranger's formulation of the invention of tradition, Pierre Bourdieu's notions of habitus, doxa, and cultural capital, which reveal how power is inscribed in the scripts of everyday life, and Jurgen Habermas's concept of the public sphere. Feminist theory has also influenced contemporary formulations by calling attention to the gendered aspects of political culture. Volumes edited by Nicholas B. Dirks, Geoff Eley, and Sherry B. Ortner (1994) and by Victoria E. Bonnell and Lynn Hunt (1999) are excellent examples of such contemporary, cross-disciplinary approaches. Their appearances in two of the half dozen important interdisciplinary series on politics and culture established by major university presses are indicative of this trend.

Although there are important exceptions to these broad generalizations, traditional and contemporary approaches can be contrasted in several major respects. Traditional approaches define political culture in terms of attitudes and values, whereas more contemporary approaches view culture in terms of scenarios and discourses. The former conceive of political culture as reified, holistic, discrete, clearly bounded, coherent systems; while the latter view boundaries as fragmented, tenuous, and contested. Older approaches view traditions as objective or 'natural' and assume continuity, while newer approaches perceive traditions as 'invented' social constructions and assume they are dynamic because actors have (limited) options. Some traditional approaches were ahistorical, while most contemporary approaches ground analysis in historical contexts. Contemporary students of political culture examine the mechanisms of transmission of political culture and strategies of boundary maintenance. Traditional approaches focus on goals, while contemporary approaches focus on strategies of action and the incompatibility of strategies of various actors. The analysis of limited repertoires of cultural options available to actors seeks to conceptualize the dialectic of system and practice.

5. Current Areas of Research and Future Directions

An important area of research is state and society relations (Migdal 1989). The disintegration of the Soviet Union and the emergence of the new states of Central and Eastern Europe give renewed currency to interest in civil society, democracy, state and institutions, political communication and meaning, and collective action (Berezin 1997). The ongoing attempt to determine whether there may be certain cultural prerequisites for democracy has been chastened by postmodern and postcolonial critiques of the Western biases in such investigations. More culturally informed reformulations of traditional areas have resulted in the study of New Social Movements and the New Institutionalism (Powell and DiMaggio 1991).

Interest in civil religion and in the relationship between religion and politics (including fundamentalism) has increased. Benedict Anderson's (1991) notion of the nation as an imagined community is one of the many publications on collective identity which, in addition to nationalism, examine ethnicity, race, gender, and sexual preference. Studies of political rituals, rhetorical strategies, strategic discourse, commemorations, museums, political funerals—including the reburial of the remains of previously buried 'heroes'—are areas of considerable innovation. The study of the cultural dimension of processes of globalization and localization (including migration) is another frontier for the study of political culture from its increasingly interdisciplinary perspective. The con-

ceptual threads connecting the different approaches to political culture are as tenuous as the webs of meaning constituting political culture. Yet, the fragility of both does not lessen their importance.

See also: Civic Culture; Civil Society, Concept and History of; Democracy; Democratic Theory; Democratic Transitions; Globalization: Political Aspects; Ideology: Political Aspects; Legitimacy: Political; Political Protest and Civil Disobedience; Power: Political; Race Relations in the United States, Politics of; Rational Choice in Politics; Religion and Politics: United States; Socialization: Political; Zionism

Bibliography

- Almond G A, Verba S 1963 *The Civic Culture: Political Attitudes and Democracy in Five Nations*. Princeton University Press, Princeton, NJ
- Anderson B 1991 *Imagined Communities: Reflections on the Origin and Spread of Nationalism*. Verso, New York
- Aronoff M J 1989 *Israeli Visions and Divisions: Cultural Change and Political Conflict*. Transaction, New Brunswick, NJ
- Berezin M 1997 Politics and culture: A less fissured terrain. *Annual Review of Sociology* 23: 361–83
- Bonnell V E, Hunt L (eds.) 1999 *Beyond the Cultural Turn: New Directions in the Study of Society and Culture*. University of California Press, Berkeley, CA
- Brightman R 1995 Forget culture: Replacement, transcendence, reflexification. *Current Anthropology* 10: 509–46
- Dirks N B, Eley G, Ortner S B (eds.) 1994 *Culture/Power/History*. Princeton University Press, NJ
- Foucault M 1966 *Les Mots et les choses*. Editions Gallimard, Paris [1970 *The Order of Things: An Archaeology of the Human Sciences*. Random House, New York, NY]
- Gendzel G 1997 Political culture: Genealogy of a concept. *Journal of Interdisciplinary History* 27: 225–50
- Inglehart R 1997 *Modernization and Postmodernization: Cultural, Economic and Political Change in Forty-three Societies*. Princeton University Press, Princeton, NJ
- Katzenstein P J (ed.) 1996 *The Culture of National Security: Norms and Identity in World Politics*. Columbia University Press, New York
- Kertzer D I 1988 *Ritual, Politics, and Power*. Yale University Press, New Haven, CT
- Kubik J 1994 *The Power of Symbols Against the Symbols of Power: The Rise of Solidarity and the Fall of State Socialism in Poland*. Pennsylvania State University Press, University Park, PA
- Laitin D 1998 *Identity in Formation: The Russian-speaking Populations in the Near Abroad*. Cornell University Press, Ithaca, NY
- Lasswell H 1960 *Politics: Who Gets What, When, How*. Meridian Books, New York
- Lustick I S 1993 *Unsettled States, Disputed Lands: Britain and Ireland, France and Algeria, Israel and the West Bank–Gaza*. Cornell University Press, Ithaca, NY
- Migdal J 1989 *Strong Societies and Weak States: State–Society Relations and State Capabilities in the Third World*. Princeton University Press, Princeton, NJ
- Ortner S B 1973 On key symbols. *American Anthropologist* 75: 1338–46

- Powell W, DiMaggio P J (eds.) 1991 *The New Institutionalism in Organizational Analysis*. University of Chicago Press, Chicago, IL
- Putnam R 1993 *Making Democracy Work: Civic Traditions in Modern Italy*. Princeton University Press, Princeton, NJ
- Ross M H 1997 Culture and identity in comparative political analysis. In: Lichbach S, Zuckerman A S (eds.) *Comparative Politics: Rationality, Culture, and Structure*. Cambridge University Press, New York
- Scott J C 1990 *Domination and the Arts of Resistance: Hidden Transcripts*. Yale University Press, New Haven, CT
- Swidler A 1984 Culture in action: Symbols and strategies. *American Sociological Review* 51: 273–86
- Thompson M, Ellis R, Wildavsky A 1990 *Culture Theory*. Westview, Boulder, CO
- Wilson R W 2000 The many voices of political culture: Assessing different approaches. *World Politics* 52: 246–73

M. J. Aronoff

Political Discourse

1. Introduction

Discourse in general is a way of organizing human experience. It establishes frames of meaning by the recounting and interpreting of events and situations. It constructs systems of order. Political discourse applies such frames to the exercise of power—including principles of hierarchy, representation, and accountability. It deals with the narrative interpretation of events and ideas, logical and mythic, and establishes criteria and contexts for comparing and evaluating political systems. In these terms, political discourse is as old as politics, its pedigree going back at least as far as Plato for whom discourse took the form of a *dialectical logic*. Plato used this method to derive a model system composed of appropriate connections between truth, knowledge, virtue, and justice that he used to evaluate and rank different polities. Plato made explicit what might best be referred to as the relationship between discourse as rationality and rationality as discourse. But he also recognized that there were limits to the role and place of rationality in human affairs. Where the rational leaves off, he invoked a revelatory higher power, religious, mythic, magical, numerological—a Pythagorean idea of myth, (Cassirer 1946). One can say that Plato defined political discourse in terms of a *mytho-logic* that he regarded foundational for the functioning of any political community. For Plato the state, or better, the city, is a discourse community.

For Aristotle political discourse was an emergent teleological process embodying final causes, the shape, purpose, and meaning of the polity. This established a basis for a purposeful history of ideas—a narrative realizing itself in the evolution of democracy of which

there are several versions. One starts with Athenian democracy as a defining and classic polity, Rome representing natural and positive law (with interpretive exegesis by Glossators and Commentators). The Schoolmen raised the issue of revelation, ecclesiastical law vis-à-vis natural law. Medieval corporatism established the principle of conciliarism (and the kings ‘two bodies’) (Kantorowicz 1981). Social contract theorists, from Hobbes, Locke, and Mill, defined a political spectrum, which at one end was autocratic, the other democratic. Another version uses ages: classic, ecclesiastical, enlightenment, romantic, democratic. The end of history is modern, secular, democratic government. Which makes a better story than a good history.

Although hardly a corpus or a stable body of knowledge, modern political discourse theory is composed of a variety of concepts and ideas from many disciplines. As a form of critique it tries to penetrate below the surface of the good stories that people tell themselves about politics. Discourse theories may differ but they are all concerned with what it is that people come to believe about power and its exercise, and how they come to think that way. Political science and sociology analyze different ideologies, dogmas, and belief systems in terms of their content and the extent and intensity of their support to understand threats and opposition to democracy and democratic principles.

As with Durkheim’s treatment of social facts, political beliefs are no different from any political perceptions and preferences, loyalties and affiliations. Examining them requires operational techniques, quantitative analysis. Belief systems are amenable to survey methods, questionnaires, and statistical techniques (Lane 1962, Converse 1964; Putnam 1973, Tilly 1978, Sidney Tarrow 1994).

In contrast, several other disciplines break up concepts like ideology, belief systems, or culture into components on the grounds that they have become ‘overkill’; that by explaining too much they explain too little. How political discourse establishes order and control and organizes and mobilizes social groups and organizations requires more differentiated categories and more interpretive knowledge. How is it that words can kill? When does ecclesiastical belief lead to murder in the cathedral? How much was Nazi ideology the cause of the Holocaust? How and when does ethnicity, religion, language, or other differentiating social criteria that establish exclusivity define ‘others’ in negative terms—the extreme consequence being genocide? Similarly, when do such circumstances reverse themselves? When do the same defining characteristics constitute pluralism, multiculturalism, or diversity? How does confession within the framework of legal discourse connect guilt with justice, and by so doing change the contours of the permissible? How was Apartheid in South Africa, with its biblical racism, its Boer myths of a chosen people replaced by a Truth

and Reconciliation Commission? These and similar questions have led to a virtual explosion of concepts, categories, and theories: structuralist, phenomenological, linguistic, psychological.

Despite differences among approaches, all favor the analysis of events as more than chance, i.e., as coherent and meaningfully and internally connected. In terms of political discourse theory, events are *social texts*. They can be read as *language* including words, of course, what people say about each other and their condition, but also as *signs, symbols, rhetoric*—as *semiotics*. Discourse theory addresses the problem of how best precisely to read events. It is less concerned with the content of particular ideologies than ‘the content of the form’ (White 1987) and the ‘narrative construction of reality’ (Bruner 1991).

Political discourse theory occupies a position on a continuum, the opposite end of which is political economy and rational choice. Like those approaches, it responds to a calculus of costs and benefits, losses and gains. Unlike them what is exchanged is outside the normal range of interest competition and bargaining—*symbolic* rather than economic capital. If so, then an important purpose of political discourse theory is to identify the coin by which different rationalities are measured.

While the substance of political narratives varies widely, they follow certain fairly standard trajectories, including the *recounting* of events in the form of *retrievals* and *projections*. Events serve as *signifiers* and *metaphors* in which meanings are transmitted in terms of past and similar situations, real or fancied; and as *metonymies* in which the event is a fragment or representation of some larger logical theoretical or revelatory belief system. By this means the past provides evidence for a projected and transcending state of affairs, including millennial outcomes, promised lands, states of grace, or a condition of political virtue or justice. This suggests a central question: when does political discourse lead away from a politics of give and take associated with rational economic behavior and when does it generate *symbolic capital*? (Bourdieu 1977). It is this question that will be discussed here.

First the intellectual pedigree of political discourse will be examined, indicating ways that discourse affects the way people think about politics, act on what they think, and, as a part of such action, how they form discourse communities using symbolic capital, in contrast to those predominantly organized around economic and political exchange followed by a critique and assessment of future trends.

2. An Intellectual Pedigree

Modern concern with political discourse might be said to begin with the *semiotic* revolution. To name only a few contributors (and leaving out forebears like Peirce,

Saussure, Wittgenstein, and others), one might include linguists like Roman Jakobson, Dell Hymes, and Umberto Eco; sociologists like Basil Bernstein, Erving Goffman, and Herbert Garfinkel; philosophers like Alfred Schuetz, John Austin, Ernest Gellner, and John Searle; anthropologists such as Claude Levi-Strauss, Clifford Geertz, Victor Turner, Mary Douglas, and Jack Goody; historians such as Robert Darnton, Pierre Nora, Le Roy Ladurie and of course many others. Political scientists include Benedict Anderson, David E. Apter, and James Scott. In literary theory there is a very large cohort concerned with *narrative* and *interpretation* (Mitchell 1982): Kenneth Burke, Roland Barthes, Hayden White, Frederic Jameson, Terry Eagleton.

Which brings up the question of the range of concepts that such thinkers apply to the analysis of political discourse theory. Walter Benjamin, examining the relation between memory and the crafted story, the potentiality for what might be called the *fictive use of recounting*, implies the significance of the *imaginary real*. Levi-Strauss laid out the structural properties or better, the isomorphic ordering properties of *myth* and *theory* in the form of binaries which literally compose the order of the universe, providing rules of appropriate *intermediation*: between man and the cosmos, men and women, kin and family, earth and sky, land and water, etc. (Levi-Strauss 1955). Similarly in Umberto Eco’s theories of semiotics, with their emphasis on *codes* and *codex, signs* and *signification*, and what he called *sign production*.

In a more literary vein Hayden White refers in his analysis of narrative to the *poetic troping of the facts* in which certain events become in effect what Eagleton called *drenched signifiers*, images and metaphors so symbolically powerful that they become incentives to action. Foucault examines the way certain *discursive codes* establish authority and power. For François Furet political discourse becomes a *nonscientific representational strategy* or a structure of time, or *time consciousness*. Ricoeur emphasizes discourse as *sin, confession, and purification, negative poles and overcoming* and the uses of metaphor. Henri Lefebvre deals with space as *terrain* and jurisdiction as *agora*—a theater of intersection between *addressor* and *addressee*. For Baudrillard an appropriate discourse produces *symbolically dense* miniaturized versions of state and society, encampments, staging areas, fortresses, neighborhoods, which form the moral center for what is yet to be. Guy Debord emphasizes political discourse as theater, drama, *spectacle*.

While each of these and many other thinkers contribute in very different ways to political discourse theory, it is possible to identify one common and underlying concern. People make stories out of events. They do so individually and collectively. Recounting individual stories makes for sociability. Collective stories have political consequences when, as myths they purport to be history, as history they are

reinterpreted as theories, and as theories they make up stories about events. Theories that become stories create *fictive truths*. Since in politics, truth-telling and story-telling are part of the same process, it becomes possible to interrogate the past in order to transform the future.

There are of course many ways to examine such matters. A useful place to begin is with the actual ways and circumstances in which individuals recount their own stories. As a process, this becomes politically interesting if such stories are *collectivized*, systematized, and formed into what have been called *master narratives*. For this purpose *agency* is required, a public figure able to play the special role of 'story-teller' whether in the form of philosopher-kings, cosmocratic figures, politicians, soothsayers, diviners, or others (Benjamin 1969).

Agency, a form of authoritative recounting, begins with some benchmark, a state of grace which then defines a *negative pole*, loss, suffering, struggle, on the basis of projecting an *overcoming project*, a preferred goal. Such recounting is intrinsically *dramatic*—exceptional experiences punctuated by *promontory events*. Events may be tragic, including suffering, torture, betrayal, murder, and death. They can also result in superior *insight*, knowledge through *struggle*, *purification*, and *redemption*. Recounting renders the unimaginable manageable. It builds into the narrative *imaginary reals*, necessary fictions which establish boundaries and borders, affiliations and loyalties, terrains and jurisdictions, defines insiders and outsiders, separates the good citizen from the pariah. The narrative establishes *boundaries* and *clienteles*, identifying relevant *markers*, ethnic, religious, racial, kin, clan, and class, or other categories. It locates *surrogates*, *negative* (but redeeming) *others*, Genet's thief, or homosexual, Fanon's colonial African, Foucault's madman or prisoner, etc. (Foucault 1979).

It deals too with *ritual endowments*, insignia, votary paraphernalia that go with them, flags, and uniforms. It emphasizes the theater of the absurd, or the solemn occasions, the mass, the funeral, the parade, votary ceremonies, candle light parades, pistols, whips, massed guns and technology; anything which permits *symbolic intensification*. Ordinary terrain, transformed into sacred space, serves as metaphors or better *simulacra* for the national patrimony (old and new Jerusalems). Politics is the stage on which leaders perform, recapitulating historical events as theater and choreography (Plant 1992).

Time becomes crucial, especially in *disjunctive* and *transformational moments*, when leaders propose a fresh start. The French revolution tried to reinvent the calendar in its desire to start the world all over again. Time foreshortened becomes space, a *geomancy of space*, enclosure as architecture, the placement of buildings, squares, parks, transportation, homes, meeting places, amphitheaters all taking on new meaning as potential rallying ground, or as Agora's or

forums, staging areas, embattled zones, etc. So too with monuments, statues, shrines, graveyards, jails, public buildings, this pub, that hideout, this square, street, neighborhood. Virtually any ordinary thing can take on significance, disrupting the daily or routine business of commerce, its taken-for-grantedness, in favor of the exceptional moment in the exceptional space (Lefebvre 1986).

Similarly with clothes, uniforms, medals, gestures (verbal and other), hair, Jacobin hats, red flags, insignia, swastikas. These emphasize the body as reference point; exaggerated posture, sexual affronts, cross dressing, racial inversions. The body politic takes on the characteristics of the individual body. It can be violated, penetrated, tortured, mobilized, transformed.

Opportunities for such *recounting* include the founding of new nations and states, revolutionary and redemptive *breaks*, a turning of the world upside down. In effect, political discourse helps produce such breaks. It also capitalizes on them when the opportunity arises. In such moments political discourse creates its own rationality, defining sharply what it is against and what it is for. It contains within itself a disordering effect in rejecting established political beliefs and institutions, and an ordering principle designed to negate and transcend the disordering condition it helped to create.

Political discourse then is bound up with history as conflict and struggle. Typical circumstances are one or other form of *inversion*; *revolutionary*, *ecclesiastical*. A classic device is some variation on Hegel's *master and slave*, or, more redemptive, the *last* becoming *first*. One can find both in the speeches and texts of early church fathers, fifteenth century mystics, and radical political revolutionaries from the Jacobins, to the Levellers and Diggers, not least of all including what Hobsbawm called *social banditry* (Hobsbawm 1959). Such situations lead to exceptional acts of daring and devotion. They involve a tension between *purity* and *danger* such that the words of the recounted narratives resonate.

Recounting is thus a way of both periodizing the past and incorporating it within a series of performances in which singular and distinctive individual events and actions stand out, the whole inspiring urgency, martyrs and heroes not to speak of loyalty, obligation, and what Bernard Yack has called political 'yearning' (Yack 1986).

Narrative is half the process. *Text* is the other. Both are required if political discourse is to become an active ingredient in political life. Required of texts in relation to symbolic capital is that they provide exegetical opportunities. *Textualization*, the transition from *orality* to *writing*, requires not only a narrator but a writer. Both are essential if those documents designated as sacral are to serve as symbolic centers of power, complete with their coterie of interpreters devoted to locating hidden meanings, logical and

magical, thus providing those esoteric nuances that make a discourse community unique and which give it its moral center. Text and coterie, with the agent at the center, suggest truth as insider knowledge. They create an *iconography of words and objects* worthy of worship and respect. So virtually all societies and social movements have certain texts which become canonical, whether they take the form of a constitution or a 'little red book.'

Whether as Marxist texts or canonical or biblical texts, the word as written embodies both logical truths and mythic or fictive ones. Some are enshrined in sanctuaries. Others as authorizing documents are on suitable display (as for example the constitution in the United States or the Magna Carta in Britain). Hence, one might say that political discourse involves the alteration of original meanings by means of mythic retrieval and logical projection, what Roland Barthes and Claude Levi-Strauss referred to as 'mytho-logics' (and what Malinowski called a mythic charter).

Which returns us to the question raised above. How can history as a representation of symbolic events become simply history—exceptional moments becoming unexceptional? Or, to put it differently, how can symbolic capital be converted into normal interest competition and bargaining within a framework of constitutional authority and law? It is a question that comes up in the context of wars, nationalist, separatist, and revolutionary movements, the rise of cults, persistent and organized protest—matters which if anything are more important today than ever before. In this sense one can say that political discourse is about how to change the world by reinterpreting it, and political discourse theories are about how to interpret the interpretations.

3. Applications and Implications

The twentieth century has perhaps been more witness to powerful and consequential political occasions than any previous one, not least of all because of the connection between terror and utopianism. If one includes the power of nationalism, the mobilization of its rhetoric, which accompanied the decline of colonialism and the founding of new states in Asia and Africa, one thinks too of a virtual explosion of political *agents* like Nkrumah, Senghor, Sekou Toure, and Nyerere, leaders all of whom, in the name of anticolonialism, nationalism, and independence, created their own political vernacular texts, doctrines by which they claimed exceptional and continuing political privileges (emulated by their putative descendants). In these terms too political discourse was very much a phenomenon of such populist dictatorships as Peron's Argentina not to speak of De Gaulle's myth of a glorious (and largely fictive) France.

Perhaps the exemplary defining event or historical moment, or disjunctive break, was the Russian

revolution with Lenin its agent, surrounded by a band of intellectuals forming a coterie. In the texts so formed the strong elements are *orality* and *textuality*. Dramatic public spaces were carefully selected for delivery. What began in speeches was translated, that is, words became converted and modified in the form of writings many of which became canonical. As the redemptive and prophetic figure of the revolution, he wrote the bible of revolutionary principles and tactics, not simply for his own followers but subsequent generations and putative descendants. Perhaps the most significant of these disciples was Mao Zedong who 'rewrote' China's history in the form of three narratives: long, intermediate, and short; each fitted inside the other, *contextualizing* in the form of radical revolution and militant nationalism, converting time into immediacy, forming a *fictive/factual* basis for *revolutionary discourse* (Apter and Saich 1994).

Political discourse not only precipitates conflict and foreshortens time but also can extend the disjunctive moment and keep it alive. Examples of conflicts of the long *duree* include the IRA in Northern Ireland, and between the Israelis and Palestinians (where two diasporic populations remain locked together in their hostility). In both cases the parties are divided by their discourses, their retrievals of the past determining their view of the future, and so making mutual accommodation difficult as long as their territorial claims are to the same physical space. Each discourse has the same contours as the other, similar themes of loss and displacement. Each makes salvational, redemptive, and recuperative national claims. For one side to give way to the other is considered betrayal.

There are innumerable examples. The rise of fundamentalism or integralism enables the preaching of the word as *scripture*, something to be pored over and through a process of *exegesis* made to reveal hidden, deeper, and higher meanings. For some such words will be profound, for others ritualized, both appearing to offer magical depth, a common sensibility. By this means, the individual is linked to the state or sect and by multiple references and *reciprocities*. Virtually any categories will do, religion, ethnicity, clan, family, or region, village, town, or cult, political program, ideology, etc. All potentially are *condensed symbols*, complex, with multiple references and signifiers, providing opportunities for *exegetical bonding*—establishing boundaries in the mind that constitute boundaries on the ground.

There are certain recurrent themes: an original and negative condition to be transcended, a mythical past to be redeemed, a frame within which certain primary symbols are characteristic, loss, defilement, sin, stain. Grievance is converted to intentionality, time into space, and space into particular places, symbolically endowed terrains. By this means individual stories are collectivized and conveyed as master narratives to a group as a whole. Hence the narrative process itself creates political leaders acting as agents who validate

preferred outcomes embodied in texts. Such narratives contain a story within a story, the storyteller or agent figuring in his or her own story. Prototypical exilic figures are Odysseus or Moses. Forced to suffer the loss of the patrimony, wandering far from their homes, they learn to outwit enemies. By means of knowledge gained from experience, they regain the power necessary to reclaim the patrimony in the first instance or a Promised Land in the second.

The more cosmocratic the impulse, the larger than life the *principal agent*, the greater the moral density of the text, the more likely are the doctrines to intensify differences, separating believers and nonbelievers; the saved and the damned, divided by a potential millennial accomplishment. Some are privileged at the expense of others. A chosen people will claim superiority, whether by class, ethnicity, religion, nationality, race, or doctrinal persuasion. Each such group becomes both *surrogate* for higher truths and a symbol or representation of them. Loyalty is thus moral obligation, affiliation deep seated, binding groups together. Hence political discourse theory constitutes as its subject matter all those cleavages of human *association*, which especially in a context of violence acquire durable properties reinforced rather than undermined in subsequent events. In these terms, discourse communities generate their own objects, symbolic capital possessing its own *exchange value*.

There are certain serious weaknesses to political discourse theory so considered. It works best in conjunction with intensive case studies in depth. It depends on depth interviewing or political anthropology. It is not very amenable to operational techniques, survey, path analysis, and other statistical methods. Its relativism is less concerned with the content of ideologies, the conventional interest of political scientists, than in the structure of beliefs, what forms them, and how they become adopted. There is also a bias that over time rationality will triumph over nonrationality, logic over revelation, theory over myth, facts over appearances.

However one approaches it, discourse theory is not likely to disappear in the foreseeable future. It has moved in several different general directions: toward *postmodernism* in the connection between *narration and legitimation* (Lyotard 1984); *cultural studies*, with respect to the symbolization of violence and the state (Bhabha 1990); *sociolinguistics*, itself divided into three main strands, psychological, linguistic, and sociological (Giglioli 1982); towards an intermediate position between comprehensive structural theory and the concrete explanations (or justifications) people give for their actions and beliefs—what has been called ‘critical hermeneutics’ (Habermas 1979). Finally, for those interested in politics, power, and rational choice theory, some attention is now being paid to how symbolic capital fits into much the same framework as economic capital, using narrative to show how the two represent different forms of the same rationality (Bates

et al. 1998). Indeed, if ‘money’ is itself the symbolic form of utility *influence* is the consequence of symbolic capital. If so, rational choice theory and discourse theory instead of occupying opposite ends of the same continuum may come to constitute their own hermeneutic circle.

See also: Discourse and Identity; Discourse, Anthropology of; Linguistic Turn and Discourse Analysis in History; Metaphor and its Role in Social Thought: History of the Concept; Narrative, Sociology of; Narratives and Accounts, in the Social and Behavioral Sciences; Political Communication; Political Culture; Semiotics

Bibliography

- Anderson B 1983 *Imagined Communities*. Verso and New Left Books, London
- Apter D E, Saich T 1994 *Revolutionary Discourse in Mao's Republic*. Harvard University Press, Cambridge, MA
- Bates R et al. 1998 *Analytic Narratives*. Princeton University Press, Princeton, NJ
- Baudrillard J 1981 *Simulacres et simulation*. Editions Galilee, Paris
- Benjamin W 1969 The storyteller. In: Arendt H (ed.) *Illuminations*. Schocken Books, New York
- Bhabha H K (ed.) 1990 *Nation and Narration*. Routledge, London
- Bourdieu P 1977 *Outline of a Theory of Practice*. Cambridge University Press, Cambridge, UK
- Bruner J A 1991 The narrative construction of reality. *Critical Inquiry* 18(1): 1–21
- Cassirer E 1946 *The Myth of the State*. Yale University Press, New Haven, CT
- Converse P A 1964 The nature of belief systems in mass publics. In: Apter D E (ed.) *Ideology and Discontent*. The Free Press, New York pp. 206–61
- Debord G 1987 *La société du spectacle*. Gerard Lebovici, Paris
- Foucault M 1979 *Discipline and Punish*. Random House, New York
- Geertz C 1973 Ideology as a cultural system. In: Geertz C (ed.) *The Interpretation of Cultures*. Basic Books, New York pp. 47–76
- Giglioli P (ed.) 1982 *Language and Social Context*. Penguin, London
- Habermas J 1984 *The Theory of Communicative Action*. Beacon Press, Boston
- Hobsbawm E S 1959 *Primitive Rebels*. University of Manchester Press, Manchester, UK
- Jameson F 1972 *The Prison-House of Language*. Princeton University Press, Princeton, NJ
- Kantorowicz E H 1981 *The King's Two Bodies*. Princeton University Press, Princeton, NJ
- Lane R E 1962 *Political Ideology*. Free Press, New York
- Lefebvre H 1986 *La production de l'espace*. Editions Anthropos, Paris
- Le Roy Ladurie E 1979 *Carnival in Romans*. George Braziller, New York
- Levi-Strauss C A 1955 The structural study of myth. *American Journal of Folklore* LXXVIII(270): 428–44

- Lyotard J-F 1984 *The Postmodern Condition: A Report on Knowledge*. The University of Minnesota Press, Minneapolis, MN
- Mitchell W J T 1982 *The Politics of Interpretation*. University of Chicago Press, Chicago
- Nora P (ed.) 1996 *Realms of Memory*. Columbia University Press, New York, 3 Vols.
- Plant S 1992 *The Most Radical Gesture*. Routledge, London
- Putnam R 1973 *The Beliefs of Politicians*. Yale University Press, New Haven, CT
- Ricoeur P 1968 *The Symbolism of Evil*. Harper and Row, New York
- Tarrow S T 1998 *Power in Movement*. Cambridge University Press, Cambridge, UK
- Tilly C 1978 *From Mobilization to Revolution*. Addison Wesley, Reading, MA
- White H 1987 *The Content of the Form*. Johns Hopkins University Press, Baltimore, MD
- Yack B 1986 *The Longing for Total Revolution*. Princeton University Press, Princeton, NJ

D. E. Apter

Political Economy, History of

The term 'political economy' first appeared during the early seventeenth century in French texts that considered the contribution that trade, agriculture or manufacturing might make in augmenting the powers of the sovereign and rendering their subjects more prosperous. It is linked to the rise of the early modern state: it assumes that the interest of the monarch lies in a strong state, and that the path to a strong state leads via the augmentation of a population, and the extension of their means of subsistence. The interests of sovereign and subject therefore coincide—the strength of a state is founded upon the well-being and numerosness of its population. During the seventeenth and early eighteenth centuries an extensive literature on the contribution of trade and agriculture to the welfare of people and ruler developed in the English, French and German languages. The diverse practical propositions encountered in such writings were systematized in the course of the eighteenth century, first through the work of François Quesnay (1694–1774) and the Physiocratic 'School' with which he was associated, and subsequently through Adam Smith's (1723–1790) critique of Physiocratic and 'mercantilist' schools of 'political economy.' Smith's *Wealth of Nations* (1776) provided the foundations for a new 'principled' science of political economy, the chief architects of this new science being Thomas Robert Malthus (1766–1834) and David Ricardo (1772–1813). The 'classical political economy' which these and their contemporaries represented elaborated a limited set of economic principles turning on the proposition that, ultimately, the exchange value of a commodity was reducible to

the amount of labor embodied in it. Price, therefore, inhered in the good, rising or falling according to the cost of labor. By the 1830s an independent American literature was developing on this basis, and in the final two decades of the century Dutch, French and English texts began to be read in Japan. In continental Europe Jean-Baptiste Say's (1767–1832) reception of Smithian principles was extremely influential; his writings laid the basis for the transformation of 'political economy', during the last third of the century, into a new 'economics' in which value was a product of subjective choices, and price the outcome of the degree of scarcity which these choices created.

1. The Derivation and Evolution of the Term 'Economy'

Although the discourse of political economy first emerged in early modern Europe, the term derives from Aristotelian beliefs concerning social and political order. Aristotle's *Politics* outlined two forms of human association: the *oikos* and the *polis*. The first, the household, was instituted for the satisfaction of private daily needs; the second, the domain of political association, structured public life. Aristotle introduced this distinction in response to Plato's argument that the art and technique of government was everywhere the same; much of *Politics* is therefore taken up with the different ways in which public life could be structured. By contrast, only one form of organization was foreseen for the household, whose head ruled over its members like a monarch over its subjects, securing the proper employment and maintenance of property and persons. Elsewhere in his *Nicomachean Ethics* Aristotle had argued that the virtuous man set no store by wealth, hence it could be no part of virtue to deliberately increase the wealth of a household. The purpose of the household was therefore the provision of an ordered material existence to its members through the allocations made by the household head. Beyond the household, in the public domain, merchants engaged in trade and money-getting, but this was activity governed by no definite conceptions of virtue or order, subordinated to the superior principles of government.

Vestiges of this way of conceiving social and political relations persisted into the eighteenth century, especially in English, French and German treatises which adumbrated the proper management of farm and field. However, in the early seventeenth century 'économie politique' adopted the central householding tenet of the Aristotelian model outlined above, projecting it into the public domain, hence creating a new 'political oeconomy'—which sought to identify the most expedient manner in which the interest of ruler and ruled might be reconciled within a state conceived as a giant household. Although this

naturally involved proposals for the augmentation of population and the means of their subsistence, the pursuit of wealth for its own ends had no definite place in this new order. In France, *commerce* was the term assigned to activity related to trade, finance, and money; literature related to this aspect of economic activity developed quite separately to that of a political oeconomy primarily focused on agricultural, subsistence activities.

2. The Sources of Power and Plenty

In the context of a European economy recovering from the ravages of the Thirty Years War, vulnerable to epidemics on the one hand and periodic poor harvests on the other, the priority given to agriculture as the foundation for any increase in numbers and welfare requires little elaboration. Overseas trade could make a small country like Holland rich, but both Britain and France were simply too large for trade to have such an impact. When in the early nineteenth century David Ricardo outlined the benefits of international trade he chose as his examples English wool against Portuguese wine—products that both countries had in abundance, of course, but where the increase in welfare consequent upon exchange had little impact upon the central issue of subsistence. And in fact the overseas trade of Britain throughout the eighteenth century facilitated the import of sugar, tea, coffee and spices, luxury goods whose significance in turn became part of a debate on the implications of the production and consumption of such goods. Not until the end of the century did grain imports exceed exports, and this reflected both an increasing population and a shift in the structure of production that formed the origins of the industrial revolution. This new pattern of international trade—in which domestic employment and a rising population increasingly depends on the export of manufactured goods in return for which food and raw materials are imported—first emerged in the nineteenth century.

States with ready access to the sea—Spain, France, Britain, and the Hanseatic cities—had the opportunity to develop overseas trade and fisheries, and an associated literature developed which both urged the extension of trade as the path to wealth, and outlined the practicalities of successful trade and commerce. Where access to the long-distance trade that the oceans offered was limited, arguments that herein lay the route to wealth and power had little force. Moreover, the profitability of merchant ventures seemed to depend upon risk and chance, not upon the wise investment of capital, the proper management of resources, and the judicious use of labor that typified a flourishing agriculture. The rhythm and progress of production in this sphere of economic activity lent itself to the conceptual framework of political

oeconomy and could without serious modification be extended to manufacturing activities. German texts of the eighteenth century routinely countered *Landwirtschaft* to *Stadtwirtschaft*, the town economy to the country economy, signifying manufacture in the former and agriculture in the latter—even though much early manufacturing actually took place in rural locations. But the contrast set up in this manner simply excludes serious consideration of trade and finance as significant elements in the welfare of the state.

3. The Enlightenment Science of Political Economy

In the mid-eighteenth century a ‘new science of political economy’ was created, based upon a conception of the ‘natural order’ of society in which agricultural activity was the unique source of a ‘net product.’ The systematic analysis of production and circulation that was developed by the so-called ‘Physiocrats’ between 1760 and 1780 owed much to contemporary enlightenment rationalism, but it was also prompted by the growing problems of the French economy, and the increasingly parlous nature of the state finances.

Physiocratic political economy identified the natural foundations of economic activity, and at the same time implied that proper management of the economy required that the ruler not ‘govern too much,’ that excessive interference would impede the formation of wealth. The source of wealth was agriculture, whose net product circulated through the various classes of society; the greater this net product, the stronger the circulation and hence the greater the wealth. Improvements required investment, which could only take place if the net produce was not in the course of its circulation diverted into unproductive ends, such as the production of luxury goods for landed proprietors. Special taxes and duties were impediments to the free circulation of goods; and since all taxes were ultimately funded by the net product, all state revenue should be drawn directly from the sole source of this net product, agriculture.

Quesnay illustrated his ideas with a circulatory diagram, the *Tableau Économique* (1758–9), depicting the flow of net product from its agricultural source, through the hands of the landlords, thence to the ‘sterile class’ of manufacturers, and then, through the agricultural purchases this last group made, back again to the source. This clear diagrammatic representation of three economic classes and the flows of goods and payments between them is now seen as the original conceptualization of circular flow in an economy, but the originality of this drew little contemporary attention. Criticism of the Physiocratic system isolated two central tenets: the proposition that

non agricultural employment was unproductive, and that a single tax upon agriculture was viable, practically or intellectually. The most prominent criticism of this system can be found in Smith's *Wealth of Nations*.

4. The Smithian Critique of the 'Mercantile' and 'Agricultural' Schools

In Book IV of *Wealth of Nations* Adam Smith arranged his intellectual forbears into two (fatally flawed) schools of thought. The first of these was the 'commercial, or mercantile system,' which, he argued, sought to enrich one state by impoverishing others, imposing duties and a range of restraints of trade designed to attract money and gold into the land. In part a rhetorical exaggeration designed to highlight his own 'system of natural liberty,' and in part a willful misreading of an extensive and varied literature more sophisticated than he would allow, his remarks laid the foundations for a critique of 'mercantilist' policy that has survived up to the present day.

The other school to which Smith addressed critical remarks was the 'agricultural' school. He rejected the idea that agricultural labor was the sole source of wealth, and that other labor was unproductive; but he warmly supported the Physiocratic representation of the annual flow of goods as the source of wealth, and the proposal that this flow should be as little impeded as possible. These were indeed central principles of Smith's own work, together with a redefinition of what counted as 'productive' as against 'unproductive' labor.

Smith defined the 'wealth' of a nation not as an accumulation of gold, nor as simple populousness, but as the annual produce of its (productive) labor. Productive labor was that embodied in capital; it left a physical trace, unlike a service (no matter how useful or vital to the state that service might be). This could be augmented through the extension of the division of labor; labor could be made more productive by subdividing tasks and rationalizing the production process. There were two provisos attached to this: the first was that this process of the increasing division of labor had greater potential in manufacturing than in agriculture; and the second was that the degree to which it could develop was limited by the extent of the market. From this it follows that the most rapid increase of national wealth would result from the augmentation and rationalization of labor applied to manufacturing, within an economy which was both internally and externally freed from impediments to trade.

Smith was writing in the context of an economy whose structure and rhythms were dominated by agricultural production. This was not only the major economic sector; it was the immediate source of food

and was prone to annual fluctuations in output and associated variations in prices. Since foodstuff represented such a major part of expenditures on the part of a majority of the population, the achievement of a low and stable price for grain was a crucial political problem throughout the eighteenth century for European governments.

Broadly speaking, there were two opposed solutions to this question. The first option was to encourage home production of grain by protecting domestic producers: high tariffs on imported grain would prevent the flooding of the market with cheap food in good years, undermining home production which would then be unable to provide adequate supplies whenever the sources of foreign grain diminished, or even dried up altogether. This option therefore secured whatever the level of home produce might be to domestic consumers, albeit at a high price. The second option also encouraged home production, but did so by permitting free import and export of grain. According to this, the availability of a market greater in extent than that offered by the national economy would stimulate domestic producers, consequently rationalizing their activities so that cheap foreign grain would not appear to be such a constant threat to domestic production. This second option proposed that the domestic population would be best served by a supply of cheap food based not upon the protection and control of the home market, but rather upon their access to an international market in foodstuffs. Smith's argument that the division of labor was a powerful engine of economic growth, and that it was limited only by the extent of the market, was therefore an argument for the benefits of free trade in all goods, augmenting the welfare of all states through the extension of the international economy.

5. The Principles of Classical Political Economy

Malthus' *An Essay on the Principle of Population* (1798) was conceived as a critique of contemporary conceptions of human betterment, and it certainly played a part in discrediting the conventional association of large populations with economic wealth and national progress. But his demographic arguments presupposed a closed agricultural economy, enjoying none of the benefits of commerce outlined by Smith. Some five years before the publication of the first edition of the *Essay* Britain had, however, become a net importer of grain. Henceforth, exported manufactures would pay for the imports of food required to subsist a large, increasingly manufacturing, population. Expanding (free) trade and commerce was to provide the solution for the 'Malthusian problem.'

Nonetheless, during the first two decades of the nineteenth century the elaboration of the classical theories of production and distribution presupposed

the conditions of an agricultural economy, producing a staple crop through the joint contribution of laborers, farmers and landowners. These three economic classes received wages, profits and rent, respectively, and the question which political economists sought to resolve was the relation between price formation and the distributive shares received by each class. Smith had assumed that prices were simply the sum of distributive shares; political economists sought a systematic explanation of the level of each of the elements within a framework that supposed that human labor was the sole creator of value.

Debate over Britain's Corn Laws during 1814–5 provided the occasion for the systematic linkage of the pattern of domestic income distribution to the trade regime. Malthus composed three pamphlets during this period which argued the benefits of limited protection, and suggested that rising rents were a reflection of human progress, since he regarded rent as a payment made subsequent to the settlement of wages and profits. Ricardo in his own pamphlet, known as the *Essay on Profits*, dissented from Malthus' protectionist sentiment and assumption that rising rents were part of the progress of wealth. Since he assumed that wages were fixed, he argued that there was an inverse relationship between rent and profits. High rents, therefore, meant low profits, they furthermore meant high prices for grain and hence increased the wages that manufacturers must pay their own workers. Since he also assumed that the rate of agricultural profit determined the rate of profit in general, protection for the landlord and high rents depressed the rate of profit in the economy as a whole. The solution to this prospect, he argued, was free trade in grain, which would force agricultural rationalization and lower the average price of grain, the basic foodstuff for the majority of the population. Ricardo's *Principles of Political Economy and Taxation* (1817) elaborated these points, and established a core of theoretical propositions that survived until late in the century, John Stuart Mill's *Principles of Political Economy* (1848) providing a lasting synthesis of classical principles in a manner that made the science accessible to an informed public readership.

6. The Diffusion of Political Economy

Smith's *Wealth of Nations* was by 1800 acknowledged everywhere as the canonical text of political economy, but nineteenth century English writing had relatively little impact outside English-language areas, and at this time anything written in French had a far bigger readership than texts written in English. Jean Baptiste Say had been greatly influenced by his reading of Smith, and in 1803 he produced his own *Traité d'Économie Politique*, followed by a number of new editions, a 'catechism' in 1815 and a new six-volume 'complete course of political economy' in 1828–9. Say conceived political economy as a significant part of an

education for republican citizenship, and his writing was deliberately directed to a large popular audience. Since his texts were readable and in French they were quickly taken up outside France, with the result that, after Smith, Say became the most widely recognized political economist of the first half of the nineteenth century.

Say is remembered today for his 'law,' that 'supply creates its own demand.' In fact, this is a travesty of his argument on 'effective demand,' that whatever the general demand for a good, the political economist was only concerned with the demand of those with means to purchase, which means had themselves been gained from previous sales of their goods or services to others. He also argued that the process of production and distribution involved utilities, not goods, and that what the process of consumption destroyed was the utility of a good or service, and not the good or service itself. When combined with his principle of effective demand, this shifted the emphasis of his work away from the (production-oriented) English Classical Political Economy and towards the sphere of consumption. In these and other respects Say's political economy laid the path which led to the new subjective, marginal economics of the 1870s.

See also: Area and International Studies: Political Economy; Capitalism; Industrial Society/Post-industrial Society: History of the Concept; Industrialization, Typologies and History of; Labor History; Labor, Division of; Malthus, Thomas Robert (1766–1834); Marx, Karl (1818–89); Mass Media, Political Economy of; Political Economy in Anthropology; Power in Society; Smith, Adam (1723–90); Wealth Distribution

Bibliography

- Faccarello G (ed.) 1998 *Studies in the History of French Political Economy. From Bodin to Walras*. Routledge, London
- Griswold C L J 1999 *Adam Smith and the Virtues of Enlightenment*. Cambridge University Press, Cambridge, UK
- Hilton B 1977 *Corn, Cash, Commerce. The Economic Policies of the Tory Governments 1815–1830*. Oxford University Press, Oxford, UK
- Hirschmann A O 1977 *The Passions and the Interests. Political Arguments for Capitalism before its Triumph*. Princeton University Press, Princeton, NJ
- Hont I, Ignatieff M 1983 (eds.) *Wealth and Virtue. The Shaping of Political Economy in the Scottish Enlightenment*. Cambridge University Press, Cambridge, UK
- Letwin W 1963 *The Origins of Scientific Economics. English Economic Thought, 1660–1776*. Methuen, London
- Peach T 1993 *Interpreting Ricardo*. Cambridge University Press, Cambridge, UK
- Thomas W 1979 *The Philosophic Radicals: Nine Studies in Theory and Practice, 1817–1841*. Oxford University Press, Oxford, UK
- Tribe K 1988 *Governing Economy. The Reformation of German Economic Discourse, 1750–1840*. Cambridge University Press, Cambridge, UK

Winch D 1996 *Riches and Poverty: An Intellectual History of Political Economy in Britain, 1750–1834*. Cambridge University Press, Cambridge, UK

K. Tribe

Political Economy in Anthropology

While in the history of economics, political economy is a widely accepted description of the general style of analysis of national wealth and international trade that emerged in the early nineteenth century in the writings of classical economists such as Adam Smith and David Ricardo, the term has come to be used in anthropology to refer to a variety of theoretical approaches that developed in the 1970s and shifted the focus of anthropological analysis away from isolatable local systems and toward 'large-scale regional political/economic systems.' In this shift, research now took 'the form of studying the effects of capitalist penetration upon these communities' (Ortner 1984, p. 141).

Although the term is sometimes applied only to developments within North American anthropology, it is misleading to see political economy in this sense as a description of a distinctively American 'school' limited to the 1970s. At least some practitioners had examined such broader links much earlier. Moreover, at around the same time, anthropologists elsewhere were also beginning to investigate more closely the links between global capitalism and the local social systems they had been trained to analyze. Doubtless American efforts in this direction were distinguishable because of their greater debt to the work of Andre Gunder Frank and the sociologist Immanuel Wallerstein on the capitalist world system, as well the influences on them of earlier materialist tendencies within American anthropology.

But, particularly in France and Britain, and often under a rather different set of theoretical influences—most notably from contemporary Marxist theorizing—anthropologists were also reexamining the relationships between capitalism and what were often described as various precapitalist modes of production or social formations in Asia, Africa, and Latin America. Any assessment of this general move on the part of anthropologists from the 1970s to engage with global economic factors and their influence on the local groups anthropologists were more used to studying must then examine both of these developments.

1. Political Economy, Anthropology, and Global Capitalism

The development of political economy in North American anthropology was linked to a more general

intellectual and cultural challenge to America's imperial expansion, itself stimulated by opposition to the war in Southeast Asia. Consequently, political economy in the American context was on the one hand theoretically somewhat eclectic, and on the other, concerned primarily with issues of global capitalism. Given the tendency of anthropologists to engage with regions, societies, and cultures outside or on the margins of North Atlantic capitalism, much of the energy of political economists in America was devoted to analyzing the impact of capitalism on individuals and communities on what was sometimes called the 'periphery' of the capitalist system. This global emphasis in their approach to more traditional anthropological concerns, moreover, made American political economists less likely to classify these marginal individuals and groups as *precapitalist*. They were consequently also less interested than their French counterparts in any attempt to develop a general theory of modes of production, seeing this as being of little relevance to the concerns of social and cultural anthropologists doing field studies in the present.

Materialist analyses were well established in American anthropology some time before the emergence of political economy, although it could hardly be argued that these were based on any genuine engagement with, or even sympathy for, the work of Karl Marx himself. This materialist current is commonly associated with the name Leslie White, who is assumed to have been familiar with work in the Marxist tradition, if not with Marx's own writings, although he was prevented from acknowledging the link explicitly because of a generalized American hostility to Marx. Marvin Harris was an influential advocate of materialist analyses from the 1960s, although, unlike White, he was openly hostile to Marxist theorizing, preferring a form of techno-ecological determinism that was rather nearer to that of White.

Closer to the interests of the new generation of political economists was the work of Julian Steward. Steward had formulated a more open, 'multilinear' theory of evolution. He also headed the Puerto Rico project, unusual for the ways it tackled issues to do with social change and colonial history as central concerns of anthropology.

But more than anyone else, a group of anthropologists studying in New York in the 1940s—including subsequently well-known figures such as Eric Wolf, Sydney Mintz, Elman Service, and Stanley Diamond—became the key figures in anthropology's direct engagement with colonialism and global capitalism from within a materialist perspective, an engagement that subsequently attracted the label 'political economy.' Wolf and Mintz in particular became increasingly concerned with the implications of European global expansion for the anthropological analysis of peasant communities in the Caribbean and South America. Both wrote anthropological monographs that were unusual for their discussion not only of local

systems of culture, kinship, and social organization, but of the formative links established in the long history of European colonization between local communities and the centers of the world economy in Europe and North America. Later, both also produced influential general works on the development of a world capitalist system, paying particular attention to the interconnections established in this process between Europe and North America on the one hand, and the only apparently isolated peoples on its periphery on the other (Wolf 1992, Mintz 1985).

Others, from both the New York group and outside it, contributed, if less directly, to the development of political economy. Unlike many materialists in American anthropology, Stanley Diamond made important contributions in Marxist theory, founding the journal *Dialectical Anthropology* in the 1970s, and working on a critical/phenomenological Marxism that contrasted rather sharply with contemporary Marxist theorizing in France. Eleanor Leacock made her own contribution, both to the development of a Marxist anthropology, through a pursuit of themes of kinship, family, and gender raised in Engels' *Origins of the Family*, and to political economy in some important work on the impact of the fur trade on indigenous North Americans.

Two further influences served to broaden the popularity of such analyses in American anthropology by the early 1970s. First, there was the publication of the very influential early works of the dependency theorist Andre Gunder Frank and, shortly thereafter, the sociologist Immanuel Wallerstein (see Frank 1969, Wallerstein 1974). What this work brought in particular was a more sophisticated understanding of the imperatives of the global economy, and the compatibility, indeed even functionality, within it of apparently noncapitalist economic and political forms. Anthropologists such as Jane Schneider and Peter Schneider, Katherine Verdery, Gavin Smith, Carol Smith, William Roseberry, and others associated with the new political economy 'school' did not necessarily accept wholeheartedly the 'capital-centrism' of these more global approaches, but instead entered into a sustained and fruitful debate with them over the global versus the local determinants of socioeconomic systems in what Wallerstein called the 'peripheries' of the world capitalist system.

The second formative influence, particularly on materialist anthropology in America in this period, was the growing movement against America's military intervention in Indochina. This contributed greatly to the radicalized political and intellectual atmosphere on North American campuses, where a new generation of anthropologists was being trained. But it also has specific implications for anthropology, raising as it did the possibility that the discipline itself could be implicated in the project of imperial. In 1968, in an article that generated a heated debate, one anthropologist went so far as to describe anthropology

explicitly as a 'child of imperialism' (Gough 1968). Similar concerns led to significant conflicts within the professional anthropological body itself (the American Anthropological Association), which debated at its annual meetings the involvement of anthropologists in covert operations.

2. The French Marxist Heritage

If in North America the main parameters of political economy were laid down by the engagement of anthropologists with a predilection for materialist explanation with issues arising out of the development of a world capitalist economy, in Europe the situation was somewhat different. This is not to say there was no interest in the relationship between global capitalism and local socioeconomic organization. On the contrary, this central concern in American anthropology found important parallels, especially in France and Britain. Here, however, the debates were informed by a far more theoretically sophisticated attempt to develop a Marxist theory appropriate to the analysis of anthropological materials, a level of theoretical concern largely absent in contemporary American anthropology. And the center of this reworking of Marxism was to be found in France.

Developments in French anthropology were subject to three distinctive influences. The first was the opening up of the Marxist left, in part spurred by the developing Marxist critique of Stalinism which, among other things, produced a space for Marxist intellectuals, both inside and outside the Communist Party, to criticize the kinds of mechanical and evolutionist interpretations of Marx that up to then were the only versions permitted by Marxist parties. Second, and related to this, the Sino-Soviet conflict inspired at least some Marxist intellectuals and political activists to search for alternative models of both socialism and the ways it could be achieved. Third, unlike the case in America, French anthropology was dominated by the imposing intellectual figure of Claude Lévi-Strauss, and thus it is not surprising that when French anthropologists turned to the work of Karl Marx for inspiration, they developed a structural version of Marxism highly critical of the existential and humanist Marxism whose best-known exponent in France was Jean Paul Sartre. It is not especially surprising that the two main 'schools' in French Marxist anthropology—the first associated closely with the theories of the communist philosopher Louis Althusser, and the second with the structural Marxism of Maurice Godelier (a student of Lévi-Strauss)—while being critical of mainstream 'mechanical Marxism,' should both also have been equally if not more critical of the main alternative. While these ideas may have been influential in the wider student movement, Althusser's anthropological admirers, most notably Emmanuel Terray and Pierre Philippe

Rey, along with Godelier and his students, all sought to reformulate a Marxism that was 'scientific,' based on Marx's 'mature' work, and avoided what they took to be the pitfalls of an approach based on humanistic assumptions about a universal human subjectivity thwarted by capitalism.

As this suggests, for structural Marxists, rethinking the Marxist heritage meant going back to the work of Marx on capitalism, and what they found in doing so was the beginnings of a general theory of historical materialism within which anthropology could play a very significant role. That role was to be the formulation of a general theory of modes of production, and the laws of transition from one to another. Contra some of their critics, this was not seen as a mere exercise in antiquarianism, since such a theory would provide, it was argued, the theoretical basis for a political practice aimed at stimulating not just a transition to communism, but the shaping the nature of the communist future as well.

It is suggested that there was no single Marxist approach in French anthropology, but instead a collection of competing and often antagonistic camps. And yet, in spite of differing allegiances—to Althusser, to Godelier's version of structuralism, etc.—at least when viewed from the outside, French Marxist anthropology did seem to constitute a fairly unified discursive field for debate.

French Marxist anthropologists made important contributions to a number of areas of concern to both Marxism and anthropology (for an overview, see Kahn and Llobera 1981). A good example is provided by the debates over the nature of lineage-based societies in Africa that were sparked off by the early publications of Claude Meillassoux (see, for example, Meillassoux 1964, 1967). Scholars, including most centrally, Meillassoux himself, Emmanuel Terray, and Pierre Philippe Rey (both former students of Louis Althusser), as well as Maurice Godelier, Arienne Deluz, Pierre Bonte, and others, used Meillassoux's material on the Gouro of the Ivory Coast, as well as ethnographic data on other uncentralized African societies, to debate the nature of most of the basic concepts of historical materialism—class, mode of production, productive forces and relations of production, exploitation, articulation, social formation, patriarchy—as well as the central theoretical tenet of historical materialism—economic determinism (or what one participant called Marx's hypothesis of economic determinism in the last instance). This important debate, more than any other, serves to illustrate much of the potential theoretical benefits, as well as some of the main pitfalls, of the project of a Marxist anthropology as it was conceived by these scholars, namely a project to recover, by means of a careful reading of Marx's own work, his 'mature' theoretical project, and to examine its more general validity in situations largely unknown to its originator. Moreover, the debate added considerably to the

understanding of more traditional anthropological problems to do with the nature of authority and social cohesion in kinship-based societies, and societies without centralized political systems.

3. *Marxism, Articulation, and Political Economy*

From the perspective of the study of the linkages between global capitalism and local social and economic systems, the most important effect of these debates was in contributing to the formation of a theory of what some participants called the 'articulation of modes of production.' Drawing on the theoretical work of Althusser, Etienne Balibar, and Maurice Godelier, anthropologists, and others with an interest in the concerns of political economy, produced analyses of the relations between the local and the global understood as an articulation of capitalism with different precapitalist modes of production. It is not often unappreciated that exploring this relation was always a central concern of Meillassoux's in a series of publications, including his early monograph on the Gouro. A sophisticated Althusserian analysis of articulation is to be found in the work of the brilliant Marxist anthropologist, Pierre Philippe Rey (see especially, Rey 1973). And other important contributions to this project were made in Britain by a variety of scholars influenced by French structural Marxism, including Ernesto Laclau, Barry Hindess, Paul Hirst, Talal Asad, and others, and by a group associated with the journal *Critique of Anthropology* including Felicity Edholm, Olivia Harris, Joel S. Kahn, John Gledhill, Josep Llobera, Stephen Nugent, and Maila Stevens. Mention needs also to be made of the important work of Jonathan Friedman on Marxist theory, anthropology, and globalization.

4. *Beyond Political Economy*

The central concerns of political economy have certainly not disappeared in anthropology. But it would be fair to say that in general within the discipline there have been significant shifts at the level of theory and analytic style, as well as at the level of the selection of problems and issues for research, so that very few anthropologists could now be described as political economists. The reasons for this declining popularity of political economy are varied. There has certainly been a recognition that earlier attempts to theorize the relationship between the global and the local in either a dependency/world systems mode or by means of a concept of the articulation of modes of production suffered from problematic theoretical tendencies. These include an overly functionalist understanding of the needs of capitalism, a unidimensional economism, an inattention to the spheres of meaning and agency, and so on. Political economists have, like many others,

been criticized for their eurocentrism, and for failing to appreciate that race, culture, and gender are not mere epiphenomena of social systems. Like those of many others, the truth claims of political economists have been subjected to wideranging critique by scholars influenced by poststructuralism, postmodernism, feminism, and postcolonialism.

Social and cultural anthropology has been affected profoundly by the critiques of those associated with the so-called 'writing culture school.' This has undermined claims about the reality of those other lives constructed in anthropological texts, by the demand for greater degrees of reflexivity, and by the placing of problems of cultural, religious, gender, racial, and national identity at the forefront of issues for investigation. Political economy which, despite the name, doubtless focused on economic issues more than any other, naturally was affected by these shifts. Finally, European versions of political economy, precisely because of their greater concern with issues of (Marxist) theory, have been damaged by the general intellectual and political rejection of Marxism, particularly since the late 1980s.

While political economy in the various forms that it took in Europe and North America in the 1970s and early 1980s will probably not regain the central role it once had in the discipline, for better or worse it would be misleading to pronounce it dead and buried. The tendency to see in the rise of Foucauldian and poststructuralist paradigms a complete rejection of, particularly, the basic conceptions of structural Marxism, is probably misleading. It is possible to argue that these merely take the structural Marxist project one step forward, from a critique of ideology, to a critique of discourse *tout court*.

Certainly, some of the central concerns of political economy have reemerged in more recent attempts by anthropologists to theorize the phenomenon of globalization. An anthropology seeking its roots in a variety of new social movements is very much on the agenda of anthropologists styling themselves as 'radical' and/or 'committed.' Some of the concerns of an earlier generation of political economists are evident in more recent feminist anthropology, as well as in various attempts to construct an anthropology of human rights. Recent attempts to forge a link between critical social theory and anthropology through various reworkings of the concept of modernity may be seen as an attempt to correct the economistic and unidimensionalist failings of the structural version of Marxism. There are even apparently once again calls to return to Marx. It remains to see how these will develop in anthropology in coming years.

See also: Dependency Theory; Economic Anthropology; Globalization and World Culture; Globalization, Anthropology of; Globalization: Political Aspects; Marxist Social Thought, History of; Political Anthropology; World Systems Theory

Bibliography

- Frank A G 1969 *Capitalism and Underdevelopment in Latin America: Historical Studies of Chile and Brazil*. Monthly Review Press, New York
- Gough K 1968 Anthropology: Child of imperialism. *Monthly Review* 19(11): 12–27
- Kahn J S, Llobera J R 1981 *The Anthropology of Precapitalist Societies*. Macmillan, London
- Meillassoux C 1964 *Anthropologie Économiques des Gouro de Côte d'Ivoire*. Mouton, Paris-Le Haye
- Meillassoux C 1967 Recherche d'un niveau de détermination dans la société Cynégétique. *L'Homme et la Société* 6
- Mintz S W 1985 *Sweetness and Power: The Place of Sugar in Modern History*. Viking, New York
- Ortner S 1984 Theory in anthropology since the sixties. *Comparative Studies in Society and History* 26: 126–66
- Rey P P 1973 *Les Alliances de Classes*. Maspero, Paris
- Wallerstein I M *The Modern World-system, Vol. 1*. Academic Press, New York
- Wolf E R 1982 *Europe and the People Without History*. University of California Press, Berkeley, CA

J. S. Kahn

Political Elites: Recruitment and Careers

The question of who rules is always central to political science precisely because power is never evenly distributed. At any given time, some individuals and subgroups in society will have more power than others. While most non-Marxist theorists tend to agree that concentrations of power are inevitable, their understandings of how individuals or groups gain, maintain, and distribute influence differ widely. Further, despite the common usage of the term elite, theorists fail to agree on an appropriate definition.

Some descriptions of elites focus on the role and effectiveness of a society's leaders. Suzanne Keller, for example, defines elites as 'a minority of individuals designated to serve a collectivity in a socially valued way' (1963, p. 4). Alternatively, Eva Etzioni-Halevy (1993), following in the tradition of Lasswell (1936), Lasswell and Kaplan (1950), and Sartori (1976), grounds her definition of elites in terms of distribution of resources, explaining that 'resources are mostly (though not exclusively) accumulated in organizations, and elites are thus mostly (though not exclusively) people who control such resources through incumbency of certain top positions in organizations' (p. 221). Taking still another tack Dogan and Higley define political elites as 'the holders of strategic positions in powerful organizations including dissident ones, who are able to affect national political outcomes regularly' (1998, p. 15, Dogan 1989). In their view it is elite adjustment and compromise which creates stable political orders, even under conditions in which the population as a whole is characterized by

sharp ethnic and cultural divisions, a perspective which is shared by Luther and Deschouwer (1999).

Etzioni-Halevy's definition, which we follow in this article, allows analysts to include among elites, movements which are attempting to change the structure of leadership in the social and/or political order. Thus African Americans holding important positions in, say, the National Association for the Advancement of Colored People (NAACP) will be included among political elites.

1. The Study of Elites

Vilfredo Pareto (1848–1923), Gaetano Mosca (1858–1941), and to a lesser extent, Roberto Michels (1876–1936) developed the underpinnings of modern elite theory. In part they were reacting to a Marxist analysis which insisted that, in bourgeois society, basic divisions occurred between two social classes, the bourgeoisie and the industrial workers (Pareto 1935, Mosca 1939, Michels 1915). Politics and political power were perceived as effects of these arrangements. Formal or nominal political leaders of all societies were agents of the dominant class and were not free to act on their own in any but matters of relatively little consequence. By the late twentieth century, Marxists had changed their orientation somewhat, allowing the state and political leadership some freedom to act upon a few important matters, but Marxist theory, even in its present tattered form continues, basically, to urge the dominance of economic class and the mode of production as the prime moving forces (Sitton 1996).

Both Mosca and Pareto insisted that political elites play an independent role in determining the nature of regimes. They argued that all societies are hierarchical, dominated at any given time by one or more elites which, eventually, will be replaced by others. Dominant elites are political elites and will remain in power as long as they retain their unity and character.

Elite theory has changed a great deal since its initial formulation but many of the same arguments continue. Thus populists and quasi-Marxist analysts such as C. Wright Mills (1959) and William Domhoff (1990) (see also Zweigenhaft and Domhoff 1998) continue to insist that America and other capitalist societies are ruled by powerful socioeconomic groups (capitalists and the military for Mills; WASP capitalists and their dependents for Domhoff). More recently feminists and race theorists have maintained that gender or race in conjunction with class are the dominant forces in society, and consequently political elites' thoughts and behavior emerges from their economic, or racial, or gender status (McClain and Garcia 1993, Carrol and Zerilli 1993).

Alternatively, scholars like Robert Dahl (1982), Nelson Polsby (1980), Theda Skocpol (1979), Evans et al. (1985), Gabriel Almond (1978), Etzioni-Halevy (1993), Dogan (1989), and Keller (1963) argue that the

political order and political elites play a significant, independent role and are not merely the captives of class, gender, or racial patterns. Indeed as noted earlier, Dogan and Higley (1998) and others go so far as to argue that it is the character of political elite conflicts and settlements which determine the nature and stability of the social order.

A number of social scientists, including Robert Putnam (1976), Samuel Eldersveld (1989, Eldersveld et al. 1995), Eric Carlton (1996), and Ronald Ingelhart (1997), have conducted empirical studies of elites around the world and through time. Their work permits some generalizations based on the evidence but does not resolve key issues concerning the structure of power. There is still little agreement as to the role of political elites as compared with other elite groups, or to what extent gender and race need to be brought into the analysis of elite organization and power.

2. What Kinds of Elites?

While Marx and Engels' notion of an anarchic, communal, totally egalitarian society at the dawn of history, in which women played key social and political roles, seems to have little basis in fact, the earliest groupings of human beings were members of relatively egalitarian, though male dominated, segmented societies. With little societal power available, power differentiation was, almost by definition, limited (Service 1975). Historically, the next human developmental stage involved the creation of pyramidal systems of social organization. This pattern characterized the most advanced societies down to the fourth or third millennium BCE and generally involved a number of groups or villages loosely organized under the authority of a paramount chief. Such chiefs often emerged as charismatic leaders who were successful in the hunt and/or war and whose power waned as they aged.

A new step in societal evolution, generally associated with the further development of agriculture and its consequent population growth, was the emergence of patrimonial societies, dominated by one individual (king or chief) whose authority was generally based on tradition supported by religious myths (Weber 1978). Such societies often attempted to institutionalize succession by hereditary right as a means, theoretically, of overcoming succession crises, and ensuring that the present chief or king could pass authority on to his male heirs. Historically, however, the hereditary mechanism for choosing rulers has not been very effective in maintaining social peace (Burling 1974).

Patrimonial kingdoms also took the first steps in the differentiation of political institutions with at least some authority devolving to subordinate officers whose positions were based on lineage and/or personal or contractual fealty. The expansion of power

available to the new social organizations led to greater asymmetries in the society. It is in patrimonial kingdoms and, even more so, in the great bureaucratic empires of the ancient world, i.e., China, Egypt, Rome, etc., that one can begin to analyse issues in terms of mutually supportive or competing elites (the army, the bureaucracy, religious elites) who shared at least some power with the ruler (Eisenstadt 1963). These societies, too, selected leaders on hereditary principles for the most part, though rulers were not infrequently deposed by the military or bureaucrats. Indeed military control of the succession came to characterize the Roman Empire by the third century CE. Military or bureaucratic leaders seized power to improve their own status and that of their clan or family. It was partly for this reason that Chinese and Byzantine emperors often assigned eunuchs to key bureaucratic and military roles in the empire (Balch 1985). Since they lacked direct descendents, eunuchs were considered more reliable instruments of the ruler.

3. Political Elites in the Modern World

In the modern world societies become more differentiated and patrimonial orders based on traditional authority gradually give way to authority derived from the beliefs in the legitimacy of 'rational laws' in the Weberian sense, agreed to by the people. Not only does the emergence of democratic structures (polyarchies to use Dahl and Lindblom's (1953) felicitous expression) increase the number of power centers and the complexity of power relations, but the issue of how power should be distributed becomes more central to political discourse. At the same time patterns of recruitment to and career development in elite institutions change dramatically.

Some groups which traditionally held political power lost it. For example at one point religious leaders had to be counted among political elites. Today religious leaders may still play a significant role on the political scene though, in the Christian world, their impact on political outcomes is far more limited than it once was. Indeed the process of secularization has proceeded in Europe, and, even the more religious United States boasts a fairly strict separation of religious and political institutions. In some Muslim countries religious leaders play a much more significant role. However, they are not, with the exception of Shi'ite Iran, members of the political elite *per se*, with the power to formulate and enforce policies (Tschirgi 1994, Zartman 1980).

There are today, also, a few patrimonial regimes dominated by particular lineages: Saudi Arabia, Jordan, and the Gulf Emirates come to mind as well as a few small African kingdoms. As traditional authority has come to be replaced by the notion that governments should be based on the rule of the people (which emerged in Europe as part of the liberal tradition),

most regimes, even when not democratic, pretend to be so, basing their authority on mass parties, often dominated by charismatic leaders such as Hitler or Mao Zedong and his successors. Charisma plays an even more important role for leaders who have not institutionalized their power. One clear example is Fidel Castro, who rules directly like a traditional Latin American *caudillo* (chief), eschewing a mass party (Aguila 1990).

Military leaders, who emerge from time to time claiming to speak for the people, and/or 'to restore order' continue to play an active role in such countries as Myanmar (formerly Burma), Pakistan, Turkey, Indonesia, and a number of African and, from time to time, Latin American countries. In Turkey the military considers itself to be the inheritor of the policies of Mustafa Kemal Atatürk and has intervened in politics on a regular basis to preserve a secular state, usually withdrawing once the crisis was over. It is currently highly suspicious of a Muslim revival with important political overtones (Gole 1997). The military has played a similar role in other Muslim countries, such as Algeria, where it has been the prime competitor of fundamentalist groups. Of course the military plays an influential role in such countries as China and North Korea though not officially part of the governing structure of either regime. Military influence and indirect connection to policy making is more likely today than direct military rule over a political system which, when it occurs, typically does not last too long. In long-established and strong parliamentary regimes (as in Europe and the United States), the military role in political decision making is very limited.

4. Elites in (Mostly) Modern Democratic Systems

In relatively polyarchic systems with successfully functioning parliamentary or presidential regimes, legislatures elected by the populace play key roles in making binding decisions about the distribution of resources. In such societies election to office is mediated by political parties and the character of the electoral system is important. In both parliamentary and presidential regimes the key elite players tend to be lawyers and other professionals such as teachers and retired bureaucrats and politics becomes a full-time profession, replacing the old English notion of the amateur politician (King 1981). In presidential regimes the chief executive (whatever the formal title) exercises considerable power independent of the party to which he/she is attached and has, in the past, based more of his authority on charisma. Indeed, in the United States, the president, while not independent of his party, draws much of his authority from his unique relation with the mass media. Arguably the political roles of American political parties have substantially diminished in the face of the media revolution.

In parliamentary regimes, however, chief executives (prime ministers, premiers) are selected from within the party and generally are leaders of a group of almost equals and exercise less personal power than presidents. Of course this depends partly on the particular nature of the party system. Within the framework of a relatively disciplined two-party system, party members are more likely to adhere to the party line than in regimes characterized by multiple, usually weaker, parties and coalition governments. Exceptionally, during their heyday, the Communist parties of France and Italy were, despite multiparty systems with relatively weak political parties in each country, quite able to retain control over party members. Even in parliamentary systems with relatively disciplined parties, the role of the media has heightened the charisma and power of the party leader (especially when in office) above that of his closest associates. This results in greater identification of the Party with the leader (Dogan 1989, Kamarck 1995, Linz and Valenzia 1994, Mainwaring and Valenzuel 1998, Patterson and Mughan 1999, Trend 1997, Yesilada 1999, Riggs 1988).

Political parties are present even in authoritarian regimes such as the People's Republic of China. They were the preferred mode of political organization in most other Communist countries and in many African states. Generally, such parties have been instruments of mobilization and control allowing relatively little freedom. They have also been the central loci of power in the society. There have been some partial exceptions. In Mexico the PRI (The Party of the Institutionalized Revolution) effectively dominated politics but opposition was permitted and the party did not attempt to control the whole society. The Mexican system has further opened up in recent years and the one-party system is probably in a state of terminal decay. Also, the Soviet Union and its former satellites have, at least temporarily, transformed themselves into more open societies often with a substantial circulation of elites. Under the Soviet regime, and still in China, advancement in political power required working one's way up through the party with the aid of powerful patrons whose fall from favor could seriously affect one's career (MacFarquhar 1993, Li and Bachman 1989, Lane 1995).

In well-established parliamentary regimes the success of those seeking elite status depends on party position, seniority, and effectiveness in parliamentary work including the give and take of debate. Party loyalty generally remains the requisite condition of advancement in parliamentary regimes with disciplined party systems. Loyalty is less important in the American presidential regime. There power in the legislature is partly a function of seniority, good committee assignments, service to constituents, and a good relationship with the media.

In most of Europe, party organizations are relatively weak and the number of active party militants small.

In less developed countries parties are even weaker and mostly quite short-lived. Parties often are the personal vehicles of individual charismatic leaders. After the leader passes from the scene, such parties usually fall apart to be replaced by still another 'personal' party. Political power in such countries is actually heavily dependent upon 'clientelist' relationships among various political families.

In modern democracies politics has increasingly become a career in which those aspiring to office start early and remain long, often beginning in local government (Eldersveld et al. 1995). With the passing of distinctly working-class socialist parties in Europe, political leaders are even more likely than they were in the first half of the twentieth century to come from middle-class backgrounds and to boast college or university degrees (Dogan 1989, Putnam 1976).

Money plays a relatively important role in elections in capitalist societies even when stringent limitations are placed upon the amounts parties are permitted to raise. However, individual wealth has not provided a passport to office, even in the United States where considerable freedom in the use of personal and contributed funds is permitted. Middle- or lower-middle-class aspirants to office can usually obtain financial support from backers who, for one reason or another are willing to bankroll them.

Among the unique elements of modern, relatively large complex societies is their heavy dependence upon highly differentiated bureaucracies. Despite important elements of bureaucratic structure in traditional empires, bureaucracies of the modern type never developed in these countries. High level bureaucratic officials, as Weber pointed out early in the twentieth century, are among key political elites (Weber 1978). They have played especially important roles in societies such as the Soviet and Chinese Communist regimes whose theoretical goal was to eliminate them. Indeed as late as the 1960s Mao Zedong's cultural revolution was designed to abolish or at least sharply reduce bureaucratic power. Mao had a point. Bureaucrats were blunting the cutting radical edge of the Chinese revolution for, as Weber argued, bureaucrats tend to routinize decision making. Mao did not succeed, for Weber was also correct in asserting, when many social theorists thought otherwise, that socialism could only increase the size and power of bureaucracy.

Weber, however, was wrong in implying the modern political order would eventually be dominated by bureaucratic elites. It is true that carrying out legislative mandates always involves administrative discretion. However, bureaucracies have never completely dominated a political system, with the possible exception of Japan in the decades after World War II (Van Wolferen 1989).

The influence of bureaucracy is heightened when the technical training of an elite corps of bureaucrats emphasizes their authority and power in the state. The

second empire in Germany (1870–1918) was characterized by the very significant role of the bureaucratic elite as was France under the Fourth Republic (1946–58). In the latter case the substantial power of the bureaucracy was partly a result of the inability of the party system to create a stable executive (Koh 1989, Peters 1978, Suleiman 1984, Farmer 1992, Thakur 1981).

For different reasons England and the United States have never created a bureaucratic elite of the continental European type. In the US, especially, the courts have added a layer of control above that created by Congressional oversight and presidential authority, and have insisted, in many cases, upon public consultation in bureaucratic decision making. Indeed, the role of courts has become so significant in many countries (though not as significant as in the United States) that some scholars now speak of the 'judicialization of politics.' Such judicialization has strongly affected bureaucratic decision making (Vallinder 1994).

In many modern political regimes, e.g., France, Japan, England, etc. higher level bureaucrats are recruited directly from the university for bureaucratic careers and often retire early to enter business or run for political office. High office in the United States federal bureaucracy is more permeable through in-house promotion and lateral recruitment. This gives American bureaucratic recruitment a more democratic cast than that of most other advanced industrial nations.

5. *Issues for the Present and the Future*

Political elites in advanced industrial societies do exercise more power in the sense of producing desired outcomes than do those in less developed countries. Elites in the poorer countries of Latin America, Africa, and parts of Asia often behave as if they are creating new public policies and yet nothing changes. This occurs because the organizational capacity of the society is limited and thus there is simply not sufficient power to carry through new initiatives. Promising beginnings frequently end in failure, too often frustrating, and eventually corrupting, those elites who begin as sincere reformers. Individuals or political groups may dominate the political orders of poorer societies but frequently they can change very little because the population lacks the capacity to respond effectively to new initiatives. Similarly efforts by richer more powerful nations to assist in or compel change are usually unsuccessful.

In general the role of women in public affairs has been very limited. In western Europe and the United States increasing numbers of women are becoming part of the political elite, though their role remains relatively small except in the Scandinavian countries.

In most of the rest of the world their role is still smaller. Indeed the representation of women in official governmental organs dropped when the regimes in Communist states were overthrown. However, the offices they had held in various ineffective legislative bodies did not entail real authority so that the loss is more symbolic than real.

The distribution of power among racial and ethnic groups in increasingly multiethnic societies is very likely to be a major source of conflict in the twenty-first century. How such conflicts are to be handled is an issue which, except in a few cases, has not been dealt with effectively.

See also: Charisma and Charismatic; Charisma: Social Aspects of; Community Power Structure; Elites: Sociological Aspects; Leadership in Organizations, Sociology of; Leadership: Political; Leadership, Psychology of; Mobility: Social; Pareto, Vilfredo (1848–1923); Pluralism; Power in Society; Power: Political; Presidential Personality; Presidential Popularity; Social Mobility, History of; Social Stratification

Bibliography

- Aguila J M del 1990 Development and revolution in Cuba. In: Wiarda H J, Kline H F (eds.) *Latin American Politics and Development*, 3rd edn. Westview, Boulder, CO, pp. 420–54
- Almond G A, Powell G B Jr 1978 *Comparative Politics*, 2nd edn. Little Brown, Boston, MA
- Balch S H 1985 The neutered civil servant: Eunuchs, celibates, abductees and the maintenance of organizational loyalty. *The Journal of Social Biological Structures* 8: 313–28
- Burling R 1974 *The Passage of Power; Studies in Political Succession*. Academic Press, New York
- Carlton E 1996 *The Few and the Many: A Typology of Elites*. Scholar Press, Brookfield, VT
- Carroll S J, Zerilli L M G 1993 Feminist challenges to political science. In: Finifter A W (ed.) *Political Science: The State of the Discipline II*. American Political Science Association, Washington, DC, pp. 55–76
- Dahl R A 1982 *Dilemmas of Pluralist Democracy: Autonomy vs. Control*. Yale University Press, New Haven, CT
- Dahl R A, Lindblom C E 1953 *Politics, Economics, and Welfare: Planning and Politico-economic Systems Resolved into Basic Social Processes*. Harper, New York
- Dogan M 1989 *Pathways to Power: Selecting Rulers in Pluralist Democracies*. Westview Press, Boulder, CO
- Dogan M, Higley J 1998 *Elites, Crises, and the Origins of Regimes*. Rowman & Littlefield, Lanham, MD
- Domhoff G W 1990 *The Power Elite and the State: How Policy is Made in America*. De Gruyter, New York
- Eisenstadt S N 1963 *The Political Systems of Empires*. Free Press of Glencoe, New York
- Eldersveld S J 1989 *Political Elites in Modern Societies: Empirical Research and Democratic Theory*. University of Michigan Press, Ann Arbor, MI
- Eldersveld S J, Stromberg L, Derksen W 1995 *Local Elites in Western Democracies: A Comparative Analysis of Urban*

- Political Leaders in the U.S., Sweden, and the Netherlands.* Westview Press, Boulder, CO
- Etzioni-Halevy E 1993 *The Elite Connection: Problems and Potential of Western Democracy.* Polity Press, Cambridge, MA
- Evans P B, Rueschmeyer D, Skocpol T (eds.) 1985 *Bringing the State Back In.* Cambridge University Press, New York
- Farmer K C 1992 *The Soviet Administrative Elite.* Praeger Publishers, New York
- Gole N 1997 Secularism and Islamism in Turkey: The making of elites and counter elites. *Middle East Journal* 51(1): 46–58
- Inglehart R 1997 *Modernization and Postmodernization: Cultural, Economic and Political Change in 43 Societies.* Princeton University Press, Princeton, NJ
- Kamarck E C 1995 Structure as strategy: Presidential nominating politics in the post-reform era. In: Pfiffner J P (ed.) *Government and American Politics, Classic and Current Perspectives.* Harcourt Brace College Publishers, Fortworth, TX
- Keller S I 1963 *Beyond the Ruling Class: Strategic Elites in Modern Society.* Random House, New York
- King A 1981 The rise of the career politician in Britain—and its consequences. *British Journal of Political Science* 11(3): 249–85
- Koh B C 1989 *Japan's Administrative Elite.* University of California Press, Berkeley, CA
- Lane D (ed.) 1995 *Russia in Transition: Parties Privatisation and Inequality.* Longman, London
- Lasswell H D 1936 *Politics: Who Gets What, When, How.* Whittlesey House, New York
- Lasswell H D, Kaplan A 1950 *Power and Society: A Framework for Political Inquiry.* Yale University Press, New Haven, CT
- Li C, Bachman D 1989 Localism, elitism, and immobilism: Elite formation and social change in post-Mao China. *World Politics* 42(1): 64–94
- Linz J J, Valenzia A 1994 *The Failure of Presidential Democracy.* Johns Hopkins University Press, Baltimore, MD
- Luther K R, Deschouwer K 1999 *Party Elites in Divided Societies: Political Parties in Consociational Democracy.* Routledge, New York
- MacFarquhar R (ed.) 1993 *The Politics of China 1949–1989.* Cambridge University Press, Cambridge, UK
- McClain P D, Garcia J A 1993 Expanding disciplinary boundaries: Black, Latino and racial minority groups in political science. In: Finifter A W (ed.) *Political Science: The State of the Discipline II.* American Political Science Association, Washington, DC, pp. 247–80
- Mainwaring S, Valenzuel A 1998 *Politics, Society, and Democracy: Latin America.* Westview, Boulder, CO
- Michels R 1915 [trans. Paul E and Paul C]. *Political Parties; A Sociological Study of the Oligarchical Tendencies of Modern Democracy.* Hearst's International Library Co., New York
- Mills C W 1959 *The Power Elite.* Oxford University Press, New York
- Mosca G 1939 [trans. Kahn H D]. *The Ruling Class.* 1st edn. McGraw-Hill, New York
- Pareto V 1935 [trans. Bongiorno A, Livingston A]. (ed.) *The Mind and Society.* Harcourt, New York
- Patterson S C, Mughan A (eds.) 1999 *Senates: Bicameralism in the Contemporary World.* Ohio State University Press, Columbus, OH
- Peters B G 1978 *The Politics of Bureaucracy: A Comparative Perspective.* Longman, New York
- Polsby N W 1980 *Community Power and Political Theory: A Further Look at Problems of Evidence and Inference,* 2nd edn. Yale University Press, New Haven, CT
- Putnam R D 1976 *The Comparative Study of Political Elites.* Prentice Hall, Englewood Cliffs, NJ
- Riggs F W 1988 The Survival of Presidentialism in America: Power-Constitutional Practices. *International Political Science Review* 9(4): 247–78
- Sartori G 1976 *Parties and Party Systems: A Framework for Analysis.* Cambridge University Press, Cambridge, UK
- Service E R 1975 *Origins of the State and Civilization,* 1st edn. Norton, New York
- Sitton J F 1996 *Recent Marxian Theory: Class Formation and Social Conflict in Contemporary Capitalism.* State University of New York Press, Albany, NY
- Skocpol T 1979 *States and Social Revolutions: A Comparative Analysis of France, Russia, and China.* Cambridge University Press, New York
- Suleiman E N (ed.) 1984 *Bureaucrats and Policy Making: A Comparative Overview.* Holmes & Meier, New York
- Thakur R 1981 *Elite Theory and Administrative System,* 1st edn. Sterling, New Delhi, India
- Trend D 1997 *Cultural Democracy: Politics, Media, New Technology.* State University of New York Press, Albany, NY
- Tschirgi D (ed.) 1994 *The Arab World Today.* Lynne Rienner, Boulder, CO
- Vallinder T 1994 The judicialization of politics. *The International Political Science Review* 15(2): 91–9
- Yesilada B A (ed.) 1999 *Comparative Political Parties and Party Elites: Essays in Honor of Samuel J. Eldersveld.* University of Michigan Press, Ann Arbor, MI
- Van Wolferen K 1989 *The Enigma of Japanese Power: People and Politics in a Stateless Nation.* Knopf, New York
- Weber M 1978 [Roth G, Wittich C (eds.); trans. Fischhoff E et al.] *Economy and Society: An Outline of Interpretive Sociology,* 2 Vols. University of California Press, Berkeley, CA
- Zartman I W (ed.) 1980 *Elites in the Middle East.* Praeger, New York
- Zweigenhaft R L, Domhoff W G 1998 *Diversity in the Power Elite: Have Women and Minorities Reached the Top?* Yale University Press, New Haven, CT

S. Rothman

Political Geography

Political geography is a field of inquiry concerned with the nature and implications of the evolving spatial organization of governmental regimes and formal political practices around the world. It is a widely recognized sub-discipline of geography, and it is increasingly an interdisciplinary pursuit as well. Political geography investigates why politically organized areas emerge in the places they do, and how the character and configuration of those areas both reflect and affect the social and environmental contexts in which they are situated. Whereas a political scientist or political sociologist might focus on the politics and social structure of a political entity, a political geographer would likely emphasize how and why a particular piece of the earth's surface came to be

organized as a discrete political unit, and what that unit represents and means for the peoples, institutions, and places affected by it.

1. *Development of Modern Political Geography, 1890s–1970*

The origins of political geography are often traced to Aristotle's effort to describe the ideal state in terms of its territory, population, and internal geographical structure (Kasperson and Minghi 1969). Its modern roots, however, lie in the writings of the German geographer, Friedrich Ratzel (1844–1904). Ratzel sought to analyze states in terms of their environmental characteristics, arguing that an understanding of the political world needed to be rooted in the human and physical character of the concrete territorial spaces occupied by states. He is most known, however, for his Darwin-inspired organic theory of the state, which treated states as competitive territorial entities vying with one another for control over parts of the earth's surface. He set forth his theory in a seminal book, *Politische Geographie* (1897), which marked the formal beginnings of the modern sub-discipline.

Ratzel's organic state theory attracted attention in both Europe and the United States (Glassner 1996). The Swedish political scientist, Rudolf Kjellén (1865–1922), used Ratzel's ideas about the territorial dynamics of state growth to develop an elaborated vision of the state as an organism, in the process giving birth to the term 'geopolitics' (see *Geopolitics*). Contemporaneously, the American naval historian Alfred Thayer Mahan (1840–1914) published an influential set of works arguing that control over sea lanes and adjacent land areas was critical for the maintenance and expansion of political power. This was soon followed by an alternative, more widely circulated geopolitical vision from the geographer-director of the London School of Economics, Sir Halford Mackinder (1861–1947). He emphasized the importance of the interior of the Eurasian landmass (the heartland) for the maintenance and spread of political power.

Given the relevance of these geopolitical contributions, and the interest they attracted, it is not surprising that the political geography of the early twentieth century was strongly identified with geo-strategic thinking (Claval 1994). Ratzel and Kjellén's work became the foundation for a school of geopolitics (*Geopolitik*) in Germany led by Karl Haushofer (1869–1946), which eventually was appropriated by the Nazi war machine. In other parts of Europe and the United States, debates over the relative merits of Mahan and Mackinder's conceptualizations attracted considerable attention in both academic and policy circles, and led to yet other geopolitical visions, most notably those of the Dutch-born, American-educated international relations scholar, Nicholas John Spykman (1893–1943), who focused attention

on the importance of the so-called rimland (coastal fringe) of Eurasia.

Despite the importance of geopolitical theorization in the early history of the sub-discipline, the first half of the twentieth century also saw expanding work on the geographical characteristics of individual states (Hartshorne 1954). The specific configuration of the world political map was undergoing substantial changes that required investigation and analysis. Geographers were being called upon to provide background on boundary disputes, the internal divisions within empires, and the impacts of new transportation technologies on political areas. World War I was a particular catalyst to such work. Indeed, the director of the American Geographical Society, Isaiah Bowman (1878–1950), was a member of the American delegation to the Paris Peace Conference where a new European political map was forged, one that reflected a combination of prevailing geopolitical visions influenced by Mackinder and understandings of the particular geographical characteristics of state territories.

During the interwar period, the study of political geography expanded modestly in colleges and universities, although relatively few geographers defined themselves primarily as political geographers. Those who did emphasized the historical territorial development and contemporary geographical characteristics of individual states. Bowman produced the first systematic political geography text of the post-World War I political order. It concentrated on the territorial problems of the new countries of the world, particularly those in Europe. Richard Hartshorne (1899–1992) saw this work as an extension of Ratzel's concern with the environmental character of state areas; his own influential work focused on the integrative and disintegrative potentials of differing state territorial characteristics and founding territorial ideologies (summarized in Hartshorne 1954).

World War II spawned increased interest in political geography offerings in some colleges and universities in the United States and parts of Europe, yet it did not lead to any fundamental reorientation of the sub-discipline. The one exception concerned the geopolitics wing of political geography, which fell into wide disfavor in some places because of its association with Nazi efforts to justify claims to territory. For decades the Soviets opposed any use of the term political geography, and in Germany itself the sub-discipline ceased to be taught altogether. In the English-speaking world, the most influential writings of the succeeding two decades dealt with geographical influences on the functional character of individual states through the study of their morphology, internal organization, social and physical homogeneity, relative location, and boundaries (e.g., Hartshorne 1954, Prescott 1965, East and Moodie 1968). Some innovative work came out of these studies, highlighting critical connections between geography and political decision making (e.g.,

Cohen 1963). For the most part, however, political geography became somewhat disconnected from conceptual and empirical developments in other parts of geography and beyond. Moreover, the positivist turn of the discipline of geography during the late 1950s and 1960s left little room for the consideration of political matters. By the end of the 1960s, these circumstances had placed political geography in a somewhat marginalized position.

2. *Political Geography in the Late Twentieth Century*

Political geography did not remain marginal for long. An important edited compilation from the late 1960s (Kasperson and Minghi 1969) provided a useful framework for teaching and research, and geographers trained in quantitative urban and economic geography began to turn their attention to the role that politics plays in the distribution of services, people, and wealth. The latter development spawned work focused at local and regional scales. Kevin Cox's (1973) seminal book on the politics of urban services gave rise to an influential set of studies in 'welfare geography' (see *Welfare State, Geography of* and *Locational Conflict (NIMBY)*) that drew on diverse theories ranging from Marxism to public choice theory. In the process, political geography came into contact with emerging 'political economy' perspectives.

A somewhat different influence on political geography came from a group of urban and economic geographers with a quantitative bent who saw the potential for the quantitative analysis of election data. Extensive work on the spatial organization of elections, voting patterns, and the geography of representation made electoral geography a widely recognized sub-specialty within political geography beginning in the 1970s (Taylor and Johnston 1979). The methods and theories of electoral geography were not closely aligned with other work in the sub-discipline (Taylor 1985), but the developing field of electoral geography widened the scope of political geography and attracted interest among a new generation of geographers. At the same time, the different scales and issues on which votes are taken encouraged a broadening of political geography's prior preoccupation with the scale of the state.

By the early 1980s, political geography was clearly emerging from its formerly marginalized status. Having shed its positivist orientation, politics was back on the agenda of geography as a whole, and political geography could therefore no longer be dismissed as peripheral to the discipline. Moreover, new work on urban, regional, and electoral themes was attracting increased attention to political geography and connecting it to issues and concerns being taken up by other social sciences. Even more tra-

ditional themes were taking on new life, with expanding interest in shifting cold-war geopolitical relations and the emergence of new regimes of control in the world's oceans (see Glassner 1996). The founding of a new journal, *Political Geography* (originally entitled *Political Geography Quarterly*), in 1982; the establishment of major research and study groups in the International Geographical Union and various national geographic organizations; and the founding of a boundary research institute at the University of Durham signaled the growing importance of the sub-discipline.

At the heart of political geography's revivification was the effort to develop a more sophisticated theoretical foundation for the sub-discipline (Reynolds and Knight 1989). Impetus for the theoretical turn came from the combination of insights being offered by the new, theoretically informed literature on urban social conflict and a growing frustration with the lack of connection between research initiatives outside political geography and more traditional work in the sub-discipline. Explicit engagement with theory took two overlapping, but sometimes conflicting forms: a concern with developing a theory of the state and a concern with theorizing the nature of the state system.

The work of Gordon Clark and Michael Dear (1984) is particularly associated with the effort to theorize the state in political geography. They promoted consideration of a variety of theories of the state and themselves advanced an approach that drew on Marxist ideas on the role of the state in the maintenance and expansion of capitalism, but which argued that the state is also the product of the institutionalized political-economic patterns and arrangements that are produced through specific state structures and functional hierarchies. Their work helped to foster wide-ranging inquiries into the spatial dimensions of state policy initiatives, as well as studies of the 'local state' and its relationship to the national state. The latter was also heavily influenced by the writings of David Harvey (e.g., Harvey 1985), who focused attention on how the capitalist accumulation process was linked to the political production of spatial inequalities within cities and sub-state regions.

Although those writing on the nature of the local and national state looked to the role of large-scale structural forces, their work focused on communities existing within states (see Reynolds and Knight 1989). Some political geographers saw this as seriously limiting, arguing that intrastate processes could only be understood with reference to the dynamics and norms of the international system of states. They also argued that political geography would have little to say about international relations if it did not confront the nature and meaning of the state system. The writings of Peter Taylor—particularly his important text, *Political Geography: World-economy, Nation-state and Locality* (Taylor 1985 and later editions)—were influential in advancing this argument.

Drawing on the world-systems work of Immanuel Wallerstein, Taylor sought to situate analyses of the changing political geography of the modern world in the context of the territorial dynamics of the emerging capitalist world economy. He argued that the forms and functions of national and sub-national units are derived from the positions territorial states occupy in the global political-economic order. Taylor's particular use of world-systems approaches was not widely emulated, but his writings helped to put the very nature of the system of states at the center of the political geographic enterprise.

The new theoretical perspectives coming into political geography during the 1980s influenced both the types of issues being studied and how they were studied. The relationship between political geography and uneven development became a particularly significant focus of attention (see Smith 1984). Political geographers also turned their attention to processes of nation-building and substate nationalism, seeing these as predicated in significant ways on the nature of the state and the system of states (e.g., Johnston et al. 1988). Even some studies in electoral geography began to look at the relationship between spatial structures of power and electoral outcomes (Johnston et al. 1990).

The closing decade and a half of the twentieth century saw the emergence of new theoretical influences and a broadening of the scope of political geography (Agnew 1997). The growing influence first of structurationist, and then of post-structuralist, theories in the social sciences had clear impacts on political geography. The concern with the relationship between structure and agency that was part of the structurationist critique prompted political geographers to focus more attention on the ways in which political contexts are created through the interplay of large-scale forces and local practices. In his influential volume *Place and Politics*, John Agnew (1987) argued that political studies are generally based on the assumption that, in the modern world, local and regional identities are being replaced by national communities segmented by socio-economic differences, not by geographic sectionalism. Yet Agnew showed that this 'nationalization thesis' failed to account for the enduring impacts of local context on social and political behavior. In the process, he made a compelling case for taking the concept of place seriously in political geography (see also Johnston 1991).

The structurationist literature also brought with it growing interest in the spatial structures and arrangements that circumscribe social and political life. This offered a potential bridge between new currents of thought and longstanding efforts to analyze the development and meaning of the world political map (Gottmann 1973). An important element in building that bridge was Robert Sack's work on *Human Territoriality* (Sack 1986). Sack argued that a critical

component of human social behavior was the effort to influence people and ideas through the control of specific geographical areas, and that a concern with 'human territoriality' could therefore provide important insights into social and political processes. Sack's work directly challenged the dominant tendency to treat political territories as *a priori* spatial givens, in the process opening an important theoretical window for political geographic work focused on the nature and meaning of specific political-territorial developments.

Work on the politics of place and on territoriality has developed in conjunction with themes from the post-structuralist literature to open new vistas for political geography. The very centrality of politics and political relations to post-structuralist thought has greatly abetted this process (Painter 1995). Indeed, at some level much of human geography as currently practiced is about politics, making it difficult to establish exactly what constitutes political geography *per se*. Yet, as Agnew (1997, p. 3) has argued, 'there is nevertheless still the need for an area of study of the more formalized and institutionalized arenas of power such as the spatiality of states, interstate geopolitics, the geographies of political and social movements, and territorial conflicts involving national and ethnic groups.' These are the matters that continue to be most closely associated with the sub-discipline of political geography, and they are not particularly confining. Indeed, the recent political geography literature addresses themes that range from the role of political geographic arrangements in the constitution of gendered identities to the implications of new telecommunications technologies for the territorial integrity of states.

3. Contemporary Approaches and Themes

The contemporary practice of political geography echoes the diverse elements that have contributed to its development. Textbooks and research literature reveal approaches that are traceable to (a) political geography's tradition of describing the spatial configuration, geographic character, and boundaries of politically organized areas; (b) the quantitative turn of political geography in the 1960s; (c) the influence of political-economy perspectives on the sub-discipline in the 1970s and 1980s; and (d) the recent interest in post-structuralist theories of society (see Agnew 1997 for a variation on this typology). Some studies can still be neatly classified in one or another of these analytical traditions, but there is also considerable blurring among them. Nevertheless, tensions persist among proponents of these different perspectives, the most acute of which is between those who posit a world of 'real' political spaces/arrangements that we can try to comprehend through some combination of empirical and critical analysis, and those post-structuralists who

view reality as unknowable outside of the individuals and societies that construct it. The opposition between these positions has created a dualism in the political geography literature that is only beginning to be addressed (Agnew and Corbridge 1995).

The substantive topics on which political geographers are working are a product of the more general effort to make sense of the rapidly changing political, social, and economic environment of the late twentieth century. The following subsections highlight the topics that are receiving the most attention. There is considerable overlap among and between these topics, but each has a core literature attached to it.

3.1 *Political Geographic Assumptions of Social and Political Theory*

There is a pervasive tendency in the social sciences to frame analyses of social and political developments in terms of the political territories in which they are situated. Yet the territories themselves are usually treated as little more than locational referents for whatever is being studied. A body of work in political geography seeks to expose the theoretical shortcomings of this state of affairs. Much of that work focuses particularly on the importance of understanding what is hidden and what is revealed when the modern territorial state is treated as an unproblematic spatial container for social and political processes. John Agnew in particular has argued that this creates a 'territorial trap' in which critical historical and geographical processes are ignored because the nature and meaning of territory are not problematized (see Chap. 4 of Agnew and Corbridge 1995). A number of political geographic studies support this line of reasoning by showing how the political organization of space is itself implicated in social and political processes, even shaping the very concepts of identity on which political communities are based. Work in this arena is thus tied to the larger effort to bring concepts of space into social theory.

3.2 *Geopolitics and Interstate Conflict*

Geopolitics is clearly back on the political geography agenda, as evidenced by the publication of a spate of recent books and the founding of a new journal, *Geopolitics*. The recent literature on the subject reveals some continuing interest in the development of formal geopolitical models, but recent work has concentrated on the changing geopolitical environment for interstate relations and on the world views and political/military priorities of key political actors. Particular attention has been paid to the emergence of the modern world economy and its impacts for the

conceptualization and pursuit of geopolitics (e.g., Agnew 1998). Political geographers have also examined the locational components of conflict itself, studying where conflicts have occurred and why some areas have experienced more conflict than others (e.g., O'Loughlin and van der Wusten 1993). The interest in discourse as a window into geopolitical motives is the focus of a new literature on 'critical geopolitics' (see Ó Tuathail 1996). Influenced particularly by deconstructionist ideas, this increasingly influential literature looks at the discourse of geopolitical practice to gain insight into the ways in which spatial understandings have been developed, deployed, and manipulated in the pursuit of particular political and military objectives.

3.3 *The Changing Role of the Territorial State in the Modern World*

There is a lively debate in political geography and beyond about the degree to which the traditional role of the state is being undermined by the increased mobility of capital and labor, the growing power of multinational corporations, the advent of transportation and communications technologies that decrease or eliminate the friction of distance, and the rise of extra-state trading blocs and political organizations (see generally Painter 1995). Within political geography this debate finds concrete expression in studies focused on the ways in which these developments have altered the internal spatial organization and territorial character of states. Political geographers are raising questions about the changing nature and function of boundaries; the distribution of powers among and between substate, state, and superstate political-territorial entities; and the means by which governments at different scales are seeking to integrate and defend their territories. Research on such matters has been spurred by concrete changes in the world political map: the rise of the European Union; the break-up of the Soviet Union, Yugoslavia, and Czechoslovakia; and the devolution of substantial powers to substate regional authorities in countries such as the United Kingdom and Spain.

3.4 *Ethno-national Challenges to the Prevailing Territorial Order*

Closely related to the prior theme is research focussed on the often discordant relationship between the map of territorial states and the territorial claims of different ethno-national groups. Political geographers have long argued that the concept of the nation-state is more an aspiration than a reality. The rise of so many visible instances of ethno-national conflict over the

past few decades has moved this issue to the center of the political geography agenda. Political geographic research has focused on the competing territorial ideologies and strategies of majority and minority groups, the ways in which the political organization of territory affects intergroup interaction and intragroup identity, and the particular notions of citizenship and nationalism that have been forged in differing historical-geographical contexts (see, e.g., Johnston et al. 1988). The infusion of post-structuralist ideas into political geography has also promoted interest in the use of cultural symbols in the pursuit of territorial strategies, as well as the ways in which meanings develop around particular territories and nationalized landscapes. Work on the political geography of ethno-nationalism has highlighted the overlapping senses of territory that exist among many ethno-national groups and the need for policy approaches that recognize this state of affairs.

3.5 The Political Geography of the Environment

One of the clearest obstacles to the effort to confront environmental problems is the fragmentation of the planet into political spaces, few of which are meaningful ecological units. The problem is not just that ecological and political boundaries do not coincide; the very roles that territorial states play in the international state system and the world economy frequently work against collective action in the interests of the environment (Johnston 1996). Building on a tradition of examining the relationship between spaces of governance and environmental geographies, political geographers are increasingly turning their attention to environmental questions. Recent work has focused on new transnational regulation regimes and the special issues presented by environments that do not lie within the territories of individual states: the world's oceans and Antarctica. At smaller scales, political geographers have examined the circumstances giving rise to the emergence of an environmental politics in particular places, the problems and prospects of collective resource management in different political contexts, and the use of environmental strategies in the pursuit of political-territorial objectives. Some of this work is closely aligned with the emerging field of political ecology, with its focus on the role of large-scale political-economic structures in the production of local environmental outcomes.

3.6 The Quest for Political Power and Influence

Studies in electoral geography have long focused on the spatial dynamics of a particular type of quest for political power: that which takes place in formal elections. The geography of elections continues to attract significant attention among political geographers, but a variety of other political movements

are commanding attention as well. This is a product of the increased visibility of such movements and the success of some of them in challenging long-standing political arrangements. Work in political geography has focussed on the political-territorial contexts that give rise to political movements, the spatial character of movement support, and the impacts of movement objectives on the political organization of space. A significant body of recent literature has considered how the economic, social, and political character of individual places is implicated in the rise and success of different types of political movements (Staeheli 1994). The focus on places, however, is not at the expense of larger-scale influences. Instead, many political geographers treat individual places as the product of the interaction between large-scale political and economic structures and local cultural, social, and material circumstances (Johnston 1991).

3.7 Political Place and the Construction of Social Identity

Closely related to the prior theme is a strand of work in political geography concerned with the ways in which social identities are shaped by the political-territorial construction of place. The plethora of new social movements that have come together around social categories such as race, gender, sexual identity, and ethnicity have attracted growing attention throughout human geography, and the recognition that such identities are rooted in territorial practices and arrangements has provided an important bridge between work in political geography and the literature of economic, social, and cultural geography (see Wolch and Dear 1989). The concept of place—as politically and socially constructed—is once again at the center of this area of research because identification with place is so often at the heart of social identity (Johnston 1991, Agnew 1997). Yet place is not necessarily presented as a benign incubator of social identity; a considerable body of literature has focused on the ways in which dominant interests create spaces/places of marginalization that reflect and reinforce existing power relations and social prejudices.

3.8 The Emergence of New Political Spaces

The sweeping economic, political, and technological changes of the late twentieth century are associated with the development of new political spaces that coexist, and sometimes compete, with territorial states. Many of these new spaces are derivative of the map of states in that they are either collections of states (e.g., the European Union) or sub-state regions that have recently achieved greater autonomy (e.g., Catalonia). Yet activities within and among these new state-derived political spaces are creating political spaces of

a completely new sort. These include new political spaces of cooperation along international boundaries and geographically distant political communities that are forging links outside the framework of states. Behind many of these developments is the explosion in communications and information technologies, which is fostering an increasingly deterritorialized politics that nonetheless has concrete territorial implications. The emergence of far-flung political communities united by ideology and technology poses new theoretical and empirical challenges for an area of inquiry concerned with the spatial dimensions of politics.

4. The Way Forward

Political geography has emerged as one of the most vibrant, active areas of inquiry within the discipline of geography, and it is likely to remain so. The political-territorial upheavals of recent years show no signs of abating. Moreover, it is likely that growing challenges to the role of the state will bring into question the tendency throughout the social sciences to treat the map of states as a static frame of reference, and will direct attention to the nature and meaning of the changing spatial organization of politics. The interdisciplinary character and impact of political geography are likely to grow with the recognition that political practices and outcomes cannot be divorced from the changing geographical spaces and contexts within which they are situated.

See also: Geopolitics; Welfare State, Geography of.

Bibliography

- Agnew J A 1987 *Place and Politics: The Geographical Mediation of State and Society*. Allen and Unwin, Boston
- Agnew J A (ed.) 1997 *Political Geography: A Reader*. Arnold, London
- Agnew J A 1998 *Geopolitics: Re-visioning World Politics*. Routledge, London
- Agnew J A, Corbridge S 1995 *Mastering Space: Hegemony, Territory, and International Political Economy*. Routledge, London
- Clark G L, Dear M 1984 *State Apparatus: Structures and Language of Legitimacy*. Allen & Unwin, Boston
- Claval P 1994 *Géopolitique et Géostratégie: La Pensée Politique, l'Espace et le territoire au XXe siècle*. Editions Nathan, Paris
- Cohen S B 1963 *Geography and Politics in a World Divided*. Random House, New York
- Cox K R 1973 *Conflict, Power and Politics in the City: A Geographic View*. McGraw Hill, New York
- East W G, Moodie A E (eds.) 1968 *The Changing World: Studies in Political Geography*. Methuen, London
- Glassner M I 1996 *Political Geography*, 2nd edn. John Wiley, New York
- Gottmann J 1973 *The Significance of Territory*. University of Virginia Press, Charlottesville, VA
- Hartshorne R E 1954 Political geography. In: James P E, Jones C F (eds.) *American Geography: Inventory & Prospect*.

- Syracuse University Press for the Association of American Geographers, Syracuse, NY
- Harvey D 1985 The geopolitics of capitalism. In: Gregory D, Urry J (eds.) *Social Relations and Spatial Structures*. St. Martin's Press, New York
- Johnston R J 1991 *A Question of Place: Exploring the Practice of Human Geography*. Blackwell, Oxford, UK
- Johnston R J 1996 *Nature, State, and Economy: A Political Economy of the Environment*, 2nd edn. Wiley, New York
- Johnston R J, Knight D B, Kofman E (eds.) 1988 *Nationalism, Self-Determination and Political Geography*. Croom Helm, London
- Johnston R J, Shelly F M, Taylor P J (eds.) 1990 *Developments in Electoral Geography*. Routledge, London
- Kasperson R E, Minghi J V (eds.) 1969 *The Structure of Political Geography*. Aldine Publishing Company, Chicago
- O'Loughlin J, van der Wusten H 1993 Political geography of war and peace. In: Taylor P (ed.) *The Political Geography of the Twentieth Century: A Global Analysis*. Belhaven, London
- Ó Tuathail G 1996 *Critical Geopolitics: The Politics of Writing Global Space*. University of Minnesota Press, Minneapolis, MN
- Painter J 1995 *Politics, Geography, and 'Political Geography'*. E Arnold, London
- Prescott J R V 1965 *The Geography of Frontiers and Boundaries*. Aldine, Chicago
- Reynolds D R, Knight D B 1989 Political geography. In: Gaile G L, Willmott C J (eds.) *Geography in America*. Merrill Publishing company, Columbus, OH
- Sack R D 1986 *Human Territoriality: Its Theory and History*. Cambridge University Press, Cambridge, UK
- Smith N 1984 *Uneven Development*. Blackwell, Oxford, UK
- Staeheli L 1994 Empowering political struggle: spaces and scales of resistance. *Political Geography*. 13: 387–91
- Taylor P J 1985 *Political Geography: World-Economy, Nation-state and Locality*. Longman, London
- Taylor P J, Johnston R J (eds.) 1979 *Geography of Elections*. Penguin, London
- Wolch J, Dear M (eds.) 1989 *The Power of Geography: How Territory Shapes Social Life*. Unwin Hyman, Boston

A. B. Murphy

Political History (History of Politics)

Political history is the study of how legitimate power has been defined, distributed, and justified in the past (see *Power: Political*). Two imperatives, both with ancient roots, link historiography and politics. The first is the need to affirm the identity of the political community by uncovering its roots in the past. This search for historical identity engaged both the Hebrews and Greeks at the very beginning of European culture; it moved the Romans to search for the sources of their state and prompted succeeding generations to study the history of their cultures, religions, and nations. 'No question,' wrote Jean Bodin in the sixteenth century, 'has exercised the writers of histories more than the origin of peoples' (Kelley 1998). The second link between history and politics is the desire to

find lessons in the past that might help solve the political problems of the present. 'I shall be content,' Thucydides remarked in the first chapter of *The Peloponnesian War*, 'if (this history) be judged useful by those inquirers who desire an exact knowledge of the past as an aid to the understanding of the future' (Thucydides 1996).

1. From the Ancient World to the End of the Middle Ages

Political history in the West begins with two masterpieces: Herodotus's history of the Persian wars, written between 445 and 425 BC and Thucydides's history of the long conflict between Athens and Sparta, which he began in 424 BC and left unfinished when he died a quarter of a century later. Products of the extraordinary burst of cultural energy that transformed fifth and fourth century BC Greece, Herodotus and Thucydides combined a deep sense of the importance of their own age—Herodotus wrote about the recent past, Thucydides about events in which he himself had taken part—with an awareness of the historian's responsibility to establish the truth about the past. Both men reflected on the problems of historical research, the reliability of eye witnesses, and the necessity of critically analyzing source material. More than any particular way of writing history, this was their most significant legacy for future historians (see *Historiography and Historical Thought: Classical Period (Especially Greece and Rome)*).

The study of history was more important in Rome than it had been in Greece, perhaps because Romans were always conscious of their complex historical relationship to the Greek culture that they had both vanquished and absorbed. Roman historians examined the problem of politics from a variety of perspectives; they wrote about leadership and institutions, military strategy and court intrigue, republican virtues, and imperial ambitions. But behind these disparate approaches was a common theme: the origins, triumph, and—eventually—the decline of the Roman state, which, as republic and empire, dominated the public life of the Mediterranean world for 700 years and then continued to haunt Europeans' historical imagination until well into the nineteenth century.

The emergence of Christianity and its adoption as the Roman Empire's official religion transformed all of western culture, including both politics and history. Christianity's most immediate contribution to political historiography was to pose the problem of church and state, which would decisively shape the theory and practice of European politics from the end of the Roman empire to the beginning of the twentieth century. Unlike the pagans whom they displaced or the Hebrew traditions upon which they built, Christians created religious institutions that were connected

to, but not coincident with, the sources of political authority. 'Spiritual and secular power,' Ranke (1962) wrote, 'could approach one another and be intimately related, but they could be fused only under exceptional circumstances and for a brief period of time. Their relationships and conflicts provide one of history's most important themes.' From Eusebius's history of the church, which reflected the triumph of Constantine's conversion at the beginning of the fourth century, to Otto of Freising's twelfth-century chronicle, *The Two Cities*, which described the conflict between empire and papacy, to Ranke's own history of the popes, which bore the marks of nineteenth-century's struggles between church and state, historians searched the past for clues to the proper relationship between God and Caesar, pope and emperor, altar and throne, religious autonomy and secular authority (see *Historiography and Historical Thought: Christian Tradition*).

2. History and the State

Beginning in the 1490s with the French invasion of the Italian peninsula, a series of international, constitutional, and religious crises disrupted European politics and helped to create a new subject for political history. This subject was, in Quentin Skinner's (1978) concise summary, 'the concept of the state—its nature, its powers, its right to common obedience'. The concept of the state emerged from the dissolution of two closely-connected assumptions that had shaped European political thought since Augustine: first, that the secular realm of politics had to be seen as part of a sacred order; and second, that Europe's components—its free cities, ecclesiastical states, and various monarchies—were part of a single polity, governed by the universal laws of the church and sustained by the institutional legacies of imperial Rome. A good deal of medieval political theory and practice had been devoted to reconciling these assumptions with the stubborn realities of political life; in the first decades of the sixteenth century, Europeans began groping their way towards new ways of thinking about politics—and, therefore, towards a new kind of political history.

One of the first signs of this can be found in the work of Niccoló Machiavelli. As a Florentine who had been forced into exile by the collapse of his city's constitutional order, Machiavelli (see *Machiavelli, Niccolò (1469–1527)*) felt the political crisis of the early sixteenth century at first hand. In *The Prince* and *The History of Florence*, Machiavelli sketched a political world that Thucydides would have recognized immediately, a world driven by struggles for power, beset by violence, subject to forces that are often beyond human capacities to control. A similar picture emerges in the work of Machiavelli's fellow citizen and contemporary, Francesco Guicciardini, whose histories

provide an even darker picture of instability and violence. But however limited people's ability to control events, history, for both Machiavelli and Guicciardini, was a totally human enterprise, untouched by the hand of God, unfettered by a transcendent order.

Machiavelli wrote about 'the state,' but without giving it a clear institutional definition. This would be the great achievement of sixteenth- and seventeenth-century legal humanists, whose study of the law provided a new foundation for both politics and history. Of course law had always been an important part of European culture, but scholars like Guillaume Budé, Jean Bodin, François Baudouin, and François Hotman combined philological methods and historical analysis to create a new science of jurisprudence that enabled them to establish the origins, nature, and limits of the state's authority. Bodin, for example, provided the classic definition of sovereignty as the supreme source of law within a commonwealth, exercised by, but clearly independent from, the ruler's person. The work of these scholars stands at the source of one of political history's richest traditions: the study of legal institutions and constitutional development, which includes among its practitioners such giants as Montesquieu, Savigny, and Maine (Kelley 1970).

Law was also important to those who sought a vision of the international order to replace medieval universalism. Hugo Grotius, for example, who wrote during the Thirty Years War in the first half of the seventeenth century, argued that the international order was based on natural laws, which were divinely ordained but could be discovered by examining the past; like the French legal humanists, Grotius regarded the study of law and history as part of a single enterprise. By the end of the eighteenth century, under the influence of a new set of crises generated by the French revolution (see *French Revolution, The*) and Napoleonic conquests, a number of scholars searched for an international system that was no longer based on God-given laws but was still subject to underlying principles that the historian could discover. Heeren (1834), for instance, wrote that the 'essential property' of the European state system was its 'internal freedom, that is, the stability and mutual independence of its members. To set forth how this was formed, endangered, and preserved ... is the great object of the historian.'

The historical image of a world of independent states struggling for existence within an international system was given its definitive formulation in the middle decades of the nineteenth century by Leopold von Ranke, perhaps the greatest political historian of the modern era (see *Historiography and Historical Thought: Modern History (Since the Eighteenth Century)*). Von Ranke is best known for his archival work and his insistence on the critical use of primary sources. But inseparable from Ranke the great empiricist was Ranke the moral philosopher of statehood, who

regarded states as the 'ideas of God' that move through the historical order 'like celestial bodies.' His work, which combined scholarly rigor with stylistic grace, was read widely and admired throughout Europe; his methods of teaching and research were adopted everywhere (see Krieger 1977).

3. History and the Nation

Of all kinds of history, the history of the nation has the deepest roots and the most widely extended branches (see *Nation: Sociological Aspects*). Its first and most persistently influential expression was in the Hebrew Bible, which recounts the archtypical story of a nation's origins, defines its special historical mission, traces its survival in the face of terrible trials, and promises its eventual triumph. Their story, the Hebrews' historians insisted, is a story like no other—among the many ways in which national histories resemble one another, none is more striking than their insistence on being unique. As for so many other nations without states, this national history sustained the Jewish people's sense of collective identity through centuries of dispersal and oppression. And once the Jewish nation finally had its state, the story of the national past became—and this is also characteristic of the genre—the site of political debates about the nation's present identity and future direction (see *Historiography and Historical Thought: Jewish Tradition*).

Although national history has a genealogy that extends to the dawn of western culture, its most intellectually vigorous and politically potent versions were written in the nineteenth century when, for the first time, political history became virtually synonymous with the history of nations and nation states (see *State, History of*). The nineteenth century was, of course, the great age of historical writing, a time when historical consciousness touched virtually every aspect of European culture, from art and literature to philosophy and the social thought and even the natural sciences. For most nineteenth-century intellectuals, to think about the world was to think about historical origins, evolution, and identity. This was a time when statesmen—men like Guizot and Macaulay—wrote history and when historians—Bancroft, Droysen, Treitschke—were active in political affairs. And this was also the era when the institutional foundation for modern historical scholarship—document collections, libraries and archives, journals and reference works—was created, much of it by, and often for, national governments (see *Historicism*).

While every nation has a story to tell about itself, every national story is different, shaped by the political imperatives, cultural climate, and institutional arrangements of each national community. Nineteenth-century French historians, for instance, whose state had a unusually long and apparently uninterrupted

existence, were able to imagine a single national history. Their problem was how to find a place in this history for the trauma of revolutionary upheavals and prolonged warfare that engulfed France after 1789. For almost two centuries, French historiography was shaped by political struggles over the revolution's legacy, just as French politics was influenced by historical struggles about the revolution's character and significance. Among the works engaged by this project are such classics of political historiography as Tocqueville's *Ancient Regime and the French Revolution* and Michelet's *History of the French Revolution* (see Tocqueville, Alexis de (1805–59)). It was not until the 1970s that François Furet could declare that the French revolution was finally over, by which he meant that it had lost its central place in the nation's political discourse and public memory.

Each national experience poses a distinctive ensemble of questions for its historians to confront. In Britain, these are most often questions about institutional continuity rather than revolutionary ruptures. In the US, the central historical issue is the question of identity itself, which was addressed in works like Frederick Jackson Turner's classic study of the frontier in American history (see *Frontiers in History*). The problem of identity also dominated German historiography in the nineteenth century: before the creation of a German nation state, this meant a competition between different versions of the national past, after national unification, it involved demonstrating that the nation created by Bismarckian 'blood and iron' was the only authentic destination for the German people's long historical journey towards political unity. In the twentieth century German history was shaped by efforts to locate the catastrophe of National Socialism in the national past: was National Socialism, as some claimed, the destination of Germany's 'special path' (*Sonderweg*) to modernization, or was it a particularly toxic example of a broader crisis within modern culture and politics (see *National Socialism and Fascism*)?

An important element in the public life of every nineteenth century state, history was even more significant for those national communities that did not have a state of their own. For the Poles, the Irish, the Czechs, and, in the course of the twentieth century, for the peoples of the world colonized by Europeans, a shared history was the surest source of present cohesion and the brightest hope for future nationhood. After all, what else did these nations possess besides their belief in a common ancestry? Often enough, this belief had to be constructed from the scattered remnants of an imagined past—literary works discovered and transcribed by patriotic scholars, ancient sites unearthed by archaeologists, memories of lost greatness brought to life in poems and plays and popular histories. Often enough the character of the national past was bitterly contested—by Czechs and Germans in Bohemia, Protestants and Catholics in

Ireland, conservatives and liberals in partitioned Poland, Muslims and Christians in the Balkans. In these parts of the world, the historical significance of John Hus, whether the town was called Derry or Londonderry, the meaning of the Polish uprising of 1863, or how to commemorate the battle of Kosovo could swiftly become matters of life or death. Every national history reflects ideological divisions and political conflict, but in nations without states, the seam between politics and history was nearly invisible, so that it was usually impossible to determine where the one began and other ended.

4. Subversive Histories

Although historians have often been the servants of those in power, there is also a long, rich tradition of subversive historiography, written by scholars who turned to the past in order to question the established order and suggest new paths for the future. In the nineteenth century, at the same time that historians were drawn into the service of the nation state, subversive historiography was given its most influential formulation by Karl Marx. For him, the political movements that had interested Thucydides, Machiavelli, and Ranke are surface phenomena, reflections of the deeper social conflicts from which history takes its shape and direction. Politics is merely an instrument of the ruling class, a way of justifying and defending its hold on the means of production. Traditional political historians, therefore, view the past through a distorting lens that turns the world upside down, obscuring the essential sources of change with their concern for artificial conflicts, empty slogans, and trivial events. The Marxist historian's purpose is to substitute for these ideological distortions the radical clarity of scientific materialism, which will reveal to the proletariat the roots of its oppression and thus help prepare them for their revolutionary role in history cosmic drama. From the last decades of the nineteenth century until the end of the twentieth, Marxism—in all its many forms—remained the most influential form of subversive historiography (see Marx, Karl (1818–89); *Marxist Social Thought, History of*).

Well before the final collapse of international communism in 1989, Marxist historiography had begun to lose its intellectual attraction and political influence. In part this was because of the diminution of ideological energies both in the communist regimes of eastern Europe and in labor movements throughout the world. Moreover, by the 1970s, the political passions that had drawn intellectuals to Marxism during the struggle against fascism in the 1930s and against colonialism in the 1960s had cooled. The Marxist view of history was also undermined by a growing skepticism about all grand narratives, a turning away from the aspiration to grasp the meaning

of human history as a whole. And finally, Marxism declined because it could not explain very well the subjects to which more and more historians now turned, especially problems of ethnicity, race, and gender, which did not fit into the categories defined by the means of production. Not surprisingly, the form of Marxism that remained most influential in Britain, the US, and most of western Europe was associated with historians like E. P. Thompson, who tried—not always with success—to fuse class analysis with a concern for culture and politics (see *Social History*).

Since the decline of Marxism, the subversive impulse in historiography has been fed from many other sources of which the most important, at least in the last two decades of the twentieth century, was certainly the work of the French philosopher and historian, Michel Foucault (see *Boserup, Ester (1910–99)*). Like Marx, Foucault's enemies were oppression and inequality, his goals, freedom and community. But unlike Marx, he did not propose an alternative view of history or a critical science of society. Foucault's project was willfully fragmented; he called his approach to the past a genealogy rather than a narrative history. Deeply concerned about what he regarded as the repressive power of organized knowledge, he was nevertheless unwilling to abandon efforts to understand and subvert the existing order of things. Above all, this meant to understand and subvert power, which was exercised not only by the agents of the state but rather was diffused 'in the multiple forms of subjugation that have a place and function within the social organism.' Emancipation, therefore, required a struggle against all of what Foucault regarded as the 'polymorphous techniques of subjugation' (Dirks et al. 1994).

Marx (and also, if much less influentially, Freud) had sought to subvert traditional political history by showing that politics was really a reflection of deeper social (or psychological) conflicts (see *Psychohistory*). Foucault subverted the conventional image of politics from the opposite direction, that is, not by showing that politics should be reduced to something else, but rather by insisting that everything—including what was once thought of as the most intimate spheres of private life—is politically charged. Both perspectives are useful since both can illuminate new problems and neglected connections. Whether they can—and should—lead to the abandonment of the subjects that have concerned political historians since classical times is much more open to question.

5. *The Varieties of Political History in the Twentieth Century*

Scholars' interest in the past has always been influenced by contemporary crises. Thucydides was moved to write his history by the Athenians' disastrous war with Sparta, Josephus recorded the destruction of

the Jewish people by the Romans, and Augustine composed his universal history under the shadow of the Vandals' sacking of Rome. In the modern era, the French revolution inspired Tocqueville, the Bolshevik revolution, Trotsky, the rise of Italian fascism, Salvemini, Nazi totalitarianism, Arendt, the world wars, Churchill—the list might easily be extended to include many other occasions when, to borrow Hegel's phrase, men's hopes were destroyed on history's slaughter bench.

The war that shattered the world order in 1914 dominated the study of international political history for a quarter of a century (see *First World War, The*). Basing their work on the documents published by all of the belligerents, historians examined both the immediate causes of the war and its deeper roots in the great power rivalries of the nineteenth century. Some of this work was obviously partisan, written to absolve the historian's own nation of responsibility; but much of it sought to strike a balance between commitment and detachment. A notable feature of this historiography was the participation of American scholars. Responding to the hopes and disappointments that attended the failure of Woodrow Wilson's vision of a conciliatory peace, these scholars—Sidney Fay, Robert Binkley, William L. Langer, Raymond J. Sontag, and many others—played an active role in controversy over the causes and consequences of the war. By the 1930s, there was a rough scholarly consensus that no one power was responsible for the war, which was thought to have come from a series of miscalculations by several different governments. In the 1960s, this consensus was shattered by the work of Fritz Fischer and his students who attempted to blame the war on Germany's reckless bid for world power. Fischer's view was bitterly attacked and then widely accepted, but no new consensus has emerged to explain what George Kennan, himself a distinguished student of international history, once called 'the great seminal catastrophe of the century' (Stevenson 1997).

In contrast to World War I, the causes of World War II did not attract much scholarly attention; virtually everyone agreed that this was Hitler's war (see *Second World War, The*). But the war's character and consequences generated an enormous historical literature, including some great works by participants such as Winston Churchill and Charles DeGaulle. Historical controversies surrounded issues such as the air war (especially the decision to use atomic weapons against Japan), wartime diplomacy, and the role of resistance and collaboration in the territories occupied by Germany and Japan. Among the historical questions about the war in Europe none are more painfully complex than those surrounding the mass murder of European Jews, conventionally known as the Holocaust. Was it, as some scholars believe, tied to a distinctive sort of German anti-Semitism? Or did it arise within the climate of violence created by the war itself? Is it best seen as a problem of German

history? Or of modernity as a whole? Could it have been prevented or at least diminished by the action of Germany's enemies? Central to the historical self-understanding of both perpetrators and victims, the Holocaust has become the century's exemplary disaster (Marrus 1987) (see *Holocaust, The*).

The origins of the Cold War, the global geopolitical rivalry between the Soviet Union and the US that dominated international politics from 1945 to 1989, were a matter of intense scholarly controversy throughout its long history. Was the Cold War, as some claimed, a defensive response by the western powers to communist aggression and expansionism? Or, as the so-called 'revisionists' argued, did the US over estimate, perhaps even manufacture, the communist threat in order to advance its own aims throughout the world? Until the collapse of the Soviet Union, both sides in this controversy depended largely upon sources from the western powers, but after 1989, the empirical basis of Cold War historiography was transformed by the opening of archives in both the former Soviet Union and its eastern European allies. As a result, historians have begun to create a new picture of international history in the second half of the twentieth century (Gaddis 1997).

Among the great themes of world history is the expansion of European power which began in the late fifteenth century and ended during the second half of the twentieth century, when the colonial states either abandoned or were expelled from their overseas possessions (see *Colonization and Colonialism, History of; Imperialism, History of*). European expansion was the subject of many important historical studies, including the analyses by J. A. Hobson and V. I. Lenin that inspired several generations of anti-imperialist writers. Decolonization also had a lively historiography, stimulated in part by the traumatic wars fought by France in Algeria and by the US in Indo-China. In the 1990s, historians became interested in what became known as postcolonialism, that is, the persistent and reciprocal relationship between colonized and colonizer both during and after the colonial experience itself. While many postcolonial studies take up issues that seem far distant from conventional political history—for instance, questions of gender, sexuality, and cultural identity—they are all connected to the ways in which political power is defined and distributed.

In the last decades of the twentieth century, political historians have become increasingly aware of the importance of regional and local allegiances and institutions. A number of recent works on the revolutions of 1848, such as Jonathan Sperber's account of the Rhineland, have shown how tensions between central and local authorities affected both the origins and outcome of the revolution. Historians have also become sensitive to the historicity of the national state: Eugen Weber, for instance, has described the long and uneven process through which the French

state imposed its values in the countryside, while Linda Colley, among many others, has uncovered the way in which Britain became a focus of people's interests and loyalties. The upheavals that destroyed Yugoslavia in the 1990s produced a cluster of new works on the historical sources of ethnic identity and changing nature of nationalism in the twentieth century (see *Nationalism: General; Nationalism, Historical Aspects of; The West*). In these studies traditional questions about state power and politics are enriched by a new concern for ethnicity, culture, and identity.

6. Conclusion: Political History in the Twenty-first Century

Although it no longer has the unquestioned historiographical primacy it once enjoyed, the history of politics remains a vigorous and remarkably diverse enterprise. For the general public, political history is a source of apparently endless fascination. Each year a staggering number of books, especially biographies and battle narratives, are published about subjects such as the American Civil War and World War II. Among professional historians there are signs that political history may have recovered some of the ground lost to social and cultural history in the 1970s and 1980s. An increasing number of scholars are now seeking to apply social scientific methods and concepts to the study of both domestic and international politics (see *History and the Social Sciences*). At the same time, insights from cultural anthropology (see *Anthropology and History*) and literary criticism have raised interesting new questions about the nature of political power and identity.

The history of politics is important because individuals' lives are shaped and limited by political institutions, while the very existence of their communities depends on how political power is exercised and restrained. The illusion that politics does not matter is a luxury that only those fortunate enough to live in a rich and orderly state can afford. Moreover, while the importance of national history may have declined in western Europe, for much of humanity the question of national identity remains vitally relevant; in many places people are still willing to die—and to kill—in behalf of their own version of the national story. And the emergence of a new set of problems and opportunities from the wreckage of the Cold War system (see *Cold War, The*) will certainly raise a new set of questions about the past and future of the international politics.

Does knowledge of the past help us, as Thucydides and Machiavelli and so many others believed, to master our own political problems? To be sure, few now share the confidence expressed by Lord Acton in his inaugural lecture of 1895 that 'the science of politics is the one science that is deposited by the

streams of history, like the grains of gold in the sand of a river; and the knowledge of the past, the record of truths revealed by experience, is eminently practical, as an instrument of action and a power that goes to making the future' (Acton 1957, pp. 25–26). And yet the impulse to extract lessons from the past seems irresistible; historical references and analogies are everywhere a prominent part of political discourse. If nothing else, the study of past politics should help to assay the ore left by the stream of history and to recognize the fool's gold so often on display in the rhetoric of public life.

See also: Aristotelian Social Thought; Citizen Participation; Democracy, History of; Democracy: Normative Theory; Democratic Transitions; Historicism; Historiography and Historical Thought; Modern History (Since the Eighteenth Century); Parliaments, History of; Participation: Political; Political Anthropology; Political Culture; Political Discourse; Political Economy, History of; Political Science: Overview; Republicanism: Philosophical Aspects

Bibliography

- Acton J 1957 *Essays on Freedom and Power*. Meridian Books, New York
- Breisach E 1983 *Historiography: Ancient, Medieval, and Modern*. University of Chicago Press, Chicago
- Dirks N B, Eley G, Ortner S B (eds.) 1994 *Culture/Power/History*. Princeton University Press, Princeton, NJ
- Gaddis J L 1997 *We Now Know: Rethinking Cold War History*. Clarendon Press, Oxford, UK
- Gilbert F 1965 *Machiavelli and Guicciardini: Politics and History in Sixteenth Century Florence*. Princeton University Press, Princeton, NJ
- Gooch G P 1959 *History and Historians in the Nineteenth Century*. Beacon, Boston
- Heeren A 1834 *A Manual of the History of the Political System of Europe and its Colonies*. Clarendon Press, Oxford, UK
- Iggers G G, Parker H T 1979 *International Handbook of Historical Studies*. Greenwood Press, Westport, CT
- Kelley D R 1970 *Foundations of Modern Historical Scholarship: Language, Law, and History in the French Renaissance*. Columbia University Press, New York
- Kelley D 1998 *Faces of History: Historical Inquiry from Herodotus to Herder*. Yale University Press, New Haven, CT
- Krieger L 1977 *Ranke: The Meaning of History*. University of Chicago Press, Chicago
- Laistner M L W 1963 *The Greater Roman Historians*. University of California Press, Berkeley, CA
- Marrus M 1987 *The Holocaust in History*. University Press of New England, Hanover, NH
- Momigliano A 1990 *The Classical Foundations of Modern Historiography*. University of California Press, Berkeley, CA
- Ranke L von 1962 *Die Römischen Päpste in den letzten Vier Jahrhunderten*. Phaidon Verlag, Vienna, Austria

Stevenson D 1997 *The Outbreak of the First World War: 1914 in Perspective*. St. Martin's Press, London

Thucydides 1996 *The Landmark Thucydides*. Free Press, New York

J. J. Sheehan

Political Lawyering

Scholarship on lawyers and politics falls into two broad classes. On the one hand, as in most theories of professions—most notably functionalist and market control theories—any politics is largely absent. In a very different way, studies of lawyers' participation in the state and political institutions, where lawyers take for granted the rules of the political game, conceal the collective action of lawyers toward a common definition of the public good. There are, however, two theoretical perspectives that engage lawyers' politics: political lawyering and cause lawyering. The first approach demonstrates that lawyers have mobilized collectively, at critical points of political transition, to advance the institution-building of liberal political regimes. The second approach indicates the impact of those lawyers who pursue political agendas in direct response to perceived inequity or injustice, and in reaction to repressive or arbitrary legal and political regimes. This article briefly reviews the first class of scholarship and then analyzes in more detail political lawyering and cause lawyering.

1. The Avoidance of Politics

1.1 Functionalism and its Critics

Functionalism dominated the sociology of professions, including law, during the years 1950–70. Since functionalists sought to understand the link between knowledge and social needs—or, more precisely, to study the social control of knowledge—the logic of the profession solved the intractable problem of the balance between esoteric, dangerous knowledge in the hands of a potentially egoistic few and democratic control of those few which could in itself be destructive. Functionalists found a solution in the logic of a profession where (a) specific social structures (socialization, ethics, rewards, etc.) oriented the use and the development of knowledge in order to produce a relationship of service between professionals and society, and (b) society delegated to a 'community' a set of duties which were dominated by 'institutionalized altruism' and whose dedication to those duties was appropriately compensated by high social rewards (self-government, prestige, power, and some-

times wealth). Yet the constellation of questions and concepts that were used to study lawyers, as a specific profession, did not refer either to politics or to collective action. History was not considered as the product of social action, but rather as the change to which a profession was compelled to adapt.

Although Talcott Parsons (1964) was the central figure of functionalism, he recognized that, due to its specific 'interstitial' position in the social structure, the legal profession cannot escape the political system. Through the different choices they make, lawyers, as intermediaries, facilitate or impede the task of political authority and favor or not transformations in legal interpretation. Yet even this analysis remains rather limited, because the terms that orient lawyers' action—political authority, law, clients—are given, and because, with some room to maneuver, the legal profession is fundamentally defined by an integrative function.

In the 1970s, functionalism (among other theories) became the target of a critical sociology that leveled three charges against the sociology of professions:

(a) theory (in this case, the functionalist theory of profession) was nothing more than the reproduction of professional ideologies and a mask for gaining or keeping symbolic and material rewards;

(b) the so-called 'service relationship' was only a social mystification and, like the ideal of independence, was not justified by the quality of the service; and

(c) sociologists, therefore, were servants, unwittingly or not, of the professional elite.

Critical sociology rejected ahistorical studies, demanded realism about professional interests, and raised new questions about power, effectively anticipating the rise of market control theory.

Magali Sarfatti Larson opened up a new research program in the late 1970s. Devoted to the comparative and historical study of several professions in the United States, *The Rise of Professionalism* (1977) brought together two intellectual traditions: the European tradition, centered on Marx, Weber, and Polanyi; and the American tradition. It combined a historical framework with collective action (the mobility project), and it integrated the transformation of the market and the change of the symbolic systems through a very original analysis of the role of universities in creating legal service as a new commodity. An interpretation of Larson's book became the central resource of a new school—market control theory—whose simplified thesis posited that lawyers are rent seekers who, to maximize their profit, use two complementary strategies: to reduce the supply of legal services by excluding competitors, and to increase demand for legal services by influencing the state. All the established and distinguishing characteristics of the legal profession—self-government, professional mortality, independence, controls over entry to the profession, moderate internal competition, and so on—still exist, but only as tools to extend market control and thus to increase

economic and social profit. This perspective, which received its most expansive application in Abel and Lewis's *Lawyers and Society* (1988a, 1988b, 1989), ignored politics, took for granted the capacity of professions to mobilize collectively, and assumed the ability of professions to impose their self-interested wills on legislatures and states. It directed scholarly attention away from lawyers acting collectively for reasons other than market control. Although erudite and fruitful in many other respects, Abbott's jurisdictional theory of professions has a similar result. While Abbott (1988) offers acute insights into ways that professions mobilize knowledge in their scramble for control over work, he does not classify either the political processes by which professions secure and maintain jurisdictional control, or the activities of professions beyond control of work domains.

1.2 Institutionalized Politics

Originating primarily from political science and studies of political recruitment, many studies since the 1960s have sought to understand the significance of a widely observed phenomenon—the overrepresentation of lawyers in the legislatures of many liberal democracies, e.g., the USA, India, Canada, Venezuela, Australia, and the Netherlands (see Miller 1995). What are the effects of background and networks on legislative process? Does the prominence of lawyers in legislatures make a difference to policy outcomes?

Studies of the compositional or recruitment politics of lawyers show limited results. Lawyers' actions in legislatures are rarely the expression of any collective professional orientations. Lawyers as legislators do not evince as a whole a coherent or distinctive substantive politics. They distribute themselves across political parties, so there is rarely a clear affinity with party ideology.

At least in the United States, the strong overrepresentation of lawyers in legislatures leads not to substantive outcomes but to a peculiarly lawyerly way of dealing with issues. According to Miller (1995, p. 174), who summarizes several decades of research as a context for his own findings on Congress and state legislatures, 'the numerical dominance of lawyer-politicians in US political institutions produces a government that is preoccupied with the procedurally-oriented, incrementalist myth of rights, instead of with broader substantive questions of public policy.' Thinking like a lawyer leads to reaction against radical change, it privileges procedural rationality over substance, and it produces a rights-focused approach to the solution of public problems. Policy-making alternatives are defined more narrowly and institutional parameters are largely taken for granted. Similar patterns may be observed elsewhere. Ledford's (1996) study of German lawyers in the last quarter of the nineteenth century finds that, through their partici-

pation in national political movements and parties, lawyers were actively committed to the building of the *Rechtsstaat*, a rule-of-law society that would implement the protection of civil rights. But he also found that this procedural orientation faltered when faced, a half-century later, with the substantive challenges posed by National Socialism.

2. Political Lawyering

Sociologists, historians, and legal scholars have demonstrated that, in Europe and North America, legal professions, acting collectively, have been periodic but by no means consistent agents of political liberalism at particular times and places from the seventeenth century onwards. These scholars generally define political liberalism along three dimensions:

(a) a moderate state, which is an amalgam of the rule of law, strong intermediate associations that exist between the state and the individual and will defend the latter against the former, a balance of national and local political authority, and an institutional separation of powers between the executive branch of government and the judiciary;

(b) a civil society, which combines the existence of voluntary associations, religious institutions, and publics, which do not owe their existence to the state and can stand against it; and

(c) citizenship, most notably the civil element of citizenship, in which individuals gain legal rights and personality.

These scholars identify historical episodes in several countries in which political lawyering has facilitated the construction of the moderate state, civil society, and civil citizenship. The findings of these studies may be summarized in five propositions.

2.1 *The Autonomy of the Judiciary and the Autonomy of the Bar are the Principal Conditions for the Fight of Lawyers on Behalf of a Moderate State*

Autonomy from the state has relevance for legal professions for two reasons. First, some independence of the judiciary from executive power is not only critical for establishment of a rule-governed state; and second, autonomy of the profession itself is the key element in this establishment, for without it no independent collective action will be possible. The fight for judicial autonomy can be found in eighteenth-century France where the *parlements* (sovereign courts with judicial and legislative functions) sought to enlarge their powers and control the person of the king and power of the royal court. Lawyers were the *parlements'* most faithful allies. They supported the independence of the *parlements* by urging effective general strikes which brought the justice system to a

halt, and by becoming the political thinkers of the struggle. Lawyers articulated the foundations of the moderate state with such slogans as 'Laws are veritable contracts between those who govern and those who are governed.' Since the 1870s, the bar in the USA has strenuously exerted itself to immunize federal and state courts from control by presidents or governors, congress or state legislatures, and federal, state and municipal bureaucracies. In Germany, where the judiciary was subordinated to the state, the lawyers' campaign for a 'free bar' began in the 1830s with the aim of obtaining the institutional separation of judicial and administrative authority and the autonomy of law through adherence to judicial precedent. This movement reached its zenith in the Imperial Justice Laws of 1877–79, only to decline thereafter until the advent of National Socialism.

Seen from the vantage point of political liberalism, the autonomy of lawyers may be one of the key distinguishing features of a liberal political society. In England, the autonomy of the bar long preceded the emergence of an English state in the sixteenth and seventeenth centuries. It thus became a pillar of English constitutionalism, and was therefore able to resist effectively the attempted incursions of Charles II and James II on such 'little commonwealths' and 'little republics' that existed independently within the English body politic. In France, that autonomy grew consistently with the nurturance of the *parlements* from the seventeenth century to the present, with short interruptions during the Revolution and 1822–30. It was the absence of long-established, independently-secured autonomy in Germany, coupled with internal and external conflicts, which ultimately brought the bar under National Socialist control in 1933 and rendered it quite unable to defend judicial autonomy.

2.2 *Lawyers' Struggles to Institute and Extend Citizenship are Diverse*

Lawyers' orientations towards the extension of citizenship can be distinguished along two dimensions: first, between substantive and procedural rights; and, second, among civil, political, and social concepts of citizenship. French lawyers have been the champions of substantive rights, such as the eighteenth-century efforts to defend the religious rights of Jansenists and to obtain property rights for peasants in Burgundy, or the nineteenth-century efforts to transform the courtroom into a political forum to fight against emergency laws, attacks on a free press, and limits on individual freedoms. In contrast, German lawyers in nineteenth or twentieth centuries did not press directly for substantive individual rights, but rather pursued them indirectly through proceduralism—which opened them to seduction by antiliberal sentiments in the 1920s and 1930s. In the United States, fringes of the

profession variously championed substantive civil rights and rights to representation on behalf of the poor, but the central tendency of the profession has been to valorize procedural limits on state authority, as in the McCarthy period of anti-Communist purges.

2.3 Lawyers' Collective Defense of Civil Rights Does Not Extend to Struggles in Favor of Economic and Social Rights, or to Fights Against Racism, Xenophobia, or Gender Discrimination

With surprising continuity across history and nation, lawyers have been strongly defined by a 'pure' civil rights model. Lawyers are not to be found collectively on the barricades of expanded welfare rights, of national political issues such as the universal ballot, or of the restriction of racism. During the civil rights era in the United States (1950s and 1960s), a few professional associations took collective action. In nineteenth-century France, few lawyers mobilized to defend the working class. Lawyers are limited liberals, most especially in relation to citizenship.

2.4 The Forms of Lawyers' Political Action are Principally Reactive and Rely on the Authority of the Public or Civil Society

Compared to social movements, lawyers react against arbitrary or illegal actions by governments and repudiations of individual rights through trials that become '*causes célèbres*.' In Britain and France, use of the courts as forums came to constitute the political public sphere. Radical barristers in eighteenth- and nineteenth-century England used the courts to rail against corrupt government, to publicize causes, and to champion rights. French barristers used their freedom of speech in the *ancien régime* to create extended legal briefs that effectively became political tracts. Not only did they stimulate debate within the enlightened public, but they claimed for themselves a representational role as the voice of the public. By contrast, lawyers in the United States used the institutions of civil society, which were well developed, to mobilize publics. In Germany, lawyers were able neither to constitute publics as an autonomous sphere, nor to mobilize the institutions of civil society, which barely existed.

2.5 Contradictory Movements Entangle Economic Globalization with Political Globalization on Behalf of Liberal Institution-building

The market has contradictory effects on political globalization of liberal political institutions. On the

one hand, the global market has homogenizing effects on nation states at the same time as it stimulates national reactions against it. The extension of the global market also introduces more contradictions within segments of legal professions and thus inhibits their capacity to act collectively. On the other hand, the conditions of financial stability and global trade require institution-building, such as the rule of law, independent courts, and private professions. Thus, differing segments of legal professions come to constitute liberal politics through institution-building that will facilitate global commerce, and through advocacy on behalf of human rights and environmental responsibility among other fields.

3. Cause Lawyering

Cause lawyering is a form of doing politics (Sarat and Scheingold 1998, 2001). Whereas legal elites separate law from politics, and legal skills from moral calling, cause lawyers seek to 'reconnect law and morality.' Unlike the classic image of the lawyer as 'hired gun,' prepared to serve a client whatever the merits of the client's case and however wide the gap between the values of lawyers and clients, cause lawyers usually identify with the values and causes of their clients. They may even join with clients in political and social movements to advance a common goal. Thus, cause lawyers have a substantive vision of the good society and they use their legal skills to produce it.

If political lawyering is defined by the defense of political rights construed narrowly, the range of causes can be enormously variegated, encompassing pay equity, animal rights, capital punishment, Bedouin land rights, struggles against military repression, women's rights, peasant movements, and human rights, among a multitude of others. (See the case studies in Sarat and Scheingold 1998, 2001.)

Scheingold (2001) shows that cause lawyering is both constant and contingent, and that cause lawyers must choose between frequently uncomfortable alternatives. There is a constancy and coherence that transcends time and place for, while 'cause lawyering is frequently directed at altering some aspect of the social, economic, and political status quo,' (Sarat and Scheingold 1998, p. 4) cause lawyers are everywhere caught between two visions of democracy—of legal and political rights versus social and economic rights. They lean towards the latter and its emphasis on 'egalitarian values and redistributive policies.' Since they face an enormous variety of political regimes (totalitarian, colonial, authoritarian, repressive, liberal democratic, neo-corporatist) with quite different sets of resources, and since their causes are highly diverse, the expression of cause lawyering and its efficacy is highly contingent.

3.1 Strategies of Cause Lawyers

Cause lawyers must usually choose between two broad strategies, although these frequently blur in practice (Scheingold 2001). To follow a *legal strategy* means operating within the very mainstream of professionalism, with which cause lawyers constantly struggle. Thus Israeli civil rights lawyers defend Palestinian 'terrorists' against Israeli security officials within Israeli courts. This may be somewhat effective on narrow grounds but leaves the larger status quo intact. To follow a *political strategy*, and to identify with the cause of clients, may take the form of a political advocacy through social and political movements that extends well beyond legal constraints, as evidenced by many cause lawyers in Latin America. This allows much greater scope of advocacy, but reduces the scope of professional collective action since it marginalizes the cause lawyers.

Similarly, cause lawyers must choose between distance from the state versus close proximity to the state. On the one hand, a close relationship with the state, through cooperation with state authorities, such as maintained by cause lawyers in post-Apartheid South Africa, may yield results while significantly constraining the scope of action, for lawyers are bound by the limits of state institutions. Yet many of the ideals advocated by cause lawyers cannot be achieved without the exercise of state authority, as in the extension and enforcement of new rights, or the complicity of state officials. On the other hand, an adversarial relationship to the state may mobilize public opinion and social movements, but it is a high-risk strategy. The closer to the state, the more the cause lawyer will be compelled to moderate the cause and compromise his or her objectives.

3.2 Resources of Cause Lawyers

Cause lawyers must often rely on resources quite different from those available to the elite or core of the profession. When cause lawyers seek to mobilize, they may be heavily resource-dependent in at least five areas.

(a) *Money*: since little funding for cause lawyers comes from third world governments, they rely heavily on western governments, development agencies and churches (see Ellman in Sarat and Scheingold 1998);

(b) *Media*: publicity, including widely distributed articles, books, print and media interviews and stories, even public demonstrations and courts, increases the impact of otherwise isolated cause lawyers (see Bisharat in Sarat and Scheingold 1998);

(c) *External validation*: in Third World countries, funding by international foundations, partnerships with international non-government organizations (INGOs), and recognition by distinguished professional bodies, cannot be underestimated;

(d) *Professional solidarity*: where cause lawyers can ally with professional legal elites or law schools, they can obtain sponsorship, protection, and legitimation at the local level; and

(e) *Churches*: in some countries, such as Brazil during the military regime of the 1970s and 1980s (see Meili in Sarat and Scheingold 1998), the church and its human rights centers offer institutional solidarity to cause lawyers (Halliday 1999).

The cause lawyering perspective also perceives an ambivalent relationship with globalization.

'For cause lawyers ... globalization and state transformation may be a mixed blessing. In some contexts it may help to provide a political structure wherein minimal rights claims and legal defenses against abuses of state power become meaningful. In other contexts, it may erode support for the kinds of social changes which these lawyers, and the movements with which they are affiliated, seek to advance' (Scheingold 2001, p. 384).

4. Conclusion

Although they have quite different theoretical origins, the two principal approaches to lawyers' politics show some convergence. Political lawyers have been studied more in the past and cause lawyers more in the present, but in fact both can be found in every period. In both cases, three questions are central: how to explain the relationships between political orientations and lawyers' involvement; how to understand the collective construction of a defense or a cause; and how to produce collective action that will change institutional rules. The main difference between political and cause lawyers lies in the conditions which facilitate or hinder the sharing of a common political vision. Cause lawyers are usually individuals or groups defending a specific case in the context of an indifferent or hostile profession. If cause lawyering is very much a creature of, and a contributor to, legal-liberal forms of democracy, cause lawyers do so as a deviant strain within the legal profession and by making appeals for support of their cause outside the profession. By contrast, political lawyering is the expression of a common orientation of lawyers and therefore, through individual causes, it deals directly with other collective actors, especially justice and the state. The more the profession is diversified, the more its political commitment is likely to take the form of the defense of a cause; and the more it is united around some specific issues, the more it is likely to become a political actor. Thus, while the value orientations of these kinds of lawyers' politics may differ, they exist in creative tension to each other, and together may come to constitute the building and maintenance of liberal politics in the modern world.

See also: Affirmative Action, Sociology of; Citizenship and Public Policy; Citizenship, Historical Develop-

ment of; Citizenship: Political; Citizenship: Sociological Aspects; Equality and Inequality: Legal Aspects; Inequality; Justice and its Many Faces: Cultural Concerns; Law and Democracy; Lawyers; Legal Professionalism; Political Science: Overview; Public Interest Law; War Crimes Tribunals

Bibliography

- Abbott A 1988 *The System of Professions*. University of Chicago Press, Chicago
- Abel R L, Lewis P S C (eds.) 1988a, 1988b, 1989 *Lawyers in Society: Vols. I, II, III*. University of California Press, Berkeley, CA
- Halliday T C (ed.) 1999 Review symposium: Lawyers and politics. *Law and Social Inquiry* 24: 1007–124 (esp. essays by T. C. Halliday, S. Scheingold, M. Burrage)
- Halliday T C, Karpik L (eds.) 1998 *Lawyers and the Rise of Western Political Liberalism*. Oxford University Press, Oxford, UK
- Karpik L 1999 *French Lawyers: A Study in Collective Action, 1274–1994*. Oxford University Press, Oxford, UK
- Larson M S 1977 *The Rise of Professionalism*. University of California Press, Berkeley, CA
- Ledford K 1996 *From General Estate to Special Interest: German Lawyers, 1878–1933*. Cambridge University Press, Cambridge, UK
- Miller M C 1995 *The High Priests of American Politics: The Role of Lawyers in American Political Institutions*. University of Tennessee Press, Knoxville, TN
- Parsons T 1964 *The Social System*. Free Press, New York
- Sarat A, Scheingold S (eds.) 1998 *Cause Lawyering: Political Commitments and Professional Responsibilities*. Oxford University Press, New York
- Sarat A, Scheingold S 2001 *Cause Lawyering and the State in a Global Era*. Oxford University Press, New York
- Scheingold S 2001 Cause lawyering and democracy in transnational perspective. In: Sarat A, Scheingold S (eds.) *Cause Lawyering and the State in a Global Era*. Oxford University Press, New York

T. C. Halliday and L. Karpik

Political Machines

Though the term ‘political machine’ is sometimes used to describe any party or electoral organization noted for its efficiency and effectiveness, its principal significance, especially in American political science, is to refer to a particular kind of party organization, that draws much of its support from patronage jobs, government contracts, and other ‘spoils’ of office. This second type of machine once dominated the political landscape in most large American cities, though by the late twentieth century most such machines had declined or disappeared.

1. Two Definitions and Their Application

The term ‘political machine’ is used in two related, but distinct ways. Particularly in the 1950s and 1960s, it was often used to refer to any party or electoral organization distinguished by its effectiveness and efficiency (see, e.g., Gottfried 1968). Usually, such an organization was headed by a leader or small leadership group and imposed some measure of discipline on its members (otherwise it would not have functioned quite so effectively), but its most important quality was its ability to acquire, maintain, and use political power on a regular, predictable basis. Hence, the comparison between a human organization and a mechanical contrivance. When used in this sense, the term has a generally neutral connotation, and the study of political machines is simply one aspect of the larger subject of political party organization.

More often, however, the term political machine has been used to refer to a particular kind of party organization: one that derives its effectiveness from the extensive use of what organizational theorists call specific, material inducements. As laid out most clearly by Banfield and Wilson (1963), a specific (as opposed to general) inducement is one ‘that can be offered to one person while being withheld from others.’ Material inducements include money and any other ‘physical “thing” to which value attaches.’ Thus, a political machine attracts support to a much greater extent than other party organizations through the use of patronage jobs, government contracts, and special favors to individual voters. Because these sorts of inducements conflict with certain widely held civic norms, and because their use frequently leads to other forms of corruption, this second use of the word machine has a decidedly pejorative connotation. Indeed, those actively involved in political combat sometimes use the term as little more than an epithet. As Theodore Roosevelt once noted, one’s own organization is invariably described as a ‘club’ or a ‘party’; it is only the opposition that has a machine.

When used in this second sense, political machines have a particularly celebrated history in American politics—and American political science. Almost every large and medium-sized city in the United States was run by one or more political machines, often for extended periods of time, during the latter half of the nineteenth century and the first half of the twentieth century. Among the cities in which machines played a particularly powerful and significant role are Chicago, New York, Philadelphia, Jersey City, Kansas City, and Memphis. Moreover, since some of the most important work in the development of political science as an academic discipline dealt with city government and politics, machines also served as the principal subject matter for what are now regarded as some of the classic early works in American political science (see especially Merriam 1929 and Gosnell 1937).

It is difficult to say with much certainty how often and in what sorts of circumstances political machines have flourished, primarily because it is difficult to get reliable information about how much various organizations used patronage or graft and whether such practices were essential to their success or just an ancillary way of enriching the leadership. There is, as has just been noted, widespread agreement about the importance of machines in American urban history. Many commentators also believe that many rural counties in the South and Midwest were once run by machines, though there seems to be little beyond anecdotal evidence to support this contention. American historians and political scientists once regularly asserted that political machines were a uniquely American institution. Over the last thirty years of the twentieth century, however, it has been recognized that machine-like organizations also exist in a number of Third World or developing countries, as well as in specific locales or regions of other Western nations, such as southern Italy (see especially Scott 1969).

2. *Explaining the Rise of Political Machines*

Why did political machines play such a dominant role in American urban politics? Three principal theories have been put forward to explain the origin and survival of machines in American cities. For shorthand purposes, these theories can be described as governmental, economic, and cultural.

According to the governmental theory, machines came into existence as a solution to the unique governing problems that beset American cities. On the one hand, cities faced an extraordinary array of difficult challenges in the late nineteenth and early twentieth centuries. Their population and geographic area grew at an explosive pace. Partly as a result, and partly as a response to new technological discoveries, cities were suddenly required to provide an array of important services—water, gas, electricity, paved streets, sewer systems, mass transit—that no American city had previously had to worry about. Unfortunately, the structure of American urban government was dramatically unsuited to these tasks. Governmental power and authority in the cities was widely dispersed among the mayor, city council, and a large number of special boards, commissions, and independently elected officials. The special function of political machines, then, was that they provided an informal way of centralizing the political power that had been fragmented among the formal organs of city government (for the classic statement of this argument, see Merton 1957).

The economic explanation for machines, by contrast, argues that machines came to power because they helped serve the needs of the large number of poor and economically insecure people who lived in

American cities. When a city's population doubled in the course of just one decade, as often occurred in the late nineteenth century, the new residents necessarily consisted primarily of immigrants: sometimes from farms and rural areas within the United States, more often, especially in northeastern and midwestern cities, from foreign countries. In either case, the newcomers generally faced severe problems adjusting to their new environment and satisfying such basic human needs as food, housing, and employment. Yet, American government at this time provided little or no help at all in the way of welfare or social services. Political machines did not completely fill the gap, of course, but they were one of the few agencies that even tried, providing at least some of the urban poor with jobs, occasional gifts of food and fuel, and help when they came in conflict with the law. One early twentieth-century social worker went so far as to call the machine 'a kind of crude socialism.' In return, the lower classes, especially those of foreign background, provided a reliable fund of votes for machine candidates, and were not particularly upset if the resources the machine distributed were ultimately derived from illegal or corrupt activities.

The cultural explanation also sees foreign immigration as a key factor in the rise of urban machines, but places emphasis not on the immigrants' economic predicament, but on their political culture. Of the immigrants who came to America after 1840, few had any experience with democracy; almost none had ever cast a meaningful vote. On the contrary, most had learned to view government only as a hostile and authoritarian presence. Against that background, most immigrants naturally saw their votes less as a way of working large-scale transformations in the urban social order than as a commodity that could be exchanged for short-term personal advantages. One widely quoted formulation of this argument has been provided by historian Hofstadter (1955), who saw the clash between immigrants and natives as having been grounded in 'two thoroughly different systems of political ethics.' Where the middle-class natives 'argued that political life ought to be run ... in accordance with general principles and abstract laws,' immigrants 'interpreted political and civic relations chiefly in terms of personal obligations, and placed strong personal loyalties above allegiance to abstract codes of law or morals.'

3. *The Decline of the Urban Machine*

If political machines were once an all-but-ubiquitous feature of American urban politics, they are no longer. It is difficult to be precise about exactly when the machines lost their capacity to deliver votes on a regular and reliable basis for their favored candidates. In Chicago, for example, Mayor Richard J. Daley continued to win easy re-election all the way up until

his death in 1976, but it is questionable if the machine *per se* had much to do with his last several victories. A good case can be made that he won less because of the organization's precinct captains and patronage army than because most city voters genuinely thought he did a good job of governing the city. There is, moreover, no easy way to say who is and who is not a 'machine candidate.' Many self-proclaimed 'reformers,' once in office, would begin rewarding their supporters with jobs and contracts in ways that were largely indistinguishable from the machines they defeated.

That said, it was primarily in the years between about 1945 and 1965 that the most celebrated machines in American history suffered a series of major defeats and lost their capacity to dominate politics on a citywide basis. By the late 1960s, it was frequently asserted that of all the large cities in the United States, only Chicago still had a functioning political machine.

In general, when a machine started to weaken, the first major office it would lose control of was the mayoralty. Yet, long after many machines could no longer dominate these sorts of highly visible, media-centered campaigns, they could still reliably elect a number of their supporters to posts on the city council and to many of the less publicized offices in city and county government. In these weakened circumstances, the vestiges of machine politics linger on in a number of cities.

Why did the machines die out? Each of the three theories developed earlier to explain the growth of machines also provides a ready explanation for their decline. According to the governmental theory, machines died out because, over time, many of the worst problems afflicting urban government became more tractable. Cities stopped expanding, they gradually learned from experience how to deal with such basic demands as paving streets and providing water and electricity, and, in many cases, city charters were changed in ways that brought a greater measure of order and coordination to the formal agencies of government.

Advocates of the economic theory, by contrast, argue that machines went into eclipse because the demand for their services declined. On the one hand, a series of laws passed in the 1920s effectively shut off the flow of foreign immigration; those immigrants who were already here—and their children—gradually learned to adjust to their new environment and, especially after World War II, escaped poverty and even made it into the middle class. Equally important, with the dramatic expansion of the American welfare state in the 1930s, government now provided many of the goods and services for which the machines had once held a near-monopoly.

The cultural theory, like the economic theory, sees the passage of restrictive immigration legislation as an important, long-term threat to the machines' existence. It would also point toward agencies, such as the

public schools and the mass media, that were attempting to assimilate the immigrants' children and grandchildren and socialize them into a new set of political norms.

According to several of these theories, an important demographic trend of the 1950s and 1960s—the massive migration of American blacks from rural southern areas to northern cities—at least had the potential to revivify the machines. Like the European immigrants of the late nineteenth and early twentieth centuries, the black migrants faced enormous social and economic problems and had little previous experience with voting, and there is little doubt that their arrival in northern cities posed huge new problems for urban governance. Yet, for a variety of reasons, blacks proved to be an uncertain and unreliable constituency for the established machines. Racial antagonism and prejudice made many white ethnic groups unwilling to seek common cause with blacks—and made blacks unwilling to take their place as just one more group in the machine coalition. Equally important, the civil rights movement encouraged blacks to think of their votes as a means of effecting social change, and not just as a way of gaining particularistic benefits.

4. *Legacy and Enduring Significance*

How well did political machines govern American cities? To answer to this question, one must pose a further one: compared to what? It is, for example, easy to criticize the aid machines gave to the urban poor: such aid was clearly inadequate to the scope of the problem and was often given out in an arbitrary fashion. And it came at a price: most of the money that the bosses skimmed from the public treasury wound up in their own pockets, not as assistance to the less fortunate. Still, when compared to the contemporary alternatives—the almost complete absence of governmental aid and the paternalistic and demeaning nature of private charity—it is hard not to feel that the machines deserve a more favorable assessment.

At the very least, one can say that there is little compelling evidence to suggest that the machines governed worse than the 'reformers' they competed against—and some reason to think the opposite. All too often, reformers proved to be personally honest but largely ineffective, and to have strikingly little appreciation of the real problems and concerns of immigrants and the urban lower classes (see in particular (Holli 1969)

Good or bad, machines are simply not a very significant presence in contemporary American politics, national, state, or local. Nor is there any obvious reason to think they will be revived at any time in the foreseeable future. Yet, for several reasons, political machines are still worth studying. Most obvious is their historic importance: for at least a century, machines were one of the principal centers of political power and political conflict in almost every large city

in the United States. No one can hope to understand American urban history without coming to terms with the machine. Second, the struggle between machines and reformers left an enduring imprint on the structures of contemporary urban government. Many of the most common institutions in city government, including nonpartisan elections and the city-manager form of government, were originally developed as ways of undercutting the power of the machine. Long after the machines have declined or disappeared, these institutions remain.

See also: Campaigning: Political; Corruption: Political and Public Aspects; Participation: Political; Political Money and Party Finance; Urban Politics: United States; Urban Sociology

Bibliography

- Banfield E C, Wilson J Q 1963 *City Politics*. Harvard University Press, Cambridge, MA
- Gosnell H F 1937 *Machine Politics: Chicago Model*. University of Chicago Press, Chicago
- Gottfried A 1968 Political machines. In: Sills D L (ed.) *International Encyclopedia of the Social Sciences*. Macmillan and The Free Press, New York, pp. 248–52
- Hofstadter R 1955 *The Age of Reform*. Knopf, New York
- Holli M 1969 *Reform in Detroit: Hazen S. Pingree and Urban Politics*. Oxford University Press, New York
- Merriam C E 1929 *Chicago: A More Intimate View of Urban Politics*. Macmillan, New York
- Merton R K 1957 *Social Theory and Social Structure*. Free Press, New York
- Scott J C 1969 Corruption, machine politics, and political change. *American Political Science Review* 63: 1142–58

W. G. Mayer

Political Money and Party Finance

Healthy democratic politics depends importantly on a free flow of information among citizens, groups, candidates for public office, and political parties. The right to speak diminishes in significance without the financial resources needed to be heard. In the legendary words of California Assembly Speaker Jesse Unruh, 'Money is the mother's milk of politics.' Yet the manner in which political money is raised and spent can undermine the legitimacy of democratic politics. Ensuring that parties and candidates have adequate resources to compete effectively in elections is problematic given the inevitable tension between the reality of economic inequality and the ideal of political equality. Efforts to prevent concentrations of wealth from undermining political equality may conflict with bedrock freedoms of speech and association. Every democracy struggles to reconcile the need for political

money with the problems it begets. None have found entirely satisfactory and long-lasting solutions. The political finance tools available to policy makers—public subsidies, limits on contributions, expenditure controls, disclosure, and regulation of campaign activity—often leave reformers with unfulfilled objectives and unanticipated consequences.

1. Institutional Context

The prominence and nature of money in politics varies across institutional settings. Party-centered electoral systems differ substantially from candidate-centered systems. In the former, most often associated with the parliamentary systems of Western Europe, political money means party finance. Most, if not all, political funds are raised and spent by political parties and the primary concerns have to do with the adequacy of funds available for routine party activities, the work of party caucuses in parliament, and election campaigns. Direct and indirect public subsidies are often provided to the parties, including cash and free broadcasting time on state-owned networks. There is substantial variation in the extent of regulation of party finance, with disclosure of large private contributions and prohibitions on paid political advertising most typical and limits on contributions the least typical. Party-centered systems have not been immune from escalating campaign costs and scandals, as the Clean Hands investigation in Italy and the more recent charges against Helmut Kohl's Christian Democrats in Germany illustrate. But the problems associated with political money in party-centered systems are relatively contained and finance practices are generally viewed as in need of monitoring and repair, not radical surgery.

In candidate-centered systems, by contrast, the focus is less on political party activity and more on campaign fundraising and spending by individual candidates. In these systems, the term of choice is campaign finance rather than party finance. The United States is, of course, the most prominent example of this class of democracies and the one most atypical in its various components. While parties once dominated the recruitment of candidates and the waging of campaigns in the US, a host of developments, including the creation of a professional civil service, the spread of the direct primary, the rise of television, the reform of party rules, and the provision of substantial staff resources to public officials, have diminished their role. Party organizations in the US have become largely appendages of incumbent officeholders—more instruments for raising money within and beyond the constraints of the law than independent political actors. Political money in the US and other candidate-centered systems is widely viewed as highly problematic and much time and energy are devoted to reporting on it and debating how best to ameliorate its harmful effects.

The demand for and supply of political money are shaped by more than the extent of party or candidate-centered elections. In the US, the sheer number of contests across the federal system, the separation of executive and legislative elections, plurality rules, constitutional protections of free speech, the high cost of reaching the large American audience, the professionalism and length of campaigns, and the permeability of the policy process all contribute to the large sums raised and spent in election campaigns. The perceived stakes of elections, the incentives and opportunities for raising and spending private resources in campaigns, and the laws and norms governing its role all condition how political money flows in democracies.

2. Problems Associated with Political Money

Rising campaign costs, scandals, and equity concerns in the early and mid-twentieth century precipitated waves of regulation of political finance and subsidization of party and campaign costs in democracies around the world (Paltiel 1981). Similar concerns kept issues of the regulation and subsidy of political finance on the public agenda in more recent years. Problems arise as parties and candidates struggle to raise adequate resources to compete effectively in times when those costs have risen sharply and when groups and individuals seek to influence election outcomes and policy decisions. At one extreme lies bribery and extortion, the illegal linkage of private resources and public authority for personal gain. Short of such explicit *quid pro quo* exchanges lie such problems as the dependence of parties and candidates on large financial contributors, conflicts of interests arising when politicians receive contributions from those who seek favorable action by a legislative or administrative body, and the impact on democratic legitimacy of the appearance of special access and influence by financial contributors. Then there are the consequences of the money chase—the unending scramble for political money—for how party and public officials allocate their time, who they see and listen to, and how both affect the deliberative processes of government.

Another set of problems are associated directly with the electoral process. Rising campaign costs and inadequate political funding depress electoral competition, discourage able people from seeking public office, give a distinct advantage to incumbent officeholders and wealthy self-financed candidates, and weaken democratic accountability by muffling the voices of individuals, groups, candidates, and parties.

A final set of problems arise from the very act of trying to cope with the others. Reforms designed to alter the flow of money in elections often produce equally problematic unanticipated consequences. Political money is fungible and legal constraints can easily divert it to less accountable passageways. Ambitious politicians naturally exploit the weaknesses and push

the boundaries of the law, threatening the effectiveness and credibility if not the complete collapse of the regulatory regime.

3. Political Finance Tools

Policy-makers have utilized a conventional toolbox in responding to the problems associated with political money, but the mix of tools and particular applications have varied widely among democracies.

3.1 Public Subsidies

In response to concerns about rising campaign costs and excessive dependence on financial contributors, public support of parties and campaigns has become nearly universal among democracies (Katz 1996). Subsidies take the form of direct cash assistance, the provision of goods and services (e.g., free broadcast time, mailings, voter brochures, and meeting space), and indirect support such as tax credits or deductions for contributions or reduced rates for political advertising. The subsidies are given to parties or candidates and directed toward party maintenance or campaign expenditures. Direct cash subsidies to parties are typically allocated in proportion to electoral strength while services and indirect subsidies are usually provided on a more equitable basis. In the United States, public matching funds are provided to candidates for their party's presidential nomination who raise the requisite number of private contributions and agree to limit their spending; full public grants are provided to the major party candidates for the general election as long as they agree to raise and spend no private funds. But unlike most other democracies, the US provides no free broadcasting time for parties or candidates and no public subsidies for candidates for Congress.

Public subsidies are not without risks. They can reinforce the dominance of established parties, increase the potential for control by the state, and weaken the connection of parties with their grassroots. Public financing conditioned on very low qualifying thresholds can also encourage fringe and patently offensive candidates and parties. Full public financing systems generate intense pressure for private resources to find alternate routes to influence elections and policy.

3.2 Contribution Limits

While public subsidies are the norm in political financing, limits on contributions are less common. Many democracies place no restrictions on the source or amount of contributions to parties or candidates. Among those that do, prohibitions on particular sources (e.g., civil servants, foreign donors, corpor-

ations, and unions) are most common. In the United States, corporations and unions have been precluded for many decades from making political contributions from their treasuries but they may form voluntary political action committees (PACs), through which they may contribute to parties and candidates. In the late twentieth century, even prohibitions on contributions from corporate and union treasuries have been rendered ineffectual through the emergence of 'soft money' donations to political parties not subject to federal regulation.

Some countries have placed statutory limits on the size of contributions from permissible sources, including France, Israel, Japan, and the United States. The contribution limits in the US are notably low, especially since they were not indexed for inflation when fixed by law in 1974. This decline in the real value of donation ceilings contributes to the fundraising demands on politicians and pressures for private donations to flow through other channels, problems associated more generally with contribution limits.

3.3 Expenditure Limits

A number of democracies place limits on spending by parties and candidates, although these limits often apply only to a brief period of time (e.g., the formal campaign in parliamentary systems), to candidates but not parties (or vice-versa), or they exempt entire categories of political activity that influence election outcomes. In Japan, candidates for the Diet face limits on campaign expenditures but these are largely irrelevant because they apply only to the 12-day formal campaign period. For well over a century, Britain limited the expenses individual candidates for Parliament could incur in their constituencies (and prohibited outside groups from spending on their behalf) but did not regulate political party expenditures at the national level, where most of the action is. However, a new law adopted in 2000 limits national campaign expenditures by political parties; it also limits the amount of spending by outside organizations and individuals for political activity during a general election.

Mandatory spending limits adopted in the US in 1974 were declared unconstitutional, while voluntary caps tied to the receipt of public funds in presidential elections remain intact. But presidential candidates, political parties, and interest groups have managed to work around these expenditure limits in the general election by fashioning political advertisements (known as 'issue advocacy' but indistinguishable from traditional electioneering) that the courts have ruled are not subject to regulation.

3.4 Public Disclosure

Transparency in the flow of political money is embraced by many democracies as a deterrent to cor-

ruption or influence peddling but the comprehensiveness and effectiveness of disclosure regimes vary widely. At one extreme are systems requiring only the filing of a single report on party expenditures well after the election; at the other are systems like the US with elaborate rules for reporting political contributions (as small as \$200) and expenditures and for making this information publicly available on a timely basis. Yet even in the US, a significant amount of electioneering falls outside the boundaries of disclosure law. While disclosure of political money is gaining in acceptance in the democratic world—Britain, for example, broke with its longstanding practice in 2000 by requiring that large donations to political parties be publicly disclosed—some Scandinavian countries continue to consider reporting requirements as an unacceptable violation of privacy.

3.5 Regulation of Campaign Activity

One way of reducing the demand for political money is to prohibit certain costly campaign activities. A surprisingly large number of countries ban paid political advertising on television, although most couple the ban with the provision of blocks of free television time to the parties. Such prohibitions are unthinkable in the United States, where free speech guarantees preclude any significant restraint on political communication. Other countries have moved to assert that a legal protection of freedom of expression limits the extent to which campaign activity can be regulated. For example, the Australian High Court declared a 1991 law that prohibited the broadcasting of political advertising during election periods unconstitutional, on grounds that it infringed the implied right to freedom of political communication.

4. Enforcement and Effectiveness

The success of any measures designed to redirect the flow of political money depends upon the ingenuity, comprehensiveness, and timeliness of the particular mix of regulatory and subsidy measures and on the mechanisms for enforcement. Short of blatant violations of criminal statutes (e.g., bribery or extortion), departures by parties and candidates from the letter and spirit of campaign finance laws often go unpunished. Politicians are reluctant to invest the requisite authority in public agencies to enforce political finance laws, a phenomenon tellingly illustrated by the structure and operation of the US Federal Election Commission. The FEC, whose six members include three from each of the two major parties, operates under very tight statutory and budgetary reins from Congress and often deadlocks on regulatory decisions. Putting in place a reasonable and workable set of regulations, appropriate criminal and civil sanctions, and administrative and judicial mechanisms has proven a daunting task in many democracies.

For all of these reasons it is not surprising that political finance reforms often fail to reduce corruption, lessen the influence of monied interests, slow the increase in campaign expenditures, or increase electoral competition. Other institutions and policies are more effective levers for combating corruption and enhancing democratic accountability. These include broad economic policies that limit the discretionary authority of public officials; institutions and practices nurturing the rule of law; an independent judiciary; adequate pay and strong ethical requirements for politicians and civil servants; transparent budget and financial control systems; and party systems supporting stable patterns of competition, accountability, and alternation of power (Kaufmann 1997, Pope 1996).

Not all political finance measures are doomed to fail, nor are they miracle cures for the ills arising from money in politics. But since neither the abolition of all private funds from politics nor the complete deregulation of political finance are feasible or sustainable, democracies will continue to struggle not to solve these problems once and for all but to contain and manage them as best as they can.

See also: Campaigning: Political; Electoral Systems; Electronic Democracy; Lobbying; Mass Media: Introduction and Schools of Thought; Mass Media, Political Economy of; Media Effects; Participation: Political; Parties: Litigants and Claimants; Party Systems; Policy Process: Business Participation; Political Machines; Voting, Sociology of

Bibliography

- Alexander H E, Shiratori R (eds.) 1994 *Comparative Political Finance Among the Democracies*. Westview Press, Boulder, CO
- Corrado A, Mann T E, Ortiz D R, Potter T 2001 *The New Campaign Finance Sourcebook*. Brookings Institution Press, Washington, DC
- Gunlicks A B (ed.) 1993 *Campaign and Party Finance in North America and Western Europe*. Westview Press, Boulder, CO
- Katz R S 1996 Party organizations and finance. In: Leduc L, Niemi R G, Norris P (eds.) *Comparing Democracies: Elections and Voting in Global Perspective*. Sage Publications, Thousand Oaks, CA, pp. 107–33
- Kaufmann D 1997 Corruption: the facts. *Foreign Policy* 107: 114–31
- Paltiel K Z 1981 Campaign finance: Contrasting practices and reforms. In: Butler D, Penniman H R, Ranney A (eds.) *Democracy At the Polls*. American Enterprise Institute, Washington, DC, pp. 138–72
- Pope J (ed.) 1996 *The Transparency International Source Book*. Transparency International, Berlin, Germany
- Report for Congress 1995 *Campaign Financing of National Elections in Selected Foreign Countries*. Law Library, Library of Congress, Washington, DC
- Sorauf F J 1992 *Inside Campaign Finance: Myths and Realities*. Yale University Press, New Haven, CT

T. E. Mann

Political Parties

1. Defining Characteristics

Political parties are a form of political grouping. The term is commonly applied in different political systems to widely different bodies, ranging from the loose, decentralized, and undisciplined associations of the American Democratic and Republican parties to the tightly bonded, highly centralized, and rigidly disciplined Communist and Nazi parties of recent memory. It therefore seems appropriate to state the characteristic that appear to qualify these variegated groupings as ‘parties’:

(a) They are groups of people to whom labels—‘Republican,’ ‘Labour,’ ‘*Rassemblement pour la République*,’ ‘Baath Socialist Arab Rebirth,’ and so on—are generally applied by themselves and others.

(b) Some of the people are organized—they consciously act together to achieve common goals.

(c) The larger society recognizes as legitimate the right of one or more parties to organize and promote their causes.

(d) In some of their goal-promoting activities, parties work through the mechanisms of representative government, such as nominations and elections.

(e) A key activity of parties is thus selecting candidates for elective public office.

Most scholars agree that, taken together, these characteristics distinguish modern parties from such historic precursors as the patricians and plebeians of the Roman Republic, the Guelfs and Ghibellines of medieval Italy, the ‘hats’ and ‘caps’ of eighteenth century Sweden and the Whigs and Tories of eighteenth century England. In his classic *The Structure of Politics at the Accession of George III*, for example, Sir Lewis Namier (1929) pointed out that ‘Whig’ and ‘Tory’ were loose labels sometimes called ‘parties,’ but, since it was uncertain which label fitted many MPs, the groupings were more ‘tendencies of opinion and personal loyalty’ than fully-fledged political parties.

The last of these defining characteristics most clearly distinguishes parties from interest groups. Typical organized interest groups, such as the American Rifle Association and Handgun Control, Inc., support, finance, and sometimes organize campaigns for political candidates who take their respective positions on gun control, but they do not *nominate* candidates who will bear their labels on general election ballots. Only parties do that.

1.1 ‘Party’

There are three levels of meaning. Many scholars of American parties have embraced a suggestion made in the first edition of V. O. Key’s *Politics, Parties and Pressure Groups* (1958 [1942]). Key observed that when most Americans talk about a particular party or

parties, they have in mind one or more of three levels of meaning of the term.

(a) *Party-in-the-electorate* refers to the degrees of attachment to particular parties held by ordinary people (a concept later developed into 'party identification,' operationalized and analyzed by the authors of *The American Voter* (Campbell et al. 1960)). People who feel some attachment to, say the Democratic party do not necessarily work for the party, contribute money to it, or even vote for its candidates.

(b) *Party organization* denotes people who hold offices in the parties or contest elections as their nominees or otherwise work for them in election campaigns.

(c) *Party-in-government* means persons who are elected or appointed to public offices commonly associated with a particular party—for example, the president, members of the cabinet, subordinate administrators serving at the pleasure of the president, members of the party's congressional party caucuses and their leaders, and the like.

There are some occasions, Key said, when some people use the term 'Democratic party' or 'Republican party' to include all three levels of meaning, although they are rarely explicit—or perhaps aware—of what they are doing.

When considering what people say about the Democratic and Republican parties, Key advised, it is well to be as precise as we can about just what meaning of 'party' we have in mind. Most political scientists (though not most journalists) concerned with the major American parties have followed his advice.

2. Party Types

Each party in any democratic system is not the same in all respects as every other party. They differ mainly along the lines covered in Sects. 2.1–2.3.

2.1 Membership

Membership in a party carries with it the right to participate in the selection of the party's candidates for public office, and it may or may not also involve the assumption of some kind of obligation to the parties. Most of the parties in most democracies have memberships that are to some degree closed: the parties' rules in such places as Great Britain, Europe, and Scandinavia usually require aspirants to apply for membership, and party officials can accept or reject the applications (they usually accept them). To continue as a party members in good standing, the person must at least pay annual party dues and perhaps also take an oath of loyalty to the party's principles and candidates. Membership in US parties, however, is open: there is no formal application for membership, Democratic and Republican presidential and other candidates are selected in direct primaries, and the

right to participate in those primaries is guaranteed by law to all registered voters. In 'closed' primaries the voters are required to register their party preferences before the election, but they and they alone control their choice of parties. In 'open' primaries they do not make any public statement of their party preferences. In 'blanket' primaries they are allowed to vote in the primaries of either or both parties, although only for one candidate for each office.

2.2 Discipline

Most human organizations have some form of discipline, in which the leaders control the members by dispensing rewards and imposing punishments for the members' behavior. The leaders of every democratic party have some disciplinary powers, but some are more effective than others.

The leaders of the British Conservative and Labour parties, for example, have strong disciplinary powers. The members of their parliamentary parties know that when 'the whip is laid on'—when the leader, through the party whips, has told the members how they should vote on a particular bill—the MPs put their political careers in great jeopardy if they disobey. The prime minister, who is leader of the party in power, controls who gets ministerial offices constituting the 'government,' and the leader of the opposition controls who gets the positions in the 'shadow cabinet.' Moreover, the leader of each party controls its national organization, which has the power to deny a rebel MP reselection as the party's official candidate in each constituency and thereby effectively end the MP's political career. Perhaps the prime minister's greatest disciplinary power stems from the fact that if his or her government is defeated on a major parliamentary vote, a new election must be held, which the party—and any of its MPs—might lose. Thus an MP's vote against the whip's instructions is, in effect, a vote to risk putting the other party in power—a powerful deterrent against defections by the majority party's MPs.

US presidents are leaders of their parties (the party that does not hold the presidency has no generally acknowledged single leader), but their disciplinary powers are much weaker than those of their British counterparts. Presidents can make campaigning for one of their party's candidates contingent upon that candidate's support of presidential programs. However, most congressional candidates' electoral success depends far more on local factors than on presidential support. Indeed, it is usually an electoral asset for a congressional candidate to be seen as independent from any 'outside' party leader, including the president. Presidential discipline in American parties thus consists of little more than asking and cajoling.

Most European and Scandinavian parties give their national organizations the power to expel from the party members of the national parliamentary parties

for bolting the party line. While such expulsion is somewhat less likely than in Britain to end a parliamentarian's political career, it is still a much greater weapon than any available to the leaders of the American Democrats and Republicans.

2.3 Cohesion

Cohesion, defined as the extent to which party members holding public office act together on policy matters, is a critical party characteristic for people concerned with party responsibility. Parties in democratic countries vary considerably in cohesion. The British Conservative and Labour parties are among the most cohesive: on almost every issue before the House of Commons the members of the parliamentary majority party vote together without exception—and usually with unanimous opposition from the other parliamentary parties.

At the other extreme, some small parties in France split in parliament on almost every issue. Other democratic parties fall between these two extremes. The larger parties in Australia, Canada, and New Zealand, as well as the leading French, German, and Scandinavian parties, are almost as cohesive as their British counterparts. The American parties in Congress fall somewhere in the middle, often mustering near-perfect cohesion on such matters as the organization of the chambers, election of presiding officers, large tax cuts, increasing social welfare expenditures, and other major issues where party positions are well entrenched. Usually one or both parties split internally on other issues such as foreign aid, abortion, and free trade.

3. Party Functions

Every democratic country has political parties at the national level, and most have them at the regional and local levels as well. Moreover, all of the nations that began their existence as democracies or replaced some form of authoritarianism with democracy have also had parties from their democratic beginnings. Evidently, then, democracy and parties are in some fashion intertwined.

Why? Most scholars would say because parties perform at least two critical functions for democratic government that no other form of political organization does or can.

3.1 Organizing Elections

Popular elections of the government's top public officeholders are traditionally the main democratic device by which a sovereign people choose their

officeholders and hold them accountable. But elections, especially in democratic nations with millions of citizens, must be organized somehow. For instance, 130 million Americans fulfill the constitutional requirements for the US presidency, and theoretically the American people have the right to pick the one person they most prefer. But that right is meaningless, for no person or electorate can choose among 130 million alternatives. Some agency must aggregate the choices and reduce the 130 million to a manageable number, say, 20 or three or two. That vital function is performed in every democracy by the political parties' activity of candidate selection. Theoretically it could be performed by another agency, such as the mass media or random selection by computers. In fact, however, democratic governments have always left the task mainly to political parties.

In addition, by putting the candidates' party labels on ballots the parties provide important cues that help voters make some sense out of the 'booming, buzzing confusion' of names, offices, and issues that elections present to them—especially where there are long ballots filled with many options. Some political scientists believe that prominent party labels explain why voter turnout in partisan candidate elections is usually much higher than in nonpartisan elections or in referendums.

3.2 Organizing Government

In every modern democracy elected officials of each party in each governing body join together in some kind of party organization. In each national legislature, regional assembly, or local council there is a comprehensive organization of all the party's legislative officeholders, often called the 'party caucus' or the 'parliamentary party.' That group usually selects some of its members as a guiding committee and one of its members as the party leader.

In the two-party systems elements of the majority party's organization become the government's 'ministry,' 'cabinet,' and 'prime minister.' In multiparty democracies the leader of one of the parties negotiates with other parties' leaders to put together a multiparty coalition government with some top positions going to leaders of all the participating parties. In both kinds of democracies the parties provide the organization that not only makes possible meaningful elections but effective and accountable government as well.

Some parties and party systems may perform these functions better than others, but even the least effective makes possible the coherent government and popular accountability that democratic government must have in modern populous nations.

See also: Civic Culture; Communist Parties; Electoral Systems; Participation: Political; Party Identification;

Party Responsibility; Party Systems; Political Culture; Political Elites: Recruitment and Careers; Political Machines; Political Money and Party Finance; Political Parties, History of

Bibliography

- Campbell A, Converse P E, Miller W E, Stokes D E 1960 *The American Voter*. Wiley, New York
- Duverger M 1954 *Political Parties, Their Organization and Activity in the Modern State* (trans. North B, North R). Methuen, London
- Eldersveld S J 1982 *Political Parties in American Society*. Basic Books, New York
- Epstein L D 1967 *Political Parties in Western Democracies*. Praeger, New York
- Epstein L D 1986 *Political Parties in the American Mold*. University of Wisconsin Press, Madison, WI
- Keefe W J 1998 [1972] *Parties, Politics, and Public Policy in America*. 8th edn., Congressional Quarterly Press, Washington, DC
- Key V O 1958 [1942] *Politics, Parties and Pressure Groups*. 4th edn. Crowell Company, New York
- Merriam C E, Gosnell H F 1949 [1922] *The American Party System: An Introduction to the Study of Political Parties in the United States*. 4th edn. 1949 Macmillan, New York
- Michels R 1958 *Political Parties: A Sociological Study of the Oligarchical Tendencies of Modern Democracy*. (trans. Paul E, Paul C). Free Press, Glencoe, IL (1st edn. in French, E. Flammarion, Paris, 1914)
- Namier L B, Sir 1929 *The Structure of Politics at the Accession of George III*. Macmillan, London
- Ostrogorski M 1902 *Democracy and the Organization of Political Parties* (trans. Clarke F). Macmillan, New York
- Ranney A, Kendall W 1956 *Democracy and the American Party System*. Harcourt, Brace, New York
- Rokkan S 1970 *Citizens, Elections, Parties; Approaches to the Comparative Study of the Processes of Development*. McKay, New York
- Sartori G 1976 *Parties and Party Systems: A Framework for Analysis*. Cambridge University Press, New York
- Sorauf F J, Beck P A 1992 *Party Politics in America*. 7th edn., HarperCollins, New York

A. Ranney

Political Parties, History of

The definition of a political party requires a historical perspective. The idea of a party is multifaceted and developed over time, in the light of the prevailing political, social, economic, and (especially) legal realities in which it is formed. In current language—and in the specialized terminology of the social sciences—the term ‘party’ has been in use for longer than the term ‘social class’. Etymologically, the term derives from the verb *partir*, which, in archaic French, meant ‘partake,’ or ‘share.’ The present article regards political parties as formal civil organizations that are

active in public life, in which they represent specific class, professional, or ideological interests, within the framework of the modern constitutional State, and whose goal is to run for election and, upon winning, to take power as the government.

1. Approaches and Definitions

In the past the expression ‘political party’ was also used to describe other political phenomena. Throughout the history of societies and of their political bodies, since antiquity, the expression to ‘be party to’ affairs of public life has been considered a common mode of human behavior. The translators of Aristotle use the term ‘party’ to designate opposing social groups in Athens. Specialists on ancient Rome have used the concepts of ‘Plebeian Party’ and ‘Patrician Party,’ and the famous *De viris illustribus Romae* relates to the party disputes that dilacerated the Roman Republic. The supporters of the Crown and the defenders of the rights of the nobility were also parties. Critics of the theological hegemony of the church and of the clergy were parties within the context of the political order of Christian Europe. Noble families in opposition to the Crown were parties in the making; religious and political reformers were parties. These are broad meanings of the term that are frequently to be found in the literature. Also under the broad heading of ‘parties’ can be categorized all manner of social organizations that seek political ends, even if they never stand for (or win) elections, as is the case for the mini-parties or revolutionary organizations that have no constituency among the electorate and that do not accept the validity of the electoral system, or even the dominant groups in totalitarian states. Thus, the expression ‘political parties’ is also often applied to interest groups; propagators of ideologies, doctrines, and political positions; political model makers; and mediators between citizens and government, including a vast multiplicity of bodies that participate in public life.

There are three nuances of the term ‘party’ from the historical point of view: rival groups, orders, states, or dynasties; theological or ideological critics; and modern formal social organization. This third interpretation has, in part, inherited the historical characteristics of earlier forms, making it possible to make such distinctions as conservative and liberal parties; religious and lay groups; corporative interests and those of the masses; revolutionaries and reformers; left and right; center and radical; and progressive and reactionary. Such distinctions are not exclusive, are directly proportional to the degree of institutional solidity and of integration which the parties enjoy in political society, and always relate to the ideas that separate—or unite—individuals or groups. Such multiplicity is evident in the classification of parties, according to the ideological density of their manifestos, the form of internal structure, competitive or

noncompetitive organization, and of party systems (single-party, two-party, or multiparty; unipolar, bipolar, or multi-polar) concerning the political and legal frame they are put in. Unlike political science, historical research does not treat these issues in isolation, but rather, seeks to reconstruct how all such aspects relate to each other.

Political parties—in the contemporary sense of the term—became common in the nineteenth century, alongside the development of representative institutions and the suffrage, becoming gradually universal, especially in Western Europe and the USA. In the aftermath of the Congress of Vienna (1814–15) political parties had not yet been formally recognized as civil associations. Any association, assembly, or grouping united by specific goal or political interest was regarded as a party. Conversely, the pursuit of private interests and the application of abusive pressures was regarded as a constraint upon political life, as the goal of winning elections and of influencing government and public policy became the established criteria for identifying a group of citizens as a political party. In the USA, by the 1830s, parties as such were fully active. In the Western European states, it was in the final decades of the nineteenth century and the first decade of the twentieth century that this development took place. Modern political parties are full players in the public life of democratic societies. In effect, only constitutional states formed and consolidated over the past two hundred years have the institutional framework necessary to ensure regular political participation on the part of their citizens. Such a framework determines the legal principles for party actions, articulates the position of the parties in relation the system of government and the electoral system, and places them within the framework of civil law. Their social origin and ideological orientation notwithstanding, it was the placing of parties within a legal framework that created the party system—understood both as a juridical and political frame of the participation of citizens in political life.

Parties began to elicit attention from philosophers and the interest of modern historians from the time of the English revolutions of the seventeenth century, on through the revolutionary turmoil of 1848, and up until the death of Marx in 1883. They were observed by Bolingbroke and Madison, Hume, Burke, Benjamin Constant and Blüntschi, and the German social-democratic thinkers. At the turn of the twentieth century, party-political phenomena began to be the object of serious scientific methodological analysis. The earliest works in this field were published after the consolidation of parliamentary forms of democratic government, which had been in place since the mid-nineteenth century. The pioneering works of Ostrogorski (1902) and Michels (1989) examined the functioning of parties in the late nineteenth and early twentieth centuries in the UK, France, and the USA. These authors sought to put forward a general theory,

beginning with highly particular experiences and specific historical developments, such as the establishment of federative republicanism of the USA or the burgeoning power of Parliament in the Government of the UK. There are many descriptive and historical studies on political parties in specific countries and also comparative studies which seek to define possible models. Research trends generally give greater emphasis to the western model of representative democracy, in which formalized parliamentary, electoral, and political-party systems play an essential role in public life and the practices of Government (Epstein 1993; Budge and Keman 1990, Fenske 1994).

With Max Weber, who couches his remarks on parties from a sociological standpoint, modern political parties came to be understood as a product of democracy and of universal suffrage. Jean Charlot (1971), according to Lapalombara and Weiner (1966), provides a systematically historical view of parties. In order for a group to be regarded as a political party, it should fulfill four requisites. It must (a) possess a visible and stable organization, for the articulation of its local, regional, and national components; (b) ensure continuity of organization, independently of its current managers; (c) have the explicit goal of achieving and holding decision-making power, either alone or as part of a coalition with others, at all levels; and (d) have organizational efficiency, with a view to capturing the vote and harnessing popular support. It is, however, a work by Duverger (1951) that systematizes and compiles knowledge on political parties. Two of his research elements remain valid: one puts forward a theory of the origin and of the proliferation of political parties, from an institutional perspective, relating to the party phenomenon and the dynamics of institutions, whereas the other resides in a typology of parties based upon the nature of their organization. After World War II, the discredit and mistrust with which people regarded the traditional political parties of Europe—owing to their failure to avoid the emergence of fascism and totalitarian dictatorships in Italy, Germany, Spain, and Russia—hampered the reintroduction of democracy, and parties were left fragmented and disunited. The understanding came up that healthy democratic life requires political parties within a system of government that establishes clear rules for the formation of governments to exercise the will of the majority, solely on the basis of the judgment expressed through the popular vote. The political instability of the Fourth Republic in France, and of the government of Italy, for example, provided Duverger with a rich vein of material upon which to base the most ambitious essay written on the theme of political parties in the mid-twentieth century. Duverger regards the parties of the masses as the inevitable product of universal suffrage and regards many of the residual parties as being archaic, especially in postwar France. His view is supported by the movements for ideological concentration in European

public life at the onset of the Cold War and has as its inspiration the two-party systems—which he regarded as being successful—in the USA, UK, and West Germany.

Although there is still a lack broad-based comparative studies, the contributions of Robert Dahl (1966), Lipset and Rokkan (1967), and Epstein (1993), take steps in this direction. The ideological nature of parties is given an innovative treatment by Stein Rokkan, comparable with the contribution of Duverger on party organization. The analysis of party systems was advanced by Sartori (1976) and Blondel (1978), whereas electoral sociology—beginning with the pioneering work of André Siegfried on electoral statistics and geography (1913, 1927), by Richard Rose (1984), David Butler (1981), and Alain Lancelot (1988). More recently, Daalder and Mair (1983), Pierre Avril (1990), Klaus von Beyme (1985), and Daniel-Louis Seiler (1986) have sought to formulate studies from an institutional and systemic viewpoint.

2. Stages of Development (Phases) Change Over Time

The first expressions of party life began with the defense of specific interests, although in different contexts and times, such as those of the aristocracy against the power of kings and governments, or of the working class against the political, economic, and social dominance of ruling groups or elites. Such defense was not passive. It was carried out by means of consistent and prolonged actions within the representative institutions, such as the estates, consultative assemblies and parliaments. It also comprised the revolutionary movements in which the respect for rights—both effectively held and merely claimed—was demanded of the political authorities, the governing establishment or the despot.

A party emerges when a political current within society finds expression through a specific (national or regional) organization. The first steps towards the organization of a political group can be observed through the 'parties' into which England polarized in the mid-seventeenth century. Although these parties were still characterized by the discipline and discourse of religion, the political effect of the change from one system of government to another imposed by those in opposition to the power of the throne, ushered in a new form of the exercise of power within society. It goes without saying that the great revolutions of the seventeenth and eighteenth centuries—in England, France, and the USA—also gave rise to factions which the majority of historians would call 'parties.'

The typical action of parties in modern societies emerged out of the political consciousness that defined and established the notion of a universal citizenship, evolves in the new space for political protagonists (both for and against the State), the emergence of

organized political parties during the nineteenth century and investiture of political parties as being the only protagonists capable of adopting practices, expressing ideas, and taking power and exercising public office within the State. The parties of the bourgeois in the early nineteenth century protagonized opposition between the aristocracy and the monarchy; a similar process can be observed among the emergent revolutionary or proletarian parties of the end of the same century. In revolutionary France there were embryonic experiences of political-party organization: the Girondes, Montagnards, and the Jacobins. The greater part of these 'parties' consisted of small groups, without formal organizational structures. It can thus be stated that no formal parties existed prior to 1820. Three phases in the development of party systems can be identified in the nineteenth century: the earliest phase between 1820 and 1830, an intermediate phase in the 1860s, and a final phase around 1880.

In the 1820s and 1830s, the British Tories and Whigs emerged and consolidated long-held positions. At the same time, in many countries of Western Europe, and notably in France, there was a movement towards a more constitutional state, expanding suffrage and consolidating the system through application the principle of majority rule in the formation of governments. It was in this phase that the earliest political parties emerged in Latin America, as in the case of the *Partido Liberal* in Colombia (1815) and the *Colorado* and *Nacional* parties in Uruguay (1836), which arose as an immediate consequence of the wars of independence against Spain and of the ideological influence of the USA. In the intermediate phase, there was a clear expansion of political parties: the Conservative (1840) and the Liberal (1860) parties in Brazil; the Liberal Party in Belgium (1846); the Christian-Democrat and Radical-Democrat parties in Switzerland (1848); the Democratic (1848) and Republican (1856) parties in the USA; the Liberal Party in Italy (1848); the *Partido Radical* in Chile (1863); and the Conservative (1854) and Liberal (1867) parties in Canada.

In Latin America, where most countries had thrown off colonial rule in the first decades of the nineteenth century, only gradually did political parties begin to gain importance. As of 1850, the national aspirations of countries of South America came under the spreading influence of North American two-party system. The first expression of political life was the affirmation of a national identity. The main banner of the 'political parties' up until the mid 1800s was nationalism, defended as the means whereby these new states could affirm their independence in face of their former colonial masters. The 1880s witnessed the emergence of 'classical' political parties, generally in twos, and of either conservative or liberal persuasion—but always nationalist—in the former Spanish colonies. It should be noted that the emergence of political parties did not necessarily mean that they assumed a decisive role in political life. Under the Brazilian Empire (1822–89), a

two-party system existed as a means for the rural aristocracy to maintain its hold on power and, under the Republic proclaimed in 1889, this oligarchic system remained essentially unchanged.

In the mid-nineteenth century, growing social tensions and bitter social criticism led people with a political outlook to express their social and economic (and occasionally revolutionary) demands through formal political action, by means of the creation of political parties. Thus began the 'socialist' or 'social-democrat' phase, typified by formation of the Danish Social-Democratic Party (1871), the Spanish Socialist Party (1879), the German Social-Democratic Party (1863, 1869, 1875), the Swiss Socialist Party (1880), the Belgian Workers Party (1885), the Indian National Congress (1885), the Austrian Social-Democratic Party (1889), the Swedish Social-Democratic Party (1889), the Italian Socialist Party (1892), and many others.

World War I forced a pause for reflection and redefinition of the party systems in Europe. On the one hand, the victory (1917) of the Bolshevik party (1903), which upon seizing power set up a revolutionary and internationalist government in Russia, and, simultaneously, the conservative bias of the Treaty of Versailles (1919) on the other, mark the polarization of two separate and irreconcilable ideological currents. On one side, moderate, conservative democratic, and liberal parties, found themselves up against hard-line right or left-wing revolutionary opponents of a nationalist or internationalist persuasion, with scant respect for suffrage, democracy, or tolerance. The twentieth century opened with the founding of a series of communist parties (Russia 1903; Austria 1918; Greece 1918; USA 1919; UK 1920; Spain 1920; Italy 1921). These new protagonists on the political scene achieved varying degrees of importance in different places and their role was to become more prominent during the Cold War. Communist ideology and the proposals of revolutionary Marxist analyses were to assume the weight of a political mortgage at two moments: in Europe in the 1920s and 1930s, and throughout the world from 1945 to 1990. Social demands and economic crises deeply disturbed the political environment. The emergence of totalitarian and fascist dictatorships were to undermine political party structures throughout Europe until 1945. Nevertheless, these dictatorships organized political life by means of an imposed adherence to a single party. Only formally such a party was maintained as an element of legitimacy in the exercise of state power, despite the fact that authoritarian regimes could dispense this.

The success of the Western model of parliamentary democracy based upon a three-legged political system (representative and participative democracy, the electoral system, and the political-party system) is thus a recent phenomenon. During the Cold War, the ideals of representative democracy as being the end product of the modern liberal constitutionalism, with roots in

the 1790s, appeared to be a heritage limited principally to Western Europe and the USA. A few other countries, notably former members of the British Empire—such as Canada, Australia, New Zealand, and India—have practiced representative democracy since achieving their independence. In countries such as Brazil and India, where democratic pluralism exists under specific historical and cultural conditions, many political activities take place outside the formal political-party structure.

Between the world wars, growing social awareness of rights and of the need to manage an increasingly literate populace led to structural modifications among the political parties. Economic crises in Central Europe were accompanied by increasing mistrust of parliamentary institutions and the emergence of manipulative right-wing political parties. Fascism and communism radically changed the way in which political parties were regarded. Into the parties of the masses flowed all the social dissatisfaction and frustration, making them easy prey for manipulation by dictators of every stripe. In Europe, as also in other regions, the trend towards mass parties, which in principal appeared to be a sign of the triumph of democracy, in the end generally resulted in quite the opposite effect as authoritarian rulers assumed power. In many regions, however, expectations with respect to the freedom to participate in party activities was enhanced, especially after the experience of World War II.

The triumph of the Western model of liberal democracy with parliamentary representation was one of the consequences of the new international order that emerged after 1945, except in regions under the yoke of communism (Eastern Europe, China, and Cuba). Communist regimes, however, do not dispense with mass parties or even with formal parliaments, although the role of these institutions is merely figurative. After World War II, a new phase of political life began, with the restoration of the values of liberal democracy, universal suffrage, civil liberties, and parliamentary representation. The Western tradition of recording political events—at least since the Code Napoleon (1804) and the British Reform Act (1832)—has provided a framework for the organization of political life in such a manner that the formal and legal structuring of parties has advanced in parallel to the campaign for constitutional government. This has caused parties to organize and represent themselves as both defenders of an overall social vision and as pursuers of political goals. Thus, parties assume the posture of political businesses whereby they compete within a legal framework to advance the political convictions, opinions, and beliefs—and to defend the preferences and class interests—of their constituents. Diverse and intersecting interests find expression among groups, clubs, associations, and, above all, political parties. The social economic and ideological cut-lines present in the social structures were key

factors throughout the nineteenth century. Conservatives and Liberals, labor parties and revolutionaries, opposed each other in Europe and also in Latin America. In the UK, the Conservatives and the Liberals succeeded the Tories and the Whigs and were in turn superseded by the Conservative and Labour parties of the twentieth century. In the USA in the period following the Civil War, the Democrats and the Republicans definitively consolidated their hegemony over the political scene (Katz and Mair 1992).

The history of political parties clearly demonstrates that the political opportunities in the social competition for power in the State has an effect upon the parties that, for their part, have an effect upon the structure of political opportunity, i.e., parties and institutions mutually affect each other. In democracies which work regularly, disciplined voters arise, as in the European Union, for example. These voters do respond to party activities, although they are increasingly critical of the means whereby parties pursue their ends. 'Belonging' to a party tends to be more a sociological identification with a political tendency, rather than involving formal party affiliation. This suggests a salutary clarification of the boundaries between parties and their environment. The process whereby a party is identified with its leader is an interesting historical phenomenon in the political organization of societies. For example, the Braybrooke's Manuscript (1788) refers to the parties 'of the Crown,' 'faithful to Mr. Pitt,' and 'of Mr. Fox.' Another interesting and more contemporary example is of the emergence of a 'Gaullist' party in France, after 1946. The tradition of having a strong personal leader continues very much alive in Latin America, even in the late twentieth century. This tradition has influenced the social practices of parties in various ways. In Mexico, the Institutional Revolutionary Party—PRI (1911) dominated the Mexican political scene single-handedly until 2000.

With the advance of democracy and of democratic negotiating practices, conducted with a view to making government of contemporary states feasible, all parties (including the new parties that emerged out of the former communist parties of Eastern Europe, and also certain parties in Western Europe) have had to become connected. The difficulty of winning elections based solely upon their different platforms and making their will prevail has, since the 1970s, obliged parties to emphasize their capacity to draw together to form and support coalitions. Finally, all modern parties are obliged to adopt business-like organizational structures or run the risk of failing to fulfill the demands of modern life, in terms of communication, social penetration, and funding.

Evidently, the aim of assuming the mantle of government remains the principal motivation of political parties. The historical trajectory of activities undertaken by political parties is closely associated with an awareness of the structure of political oppor-

tunities within a given political system. Revolutionary organizations that reject existing institutions, social clubs, debating societies, pressure or advocacy groups with independent objectives can be regarded as entities that discreetly play significant political roles without being formally political parties. Formal political parties must, as such, compete to control a political process whose agenda they do not establish by themselves, but which cannot function without their participation. This experience is deeply ingrained in the political parties of the late twentieth century.

3. Unanswered Questions and Issues for Debate

Some important questions remain unanswered. For example, the emergence of political participation—as an alternative to traditional representation in liberal democracies—is often treated both as an essential element for modernization and as its critical consequence. To assume the mantle of government and to perpetuate the values of parliamentary democracy, strong parliamentary institutions are necessary if modern political parties are to perform their roles. This is true even in countries such as the USA—and many others in Africa and Latin-America—in which the presidency is a strong and 'personalized' institution. This applies even in cases where presidents are elected by direct vote and are endowed with special powers, as occurs in France and in Portugal.

Although much is known with respect to the details of the emergence of specific party systems, the capacity to generalize as to the reasons why different types of parties emerge in the ways and forms in which they have done so, remains incipient. There are few works that examine or seek to explain the vast variety of different contexts, and comparative studies among regions with different levels of development are rare. In the 1980s, the political development of multinational regional institutions—which initially had a solely economic focus—such as the European Economic Community, ASEAN, and (in the 1990s) MERCOSUR, has unveiled a whole new field for research into the political life of such institutions, their relations with member states, and the entire theme of supranational integration. From a historical-research perspective, only now are the first comparative studies of the political parties being carried out. This applies not only to research on existing parties in specific countries or groups of countries (as is the case in the European Union), but also with respect to an examination of the historical roots of contemporary political parties.

Not only parties, but also party systems can be analyzed from the institutional standpoint. Both the parties and the party systems of the 1970s were little different from those of the 1920s. With the exception of the ecological movement, the majority of the party groupings, i.e., similar parties, can be presented

according to their similarities. Practically all of the present parties—of the left, of the right, of the center; liberal and revolutionary parties—consolidated themselves in the 1920s, and some much earlier. Political parties are present at the heart of all activities of the state in liberal democracies. However, even under communist regimes the role of the political party is performed in a well-established manner. In the 1990s, however, it could not safely be predicted that political parties would face a future with development and prestige, similar to that they enjoyed in the recent past. At the present time, a new type of political culture is emerging. Political awareness and politics—as subject for discussion—open up a new avenue of research into general participation in political life on the threshold of the twenty-first century. Culture may help a understanding as to why certain social cleavages are successfully transposed by a stable party system, whereas others are not. Political culture and the manner in which it is created is responsible for some significant differences in the traditional organization of states, or of bodies associated to them. Political culture includes not only ideals, such as values, beliefs, attitudes, but also political practices and societal behavior.

There are certain themes in the study of political parties that still require further research from a historical point of view. Politics, with the constrained use of political power, and the politics of containment have been recurrent themes in political science since the 1970s. One of the most important contentions of political activity is the institutional framework of the state. Contemporary political research has made clear the importance of institutional factors in political life. Aside from this new institutional focus, the study of political parties has demonstrated a renewed appreciation of history, of events, of the rules and of political regimes as the active forces of political life. It is perhaps commonplace to state that 'history matters.' The 'frozen cleavages' of Lipset and Rokkan, the development models of communism, fascism and the parliamentary democracy of Moore (1968), the theories of 'critical realignment' of Burnham (1970), are good examples of historical reviews and models. The legacy of history is thus one of the limitations to which institutionalism is subject.

Socio-economic constraints are generally accepted without going into the mythical and overworked concepts that have been served up since the time of Marx. Rapid changes in international finances and economic scenarios have produced intriguing works. The exercise of social power is also constrained by cognitive means, in the use of pure reason (and more specifically, of the practical one). The knowledge society and information society will have to take into consideration non-rational features of political life, and the social and ideological factors that influence the belief systems of the masses (Meny 1989). The final two decades of the twentieth century witnessed a swift

retreat in the positivist distinction between fact and value. Specialists recognized that political agents were also ethical players (Rawls 1993), who have their own values and act in accordance with them, and that occasionally they may be persuaded (by political philosophers, perhaps) to adopt other, better values. Any attempt to divorce facts and values from the mental processes and from the underlying political dynamics, would be doomed to failure. Similarly, political science increasingly seeks to develop studies that take into account the complexity of the structures, processes, and results that it investigates. In order to do this, one must have a theoretical framework within which the history of the processes and the prevailing political philosophy must occupy a space which formally would have drawn opposition from previous generations.

See also: Conservatism: Historical Aspects; Democracy, History of; Interest Groups, History of; Liberalism: Historical Aspects; Parliaments, History of; Party Systems; Political History (History of Politics); Political Parties; Socialism: Historical Aspects

Bibliography

- Avril P 1990 *Essais sur les Partis Politiques*. Payot, Paris
- Budge I, Keman K 1990 *Parties and Democracy. Coalition, Formation and Government Functioning in Twenty States*. Oxford University Press, Oxford, UK
- Butler D 1981 *Democracy at the Polls*. AEI, Washington DC
- Charlot J 1971 *Les Partis Politiques*. Payot, Paris
- Duverger M 1951 *Les Partis Politiques*. Colin, Paris
- Epstein L D 1993 *Political Parties in Western Democracies*. Transaction Publishers, New Brunswick NJ
- Fenske H 1994 *Deutsche Parteigeschichte*. Schöningh, Munich, Germany
- Galli G 1990 *Storia dei Partiti Politici Europei*. Rizzoli, Milan, Italy
- Goodin R E, Klingemann H-D (eds.) 1966 *A New Handbook of Political Science*. Oxford University Press, Oxford, UK
- Graham B D 1993 *Representation and Party Politics. A Comparative Perspective*. Blackwell, Cambridge, UK
- Huntington S P 1968 *Political Order in Changing Societies*. Yale University Press, New Haven, NJ
- Kaase M, Newton K (eds.) 1995 *Beliefs in Government*. Oxford University Press, Oxford, UK
- Katz R S, Mair P (eds.) 1992 *Party Organizations. A Data Handbook on Party Organizations in Western Democracies, 1960–90*. Sage Publications, London
- Lancelot A 1988 *Les élections sous la Vème République*. PUF, Paris
- LaPalombara J, Weiner M (eds.) 1966 *Political Parties and Political Development*. Princeton University Press, NJ
- Lawson K (ed.) 1994 *How Political Parties Work: Perspectives from Within*. Praeger Publishing, Westport, CT
- Lipset S M, Rokkan S 1967 Cleavage structures, party systems, and voter alignments: an introduction. In: S. Lipset S M, Rokkan S (eds.) *Party Systems and Voter Alignments: Cross-national Perspectives*. Free Press, New York

- Lloyd C (ed.) 1983 *Social Theory and Political Practice*. Wolfson College Lectures. Clarendon Press, Oxford, UK
- Lloyd C 1993 *The Structures of History (Studies in Social Discontinuity)*. Blackwell, London
- Lovenduski J, Pippa N (eds.) 1993 *Gender and Party Politics*. Sage Publications, London
- Maisel L, Cooper J (eds.) 1978 *Political Parties: Development and Decay*. Sage Publications, London
- Maisel L, Sacks P M (eds.) 1975 *The Future of Political Parties*. Sage, Beverly Hills, CA
- Mayhew D R 1986 *Placing Parties in American Politics: Organization, Electoral Settings, and Government Activity in the 20th Century*. Princeton University Press, NJ
- Meny Y (ed.) 1989 *Idéologies, Partis Politiques, et Groupes Sociaux*. Presses de la FNSP, Paris
- Merton R K 1968 *Social Theory and Social Structure*. Free Press, New York
- Michels R 1989 *Soziologie des Parteiwesens*. Kröner, Stuttgart, Germany
- Moore B Jr 1968 *Injustice. The Social Bases of Obedience and Revolt*. Princeton University Press, NJ
- Moore B Jr 1998 *Moral Aspects of Economic Growth and Other Essays*. Cornell University Press, Ithaca, NY
- Norton P (ed.) 1996 *The Conservative Party*. Oxford University Press, Oxford, UK
- Offerlé M 1997 *Les Partis Politiques*. PUF, Paris
- Ostrogorski M 1903 *La démocratie et l'Organisation des Partis Politiques*. Calmann-Lévy, Paris
- Pippa N (ed.) 1999 *Critical Citizens: Global Support for Democratic Government*. Oxford University Press, Oxford, UK
- Pombeni P 1994 *Partiti e Sistemi Politici Nella Storia Contemporanea*. il Mulino, Bologna, Italy
- Rawls J 1993 *Political Liberalism*. Columbia University Press, TN
- Rohe K (ed.) 1990 *Elections, Parties and Political Traditions. Social Foundations of German Politics and Party Systems, 1867–1987*. Berg, New York
- Rose R 1984 *Do Parties make a Difference?* MacMillan, London
- Schumpeter J A 1942 *Capitalism, Socialism, and Democracy*. Harper and Row, New York
- Seiler D L 1986 *De la Comparaison des Partis Politiques*. Economica, Paris
- Skocpol T (ed.) 1999 *Democracy, Revolution, and History*. Cornell University Press, Ithaca, NY
- Tufte E R 1980 *Political Control of the Economy*. Princeton University Press, NJ
- Ware A 1996 *Political Parties and Party Systems*. Oxford University Press, Oxford, UK
- Weber M 1919 *Politik als Beruf*. J. B. Mohr, Tübingen, Germany

E. C. de Rezende Martins

Political Protest and Civil Disobedience

Political protest refers to a multitude of methods used by individuals and groups within a political system to express their dissatisfaction with the status quo. Although protest can be registered through voting, letter writing, campaign contributions, participation

in interest groups and political parties, and other traditional means, the term more commonly refers to those kinds of mass political activism that occur outside of conventional parliamentary channels, such as demonstrations and marches, sit-ins, blockades, labor stoppages, boycotts, and other methods that combine communication and persuasion, direct action, and noncooperation with political, economic, and social institutions.

Much social science research on political protest has concentrated on explaining organized, sustained forms of mass collective action such as social movements. A social movement consists of an array of organizations and individuals who share a common identity and agenda of social and political goals, and who are regarded as a collective entity by other political actors. Social movements typically pursue their goals by mixing protest actions with conventional political strategies such as lobbying, litigation, public education, and electoral mobilization. Discontented groups often resort to direct action when they are denied equal access to conventional means of participation or when such avenues are regarded to be ineffectual or inadequate for registering the intensity of their grievances. Theoretical and empirical studies of political protest have analyzed the emergence of group consciousness and solidarity, the framing of grievances, the role of social networks and community institutions in fostering cooperation, the formation of social movement organizations, the psychological make-up and motivations of individuals who engage in protest, the evolution of strategy and tactics in social movements, the influence of changing political opportunities on the course of a movement, and the efficacy of violent and nonviolent methods of protest.

Unlike protest marches, picketing, and other kinds of legal protest, civil disobedience is a particular form of political protest that is marked by the deliberate violation of the law for social purposes. Civil disobedience is not merely conscious law breaking, but rather it is law breaking with the express goal of correcting a perceived injustice. Action that would now be regarded as civil disobedience has been a regular feature of social protest throughout history even though the term itself, attributed to an 1849 essay by Henry David Thoreau, is a relatively recent creation.

The act of civil disobedience is fundamentally an act of political protest because the noncompliance is intended to convey a message of discontent to the government, media, and public. The acts of defiance are typically overt, public acts aimed at marshaling collective opposition to unjust or unconstitutional laws and policies. Civil disobedience is not restricted to protests over laws and public policies, but is also employed against the practices of private institutions such as corporations and universities, as occurred for example in the divestment campaign on American university campuses in the early 1980s, which pres-

sured universities (both private and public) to cut their financial ties to companies that conducted business in South Africa in an effort to bring an end to apartheid.

1. *Forms of Civil Disobedience*

Within this broad conceptualization, civil disobedience can emanate from different motivations and assume a number of forms. Civil disobedience can be active or passive, direct or indirect. An individual can either actively commit a prohibited act or passively refuse to conform to a prescribed action. For example, an anti-war protester may destroy his draft card or, alternatively, he may refuse to register for the draft. Both actions qualify as civil disobedience because the protester intentionally violates the law as an expression of his opposition to the war, knowing the legal consequences of his actions. Thoreau himself engaged in a form of passive civil disobedience when he chose to register his opposition to the institution of slavery by refusing to pay the Massachusetts state poll tax.

Civil disobedients often break laws that are unrelated to the law or policy they are protesting. For example, peace activists who disrupt operations at defense laboratories are not protesting laws against trespassing, but are using civil disobedience to dramatize their opposition to nuclear weapons. Supreme Court Justice Abe Fortas objected to the indirect use of civil disobedience on the grounds that it would give licence to people to break any law they wished in the service of a goal that they perceived to be more noble. But few scholars have subscribed to this restriction because, on many issues, there is no feasible direct method to engage in civil disobedience. There may, for example, be no direct method by which activists can use civil disobedience to protest the government's foreign policy or its failure to pass laws addressing issues of inequality and poverty.

A narrower definition of civil disobedience would exclude illegal actions taken by individuals who do not accept the constitutionality of the law they are violating. By this definition, an activist who breaks the law by setting fire to the American flag is *not* engaging in civil disobedience if he believes that his actions are protected by the First Amendment, but a flag burner who accepts the constitutionality of a law against flag burning *is* engaging in civil disobedience. For the purpose of defining civil disobedience, however, it may be unnecessary to make such a distinction based on the activist's frame of mind, because the crucial assumption underlying civil disobedience is that the law being broken is generally viewed to be valid by those in authority, so the actions of the civil disobedient are generally viewed to be illegal. Therefore, both flag burners are committing civil disobedience in the sense that they are breaking what is generally understood to be the law in an effort to change it, and the main

difference between them is whether they are appealing to the courts or to the executive and legislature for remedial action.

Although Thoreau assumed that 'civil' protest would be conducted in an orderly and peaceful fashion, scholars disagree over whether civil disobedience is necessarily nonviolent by definition. Some commentators insist that civil disobedience must be nonviolent or else it cannot be distinguished from rebellion, rioting, and other types of violent direct action. The two individuals who are most responsible for the association of civil disobedience with nonviolence are Mohandas Gandhi (see e.g., Gandhi 1961) and Martin Luther King, Jr. (see e.g., King 1964). Indeed, contemporary understandings of civil disobedience have been substantially shaped by the mass campaigns for equal rights organized by Gandhi in South Africa and India and by King in the United States.

Both Gandhi and King possessed the extraordinary combination of political 'savvy' and unwavering moral and spiritual commitment to the philosophy of non-violence. In his teachings, Gandhi insisted that civil disobedients must scrupulously obey laws that are unrelated to the target of the disobedience campaign and submit themselves to the punishment of the state for breaking a law they believed to be unjust. Civil disobedients do not flee or resist arrest, for in accepting penalties for their actions, they bring attention to the injustice they wish to correct and arouse the sense of injustice in others. Gandhi's concept of *satyagraha* or 'truth-force' refers to the moral conversion of one's adversary. In their quiet suffering, civil disobedients demonstrate their honor and integrity and their capacity to differentiate between just and unjust laws.

King regarded Gandhi as his mentor and shared his belief that the purpose of civil disobedience is to persuade the state and its citizens that an existing law violates a higher moral principle. Violence, therefore, must be avoided because it potentially undermines the communicative intent of civil disobedience. Adherence to nonviolence, on the other hand, indicates that one bears no malice or hatred for one's opponent. Like Gandhi, King believed that the ultimate goal of a campaign of nonviolent protest is to achieve a new moral understanding with one's opponents. Nonviolent resistance works, not by defeating and humiliating the opponent, but by winning the opponent's empathy and friendship.

2. *Justification of Civil Disobedience*

In addition to disagreements over what kinds of actions qualify as civil disobedience, there is debate over the circumstances in which civil disobedience can be justified. Civil disobedience raises issues about the nature of authority, obligation, and consent in society.

The practice originates from the idea that a person must be true to his conscience and that he is not obligated to submit to the laws or commands of the state when they violate his conception of justice. Civil disobedience therefore assumes both a right to individual freedom and the existence of a natural law that takes precedence over civil law.

Philosophical discussions of the justification for civil disobedience generally fall between legal absolutism and moral individualism. In justifying civil disobedience, it is necessary to accept that there are standards against which governments can be evaluated, and that citizens have the right to exercise such judgments. Few would argue that the law must always be obeyed; and similarly, few would contend that obeying the law is entirely a matter of individual judgment about what is morally right and wrong. One extreme supports despotic governments, while the other suggests anarchy. Whether civil disobedience undermines respect for the law is a matter for empirical inquiry. The penalties for breaking the law are usually sufficiently great that it is unlikely that people will engage in civil disobedience unless they are intensely motivated by a sense of injustice.

But do the moral criteria for civil disobedience reside only in conscience? Or are there intersubjective criteria that permit rational assessment of the morality of a law or policy? Without such objective criteria, we are in danger of every individual invoking arbitrary subjective criteria for disobeying different laws. We invite moral anarchy unless there are some agreed-upon criteria for evaluating whether a law or policy is unjust or immoral. At a minimum, the civil disobedient must be able to demonstrate that the law or policy violates principles of welfare, equality, or justice. The law or policy must be shown either to do more harm than good, to distribute benefits unequally, or to operate without the consent of certain groups. Legal methods of petition to change the law must be tried first, and the activists must grant the same rights of civil disobedience to others who believe that another law is unjust. Some would argue that the civil disobedient must also have a reasonable chance of overturning the law. According to these normative criteria, only those who believe in reciprocity and adhere to a sophisticated set of general rules of political conduct can rightfully be said to be engaged in justifiable acts of civil disobedience.

3. Effectiveness of Civil Disobedience

What are the mechanisms that make civil disobedience effective in practice? Gandhi recognized that civil disobedience could not be effectively employed for a global, distant goal such as national independence, but had to be used for specific, local issues on which the opposition had the power to compromise and make concessions. True to form, civil disobedience

campaigns led by Gandhi were most successfully applied in local disputes concerning such issues as land revenue assessments and mill workers' wages, in which small numbers of close knit, disciplined protesters were united by their interest in the issue and capable of conforming to a norm of nonviolence.

In theory, if civil disobedience is to work through persuasion, then the protesters must invoke principles or values that are shared by a large proportion of the population. In practice, however, the preferences that are converted are more likely to be those of the spectators to the conflict than those of the adversaries of the protesters. A small number of disobedients can seldom, by their actions alone, render a law inoperative, but they may be able to disrupt the social order sufficiently to draw public attention to their grievances. Although Gandhi believed that the suffering of protesters in jail would apply pressure on the government, he also recognized that the state's use of violence would serve the interests of the protesters by influencing third parties who were witnesses to the conflict. British repression of Indian dissenters confirmed to onlookers that the British ruled by force and not through the consent of the Indian population. Similarly, in the course of the civil rights movement, it was the contrast between the nonviolent protesters and the violent response of the southern police that helped to galvanize public opinion and mobilize federal government intervention in support of racial equality.

In response to a campaign of civil disobedience, the defenders of the status quo must decide whether to offer concessions to the protesters or to escalate the conflict by using force or summoning the power of the state. It is the threat of continued disruption of the normal operations of business and government that 'persuades' the targets of the protest to negotiate with the activists and to yield to their demands. During the civil rights movement, the early success of direct action campaigns in some southern cities aimed at forcing businesses to provide equal treatment to blacks made it easier for similar campaigns to succeed elsewhere because businesses foresaw the normal course of these protests and sought to cut their losses sooner rather than later. There is little evidence that the attitudes of white southerners were converted by direct action; indeed, the attitudes of many were hardened by the confrontational tactics of the movement.

Civil disobedience thus rarely works through persuasion and attitude change alone, but also depends significantly on pressure and coercion. Actions that are aimed at transforming preferences may leave attitudes unchanged, but still be effective at promoting social change because the opponents are pragmatic enough to calculate that they are better off making concessions than dealing with the disorder created by continual protests. Civil disobedience in this respect resembles a pragmatic tactic within a repertoire of potential strategies that are available to a protest

group rather than an option that is resorted to only when the conventional channels of participation have been exhausted.

See also: Citizen Participation; Civil Rights Movement, The; Civil Society, Concept and History of; Collective Behavior, Sociology of; Nonviolence: Protective Factors; Peace and Nonviolence: Anthropological Aspects; Peace Movements, History of; Social Movements: Resource Mobilization Theory; Social Movements, Sociology of

Bibliography

- Bay C 1968 Civil disobedience. In: Sills D L (ed.) *International Encyclopedia of the Social Sciences*. MacMillan, New York
- Chong D 1991 *Collective Action and the Civil Rights Movement*. University of Chicago Press, Chicago
- Gandhi M K 1961 *Non-violent Resistance*. Kumarappa B (ed.). Schocken Books, New York
- Gewirth A 1982 *Human Rights: Essays on Justification and Applications*. University of Chicago Press, Chicago
- Greenawalt K 1987 *Conflicts of Law and Morality*. Oxford University Press, New York
- King M L 1964 *Why We Can't Wait*, 1st edn. Harper & Row, New York
- McAdam D, McCarthy J D, Zald M N 1988 Social movements. In: Smelser N J (ed.) *Handbook of Sociology*. Sage, Newbury Park, CA
- Powers R S, Voge W B 1997 *Protest, Power, and Change: An Encyclopedia of Nonviolent Action from ACT-UP to Women's Suffrage*. Garland Publishing, New York
- Rawls J 1971 *A Theory of Justice*. Belknap Press of Harvard University Press, Cambridge, MA
- Sharp G 1973 *The Politics of Nonviolent Action*. Porter Sargent, Boston
- Zinn H 1968 *Disobedience and Democracy: Nine Fallacies on Law and Order*. Random House, New York

D. Chong

Political Representation

Political representation is a process in which one individual or group (the representative) acts on behalf of other individuals or groups (the represented) in making or influencing authoritative decisions, policies, or laws for a polity. The *representative* may hold any of a variety of offices (executive, administrator, ambassador, judge, lobbyist, party leader) and may even be a collectivity (a governmental agency or political organization), but the modern concept most commonly refers to the role of a legislator. Similarly, the *represented* may be almost any individual or group, but is usually understood to be those individuals

or groups, typically called constituents, who are entitled to select the legislators. Some important work examines the relationship between legislative and other forms of representation (Stimson et al. 1995), but most has focused on one or more of three dimensions of legislative representation: relationships between the representatives and the represented; among the represented; and among the representatives.

1. Origins

Political representation in its modern form evolved over a long period from about the fourteenth to the seventeenth century as kings and nobles convened assemblies to collect revenue and conduct wars. The bodies, usually composed of members from various estates or classes, gradually assumed independent power, and came to be thought of as representative institutions. In the debates during the Civil War period in England, defenders of Parliament connected the idea of representation clearly to the idea of consent of the governed. In 1651, the members of Parliament were for the first time referred to as 'representatives' (Pitkin 1967; also see Larsen 1966, Manin 1997).

By the eighteenth century the link between representation and popular rule had become sufficiently close that Baron de Montesquieu could praise the English constitution for establishing a republic in a large state by giving the people a voice through their representatives in the legislature. Defending the new American constitution, James Madison argued that representation not only made republics possible in large territories, but also mitigated their dangers, especially the influence of factions (Wood 1972). Representative government was not yet representative democracy, because much of the adult population could serve as neither representatives nor the (actually) represented. It was John Stuart Mill who in the late nineteenth century provided the framework for the theoretical synthesis that combined the traditional virtues of representation and the emerging ideals of democracy (Thompson 1976).

2. Legislators and Constituents

At least three aspects of the role of the representative should be distinguished. The first concerns the *mode* of representation—how constituents are represented. It expresses a continuum in which the limits are specified by the traditional contrast between trustee and delegate. The second aspect is the *locus* of representation—which citizens are represented. Among the possibilities are the nation as a whole, a political party, the state or district, and various groups and organizations. The district is more complex than is often assumed. For example, many legislators are bound to

local elites and national interest groups whose interests may diverge from those of other groups in the district, and some legislators see themselves as representatives of under-represented groups, such as blacks and women. The third aspect refers to variation in the *content* of representation—the types of issues with which a representative deals. A legislator may act, for example, as a delegate of the district majority on welfare issues, as a delegate of the national majority on civil rights, but as a trustee of both on foreign policy.

Given this complexity, it is not surprising that the most plausible normative theories refuse to privilege any one of these roles; what each representative should do depends on the nature of the particular issues, what other representatives are doing, and the biases in the political system. Similarly, when political scientists use the simpler conceptions of role, they rarely find any general correspondence between a representative's conception of role and actual behavior. The most robust accounts of legislators' conceptions posit decision trees with multiple branches and a complex set of factors (Kingdon 1981, p. 244).

On any account of the role of the representative, there is the further question of how representatives can be held accountable. The methods vary, depending on the political system, and on assumptions about what political accountability means. The form of accountability that has received the most attention is the electoral connection between representatives and constituents. Much of the empirical literature on the linkages between the policy preferences of constituents and legislators has been influenced by the classic model presented by Miller and Stokes (1963), which highlights four major variables: legislators' personal attitudes, roll call votes, constituency opinion, and legislators' perceptions of constituency attitudes. The vast literature on this linkage shows that the conditions under which each has an effect are quite variable. It is now recognized that roll call votes fail to capture much important legislative activity, and do not always reflect the actual views of legislators. A still unresolved puzzle is why individual representatives (even in safe seats) exert so much effort to please their constituents and collectively act in ways that track mass public opinion even while the evidence indicates that constituents and the mass public are not well informed about what their representatives believe and do (see Uslander 1986).

A related trend is the growing importance of constituent service (a practice in which representatives help constituents and groups in dealing with their particular problems with government, such as immigration and state and federal grants). The electoral effect of constituency service remains a subject of some controversy among political scientists, but there is agreement that the practice is now more prevalent (Cain et al. 1987). Another effect that has provoked important work is pecuniary. Since at least the eighteenth century students of legislatures have stud-

ied and often criticized the influence of money and wealth on the behavior of representatives. Political scientists have not generally found strong correlations between contributions and roll call votes on broad policy areas even in the US system. But they have found that money may buy access—marginal time, energy, and legislative support.

3. *Constituents*

All legislative representation is group representation. Even in societies governed by individualist norms such as 'one person-one vote,' the definition of the constituency implicitly identifies a group or groups that are to be represented.

Party-based representation is associated with a modified form of proportional representation in which the share of votes that each party receives determines how many of the candidates from its lists win seats in the legislature. The system usually produces multiple parties, and thus gives voters a wide range of ways to express the interests they wish to have represented. However, it gives voters less choice about which particular candidate should serve as their representative, and in this respect differs significantly from the purer form of proportional representation, the system of the single-transferable vote, in which every successful candidate has a unanimous constituency.

Geographical representation usually relies on single-member plurality systems, and tends to produce only two or three parties. The parties represent broader interests than do the parties in modified proportional systems, but if party discipline is weak, as in the USA, the geographical basis encourages the representation of particular regional and other interests. No single type representational system has been shown to be superior in all or most circumstances, and the choice generally involves a trade-off: the wider the range of interests represented in the legislature, the less the likelihood that any particular interest will be fully represented in legislation actually adopted. The most useful scholarship on these and related questions has been comparative (Klingemann et al. 1994, Converse and Pierce 1986).

In systems such as the USA, where many of the districts that define constituencies do not follow any natural geographical boundaries, the issues of how the districts are drawn and whether they produce fair representation have stimulated important normative and empirical research. Political scientists have examined the partisan and incumbency biases in various redistricting schemes. Some of the most interesting theoretical work in this area combines empirical and normative approaches. One study has developed more analytically robust concepts of 'fair representation' in electoral systems (Beitz 1989), and another has revived the theory of descriptive representation for 'marginalized groups' (Williams 1998).

4. Legislators

Much of the research on this element of representation has dealt with the tension between individual and collective behavior in the legislature. In the USA a phenomenon that has received considerable attention is what is sometimes called legislative individualism, the tendency of legislators to act in ways that promote their own re-election without regard to the effect on collective product of the legislature. Such individualism has been shown to produce a pattern of 'particularized benefits' (dams in certain regions, tariffs for local industries, and military bases in the district) and the neglect of more general public benefits (such as education, health, and the environment). The explanations for these outcomes vary, but they are generally structural, locating the problem in the incentives of the system, rather than in the character of legislators. They generally assume that, given the structure of the political system, individually rational behavior by legislators produces collectively suboptimal outcomes in the legislature.

If all legislators are pure individualists seeking only particularized benefits for their own constituents, how (under conditions of majority rule and weak parties) can the legislature adopt any legislation at all, even the pork barrel kind? This puzzle has been extensively studied by formal theorists, who propose various solutions for achieving policy equilibria in legislative institutions. One of the most plausible relies on a committee system to set an agenda and limit the scope of conflict. But shared interests and cross-cutting cleavages would still be necessary in any actual system. These are more likely to exist or to arise in legislatures that encourage or at least permit representatives to change their positions through interaction with their colleagues. This is one of the advantages of deliberative theories of legislative representation, which allow for changes in the preferences of both representatives and the represented over time (Gutmann and Thompson 1996, Chap. 4).

See also: Accountability: Political; Apportionment: Political; Campaigning: Political; Cleavages: Political; Consociationalism; Deliberation: Political Aspects; Electoral Systems; Feminist Political Theory and Political Science; Interest Groups; Legislatures: United States; Participation: Political; Political Parties; Proportional Representation.

Bibliography

- Beitz C R 1989 *Political Equality*. Princeton University Press, Princeton, NJ
- Cain B, Ferejohn J, Fiorina M 1987 *The Personal Vote: Constituency Service and Electoral Independence*. Harvard University Press, Cambridge, MA
- Converse P E, Pierce R 1986 *Political Representation in France*. Belknap Press of Harvard University Press, Cambridge, MA

- Gutmann A, Thompson D 1996 *Democracy and Disagreement*. Harvard University Press, Cambridge, MA
- Kingdon J W 1981 *Congressmen's Voting Decisions*, 2nd edn. Harper and Row, New York
- Klingemann H-D, Hofferbert R I, Budge I 1994 *Parties, Policies and Democracy*. Westview Press, Boulder, CO
- Larsen J A O 1966 *Representative Government in Greek and Roman History*. University of California Press, Berkeley, CA
- Manin B 1997 *The Principles of Representative Government*. Cambridge University Press, Cambridge, UK
- Miller W E, Stokes D 1963 Constituency influence in Congress. *American Political Science Review* 57
- Pitkin H F 1967 *The Concept of Representation*. University of California Press, Berkeley, CA
- Stimson J A, Mackuen M B, Erikson R S 1995 Dynamic representation. *American Political Science Review* 89: 543–65
- Thompson D F 1976 *John Stuart Mill and Representative Government*. Princeton University Press, Princeton, NJ
- Uslaner E M 1986 Legislative behavior: the study of representation. *Annual Review of Political Science* 1: 205–37
- Williams M S 1998 *Voice Trust and Memory: Marginalized Groups and the Failings of Liberal Representation*. Princeton University Press, Princeton, NJ
- Wood G S 1972 *The Creation of the American Republic 1776–1787*. Norton, New York

D. F. Thompson

Political Science: Overview

Political Science is an academic discipline that has been entrenched in university curriculums for about a hundred years. It takes as its subject matter government and politics very broadly defined, meaning the study (in whatever setting) of outcomes deemed important by participants or observers. This means that there is politics, and therefore in principle political science, concerned with, *inter alia*, organized crime, organized sports, organized religion (Reese), business firms (Bauer et al. 1963), family life, and markets (Dahl and Lindblom 1953). Most commonly political scientists study social patterns and practices associated with the apparatus of the state, identified by Max Weber as an institutional complex located within a specific territory that monopolizes the legitimate use of force.

One formula describing the scope of the field is contained in the full title of one of Harold Lasswell's path-breaking books: *Politics: Who Gets What When How*. This definition, by focusing on what people 'get,' leaves uncertain the demarcation between political science and economics, and may lead to the impermissible inference that answers to the question 'who benefits' are by definition the same as answers to the question 'who governs.' Perhaps a more useful des-

cription of the scope of the discipline is David Easton's formulation of politics as 'the authoritative allocation of values' in society, entailing the study of authority and its bases, allocation and its modes, and the shaping and sharing of values. Political scientists adopt many different approaches to the study of politics, and so it may be helpful to contrast political science, which is a discipline whose practitioners are united by very little other than a focus on a population of dependent variables or effects of human intention or organization or action with, e.g., sociology or psychology, which are unified around a preferred set of independent variables or causes of varied social and personal outcomes. This is not to say that political scientists do not write about causes as well as effects; rather, when they do so, they borrow freely from other disciplines. There are no psychological or sociological mechanisms that are the monopoly of political science. What is distinctive to political science is what it is about: politics. Political science draws on and elaborates ideas from other disciplines in seeking to account for, or explain, or understand political life. This helps us understand why no clear-cut mainstream tradition exists in political science, rather a number of different traditions of inquiry, organized in part around the differing scale and character of different topics within the field, of which I shall mention several.

Political scientists draw on several branches of psychology in attempting to understand attitude formation and the preferences of citizens and individual voters, as well as the choices of leaders. The development of sophisticated survey instruments for the purpose of understanding individual attitudes and opinions has taken place almost wholly outside political science, but political scientists have been quite successful in adapting these tools to the tasks of understanding how individuals think about political outcomes, make voting choices, and form views of their political world (Converse 1987, Zaller 1992). Social psychology and the social psychological branch of cognitive science have also framed and explored principles of human behavior (some of which are called 'heuristics') that shape political scientists' explanations of political outcomes (Tversky and Kahneman 1974, Abelson et al. 1968, Simon 1957). Finally, the multifaceted Freudian and neo-Freudian traditions of inquiry have had an impact on the study of the behavior both of masses and elites (Greenstein 1969, George and George 1956).

A second focus of research for political science has been the study of social organizations, groups, and classes, in which mainstream sociology has provided intellectual inspiration. In some cases, despite the restricted domain of political scientists' interests, political scientists have entered into protracted and serious dialogue with sociological students of such topics as organizational design and reform, community power, and social capital (Wilensky 1967, Polsby 1963, Putnam et al. 1993).

On a still broader scale, political scientists study institutions as large and consequential as nation-states and their constituent elements: executives, legislatures, courts, parties, elections, interest groups, drawing on a great many specialized studies by historians, legal scholars, and journalists as well as other social scientists (Greenstein and Polsby 1975).

Relations among nation-states is a fourth domain in which political scientists work, exploring questions like the extent to which inter-state cooperation on small or less contentious matters such as the regulation of financial markets or trade increases the probability of harmony on larger matters touching national security interests and forestalls hostility (Haas 1968, Deutsch et al. 1957).

In the study of international affairs (as is also the case in other domains) we may note a significant influence on the agendas of political scientists, namely the preoccupations of politicians who are themselves the objects of study. Just as diplomats historically have sought to find a balance of power among nation-states, students of international affairs have concerned themselves with issues of international political stability, the formation of security communities, and the causes of war (Waltz 1959, Van Evera 1999). This illustrates the fact that the intellectual agendas of political scientists are made in part from political events and not only from problems posed by the ongoing literature of political science.

A fifth focus for political scientists has been the history of ideas about social and political organization, sometimes studied through the elucidation of great books in the heritage primarily of Western Civilization, sometimes through the examination of prescriptions about social organization and the textual or empirical study of prerequisites of their fulfilment (Skinner 1978, Berlin 1969, Wolin 1960, Dahl 1956). Thus a political scientist may be a historian of ideas, or a philosopher, or draw inspiration from those disciplines.

As a professional discipline organized into the curriculums of the world's universities, Political Science is, like the other social studies, only about a century old. Systematic political study, to be sure, is as old as the Greek philosophers of the fourth century BC, notably Plato and Aristotle. The oldest University chair in the subject is at the University of Uppsala which in 1622 established the Johan Skytte chair of Eloquence and Government, helpfully mentioning one of the classic disciplines (rhetoric; moral philosophy and law are two others) from which late nineteenth century political science emerged as a separate subject offered on a regular basis in the newly professionalizing universities of Europe and, especially, North America.

The editors of a new British journal of politics and international relations estimate that today about 85 percent of all the political science being studied worldwide is located in the USA. Where the discipline

flourishes elsewhere it is more likely to be limited to studies in public law or philosophy and to lack the autonomous ability to sponsor the discovery and pursuit of its own program of study as has been possible for political scientists in the US, at least since the formation of the American Political Science Association in 1903. The Association now has about 15000 members, the largest aggregation of political scientists in the world. It sponsors an annual meeting and two journals, one (*American Political Science Review*) scholarly and one (*Political Science*) professional (Somit and Tanenhaus 1964).

Like other social sciences, the pursuit of disinterested inquiry into political life requires a political atmosphere that tolerates disinterested inquiry. Without this atmosphere political science is suffocated. It can be no accident that contemporary political science is primarily to be found in those nations most advanced in their incorporation of the values of the Enlightenment, which underwrite academic freedoms.

Political science as knowledge can be distinguished from political lore or political news primarily by reference to the social organization that produces, conserves, and criticizes it. Political news is the property of a nation's political class and it is spread by journalists or chains of gossip. Political lore consists of precepts codifying and sustaining the practices of political practitioners. Political science may incorporate content from either of these bodies of knowledge but the social organization that certifies knowledge as political science at any given time and sponsors criticism and new exploration in the field is the congeries of academic departments variously called political science, political studies, politics, or government in colleges and universities worldwide. These departments promote inquiry and conserve knowledge by educating undergraduates, training graduate students, producing texts, announcing discoveries, criticizing work, and maintaining a community of mutually attentive scholars (or, more accurately, a set of sub-communities) who participate in a loosely articulated status system, recognize common intellectual ancestors, promote fads, award prizes, and otherwise establish and enforce social boundaries.

In the USA alone, some 125 departments grant the Ph.D. in political science, and roughly five times that number of colleges offer undergraduate majors in the topic. Substantial numbers of universities in Canada, Europe, India, Japan, and Australia do likewise, but graduate studies in political science in other places, where they exist, are more likely to be subsumed under law, history, or philosophy as, for example is the case at Cambridge University in the UK.

The incorporation of the term 'science' in the disciplinary label political science has from time to time provoked controversy on the ground that no inquiry into the behavior of sentient beings can possibly approximate the predictive capacities of

physics or chemistry. This charge is made against all the human sciences but because of the discipline's eighteenth century—largely pre-scientific—name it is sometimes assumed that in the case of political science ambitious claims for prediction or control of human behavior are being advanced. There are several standard answers: (a) That the term 'science' in historical context makes reference not to some social analogue to atomic physics but simply to 'knowledge'; (b) that the application of political knowledge was thought to be desirable in the creation of new constitutions, notably the US Constitution of 1787; (c) that political knowledge takes many forms requiring only the recognition of patterns of social behavior. This facilitates the codification for example, of comparative information about analogous activities (voting, budget-making, social/political protest) in different political systems; (d) that the goals of prediction and control as a means of understanding social behavior are best understood as applying not to historical events but rather to variables abstracted from events which may lend themselves to historical interpretation; (e) and that social patterns and other manifestations of human behavior can be more competently observed and explained by reference to standards and canons of inquiry (especially recourse to evidence and codification of findings into propositions of some generality) associated with more scientifically advanced disciplines.

Political scientists themselves vary in the extent to which they pay explicit attention to rules and methods of scientific inquiry. As the twentieth century advanced, more and more did so, creating among other things, sub-fields devoted to methodology—mainly adapting to political materials the methods of survey research and econometrics—and to specialties sometimes called rational choice or formal or positive theory—adaptation of the economist's propensity to describe simplified versions of human activity by means of mathematical models. (Solow 1997).

Earlier, with growing influence from the 1950's onward, a so-called 'behavioral revolution' took place in which political scientists in many sub-fields of the discipline sought to introduce into their work the rigorous intellectual habits of the other social sciences with their concern for evidence and for structuring their inquiries so as to alert investigators to unexpected or generally applicable outcomes—e.g. by referring to null hypotheses, or by adhering to falsificationist standards of proof. This has facilitated storage and retrieval of information and led to the cumulation of findings in several branches of the discipline. While these practices have spread in the last half-century, they have not wholly supplanted more traditional modes of inquiry such as the close reading of classic texts or the clarification—or at least discussion—of key concepts, such as 'democracy' or 'power' or 'coalition' or 'party.' (Lasswell and Kaplan 1950, Sartori 1984).

See also: Aristotle (384–322 BC); Communitarianism; Political Theory; Governments; Political Anthropology; Political Geography; Political Sociology; State and Society; State, History of

Bibliography

- Abelson R P et al. (ed.) 1968 *Theories of Cognitive Consistency: A Sourcebook*. Rand McNally, Chicago
- Bauer R A, Pool I de Sola, Dexter L A 1963 *American Business and Public Policy*. Atherton, New York
- Berlin I 1969 *Four Essays on Liberty*. Oxford University Press, London
- British Journal of Politics and International Relations* 1999 "Editorial: Studying British Politics," 1(1): 2
- Converse J M 1987 *Survey Research in the United States: Roots and Emergence 1890–1960*. University of California Press, Berkeley, CA
- Dahl R A 1956 *A Preface to Democratic Theory*. University of Chicago Press, Chicago
- Dahl R A, Lindblom C E 1953 *Politics Economics and Welfare: Planning and Politico-Economic Systems Resolved Into Basic Social Processes*. Harper, New York
- Deutsch K W et al. 1957 *Political Community and the North Atlantic Area: International Organization in the Light of Historical Experience*. Princeton University Press, Princeton
- Easton D 1953 *The Political System: An Inquiry into the State of Political Science*. 1st Edn. Knopf, New York
- George A L, George J L 1956 *Woodrow Wilson and Colonel House: A Personality Study*. J. Day Co., New York
- Greenstein F I 1969 *Personality and Politics: Problems of Evidence, Inference and Conceptualization*. Markham, Chicago
- Greenstein F I, Polsby N W 1975 *Handbook of Political Science* (Volume I). Addison-Wesley Publishing Company, Reading, MA
- Haas E B 1968 *The Uniting of Europe: Political, Social, and Economic Forces, 1950–1957*. Stanford University Press, Stanford, CA
- Lasswell H D, Kaplan A 1950 *Power and Society: A Framework for Political Inquiry*. Yale University Press, New Haven
- Lasswell H D 1936 *Politics; Who Gets What When How*. Whittlesey House, New York
- Polsby N W 1963, 1980 *Community Power and Political Theory*. Yale University Press, New Haven, CT
- Putnam R D, Leonardi R, Nanetti R Y 1993 *Making Democracy Work: Civic Traditions in Modern Italy*. Princeton University Press, Princeton, Jamaica
- Reese T J 1989 *Archbishop: Inside the Power Structure of the American Catholic Church*. 1st Edn. Harper and Row, San Francisco
- Reese T J 1992 *A Flock of Shepherds: The National Conference of Catholic Bishops*. Sheed and Ward, Kansas City, MO
- Reese T J 1996 *Inside the Vatican: The Politics and Organization of the Catholic Church*. Harvard University Press, Cambridge, MA
- Sartori G (ed.) 1984 *Social Science Concepts: A Systematic Analysis*. Sage, Beverly Hills, CA
- Simon H A 1957 *Models of Man: Social and Rational*. Wiley, New York
- Skinner Q 1978 *The Foundations of Modern Political Thought*. Cambridge University Press, Cambridge, UK

- Solow R M 1997 How Did Economics Get That Way and What Way Did It Get? *Daedalus* 126: 39–58
- Somit A, Tanenhaus J 1964 *American Political Science: A Profile of A Discipline*. Atherton Press, New York
- Tversky A, Kahneman D 1974 Judgement Under Uncertainty: Heuristics and Biases. *Science* 85: 1124–31
- Van Evera S 1999 *Causes of War: Power and the Roots of Conflict*. Cornell University Press, Ithaca
- Waltz K 1959 *Man, the State, and War: A Theoretical Analysis*. Columbia University Press, New York
- Weber M 1946 *From Max Weber: Essays in Sociology*. Oxford University Press, New York
- Wilensky H 1967 *Organizational Intelligence: Knowledge and Policy in Government and Industry*. Basic Books, New York
- Wolin S S 1960 *Politics and Vision: Continuity and Innovation in Western Political Thought*. Little, Brown, Boston, MA
- Zaller J 1992 *The Nature and Origins of Mass Opinion*. Cambridge University Press, Cambridge

N. W. Polsby

Political Sociology

Political sociology is a border field between political science and sociology, and the term encompasses the overlap between these two neighboring disciplines. It denotes the analysis of the inter-relationship between the social and the political, social structures and political institutions, between the society and the state. There is no stable consensus of what counts as political sociology in contrast to sociology and political science proper. There have been great variations over time in the popularity of political sociology and in the tendencies to emphasize it as a genuine field of its own. Nevertheless a common element is that political sociology is related to the distinction between the social and the political.

1. Historical Roots of Political Sociology

The historical development of political sociology can be roughly outlined by emphasizing three periods of special importance.

(a) The birth of political sociology can be dated to the mid-nineteenth century, when the founding fathers of social science began to conceive of the social and political, the society and the state as separate entities, and to analyze their relationships (Runciman 1963, pp. 22–42, Pizzorno 1971, pp. 7–18). This type of inquiry had its intellectual roots in the writings of earlier philosophers about the state and the social life, however, as a broad and joint interest of an academic community, political sociology emerged in the schol-

arship of the nineteenth-century fathers of social science.

(b) The post-World War II studies of the conditions for and the citizen activities in political democracies with a special focus on parliamentary elections conceived them as expressions of the political competition between different social groups. The joint focus on the prerequisites of democracy and voting patterns proliferated during the 1950s and 1960s, receiving institutionalized expression in the work and activities of the Committee of Political Sociology, founded in 1959 with Seymour Martin Lipset as chairman and Stein Rokkan as secretary and functioning under the auspices of both the ISA (International Sociological Association) and the IPSA (International Political Science Association) (Rokkan 1970, pp. 1–20).

(c) The late twentieth-century studies of nation and state building placed a special emphasis on great political transformations and breakdowns of regimes. This type of inquiry received a strong impetus from the renaissance of Marxist thinking in the 1960s and 1970s, but it was not restricted to social scientists with a leftist orientation (Linz and Stepan 1978, Skocpol 1979).

1.1 The Establishment of a Distinction between Society and State

In a great number of their texts, most of the path-breaking social science theoreticians of the second half of the nineteenth century and the early twentieth century analyzed the relationships between the social and the political. One of the first thinkers to delineate an empirical political sociology was Karl Marx in his systematic insistence on seeing politics as an expression of conflicts between different social classes. His theorizing had a factual historical background in the emergence of the bourgeoisie and the industrial working class as distinct and dominant social groups (Pizzorno 1971, pp. 7–9).

Max Weber on his part wrote many of his works as rejoinders to Marx, but even apart from his counterclaims against Marx, there is in Weber's writings continuous analyses of the social institutions and the types of authority existing in modern society (Runciman 1963, pp. 53–7).

Among the great founders of sociology Émile Durkheim is not normally thought of as a scholar of politics, but his preoccupation with order, social integration, and social solidarity had considerable implications for the analysis of the relationship between society and politics.

Vilfredo Pareto was a pioneer both in econometrics and in the sociological study of elites. His ability to relate the circulation of elites to large-scale social and political changes makes him one of the founding scholars of political sociology.

Karl Marx, Max Weber, Émile Durkheim, and Vilfredo Pareto were not primarily political sociologists, but instead more generally the great founding fathers of social science theory building. There are, however, some nineteenth-century scholars who can be labeled as specialists in political sociology. They are not unknown as scholars in history and in the study of politics, but their fame rests basically on the fact that they were important creators of the tradition of political sociology with its focus on the relationship between society and politics. Among them Lorenz von Stein, Alexis de Tocqueville, Mosei Ostrogorski, and Robert Michels deserve special mention.

Lorenz von Stein (1815–90) was a forerunner of Marxism in the sense that he explicitly emphasized how material and social conditions mould political development. His *Geschichte der sozialen Bewegung in Frankreich* (1850) contains not only a history of the French Revolution but also analyses of the importance of the property distributions and educational opportunities for political activities, and of the phenomena of class struggles and sudden outbursts of social change. It has been emphasized how some of Karl Marx's ideas were influenced by the writings of von Stein.

Alexis de Tocqueville (1805–1859) descended from an ancient and distinguished noble French family. His visit to the United States in 1831–2, which was of decisive importance for his scholarship, resulted in a study of American democracy published in two volumes *De la démocratie en Amérique* (1835, 1840). Although the work made him well known, his great scholarly fame was established only after the Second World War and to a considerable degree by the then growing political sociology. Especially the English translations of his works, *Democracy in America* (1945) and his study of the French revolution, *The Old Regime and the Revolution* (1955; originally published in French in 1856), became standard works for political sociologists. His great contribution was that he specified the conditions for pluralism and democratic systems. He argued that social cleavages and competition between social groups and strata are necessary conditions for consensus in industrially developed societies. He emphasized voluntary associations and active, partially autonomous local communities as crucial in creating and sustaining a plural democracy. In the texts of political sociology, de Tocqueville has often been contrasted to Marx. While Karl Marx envisaged conflicts and social cleavages as one-sidedly destructive, de Tocqueville emphasized their importance for obtaining consensus in well-functioning democracies (Lipset 1960, pp. 4–9).

Although Mosei Ostrogorski (1854–1919) was born in Tsarist Russia and died in the Soviet Union, during important parts of his life he worked in Paris and London. Early in his career he published studies of the suffrage demands of the women's movement, but his main work was a critique of party democracy in

Democracy and the Organization of Political Parties (1902). Although it was an empirical study of the political party organizations in Britain and the United States, its main theme was that the mechanical and massive organization of political life into bureaucratized political parties, aimed at manipulating individuals and masses, was harmful for real democracy.

As was Ostrogorski, Robert Michels (1876–1936) was a critic of the emerging party bureaucracies in modern states. Through his education and activities Michels had background liaisons with Germany, France, and Italy. Politically he started as a socialist with a syndicalist orientation but developed later profascist opinions. He took Italian citizenship in 1913, and although he was active as a professor and scholar in many countries he regarded Italy as his home country. His major work *Zur Soziologie des Parteiwesens in der modernen Demokratie* (1911) is considered a classic in political sociology. It treats many of its central themes, including the theory of elites, political movements on the left, processes of bureaucratization, and the gap between democratic theory and practice. His central thesis, known under the name of *the iron law of oligarchy*, is that organization necessarily leads to oligarchy, or as Michels summarized it, ‘it is organization which gives birth to the dominion of the elected over the electors, of the mandataries over the mandators, of the delegates over the delegators. Who says organization, says oligarchy.’ Despite Michel’s pessimism and one-sidedness, his work and the iron law of oligarchy has inspired a great number of studies of political and other civic organizations.

1.2 The Institutionalization of Political Sociology after World War II

The period between World War I and II, and especially the 1930s, were poor as regards studies in political sociology. The European political scene was turbulent, little support was given for studies in the social sciences, and in the 1930s the Nazi, Fascist, and Communist regimes ended most social science research in a number of countries. During the period between the world wars there were nevertheless some pathbreaking studies of voting and electoral behavior. Already before World War I André Siegfried had introduced the analysis of geographical variations in electoral behavior by studying voting patterns in French constituencies. A summary of his results was published in *Tableau des partis politiques en France* (1930). His central conclusion was that electoral behavior was steered by stable, deep-seated, regional traditions. After Siegfried the concept of *political traditions* became central in descriptions of voting patterns. Another innovator was the German-American Rudolf Heberle, who had started his career with an analysis of the tendencies to radicalization of the Swedish labor movement. His leading

work *From Democracy to Nazism* (1945) about the appeal of Nazism to the rural population in the German province of Schleswig Holstein was published when he already had settled as professor in Baton Rouge, Louisiana, USA.

Large-scale studies of voting behavior, however, were not undertaken until after World War II, and this subject became a leading problem area of political sociology during the first postwar decades. It had a background in the postwar situation, as great hopes were attached to the possibility of building a better world with the aid of social science and research. Studies of the prerequisites of democracy became important, and parliamentary elections were seen as the major mechanism for establishing and securing democratic rule. Studies of the social background of political behavior represented an almost empty space between political science and sociology. Political scientists had been mainly preoccupied with constitutional problems and methods of steering the states, whereas sociologists had focused their interest on studies of social structure and behavior in social groups. Political sociology moved into a vacuum, and it became soon both an important field of research and an activity of interest to the general public. Although the study of electoral behavior later became mainly a field within political science, the first pathbreaking studies were made by sociologists. Theoretically fruitful conceptions and hypotheses, such as the tendency of cross-pressures to lead to political passivity, were developed at the Bureau of Applied Social Research at Columbia University under the inspiration and leadership of Paul Lazarsfeld (Lazarsfeld et al. 1944).

It was, however, through the founding of the Committee of Political Sociology at the International Sociological Association’s Fourth World Congress in Stresa in September 1959 that political sociology became an established and institutionalized field of research. It led to a new kind of cooperation between political scientists and sociologists, and the committee also became a research committee of the International Political Science Association. The Committee’s first chairman, Seymour Martin Lipset, had already published several innovative books in political sociology, and its first secretary, Stein Rokkan, had started to develop means for comparative, cross-national research. The year 1959 represented in many senses an important mark in the development of political sociology. In 1959 Lipset obtained the copyright for his *Political Man. The Social Bases of Politics* (1960), which became the leading text of the period of birth of an institutionalized political sociology.

The references and data sources in Lipset’s book reflect the fact that at the end of the 1950s political sociology had become a small industry with young researchers all over the industrially developed world digging up data about the social bases of politics. The work of the Committee led to new developments. This is evident by a comparison of Lipset’s *Political Man*

and the first comprehensive book produced by co-operation within the Committee, namely Lipset's and Rokkan's *Party Systems and Voter Alignments: Cross-National Perspectives* (1967). It contains a lengthy and theoretical introduction by Lipset and Rokkan, and chapters about individual countries, including the United States, Canada, Great Britain, Australia, New Zealand, France, Italy, Spain, West Germany, Finland, Norway, Japan, Brazil, and some West African states. Both the theoretical introduction, with its application of Talcott Parson's A-G-I-L scheme, and most of the country chapters, reflect attempts to combine a structural and institutional approach and to study both the social bases of politics and the effects of politics on social structures. One of the members of the Committee, Giovanni Sartori, described the change in orientation in a paper entitled *From the Sociology of Politics to Political Sociology* (1969). Political sociology had ceased to be a subfield of sociology and had become a genuine border field of its own.

Studies of voting patterns were by no means the only ones conducted and presented at the conferences of the Committee and by its members. The first part of Lipset's book on the *Political Man* consists of an analysis of the prerequisites of the democratic order. At the turn of 1950s and 1960s several studies analyzed the conditions of democracy. In addition to the formal criteria of democracy such as popular representation and individual freedoms, political sociologists were keen to stress some factual social conditions as prerequisites of a democratic order. Of particular importance were *effectiveness*, the government's capacity to implement policy, *efficacy*, peoples' subjective feelings of being important for the politicians, and *legitimacy*, beliefs that the existing political institutions are appropriate and lawful. Some of the studies were particularly focused on breakdowns of democracy and the rise and appeal of Fascism and Communism, such as Raymond Aron's *L'Opium des intellectuels* (1955) and William Kornhauser's *The Politics of Mass Society* (1959). Democracy was studied from many angles, such as in the study done by Lipset and his associates on the conditions for democracy in labor unions, Robert Dahl's studies of democracy on the local community level, and Ralf Dahrendorf's analyses of the conditions of conflict regulation in industrial society.

Even if studies of and data about voting and electoral behavior formed a central point of departure for the political sociologists in the 1950s and 1960s, their aim was more general than to give information about how people behave during elections. The basic interest was in studying conditions for democracy, breakdowns of regimes, nation and state building, and master processes of societal and political change. From the 1970s onward the electoral studies became methodologically more sophisticated at the same time as they increasingly became a preoccupation of spec-

ialists of political issues and mass media effects within the field of political science. Initially such studies had a center at the University of Michigan under the leadership of Angus Campbell.

1.3 The Re-entry of the State and its Transformations

The disappearance of electoral studies from the central agenda of political sociology was not an isolated phenomenon. From the beginning of the 1970s many of the original topics of political sociology lost much of their popularity. One of the reasons was that the student rebellions and leftist radicalism in the 1960s and 1970s had severely criticized the agenda of the prevailing political sociology. Another reason was that some fields of research had become routine and were incorporated into either sociology or political science.

Despite a decline of the distinctiveness of political sociology, some research themes have remained central and typical for political sociology. The problem sphere of political sociology comprises increasingly master patterns of societal change, i.e. simultaneous changes and pressures to change in the political order, the social system, and the socioeconomic composition of societies. Such studies had been initiated by members of the Committee of Political Sociology already during its preoccupation with voting studies. Thus, there were studies of the formation and fall of empires by S.N. Eisenstadt (1963), studies of the historical formation of European centers and peripheries by Stein Rokkan (Rokkan and Urwin 1983), and studies of the breakdown of democratic regimes by Juan Linz (Linz and Stepan 1978). Also many scholars who never had been associated with political sociology as such wrote influential texts. An outstanding representative in this category is Barrington Moore's often-quoted major work *The Social Origins of Dictatorship and Democracy* (1966).

In the 1980s and 1990s interest in master patterns of change, and in state and nation building proliferated. The state was reintroduced as a truly central concern. Major representatives of the new orientation were especially Theda Skocpol (1979), who penetratingly emphasized the state as the originator of social outcomes, and Charles Tilly (1975), who rather than seeking one particular important unit, has stressed how national developments were formed by a complicated web of warfare, fiscal policy, bureaucratization, and state-making. Göran Therborn's (1995) studies of modern developments in the European states have also emphasized the multidimensional nature of national development and warned against beliefs in irreversible tendencies. The increased interest in state and nation building also went beyond a focus on national states by studies of worldwide tendencies of development, as seen in the work of Immanuel

Wallerstein (1974) about the origins of the European world economy in the sixteenth century.

The developments from the 1970s onward made it increasingly difficult to define the borders of political sociology. There was for instance a recrudescence of studies of ethnicity, ethnic groups, and the role of ethnic identities in the formation of nations. The studies of ethnicity are sometimes counted as part of political sociology, sometimes as a field of its own. Yet, the borderline problem is fairly clear. It is a question of where the main emphasis is. When the aim is to describe the role of ethnicity and ethnic identities in the formation of nations and political development on the macro level, as in Stein Rokkan's (Rokkan and Urwin 1983) studies of the formation of European centers and peripheries, they are of great relevance for political sociology.

It is much more difficult to evaluate the character and place of Marxist-oriented and radically critical studies which multiplied from the second half of the 1960s onward. Earlier this type of scholarship had been mainly carried out by specific individuals. In studies of the power elite, C. Wright Mills definitely treated the relationship between the social and political, but he was not generally counted among the political sociologists (Kimmerling 1996). In the 1960s there was a rise—indeed a renaissance—of whole schools of critical sociologists such as the Frankfurt School of dialectical sociology and the dependence theorists who emphasized the hopeless fate of the underdeveloped countries in a world dominated by international capitalism. It is an open question whether these and kindred groups should be included in political sociology. Usually they were not counted as such. It is again a question of how one conceives of their main emphasis. If the main intent is to present social critique, the studies are usually not counted as political sociology, however, if the studies present hypotheses and concepts useful in most studies of societal and political development, they are regarded as political sociology. In the 1980s the radical thrust in the social sciences came to an end. The studies in political sociology began to combine elements from studies of very different political shades. It is characteristic of most of the studies of nation and state building mentioned above that they had assimilated elements from both Marxist and liberal thinking and scholarship.

2. The Consciousness-creating Functions of Political Sociology

Whatever solution is reached as regards the labeling of radical social critiques, it is obvious that political sociology is always directed towards expanding and creating awareness of macro changes and fundamental problems of societies. Nevertheless, it is seen as important that political sociology should not basically

serve political aims. There are strong claims that political sociology should be empirically oriented, aim at theory building, employ a comparative approach, and systematically consider historical factors. The combination of the two claims, the consciousness-expanding function on one hand, and the quest for systematic comparative, historically based, and theoretically oriented studies on the other, has been astutely analyzed by Birgitta Nedelmann (1997) in a critique of the failure of German postwar political sociology to account for the period of National Socialism in German history.

Even if a successful delineation of political sociology never emerges, some main problem areas can be listed with the help of the two claims and objectives listed above. Political sociology studies (a) master patterns of simultaneous social, cultural, and political change in society, (b) dominant sociopolitical and emerging new cleavages, (c) the formation of states and nations as well as breakdowns of social and political orders, (d) prerequisites of and threats to democratic regimes, (e) problematic features of modern states such as bureaucratization, centralization, tendencies to corporatism, and (f) the political consequences of globalization such as the formation of continent-based political units and institutions.

See also: Marx, Karl (1818–89); Pareto, Vilfredo (1848–1923); Political Parties, History of; Political Science: Overview; State and Society; State, Sociology of the; Tocqueville, Alexis de (1805–59); Voting: Class; Voting, Sociology of; Weber, Max (1864–1920)

Bibliography

- Aron R 1955 *l'Opium des intellectuels*. Calman Levy, Paris
- Eisenstadt S N 1963 *The Political System of Empires. The Rise and Fall of the Historical Bureaucratic Empires*. The Free Press, Glencoe, IL
- Heberle R 1945 *From Democracy to Nazism: A Regional Case Study on Political Parties in Germany*. Louisiana University Press, Baton Rouge, LA
- Kimmerling B 1996 Changing meanings and boundaries of the 'political'. *Current Sociology* 44: 152–76
- Kornhauser W 1959 *The Politics of Mass Society*. The Free Press, Glencoe, IL
- Lazarsfeld P F, Berelson B, Gaudet H 1944 *The People's Choice. How the Voter Makes up His Mind in a Presidential Campaign*. Columbia University Press, New York
- Linz J J, Stepan A (eds.) 1978 *The Breakdown of Democratic Regimes*. Johns Hopkins University Press, Baltimore, MD
- Lipset S M 1960 *Political Man. The Social Bases of Politics*. Doubleday, Garden City, New York
- Lipset S M, Rokkan S (eds.) 1967 *Party Systems and Voter Alignments. Cross-National Perspectives*. The Free Press, Glencoe, IL
- Michels R 1911 *Zur Soziologie des Parteiwesens in der modernen Demokratie. Untersuchungen über die oligarchischen Tendenzen des Gruppenlebens*. Klinkhardt, Leipzig, Germany

- Moore B Jr 1966 *The Social Origins of Dictatorship and Democracy: Lord and Peasant in the Making of the Modern World*. Beacon Press, Boston, MA
- Nedelmann B 1997 Between national socialism and real socialism. Political sociology in the Federal Republic of Germany. *Current Sociology* 45: 157–86
- Ostrogorski M Y 1902 *Democracy and the Organization of Political Parties*. MacMillan, London
- Pizzorno A (ed.) 1971 *Political Sociology. Selected Readings*. Penguin Books, Harmondsworth, UK
- Rokkan S 1970 Introduction: international cooperation in political sociology. In: Allardt E, Rokkan S (eds.) *Mass Politics. Studies in Political Sociology*. The Free Press, New York
- Rokkan S, Urwin D W 1983 *Economy Territory Identity. Politics of West European Peripheries*. Sage, London
- Runciman W G 1963 *Social Science and Political Theory*. Cambridge University Press, Cambridge, UK
- Sartori G 1969 From the sociology of politics to political sociology. *Government and Opposition* 4: 195–214
- Skocpol T 1979 *States and Social Revolutions: A Comparative Analysis of France, Russia and China*. Cambridge University Press, New York
- Siegfried A 1930 *Tableau des partis politiques en France*. Grasset, Paris
- von Stein L 1850 *Geschichte der sozialen Bewegung in Frankreich von 1789 bis auf unsere Tage*. Leipzig
- Therborn G 1995 *European Modernity and Beyond. The Trajectory of European Societies 1945–2000*. Sage, London
- Tilly C (ed.) 1975 *The Formation of National States in Western Europe*. Princeton University Press, Princeton, NJ
- Tocqueville A C de 1835, 1840 *De la démocratie en Amérique I–III*. Paris. [Published in English 1945 *Democracy in America*. Knopf, New York]
- Tocqueville A C de 1955 *The Old Regime and the Revolution*. (trans. from French), Doubleday, New York
- Wallerstein I 1974 *The Modern World System. Capitalist Agriculture and the Origins of the European World-Economy in the Sixteenth Century*. Academic Press, New York

E. Allardt

Political Thought, History of

Reflection about political phenomena in a wide sense of the term is as old as human community itself. It focuses on a triad of related problems that might be broadly delimited as follows. First, there is the problem of constituting a community bound together by a normative order, secondly that of establishing the legitimacy and the terms of enforcement of such an order, and, thirdly, that of finding criteria for a proper and just allocation of resources as well as of entitlements. Jointly these three sets of problems entail that political theory and the history of political theory, as distinct from, say, a theory of scientific validation or a theory of argumentation, will be characterized by an inherent duality. Thus it will not only, as these other types of theory, be an account of rules of contestation and signification. It will also have to involve an account of which such rules are and which are not

upheld by a normativity that is, in the last instance, enforceable by violent means. Put in everyday language, political theory will in its core be a theory both about meaning and communication and a theory about power, enforcement, and, ultimately, violence. It is easy to see that different political theorists have come to emphasize different strands in this duality, but the duality itself is irreducible and constitutive of political thought.

1. Political Thought in Historical Context

Thought about such matters is found in clearly articulated form from the middle of the first millennium BCE onwards in all the major civilizations of Afro-Eurasia. Thus it is manifested in written texts in the sixth century by Confucius and two centuries later by Mencius, in China and the writings during the same period in India, as well as in the prophetic age of Jewish religion and during the era of classical philosophy in Greece. The emergence of imperial-like political orders across Afro-Eurasia in the last centuries of the first millennium BCE also gave rise to a range of texts about the nature of political order. This is equally valid for texts about the Roman and Byzantine empires in the Mediterranean region to texts highlighting features of rulership in Parthian and later also in Sassanian Persia, in Maurya India, and in the Han Empire of China.

The recording of political thought in the form of the writing of histories of such thought is however a much more recent phenomenon. At its focus it is the relationship of community-wide rules to forms of enforcement and an effort to arrive at an understanding of the historical context in which particular types of such relationships occur. The emergence of clearly articulated forms of historical thought is largely coterminous with the emergence of history and social science as forms of scholarly activity performed within reformed or newly created universities and higher education institutions from the early nineteenth century onwards. In order to understand the key features and characteristics of the writing of histories of political thought it is first of all necessary to bring out key conditions of political thought in the region in which such an early disciplinary condition first occurred, namely in Europe.

2. Political Thought and Early Modernities

In a broad sense political thought in early modern Europe was a way to understand the nature of ruling in human communities and to grasp the conditions of normative argumentation and enforcement in such communities. The need for new understandings of these types of phenomena was itself directly related to the gradual decline and eventual demise of the two

most prominent institutional forms of medieval Europe. Those were, first, the cultural world of Western Christendom, epitomized first by the all-embracing Christian Church and second by its secular counterpart, namely the pervasive, but elusive, political order of the Holy Roman Empire. By the twelfth and thirteenth centuries, however, city republican rule, particularly in the northern half of Italy, had emerged as a new form of political order. Originally it was often modeled on trading communities and was created out of a sense of the necessity of mutual support and protection in areas exposed to incursions and with little or no protection provided by the nominal Imperial rulers.

Similarly, incipient national monarchies were starting to assert their authority vis-à-vis both pope and emperor. In this secular process the French king played a pioneering role. In particular, he had earlier than most other kings and rulers demonstrated his ability to strengthen his own position and financial basis, e.g., by confiscating lands and goods of the order of the Templars. This example was to be followed by kings across Europe in their confiscating the lands and riches of monasteries and religious orders in centuries to come.

In this sense, the Reformation and the Counter-reformation meant that in Protestant and Catholic countries alike, national monarchies, normally of a composite nature (cf. Elliott 1992), were created or reformed. This occurred on the basis of the previous dominant ideals of church, empire, and city republics, but in a form that synthesized parts that had previously been separated. Furthermore, the new national and composite monarchies almost invariably involved efforts on the part of rulers to appropriate lands and resources that had previously been controlled by the church or religious orders. Thereby they were also able to secure a fiscal basis for the new forms of political ruling, if often at a horrendous price of protracted religious civil wars and cleansings.

Under these conditions the pre-eminent political questions became that of grasping the nature and legitimacy of a normatively binding communal order and to outline the range of legitimate means for upholding and enforcing that order. Many of the early modern classical political thinkers, from Machiavelli to Jean Bodin, provide answers to these very questions. They have, however, significantly less to say about the two other sets of questions discerned above and that were to become prominent in later political thought. These were the questions of the very constitution of a community of human beings and of the principles underlying decisions about the allocations of resources or the granting or withholding of entitlements.

All this was to change with the final breakdown of the old political and cultural order in Europe in the middle of the seventeenth century. In China, even in the embattled transition from the Ming to the Ching dynasty at roughly the same time, the general linguistic

and political order was preserved. In Japan a long period of internal warfare and turmoil came to an end with the establishment of one of the most effective forms of absolutist rule, namely the Tokugawa shogunate. Across the Islamic world, the three parallel empires of the Ottomans, the Safavi, and the Mughals, constituted stable political entities. In Europe, however, a new system of states took form and was promulgated in the Westphalian peace treaty of 1648, based on an explicit recognition of religious diversity and monarchical pre-eminence. It was also based on the tacit recognition that an all-encompassing imperial polity was no longer even possible as a viable political ideal and that religious diversity was not to be overcome.

3. Political Thought as the Constitution of a Political Community

In this unfamiliar context, there was an urgency about finding ways of granting legitimacy and meaning to the new order that emerged in the wake of the Westphalian peace treaty and the provisional end of a century and a half of bloody civil wars in Europe. Thereby a sea change occurred in political thought, namely a deep-seated shift from a concern with the enforcement of normative political order to a concern with the nature and very emergence of political community in the first place.

The concern with the constitution of political community becomes the central one in political thought from Hobbes and Pufendorf to Locke all the way to Rousseau. Within this wide range of answers there is, however, a recurring and predominant concept, namely an inquiry into the conditions for the establishment of a social contract. The existence of such a social contract was normally seen also to entail an answer to the question of which types of enforcement of the normative order constitutive of the political community are legitimate. More specifically only those types of enforcement are sanctioned that, implicitly or explicitly, follow from the terms of the social contract itself.

Political theory from Hobbes onwards is centrally preoccupied with the nature of and the consequences of a transition from an imaginary state of nature to the contractually constituted political community. Ultimately all such forms of theorizing seem to depend on a limited set of assumptions about human beings and the nature of their joining together in societies. One such set of assumptions refer to the assignment of some kind of human rights, i.e., rights that accrue to human beings irrespective of their origins and other accidental traits and characteristics. Furthermore, there is a set of assumptions that refer to the thought-experiment of society being created through a process, normally described in terms of a contract, whereby

human beings agree about their association in political form.

Questions about enforcement, separate from an answer to the question of the constitution of a political order, based upon legitimately agreed upon contractual arrangements, were of subordinate and derivative importance relative to this overriding concern. The same is true of questions pertaining to the allocation of resources and the assignment of entitlements to the members and subjects of a political community. Thus such matters have to be dealt with and negotiated in accordance with the terms of those agreed upon in the original contractual arrangements.

Thus European political theory in this classical form emerged in a particular historical context both in European and world history. It also involved forms of theorizing that portrayed an imaginary contractual past as the point of constitution of polities. However, it was not in itself characterized by an effort to reflect on the history of political theory as a key component of developing such a theory. This form of reflexive consciousness was to become a key characteristic of yet another major transformation in intellectual and political history.

4. The Formation of Modernity and the Transition from Political Philosophy to Social Science and History

The late eighteenth and early nineteenth centuries are often described in terms of the coterminous occurrence of major transformations of both technological and economic practices and of political order. Often these transformations are captured with terms such as the industrial revolution and the political transformations inherent in the French and American revolutions. However, in this period there also occurs a major restructuring of ideas about nature and society and, more broadly, in the forms of knowledge and beliefs about the world. This intellectual transformation had deep repercussions for the emergence of new forms of political thought and reflection about the historical context of such forms of thought and for the decline of other forms of political thought.

In this period broad genres of thought about the world gradually gave way to more differentiated forms of knowledge. Slowly natural philosophy and political philosophy came to be replaced by a variety of different natural and social sciences. At the core of this transformation were epistemic shifts where ideas about biological and historical processes, about language and interpretation, about human beings and their capacities and bonds were created or reformulated. Some scholars have even spoken of this period as a second scientific revolution, where modern disciplines emerge in rudimentary form and where there is a shift in the institutional locations of scholarly and scientific activities toward new research laboratories

and universities. In this process, society itself may even, to quote one scholar (Manent 1994, pp. 80–3), be said to have been discovered in its modern form, as distinct from the political order of the prerevolutionary period, whether it had the form of an absolutist monarchy or a constitutional one. Contrary to older forms of contractual reasoning, which postulated social contract as the hypothetical medium that had created political order, the new forms of social thought were distinctly historical and could also draw on anthropological experiences of human beings and societies in different parts of the world (Fox et al. 1995).

Broadly speaking, four distinct forms of thinking about human beings, societies, and political order emerged. These new forms of reasoning and the transition from generalized moral and political philosophy to social sciences had deep repercussions for political thought.

First, there was, with roots in the political economy and moral philosophy of the Scottish Enlightenment, a view of human beings as rational and deliberative individuals and of society as a compositional collective. This also made possible a new way of thinking about political community and its legitimacy. More precisely, it provided the possibility for thinking about political order as formed through the consent of individual human beings. However it was not, as in older contractual thinking, a consent that would have ushered in a distant and hypothetical social contract but rather a consent that could be assessed and rationally evaluated in terms of its reasonableness in its present and historical context.

Second, there emerged, often enough out of the experiences of the relative inability of the French Revolution to create stable institutions, a conception of human agency as formed and constrained by structural and systemic properties and of society as an organic totality. Such a conception would in a conservative version lead to theories of the necessity to restore an organic political order. In a radical version it could, however, also usher in demands for the overthrow of unnatural and illegitimate order. In a more broadly defined liberal or republican tradition, it might and would be compatible with ideas of the type that thinkers, often inspired by Durkheim, would propound toward the end of the nineteenth century in their combat against the ideological enemies of the republic. Thus it could support the idea that modern society was not an atomistic and distorted form of human community but could relate to an organic solidarity compatible with the life of free citizens in a free republic.

Third, and often in relation to historical and linguistic thought among representative of German philosophical idealism, a conceptualization of human beings took form in terms of the linguistic-interpretative nature of human agency and of society as an historically emergent totality. In this sense, political

community did not have to derive its legitimacy from dynastic tradition but from the historically grown experience of a community. Such communities in turn were constituted by human beings sharing a language and a history and where the community of shared life forms was seen as deeper, more genuine, and more real than the artificial boundaries of a minor princely state. It is easy to see that demands for national self-determination and for both popular rule and a nation-state easily flow from a conception along these lines.

Finally, in parallel to the dramatic development of forms of statistical analysis, there emerged a statistical-inductive way of representing the interactions of human beings and of society as a kind of systematized aggregate. Such views did not immediately entail a specific form of legitimate political community, but it would mean that any such community would have at its disposal infinitely more detailed knowledge of the polity and its members than even the most absolutist monarchy (cf. Brian 1994).

Thus the great intellectual transition entailed a deep shift in the history of political thought which affected both conceptions of the nature of a political community and the nature of legitimate and illegitimate enforcement. By virtue of this it also had profound implications for conceptions of the proper distribution of entitlements. Thus it was simply no longer possible, as observed most clearly perhaps by Tocqueville (1964, [1850]), to deny demands for popular participation with reference to tradition or dynastic legitimacy. The thought of a right to the political entitlements belonging to all members of a society, that had been manifested in the proclamations of the revolutionary upheavals, could never again, despite all efforts of the Vienna Congress and the Holy Alliance, be unthought. They were to pervade the political struggles first of European and American societies throughout the nineteenth and the early twentieth centuries and eventually in the world at large.

5. Triumph and Crisis of Liberal Political Thought

The end of World War I marked a sea change in world history. It spelt the end of European global pre-eminence. It also meant that the transition in Western Europe, normally from forms of constitutional monarchical rule, to parliamentary democratic rule was dramatically increased. In the Eastern half of Europe all of the four imperial political entities, the Russian, the Habsburg, the Wilhelminian, and the Ottoman, crumbled under the impact of external military defeat and internal revolutionary upheavals. However, the very triumph of liberal principles of representative democracy was coterminous with the emergence of new forms of political thought that transcended the traditions of the prewar world.

Leninist political thought, which before and during the war had been a radical version of classical Marxism, gradually came to take on the features of an official ideology of the new Soviet state. Precisely in its historical imaginary, Leninism differed from both revisionist and orthodox thought in a Marxist tradition, as epitomized by Bernstein and Kautsky respectively, of the Social Democratic mainstream of the Second International in the late nineteenth and early twentieth centuries. Thus Leninism cast its own claims to legitimacy and its right to use forms of violent enforcement, previously unthought of in late nineteenth century socialist thought, in terms of an historical account and of a projected visionary transition to a stateless society of communism.

Experiences of the mass violence of the war itself, as well as of the violent national and social struggles of the immediate postwar period, formed the environment out of which fascism emerged as a violent rejection of both Marxism and of the whole mainstream tradition of liberal political thought. Even if fascism had roots in prewar forms of radical conservative thought, it marked a form of transcendence of such traditions and indeed of the legacy of modernity itself. Thus even if fascism in power, and its radicalized form of National Socialism, embraced technological modernization, it represented a rejection of basic premises of political thought of modernity in the form it had taken since the late eighteenth century. In particular this is true of the basic presupposition of political thought of modernity of a principled equal right to participation in governance. It is also true of fascism's rejection of another basic presupposition of modernity, namely the legitimacy of a public sphere with free access for all and a principled freedom of every member of a polity to speak not only about but also to the rulers.

In a sense fascist political thought was a form of political thought that invoked history, or rather historical mythology. However, it was not in itself characterized by historical reflection. In Germany, the challenge of fascism to the political order of the Weimar republic was manifested in a range of debates, one of the most important of which concerned the nature of legal order and of the state itself. It was in this debate that the contesting positions of the radical and pro-Nazi decisionism of Carl Schmitt was confronted by the legal positivism of Hans Kelsen and the complex and sophisticated thought of Hermann Heller.

The outlines of a radical rethinking of the history of Western philosophical and political thought can be discerned in the elective affinity between fascism and some forms of philosophical thought, in particular that of Heidegger of the late 1920s and early 1930s. However, such a program was never elaborated or extended, nor was the decisive response and resistance to it on the part of a thinker such as Ernst Cassirer. Instead the military defeat of fascism in World War II

was seen to entail the definitive demise of features of political thought characteristic of fascism.

In the French context, the response to the challenges of fascism was to be found in several quarters. These would range, in political terms, from Marxism to radical Catholic political thought, and in scholarly terms, from sociology in the tradition of Durkheim to the profound renewal of historical studies of the Annales School. In the period after World War II, the French intellectual scene may have been the most important one in Europe in terms of a renewal of political thought. Thus it would comprise the nexus of contributions made by scholars elaborating a Marxian philosophical existentialist position, most importantly of course Sartre, but also thinkers such as Claude Lefort, who, starting out from such a position came to question the core of Leninist political thought. It would also comprise some of the most articulated liberal political thought as epitomized by Raymond Aron.

In the interwar years, the new challenges to liberal political thought had already given rise to a profound renewal of political thought in historical context. These were the years when new standard histories of political thought were being written. The most prominent example in the Anglo-Saxon world might have been a book that was to become a classic for generations, namely George Sabine's (1937) majestic volume *A History of Political Theory*. Refugees from continental Europe made key contributions to this debate. A classical example is the analysis of the meaning of democracy in historical context that was articulated by Joseph Schumpeter (1943) in *Capitalism, Socialism, and Democracy*. Most importantly, perhaps, Karl Popper transcended the bounds of philosophy of science and wrote one of the most quoted works in political thought during the twentieth century in his two-volume opus on *The Open Society and Its Enemies* (Popper 1966, [1945]). This was, of course, a work that was pre-eminently characterized by the links between historical analysis and political philosophy of immediate and contemporary relevance.

In parallel, leading representatives of Scandinavian political and legal science made some pathbreaking contributions both to an understanding of fascism and of the nature of democracy. In this respect the names of Herbert Tingsten and Alf Ross stand out, particularly with Tingsten's pioneering studies of fascism and his and Ross's contributions to a political theory of democracy.

One prominent characteristic of antiliberal political thought of the first half of the twentieth century was its emphasis on the right of the polity to resort to enforcement and violence against the enemies of the polity without being bound by any other constraints than those of pure convenience. Thus the constitutive boundary between members and enemies of the political community, between those chosen to fulfill the destiny of that community and those deemed to

stand in the way and worthy of being removed from community, and ultimately from life, became the most crucial one. The political community itself was seen as constituted in terms of a historical account, whether of the struggle of a class for its liberation against oppression or of a people for its legitimate sphere of life and domination. Resistance to such types of political thought of a historicist nature formed the core of the argument of Popper's political philosophy. It also formed the focus for a range of contributions criticizing totalitarian political thought. Among the main contributors to this critique were thinkers as different as Jacob L. Talmon, Hannah Arendt, and Claude Lefort.

6. *Renewal and Critique in Political Thought*

The history of political thought in the first part of the latter half of the twentieth century is characterized by three parallel developments. First, political thought in this period cannot but reflect the deep tensions in a world with two superpowers in a state of cold war. At the same time large parts of the Third World were emerging out of the experiences of a colonial past and seeking a place in this new world of fundamental tensions. One manifestation of this historical context, was the absolute absence of publications in the tradition of liberal political thought in the countries under Soviet dominance, but also the relative absence of Marxist political thought at many universities in the English-speaking world. Another indirect consequence was a persistent interest in exploring the empirical basis of liberal democracy. In this sense, the contributions by scholars such as Robert Dahl and some of his colleagues at Yale university, not least Charles Lindblom, stand out as exemplary. Second, a classical mainstream tradition of historically orientated political philosophy made significant advances. Isaiah Berlin and Judith Shklar are pre-eminent representatives of this type of scholarship, and the works of Charles Taylor and Norberto Bobbio should also be included here. Third, however, philosophical and theoretical contributions were made that were to reshape the history of late twentieth century political thought. I shall highlight some of these contributions and their consequences in terms of two of the major conceptualizations of human agency and of society that were discerned in the section above about the rise of the social sciences at the turn of the eighteenth century.

In a rationalistic-deliberative tradition, a landmark was established by the publication of Kenneth Arrow's (1951) *Social Choice and Individual Values*. A few years later, economic theorizing would enter political science at large with Anthony Downs's (1957) book *An Economic Theory of Democracy*, to be followed in the early 1960s by a range of books applying game theory and social choice theory to political phenomena.

(Incidentally, as testified by Gabriel Almond (1990, p.117) the institutional role of the Center for Advanced Study in the Behavioral Sciences at Stanford was a crucial one for some of the interactions involving Arrow, Dahl, and others. These interactions were of immediate importance to the very publication of Downs's book but also more generally to the breakthrough of rationalistic forms of analysis).

A few decades later, what had once appeared as an interesting but rather marginal concern of some specialized political theorists had evolved into a broad tradition of rational choice analysis with pervasive influence on both the theory and practice of political thought, particularly in North America. Society as a compositional collective was described not just as an analytically convenient metaphor but as a representation of societal order and with consequential claims concerning the nature of legitimate political action. One thrust of theorizing in this tradition was to demonstrate the limitations, or even the impossibility, of delimiting an unambiguous common good, whether arrived at with the help of a social welfare function or of a majority decision. It is in this context that John Rawls's (1971) *A Theory of Justice* appeared as a pathbreaking book. Drawing on an individualistic and contractual form of reasoning, succinctly criticised by several scholars, most notably perhaps Michael Sandel, Rawls was able to reintroduce discussions of justice and just distributions and allocations into the core of political philosophy. Even after decades of discussions and critique, both Arrow's and Rawls's contribution retain their stature. However, both of them signal a rupture with historical and contextual reasoning rather than the opposite.

In the broad tradition of linguistic-interpretative conceptions of human action and of society as a form of emergent totality in historical context, important new developments also occurred. Thus contributions by continental European phenomenological and hermeneutic philosophers in the 1950s and 1960s came to merge with influences from Anglo-Saxon linguistic philosophy in the tradition of Wittgenstein and Austin. Three forms of such thought, in particular, emerged and have exerted a deep influence on the ways of studying the history of political thought.

In 1959, Reinhart Koselleck (1987), drawing on a tradition of hermeneutic and historical scholarship, published a book on critique and crisis. This book had the deep transformation of intellectual and political order at the turn of the eighteenth century as its focus but also contained the core of a research program on the history of key concepts of society and political order in a long-term perspective. This program was realized through the publication during the following decades of a set of nine massive volumes tracing the history of a limited number of such concepts, primarily in continental Europe.

Roughly at the same time, Jürgen Habermas elaborated another research program, drawing on the joint

traditions of critical Marxist thought, hermeneutics, and speech act theory. In 1962 Habermas published a volume that had an analogous empirical focus as that of Koselleck (Habermas 1989). Thus it was a reinterpretation of the emergence of a new type of public sphere in the late eighteenth century. Habermas's subsequent long-term program has consistently linked these different traditions and may in one sense be seen as a form of historical positioning but also a radical generalization of the philosophical insights of linguistic philosophy and speech act theory.

This later tradition also inspired Quentin Skinner's (1978) two volume major opus *The Foundations of Modern Political Thought* (a work, incidentally, written largely at the Institute for Advanced Study at Princeton). Skinner came to be the major representative of a new school, often identified with his—and John Pocock's old—university, namely Cambridge. It involved the setting of new and significantly more rigorous standards for the writing of the history of political thought, particularly in its demand that the linguistic and intellectual context of the actual writing of major works be carefully brought to the fore of analysis. After decades of discussions, there can be little doubt that intellectual history has made advances in this period, inspired by the works of Skinner and several of his colleagues, that still by no means have been fully explored. At the same time, it is also clear that these works still leave room for a more extensive inquiry into the relationship between long-term institutional and intellectual change.

Finally, links between political thought and political agency are prominent in studies of mechanisms of inclusion and exclusion and on the granting and withholding of entitlements for human beings. Gender studies has clearly entailed a renaissance of the writing of the history of political thought with pioneering contributions by Carole Pateman and Seyla Benhabib, just to mention two of the most distinguished scholars on the late twentieth century.

Historically orientated scholarship has in recent years seemed to open up the possibility for situating political thought in a much more encompassing context than has hitherto been customary. Not least, scholars writing in countries outside of the North Atlantic and West European context seem now to be about to help bring about a more truly global effort to rethink the history of political thought. That would also, almost inevitably, contribute to a transcendence of some of the most severe limitations of contemporary political thought, in particular the following ones:

The relative inability of political thought in the rationalistic tradition to transcend a theory of agency that is manifestly impoverished vis-à-vis available historical research but also vis-à-vis demands that social and political thought itself cannot but make concerning a theory of agency (cf. Joas 1996, 2000).

The relative inability of the so-called communitarian critics of this tradition to supply a sustained historical account of their

preferred community and to escape a tendency of entering into circular justifications of those very societies which they themselves happened to be located or identified with.

The relative inability of research in the traditions of conceptual history and linguistic contextualism to contribute to an understanding of the institutional formation of political order in a long-term and also extra-European perspective.

In order to be able to advance beyond these limitations, it seems that political thought must become significantly more prepared to engage in studies of long-term processes of intellectual and institutional transformations. This will in all likelihood entail efforts to invigorate historical social science of a type that was characteristic of the classics of social science, not least Weber (cf. Schluchter 1996). It would then also stimulate culturally and contextually sensitive social science of a type that has been advocated most ambitiously and consistently by scholars across a range of the social and historical sciences and with S. N. Eisenstadt (1995) as one of its most prominent representatives. The history of political thought in the last three decades of the twentieth century involved impressive advances. At the turn of the millennium, however, the need for historically and comparatively orientated research programs able to transcend the limitations of contemporary political thought also stand out as possibly more urgent than at any time since the emergence of social science itself in the end of the eighteenth century.

See also: Communitarianism: Political Theory; Democratic Theory; Enlightenment; Hermeneutics, History of; Integration: Social; Intellectual History; Legitimacy: Political; Liberalism: Historical Aspects; Locke, John (1632–1704); Marxist Social Thought, History of; National Socialism and Fascism; Rational Choice in Politics; Scientific Disciplines, History of; Solidarity: History of the Concept

Bibliography

- Almond G A 1990 *A Discipline Divided: Schools and Sects in Political Science*. Sage, Newbury Park, CA
- Arrow K 1951 *Social Choice and Individual Values*. Wiley, New York
- Benhabib S 1992 *Situating the Self: Gender, Community and Postmodernism in Contemporary Ethics*. Routledge, New York
- Brian E 1994 *La Mesure de l'Etat: Administrateurs et Géomètres au XVIII^e Siècle*. Albin Michel, Paris
- Downs A 1957 *An Economic Theory of Democracy*. Harper, New York
- Eisenstadt S N 1995 *Power, Trust and Meaning*. University of Chicago Press, Chicago
- Elliott J H 1992 A Europe of composite monarchies. *Past and Present* 137: 48–71
- Fox C, Porter R, Wokler R (eds.) 1995 *Inventing Human Science: Eighteenth Century Domains*. University of California Press, Berkeley, CA
- Habermas J 1989 *The Structural Transformation of the Public Sphere: An Inquiry into a Category of Bourgeois Society*. MIT Press, Cambridge, MA

- Harré R 2000 *One Thousand Years of Philosophy: From Ramanuja to Wittgenstein*. Blackwell, Malden, MA
- Heilbron J 1995 *The Rise of Social Theory*. Polity Press, Cambridge, UK
- Heilbron J, Magnusson L, Wittrock B (eds.) 1998 *The Rise of the Social Sciences and the Formation of Modernity: Conceptual Change in Context, 1750–1850*. Kluwer, Dordrecht, The Netherlands
- Joas H 1996 *The Creativity of Action Polity*. University of Chicago Press, Chicago, IL
- Joas H 2000 *The Genesis of Values Polity*. University of Chicago Press, Chicago, IL
- Koselleck R 1988 [1957] *Critique and Crisis: Enlightenment and the Pathogenesis of Modern Society*. MIT Press, Cambridge, MA
- Lefort C 1988 *Democracy and Political Theory*. University of Minnesota Press, Minneapolis, MN
- Manent P 1994 *An Intellectual History of Liberalism*. Princeton University Press, Princeton, NJ
- Pagden A (ed.) 1987 *The Languages of Political Theory in Early Modern Europe*. Cambridge University Press, Cambridge, UK
- Pocock J G A 1975 *The Machiavellian Moment*. Princeton University Press, Princeton, NJ
- Popper K 1966 [1945] *The Open Society and its Enemies*, 2 vols. Routledge, London
- Rawls J 1971 *A Theory of Justice*. Belknap Press of Harvard University Press, Cambridge, MA
- Sabine G 1937 *A History of Political Theory*. Holt, New York
- Schluchter W 1996 *Paradoxes of Modernity: Culture and Conduct in the Theory of Max Weber*. Stanford University Press, Stanford, CA
- Schumpeter J A 1943 *Capitalism, Socialism and Democracy*. Allen & Unwin, London
- Shklar J N 1998 *Political Thought and Political Thinkers*. University of Chicago Press, Chicago
- Skinner Q 1978 *The Foundations of Modern Political Thought*, 2 vols. Cambridge University Press, Cambridge, UK
- Skinner Q 1996 *Reason and Rhetoric in the Philosophy of Hobbes*. Cambridge University Press, Cambridge, UK
- Talmon J L 1960 *The Origins of Totalitarian Democracy*. Praeger, New York
- Tocqueville A de 1964 [1850] *Oeuvres Complètes, Vol. XII, Souvenirs*. Gallimard, Paris

B. Wittrock

Political Trials

The term 'political trial' is popularly associated with 'show trials' like the Moscow Purge trials that were conducted by Stalin during the 1930s. Such trials represent an extreme example of political repression and are viewed as sham legal proceedings designed by the authorities to dramatize specific political campaigns and/or to eliminate prominent individuals. In this context the political trial is taken to be a 'show' in which politicians and not jurists are those who pull the strings and determine the results of the trial in advance. This pejorative use of the term also helps designate regimes that resort to such techniques as the antithesis

of liberal democracies. Yet, as Otto Kirchheimer observed, the most effective usage of political trials occurs precisely within liberal democracies, in societies committed to the rule of law. What can explain this gap between popular perceptions and the actual practice of political trials? Can we distinguish a political trial from a 'show trial'? Should we continue to view political trials as opposed to the very essence of liberal democracies, or should we accept them as part and parcel of this system, deserving a sustained theoretical elaboration?

In his seminal book *Political Justice* Kirchheimer defines political trials as instances in which 'court action is called upon to exert influence on the distribution of political power' (Kirchheimer 1961). This broad definition includes the use of the court by the ruling elites or by outgroups and dissenters to achieve political goals by judicial means. Many writers regard the central factor in such trials as the political motive behind them. The connecting thread between the trial of Socrates for corrupting the youth of Athens (399 BC), the trial of Jesus for blasphemy and sedition (AD 30), the trial of Joan of Arc (1431) for heresy and witchcraft, the trial of Thomas More (1534) for remaining silent when asked about Henry VIII's supremacy in religion, the trial of Galileo (1633) for heretically suggesting that the earth moved around the sun, and many other trials is that, in each case, men in power believed that the defendant was a threat to them (Becker 1971, Belknap 1981). For this reason the prototype of a political trial has been the criminal trial of a political adversary for political reasons. This broad definition encompasses trials in which the defendant has directly attacked the established order by acts such as treason or sedition, but also trials involving common crimes in which only the personality or the motive of the offender indicates their political nature. Other writers try to refine this definition by expanding its scope to legal proceedings other than criminal trials such as impeachment, congressional hearings, and truth commissions. It has also been suggested that the definition be extended to include trials involving larger group conflicts such as ethnic, race, or labor struggles. Some move the emphasis from the political motive of the authorities to the political identity of the defendant. Yet others conclude that the important factor is not the motive but the competition of stories over the identity of the polity that distinguishes such trials. Notwithstanding their differences, all writers on the subject agree that political trials are inevitable not only in authoritarian regimes but also in constitutional democracies, although this is generally denied. Critical of such denials, they set out to replace them with an honest recognition of the existence and function of political trials in both types of regimes.

Why should the term political trial cause such anxiety and widespread denial? Since law tends to stabilize the status quo it is not surprising that there

are always people who feel justified in ignoring the law in order to bring about drastic and necessary changes. This is especially true for powerless groups that are prevented from participating effectively in important institutional decisions affecting their lives (Hakman 1972, Fletcher 1995). When the authorities put on trial those who challenge the basic value distribution in society, especially those who are willing to disregard existing laws to this end, the result is a political trial. Despite this regular occurrence of political trials within democratic societies the term is widely resisted because of its connotation with notions of 'show trials' in which either the charge is a 'frame-up' to conceal the political motivation of the prosecution, or the hearing itself fails to meet minimum standards of judicial impartiality and independence. For this reason, writers attempt to distinguish among different types of political trials. One classification distinguishes among the politically motivated trial, politically determined trial, and trials with substantial political consequences (Friedman 1970). Another differentiates between political 'trials' (abridgment of due process guarantees) the 'political' trial (camouflaged as a criminal trial) and the political trial (politically motivated) (Becker 1971). These efforts at classification have proved frustrating because of the immense historical variations and the widespread existence of political trials in very different regimes. Writers have identified a variety of characteristics of political trials but none has proved to be a sufficient and necessary condition. This literature has largely been descriptive but has failed to assess the compatibility of political trials with liberal values and thus to overcome the negative connotations the term arouses. In order to address this question it is necessary to expose the hidden engine of political trials—their legitimizing function—and its relation to democracy.

The legal system offers a tempting opportunity for those in power to damage enemies, tarnish their image, and isolate them from potential allies by casting them as criminals. This temptation becomes especially strong in democracies that constrain the ruling elite from resorting to more direct devices of repression. What distinguishes political trials from other political devices is their legitimizing function, i.e., their ability to turn a political adversary into a criminal defendant and thereby reinforce the legitimacy of those in power. However, this function can be served only if the public perceives the trial as providing the defendant with a 'fighting chance.' In constitutional democracies this 'fighting chance' is generally assured by the relative autonomy of the judges and the constraining effects of the rule of law (Kirchheimer 1961). Moreover, these guarantees introduce an element of uncertainty into the trial, and encourage political defendants to use it to expose the illegitimacy of the policy or value system advanced by the authorities. Indeed, from Socrates's trial (399 BC) to the Chicago Conspiracy trial (1969), there have always been defendants who have exploited

the limited forum that the political trial provided them to articulate their oppositional views. Some have adopted a defiant stance, proclaiming their right to commit the crime (Solomon Tellirian—as a protest against the Armenian massacre by the Turks in 1921; Dr Jack Kevorkian—advocating euthanasia in 1999), or refusing to respect the traditional decorum of the courtroom altogether (Bobby Seale–Black Panthers in 1969; Klaus Barbie–Nazism and Vichy France in 1987); others have played by the rules in order to win an acquittal that would confirm their claim that they were being prosecuted for their political beliefs (Dr Spock–Vietnam War in 1968). Insight into this legitimizing function and the element of uncertainty that it introduces into political trials in democratic regimes exposes the tension between politics and justice inherent in them. This has made political trials a fascinating area for research since the 1960s.

1. Political Trials in the Transition to Democracy

Political trials are most salient in times of transition between regimes, especially when a new democratic regime confronts crimes of the old regime. Indeed, the first attempts to seriously consider the compatibility of political trials with liberal–democratic values appear in the literature on the transition to democracy. Unlike a military revolution that sustains its authority by brute force, democratic regimes are committed to the rule of law and are inclined to address the evils of the previous regime with the help of legal devices. However, the new regime's commitment to the rule of law also makes it aware of the dangers of using *ex post facto* laws and indulging in 'victor's justice.' At such times, the various expectations from the law—to punish the guilty, to ascertain the truth about the old regime, and to enhance reconciliation in society—seem to overwhelm the legal system and to push it in opposite directions. Moreover, the forward-looking direction of the architects of democracy who are concerned with the efficacy of the transition often conflicts with the backward-looking direction of legal proceedings and their narrow focus on individual guilt (Osiel 1997, Nino 1991, Teitel 1997). For these reasons trials of transition bring to the foreground the clash between politics and justice. Two main approaches to the problem have evolved since World War II: Exemplary Criminal Trials and Truth Commissions, both of which can be considered political trials of sorts and have received extensive theoretical elaboration.

Following the end of World War II the Allies established an international military tribunal in Nuremberg (1945) to judge the leaders of the Nazi regime. This was the first time in history that the leaders of a defeated country had faced criminal prosecution for war crimes, crimes against the peace, and crimes against humanity by an international tribunal. The charge that was applied in the trial—conspiring to

wage an aggressive war—allowed the prosecution to tell a general historical narrative. Could the obvious political end of using the court to teach a history lesson (through *ex post facto* laws, and victors' judges) be reconciled with the demands of liberalism? Confronted with this dilemma, Judith Shklar made the first serious effort to reconcile the legacy of liberalism with political trials occurring in the transition to democracy (Shklar 1964). A legalistic understanding of liberalism, she argued, could justify the Nuremberg trials only at the high cost of denying their political nature altogether. In her view, procedural safeguards to defendants could only guarantee that the trial would not deteriorate into a show trial, but this in itself could not justify deviating from basic liberal demands such as an independent judiciary, established court, nonretroactive laws etc. Paradoxically, Shklar came to the conclusion that the political dictate of founding a democracy justified the divergence from strict legalism in the Nuremberg trials. In other words, it was only an honest recognition of the political aspects of the trials, their educational and symbolic contribution to the building of democracy in postwar Germany, that could justify them. Later writers accepted Shklar's consequentialist approach to transitional justice but limited its applicability to trials that guarantee the due-process rights of the defendants (Osiel 1997).

A very different attempt to address the crimes of the Nazi regime was made in the Eichmann trial (1961). This time it was not the international community but an Israeli court that undertook to judge one of the central figures in the administrative massacre of the Jews. The political nature of the trial was evident, among other things, from its reliance on the particularistic law of 'crimes against the Jewish people.' Likewise the heavy reliance on victims' testimonies about their suffering (and through them the telling of the Jewish Holocaust) conflicted with the liberal demand that the trial concentrate on proving the deeds of the accused. Hannah Arendt and other writers criticized the Israeli prosecution for politicizing the trial in this way (Arendt 1963, Segev 1993, Lahav 1992). Others see the decision as stemming from the general dilemma of political trials of how to strike a balance between the need to establish the truth about an oppressive regime, the need to give voice to the victims, and the need to ascertain the guilt of the defendants (Bilsky 2001).

This dilemma is compounded when the former authoritarian regime retains enclaves of power and opposes any attempt to bring to justice the officials of the former regime. In South America, Argentina remains the only nation that has indicted and prosecuted military officers in public trials for their role in acts of repression carried out under the former regime. The political nature of these trials was evident in the decision to prosecute only the masterminds and the worst offenders, and in the greater priority given to the value of learning about the evils of the authoritarian

past than to prosecuting every individual who was involved. Again, the justification for such selective prosecution is given in consequentialist terms—to what extent the trial serves to strengthen the emerging democracy (Nino 1991). A central dilemma that emerges in transitional trials of this kind is the breakdown of common discourse which undermines the legitimacy of the trial for some segments of society. Thus, the Argentinian court's decision to limit the evidence to proving the deeds of the accused, without allowing them to be situated within the wider historical narrative about an alleged war of self-defense against attacks on the regime by guerrilla forces, was deemed political and undermined the legitimacy of the trial in the eyes of the military (Osiel 1997).

Side by side with such efforts to bring to trial leaders of former authoritarian regimes for gross violations of human rights, there has developed the institution of truth commissions established by national or international organizations. Truth commissions reduce the tensions that arise in transitional trials by separating the various functions of learning the truth about the past, of ascertaining individual guilt, and of giving voice to victims. The Truth and Reconciliation Commission in South Africa, for example, was divided into a Committee on Human Rights Violations (victims' testimonies), a Committee on Amnesty (offender's testimonies), and a Committee on Reparation and Rehabilitation. In each committee different procedures were used (1995 Act). It further made the granting of individual amnesty conditional upon providing a full account of the offenses that were committed (Minow 1998). Truth commissions have proved to be far more effective than court proceedings in furnishing a dramatic medium for theatricalizing the new official history since the narratives of victims are rarely interrupted by lawyers and there are incentives for offenders to relate their part in the repression. Indeed, in several countries the commission's proceedings have been broadcast daily, and in others the final report, *Nunca Mais* (Never Again), has become a best seller (Brazil, Argentina.) Truth commissions present us with a kind of political trial in which the balance between politics and justice has been struck in ways more compatible with liberal concerns.

2. *Political Trials in Established Democracies*

The political nature of transitional justice seems from the fact that there are no overarching legal norms that are accepted as legitimate by the two successive regimes. Moreover, in these cases the trials have overwhelmingly been used to fulfill social and political functions other than ascertaining individual guilt. Liberal scholars, as we have seen, have been willing to recognize the political nature of such trials and to

justify it, but have not extended their theories to trials in established democracies. The general view is that in normal times law should endeavor to keep politics out of the courtrooms so that every political trial is perceived as a corruption of the rule of law (which may be termed the pathology thesis). This distinction, however, fails to account for the striking similarities between transitional situations and periods in which there has been a significant increase in political trials in democratic societies.

In the United States political trials have occurred whenever the status quo has been challenged, generally during periods of social and political ferment unleashed by such forces as war, economic conflict, or racial discord (Belknap 1981). For example, the decade that began with the escalation of the Vietnam war in 1965 brought an epidemic of political trials culminating in the Chicago Conspiracy trial (1969–70). This case represents a microcosm of American political justice of the era, as it was directed against representatives of major antiwar groups, youth counter-culture, and the Black Panther Party who all believed that the war was illegal and sought to mobilize the public against it. Although the charges were clearly political (conspiracy to cross state lines with the intent to incite riot), the trial also demonstrated the ability of the defense to politicize the trial and turn it into a forum for social protest by flouting the norms of the courtroom and ridiculing the judge (Ely 1981). In such trials as the Chicago Conspiracy trial the same breakdown of common discourse was manifested as in political trials during the transition to democracy.

The social turmoil and political trials of the late 1960s influenced the development of radical theories of law (Gordon 1982) scholars in critical legal studies, critical race theory; and feminist legal studies began to question the plausibility of the liberal ideal of separating law from politics. For these scholars the very attempt to delimit the boundaries of the category 'political trials' obscures the way in which politics enters every trial and every field of law. They argue that the liberal reliance on a distinct category of political trials often serves to legitimize the status quo. Critical theory understands politics not in the narrow sense of the term (motive of authorities, identity of defendant, etc.) but as the hegemonic ideology that shapes the interpretation of law while presenting it as neutral. The main contribution of critical writings to understanding political trials has been their rejection of the 'pathology thesis' about the relation between law and politics and their systematic efforts to uncover the ideological structures that shape different areas of law (Kelman 1987, Kairys 1982). This broad definition of political trials, however, had its drawbacks because it no longer called for a systematic investigation of the unique features of classic political trials, and shifted the attention of scholars from the court drama and social reception of the trial to the appellate court interpretation of the law.

Other scholars, still committed to the liberal framework, have identified a new form of political trial that emerged during the 1980s and 1990s. While in the old political trial the ruling authorities selected certain individuals to stand for an opposition the state wanted to eliminate, in the new political trial a section of the public turns the trial into a political trial by identifying with the victims (who are not a formal party to the trial) or with the defendant. In these cases the state does not intend a political trial and has very little control over its politicization (Fletcher 1995). The people mobilized around these trials are usually outgroups protesting their marginalization by the legal and political institutions of the state. Examples are famous rape trials and self-defense trials of battered women who killed their abusive spouses in which women's group identify with the victim and politicize the trial. The Rodney King trial and O. J. Simpson trial likewise became political trials in which African-American groups were mobilized to protest against white justice in America. The study of these trials draws particular attention to the role of the media in these cases.

The focus on group conflict and on the narrative and rhetorical aspects of political trials in on-going democracies provides scholars with the key to understanding their dynamics from a pluralistic perspective. The famous trials of Socrates, Jesus, Dreyfus, and others are examples of heroes unjustly prosecuted, but they can also be viewed as major junctions in the life of the republic where society's conflicting values are played out not only, and not even mainly, through the learned interpretation of the law, but through the human drama in and around the courtroom. In these trials the social conflict is transformed into competing narratives that capture the public's attention and offer an opportunity for collective self-reflection. Through an examination of competing values and loyalties they bring together for public consideration society's basic contradictions (Christenson 1999). From this perspective we can see how political trials, while threatening the rule of law, at the same time may contribute to the development of a more critical and democratic society.

See also: Political Lawyering; Political Protest and Civil Disobedience; Trials: Cultural

Bibliography

- Arendt H 1963, 1964 *Eichmann in Jerusalem, A Report on the Banality of Evil*. Penguin Books, New York
- Barrett W P 1931 *The Trial of Jeanne d'Arc*. George Routledge & Sons, London
- Becker T L (ed.) 1971 *Political Trials*. Bobbs-Merrill, New York
- Belknap M R 1981 *American Political Trials*. Greenwood Press, Westport, CT
- Bilsky L 1999 *Political Trials: A History of Misrecognition*. Unpublished manuscript
- Bilsky L 2001 Between justice and politics: The competition of storytellers in the Eichmann trial. In: Aschheim S (ed.) *Hannah Arendt in Jerusalem*. University of California Press
- Boyle J (ed.) 1994 *Critical Legal Studies*. New York University Press, New York
- Christenson R 1999 *Political Trials, Gordian Knots in the Law*. Transaction Publishers, New Brunswick, Canada
- Conot R 1983 *Justice at Nuremberg*. Harper & Row, New York
- Cover R M 1983 The Supreme Court, 1982 Term—Foreword: nomos and narrative. *Harvard Law Review* 97: 4–70
- De Brito A B 1997 *Human Rights and Democratization in Latin America: Uruguay and Chile*. Oxford University Press, Oxford, UK
- Dorsen N, Friedman L (eds.) 1974 *Disorder in the Court, Report of the Association of the Bar of City of New York, Special Committee on Courtroom Conduct*. Pantheon Books, New York
- Du Plessis L 1994 Amnesty and transition in South Africa. In: Boraine A, Levy J, Scheffer R (eds.) *Dealing with the Past: Truth and Reconciliation in South Africa*. IDASA, Capetown, South Africa
- Ely J W 1981 The Chicago conspiracy case. In: Belknap M R (ed.) *American Political Trials*. Greenwood Press, Westport, CT
- Fletcher G F 1995 *With Justice for Some, Victim's Rights in Criminal Trials*. Addison-Wesley, Reading, UK
- Friedman L 1970 Political power and legal legitimacy: A short history of political trials. *Antioch Review* 30: 157–71
- Gagne P 1998 *Battered Women's Justice*. Twayne, New York
- Gordon R W 1982 Some critical theories of law and their critics. In: Kairys (ed.) *The Politics of Law*. Basics Books, New York
- Hakman N 1972 Political trials in the legal order: A political scientist's perspective. *Journal of Public Law* 21: 73–126
- Hariman R (ed.) 1990 *Popular Trials, Rhetoric, Mass Media and the Law*. The University of Alabama Press, Tuscaloosa, AL
- Harris W R 1954 *Tyranny on Trial, The Trial of the Major German War Criminals at the End of World War 2 at Nuremberg, Germany, 1945–1946*. Southern Methodist University Press, Dallas, TX
- Hesse C, Post R (eds.) 1999 *Human Rights in Political Transition*. Zone Books, New York
- Kairys D (ed.) 1982 *The Politics of Law, A Progressive Critique*. Basic Books, New York
- Kelman M 1987 *A Guide to Critical Legal Studies*. Harvard University Press, Cambridge
- Kirchheimer O 1955 Politics and justice. *Social Research* 22: 377–398
- Kirchheimer O 1961 *Political Justice, The Use of Legal Procedure for Political Ends*. Princeton University Press, Princeton, NJ
- Lahav P 1992 The Eichmann trial, the Jewish question and the American-Jewish intelligensia. *Boston University Law Review* 72: 555–575
- MacKinnon C A 1989 *Toward A Feminist Theory of the State*. Harvard University Press, Cambridge, MA
- Minow M 1998 *Between Vengeance and Forgiveness, Facing History after Genocide and Mass Violence*. Beacon Press, Boston
- Nino C S 1991 The duty to punish past abuses of human rights put into context: The case of Argentina. *Yale Law Journal* 100: 2619–640
- Nino C S 1996 *Radical Evil on Trial*. Yale University Press, New Haven
- Orentlicher D F 1991 Settling accounts: The duty to prosecute human rights violations of a prior regime. *Yale Law Journal* 100: 2537–615

- Osiel M 1997 *Mass Atrocity, Collective Memory and the Law*. Transaction Publishers, New Brunswick, Canada
- Segev T 1993 *The Seventh Million: The Israelis and the Holocaust*. Hill and Wang, New York
- Shklar J N 1964 *Legalism, Law, Morals and Political Trials*. Harvard University Press, Cambridge, MA
- Stone I F 1989 *The Trial of Socrates*. Doubleday, New York
- Teitel R 1997 Transitional jurisprudence: The role of law in political transformation. *The Yale Law Journal* **106**: 2009–80
- Zalaquett J 1988 Confronting human rights violations by former governments: Principals applicable and political constraints. In: Henkin A (ed.) *State Crimes: Punishment or Pardon*. Aspen Institute, Queenstown, MD, 23–71

L. Bilsky

Politics, Formal Modeling of

Positive theories of politics provide explanations for why political phenomena occur in the manner that they do. Examples of such phenomena include which parties or candidates are elected at certain times, what policies legislative bodies adopt, and when and how wars are fought between countries. Formal models complement these explanations by using the language and analysis of mathematics to derive the phenomena as necessary consequences of certain sets of underlying factors. As with mathematical models in other disciplines, the main advantage of such an approach is twofold: to give an unambiguous representation of the relevant moving parts of a theory (the clarity of the argument) and to enforce a logical coherence on a theory (the consistency of the argument). The main drawback is found in the required simplification of the phenomena down to a manageable number of moving parts, and the subsequent possibility of simplifying away the very essence of the problem.

1. The Basic Structure

Most formal models of politics fall in the category of rational choice (e.g., see *Rational Choice in Politics*), where the presumption is that observed outcomes are the result of decisions made by a certain relevant set of individuals. These could be voters and candidates in the first example above, elected representatives and appointed ministers in the second, or heads of state in the third. The general behavioral postulate is that individuals attempt to influence the outcome in a manner reflecting their preferences (tastes, values, etc.) over the set of possible outcomes. For instance, in selecting campaign platforms, a candidate who only cares about winning the current election might look to adopt policies appealing to the greatest number of voters; conversely, a candidate who cares intrinsically about the policies to be implemented might shy away

from politically popular but personally unpalatable alternatives. Whatever their specific form, individuals' preferences are taken to be primitives of the model, and so are held fixed throughout the analytic determination of an outcome. However, different specifications of preferences can give rise to different outcomes. Indeed, to the extent the predicted outcome varies with changes in the underlying individual preferences, one has a potential explanation for any variation in observed outcomes as being due to variation in these preferences. The extent to which these preferences can themselves be modeled as functions of observable parameters, such as wealth or legislative seniority, permits one to generate a theoretical relationship between these parameters and observable political outcomes as well.

The influence of preferences on outcomes occurs through the choices the individuals make. The most common assumption is that each individual's choice is optimal with respect to their preferences, for example, in the above scenario the candidate selects a vote-maximizing policy. This optimality assumption may be far from a realistic representation of individual decision-making, implying as it does that individuals can effortlessly solve what are at times relatively difficult mathematical problems. However, assuming otherwise typically requires the presence of auxiliary factors (for instance, decision costs) to rationalize *non*-optimizing behavior, factors which, while reasonable, tend to create additional complications in the model. Also the optimality assumption allows one to import existing results from optimization theory to aid in the analysis.

As seen in the examples above, most formal models of politics involve more than one individual, and so this individual level of optimization often takes place within the confines of non-cooperative game theory. Such models explicitly describe the interaction among the individuals that ultimately produces a collective decision. This description begins with a 'game form' consisting of (a) the set of strategies available to each individual, and (b) the specification of which outcome occurs when a given profile of strategies is chosen (with each profile consisting of one strategy per individual). Continuing with the electoral setting, one example of a game form is where the set of strategies for a candidate is equal to the set of available policies, a voter's strategy describes which candidate to vote for based on the candidates' chosen policies, and an outcome, consisting of who wins the election and with what policy, is determined by the candidate receiving the most votes. The remaining piece of the puzzle then is the derivation of the individuals' strategies, with the fundamental concept being that of the 'Nash equilibrium.' Note that, even assuming all individuals have well-defined preferences over the set of outcomes, the notion of optimality in this multiperson environment is necessarily conditional, as an individual's best choice of strategy may depend nontrivially on the strategy

choices by the other individuals. For instance, in the electoral game above, the location of a candidate's optimal policy may depend on the voters' behavior as well as the policies endorsed by rivals. A Nash equilibrium is a profile of strategies wherein each individual's strategy is optimal, given the strategies of the others, that is, conditional on the choices by the other individuals, no one individual has an incentive to alter her choice. A Nash equilibrium outcome is thus a collective decision consistent with individually optimal behavior and constitutes the basic object of analysis in any noncooperative model. (In certain game forms, some Nash equilibria are inherently more plausible than others and so refinements of Nash equilibria, such as subgame perfection, are used.)

As with the choice of preferences, the choice of game form is in the hands of the modeler. This choice may be influenced by observable features of the political phenomenon in question, as in the above electoral example. In such situations, there will invariably be a trade-off between making the game form descriptively accurate and keeping the game form analytically tractable. Thus while the electoral rule may easily be incorporated, simplifying the campaign stage of the process to a single choice of policy, thereby avoiding the complexities found in actual elections, may be necessary to guarantee the existence of and to solve for equilibrium outcomes. At other times, there may not exist any such observable features to leverage, an example of which would be when two heads of state are bargaining over land and other resources in the shadow of war. Here one would be hard-pressed to assert that the unstructured negotiations taking place can be definitively captured in any one particular game form. Hence descriptive accuracy is not an issue, and the modeler is relatively unconstrained in the search for a 'good' game form, that is, a game form capable of generating new and valuable insights into the political phenomenon in question.

Yet to the extent that the game form incorporates observable features of the political process and that the equilibrium outcomes vary with these features, one has a second source of variation in outcomes above and beyond changes in preferences, namely, changes in the political process. For instance, the equilibrium outcomes under the above plurality electoral rule may differ from those associated with a run-off rule (where, if no candidate receives a majority of the votes, the two candidates receiving the highest number of votes are paired in a second election, in which majority rule determines the winner) even if the candidates' and voters' preferences are held fixed. This allows one to make statements about the degree to which differences in observed collective choices are due to differences in preferences, or to differences in political institutions. More ambitiously, this type of comparative institutional analysis can be used to predict the effect changes in existing political institutions will have on collective outcomes.

2. An Alternative Approach

While noncooperative game theory has tended to dominate the modeling landscape in recent years, much of the early formal work employed the techniques of cooperative game theory. Rather than explicitly modeling outcomes as the aggregation of individual decisions, under this approach one looks for outcomes that are stable in some sense, with this 'sense' implicitly characterizing the (unmodeled) underlying strategic interaction. Given the normative appeal and frequent use of majoritarian-based political procedures, the focus was primarily on identifying outcomes which are majority-preferred to all others in pair-wise comparisons. Such an outcome would then be immune to changes proposed by any majority coalition, whereas all other outcomes would not.

It turns out that a majority-stable outcome necessarily exists in certain environments, the most compelling being when the collective choice concerns the selection of a single number. For example, suppose the issue before a legislature is the size of the nation's military budget, the set of possible outcomes can be taken to be the set of nonnegative real numbers. Suppose further that the legislators have 'single-peaked' preferences: each has some best number, and the value to them of the outcome falls as one moves away from this best number. Then a stable outcome will exist and, if the number of individuals is odd, will be unique and equal to the median of these best numbers. This 'median voter theorem' has been at the core of numerous and varied empirical studies of collective choice, from local school budgets in the United States to European Union decision-making. Its comparative advantage as a formal model is obviously its parsimony, requiring as it does only a description of how individual preferences over outcomes are related to any underlying parameters of interest (e.g., household income in the local school budget case). Further, by imposing additional assumptions on preferences, one finds that the median voter's preferences coincide with those of the majority over the entire set of outcomes, implying that the median voter is in a literal sense representative of the collective as a whole. The benefits from this result accrue principally to the noncooperative modeler: suddenly a one-dimensional model of the interaction between the US President and the 435 members of the US House of Representatives over the size of the military budget, with majority rule dictating the decision-making among the latter, can be simplified to a two-person interaction between the President and the median member of the House.

The main shortcoming of the median voter theorem of course is its restriction on the set of outcomes to a single dimension. In certain situations this is benign, as in the above example when the policy choice in question consists of selecting the military budget.

However, a policy choice consisting of the distribution of this budget across more than two groups is inherently a multidimensional problem and the conclusions about majority rule for multidimensional problems, relative to their single-dimensional cousins, could hardly be more dramatic. In particular, for most descriptions of preferences, no majority-stable outcome exists (where 'most' has a mathematically precise definition in this setting). As an example of such non-existence, suppose in the above scenario that each legislator cares only about how much their favorite group receives, with one-third of the legislators associated with each of three groups. Then for any distribution of the budget, the legislators for two groups (which together constitute a majority) prefer to take away any amount going to the third group and redistribute it amongst themselves, thereby guaranteeing that no distribution is majority-stable. Furthermore, when no stable outcome exists, not only is there no representative voter but in fact one can construct majority preference cycles throughout the policy space (i.e., policy *a* is majority-preferred to policy *b*, which is majority preferred to policy *c*, which ... which is majority-preferred to policy *z*, which is majority-preferred to policy *a*). Thus when majority rule breaks down, it breaks down completely. Subsequent research has shown that these negative conclusions continue to hold, to a greater or lesser degree, for most non-majoritarian rules as well.

As with the median voter theorem, it is difficult to overstate the impact these 'chaos theorems' have had on the formal modeling of politics. Clearly, in a multidimensional world, the modeler has lost the analytic shortcut of representing a collective's decision-making as being governed by a single individual (or, at least, the logical rationale for doing so). More importantly, the theorems gave rise to a new-found appreciation for political institutions, such as committee systems, restrictive voting rules, and the like as key elements in the determination of collective outcomes. Indeed, such institutions are often depicted as necessary ingredients in any multidimensional model of politics, in that, but for such institutions, collective decision-making would fall prey to the inevitable onslaught of cycling and chaos. This interpretation, while quite popular, appears inconsistent with the central tenet of formal modeling, that of equilibrium analysis, as the chaos theorems are properly seen as nonequilibrium results. Put differently, these theorems do not predict that (as commonly phrased) 'anything can happen,' for as nonequilibrium results they do not predict anything at all. Rather, they demonstrate the impossibility of any general theory of collective behavior based solely on the (majority) aggregation of preferences, and hence the necessity of being more explicit about the decision-making process so as to provide the analytical traction required to formulate a well-posed model. From this perspective, the lesson to be learned from the chaos theorems is one for the modeler of collective

decision-making, rather than as one about collective decision-making itself.

See also: Action, Collective; Action, Theories of Social; Behavioralism: Political; Bounded Rationality; Game Theory: Noncooperative Games; Model Testing and Selection, Theory of; Systems Modeling

Bibliography

- Austen-Smith D, Banks J S 1999 *Positive Political Theory I: Collective Preference*. University of Michigan Press, Ann Arbor, MI
- Morrow J D 1994 *Game Theory for Political Scientists*. Princeton University Press, Princeton, NJ
- Ordeshook P C 1986 *Game Theory and Political Theory: An Introduction*. Cambridge University Press, Cambridge, UK
- Shepsle K A, Bonchek M S 1997 *Analyzing Politics: Rationality, Behavior, and Institutions*. Norton, New York

J. S. Banks

Polling

Opinion polls or surveys are used widely to study the behavior of individuals. Survey researchers select and ask questions concerning individuals' attitudes, behavior, personal characteristics, or other types of information that they can provide as respondents. These responses are counted, analyzed statistically, and evaluated or interpreted in different ways, depending on the types of questions asked.

Academicians and practitioners engaged in opinion polling have preferred to call themselves 'survey researchers,' in contrast to 'pollsters,' who carry out nonacademic polls to predict election results, or report the mass public's or others' opinions and attitudes in the news media, or who work as consultants doing proprietary research, such as market research, for political candidates, political parties (including polls for their presidents), or other clients (Converse 1987, Moore 1992, Jacobs and Shapiro 1995). Polls and surveys are synonymous, but their quality (validity and reliability) can vary widely.

1. 'Straw Polls'

The use of polling had expanded greatly worldwide by the late twentieth century. Its origin and impetus occurred in the United States beginning with election 'straw polls' in the nineteenth century, and developing into modern-day political polling and academic surveys used widely to study social, economic, and political phenomena. Polling was arguably a major part of the rise of quantification in society consistent with Max Weber's writing about the role of calculation and rationality in social life (Herbst 1993). The earliest straw poll counterparts of modern opinion surveys

occurred in the 1824 presidential election, and the use of different types of these in politics and journalism increased until the emergence of more consciously scientific polling in the 1930s (Smith 1990).

These polls took many different forms, including oral counts, handraising, paper ballots, and writing preferences in 'poll books.' The participants were 'self-selected' in ways that were likely to produce a problem known as 'selection bias,' in which the results of these polls would be influenced by the characteristics of those who were able to participate, which might be different from the electorate or population to which the polls' results would be generalized. Reporters and partisans nonetheless treated these polls as indications of candidates' election chances.

2. 'Scientific Polls'

The problems associated with the overall representativeness and accuracy of such polls became clear publicly, when the accuracy of the *Literary Digest's* paper ballot poll, which had predicted the winners in US presidential elections from 1920 to 1932, was called into question by George Gallup and other new and more scientific-minded pollsters, statisticians, and social scientists. Gallup was the most widely known, if not the most important figure (along with Elmo Roper, Archibald Crossley, Paul F. Lazarsfeld, and Hadley Cantril) in 'scientific polling,' and approached polling with a democratic philosophy (Gallup and Rae 1940). While the first national Gallup election poll, done through what became standard in-person interviews did not predict the margin of President Franklin Roosevelt's 1936 election victory over Alfred Landon accurately, it triumphed over the *Literary Digest* in predicting the winner.

The Gallup poll represented the actual electorate far better than did the *Digest*, confirming the superiority of scientific sampling methods. In principle, initially, these methods allowed each potentially eligible adult voter (or eligible respondent, given the research objectives of a particular poll) an equal probability of selection, which was not the case in the *Digest* or other straw polls that allowed self-selection (Squire 1988). The early scientific polls involved sending interviewers to randomly-selected locations and then allowed interviewers largely to select their respondents to fill 'quotas' based on population characteristics to be represented in the poll. This led to selection biases.

These and other problems in the early polls, including ceasing to interview weeks before an election, were debated widely among pollsters and social scientists when the polls predicted incorrectly that Thomas Dewey would defeat President Harry Truman in the 1948 presidential election (Mosteller et al. 1949). Gallup and other pollsters improved their sampling methods, and the pollsters and academic researchers,

who were beginning to engage in large-scale public opinion and voting studies at this time, engaged in a concerted and continuing effort to improve and defend their research (see, e.g., Sheatsley and Mitofsky 1992, Mitofsky 1998).

3. Academic Research

University-based survey research had its beginnings in 1939–40 with sociologist Paul F. Lazarsfeld's Office of Radio Research, later renamed the Bureau of Applied Social Research. This history is described by Rossi (1959), Natchez (1985), and especially Converse (1987). Lazarsfeld and his colleagues in *The People's Choice* (Lazarsfeld et al. 1944) did the first sophisticated survey study of presidential elections in 1940 in Erie County, Ohio. This was followed by *Voting* (Berelson et al. 1954), a study of the 1948 election in Elmira, New York. Two of the largest and most widely known research centers that developed the capacity to do long-term national surveys are the National Opinion Research Center (NORC), which was established in 1941 at the University of Denver and moved to the University of Chicago in 1947, and the University of Michigan's Survey Research Center (SRC, Institute for Social Research (ISR)), established in 1946. With the publication of *The Voter Decides* (Campbell et al. 1954), the funding and locus for national election surveys shifted to Michigan's SRC, which led to the publication of *The American Voter* (Campbell et al. 1960) and the continuing series of American National Election Studies and studies based upon them (Nie et al. 1976/1979, Miller and Shanks 1996).

4. Further Expansion

Political polling and market and academic survey research increased after 1948. The early and subsequent public pollsters who provided their survey results to the mass media and other subscribers (for example, Louis Harris and Daniel Yankelovich, among others) had used their political polling as advertising to generate business for their proprietary market and other types of research. Market researchers and other pollsters expanded their work into polling and consulting for political parties and candidates (Moore 1992). Polling experienced its greatest expansion beginning in the 1970s, when it was found that sufficiently representative and accurate polling could be done by telephone (based on experiments comparing polls conducted by phone with those done in-person), eventually leading all the major pollsters to do most of their polling by phone. The most important national news media, which had relied exclusively on polling by public pollsters, began sponsoring and conducting their own polls, establishing their own,

often joint, polling operations or making arrangements to contract-out polls in which they controlled the content and reporting of results (Gollin 1980, Mann and Orren 1992).

5. *Methodology and Technology*

While social scientific polling depends on an understanding and use of scientific sampling to draw inferences about a target population (e.g., national or local, the entire adult population or particular sub-populations, ordinary citizens or elites) polls can be conducted in different ways depending on the research goal, the kind of survey questionnaire to be administered, and the available resources (money and time). In-person national polling (typically randomly sampling geographic locations, then households in these locations, and then random sampling within households) is the most expensive and time-consuming because of travel costs, though these can involve longer surveys and the use of visual aids and secret ballots for responses to sensitive questions. In contrast, election 'exit polls' that are used to project election results and provide the basis for more in-depth statistical analysis of voters' behavior, are short and efficient surveys, based on random samples of voters exiting voting places. Exit polling has occasionally been controversial due to the 'early' reporting of results that might affect the behavior of those who vote later in the day.

Telephone polling (typically using Random Digit Dialing, RDD) is less costly, as it eliminates travel costs and is easier to monitor and validate than in-person interviews; however, it excludes households without telephones, and it has higher rates of nonresponse due to households that are not reached (no answers, busy signals, answering machines, and refusals to participate); it is easier to make contact with respondents and establish rapport through in-person contact. Mail surveys are the least expensive but produce the lowest response rates unless extensive follow-ups are made (Dillman 2000). Low response rates, with their increasing potential for selection or non-response bias, has emerged as the most pressing professional and methodological problem currently facing survey researchers as the result of the proliferation of telephone surveys, telemarketing, and solicitations that interrupt and impose upon people's time (Groves and Couper 1998).

Problems of survey costs and response rates have led surveyors to turn to better technology. The use of computerized RDD and computer-assisted telephone interviewing (CATI) was a major development in improving surveys (making complex surveys, including experimental designs, easier to administer, and eliminating interviewer and data processing errors) and expanding their use. The use of computer-assisted

personal interviewing (CAPI) followed, including the use computers for self-administered surveys allowing for greater assurance of confidentiality for sensitive questions. Recently, there have been increasing efforts to expand sampling and polling through the Internet. While the Internet provides a way of increasing response rates through ease of contact without disruption, for potential respondents whose electronic addresses have been obtained, and through ease in administering surveys, the samples obtained may not be sufficiently representative of the target population, including those without Internet access, or who refuse to make their electronic addresses available for polling (thus echoing the limitation of the straw polls).

6. *Sources of Error*

Beyond statistical sampling error and nonresponse bias, there are other sources of error in surveys that are not easily quantified. The responses to survey questions and measurement of opinions and behavior can be affected by how questions are worded, what response categories are offered (e.g., whether a middle category or a don't know or no opinion response is offered or allowed), and whether questions should have fixed or forced choices, or whether questions should be asked and responses recorded as open-ended questions in which respondents can offer any answer they choose. Further, there may be 'context effects' that are produced by the order in which questions are asked, so that the way respondents answer questions may be influenced by what they had thought about and responded to in previous questions (Schuman and Presser 1981, Asher 1998).

A major source of error can occur depending on how research problems are formulated or specified, as when researchers assume respondents have familiarity with the object about which they are asked to offer an opinion or response. Care needs to be exercised in drawing inferences about actual behavior from survey measures of opinion, and self-reports of future—or even past—behavior. Reported voting in an election after the fact may not match actual behavior based upon data on 'validated' votes. Difficulties may also arise in how poll results are reported, such as when journalists or researchers themselves do not report enough information about their survey data and methods to allow their audiences or readers to evaluate their results (Cantril 1991, Traugott and Lavrakas 2000).

7. *Contributions to Knowledge*

The uses of surveys, how to improve polling methods, and how survey researchers should practice and report about their craft are matters that have been the

concern of the professional associations and journals established by those actively involved in the field of survey research. These include the American Association for Public Opinion Research and its journal, *Public Opinion Quarterly*; the World Association for Public Opinion Research and its journal, *International Journal of Public Opinion Research*; the Council of American Survey Research Organizations; and the National Council on Public Polls. In addition, political science, sociology, and other social science journals also publish survey studies widely.

The early voting and public opinion studies by Lazarsfeld and others were highly sensitive to methodological issues in survey research, and this led to the development and use of panel surveys (the interviewing of the same respondents over two or more points in time) and complex research designs. In addition to contributing to a large body of knowledge about the influences on voting, survey research contributed from its inception to the understanding of other forms of social and political attitudes and behavior, ranging from Samuel Stouffer's studies of *The American Soldier* (Stouffer et al. 1965 [1949]) and attitudes toward communism (Stouffer 1955), to studies of long-term trends in public opinion which utilize all the available polling data since the 1930s (Page and Shapiro 1992, Mayer 1993).

While academic researchers have made substantial use of data collected by academic survey organizations (especially the NORC General Social Surveys conducted since 1972, and the American National Election Studies covering every presidential and congressional election since 1952), increasingly they have used data made available by the public and mass media pollsters. The major archives for these data are the Roper Center for Public Opinion Research (University of Connecticut), the Inter-university Consortium for Political and Social Research (ISR, University of Michigan), and the Louis Harris survey and state poll archive at the Institute for Research in Social Science (University of North Carolina). These archives and the many polling organizations nationwide have also established and expanded access to their survey data through websites on the Internet.

See also: Democracy; Opinion Formation; Political Representation; Public Opinion: Political Aspects

Bibliography

- Asher H 1998 *Polling and the Public: What Every Citizen Should Know*, 4th edn. Congressional Quarterly, Washington, DC
- Berelson B R, Lazarsfeld P F, McPhee W N 1954 *Voting: A Study of Opinion Formation in a Presidential Campaign*. University of Chicago Press, Chicago, IL
- Campbell A, Converse P E, Miller W E, Stokes D E 1960 *The American Voter*. University of Chicago Press, Chicago, IL

- Campbell A, Gurin G, Miller W E 1954 *The Voter Decides*. Row, Peterson, White Plains, NY
- Cantril A H 1991 *The Opinion Connection: Polling, Politics, and the Press*. Congressional Quarterly, Washington, DC
- Converse J M 1987 *Survey Research in the United States: Roots and Emergence, 1890–1960*. University of California Press, Berkeley, CA
- Dillman D 2000 *Mail and Internet Surveys: The Tailored Design Method*, 2nd edn. Wiley, New York
- Gallup G, Rae S F 1940 *The Pulse of Democracy*. Simon & Schuster, New York
- Gollin A 1980 Exploiting the liaison between polling and the press. *Public Opinion Quarterly* **44**: 445–61
- Groves R M, Couper M P 1998 *Nonresponse in Household Interview Surveys*. Wiley, New York
- Herbst S 1993 *Numbered Voices: How Opinion Polling has Shaped American Politics*. University of Chicago Press, Chicago, IL
- Jacobs L R, Shapiro R Y 1995 The rise of presidential polling: The Nixon White House in historical perspective. *Public Opinion Quarterly* **59**: 163–95
- Lazarsfeld P F, Berelson B, Gaudet H 1944 *The People's Choice: How the Voter Makes Up His Mind in a Presidential Campaign*. Columbia University Press, New York
- Mann T, Orren G (eds.) 1992 *Media Polls in American Politics*. Brookings Institution, Washington, DC
- Mayer W G 1993 *The Changing American Mind: How and Why American Public Opinion Changed Between 1960 and 1988*. University of Michigan Press, Ann Arbor, MI
- Miller W E, Shanks J M 1996 *The New American Voter*. Harvard University Press, Cambridge, MA
- Mitofsky W J 1998 Was 1996 a worse year for polls than 1948? *Public Opinion Quarterly* **62**: 230–49
- Moore D W 1992 *The Superpollsters: How They Measure and Manipulate Public Opinion in America*. Four Walls Eight Windows, New York
- Mosteller F, Hyman H, McCarthy P J, Marks, E S, Truman D B 1949 *The Pre-election Polls of 1948*. Social Science Research Council, New York
- Natchez P B 1985 *Images of Voting: Visions of Democracy*. Basic Books, New York
- Nie N H, Verba S, Petrocik J R 1976/1979 *The Changing American Voter* (enlarged edn.). Harvard University Press, Cambridge, MA
- Page B I, Shapiro R Y 1992 *The Rational Public: Fifty Years of Trends in Americans' Policy Preferences*. University of Chicago Press, Chicago, IL
- Rossi P H 1959 Four landmarks in voting research. In: Burdick E, Brodbeck A J (eds.) *American Voting Behavior*. The Free Press, Glencoe, IL
- Schuman H, Presser S 1981 *Questions and Answers in Attitude Surveys: Experiments in Question Form, Wording, and Context*. Academic Press, New York
- Sheatsley P B, Mitofsky W J (eds.) 1992 *A Meeting Place: The History of The American Association for Public Opinion Research*. American Association for Public Opinion Research, Ann Arbor, MI
- Smith T W 1990 The first straw: A study of the origins of election polls. *Public Opinion Quarterly* **54**: 21–36
- Squire P 1988 Why the 1936 *Literary Digest* poll failed. *Public Opinion Quarterly* **52**: 125–33
- Stouffer S A 1955 *Communism, Conformity, and Civil Liberties: A Cross-section of the Nation Speaks its Mind*. Doubleday, Garden City, NY

Stouffer S A, Suchman E A, DeVinney L C, Star S A, Williams R M Jr 1965 [1949] *The American Soldier: Adjustments During Army Life*. Wiley, New York, Vol. 1
Traugott M W, Lavrakas P J 2000 *The Voter's Guide to Election Polls*, 2nd edn. Chatham House, New York

R. Y. Shapiro

Polynesia and Micronesia: Sociocultural Aspects

The two geographical areas commonly referred to as 'Polynesia' and 'Micronesia' consist of a scattering of relatively small and widely disseminated islands across the central, south, and northwestern Pacific Ocean. The two areas are complementary to Melanesia, the third area customarily recognized in the insular Pacific. Together, the three areas are commonly referred to as the 'Pacific Islands' or 'Oceania,' although these categories, particularly the latter, occasionally include Australia and Tasmania, and occasionally the larger islands of Southeast Asia. While some geographical, social, cultural, linguistic, and historical generalizations can be made about each of them, Polynesia and Micronesia are meaningful entities beyond simple areal demarcations only as a result of historical contingencies, in which the history of Western colonialism in the Pacific figures prominently.

1. General Identification

The terms 'Polynesia' and 'Micronesia' are convenient labels for geographical areas whose social, cultural, archeological, historical, political, and linguistic significance is somewhat arbitrary. In modern-day usage, the term 'Polynesia' refers to all island and island groups falling within a large triangular area whose apexes are New Zealand to the south, Hawaii to the north, and Rapanui (Easter Island) to the west. The largest islands and island groups of the region are Tonga, Samoa, Rarotonga and the Southern Cooks, Tahiti and the Society Islands, the Marquesas, and Hawaii, in addition to the much larger New Zealand. Politically or culturally notable smaller islands include Niue, Wallis and Futuna, Tuvalu, Tokelau, the Northern Cooks, Rapanui, Pitcairn Island, and the Tuamotus, Austral, and Gambier Islands.

In addition, a geographically heterogeneous group of about 18 islands and sections of islands are commonly identified as 'Polynesian Outliers' societies because they are located outside of this triangular area: in Melanesia to the west (e.g., Takuu off Bougainville Island, Sikaiana and Rennell in the Solomon Islands, Mae and Mele in Vanuatu) and, in the case of two islands (Nukuoro and Kapinga-

marangi), in Micronesia to the northwest of the Polynesian Triangle.

Micronesia forms a wide arc of small islands spanning from the western region of Insular Southeast Asia to the Central Pacific, comprising the following island groups: Palau and outlying islands, Guam and the Mariana Islands, Caroline Islands, Marshall Islands, Gilbert Islands, and two isolated islands, Nauru and Banaba (Ocean Island). Contemporary political considerations may include island groups of the Central Pacific like the Phoenix and Line Islands; although geographically located in the Polynesian Triangle, these islands, which were not permanently inhabited until the modern age, are governed by Kiribati, the modern state based in the Gilbert Islands. With the exception of Guam, the islands of Micronesia are small, although some of the largest atolls in the world are found in the region (e.g., Kwajalein in the Marshalls).

Islands of many different types are found in Polynesia and Micronesia: a few continental islands (e.g., Guam); volcanic structures, some of which can be substantial in size (e.g., Hawaii); raised coral islands of different types (e.g., Tonga); atolls (e.g., Kiribati); and combinations of these.

2. The Historical Contingency of Labels

The societies of each of the two regions share a number of characteristics with one another, although it is impossible to arrive at a list of necessary and sufficient conditions that would determine whether a given society should be identified as Polynesian or Micronesian, or as something else. The reason for this is twofold. First, commonalities among the societies of both areas are balanced by important patterns of variation (to the extent that no feature of social organization, for example, will be found in all societies in question). Second, many societies in other areas of Oceania display sociocultural characteristics identical to those prevalent among Polynesian or Micronesian societies. Perhaps the least controversial criterion for 'Polynesian-ness' is language: all languages spoken natively in the islands are more closely related to one another than to any other language. Indeed, the reason for identifying Outlier societies as Polynesian is primarily linguistic: all Outlier communities speak languages that are most closely related to the languages spoken on islands of the Polynesian Triangle. In terms of social organization and culture, however, Outlier communities vary widely, from bearing considerable similarity to the rest of Polynesia (e.g., Tikopia) to having much more in common with their more immediate non-Polynesian neighbors (e.g., the Polynesian-speaking villages of Ouvea, Loyalty Islands).

Yet even this criterion is not devoid of caveats: on the boundary between Melanesian Fiji and Polynesian Tonga, historical linguistic evidence suggests a con-

tinuum of gradual linguistic differentiation rather than a clean break. Furthermore, linguistic factors are of little use as determiners of what Micronesia includes and what it does not: all languages spoken in the area are related to one another, but some (e.g., Chamorro of the Marianas and Palauan) are historically closer to languages spoken outside of Micronesia (e.g., in the Philippines) than to the other languages of Micronesia.

Two additional issues further complicate the problem. First, many Polynesian and Micronesian societies are increasingly diasporic, and significant communities of Samoans, Wallisians, Cook Islanders, Caroline Islanders, for example, are well established in New Zealand, Australia, the United States, and the metropolitan centers of Oceania. In some cases, emigrant communities are much more populous than the island-based communities. Second, the characterization of some areas of the region as 'Polynesian' or 'Micronesian' is historically well motivated but, from a contemporary perspective, is part and parcel of acrid political debates. Such is the case of Hawaii, New Zealand, and Guam, where the original Polynesian or Micronesian inhabitants, as well as recent migrants from other Pacific Islands, today form numerical and political minorities, albeit vocal ones in terms of identity politics. These observations highlight the problems associated with attributing human significance to the definitions of Polynesia and Micronesia.

To say that the regions are arbitrarily defined does not mean that their characterization is haphazard. Rather, the definition of sections of the globe as 'Polynesia' and 'Micronesia' is deeply embedded in a history of elaboration of certain differences and similarities and the obscuration of others to suit politically dominant agendas, of the kind that suffuses any characterization of the 'other.' In this case, this history is that of Western colonial hegemony and of the intellectual endeavors that ran alongside it. When Enlightenment-era Europeans invented it, the term 'Polynesia' ('many islands') was applied to the entire Pacific region. In the course of the nineteenth century, it became more common to restrict the term to its current referent, and to contrast it with the newly coined 'Micronesia' ('small islands') and 'Melanesia' ('black islands'). As the etymology of the latter term indicates, race and its various associations figured prominently in the newly created finer distinctions. The term 'Polynesia' came to embody (mostly positive) images of noble fair-skinned otherness already prominent in the Enlightenment and early Romantic intellect, in contrast to Melanesia in particular, which Westerners deemed to be dominated by dark-skinned and uncivilized savagery (Smith 1985, Thomas 1989).

3. Initial Human Settlement

The prehistory of Polynesia can only be understood in the context of human movements in the Pacific region since about 3500 BP. Around that date, a wave of

people appears to have traveled gradually from west to east from Southeast Asia, settling islands of Melanesia, some of which were already inhabited while others were not. The most salient archeological evidence for this conjecture is a lowly style of decorated pottery referred to as 'Lapita,' fragments of which are found in insular Melanesia and Western Polynesia, carbon-dated to 3500–2000 BP. Lapita pottery makers and users appear to have been accomplished long-distance sailors, fisherfolk, and agriculturalists, to have organized their communities in hierarchical fashion, and to have spoken languages ancestral to most languages of insular Melanesia and Polynesia.

They reached Fiji, Tonga, and Samoa around 3000 BP, an area in which, according to some researchers, migrations may have 'paused' for a while (while maintaining active contact both externally and with the West from where they had come). Such a 'pause' would have allowed a social, cultural, and linguistic distinctiveness to emerge that would eventually become what is now recognized as Polynesian distinctiveness.

From the Fiji–Western Polynesia area, early Polynesians settled the rest of Polynesia in the course of the following two millennia, finally reaching Hawaii in about 650 AD and New Zealand around 1000 AD (Bellwood 1979, Kirch 1996). There is no convincing evidence of any subsequent prehistorical human settlement in Polynesia other than the Lapita potters and their descendants, although they intermarried with their non-Lapita-making neighbors, forming as diverse a genetic pool as is found in any other part of the world.

Micronesia's prehistory is much more heterogeneous and complex than that of Polynesia, and less well understood. Evidence of human settlement in the Mariana Islands dates back to approximately 4000 BP. The archeological record indicates a Southeast Asian connection for this early population, as does relatively more recent evidence gathered on Palau and Yap. The rest of the Caroline Islands, the Marshall Islands, and the Gilbert Islands were settled by a northwestward back-migration from eastern Melanesia. Languages spoken in this region, commonly referred to as 'Nuclear Micronesian,' exhibit greater linguistic homogeneity than other Micronesian languages, suggesting a more focused history of prehistoric settlement.

4. Contacts and Movements

It is difficult to issue brief general statements about the history of Polynesia and Micronesia since colonial encroachment because of the diversity of experiences and heterogeneity of influences. European travelers began sailing through Polynesia and Micronesia early in the Age of Discovery, making contacts with Islanders that were sporadic except in Guam and the Marianas, which were situated strategically on the sea

road between the Philippines and the New World. Spain claimed the islands in the mid-sixteenth century, a move which was followed by the virtual decimation of the Islanders, through disease, slaughter, and forced resettlement.

Contacts between Westerners and Islanders in both Polynesia and Micronesia began earnestly at the end of the eighteenth century, and often had equally tragic consequences on the island populations. Enlightenment-era travelers, spurred on by complex and sometimes contradictory motivations, roamed Polynesia in particular, making contact with Islanders in a variety of fashions that ranged from the hostile to the very friendly. They were soon followed by Christian missionaries (who would eventually succeed in converting all of Polynesia and Micronesia), adventurers and traders, developers seeking to establish plantations and other large-scale ventures, whalers, and of course governments seeking to establish colonies and spheres of influence, often under the pretext of protecting their citizens. In Polynesia, the key players (big and small) in the nineteenth century were predominantly British and French, to which American and German interests added competition later in the century. In Micronesia, Spain continued to hold political and economic sway, until it was completely displaced at the end of the century by the United States and Germany.

A notable consequence of Westerners' increasing visibility in nineteenth century Polynesia was the consolidation of hitherto politically fragmented chiefdoms. In Tonga, Tahiti, and Hawaii (and to a lesser extent Samoa), ambitious and astute chieftains utilized the new tools that Westerners introduced to the islands (trading goods for firearms, protection for religious conversion) to defeat their contenders and impose their rule over entire islands and island groups, establishing themselves as sovereigns under the protection of Western powers. Of these kingdoms, only Tonga remains to this day.

By the dawn of the twentieth century, European powers had divided Polynesia and Micronesia (as well as Melanesia) into sphere of colonial domination and influence. The end of World War I somewhat reshuffled the distribution of these spheres, introducing Japan as a colonial power in Micronesia (except the Gilbert Islands, which by then had become a British colony). World War II had a profound and dramatic effect on all of Micronesia, which was the stage of some of the fiercest battles, as was Melanesia. In contrast, Polynesia (other than Pearl Harbor in Hawaii) remained in the background of the conflict.

powers, particularly Great Britain, which by then had also delegated some of its colonial authority to Australia and New Zealand. In Polynesia, for example, Western Samoa (now Samoa) was granted independence in 1962, and Tonga ceased to be a British protected state in 1970. Decolonization in island groups held by the United States and France (Micronesia and French Polynesia, respectively, in particular) was and still is a slower, more complex, and conflictual process. A notable factor in both cases is the use by both powers of islands for nuclear testing until well into the 1990s.

Today, France maintains a colonial presence in French Polynesia and Wallis and Futuna (both Polynesian groups), which is contested in the former but receives general local approbation in the latter. The United States signed Compacts of Free Association with the Federated States of Micronesia and the Marshall Islands in 1986, and with Palau in 1994, and since then these entities have been recognized as sovereign nations despite some continuing US security interests in the islands. American Samoa and Guam are unincorporated US territories in which the federal government continues to maintain control of many functions, in spite of some local opposition in the case of Guam. The Northern Mariana Islands is a Commonwealth with more control of local affairs. Rapanui is administered by Chile, despite some political tension, and tiny Pitcairn Island (Polynesian from a geographical point of view) by Britain. New Zealand holds special political ties with three now self-governing territories, namely Niue, Tokelau, and the Cooks. Hawaii, a state of the United States, and New Zealand are both postcolonial entities in which the politics of indigeneity are tense. These varied historical connections, as well as the different configurations in which domination, resistance, and everything in between are embedded, have contributed to the socio-political diversity of both regions.

The states and territories of Polynesia and Micronesia include some of the tiniest political entities of the world, such as Tuvalu (10,000 inhabitants, 26 km²). Outside of New Zealand, most are heavily dependent on world powers, relying on foreign aid, migrations, and access to preferential markets to ensure at least a semblance of economic viability and, in some cases, political stability. The notable exception is Nauru, which is one of the richest countries of the world because of decades of phosphate mining.

5. Political Re-emergence

In tune with the worldwide decolonizing trends of the second half of the twentieth century, island nations began obtaining their independence from world

6. Society and Culture

In precontact times, the societies of Polynesia and Micronesia displayed a certain degree of sociocultural commonality, as witnessed, for example, in a tendency to have stratified political systems and in the elaboration of certain symbolic complexes such as *tapu*

(roughly, 'religiously-based prohibition') and *mana* (roughly, 'efficacy of divine origin') in Polynesia. However, the recognition of patterns of commonality must always be qualified in at least two ways. First, one must be attentive to the dynamics through which certain characteristics have come to be identified as 'typically' Polynesian or Micronesian in spite of the fact that these characteristics are absent in some societies of the region. Second, 'precontact times' is a vague characterization at best, and all island societies experienced radical changes over the centuries, some of which can be identified archeologically, in the course of which 'typical' Polynesian or Micronesian characteristics appeared in and disappeared from particular societies.

A significant number of societies in both regions appear to have been stratified political entities, the more so in Polynesia than in Micronesia. Particularly on large islands and island groups, chiefdoms were organized around a leadership that derived power through a combination of genealogical ties to both sacred and profane entities that confirmed claims to power (ascription) and personal performance that hopefully confirmed these claims (achievement). In some of more stratified societies, persons were also ranked into categories ranging from the highest ranks, which were commonly sacralized, to the lowest ranks, whose members were often not considered human. Other societies, in contrast, displayed little stratification and emphasized egalitarianism and consensual decision-making. Political organization on some islands was centered on localized kinship units, particularly in Micronesia but also on the smaller islands of Polynesia. The rough correlation between the amount of resources to be produced and organized and the degree of social stratification, at least in Polynesia, led some early researchers to propose that one engenders the other: chiefs, as resource managers, increase in importance as resources increase in volume (Sahlins 1958). Few anthropologists today would give such determinative power to resource management (and indeed the empirical evidence does not support the proposal), although it certainly plays a role in the development and elaboration of social stratification.

Prevalent patterns throughout the area centralized the identification of a common founding family ancestor, creating in some societies a pyramid-shaped kinship structure. In Polynesia, the most important branch was the senior patrilineal branch, and members of other branches reckoned their relative status according to their relative distance from the main descent branch, measured through relative patrilineality and seniority. Micronesian societies are predominantly matrilineal. There, matrilineal clans have historic claims to particular pieces of land and are the basis for social organization. Rank is based on claims of founding settlement of particular pieces of land. Probably all Polynesian and Micronesian societies also offered the possibility of bilateral affiliation,

whereby kinship could be based on either patrilineal or matrilineal principles, thus providing room for maneuvering for individual advantage. In Polynesia, there emerges an east-west contrast between societies in which gender and its associations constitute a pivotal organizational principle for kinship organization (as in Tahiti and Hawaii), and societies in which seniority plays a greater role (as in Samoa and Tonga). Adoption in its various guises was and is also prevalent throughout Polynesia and Micronesia.

Many aspects of precontact (or early contact) society and culture have reproduced over time and continue to characterize contemporary circumstances. However, any statement about contemporary Polynesian and Micronesian societies must take into consideration the enormous transformations that these societies have undergone in both bygone days and more recent times, introducing dynamics and agents that have been variously resisted, localized, and transformed. Two examples will suffice. The first is the recognition that many contemporary Polynesian and Micronesian societies are vigorously diasporic. While strong sociopolitical forces continue to ground society and culture in the 'islands of origin,' these forces are increasingly subject to contestation as second- and third-generation expatriates constantly negotiate their place across different identities and allegiances. The second illustration is the fact that Christianity, which is two centuries old in some area of Polynesia and Micronesia and much more recent in others, is almost universal and highly visible in both regions. In some parts of Polynesia, it has become one of the main backbones of hierarchy, exchange systems, and kinship, and is viewed as one of the most fundamental tenets of 'tradition.' In short, like all social groupings around the world, Polynesian and Micronesian societies are dynamic and adaptable entities, whose boundaries, core cultural values, and principles of social organization are subject to change, reassessment, and potential contestation.

See also: Colonialism, Anthropology of; Colonization and Colonialism, History of; Cultural Assimilation; Melanesia: Sociocultural Aspects

Bibliography

- Bellwood P 1979 *Man's Conquest of the Pacific: The Prehistory of Southeast Asia and Oceania*. Oxford University Press, New York
- Denoon D (ed.) 1997 *The Cambridge History of the Pacific Islanders*. Cambridge University Press, Cambridge, UK
- Howard A, Borofsky R (eds.) 1989 *Developments in Polynesian Ethnology*. University of Hawaii Press, Honolulu, HI
- Kirch P V 1996 *The Lapita Peoples: Ancestors of the Oceanic World*. Basil Blackwell, Oxford, UK

- Kirch P V 2000 *On the Road of the Winds: An Archaeological History of the Pacific Islands Before European Contact*. University of California Press, Berkeley, CA
- Kiste R C, Marshall M (eds.) 1999 *American Anthropology in Micronesia: An Assessment*. University of Hawaii Press, Honolulu, HI
- Lal B V, Fortune K (eds.) 2000 *The Pacific Islands: An Encyclopedia*. University of Hawaii Press, Honolulu, HI
- Sahlins M 1958 *Social Stratification in Polynesia*. American Ethnological Society, University of Washington Press, Seattle, WA
- Smith B 1985/1969 *European Vision and the South Pacific*. Yale University Press, New Haven, CT
- Thomas N 1989 The forces of ethnology: origins and significance of the Melanesia/Polynesia division. *Current Anthropology* 30: 27–41

N. Besnier

Popper, Karl Raimund (1902–94)

1. Popper's Life

Karl Raimund Popper was born in Vienna, Austria, on July 28, 1902. His father, Simon [Siegmund Carl] Popper (1856–1932), who had come from Bohemia to Vienna, was a well-established lawyer with his offices and family residence in the center of Vienna. Karl Popper's mother, Jenny Popper (née Schiff) (1864–1938) came from a musical family and was a gifted pianist herself. Both Simon and Jenny Popper were Jews by origin, but converted later in their lives to Protestantism. Besides their son Karl, they had two older daughters, Dora (1893–1932) and Anna Lydia (1898–1975).

Karl Popper left high school (Franz Josef Gymnasium) at the age of 16 without a degree, moved out of his parents' home and attended lectures at the University of Vienna on various subjects, in particular mathematics, philosophy, physics, and psychology. Besides occasional jobs to earn his living (e.g., coaching American students), Popper was engaged in social activities. For some time, he worked as an assistant of Alfred Adler on neglected children. Hints in Popper's autobiography indicate that this connection (through which he got acquainted with Adler's 'individual psychology') contributed to Popper's later critique of psychoanalysis as a 'pseudoscience.' In his youth, Popper had a strong favor for democratic socialism. For a few months, he even considered himself a communist, from which he refrained, in 1919, after witnessing a shooting incident in which socialist workers were killed by police forces. This experience of a useless sacrifice of lives demonstrated to him the futility of ideologies. A second, more intellectual

experience in 1919 was his taking interest in Eddington's test of Einstein's eclipse predictions carried out at an expedition of the Royal Astronomical Society, whose results gained widespread publicity at the time. During a total solar eclipse which was observed from West Africa and Brazil, a shift of apparent position of stars in the presence of the sun's gravitational field could be established, as predicted by the general theory of relativity. This coined Popper's later insistence that scientific theories develop by exposing them to possibly falsifying experience. Around that time, Popper visited a lecture by Einstein in Vienna.

The young Popper also tried manual work. From 1922 to 1924 he was an apprentice to a cabinet maker. At the same time, he passed the high school exams ('Matura') as an external participant and formally enrolled as a student at the University of Vienna. He also considered becoming a musician. After being disappointed with the sort of music cultivated in the Schönberg circle, he enrolled as a student of Church Music at the Vienna Academy of Music ('Konservatorium') and stayed there for one year. He was admitted upon submission of a fugue for organ in F sharp minor. Besides his scientific studies at the University of Vienna, in 1924 he qualified as a primary school teacher. Since no appropriate teacher's post was available at the time, Popper worked as a social worker with neglected children.

In 1925 Popper became a student at the Pedagogic Institute of the City of Vienna, where he stayed until 1927. This institute, which Popper was admitted to due to his status of a social worker, had been newly founded. It was linked to the University, though formally independent. In today's terms, its purpose was to scientifically accompany and evaluate the school reform movement. This school reform movement had been politically supported by the Austrian social democrats since 1919 and was in 1925 still very strong in Vienna. It was characterized by ideas of abolishing school learning as an accumulation of facts in favor of developing a more autonomous and participating attitude. The influence of these ideas on Popper should not be underestimated. His later criticism of what he calls the 'bucket theory of knowledge,' i.e., the inductive generation of knowledge by collecting facts, is related to pedagogic proposals on how learning proceeds or should proceed. Actually, Popper's first published papers were in this field and appeared in journals belonging to the school reform movement. The very first one *Über die Stellung des Lehrers zu Schule und Schüler (On the teacher's attitude towards school and students, 1925)* is very significant in this respect.

At the pedagogic institute, Popper met Josefine Henninger (1906–85), who was training as a teacher and who was to become his wife in 1930. (They had no children.) He also met the psychologist Karl Bühler and the philosopher Heinrich Gomperz, both of them

professors at the University. They were of crucial importance to Popper's intellectual development. It was through discussions with Gomperz and reading his works that Popper shifted away from any psychological approach to the philosophy of science. From Bühler he adopted the distinction of expressive, signaling, and descriptive functions of language, which still figured prominently in Popper's late metaphysical writings.

In 1928 Popper completed his Ph.D. in Psychology with Karl Bühler and Moritz Schlick as examiners. The subject of the thesis was *Zur Methodenlehre der Denkpsychologie* (*On the Methodology of the Psychology of Thinking*, unpublished), a work which combined Popper's interest in psychology with his methodological bias that he had finally shifted to during his time at the Pedagogic Institute. In 1929 Popper qualified as a teacher for secondary schools ('Hauptschulen'), and in 1930 he became a teacher in Vienna.

Since 1922 Popper had been in contact with members of the Vienna circle, a group of philosophers around Moritz Schlick who propagated what they called 'logical positivism,' a version of empiricism which made strong use of modern mathematical logic as a tool of philosophical analysis. Mathematical logic had already been developed by Gottlob Frege (1848–1925) and Bertrand Russell (1872–1970) decades ago, but had not been appropriately appreciated within the established academic philosophy. Though never a member of the circle, Popper attended its sessions occasionally and was in contact with several of its members, in particular Rudolf Carnap, Victor Kraft, and Herbert Feigl. Although he shared certain basic attitudes of the Vienna Circle, in particular the plea for a scientifically oriented philosophy, he was highly critical of other tenets such as the theory of meaning and the form of empiricism propagated. Motivated by Feigl, he put this criticism into a large manuscript in two volumes *Die beiden Grundprobleme der Erkenntnistheorie* (*The Two Basic Problems of Epistemology*, published in 1979, the second volume only as a fragment). A shorter version of the second volume of this manuscript was published in 1934 as *Logik der Forschung* (*Logic of Scientific Discovery*), which was to become his main work.

The book was well received. Reviews appeared in many journals. Even Einstein wrote Popper a letter after the book was sent to him through intermediates. In 1935 and 1936 Popper received invitations to England, where he lectured at Bedford College and Imperial College, London, at the London School of Economics and at Cambridge. He made a lot of contacts. Of most importance for his later career were those with Ernst Gombrich and Fritz Hayek. However, attempts to find a position in England did not materialize, and in 1937 he emigrated with his wife to New Zealand, accepting an offer of a lectureship at Canterbury University College.

During the time in exile Popper intensified his interest in social and political philosophy and in the philosophy of history, besides continuing his research in logic and philosophy of science. From that time on, he published mainly in English. He wrote a series of articles *The Poverty of Historicism* (Popper 1957a originally 1944–5); and, more importantly, his two-volume classic *The Open Society and Its Enemies* (Popper 1945). For both works he had to struggle to find a publisher. Due to the *Open Society*, he acquired his initial postwar reputation and, in particular, his post, from which he carried out all subsequent work. Through the intervention of Hayek who, together with Gombrich, had made the publication of the *Open Society* with the publisher Routledge possible, he was offered a readership in 'Logic and Scientific Method' at the London School of Economics in 1945, which Popper took up early in 1946, shortly after publication of the *Open Society*. In 1949 he received a professorship of the same title at the London School of Economics, which he held until his retirement in 1969.

During the first five years of his stay in Britain, he was sort of a 'star.' The *Open Society* made him a quite popular theoretician of political liberalism. It has often been said that this book filled some intellectual vacuum in the postwar period. Popper was invited to many places and, at the same time, had a substantial audience at the London School of Economics. In the 1950s, he shifted more to working from home with less institutional involvement. At the same time, his research interests moved away from political and social philosophy and back towards the subject of the *Logik der Forschung*, i.e., the methodology of the natural sciences. However, the first English edition of this book appeared only in 1959, with a lot of material added compared with the German original. Many other books and plenty of papers followed. Up to his death, Popper produced an enormous scientific output and had a vast correspondence. From the mid-1960s on, Popper received a great number of prizes and honorary degrees in many countries, including Knighthood and Membership of the Royal Society in Britain. He died on 17 September 1994 in London. His ashes were buried in Vienna, where his wife had died in 1985.

2. Popper's Work

Popper's work covers nearly all areas of theoretical philosophy. His overall profile is that of a great philosophical classic. However, most important are his contributions to the methodology of science, where he established falsificationism as one of the basic orientations in the modern philosophy of science.

The basic claims of falsificationism are put forward in the *Logik der Forschung* (1934, official year given in

the book: 1935). Most of Popper's later methodological contributions can be viewed as elaborations or amendments of what was already stated in his first book. Even Popper himself treated the *Logik der Forschung* like a classical text, with all changes in the first English translation (*The Logic of Scientific Discovery* 1959) as well as in the later 10 German and 10 English editions, as a system of footnotes and appendices to the 1934 edition.

Falsificationism can be described as a 'why not' methodology in contradistinction to a 'why' methodology with respect to scientific theories. According to Popper, a scientific theory is posed without any prior justification. It gains its confirmation, or 'corroboration,' which is Popper's preferred term in later writings, not by experimental evidence demonstrating *why* it should hold, but by the fact that all attempts at showing *why* it does *not* hold have failed. The main activity of an experimental researcher is trying to refute or '*falsify*' a scientific theory, rather than to justify or '*verify*' it. Theoreticians who design theories should do this in such a way that they can be refuted most easily. In no case must a theory be immunized against possible falsification, e.g., by the introduction of *ad hoc* hypotheses. We rely on a theory only because we have not managed to falsify it (so far), not because we have any positive or direct reasons in its favor.

His own theorizing, and especially the idea of falsificationism, is considered by Popper as proposing 'methodological rules.' These methodological rules have to prove fruitful in the understanding of scientific research and, particularly, of scientific progress. The main argument in favor of his falsificationist proposal is that the converse proposal, which is called 'verificationist,' is bound to fail. Verificationism, which underlies the 'inductivist' methodology, claims that a scientific theory is supported by a collection of verifying instances. However, as has already been established by David Hume (1711–76), the inductive reasoning from a finite set of observations to a general statement is not logically valid if the domain of reasoning is infinite or at least indefinite. Furthermore, any attempt at creating a special nondeductive form of reasoning (an 'inductive logic') can only be justified by means of induction itself, thus ending up in a vicious circle. To quote Popper's famous example, no observation of white swans can verify the general statement that all swans are white.

Popper's falsificationist methodology rests on the assumption that unlike verification, falsification of a general statement by experimental or observational means is logically without any problem. The statement 'All swans are white' can be falsified by the observation of a single black swan. The logic involved here is just ordinary deductive logic, due to which the falsity of an instance derived from a general hypothesis falsifies the hypothesis itself. Seemingly singular statements such as 'The planet Pluto has a moon,' which at first glance seem to be verifiable but not falsifiable, are viewed by

Popper not in isolation but as embedded in a general theory and therefore are no counterexample to what he calls the 'asymmetry between verification and falsification.'

More complicated is the example of statistical theories where definite falsification is not possible. Here Popper applied as an additional methodological rule that what is extremely unlikely can be neglected as something which cannot be reproduced deliberately. This allows Popper to cope with the standard way of rejecting statistical theories by means of a critical test (usually defined by stipulating a level of significance).

In traditional methodology, particularly that of the Vienna Circle, inductivism was tied to empiricism and guaranteed the connection between theoretical statements and their empirical basis. In falsificationism, the empirical significance of theoretical statements is kept by relating them to basic statements which possibly falsify them. However, these basic statements are no longer purely observational in some epistemological sense, but are themselves theory-laden from the very beginning. Even elementary perceptual evidence has to rely on theoretical assumptions. Thus Popper rejects any semantic division between theoretical and observational statements and, especially, the claim that a particular sort of certainty comes with the latter. According to Popper, the validity of basic statements can only be fixed in the course of the testing of a theory, and, in the end, this fixing is a matter of convention. This variant of conventionalism, which enters Popper's approach in addition to the conventionalism inherent in the choice of methodological rules, has repeatedly been the subject of criticism.

For Popper the distinction between a theory which can be falsified by means of basic statements and one which is not falsifiable serves as a demarcation criterion for empirical science vs. metaphysics. However, metaphysics, though not falsifiable, is not meaningless. Metaphysical statements may serve useful purposes, even within science. For example, the principle of causality ('any event can be causally explained'), which cannot be falsified and is therefore metaphysical, can be viewed as the useful methodological rule never to abandon the search for laws which explain observed facts. This view of metaphysics is less rigid than that of the Vienna Circle, who considered metaphysics to be meaningless throughout. It allows Popper to sharply criticize certain nonfalsifiable theories as 'pseudosciences' (such as psychoanalysis) but nevertheless to make sense of certain traditional philosophical claims or to appreciate the fruitful role of metaphysical research programs in the history of physics ('A metaphysical epilogue,' in *Quantum Theory and the Schism in Physics* 1956, published 1982b).

In his later works, Popper extended the methodological approaches of the *Logik der Forschung* towards a general account of the dynamics of scientific theories ('Truth, rationality, and the growth of scientific knowledge,' in *Conjectures and Refutations* 1963).

According to this view, theoretical progress is made by successive critique and revision of existing theories, which is governed by the idea of objective truth. Scientific progress works by replacing a falsified theory with a new one which, in the ideal case, is more encompassing than the previous one in that, while coping with the new results, it also covers the unfalsified instances of the old theory. This appeared to be the case with classical and relativistic gravitational theories in the presence of the results of the 1919 solar eclipse observations, which were so impressive to the young Popper. His later claim is that the new and more encompassing theory is nearer to the truth than the older and more special one, even if it should at some point become falsified itself. Popper has tried to characterize the distance of a theory from truth, its 'truthlikeness' or 'verisimilitude,' in numeric terms. Considerable work has been carried out in recent decades to develop improved measures.

Within the revision of scientific theories, Popper strongly distinguishes between hypotheses being tested and background knowledge which, though testable in principle, is not being questioned at the same time. This distinction makes the falsification of parts of a theory possible, and not only the falsification of a theory as a whole. Therefore Popper pleads to formulate theories in a modular way such that its parts can be tested independently of each other. He is opposed to 'holistic' views of theories, according to which, in a single test, the whole theory (Pierre Duhem) or, even more radically, all our theoretical knowledge including logic and mathematics (W. V. O. Quine) is at stake. Popper defended his model of scientific progress as a change of scientific hypotheses by means of falsification against historic approaches which denied the rational character of scientific change. So he was strongly opposed to Thomas S. Kuhn's account of scientific revolutions as irrational changes without the concept of 'progress' being applicable. Similarly, he argued with Imre Lakatos (1922–74) and Paul K. Feyerabend (1924–94), who started as Popperians, but later on made changes in falsificationist methodology by weakening and relativizing Popper's strict idea of progress by critical testing.

In *Evolution and the Tree of Knowledge* (1961, published 1972 in *Objective Knowledge*), Popper embedded the idea of theory change by falsification and revision within a general theory of the evolution of knowledge. Here 'knowledge' is viewed as a form of adaptation of an organism to the environment. Popper's schema of progress from problems via falsification to new problems is interpreted as the evolutionary transition from organisms via natural selection to organisms which are better adapted. By viewing all life as problem solving (*All Life is Problem Solving* 1994c), Popper claimed to combine theories of biological evolution with theories of scientific progress, relating his approach to evolutionary epistemology in the sense of D. Campbell. However, it must be

emphasized that, unlike most evolutionary epistemologists, Popper was not a naturalist. He insisted that the cultural-scientific evolution is rational and based on conscious and deliberate criticism, not on natural selection governed by chance.

In his later work, Popper also replaced the narrow concept of 'falsification' with a broader concept of 'criticism' or 'critical examination.' This broader concept not only applies to scientific theories which are formulated in a logical language but to philosophical claims which are metaphysical. It involves tests of consistency and the investigation of how far the theory in question is able to solve particular problems. This orientation, with critical examination as its cornerstone, was called 'critical rationalism' by Popper and has since become the standard denomination of the Popper school. The main examples of Popper's dealings with metaphysical problems are his treatment of determinism and his philosophy of mind. Quite independent of problems of quantum mechanics, which often serve as arguments against determinism, in *The Open Universe: An Argument for Indeterminism* (1956, published 1982a), he developed an indeterministic world view. In 'Of clouds and clocks' (1966, reprinted in *Objective Knowledge* 1972), he extended this view with arguments for the openness of the future for free action. In the mind–body debate, Popper defended a mind–body dualism ('Language and the body–mind Problem,' 1953, reprinted in *Conjectures and Refutations* 1963), which he later extended to a model of three worlds of (a) physical entities, (b) mental states, and (c) objective contents of thought, which causally interact. This approach gained much attention through his discussions with J. R. Eccles which, together with their separate contributions, are presented in their joint book *The Self and its Brain* (1977).

Epistemologically, Popper was an outspoken objectivist and realist. On the Popperian view, our theories are not just methodological instruments or means of acquiring (some degree of) certainty about something, but attempts at explaining the real world. As reality is described by means of true sentences, this is another way of claiming that the goal of science is truth. Part of Popper's objectivism and realism is his rejection of any subjective interpretation of probability. In particular, he was strongly opposed to Bayesianism and its view that scientific progress is just changing the probabilistic degree of belief in certain propositions, which means that scientific dynamics is probability dynamics (*Probability Magic or Knowledge out of Ignorance* 1957b). Both the probability of hypotheses, for which Popper coined the term 'logical probability,' and statistical probability, which may be part of a hypothesis, are objective probabilities reflecting properties of what they are attributed to. Logical probabilities are inversely proportional to the content of a theory, the less probable theory having a higher content. Statistical probabilities which, in the *Logik*

der Forschung, were interpreted as relative frequencies, were later viewed as objective 'propensities' of experimental designs to develop certain outcomes (*Realism and the Aim of Science*, 1956, published 1983). This propensity interpretation of probability claims to be a solution of the problem of assigning probabilities to singular events, rather than only to events as members of a given sequence or set of related events.

Popper's attitude against probabilities as degrees of belief does not contradict the fact that he attempted to measure the degree of corroboration of a theory in probabilistic terms. However, first of all, Popper's degree of corroboration, though relying on probabilities, is not a probability itself, i.e., does not obey the laws of the calculus of probability. Second, the degree of corroboration is intended as a strictly objective measure which reflects the way the hypothesis in question was subjected to falsification, and is definitely not meant as a degree of rational belief in a hypothesis. According to Popper, epistemology is the methodological treatment of objective scientific contents and not a theory of the knowing subject ('Epistemology without a knowing subject,' 1967, reprinted in *Objective Knowledge* 1972).

Popper's social and political philosophy which was developed during his emigration in New Zealand and published in *The Poverty of Historicism* (1944–5) and *The Open Society and its Enemies* (1945) can, to some extent, be viewed as an application of falsificationism to the area of social interaction. By 'historicism' Popper denoted the claim that the methodology in the historical and social sciences differs fundamentally from those in the natural sciences but, nevertheless, allows one to formulate objective laws of historical development, which cannot be changed by subjective action. The standard example is Marxist philosophy of history (historical materialism). In contradistinction to that, Popper put forward the idea of 'piecemeal social engineering,' i.e., of a development of social institutions in small steps which can be revised and whose potential damage can be controlled. According to Popper, piecemeal social engineering can be combined with a falsificationist form of social research, which is based on trial and error as it generates models for social development. For areas of historical and sociological research which are interested in the explanation of single events rather than the testing of general laws, Popper proposed the procedure of a 'logic of the situation' which describes the acting of individuals under the hypothesis of rational purpose-oriented behavior within a certain social situation. In this context he defends the autonomy of sociology against psychologism, as this situation is described in objective terms by reference to social institutions, etc. Since 1961, Popper defended this account of social theorizing against representatives of the neo-Marxist Frankfurt school in Germany, in particular, Theodor W. Adorno and Jürgen Habermas, in a dispute which became known as the 'positivist dispute in German

sociology,' and which made Popper well known to a wider audience in the German-speaking countries.

The *Open Society* extends the critique of historicism to a critique of the philosophical theories of state and society by Plato, Hegel, and Marx, which Popper considers to be philosophical precursors of political totalitarianism. According to Popper, they advocate the idea of a closed society, which is a collectivistic and organismic whole, without the distinction between natural and cultural laws. In an open society, which Popper strongly favors, individuals are personally responsible for their decisions and can therefore be critical to social regulations. Popper argues that the traditional question 'who should rule' ought to be replaced by the question of how to minimize the risk of bad rulers and how to get rid of them once they are in power. For Popper, the long-standing philosophical discussion about the best rulers belongs to justificationist thinking that has to be abandoned in favor of critical thinking, just as in general methodology. The utilitarian principle of maximization of happiness should give way to the idea of minimizing suffering.

Although there appears to be a lot of moralizing both in the *Open Society* and in many short papers and interviews on questions of morals and politics given towards the end of his life, this should not be understood as fragments of an ethical theory. For Popper, there is no such thing as ethics, because ethical sentences are not propositions in the logical sense that one could argue for or that one could reject. However, there are moral actions. These are actions which are carried out by someone who is aware of their consequences. Since the consequences of actions can be critically examined in the sense of critical rationalism, moral decisions are not irrational.

3. Relevance and Impact of Popper's Philosophy

Popper's methodological program was revolutionary as it reverted the established conceptual order between certain fundamental notions. He put 'negative' concepts such as 'falsification' and 'rejection' first, and made 'positive' concepts such as 'acceptance' or 'corroboration' dependent on them. Whereas according to the traditional philosophical view of science established ever since Bacon, acceptance of a theory is based on the direct generalization from observation, according to Popper, it can only rely on failed rejection. It is therefore correct to speak of Popper's 'negative methodology.' As Popper tried to show in his later writings, this is not just a program for the natural sciences, but applies to any area of scientific reasoning and even to metaphysics, where negative concepts such as that of 'criticism' come first. The credit we give to a conjecture is based on the fact that refuting it by criticism has been unsuccessful. This program is even the basis of Popper's political phil-

osophy. The negative question of how to restrain political power and to remove bad rulers from power is of primary concern compared with the positive one of whom to give power. Popper's negative methodology is a unified view of rationality under the header 'critical rationalism.'

This comprehensive perspective of theoretical reasoning and social and political action makes Popper's philosophy comparable with that of great classics like Aristotle or Kant. It was also one of the sources of the great success which Popper enjoyed still during his lifetime. As a general and unifying view, 'critical rationalism' is, of course, liable to oversimplification and, among certain adherents of Popper there has, in fact, been a considerable amount of that. On the other hand, it was only critical rationalism as a general orientation that could publicly compete with the eloquence of hermeneutic or neo-Marxist approaches. Philosophical theories like logical empirism, though with respect to intellectual clarity and responsibility by no means behind Popper's, would have been too academic to develop any major public significance.

In fact, it is not very difficult to give an outline of the basic principles of critical rationalism to somebody who is not a trained philosopher. Critical rationalism is an approach which offers something at any depth of understanding. Popper's works both cover extremely intricate discussions, e.g., of the methodological role of probability, whose significance can only be judged by highly specialist researchers, and general questions of reasoning, society, and politics at a level accessible to anybody who is really interested. With the exception of the chapters on probability and on quantum mechanics and of the appendices, the *Logik der Forschung* can easily be studied without deep expertise in academic philosophy. Even from a first reading, the average student would gain a preliminary understanding of falsificationism. Matters of presentation are very important in philosophy, and Popper was a master in that. His language is extremely clear and nonpretentious, which distinguishes him from many other philosophers, particularly in the German tradition.

Popper made a strong impression on politicians of various directions, although he was himself never involved in party politics. Both conservative and social democratic or labor politicians have referred to Popper to justify their respective programs and proposals. The term 'open society,' which Popper coined, has become widely accepted as a designation of basic features of a democratic state.

Popper's lasting impact on the shape of philosophy of science was not only due to his methodology in the narrower sense, but also due to his drawing attention to the *dynamics* of science. It is significantly due to Popper's influence that the philosophical and logical reconstruction of scientific reasoning is not just a modeling of static theories in logical terms but, to a great extent, an account of theory dynamics, i.e., of

what it logically means that a theory is being given up or revised. The combination of philosophy of science with the history of science, which is reflected in many divisions of university departments, denominations of chairs, etc. emerged on this background, as philosophical theories of theory dynamics have to be tested against the historical development of theories, and, conversely, the proper understanding of historical developments requires appropriate methodological tools. Although today's theorizing differs in many respects from what Popper argued for and often contradicts his claims (i.e., Bayesian approaches in the field of belief revision), the research program of putting dynamics of theories first in the methodology of science is unthinkable without Popper's revival of the induction problem and his negative methodology as a proposed solution.

The influence of Popperian theories on the 'working scientist' should not be underestimated. If scientists have any interest in the methodology of science at all, they would know of Popper's approach. This is particularly true of the social and behavioral sciences. There, falsificationism figures prominently in textbook chapters dealing with methodological questions. One may speculate why falsificationism is so appealing to social scientists. A possible answer might be that statistical methodology, which lies at the heart of the social sciences, has a bias towards falsification. According to the most widespread statistical methodology, which is a combination of accounts by R. A. Fisher and by J. Neyman and E. Pearson, the rejection or falsification of a statistical null hypothesis at a certain level of significance is a direct and uncontroversial result, whereas the acceptance of a positively formulated statistical hypothesis is only indirect and preliminary (as it depends on features such as sample size) and, in any case, is technically more involved.

Certain results of modern computer science have challenged Popper's view on induction. His methodology comes along with the claim that we always start with theories or hypotheses which, in the course of time, are then revised to better hypotheses by trial and error. How a new hypothesis is formed was for Popper not the subject of rational investigation. In artificial intelligence and, in particular, in learning theory, algorithms have been developed to generate theories that fit given data. This does not contradict falsificationism, as these heuristically developed hypotheses would then have to be tested, just as falsificationism claims. However, the availability of such procedures destroy the myth that a theory is the extrarational product of a genius or some sort of accidental mutation, and lays the foundation for algorithmic inductive heuristics as a rational enterprise.

There are a lot of questions in the technical development of the Popperian approach, e.g., his treatment of probability and of logic, whose significance for future developments has still to be proven.

Something similar holds for Popper's metaphysical theories. However, these are less important details compared with his basic methodological claims, which will remain a standing challenge.

See also: Causation (Theories and Models): Conceptions in the Social Sciences; Causation: Physical, Mental, and Social; Explanation: Conceptions in the Social Sciences; Hayek, Friedrich A von (1899–1996); Historicism; Liberalism; Liberalism: Historical Aspects; Liberalism: Impact on Social Science; Positivism, History of; Positivism: Sociological; Scientific Reasoning and Discovery, Cognitive Psychology of

Bibliography

- Corri R 1996 *An Introduction to the Thought of Karl Popper*. Routledge, London
- Geier M 1994 *Karl Popper*. Rowohlt, Hamburg
- Hacohen M H 2000 *Karl Popper—The Formative Years, 1902–1945: Politics and Philosophy in Interwar Vienna*. Cambridge University Press, Cambridge, UK
- Keuth H 2000 *Die Philosophie Karl Poppers*. Mohr Siebeck, Tübingen
- Magee B 1985 *Philosophy and the Real World: An Introduction to Karl Popper*. Open Court, La Salle
- O'Hear A 1995 *Karl Popper: Philosophy and Problems*. Cambridge University Press, Cambridge, UK
- Popper K R 1925 Über die Stellung des Lehrers zu Schule und Schüler: Gesellschaftliche oder individualistische Erziehung? *Schulreform* 4: 204–8
- Popper K R 1934 *Logik der Forschung: Zur Erkenntnistheorie der modernen Naturwissenschaft*. Julius Springer, Vienna (10th edn. Mohr Siebeck, Tübingen 1994) [*The Logic of Scientific Discovery*. Hutchinson, London 1959 (10th edn. 1980)]
- Popper K R 1945 *The Open Society and its Enemies, Vol. I: The Spell of Plato, Vol. II: The High Tide of Prophecy: Hegel Marx, and the Aftermath*. George Routledge, London (2nd edn. Routledge & Kegan Paul, London 1952, many reprints)
- Popper K R 1957a *The Poverty of Historicism*. Routledge & Kegan Paul, London [originally as a series of papers in *Economica* 11 (1944) and 12 (1945)]
- Popper K R 1957b Probability magic, or knowledge out of ignorance. *Dialectica* 11: 354–72
- Popper K R 1963 *Conjectures and Refutations: The Growth of Scientific Knowledge*. Routledge & Kegan Paul, London
- Popper K R 1972 *Objective Knowledge: An Evolutionary Approach*. Clarendon Press, Oxford, UK (2nd edn. 1979)
- Popper K R 1976 *Unended Quest: An Intellectual Autobiography*. Fontana, London (= *Intellectual Autobiography* In: Schilpp 1974)
- Popper K R 1979 In: Hansen T E (ed.) *Die beiden Grundprobleme der Erkenntnistheorie*. Mohr Siebeck, Tübingen
- Popper K R 1982a The open universe: An argument for indeterminism. In: Bartley W W III (ed.) *Postscript to the Logic of Scientific Discovery*. Hutchinson, London
- Popper K R 1982b Quantum theory and the schism in physics. In: Bartley W W III (ed.) *Postscript to the Logic of Scientific Discovery*. Hutchinson, London
- Popper K R 1983 Realism and the aim of science. In: Bartley W W III (ed.) *Postscript to the Logic of Scientific Discovery*. Hutchinson, London

- Popper K R 1984 *Auf der Suche nach einer besseren Welt: Vorträge und Aufsätze aus dreißig Jahren*. Piper, München (*In Search of a Better World: Lectures and Essays from Thirty Years*. Routledge, London 1992)
- Popper K R 1985 In: Miller D (ed.) *Popper Selections*. Princeton University Press, Princeton, NJ
- Popper K R 1990 *A World of Propensities*. Thoemmes, Bristol, UK
- Popper K R 1994a The myth of the framework. In: Notturmo M A (ed.) *Defence of Science and Rationality*. Routledge, London
- Popper K R 1994b Knowledge and the body–mind problem. In: Notturmo M A (ed.) *Defence of Interaction*. Routledge, London
- Popper K R 1994c *Alles Leben ist Problemlösen: Über Erkenntnis, Geschichte und Politik*. Piper, München (*All Life is Problem Solving*. Routledge, London 1999)
- Popper K R 1996 *The Lesson of this Century: Karl Popper Interviewed by Giancarlo Bosetti*. Routledge, London
- Popper K R 1998 In: Petersen A F (ed.) *The World of Parmenides. Essays on the Pre-Socratic Enlightenment*. Routledge, London
- Popper K R, Eccles J C 1977 *The Self and its Brain: An Argument for Interactionism*. Springer International, Berlin
- Sandra Pralong I J (ed.) 1999 *Popper's Open Society After Fifty Years*. Routledge, London
- Schilpp P A (ed.) 1974 *The Philosophy of Karl Popper*. Open Court, La Salle, 2 Vols.
- Shearmur J 1996 *The Political Thought of Karl Popper*. Routledge, London
- Simkin C G F 1993 *Popper's Views on Natural and Social Science*. Brill, Leiden, The Netherlands

P. Schroeder-Heister

Popular Culture

For most of us, popular culture is a source of entertainment. It encapsulates mass media (such as film, television, print media), and includes sports events, advertisements, street fairs, and tourism. We engage with popular culture because we get pleasure from it; at the same time, we are also informed by it. We usually associate information with news rather than with entertainment. But other forms of popular culture that we often associate with 'mere entertainment'—for instance, tourist practices, popular music, films, sports, and television programs—also inform us in profound ways. For example, the US sitcom, the *Cosby Show*, may shape our ideas not just of what an ideal family is, but also about the relationship between race and upward mobility; a romance novel might instill in us expectations about intimacy as well as about gender and sexuality. Popular culture plays an important role in shaping our personal and collective identities. It also provides us with an analytic lens to understand sociohistorical processes such as class, gender and sexuality, nationalism, and transnationalism.

1. What is Popular Culture?

In the modern world, popular culture is available to what we may call 'the public' or to 'masses' of people. For this to happen, there have to be established means of production, distribution, and reception. For example, for a soap opera to exist there have to be a crew of script writers, actors, a director, a producer—in other words, a crew of people who work behind the scenes and in front of the camera. There also has to be a means to finance this production: in capitalist societies, soap operas are financed by sponsors who pay for the production and telecast of a soap in exchange for being able to advertise on it. This is how the term 'soap opera' originated: soap operas were sponsored originally by manufacturers of detergents. Even today, the manufacturers of detergents and other household goods like washing machines, kitchen appliances, and cleaning agents are key advertisers on soap operas. There also needs to be an infrastructure for its distribution. This is where television networks come in. Very often, television networks are part of larger media conglomerates or groups of companies. Television production is, therefore, part of a larger system of industrial production.

Popular culture presents us with a vantage point on the world: it is a domain that shapes our sense of self or our identity. Not only is it available generally to a large number of people, it also responds to and creates points of identification among large groups within a community or nation. Popular culture plays an important role in constituting and representing the public to itself. It is public in two senses: not only does it circulate widely and is often associated with public exhibition and display, it also configures 'the public,' 'the nation,' 'society,' 'culture,' 'family,' and other collectivities. From a sociological viewpoint, popular culture is probably the most important influence on the lives of millions. However, it is important to recognize that viewers are not just unthinking targets of popular culture. Rather, in order to understand the cultural significance of popular culture, we need to examine how its consumers respond to, reinterpret, and employ its messages.

Popular culture is a site of struggle between dominant discourses and forces of resistance. It contains 'points of resistance' as well as 'moments of super-session;' it forms a 'battlefield where no once-for-all victories are obtained but where there are always strategic positions to be won and lost' (Hall 1981, p. 233). Mass media may open up a 'zone of debate' about issues current in a given community (Gupta 1995). For instance, television shows like the US sitcom 'Ellen' enabled the rearticulation of debates around homosexuality. Popular culture, then, does not simply 'reflect' the lives and worlds of audiences; nor does it 'impact' these worlds or lives in a simple way. Instead, popular cultural texts are social formations in and of themselves.

Our subjectivities are constituted in the process of making sense of the world around us and in our interaction with others. In capitalist and postcapitalist societies, these processes are refracted by mass media. Film, television, popular music, and mass market fiction constitute and represent the public realm in cognitive, moral, and emotional terms. But these representations do not impact us as if we are blank screens—instead, 'they occupy and rework the interior contradictions of feeling and perception ... they do find or clear a space of recognition in those who respond to them' (Hall 1981, p. 232). Thus, although representations are neither all-powerful nor all-inclusive, they have material effects on how people see themselves and lead their lives.

2. Scholarly Perspectives on Popular Culture

Definitions of popular culture have shifted according to the disciplinary emphases of its students. Historians of early modern Europe (e.g., Peter Burke (1978) and Robert Darnton (1984)) stress how members of different classes created and participated in popular culture and, simultaneously, were shaped and constrained by it. These historians analyze the factors influencing the popular cultures of early modern Europe, chiefly, the rise of cities and the advent of printing. They supplement thematic analyses of books with information about audiences and, thus, aim at providing the context for the meaning and uses of books.

Some of the foundational work on the relationship between capitalism and mass media is Max Horkheimer and Theodor Adorno's *The Culture Industry: Enlightenment as Mass Deception* (1944). Horkheimer and Adorno focused on how culture and the individual are 'produced' by the culture industry. They claimed that the culture industry's power lies in its ability to disseminate ideas and beliefs on a mass scale. They were concerned about how the management and control of leisure time and the socialization of the ego by the culture industry leads to a 'mass' culture that undermines the private realm.

The relationship between modern forms of popular culture and the growth of capitalism is also examined by some cultural studies scholars. Some of the most influential work in this area was produced (and has been influenced by) the Centre for Contemporary Cultural Studies at the University of Birmingham, UK. These scholars are concerned with how mass media shape dominant ideologies and discourses. Their central project is to examine the relationship between the production and consumption of popular culture and social axes of inequality, such as class, race, and gender. Although cultural studies' earliest borrowings were from literary criticism, chiefly the work of Raymond Williams and Richard Hoggart, its primary focus has been on practices of everyday life.

The greatest debts of this genre of cultural studies are to Marxist theories of culture. Its basic premise is that cultural production often takes the form of commodities. Other important propositions are: (a) popular culture is inextricably tied up with social relations; (b) popular culture can both enable social groups to define and realize their needs and reproduce asymmetries; and (c) popular culture is a terrain where social differences and struggles are articulated. These concerns reflect the influence of Gramsci on the ideological and popular struggles involved in the construction of common sense. In addition, some cultural studies scholars influenced by Lacanian analyses of film and television reconceptualize ideology in terms of how texts construct subject positions. Others have rethought the role of ideology in popular culture in terms of Foucault's analyses of discursive practices. Finally, breaking with Horkheimer and Adorno's assumption of the audience as a passive and undifferentiated mass of receivers, many cultural studies scholars conceive of audiences or consumers who actively 'decode' and interpret popular culture texts.

In addition to studying everyday practices, some cultural studies scholars also use semiotic methods of textual analysis. They treat cultural products as 'texts' to be 'read' according to interpretive strategies, including the literary analysis of narrative, semiology, and deconstruction. These analytical strategies reveal that literary conventions and forms have greater social and cultural significance than we might first suspect. For example, feminists have pointed to the place of fictional representations of romantic love in cultural constructions of femininity. Far from being 'mere' literary constructions, these representations play a crucial role in the constitution of dominant practices of gender and sexuality.

Anthropologists bring important theoretical and methodological insights to the study of popular culture because of their commitment to ethnographically analyzing the relationship between popular culture and other social institutions. Initially, mass media was neglected by anthropologists (exceptions include Bateson 1943 and Powdermaker 1950) because it was deemed 'too redolent of western modernity and cultural imperialism for a field identified with tradition, the non-western and the vitality of the local' (Ginsburg 1999, p. 297). Since the 1980s, however, many anthropologists have analyzed popular culture by studying mass media. They have examined the cultural implications of popular culture through an ethnographic attention to the everyday practices and lives of producers and/or consumers of media, and have traced how popular culture participates in the formation of identities, collectivities, and imaginaries.

Anthropological perspectives underscore that popular culture cannot be generalized across different historical and cultural contexts simply on the basis of the mass media apparatuses. Producers and consumers of mass media are, fundamentally, social and historical

subjects and mass media are, primarily, cultural institutions that need to be studied in their concrete particularity. Hence, we cannot conceptualize relationships between popular culture and the social actors who engage its myriad texts in the abstract: neither can be studied outside their location in specific socio-historical contexts.

Several anthropologists have engaged in analyses of mass media texts. These scholars, together with those who study the reception of mass media, carefully situate popular culture in the discursive and historical contexts in which it is produced and consumed. These studies undermine unilinear models of 'communication flows,' and foreground the importance of examining the specific conditions in which popular culture circulates.

3. *Popular Culture as a Lens for Studying other Sociocultural Processes*

Popular culture is a terrain where class identities are articulated, negotiated, and consolidated. The production and consumption of popular culture is mediated by class. For example, class provides individuals with 'cultural constraints' and a common 'outer frame of meaning' (Darnton 1984, pp. 6–7). A mutually influencing relationship exists between elite culture and popular culture. Popular culture also provides a site for the negotiation and articulation of class identities and hegemonies. Thus, in nineteenth-century India, Bengali elites took upon themselves the task of 'purifying' popular culture produced by working-class women: while their objective was ostensibly to 'emancipate' women of their class, their efforts resulted in the consolidation of their own middle- and upper-class identities through the marginalization of working-class popular culture (Bannerjee 1989).

The role of popular culture in mediating discourses of gender and sexuality has long been acknowledged. Two pioneering works in this regard are Laura Mulvey's *Visual Pleasure and Narrative Cinema* (1975) and Janice Radway's *Reading the Romance* (1984). Laura Mulvey examined how the cinematic forms and narratives of classic Hollywood film 'suture' viewers into 'spectatorial positions' that reinscribe dominant ideas about gender and sexuality. Mulvey's work provided feminist critics of popular culture with a paradigm for examining the role of mass media in the constitution of discourses of gender and sexuality. Janice Radway's work laid the foundation for studying the role of a specific genre of texts, romance fiction, in the constitution of discourses of gender and sexuality. Radway situated women's reading practices in their social relationships and everyday lives, and studied their interpretations of different romances in order to analyze what the books mean to their readers.

Since the publication of these works by Mulvey and Radway, the relationship of popular culture to gender

and sexuality has been analyzed by scholars too numerous to cite. Some of the most noteworthy studies in this regard, however, come from scholars, chiefly anthropologists, who have examined media in non-Western contexts. These include the analysis of the relationship between popular culture, gender and sexuality in Japan (Allison 1996), and television and notions of modern womanhood in Egypt and India (Abu-Lughod 1993, Mankekar 1999, respectively). The significance of these studies lies in their decentering of Eurocentric perspectives dominant in cultural studies, media studies, and feminist studies.

Some scholars have argued that the conceptualization of popular culture was itself engendered by the emergence of nationalism (Burke 1978). Building on Benedict Anderson's insights about the role of print capitalism in the construction of nations as imagined communities (1983), several scholars have examined the role of popular culture in the formation of national identity. For instance, diverse forms of popular culture, including tourism, exhibitions, material culture and consumption, have enabled the constitution of the national imaginary in China (Anagnost 1997). In some parts of the developing world, television has been consciously deployed to construct a 'national community': in Mexico and Venezuela, indigenous soap operas called telenovelas have combined entertainment with 'social messages' about development. In India, state-owned television was harnessed to the explicitly nationalist goals of development and national integration (Mankekar 1999) and, furthermore, participated in the consolidation of an exclusionary and strident Hindu nationalism (Rajagopal 2000).

Yet, at the same time that popular culture enables the constitution of local and national communities, some of its forms are inherently transnational. Thus, the telenovelas described above circulated not only within but across national borders in Latin America, and were adopted as models for the television programs in India. In addition, the transnational circulation of popular culture has resulted in the creation of new forms of diasporic identities. The transnational circulation of newspapers, film, television programs relayed via satellite, advertisements, and mass market fiction have created, what anthropologist Arjun Appadurai terms, 'mediascapes' (1996). Mediascapes 'refer both to the distribution of the electronic capabilities to produce and disseminate information ... now available to a growing number of private and public interests throughout the world, and to the images of the world created by these media' (1996, p. 35).

The boundaries of cultures and nations have thus proved to be permeable to globalized and transnational forms of popular culture. Yet, globalization has not resulted in a homogenization of culture. At different historical moments, globalization has resulted in a recasting of 'tradition' and processes of

creolization (Hannerz 1996, pp. 67–77). Transnational popular culture thus forces us to revisit previously held notions of culture as bounded and systemic (Appadurai 1996, Gupta and Ferguson 1997).

See also: Adolescents: Leisure-time Activities; Critical Theory: Contemporary; Critical Theory: Frankfurt School; Cultural Assimilation; Fashion, Sociology of; Fast Food Enterprises; Film: Genres and Genre Theory; Film: History; Globalization and World Culture; Mass Society: History of the Concept; Music as Expressive Form; Television: Genres; Television: Industry

Bibliography

- Abu-Lughod L 1993 Finding a place for Islam: Egyptian television serials and the national interest. *Public Culture* 5: 493–514
- Allison A 1996 *Permitted and Prohibited Desires*. Westview Press, Boulder, CO
- Anagnost A 1997 *National Past-times: Narrative, Representation, and Power in Modern China*. Duke University Press, Durham, NC
- Anderson B 1983 *Imagined Communities*. Verso, London
- Appadurai A 1996 *Modernity at Large*. University of Minnesota Press, Minneapolis
- Bannerjee S 1989 Marginalization of women's popular culture in nineteenth century Bengal. In: Kumkum S, Sudesh V (eds.) *Recasting Women: Essays in Colonial History*. Kali, New Delhi, India, pp. 127–80
- Bateson G 1943 Cultural and thematic analysis of fictional films. *Transactions of the New York Academy of Sciences* 5: 72–8
- Burke P 1978 *Popular Culture in Early Modern Europe*. Harper and Row, New York
- Darnton R 1984 *The Great Cat Massacre and Other Episodes in French Cultural History*. Vintage, New York
- Ginsburg F 1999 Shooting back: From ethnographic film to indigenous production/ethnography of media. In: Miller T, Stam R (eds.) *A Companion to Film Theory*. Blackwell, Malden, MA, pp. 295–322
- Gupta A 1995 Blurred boundaries. *American Ethnologist* 22: 375–402
- Gupta A, Ferguson J (eds.) 1997 *Culture, Power, Place*. Duke University Press, Durham, NC
- Hall S 1981 Notes on deconstructing 'the popular'. In: Samuel R (ed.) *People's History and Socialist Theory*. Routledge, London, pp. 227–40
- Hannerz U 1996 *Transnational Connections*. Routledge, London
- Horkheimer M, Theodor A 1944 *Dialectic of Enlightenment*. Continuum Press, New York
- Mankekar P 1999 *Screening Culture, Viewing Politics*. Duke University Press, Durham, NC
- Mulvey L 1975 Visual pleasure and narrative cinema. *Screen* 16(3): 6–18
- Powdermaker H 1950 *Hollywood: The Dream Factory*. Houghton Mifflin, Boston
- Radway J 1984 *Reading the Romance*. University of North Carolina Press, Chapel Hill, NC

Rajagopal A 2000 *Retailing Hinduism*. Cambridge University Press, Cambridge, UK

P. Mankekar

Population Aging: Economic and Social Consequences

The notion of aging is quite natural for human beings: from the age of reason, we understand that with the passage of time, every living thing ages and finally dies. Young children experience this with their goldfish or dog, adults with their parents and old people with their friends. By analogy, we apply the notion of aging even to nonliving things: for example, as a car deteriorates, we say that it is getting old and eventually we relegate it to the scrap lot. Although a good wine may improve with age, in the end it becomes undrinkable.

From the individual, the notion of aging is extended to the group formed by the individuals: thus we speak of an aging fleet of cars or an aging population, meaning that the average length of life of the entire group of these individual units changes with the passage of time. However, unlike the case for the individuals making up the group, the group itself is able not to age, or even to get younger. This is the case for human populations. For them the aging process is strictly structural, not biological. The decay generally associated with the aging of individuals should not be assumed to apply to aging populations.

1. Population Aging: A Modern Demographic Phenomenon

A population changes through the interaction of the fertility, mortality, and migration of its members. In recent centuries, we have witnessed remarkable changes in human fertility and mortality, which, uncontrolled since time immemorial, had always been high. In the past, there was a 'natural' fertility and mortality, and population numbers throughout the world were relatively low. Eventually, people found ways to reduce first mortality and then fertility after a certain time lag. In theoretical terms, this historical process is known as the demographic transition (Chesnais 1986). During this process, the population grows at a higher or lower rate, depending on the extent of this time lag. At the end of the process, according to the theory, both fertility and mortality are low and the population stops growing. But an inescapable law of demography says that a population either grows or it ages (Coale 1987). As all populations around the world have reduced or stopped their

growth, they are all bound to age. In some countries the fertility level is well below the replacement level, and thus their populations will decrease unless they experience in-migration, a possibility not foreseen by classical transition theory.

There are several ways of measuring the aging of a population. Instinctively, one could say that a population is aging if the average—or median—age of its members as a group increases over time. Alternatively, one could look for an increase in the proportion of old people in the population. But at what age does a person become old? We do not want to enter into a debate over the exact threshold of old age or retirement, but for purposes of illustration we will use the percentage aged 65 and over in the population. To avoid delving too far into the past or future, we will restrict our discussion of demographic aging to the period from 1950 to 2050.

During the twentieth century, the growth in the number of people aged 65 and over was relatively small and there were roughly as many people old (according to this definition) in industrialized countries as in less developed countries. However, from now until 2050, this number will grow spectacularly, mainly because of the increase in less developed countries (Table 1a). But if we look at the proportion of people aged 65 and over in the population, as a measure of demographic aging between 1950 and 2050 (Table 1b), we see that the situation differs greatly from one continent to another. Africa will remain young while the more developed countries as a whole will have more than one-quarter of their populations in the 'old' category. Asia will age at an increasing rate: in 2050, it will have more than 900 million people aged 65 and over. This is two-thirds of the world's elderly population.

In the industrialized countries, the process of population aging is relatively slow, but in developing countries it will be more intense and rapid. This can be clearly shown by a comparison of Canada and China (Fig. 1). In 2050, China will be a world leader in demographic aging with almost 450 million people aged 60 and over. This is close to 30 percent of its total population.

In the face of globalization, the changes occurring over this century will give rise to a number of socioeconomic issues that would not have made themselves felt if population aging affected only the currently industrialized countries. There will be both demographic and economic and social consequences for our societies.

2. Demographic Consequences

As is easily demonstrated, in the early stages of the demographic transition aging is much more the result of a decline in fertility than in mortality. When mortality declines from a high level, it is primarily

Table 1a/b
Population aged 65 and over, 1950–2050

Regions	1950	1975	2000	2025	2050
<i>(a) Population aged 65 and over, in thousands</i>					
World Total	131 118	232 258	417 798	813 665	1 461 092
More Developed Regions*	64 202	112 177	171 069	253 912	299 249
Less Developed Regions*	66 644	118 027	248 221	561 749	1 163 054
Africa	7 070	11 068	25 102	51 932	141 287
Latin America and the Caribbean	6 179	13 841	28 034	66 183	135 897
Northern America	14 073	25 079	38 704	69 086	85 800
Asia	57 483	101 037	217 270	477 037	911 442
Europe	44 880	77 108	107 146	147 490	173 243
Oceania	933	1 608	2 820	5 788	8 636
<i>(b) Population aged 65 and over, percentage of total population</i>					
World Total	5.2	5.7	6.9	10.4	16.4
More Developed Regions*	7.9	10.7	14.4	20.9	25.9
Less Developed Regions*	3.9	3.9	5.1	8.5	15.0
Africa	3.2	3.1	3.2	4.0	8.0
Latin America and the Caribbean	3.7	4.3	5.4	9.5	16.8
Northern America	8.2	10.3	12.5	19.0	21.9
Asia	4.1	4.2	5.9	10.1	17.3
Europe	8.2	11.4	14.7	21.0	27.6
Oceania	7.4	7.5	9.9	14.6	18.7

* More developed regions comprise Northern America, Japan, Europe, and Australia/New Zealand. Less developed regions comprise all regions of Africa, Latin America and the Caribbean, Asia (excluding Japan), and Melanesia, Micronesia, and Polynesia.

Source: United Nations Department of Economic and Social Affairs 1999

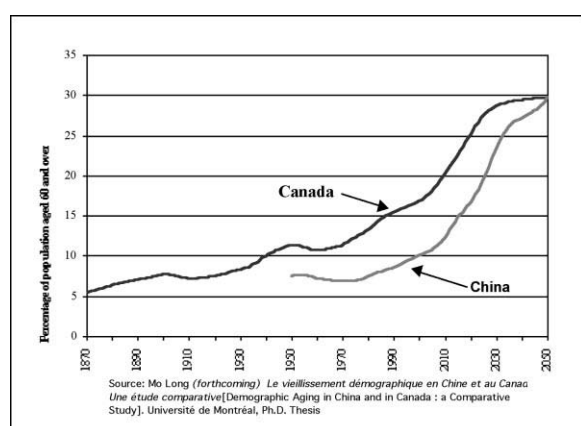


Figure 1
Population aging in Canada and China, 1870–2050

early deaths that are eliminated, in particular deaths of children and young adults. The result is a younger (not older) population, as was seen in Africa in the 1950s and 1960s (Table 1b).

But when life expectancy is already very high, as is the case now in industrialized countries, mortality becomes an important factor in demographic aging. If mortality is low and decreasing, we get a double effect.

First, increasing numbers of people reach the old age threshold—because of lowered mortality before that age. Second, people live for increasing lengths of time beyond this threshold—owing to a continued decline of mortality at old age (Bourbeau et al. 1997). The main consequence of this development is that old people no longer form a homogenous group: we now have the young old, who generally are in quite good health, and the oldest old, who generally are more dependent for income security and health. Thus we have moved from having three main lifecycle stages to four. There is a new third age that extends from retirement to the oldest-old stage. This is an entity that never existed before in human history (Laslett 1989). Another important feature in mortality changes in many societies has been an increased mortality gap between men and women. Very frequently old people in these societies are and will continue to be primarily old women.

Like any other population, the elderly population is subject to a process of incessant inward and outward movement of cohorts. In most industrialized countries the incoming generations will be clearly differentiated with respect to their numbers and their characteristics. The generations now approaching retirement are the Baby Boomers born after World War II. They will be followed by the Baby Busters, who are far less numerous. These generations, particularly the women,

will have characteristics very different from those of today's old people (Marcil-Gratton and Légaré 1987), especially in terms of education.

The changes in the age structure of our societies will cause major realignments that affect not only the elderly but the entire population. In terms of health and financial security, will old people be more independent or more dependent than before? The increasing presence of very old people tends to suggest dependence, while the characteristics of the future younger elderly appear to suggest greater autonomy. Will the younger adults, who are far less numerous than the Baby Boomers, be willing to pay for a golden twilight for tomorrow's seniors? We now turn to these questions in the following sections.

3. Economic Consequences

3.1 The Aging Process and the Working-age Population

At the beginning of the population aging process, a standard age pyramid shows the population cohorts by age increasingly reduced by mortality. As the process gradually proceeds, positive impacts on the working-age population appear. The decline in mortality eliminates early deaths and increases population numbers. Then, as the effects of fertility decline come to the fore, cohort size is increasingly reduced. When mortality improvements lead to significant reductions in old-age deaths, the members of the retired cohorts become as numerous as, and sometimes more numerous than, those of working-age cohorts. The pyramid changes into a top and finally into a cylinder. This means that at the end of the aging process, all generations are of approximately the same size except for the very old. This process is reflected in the variations of what are commonly called the dependency ratios, one of which is the ratio of people aged 65 and over to those aged 19–64. Unfortunately, the use of the adjective 'dependency' may unduly give an impression of economic dependence. These indicators are merely indices of the age structure and it would be more appropriate to speak of demographic dependency ratios.

In studies of aging we frequently find the members of the working-age population equated to actual workers. While in the past this was true for men when their labor force participation rate was close to 100 percent, it is now less so, in particular for women. To assess the economic impacts of aging and our ways of dealing with them, we needed to take into account not only population numbers, but perhaps more importantly labor force participation rates and worker productivity (Fellegi 1988). If the shift from agricultural to industrial societies had great impacts on the world of work, the shift to aging post-industrial

societies will have an equally great impact. This is particularly true for the change in the rate of labor force participation by women and by older workers.

The radical decline in fertility and otherwise changing family arrangements has freed women from a myriad of family duties and made them available for increased participation in the paid labor market. The women's liberation movement also contributed to changes in attitudes and women are becoming increasingly independent, not only at working age, but also, and perhaps especially, during retirement.

One of the corollaries—and achievements—of the development of industrial societies has been the right to retirement and associated pensions. In this century we have witnessed a decline in the labor force participation rate of older workers, along with a systematic decline in the age of leaving the labor market. This ever-earlier retirement from the labor market is now increasingly being questioned. As life expectancy increases, especially at old age, and as the time spent taking education is extended, the life segment in retirement is growing longer and the segment in which one works is getting shorter. When a large and growing share of the costs of social security, including pensions and health insurance, is borne by those who work, some adjustments must be considered. It is not possible to maintain contribution and benefit levels and, simultaneously, to bring down, or even maintain, the age of access to benefits.

There is a need for greater flexibility, both in wages and responsibilities for the population approaching retirement. For older workers, one should consider a gradual progression from wages to pension income, perhaps following a normal rather than S-shaped curve. This may call for a change in attitudes, not only among the workers themselves but also among employers and trade unions.

Solidarity among generations presupposes a minimum of equity between groups. When social security programs in the industrialized countries were re-organized after World War II, both the population and the economy were growing: the populations were young. Pay-as-you-go financing did not create any visible problems and it eliminated some perverse effects of funded programs, many of which had collapsed in Europe during the Depression and the war. In aging societies—particularly those which have experienced a large baby boom followed by a baby bust—a hybrid system combining pay-as-you-go and funded elements is likely to be more equitable across generations.

At present, the labor market is undergoing major changes. Workers will no longer hold full-time jobs for life the way they did previously in industrial societies. New workers will enter the labor market later in life. They will hold part-time jobs with varying employers, more often and possibly for a significant part of their working life, or be self-employed. From time to time they will have periods of unemployment or of retrain-

ing. Moreover, such patterns, which formerly applied mainly to women, will now apply to men as well. Consequently, retirement systems based on previous labor market participation and involving individual employers will be subject to in-depth revision. In addition, one should take the value of unpaid work into account. In particular, work performed to raise children or to care for aging parents needs to be recognized.

3.2 An Aging Society: A Burden?

Because population aging is equated to individual aging all too often, one tends to highlight problems for aging societies, in particular problems related to the costs of health and pension programs for the elderly. On this subject some issues need to be clarified and some myths should be dispelled.

First, the situation of old people in most countries has improved greatly largely as a result of post-war social programs. There have even been hard economic times when their situation has deteriorated less than that of younger people.

Second, health and pension costs will certainly continue to increase in aging societies, but at what rate will this happen? With regard to health Zweifel et al. (1999) have shown that increases in medical costs are more closely linked to the number of old-age deaths than to the size of the elderly population, and the number of old-age deaths is increasing much less rapidly than the size of the elderly population. The same is not true for pensions, whose costs are linked directly to the number of old people. However, one should not forget that retired people also pay taxes, both on their income and on consumption. As people age, their material needs (other than those related to health) decrease, however, and this might be reflected in their pension requirements.

Third, although incomes are on average lower at old age than at other ages, the elderly often hold a large share of the collective wealth through their assets, in particular by owning of their dwellings.

In brief, far from being an impediment to economic development, elderly people in aging societies can sometimes serve as an important actor if not the engine of such development.

4. Social Consequences

Increasingly, the group of old people is not homogeneous with regard to health status and wealth, and this has numerous consequences for their life-style. First, let us recall that to a large extent the world of the elderly is a world of old widows. This feature has become increasingly prominent over the twentieth century, as excess male mortality has grown while the difference in age at marriage has remained stable. In

addition, this century has been characterized by the quest for greater independence at all ages. Old people in industrialized countries prefer to live alone if they do not live with their spouse, particularly when they are in good health. But this choice does not necessarily imply isolation. Alternatively, living together with people other than one's late spouse may become a strategy for dealing with financial or health problems, perhaps as a stage that precedes institutionalization.

Multigenerational households are much more common in less industrialized countries, mainly for cultural and economic reasons. We can only wonder what will happen in the future in these societies. In a context of rapid urbanization and scarcity of dwellings, there will be an acceleration in aging following the sharp decline in fertility. With its one-child policy, China is a particularly poignant case. Demographic constraints may become apparent as daughters and daughters-in-law become increasingly less numerous and less often available to take care of old parents.

In societies with low mortality (or equivalently, long life expectancy), early deaths are reduced to a minimum and are usually related to life-style rather than caused by exogenous diseases. This produces a great deal of interest in the population's health status, and public health policy objectives may concentrate on insuring quality of life for the aging population more than on preventing early adult deaths. The focus moves from mortality to morbidity, from life expectancy to healthy-life expectancy (Robine and Romieu 1998). By means such as surveys on seniors' functional capacities, it is possible to identify impediments to independent existence and to develop policies that will minimize the adverse effects of biological and social aging.

Such policies are centered on support networks. Assistance provided to old people may come from professionals hired by the state, the community, or private enterprise. Alternatively, they may be offered by the family, friends, or neighbors. The expanding need for services provided through such formal and informal networks is an increasing source of concern. For the informal networks, the many changes that have occurred in family structures leave future prospects unclear; for the formal networks, it is mainly the cost that worries public policy makers. In this picture one should not forget that seniors are not only receivers of assistance, but very often also providers.

5. Conclusion

Some believe that the twenty-first century will see the emergence of intergenerational conflicts similar to the gender conflicts of the twentieth century. Major transformations of the population age structure have altered the ratio between the retired and the working population and will continue to do so. In addition to solidarity among generations, any reforms to social

security programs taking account of the new demographic reality need to aim to promote inter-generational equity. For this purpose, we have seen the development of new measurement techniques, known as demographic accounting (Auerbach et al. 1999). By making use of individual-level data and by making adequate long-term projections we will be able to benefit from these new methodologies.

In conclusion, we observe that for centuries, life in all its forms was regulated by nature. Gradually, human societies became able to control their destiny by largely controlling their mortality and fertility. This resulted in great changes for the societies, and in particular it accelerated their aging. The fact that most now live to old age and that our populations progressively age can only be regarded as great cultural progress (Loriaux et al. 1990). However, like all progress, individual and societal aging must be managed skillfully. Moreover, we are in the midst of a process of shifting control from society to the citizen (Légaré and Marcil-Gratton 1990). Decisions concerning demographic behavior, like whether or not to get married, whether to have children, how many and when, whether to leave an existing marriage and contract another one, and so on, are now made on the individual level more than ever before. In a world where centenarians will become rather common, people will doubtless also want to have a say on the timing of their departure from life.

See also: Age Policy; Age, Sociology of; Age Stratification; Age Structure; Aging and Health in Old Age; Cohort Analysis; Old Age and Centenarians; Retirement and Health; Retirement, Economics of; Social Security

Bibliography

- Auerbach A J, Kotlikoff L J, Leibfritz W (eds.) 1999 *Generational Accounting around the World*. University of Chicago Press, Chicago
- Bourbeau R, Légaré J, Emond V 1997 *New Birth Cohort Life Tables for Canada and Quebec, 1801–1991*. Statistics Canada, Ottawa, Canada
- Chesnais J C 1986 *La transition démographique: étapes, formes, implications économiques: étude de séries temporelles (1720–1984) relatives à 67 pays*. Presses universitaires de France, Paris [1992 *The Demographic Transition: Stages, Patterns, and Economic Implications*. Oxford University Press, New York]
- Cliquet R, Nizamuddin M (eds.) 1999 *Population Ageing. Challenges for Policies and Programmes in Developed and Developing Countries*. United Nations Population Fund, New York
- Coale A 1987 How a population ages or grows younger. In: Menard S W, Moen E W (eds.) *Perspectives on Population: An Introduction to Concepts and Issues*. Oxford University Press, New York
- Fellegi I P 1988 Can we afford an aging society? *Canadian Economic Observer* **1**,10: 4.1–4.34
- Laslett P 1989 *A Fresh Map of Life. The Emergence of the Third Age*. Weidenfeld and Nicolson, London

- Légaré J, Marcil-Gratton N 1990 Individual programming of life events: A challenge for demographers in the 21st century. *Annals of the New York Academy of Sciences* **610**: 98–105
- Loriaux M, Remy D, Vilquin E (eds.) 1990 *Populations âgées et révolution grise: les hommes et les sociétés face à leurs vieillissements* [Aged Populations and Grey Revolution: Humans and Societies Facing their own Ageing]. CIACO, Louvain-la-Neuve, Belgium
- Marcil-Gratton N, Légaré J 1987 Being old today and tomorrow: A different proposition. *Canadian Studies in Population* **14**,2: 237–41
- Martin L G, Preston S H (eds.) 1994 *Demography of Ageing*. National Academy Press, Washington, DC
- Mo Long (forthcoming) Le vieillissement démographique en Chine et au Canada: une étude comparative [Demographic Aging in China and in Canada: a Comparative Study]. Ph.D. thesis, Université de Montréal
- Murray C J L, Lopez A D (eds.) 1996 *The Global Burden of Disease: A Comprehensive Assessment of Mortality and Disability from Diseases, Injuries, and Risk Factors in 1990 and Projected to 2020* (on behalf of the World Health Organization and the World Bank). Harvard University Press, Cambridge, MA
- Organisation for Economic Co-operation and Development (OECD) 1988 *Ageing Populations. The Social Policy Implications*. Organisation for Economic Co-operation and Development (OECD), Paris
- Organisation for Economic Co-operation and Development (OECD) 1998 *Maintaining Prosperity in an Ageing Society*. Organisation for Economic Co-operation and Development (OECD), Paris
- Organisation for Economic Co-operation and Development (OECD) 2000 *Reforms for an Ageing Society*. Organisation for Economic Co-operation and Development (OECD), Paris
- Robine J M, Romieu I 1998 *Healthy Active Ageing: Health Expectancies at Age 65 in the Different Parts of the World*. REVES/INSERM, Montpellier, France
- Stolnitz G J (ed.) 1992 *Demographic Causes and Economic Consequences of Population Ageing: Europe and North America*. UN Economic Commission for Europe (ECE), Geneva, Switzerland
- Stolnitz G J (ed.) 1994 *Social Aspects and Country Reviews of Population Ageing*. UN Economic Commission for Europe (ECE), Geneva, Switzerland
- Suzman R, Willis D P, Manton K G 1992 *The Oldest Old*. Oxford University Press, New York
- United Nations Department of Economic and Social Affairs 1956 *The Ageing of Populations and Its Economic and Social Implications*. United Nations, New York
- United Nations Department of Economic and Social Affairs 1988 *Economic and Social Implications of Population Ageing*. United Nations, New York
- United Nations Department of Economic and Social Affairs 1999 *World Population Prospects: The 1998 Revision*. United Nations, New York
- World Bank 1994 *Averting the Old Age Crisis: Policies to Protect the Old and Promote Growth*. Oxford University Press, New York
- Zweifel P, Felder S, Meiers M 1999 Ageing of population and health care expenditure: A red herring? *Health Economics* **8**,6: 485–98

J. Légaré

Copyright © 2001 Elsevier Science Ltd.
All rights reserved.

11741

Population and Technological Change in Agriculture

Induced innovation refers to the process by which societies develop technologies that facilitate the substitution of relatively abundant (hence cheap) factors of production for relatively scarce (hence expensive) factors in the economy. For example, the choice between land saving vs. labor saving technological change has been largely determined by the relative scarcity of land and labor in a particular society (Hayami and Ruttan 1985). Population pressure has played a defining role in technology development and farmer choice of technologies. Ester Boserup (1965) argued that societies across the world have responded to increasing land scarcity in a strikingly similar and predictable manner, by increasing land productivity through agricultural intensification. The transition from low-yield, land-extensive cultivation systems to land intensive, double and triple crop systems has only been profitable in societies where the supply of uncultivated land has been exhausted (Boserup 1965, Hayami and Ruttan 1985, Pingali et al. 1987). The process of agricultural intensification has been observed in traditional as well as modern agricultural societies.

The movement from forest and bush fallow systems of cultivation to annual and multicrop cultivation systems, whereby plots of land are cultivated one or more times per year, has generally been influenced by rising population densities in traditional societies (Boserup 1965). It is no accident that the modern seed-fertilizer revolution has been most successful in densely populated areas of the world, where traditional mechanisms for enhancing yields per unit area have been exhausted.

Kikuchi and Hayami (1978) document the process of land augmentation as the cultivation frontier closes for Japan, Taiwan, Korea, and the Philippines. 'As population pressure pushes the cultivation frontier into marginal areas, we expect the marginal cost of agricultural production via expansion of cultivated area to rise relative to the marginal cost of production via intensification. Eventually the economy will reach a stage at which internal land augmentation becomes a less costly means of increasing agricultural output than external land augmentation' (Hayami and Ruttan 1985, p. 310).

In addition to high population densities, intensive cultivation will also be observed in areas with better access to markets, provided soil conditions are suitable. Intensification occurs in the less densely populated areas for two reasons: (a) higher prices and elastic demand for output imply that the marginal utility of effort increases, hence farmers in the region will begin cultivating larger areas; and (b) higher returns to labor encourage migration into well-connected areas from neighboring regions with higher

transport costs. In other words, where population densities are low, the returns to intensification are high only if markets, domestic and/or export, are easily accessible, such as the Central Plains of Thailand or parts of South America's Southern Cone. If these conditions are not present, labor and other costs associated with intensive agriculture are substantially higher than its incremental economic returns.

1. *Farmer-induced Technological Change in Traditional Farming Systems*

Pingali et al. (1987) documented, in the case of Africa, the movement from shifting cultivation to permanent agriculture with increases in population densities and improvements in market infrastructure. As land became scarce, traditional farming communities across sub-Saharan Africa extracted higher and higher levels of output from their land through investments in land improvements and soil fertility management. Erosion control, irrigation, and drainage are the most commonly made land investments for enhancing productivity. Tiffen et al. (1994) provided an excellent case study from the Machakos District of Kenya on the process of intensification and farmer investment in land improvements.

The extent of intensification is conditional on the relative responsiveness of the soils to inputs associated with intensive production such as land improvements, manure, and fertilizers. As population densities rise, one observes the cultivation of land that requires substantially higher labor input but that is also more responsive to the extra inputs. The responsiveness to intensification is generally higher on soils with higher water and nutrient holding capacity; these are usually the heavier, clayey soils, rather than the lighter sandy soils (Pingali et al. 1987). The movement from communal access, to individual property rights, with population growth induced increases in land values, provide individuals the incentive to invest in land improvements and other technologies necessary for intensive agricultural systems (Binswanger and Deininger 1997). Formal land ownership as characterized by the possession of a title also helps the farmer in acquiring credit for making the necessary investments in land and technology (Pingali 1990).

In using population densities as a basis for predicting the level of agricultural intensification, one ought to standardize arable land by soil quality and climate. Binswanger and Pingali (1988) provide a standardized population density measure: the number of people per million kilocalories of production potential. This measure is called the agroclimatic population density. When countries are ranked by agroclimatic population densities, Niger and Kenya turn out to be more densely populated than Bangladesh, and India ranks only twenty-ninth rather than seventh as in the traditional

population rankings. This explains why one observes intensive cultivation in some sparsely populated countries with poor market infrastructure—Niger and Botswana are examples.

2. Adoption of Modern Scientific Innovations

The intensification of traditional farming systems was a process that the more densely populated regions of Asia had been through several decades, and in some cases centuries, earlier. The application of modern scientific innovations in the form of high yielding seeds and fertilizer allowed the extremely land scarce regions of Asia achieve levels of land productivity that were not possible through the exclusive reliance on traditional farmer innovation systems.

A cross-sectional comparison across Asian countries shows that the proportion of irrigated area and the percentage of area under high yielding varieties of crops (HYVs) and fertilizer use is positively related to population densities (Pingali et al. 1997). It was not a historical accident that research for enhancing the yields of the two major food crops of Asia, rice and wheat, began in the 1950s. Research investments for enhancing yields were made in response to declining opportunities for area expansion, which was the main source of output growth in the pre-Green Revolution Asia. A comparison of sources of rice output growth during the early and late Green Revolution periods shows that the overall rate of growth was similar for the two time periods. In the early period, one-third of the growth came from an increase in cropped land, while in the latter period almost all of it was due to yield increases. The adoption of the Green Revolution technologies has been spotty in other parts of the world; this reflects in part the differences in land and labor endowments relative to Asia (Pingali and Heisey 1999). Research on induced intensification has not been restricted to the agricultural economics literature alone; a substantial number of comparative studies that look at the process of intensification, over time and over space, can be found in the geography–anthropology literature; see for example, Turner and Brush (1987) and Turner and Ali (1996).

3. Labor Use and Employment and Labor Saving Technologies

In both traditional and modern agricultural systems, intensification of agriculture leads to an increase in labor use per hectare, due to an increase in the number of tasks performed, such as irrigation, drainage, etc., and due to an increase in the intensity with which some tasks have to be performed, such as more intensive manure use and weeding (Pingali et al. 1987). Intensification therefore leads to an increase in agricultural employment as well as an increase in yields per hectare.

Labor productivity can only be sustained, as intensity rises, if labor saving technologies are available.

Pingali et al. (1997) have argued that the adoption of the plow in African agriculture has been closely associated with the level of intensification. Similarly, increases in labor productivity in Asia are associated with the increasing adoption of labor saving technologies, both as a means of saving on the higher labor needs due to intensification as well as a means of economizing on labor as its opportunity cost rises with economic growth. The movement from single-crop cultivation systems to a double- and triple-crop systems shifts out the demand for labor and also increases the demand for timely completion of operations. Initially, the switch to modern varieties was profitable, even with higher labor requirements per hectare, because the labor requirements per ton of paddy were lower compared with that of traditional varieties (Barker and Cordova 1978). Over time however, increasing labor demand for peak period operations led to a rise in real wages for these operations, even in densely-populated labor surplus countries such as India and Indonesia. The rise in wages was further exacerbated by the concurrent growth in the rural nonfarm sector and increasing employment opportunities in the urban sector, as has happened in Japan and Korea and now being witnessed in Southeast Asia.

Widespread adoption of labor-saving mechanical and chemical technologies has alleviated the growing labor constraints, and contributed substantially to overall productivity growth (Sidhu and Byerlee 1992). The most notable confirmation of this hypothesis comes from the now developed country, Japan, where graphs plotting the log of land productivity vs. the log of labor productivity show almost no increases in land productivity but substantial gains in labor productivity in the last 30 years (Hayami and Ruttan 1985, Craig et al. 1997).

4. Intensification and Degradation—When Does the System Break Down?

The process of population induced technological change is by no means automatic; there are several instances where societies faced with growing population pressure have not been able to achieve sustainable intensification. Failures in agricultural intensification can be attributed to one or more of the following reasons: persistence of uncertain long-term rights to land; encroachment of cultivation onto marginal lands; collective effort required for watershed-level protective investments; and an inappropriate policy environment.

Farmer incentives for regenerating soil fertility and for making erosion-prevention investments are high if secure long-term land rights exist, both to currently cultivated land and to fallow land. Where farmers do

not have long-term rights to land, the private rate of discount is higher than the social rate of discount, hence private investments in land productivity will be lower than the societal optimum. In societies with slowly and steadily growing populations, property rights to land evolved systematically, being induced by growing land scarcity. However, the evolution to secure land rights has not been successful in societies experiencing rapid population growth (through natural increase or through migration), nor in societies where this induced institutional change is circumvented by government policies (e.g., land expropriation by colonial authorities or socialist governments).

Given secure rights to land, the rate of return to land investments varies by agroclimatic zone and soil type: degradation problems are most severe where the returns to land investments are lowest. The arid fringe areas, upper slopes in the semi-arid and the humid zones, and shallow sandy soils exhibit the highest levels of erosion, other things being equal. Degradation problems are likely to be most severe in regions with relatively large endowments of marginal lands.

Degradation of marginal lands is also likely to be severe where institutional arrangements are not available for controlling access to grazing and forestry. Where privatization is infeasible or undesirable, community action or government intervention is required both for control of access and for investments in erosion control and forest replanting. Difficult dilemmas frequently arise as traditional users are threatened with a loss of their use rights. Societies are often incapable of solving these problems of ownership or controlled access before serious damage has occurred.

In several instances, inappropriate policies, both micro- and macroeconomic, have created disincentives for farmer investments in land productivity enhancing and/or resource conserving technologies (Heath and Binswanger 1996). Pingali et al. (1997) argue that even in the intensive lowland cereal production systems, where Green revolution technologies were most successfully adopted, price and trade policy distortions are leading to significant levels of environmental and ecological degradation.

The potential problems of declining labor productivity and environmental degradation are not problems of levels of population densities. Given sufficient time, it is likely that a combination of farmer innovations, savings, and the development of research facilities and institutions for dealing with soil degradation issues will be able to accommodate much more than the current population in most countries, especially in many of the less densely populated ones. However, if all these changes are required quickly and simultaneously because of high population growth rates, they may emerge at too slow a pace to prevent a decline in human welfare.

See also: Agricultural Change Theory; Agricultural Sciences and Technology; Agriculture, Economics of;

Development and the State; Development: Rural Development Strategies; Development: Sustainable Agriculture; Food Production, Origins of; Industrialization; Industrialization, Typologies and History of; Innovation and Technological Change, Economics of; Innovation, Theory of; Technology Transfer

Bibliography

- Barker R, Cordova V 1978 Labor utilization in rice production. In: *Economic Consequences of the New Rice Technology*. International Rice Research Institute, Los Baños, Philippines
- Binswanger H, Deininger K 1997 Explaining agricultural and agrarian policies in developing countries. *Journal of Economic Literature* 35: 1958–2005
- Binswanger H, Pingali P 1988 Technological priorities for farming in sub-Saharan Africa. *World Bank Research Observer* 3: 81–98
- Boserup E 1965 *Conditions of Agricultural Growth*. Aldine Publishing Company, Chicago
- Craig B J, Pardey P G, Roseboom J 1997 International productivity patterns: accounting for input quality, infrastructure, and research. *American Journal of Agricultural Economics* 79(4): 1064–76
- Hayami Y, Ruttan V W 1985 *Agricultural Development: An International Perspective*, rev. and expanded edn., Johns Hopkins University Press, Baltimore, MD
- Heath J, Binswanger H 1996 Natural resource degradation effects of poverty and population growth are largely policy-induced: the case of Colombia. *Environment and Development Economics* 1: 65–83
- Kikuchi M, Hayami Y 1978 Agricultural growth against a land resource constraint: a comparative history of Japan, Taiwan, Korea, and the Philippines. *Journal of Economic History* 38: 839–64
- Pingali P 1990 Institutional and environmental constraints to agricultural intensification. In: McNicoll G, Cain M (eds.) *Rural Development and Population (Institutions and Policy)*. Population and Development Review, a supplement to Volume 15, 1989. Oxford University Press, Oxford, UK
- Pingali P, Bigot Y, Binswanger H 1987 *Agricultural Mechanization and the Evolution of Farming Systems in Sub-Saharan Africa*. Johns Hopkins University Press, Baltimore, MD
- Pingali P, Heisey P W 1999 *Cereal Crop Productivity in Developing Countries*. CIMMYT Economics Paper 99-03, CIMMYT, Mexico
- Pingali P, Hossain M, Gerpacio R V 1997 *Asian Rice Bowls: The Returning Crisis?* CAB International, Wallingford, UK
- Sidhu D S, Byerlee D 1992 *Technical Change and Wheat Productivity in the Indian Punjab in the Post-Green Revolution Period*. CIMMYT Economics Program Working Paper 92-02, CIMMYT, Mexico
- Tiffen M, Mortimore M, Gichuki F 1994 *More People, Less Erosion: Environmental Recovery in Kenya*. Wiley, Chichester, UK
- Turner B L II, Ali A M S 1996 Induced intensification: agricultural change in Bangladesh with implications for Malthus and Boserup. *Proceedings of the National Academy of Sciences* 93: 14984–91
- Turner B L II, Brush S B (eds.) 1987 *Comparative Farming Systems*. Guilford Press, New York

P. Pingali

Population Composition by Race and Ethnicity: North America

In the three nations of North America—Canada, the United States, and Mexico—ethnicity has different connotations and is officially recorded using a variety of measures. Weber's (1978) definition of ethnic groups as those that share 'a subjective belief in their common descent because of similarities of physical type or of customs' permits various attributes to be used to characterize ethnic boundaries. Not only do these traits vary from place to place, but they change over time as well. In contemporary Mexico, the primary ethnic distinction is between the indigenous and non-indigenous populations. Statistics Canada records national origin, minority and aboriginal status, and religion. Finally, the US Census Bureau solicits race and Hispanic origin, and for a subset of the population, inquires about ethnic ancestry.

1. Historical Perspective of North America's Ethnic Composition

The peopling of North America has been shaped by the presence of indigenous Americans, European colonization, African slavery, voluntary immigration, and for Mexico, considerable emigration. Along with differential fertility and mortality rates, these factors have dramatically transformed the continent's ethnic landscape over time.

The size of the indigenous population of North America at the time of European colonization is unknown. Estimates for the pre-Contact population of Canada and the US range from 900,000 to 18 million (Snipp 1989), while similar estimates for Mexico range from 4.5 million to 25 million (Alba 1977). Despite the lack of consensus about fifteenth-century population levels, there is widespread agreement that in the wake of European colonization, indigenous depopulation was severe. Although 200,000 Native people were estimated to inhabit Canada when the French began settlement in the early seventeenth century, less than 100,000 remained by 1867 (Anderson and Frideres 1981). In 1548, 6 million indigenous people were estimated to inhabit Mexico, but by 1605, only 1 million remained (Alba 1977). Depopulation estimates for North America as a whole have been put at 1.75 million to 4.75 million between 1500 and 1890 (Snipp 1989).

As Native American population levels dropped, the arrival of Europeans, and to a lesser extent Africans, began a repopulation process that transformed the demography of North America. Until the twentieth century, Canadian immigration would be dominated by the arrival of English and French settlers, while in the US, English colonization followed by German, French, Dutch, Irish, and Scotch-Irish immigration

was accompanied by the coercive importation of African slaves. Although Africans were a feature of colonial Mexican society, their shares were extremely small and highly concentrated geographically.

Despite some similarities in their peopling processes, the three nations of North America have experienced distinct immigration trends following the European colonization. In 1870—soon after Confederation and almost 400 years after Christopher Columbus arrived in the Americas—Canada's population was overwhelmingly European-origin; 61 percent of approximately 3 million claimed roots in the British Isles and 31 percent reported French ancestry (Anderson and Frideres 1981). In contrast, the US population of over 38 million was more ethnically diverse: nearly 13 percent were 'colored' (i.e., black) (US Census Office 1883).

Three waves of immigrants diversified Canada's population during the twentieth century. Around the turn of the century, the European origins of Canadian immigrants changed, but remained overwhelmingly white. From 1896 to 1914, 3 million new immigrants responded to government incentives to settle the western prairie provinces, and the majority were German, Slavic, or Scandinavian. In a second wave from approximately 1946 to 1961, Canada absorbed 2 million newcomers, mainly Europeans who were fleeing the aftermath of World War II. Finally, in the wake of immigration reforms after 1970, the share of European immigrants dropped sharply from 76 percent in 1966 to 39 percent in 1973, while arrivals from Asia jumped from 6 percent to 23 percent (Anderson and Frideres 1981).

Like Canada, the volume and national origin composition of US immigration flows has varied over time. In the early years of the republic, the population was largely of English, German, Irish, and African origin. When the legal importation of slaves was halted in 1808, the black population stood at over 16 percent of the total (US Census Bureau 1976). Concurrent and continued immigration for Italians, Russians, Poles, and Scandinavians boosted the white population during the nineteenth and twentieth century. Following the 1965 Amendments to the Immigration and Nationalization Act, immigrants from Asia and Latin America have dominated flows. As a result, the US has become both less white and black than it was at its founding.

Mexico's demographic evolution has differed in several ways from its neighbors to the north. After the decline of the indigenous population that characterized the early colonization, and the subsequent period of Spanish settlement and African slavery, the Mexican population grew slowly by natural increase, as it was untouched by the large-scale immigration to the US and Canada. From the sixteenth through the seventeenth centuries, admixture between Europeans, Indians, and Africans produced the *mestizo* population that characterizes Mexico today. Both fertility

and mortality rates were high throughout this period, and remained so well into the twentieth century. At the beginning of the nineteenth century, when the colonial period came to a close, Mexico's total population stood at 6 million—roughly equivalent to the indigenous population prior to colonization (Alba 1977). Mexico's indigenous population rebounded following a decline between 1900 and 1920 (as did the national population in the wake of the Mexican Revolution, emigration, and an influenza epidemic).

2. Contemporary Population Composition in North America

Recent changes in the ethnic composition of North America derive from three processes, namely immigration, intermarriage across race and ethnic lines, and differential fertility rates of natural increase. Inconsistent reporting categories over time preclude precise estimates of the magnitude and components of change, but general inferences about the relative importance of each are possible. Accordingly, we discuss the contemporary ethnic landscape of Canada,

the US, and Mexico and conclude with a broad comparison of the primary axes of ethnoracial differentiation among the three countries.

2.1 Canada

Table 1, which displays Canada's population composition in 1961 and 1996, reveals both increasing ethnoracial diversification over the 35-year period and also persisting dominance of European ancestries. Despite comparability problems over time, the two most salient differences are the absence of 'Canadian' as an ethnic category in 1960, and the allowance for multiple ancestries. In 1996 nearly 18 percent of all residents identified as 'Canadian'; this figure approaches one in five if the aboriginal population is added. Between 1961 and 1996 the aboriginal population increased its share slightly, but remains a tiny share of the total population. In 1961, virtually all Canadians (97 percent) traced their origins to a European ancestry, and of these, nearly three in four identified as either British or French origin. As recently as 1996, the majority of ethnic Canadians identify with

Table 1
Canada: population by ethnic origin, 1961–96

Ethnic origins	1961		1996	
	<i>N</i>	%	<i>N</i>	%
<i>Single origins</i>	18,238,247	100.0	20,272,095	66.5
British Isles	7,996,669	43.8	3,267,520	10.7
French	5,540,346	30.4	2,683,840	8.8
Other European	4,116,849	22.6	3,742,895	12.3
Western European			1,126,095	3.7
Northern European			167,285	0.5
Eastern European			867,055	2.8
Southern European			1,376,935	4.5
Other European			205,525	0.7
Arab			188,435	0.6
Asian	121,753	0.7	1,968,465	6.5
West Asian			106,870	0.4
South Asian			590,145	1.9
East and Southeast Asian			1,271,450	4.2
African	32,127	0.2	137,315	0.5
Pacific Islands			5,765	0.0
Latin, Central and South American			118,640	0.4
Caribbean			305,290	1.0
Aboriginal	220,121	1.2	477,630	1.6
Canadian			5,326,995	17.5
Other	210,382	1.2	80,840	0.3
<i>Multiple origins</i>	<i>NA</i>	<i>NA</i>	10,224,495	33.5
Total	18,238,247	100.0	30,496,590	100.0

Sources: Statistics Canada, *Population by ethnic origin, 1996 Census*, www.statcan.ca.
Dominion Bureau of Statistics, 1961 Census of Canada, Population: Ethnic Groups (Bulletin 1.2–5).
Notes: From the 1961 data, the category 'Negro' has been assigned to African origins.

Table 2
United States: population by race and Hispanic origin, 1960–99

Group	1960		1990		1999	
	Total (thousands)	Percentage of population	Total (thousands)	Percentage of population	Total (thousands)	Percentage of population
White	151,932	84.7	188,307	75.7	196,043	71.9
Black	18,872	10.5	29,299	11.8	33,088	12.1
Hispanic ^a	6,900	3.9	22,372	9.0	31,265	11.5
Asian and Pacific Islander ^b	878	0.5	6,992	2.8	10,219	3.7
American Indian, Eskimo, and Aleut	524	0.3	1,796	0.7	2,022	0.7
Other	218	0.1	—	—	—	—
Total	179,323	100.0	248,766	100.0	272,637	100.0

Sources: US Bureau of the Census, Census of Population 1960, Vol. 1: Characteristics of the Population, Part 1: United States Summary, Table 56. US Bureau of the Census, *Resident Population Estimates of the United States by Sex, Race, and Hispanic Origin: April 1, 1990 to June 1, 1999*.

Notes: a The estimate of the 1960 Hispanic population is from Cary Davis, Carl Haub, and JoAnne Willette, 1983. 'US Hispanics: Changing the Face of America.' *Population Bulletin*, Vol. 38, No. 3, p. 8, Table 2.

b For 1960, the Asian population comprises the categories 'Japanese,' and 'Chinese,' and 'Filipino' only. In subsequent years, Koreans, Vietnamese, Asian Indians, Samoans, Guamanians, and Hawaiians were also included.

European ancestries; nearly half (48 percent) of all single origin ancestries reported in 1996 were from Europe.

Among Canadians who report a single ancestry, the major change in ethnic composition derives from shifts in the source countries of immigrants. Prior to 1961, fewer than 1 percent of Canadian residents derived from the Middle East, Latin America, or Asia, compared to 19 percent by 1996. Canadian immigration from these regions is relatively recent, which explains their small population shares at the turn of the twenty-first century. Among the million foreign-born Canadian residents admitted before 1961, less than 5 percent were from Asia, Africa, or Latin America. But, as the volume of immigration to Canada increased, so too has the representation of these groups among new arrivals. Should current immigration trends continue, Canada's ethnic mosaic will be very different in the middle of the twenty-first century. However, the homogenization of Canadians through assimilation and intermarriage may also swell the numbers who identify in nonethnic terms as 'Canadian.'

2.2 United States

Table 2 presents the ethnoracial composition of the US in 1960 and 1990 based on the last two censuses, and for 1999 based on Current Population Survey estimates. The broad groups are the official categories approved by the Office of Management and Budget. Like Canada, the US racial classification scheme evolved to reflect the rapid growth of the Hispanic population and increasing diversification of Asian populations. Although Hispanics can be of any race,

the vast majority is considered white because *mestizo* has not been an acceptable racial category for administrative reporting purposes. In 1960, the Asian and Pacific Islander category included only Japanese, Chinese, and Filipinos, but when the racial classification was revised and expanded in 1980, Vietnamese, Asian Indians, Samoans, Guamanians, and Hawaiians were added as separate racial groups. In 1990, the US Census identified nearly 1,000 unique racial and ethnic codes, which included nearly 400 American-Indian tribes, hundreds of detailed Asian and Pacific Islander groups, and many more self-identified nationalities.

Changes in the US racial classification system notwithstanding, trends in ethnic composition are clear. The white share of the total population has been falling steadily, from 85 percent in 1960 to 76 percent in 1990 and 72 percent in 1999; the black share grew slowly from 10.5 percent in 1960 and remained stable around 12 percent during the 1990s, while Hispanics and Asians increased their population shares. Census projections indicate that the Hispanic population will surpass blacks by 2003 or sooner. Less than 1 percent of the 1960 US population was Asian, but by 1999 Asians comprised almost 4 percent of all US residents—a seven-fold increase since 1960. Although Asians are the fastest growing minority population in the US, their population share remains low for historical reasons, primarily due to laws restricting their admission as immigrants until 1968. Compared to blacks, Hispanics and Asians, the indigenous population of the US remains tiny. In 1960 less than 0.3 percent of the total population consisted of indigenous groups, and by 1990 this share had increased to 0.7 percent. Part of this increase is an artifact of changed reporting criteria that increased incentives and opportunities for individuals of mixed

Table 3
Ten most frequently cited ancestries, 1990

Ancestry ^a	Percentage of reported ancestries
German	19.7
Irish	13.1
English	11.0
Afro-American	8.1
Italian	4.9
American	4.2
Mexican	4.0
French	3.5
Polish	3.2
American Indian	3.0

Source: University of Minnesota, Integrated Public Use Microdata Series, 1990 Census 1 percent Sample

a The ancestry data includes multiple-ancestry reports. A person who described herself as 'French' and 'English' would be counted twice, once under each ancestry listing.

racial origin to self-identify as Native American (Snipp 1989).

Because the US reports race and ethnic origin separately, ancestry data are instructive about the impact of historical and recent immigration trends on the ethnic landscape. Table 3 reports the 10 most frequently reported ancestries in 1990. Multiple responses are double counted; therefore the white population is more heavily weighted because they tend to report more ancestries (resulting from intermarriage) and because inter-racial marriages are typically rare. Germans, Irish, and English are the top three most frequently reported ancestries, followed by African-Americans and Italians. Only 4.2 percent of the US population identified as 'American,' but this figure would be higher if multiple ancestries were not double

counted. The prominence of Mexicans among the top 10 ancestries reflects the dramatic growth of this population through immigration since 1960. Between 1970 and 1995, for example, the Mexico-born population of the US increased eight-fold, rising from 750 thousand to 6.2 million. Currently, Mexico is the largest single source of US immigrants.

Census Bureau projections using moderate assumptions about mortality, fertility and intermarriage patterns and current immigration levels indicate that by the year 2020, 65 percent of the US population will be white and 13 percent black. However, the two groups most affected by immigration, Hispanics and Asians, are projected to comprise 16 and 7 percent, respectively, of the total population. This is a marked change from 1960, when Hispanics and Asians combined made up less than 5 percent of the total. In recognition that intermarriage will continue to blur ethnic and racial boundaries, plans for the US Census 2000 will allow Americans to identify themselves using multiple races.

2.3 Mexico

Mexican ethnoracial categories are less differentiated than either Canada or the US mainly because it is a country of emigration rather than immigration. In Mexico the key racial division is between the indigenous and *mestizo* populations, and use of an indigenous language is the only measure the government now uses to draw this racial distinction. This is because designating members of indigenous groups hinges much more on sociocultural characteristics (like language) than phenotype or precise blood lineages, as practiced in the US.

Table 4
Mexico: indigenous language speakers, 1900–90 population aged 5 and above

Year	Spanish	Indigenous language*	Total population	Percentage indigenous language
1900*	11,528,345	2,078,914	13,607,259	15.3
1910**	13,200,063	1,960,306	15,160,369	12.9
1930	11,632,660	1,185,162	12,817,822	9.2
1940	18,410,534	1,237,018	19,647,552	6.3
1950	19,373,423	795,069	20,168,492	3.9
1960	33,818,174	1,104,955	34,923,129	3.2
1970	45,113,823	3,111,415	48,225,238	6.5
1980	61,665,795	5,181,038	66,846,833	7.8
1990	65,279,855	5,282,347	70,562,202	7.5
1995	74,735,782	5,483,555	80,219,337	6.8

Sources: Instituto Nacional de Estadística, Geografía e Informática, Estadísticas Históricas de México, Volume 1 (1985), Table 2.4.

INEGI, *Censo '95* (1995), Table 6.

INEGI, *XI Censo General 1990* (1992), Table 6.

Pick, James G. and Edgar W. Butler. 1994. *The Mexico Handbook* (1994).

Secretaría de Industria y Comercio, Dirección General de Estadísticas, *VII Censo General de Población 1950* (1952), Table 6; *VIII Censo General 1960* (1962), Tables 3, 14 and 15; *IX Censo General 1970*, Tables 4, 17; *X Censo General 1980*.

* From 1950 forward, the 'indigenous language speakers' includes both those who do and do not speak Spanish. In earlier years, it is unclear if both groups are included.

** The 1900 and 1910 figures are for the entire population, not just those aged 5 and above.

Table 4 provides a glimpse of Mexico's indigenous population and its evolution since the turn of the twentieth century. The series reported in Table 4 is not totally consistent because the criteria used to enumerate the indigenous populations at times were restricted by age (5 years and over) and at times combined speakers of indigenous languages who did and did not speak Spanish. In 1990, the majority of the indigenous population spoke over 11 different languages, but Nahuatl, Maya, and Zapoteco were the three most commonly spoken. Mexico's indigenous population is concentrated in the southern and least developed states, with Quintana Roo, Yucatan, Campeche, Chiapas, Oaxaca, Guerrero, Puebla, Hidalgo, and Tlaxcala housing the vast majority of indigenous peoples.

Despite changes in enumeration criteria, the historical record shows a clear decline in the share of indigenous people in Mexico's population. The indigenous population has grown strongly in absolute terms, but this is due partly to improvements in enumeration and coverage over the period. At the turn of the century, 15 percent of Mexicans were indigenous language speakers but by 1995 this share had been halved. In relative terms, the indigenous population of Mexico is shrinking. Mexico's declining linguistic diversity is due to social and economic integration of indigenous minorities through internal migration, urbanization, and modernization processes. Further shrinkage of Mexico's indigenous population is likely in the future, but linguistic diversity is likely to persist much longer than in Canada and the US owing to the larger population base and the persisting rural isolation of many groups.

2.4 Comparative North American Perspective

Canada, the US, and Mexico have some similarities and many differences in the composition and evolution of their ethnoracial landscapes. First, despite rising immigration to Canada and the US from Asia, Africa, and Latin America, both countries have witnessed an increase in the population shares that do not self-identify in ethnic terms, but report themselves as Canadian or American. Second, by comparison to Mexico, Canada and the US have much smaller indigenous populations. Third, unlike the US, Canada and Mexico have negligible shares of residents who trace their ancestries to Africa; owing largely to the legacy of slavery, blacks comprise over 12 percent of US residents. Finally, immigration has diversified the ethnoracial landscape of Canada and the US since the 1960s, but Mexico has witnessed a negligible impact of foreign admissions. As an emigrant country, Mexico's population composition will be driven by the assimilation of indigenous populations into the national population, a process that is shaped largely by social and economic change, along with internal migration and intermarriage patterns.

See also: Ethnic Groups/Ethnicity: Historical Aspects; Ethnicity: Anthropological Aspects; Ethnicity, Race, and Health; Ethnicity, Sociology of; Gender, Class, Race, and Ethnicity, Social Construction of; Population Cycles and Demographic Behavior; Population Ecology; Race and Urban Planning; Race: History of the Concept; Statistical Systems: Censuses of Population

Bibliography

- Alba F 1977 *La Población de México: Evolución y Dilemas*. El Colegio de México, Mexico City
- Anderson A B, Frideres J S 1981 *Ethnicity in Canada: Theoretical Perspectives*. Butterworth & Co, Toronto
- Bean F D, Tienda M 1987 *The Hispanic Population of the United States*. Russell Sage Foundation, New York
- Davis C, Haub C, Willette J 1983 US Hispanics: Changing the face of the nation. *Population Bulletin*, Vol. 38, No. 3. Washington, DC: Population Reference Bureau
- Instituto Nacional de Estadística, G. e. I. 1996 *Conteo de Población y Vivienda 1995: Resultados Definitivos, Tabulados Básicos*. Aguascalientes
- Smith J P, Edmonston B (eds.) 1997 *New Americans: Economic, Demographic and Fiscal Effects of Immigration*. National Academy Press, Washington, DC
- Snipp M C 1989 *American Indians: The First of This Land*. Russell Sage Foundation, New York
- US Census Office 1883 Compendium of the Tenth Census (June 1, 1880). Washington, DC
- US Census Bureau 1976 *The Statistical History of the United States From Colonial Times to the Present*. Basic Books, New York, Tables A 1–5 and A 91–104
- Weber M 1978 *Economy and Society*. University of California Press, Berkeley, CA

M. Tienda and A. Morning

Population Cycles and Demographic Behavior

1. Population Cycles in the Context of Intellectual History

Population studies and economic theory have been developed and discussed in conjunction with one another since the early nineteenth century. At this time Thomas Robert Malthus devised the ingenious theory, influenced by religion and controversial in its time, that bears his name. Ever since then it has been assumed within the social sciences almost as a matter of course that a strong linkage exists between demographic and economic factors. Malthus elaborated his classical theory around the year 1800. There were few new theoretical contributions for a long time afterwards, and those that came forth were not especially

significant. Until well into our own day, therefore, Malthus' basic ideas dominated the scene. According to his theory the causal relation between population and the economy goes chiefly in one direction—population changes play only a passive role and are determined by changes in the means of subsistence. Furthermore, in his system a single demographic variable, i.e., mortality, is mainly influenced by the economy. Although Malthus points out that moral restraint and the postponement of marriage may counteract excessively strong population growth, mortality is the dominant force adapting population size to the existing means of subsistence.

By contrast, theoretical developments have proliferated rapidly during recent decades. Entirely new models have emerged to elucidate the interaction between the population and the economy. This has given rise to population economics as a new and vigorously expanding field of scholarly endeavor. Its reasoning is usually strongly rooted in traditional economic theory and often bears the imprint of Malthusian thinking. These developments have meant that the shaping of theory has taken quite another turn and has also become more diversified, with much greater breadth and depth than before. The major characteristics of modern population economics are discussed below.

First, interest has shifted from one demographic variable to another. The strong link which Malthus believed to exist between changes in the economic environment and mortality has been called into question. Mortality variations are influenced by a number of other factors, the means of subsistence and standard of living forming only one of these, and not always the most important one. Fertility—along with the factors which determine it, of which particular mention may be made of varying age at marriage, the use of birth control, abortion, and nutrition—has been regarded instead as being determined by economic factors in a way quite different from that found in Malthus. This does not apply only to modern society. Several studies have shown that even in preindustrial society, the link between fertility and economic factors was often stronger than the relation between mortality and the economy, in both the short and the long term.

Second, population variables have been linked more closely to economic theory, especially neoclassical microtheory, and have then usually been analyzed in a more stringent manner than is the case with the Malthusian theory. Here we may point to the 'micro-economic theory of fertility' or, as it is sometimes termed, the 'new household economics' or 'new home economics.' In this theoretical construction, fertility is viewed as the result of economically rational decisions (the demand for children). Established economic methods and conceptual mechanisms are employed in order to explain variations in fertility. For example, models are devised in which fertility is regarded as a function of income and the 'price' of children in

relation to the price of other goods and services. In this reasoning children are regarded as ordinary goods, exchangeable in the choice situation for other goods and services. Fluctuations in fertility are explained using only prices and incomes. The pioneering theoretical work in this area was done by Gary Becker (1960) and was then elaborated, for example by Becker and Lewis (1973); it has often been cited as one of the basic explanations of fertility variations in the Western world since World War II. Models have also been formulated along highly neoclassical lines and have been based on a traditional investment approach: cost-benefit analysis. This method regards fertility as being determined by how costs and revenues associated with children change over time. From this standpoint, accordingly, the strong decline of Western fertility which set in at the end of the nineteenth century should be regarded as a consequence of the falling revenues and rising costs which having another child would entail for the parents.

Third, Malthus' view that there is a one-way causality between the population and the economy has been abandoned. What is now emphasized is that the interaction between the population and the economy is considerably more complex than Malthus believed. In many instances, demographic events can be one of the most important factors underlying the process of economic change. In other words, the population factor may also, in many cases, be the independent variable, which in its turn influences the economy and the means of subsistence. According to this approach, numerous other elements of social change must be viewed in the context of changing population size and composition (e.g., Boserup 1965, Simon 1977, Easterlin 1980).

Fourth, the various models which have been developed to represent the interaction between the population and the economy have become considerably more dynamic in character. One important element marking this trend has been the wide-ranging research conducted into the incidence and causes of various population cycles and into their effects on social change. Intensive discussions have been conducted concerning the extent to which population cycles with varying durations and strengths can be identified; what demographic or other factors can be cited as explanations of their genesis; and what effects the various population cycles have had on demographic behavior and on the national economy as a whole.

2. The Mechanisms of Population Cycles

The study of population cycles goes a long way back in time. Its origins can be traced back to the northern European countries in the nineteenth century. As early as in 1855 Ejler Sundt, a Norwegian embryonic sociologist and statistician, observed that a strong rise

and fall in a country's birthrate is repeated during the next generation, though in a less marked fashion. On reaching marriageable age an extraordinarily large birth cohort will bring into the world more children than an age group of normal size. The same law applies in reverse: after one generation an undersized cohort will produce a new undersized group. As a result, regularly recurring waves or 'echo effects' are generated in the size and age composition of the population.

Ten years later, in 1865, the Swedish demographer Fredrik Berg paid tribute to Sundt for the idea of cohort analysis, and emphasized the importance of keeping different birth cohorts apart and studying them separately. He also argued that it ought to be possible to use variations in birth cohort sizes to compute the future number of births. Berg foreshadowed modern population projection techniques.

The Swedish demographer Gustav Sundbärg was a pioneer in expressing the view that a substantial change in birthrates may have far-reaching repercussions, not only on the economy but also on institutional and social change for a long time ahead. He may also have been the first to point out that *the size of a cohort* can have an impact not only on individuals' demographic behavior, but also on their behavior in numerous other spheres. Sundbärg developed these ideas in his analysis of the strong rise in the number of births in Sweden between 1815 and 1830, and its after-effects. Little by little the rise in births transmitted itself to higher age levels and grew in strength compared with earlier cohorts because mortality was also falling dramatically. Sundbärg believed that this rise had a far-reaching influence in many areas of social and economic development in Sweden all through the nineteenth century. He wrote in 1903:

As they reached school age these cohorts filled up all the school rooms so that their education and upbringing became a burning issue. As they reached working age they shattered the guilds (in the 1840s) and compelled the introduction of a more liberal trade legislation, without which these masses of people would have been unable to survive. As they reached marriageable age their large numbers enforced delayed marriage, and in the end (though not until the second generation) the great emigration movement. The immense class of 'all others' who were excluded from democratic representation at the time destroyed the constitution based on the four estates (in the 1850s and 1860s). And in our own day, having reached old age they fill our poorhouses and make the problem of disability and old age a matter of topical concern.

In 1937, the German economist August Lösch noticed that sudden and dramatic changes in the population structure, caused by war for example, generated long waves in population growth, which in turn brought about far-reaching changes in the economy. Great wars are characterized by a deficit of births during the war and a surplus of births in the immediate postwar period. In a now-classical article in the

Quarterly Journal of Economics, 'Population Cycles as a Cause of Business Cycles,' Lösch described how these features caused a population cycle which in turn gave rise to strong fluctuations in German economic growth and in the business cycle.

This discussion of the origins of population cycles and their influence on demographic and economic factors transmitted itself to the American scholarly arena after World War II. American researchers had been able to identify long cycles in US economic activity since the middle of the nineteenth century. How these cycles were generated and propagated became the object of extensive empirical and theoretical studies by American scholars during the 1950s and 1960s. They found that rather than being a passive response to changes in the economy, population cycles played an active role in determining the long cycles that had been observed in US economic activity. These long cycles could be found not only in aggregate production growth but also in most of its sub-components. This made it possible to identify long cycles in the growth of the labor force, in capital formation, in productivity, and in utilization of capacity. In particular, there was an extensive discussion on the existence of, and mechanisms underlying, Kuznets cycles—named after Simon Kuznets—or 'long swings,' as they are often called. Models were devised to show how changes in the age structure and population growth over about 20 years generated long-term fluctuations in investment activity, mainly in building construction, and corresponding fluctuations in economic growth. In explanations of the underlying mechanisms the primary stress was laid on the interaction between demographic and economic factors. In this way long cycles in marriage frequency, birthrate, and immigration were seen to give rise to long cycles in both the growth and the utilization of resources through their effect on such factors as capital formation, housing construction, and labor supply. The causal relation in the reverse direction was also important. Through the link with unemployment and effective demand, changes in the utilization of resources intensified immigration and gave rise to long-term cycles in the pattern of immigration. In the same way, the variations in unemployment influenced job opportunities for the labor reserve (mainly married women, ethnic minorities, and older workers). In other words, there was a long-term cycle in the size of the labor supply. A side result of this wide-ranging discussion of the interaction between demographic and economic factors was that economists increasingly incorporated demographic variables into their reasoning.

After World War II the long-term cycles assumed a different character both in duration and strength. A number of the suppositions applied to earlier periods of history were changed. Some of the mechanisms which formerly generated Kuznets cycles changed character and others weakened or vanished. In par-

ticular, the great influence that immigration to the USA had exerted in the generation of Kuznets cycles after 1850 was modified. When immigration was restricted during the interwar period, its cycle-generating function vanished. Instead, this function was replaced by the marked rise in the number of births during the 1940s and 1950s and by an equally marked decline during the next two decades (e.g., Kuznets 1958, Abramovitz 1961, 1968, Easterlin 1968).

2.1 The Easterlin Hypothesis

One of the most prominent figures in this earlier research into the genesis of population cycles and their influence on demographic and economic phenomena was the American economic historian Richard Easterlin. It was also Easterlin who drew attention to the changed demographic conditions that applied to the Kuznets cycle after World War II. It therefore seems quite appropriate that it would be Easterlin who made the most trail-blazing analysis of the causes of the dramatic changes in births in the USA after World War II and their effects on the demographic and economic behavior of individuals. A number of other features of modern social change must also be viewed, he argued, in the context of demographic events.

A group of people born during a given period are marked by the same code of values and attitudes in many respects. Easterlin believes it possible to identify an important underlying reason why values and attitudes may vary from one cohort to another. Accordingly he asserts that an individual's behavior in a number of areas will be influenced by the size of the cohort to which he or she belongs, since it determines economic situation and future prospects to such a large extent. At this point Easterlin ties in with earlier thinking by Gustav Sundbärg and others, but he conducts the analysis with considerably greater precision and stringency (e.g., Easterlin 1968, 1978, 1981, Artzrouni and Easterlin 1982.)

For young individuals who belong to a large cohort, prospects for the future appear somber. Their incomes will be lower in general and the likelihood of becoming unemployed is higher than for individuals belonging to a small cohort. As a consequence they marry later and also divorce more frequently. Besides they have a greater propensity to commit suicide and to manifest higher frequency of criminality. Married women in a large cohort will participate in the labor force more often because of the family's difficult economic situation. Unmarried women have high nonmarital fertility, also resulting from social stress factors. Alcoholism is also regarded as more common in a large cohort. Values, too (political opinions, attitude to work, savings, etc.), depend on cohort size.

One key element of Easterlin's reasoning is the 'relative income hypothesis.' This holds that the individual's behavior is determined by the relation

between, on the one hand, the material standard and position in society which he or she has attained, and on the other hand, what he or she had expected. The expectations are in high degree determined by family background, i.e., by the parents' economic situation, while the possibilities of translating them into reality are dependent on the prevailing labor market situation and relative wages. In other words, it is not the absolute level of living standard that determines behavior, but the relative level, in which the crucial factor is the relation between expectations and the possibilities of realizing them (material aspirations). During a period when high aims are set but the possibilities of realizing them are low, as is the case with a large cohort, the social stress factors are magnified. The stresses and strains associated with large cohort size are likely to continue throughout the whole course of a cohort's life.

In a small cohort the reverse applies. Firms have to compete for the relatively few new entrants to the labor market, which means that starting wages become high and opportunities of advancement good. Social stress is reduced considerably, and divorce, suicide and criminal frequencies become low. Moreover, in a small cohort expectations are low, because they are determined by the parents' (the preceding large cohort's) economic situation. On the other hand the possibilities of realizing expectations—and even of surpassing them—are high. Here, too, it happens that these favorable circumstances follow the small cohort all through its life.

Easterlin makes a particular point of analyzing how cohort size influences childbearing. He observes that a large cohort will bear relatively few children per woman while a small cohort will give birth to many. The linkage between birthrates and adverse economic and social effects causes the emergence of 'crowding mechanisms' operating within three major social institutions: family, school, and labor market. The result of the 'crowding mechanisms' is that the appearance of social stress factors are quite different in large and small cohorts. A large cohort adapts in order to avoid deterioration of its living standard compared with that of its parents. One such adaptation is that family formation takes place later in life. Another is that a woman in such a situation finds herself compelled to combine labor force participation with childbearing, which reduces her natality. By means of such demographic adaptation the large cohort strives to maintain its economic status compared with that of its parents. For a small cohort the situation is again reversed. Its members have no need to make any demographic adaptation in order to preserve an economic status that is favorable from the outset.

These demographic mechanisms produce what is called the 'Easterlin effect,' and they have been used as an alternative to the explanation suggested by the 'microeconomic theory of fertility' for the large fertility variations observed in the USA and Western

Europe during the postwar period. The most essential difference between the two approaches is that, according to the Easterlin model, demographic behavior is determined by incomes and consumption relative to expectations, whereas in the neoclassical theory demographic behavior is determined by absolute incomes and actual levels of consumption. Where the neoclassical economists regard the cost of children as a response to exogenous changes in female wages, Easterlin regards tastes as response to exogenous changes in relative income status. Instead of assuming a constant preference structure as in Becker's theory, Easterlin introduces a shifting one; the shift being determined by incomes and prices. The Easterlin model is regarded by many as an attempt to introduce a sociological mode of thought into economic theory.

The observation that large cohorts would give birth to relatively few children and small cohorts to many children also creates the potential for continuing fluctuations in the relative size of birth cohorts. As a result there emerges a self-generating and to a certain extent self-correcting fertility cycle of around four to five decades in length and with associated changes in a wide range of social and economic phenomena induced by fluctuations in relative cohort size. Pre-World War II Kuznets cycles of 15–25 years caused by aggregate demand fluctuations would be replaced, according to Easterlin's argument, by post-World War II swings of double the length originating in fertility changes operating with a lagged effect on the age structure of labor supply. Some scholars have included such population cycles in mathematical models (e.g., Lee 1974, Samuelson 1976, Chu and Lu 1995). With the relative income hypothesis as a starting point, Easterlin describes *inter alia* important elements of social change in the USA during the postwar era, his main objective being to explain the baby boom of the 1940s and 1950s and the baby bust of the 1960s and 1970s. As a result of low birthrates during the interwar period—and of low immigration—relatively few young people entered the labor market during the 1940s and 1950s. Labor market conditions were favorable, since firms were competing fiercely for the limited numbers on the labor market by such means as increased starting wages. The future beckoned brightly too, for the prospects of advancement appeared good in the prevailing conditions of labor shortage. As a result of the fall in the proportion of young people in the total labor force during the 1940s and 1950s, the relative incomes of the young cohorts increased during this period at the same time as fertility was rising.

Conditions for the next generation were quite the opposite. Because of the high relative incomes of their parents and the subsequent rise in fertility, the new generation was comparatively large and at the same time entertained expectations of a high material standard. Thus, when the new generation emerged on the labor market during the 1960s and 1970s its expectations were high. Because of the cohort's size,

starting wages were low, the likelihood of becoming unemployed was high, and opportunities for advancement seemed rather limited. The chances of fulfilling their high expectations therefore appeared small. As a consequence, families were formed later, and childbearing was reduced. In other respects, too, as regards marriage and divorce patterns for example, female labor force participation rates, nonmarital fertility, suicide and criminal frequencies, and political alienation (students riots in the late 1960s, etc.), the 1960s and 1970s exhibited characteristics quite different from those of the previous two decades. Easterlin also predicted that a repetition of the baby boom would take place during the 1980s and early 1990s. In reality, however, fertility remained largely unchanged in these small cohorts.

2.2 Criticism of the Easterlin Hypothesis

Though much admired, the Easterlin hypothesis has not been exempt from criticism. Some scholars contend that the analysis is conducted at far too high a level of aggregation. They assert that the information available on the individual level concerning numbers of children and relative wages at a given time does not substantiate the Easterlin hypothesis, though there also are micro studies that report findings consistent with it. A number of scholars maintain that the theoretical ground for believing that changed relative incomes should influence fertility, labor force participation rate, and so on, are not particularly well developed either. Easterlin is also said to fail to take account of women's contribution to family income. If men's relative incomes are low at the same time as the labor force participation rate of women is high, it is fully conceivable that household incomes—which are the relevant factor at this point—are even higher than if the man had a relatively high income and the woman participated little in the labor force.

Other scholars argue that Easterlin's view of the modus operandi of the labor market is too simple and not rooted in any accepted labor market theory. In particular, one has criticized the idea that substitutability between older and younger workers is as limited as the Easterlin hypothesis holds. Easterlin is even accused of being careless with empirical data. He is said to pay too little attention to alternative explanations and to let his hypothesis rest on fragile empirical foundations, underpinning it only by *one* observed cyclical covariance between age structure relative incomes, and fertility.

Furthermore, the efforts made by Easterlin to link varying economic conditions to individual behavior in fields other than that of childbearing have not had any profound impact on current research. The studies which have been carried out give incomplete or conflicting results.

Nevertheless, the view that generation size significantly affects the attitudes and actions of individuals in a variety of areas, especially fertility behavior, has formed the basis of a number of international studies. As noted above, the relationships postulated by Easterlin seem to fit rather well for the USA. Some other countries (Australia, New Zealand, Norway, Canada) show similar patterns. In other countries there is a connection, but it is considerably weaker (Finland, France, The Netherlands, England and Wales, Sweden). For certain countries (Germany, Italy, Denmark) no such connection has been found at all. It should be observed, however, that usually it is only the reduced relation between age structure and fertility that has been studied, which is a 'second best' solution in this connection. Empirical studies have had to be based on age-specific incomes of different age groups to test the hypothesis properly. (For an overview of the 76 empirical analyses of the 'Easterlin hypothesis' that have been undertaken so far, and also of the criticism against the hypothesis, see Macunovich 1998.)

The first link of Easterlin's key relation between age structure relative incomes and fertility, has shown itself to be the more tenable and has inspired some scholars to develop the basic lines of thought both empirically and theoretically. In Easterlin's original analysis the relation is constructed on the basis of a relatively simple model of how the labor market functions. Thus, Easterlin does not pay much attention to possibilities for substitution between younger and older workers, nor does he differentiate between educational categories. An increased supply of younger labor will, therefore, result automatically in reduced relative incomes and increased unemployment for that younger element, all other things being equal. Later studies have shown with greater precision and detail how changes in the age structure have influenced relative incomes, and have established *inter alia* that cohort effects on incomes have been stronger for higher-educated categories than for lower. It has further been shown that the cohort effect on incomes has a tendency to wane over time, and that it is essentially absent for female workers. This suggests that the possibilities of substitution between younger and older women are large.

Easterlin's thesis that economic conditions can vary very widely between different demographic groups, and that a group's size influences its relative economic situation, has been clearly substantiated. But studies concerning the period after 1980 suggest that the effect applies only under special conditions. According to Easterlin, the relative income of the young ought to increase when the 'baby bust' cohorts entered the labor market in the late 1980s. Yet what one could observe was a continued decline in the relative wages of the young. It seems that changes in the supply of male workers may be able to explain the pattern of wages prior to 1980, but that a change in the structure

of labor demand—in particular an increase in the demand for more educated labor—offers a better explanation of wage patterns after 1980.

Easterlin himself has noted that the fortunes of a cohort depend on a number of other factors besides cohort size alone. One would therefore expect cohort size effects to dominate a cohort's experience only when other factors are relatively constant. This was the case in the period after World War II, when governmental monetary and fiscal policies stabilized growth of aggregate demand. In addition, the relative impact of immigration on the labor force was greatly reduced as a result of restrictive legislation adopted in the 1920s. In Easterlin's view the general slowdown of economic growth after 1973 and the widespread retreat from full employment as an objective of macro-economic policy are more important than changes in cohort size as factors explaining observed changes in fertility during the most recent decades. Furthermore, high immigration and the rise in female labor force participation also tend to limit the importance of birth cohort size, and so will government welfare programs that offer policies of full employment. According to Easterlin this is the reason why stronger cohort effects can be observed for the USA than for European nations, since the latter have more general welfare systems (see also Pampel and Peters 1995).

Just the same, it has proved possible to single out one factor—cohort size—that underlies variations in relative income and in fertility during certain periods and in some countries after World War II. This finding is moderate in scope, but solid and robust in its moderation. Only in exceptional cases can it be adduced as a dominant explanation of the variations observed. Nevertheless, the 'Easterlin hypothesis' and the many studies it has inspired, has given us new insights into the kinds of different connections and responses of economic, sociological, demographic, and psychological character which generate population cycles.

See also: Baby Booms and Baby Busts in the Twentieth Century; Cohort Analysis; Fertility Change: Quantum and Tempo; Malthus, Thomas Robert (1766–1834); Population Cycles, Formal Theory of

Bibliography

- Abramovitz M 1961 The nature and significance of Kutznets cycles. *Economic Development and Cultural Change* 9: 225–48
- Abramovitz M 1968 The passing of the Kutznets cycle. *Economica* 35
- Artzrouni M A, Easterlin R A 1982 Birth history, age structure, and post World War II fertility in ten developed countries: An exploratory empirical analysis. *Genus* 38(3–4): 81–99
- Becker G S 1960 An economic analysis of fertility. In: Universities-National Bureau, Committee for Economic Research. *Demographic and Economic Change in Developed Countries*. Universities-National Bureau Conference Series No. 11. Princeton University Press, Princeton, NJ

- Becker G S, Lewis H G 1973 On the interaction between the quantity and quality of children. *Journal of Political Economy* **81**(2): S279–88
- Boserup E 1965 *The Conditions of Agricultural Growth: The Economics of Agrarian Change Under Population Pressure*. Aldine, Chicago
- Chu C T C, Lu H C 1995 Toward a general analysis of endogenous Easterlin cycles. *Journal of Populations Economics* **8**: 35–57
- Easterlin R A 1966 Economic-demographic interactions and long swings in economic growth. *American Economic Review* **16**: 1063–104
- Easterlin R A 1968 *Population, Labor Force and Long Swings in Economic Growth: The American Experience*. Columbia University Press, New York
- Easterlin R A 1978 What will 1984 be like? Socioeconomic implications of recent twists in age structure. *Demography* **15**(4): 397–432
- Easterlin R A 1980 *Birth and Fortune*. Basic Books, New York
- Easterlin R A 1987 Easterlin hypothesis. In: Eatwell J, Milgate M, Newman P (eds.) *The New Palgrave: A Dictionary of Economics* 2. The Stockton Press, New York, pp. 1–4
- Easterlin R A, Schaeffer C M, Macunovich D J 1993 Will baby boomers be less well off than their parents? Income, wealth and family circumstances over the life cycle in the United States. *Population and Development Review* **19**(3): 497–522
- Kuznets S 1958 Long secular swings in the growth of population and in related economic variables. *Proceedings of the American Philosophical Society* **CII**: 25
- Lee R 1974 The formal dynamics of controlled populations and the echo, the boom and the bust. *Demography* **11**: 563–85
- Macunovich D J 1998 Fertility and the Easterlin hypothesis: an assessment of the literature. *Journal of Population Economics* **11**: 53–111
- Pampel F C, Peters H E 1995 The Easterlin effect. *Annual Review of Sociology* **21**: 163–94
- Samuelson P A 1976 An economist's non-linear model of self-generated fertility waves. *Population Studies* **30**: 243–7
- Simon J S 1977 *The Economics of Population Growth*. Princeton University Press, Princeton, NJ

R. Ohlsson

Population Cycles, Formal Theory of

A population cycle is a roughly periodic variation over time in the size or composition of a population. Here, composition refers to the component structure of the population, for example into distinct age classes. The formal theory of population cycles consists of mathematical tools that yield qualitative and quantitative information about the characteristics of cycles, such as their amplitude or period. There are three broad types of cycle: *transient* cycles are a response to a perturbation of finite duration such as a fertility transition, *externally driven* cycles are triggered and sustained by a persistent external perturbation such as an

economic cycle; *internally driven* cycles are maintained by interactions within the population such as a feedback between cohort size and cohort fertility. This article discusses the theory for these three types of cycles, dealing mainly with age-structured populations, although other aspects of population structure are considered briefly.

1. Mechanisms Driving Population Cycles

To explain the different mechanisms that drive cycles, and the kinds of formal theory that are required, it is useful to begin with Lotka's classical demographic equation which relates births $B(t)$ at the present time $t > 0$ to past births, by summing the reproductive contributions $\phi(a, t)$ made by individuals of all ages a , as

$$B(t) = \int_0^t da \phi(a, t) B(t-a) + H(t). \quad (1)$$

Here $H(t)$ specifies births to individuals who were present at $t = 0$, and immigration is ignored. The theory of this classical demographic model is discussed in *Population Dynamics: Theory of Stable Populations*; *Population Dynamics: Classical Applications of Stable Population Theory*; *Population Dynamics: Theory of Nonstable Populations*.

Transient cycles typically occur in a population whose vital rates, i.e., mortality and fertility, change over a time interval, say from $t = 0$ to $t = t_1$, and are constant thereafter. A practical example of such a perturbation is a decline from high fertility to replacement fertility that starts at $t = 0$ and is completed at $t = t_1$. After time t_1 , the population is described by classical demographic theory and undergoes damped oscillations towards the classical stationary state. But the dynamics during the transition are also oscillatory and require a different treatment.

Externally driven cycles result from a persistent perturbation of vital rates. One example is a periodic variation over time in fertility that is generated by an exogenous force, such as an economic cycle in wages. Another example is an irregular, stochastic variation in fertility or mortality that produces an ongoing stochastic response that has cyclical components. Formally, such perturbations are described by an appropriate time-dependence in the function $\phi(a, t)$.

Internally driven cycles result from changes in vital rates that are driven by changes in population size and composition. In such cases, the net reproduction $\phi(a, t)$ has a functional dependence on the past trajectory of births, or on past vital rates. One example is a relationship between the fertility of a cohort in its childbearing years with the size of the cohort: for a substantive discussion of such relationships, see *Ferti-*

lity Transition: Economic Explanations; Baby Booms and Baby Busts in the Twentieth Century; Population Cycles and Demographic Behavior. Another example is provided by inverse projection in historical demography (Lee 1977), in which observed counts of births and deaths over time are used in conjunction with the Lotka equation to estimate the time-variation in the level of mortality rates. Here, the constraint of matching observed counts results in a feedback between the level of mortality and population counts (Wachter 1986).

The formal theory of cycles driven by each of these mechanisms is discussed below, followed by a brief consideration of future directions.

2. Transient Cycles

A perturbation such as a fertility transition may take decades to complete, and a formal theory must provide formulas for the birth sequence $B(t)$ both during and after the perturbation. Li and Tuljapurkar (1999b) provide an explicit if cumbersome description of the decaying population cycles that result from a perturbation. This theory shows that the population undergoes damped cyclical changes during and after the transition. The characteristics of these cycles have not been explored except in particular cases; but this analysis makes possible the computation of population momentum for all gradual transitions (Li and Tuljapurkar 1999a). It is worth noting that the analytical results for this case rest on properties of the roots of Lotka's equation. The relationship between these roots and the age-pattern of reproduction has been studied by approximate (Keyfitz 1985) and exact (Wachter 1984) methods that are also useful in the analysis of externally and internally driven cycles.

3. Externally Driven Cycles

An illuminating treatment of population cycles driven by external periodic variation in reproduction was given by Coale (1972), who showed that the population responds with periodic variation in the numbers and relative sizes of its age components. The challenge for the theory lies in the nonlinearity of eqn. (1) when $\phi(a, t)$ contains a periodic component. To see this, suppose that $\phi(a, t) = \phi(a) + \psi(a, t)$ with ψ being periodic in time t , and suppose that the resulting birth sequence in eqn. (1) has the form $B(t) = B + b(t)$, with b also a periodic function. Substitution of these forms into eqn. (1) shows that the unknown $b(t)$ depends on a nonlinear product $\psi(a, t - a) b(t - a)$. General results for nonlinear renewal eqns. of this type are not known. There are analytical approximations which employ an expansion in the amplitude of the periodic perturbation ψ ; these describe the amplitude response of the

population and the phase difference between population cycles and the external driving cycle (Tuljapurkar 1985). An interesting class of periodically perturbed Lotka models has been described and analyzed exactly by Kim and Schoen (1997), Schoen and Kim (1994, 1997). The latter work demonstrates that it is sometimes possible to solve exactly the homogeneous form of equation (1), i.e., with $H = 0$, for an explicitly time-dependent $\phi(a, t)$. Such analytical examples can be valuable as a source of insight into the dynamics, and as a check on approximate theories intended for less tractable problems.

The analysis of periodic perturbations suggests that a persistent stochastic perturbation of vital rates will also generate a superposition of cycles of population change. The dynamics of a population subject to stochastic perturbation are also nonlinear, so again the theory has relied upon approximations. Lee (1975) pioneered the application of time-series methods to the linearized analysis of stochastic perturbations. To go beyond linearization, e.g., if the perturbations are large or have strong positive serial correlations, a higher order approximation may be used as discussed by Tuljapurkar (1990), Tuljapurkar and Lee (1997).

4. Internally Driven Cycles

The formal analysis of internally driven cycles is quite distinct from the analyses used in classical demography, and draws on developments in nonlinear dynamics (for a review of modern nonlinear techniques for dynamics, see Guckenheimer and Holmes (1983)). To start with, it is important to recognize that there are different forms of internal feedback. One class of feedback is exemplified by cohort size effects on fertility: large birth cohorts go on to have lower fertility than their parents and the opposite for small birth cohorts (Lee 1974, 1976, Wachter and Lee 1989). A second class of feedback is generated by social norms or expectations: for example, individual fertility is influenced by the average fertility in the population (Kohler 2000). A third class of feedback arises through an interaction between a population and external variables—for example, economic conditions, environmental conditions, or resources such as land holdings—and these external variables in turn influence fertility or mortality (Bonneuil 1990, 1992, 1997, Chu 1998, 1999, Feichtinger and Sorger 1990, Feichtinger and Dockner 1990, Lee 1987, Milik and Prskawetz 1996).

The formal theory of such feedbacks shares several elements that may be illustrated by considering a nonlinear version of Eqn. (1). Suppose that the net reproduction $\phi(a, t)$ in that equation changes with the population structure: for example, the fertility of individuals aged a at time t depends on the size of that cohort relative to the size of their parents' cohort.

Such feedbacks result in a functional relationship of the form

$$\phi(a, t) = \phi(a)h(g, a, \{B(t-s), 0 < s < S\}). \quad (2)$$

Here the function h on the right describes the effect of the other birth cohorts, born as long as $S > 0$ years ago, on the fertility of the cohort aged a . The parameter g controls the strength of the feedback, for example, by controlling the gradient of h . Lotka's equation is now a nonlinear renewal equation, which can be written in abstract form as

$$B = \mathcal{L}(\mu, B). \quad (3)$$

The nonlinear operator $\mathcal{L}$ maps past values of births into their current value. This verbal description can be formalized suitably; for an example of the mathematical details see, e.g., Cushing (1978).

In most applications the nonlinear eqn. (3) has a constant solution, $E(\mu)$. The dynamic stability of this solution is examined by studying small deviations $x(t) = \{B(t) - E(\mu)\}$: stability means that $x(t)$ should decrease over time. An expansion of the nonlinearity in eqn. (3) yields a series of terms of the form

$$x = \mathcal{H}(\mu, x) + \mathcal{G}(\mu, x, x) + \mathcal{R}. \quad (4)$$

The operator $\mathcal{H}$ is a linear operator, $\mathcal{G}$ is quadratically nonlinear, and $\mathcal{R}$ involves higher-order nonlinear terms. Local stability is determined by the eigenvalues of the linear operator $\mathcal{H}$, and these in turn depend on the elasticity of the linear response (Wachter 1991a). In typical examples, the equilibrium is locally stable when $g < g^*$, say, but becomes unstable when $g \geq g^*$. Local instability means that small deviations from the equilibrium births will change with time as $\exp\{(p+iq)t\}$ where $p > 0$ and q are given by eigenvalues of the linear operator $\mathcal{H}$. The local change in $x(t)$ follows a cycle of period $(2\pi/q)$. The central question now is: when such local instabilities grow and experience the nonlinear forces in equation (4), will they turn into sustained, stable cycles? If they do, increases in the parameter g past the threshold value g^* will generate a sustained population cycle, called a *limit cycle*. Since $E(\mu)$ is still a constant solution, eqn. (3) (or its solution) is said to have undergone a bifurcation at $g = g^*$.

For any value of g above but not far from the threshold g^* , the nonlinear limit cycle may be represented by a superposition of periodic solutions to the linear part of eqn. (3). Numerical illustrations of such limit cycles are given by Frauenthal and Swick (1983). A systematic procedure for computing the amplitude, period, and stability properties of this limit cycle beyond the bifurcation point is described by Tuljapurkar (1987), and substantive applications that illustrate these computations are given by Wachter

(1991b) and Chu (1998). These procedures may be extended to many other models with feedback, and provide an important part of the theory of population cycles. Other feedback mechanisms may result in a different approach to bifurcation (e.g., bifurcation in fertility level, Kohler 2000).

5. New Directions

The formal theory of cycles has moved considerably beyond classical demographic theory. The nonlinear bifurcation analysis discussed above can be considerably extended. Wachter (1994) has explored global properties of feedback models, addressing the behavior of cycles far above a bifurcation threshold. Bonneuil (1992), Prskawetz and Feichtinger (1995), and Tuljapurkar et al. (1994) show that increasing feedback can lead to more complex types of population cycles including aperiodicity and chaos. Purely internal feedbacks may not be strong enough to generate demographic cycles in human populations, but feedbacks mediated by social or external interactions may well be strong enough. The challenges that face the theory now lie in two areas: confrontation with data to test and modify theories, and the development of new models that incorporate a wider range of demographic dynamics and interaction. As yet, nonlinear cycles in human populations have not been examined in the realistic and rich way that biologists have explored cycles in the structure of laboratory populations (de Roos 1997, Cushing 1997).

Bibliography

- Bonneuil N 1990 Turbulent dynamics in a 17th century population. *Mathematical Population Studies* 2: 289–311
- Bonneuil N 1992 Attractors and confiners in demography. *Annals of Operations Research* 37: 17–32
- Bonneuil N 1997 *Transformation of the French Demographic Landscape, 1806–1906*. Oxford University Press, Oxford
- Chu C Y C 1998 *Population Dynamics: A New Economic Approach*. Oxford University Press, Oxford
- Chu C Y C 1999 Predator-prey models with endogenous decisions. *Mathematical Population Studies* 8: 55–71
- Coale A J 1972 *The Growth and Structure of Human Populations: a mathematical investigation*. Princeton University Press, Princeton, NJ
- Cushing J M 1978 Nontrivial-periodic solutions of some Volterra integral equations. *Lecture Notes in Mathematics* 737: 50–66
- Cushing J M 1997 Nonlinear matrix equations and population dynamics. In: Tuljapurkar S, Caswell H (eds.) *Structured Population Models in Marine, Terrestrial and Freshwater systems*. Chapman and Hall, New York
- de Roos A M 1997 A gentle introduction to physiologically structured models. In: Tuljapurkar S, Caswell H (eds.) *Structured Population Models in Marine, Terrestrial and Freshwater systems*. Chapman and Hall, New York

- Feichtinger G, Dockner E J 1990 Capital accumulation, endogenous population growth, and Easterlin cycles. *Journal of Population Economics* 3: 73–87
- Feichtinger G, Sorger G 1990 Capital accumulation, aspiration adjustment, and population growth: limit cycles in an Easterlin-type model. *Mathematical Population Studies* 2: 93–103
- Frauenthal J C, Swick K E 1983 Limit cycle oscillations of the human population. *Demography* 20: 285–98
- Guckenheimer J, Holmes P 1983 *Nonlinear Oscillations, Dynamical Systems and Bifurcations of Vector Fields*. Springer, New York
- Keyfitz N 1985 *Applied Mathematical Demography*. Springer, New York
- Kim Y J, Schoen R 1996 Populations with sinusoidal birth trajectories. *Theoretical Population Biology* 50: 105–23
- Kohler H P 2000 Social interactions and fluctuations in birth rates. *Population Studies* 54: 223–7
- Lee R D 1974 Formal dynamics of controlled populations and the echo, the boom and bust. *Demography* 11: 563–85
- Lee R D 1975 Natural fertility, population cycles, and the spectral analysis of births and marriages. *Journal of the American Statistical Association* 70: 295–304
- Lee R D 1976 Demographic forecasting and Easterlin hypothesis. *Population and Development Review* 2: 459–68
- Lee R D 1977 Methods and models for analyzing historical series of births, deaths and marriages. In: Lee R D (ed.) *Population Patterns in the Past*. Academic Press, New York
- Lee R D 1987 Population cycles. In: Eatwell J, Milgate M, Newman P (eds.) *The New Palgrave; A Dictionary of Economic Theory and Doctrine*. MacMillan, UK
- Li N, Tuljapurkar S 1999a Population momentum for gradual demographic transitions. *Population Studies* 53: 255–62
- Li N, Tuljapurkar S 1999b The solution of time-dependent population models. *Mathematical Population Studies—J. Demog.* 7: 311–29
- Milik A, Prskawetz A 1996 Slow-fast dynamics in a model of population and resource growth. *Mathematical Population Studies* 6: 155–69
- Prskawetz A, Feichtinger G 1995 Endogenous population growth may imply chaos. *Journal of Population Economics* 8: 59–80
- Schoen R, Kim Y J 1994 Cyclically stable populations. *Mathematical Population Studies* 4: 283–95
- Schoen R, Kim Y J 1997 Exploring cyclic net reproduction. *Mathematical Population Studies* 6: 277–90
- Tuljapurkar S 1985 Population dynamics in variable environments. 6 Cyclical environments. *Theoretical Population Biology* 28: 1–17
- Tuljapurkar S 1987 Cycles in nonlinear age-structured models. I. Renewal equations. *Theoretical Population Biology* 32: 26–41
- Tuljapurkar S 1990 *Population Dynamics in Variable Environments*. Springer, New York
- Tuljapurkar S, Boe C, Wachter K 1994 Nonlinear feedback dynamics in fisheries: analysis of the DeRiso–Schnute model. *Canadian Journal of Fisheries and Aquatic Science* 51: 1462–73
- Tuljapurkar S, Lee R 1997 Demographic uncertainty and the stable equivalent population. *Mathematics and Computer Modelling* 26: 39–56
- Wachter K W 1984 Lotka's roots under rescalings. *Proceedings of the Natural Academy of Sciences of the USA* 81: 3600–4
- Wachter K W 1986 Ergodicity and inverse projection. *Population Studies* 40: 275–87
- Wachter K W 1991a Pre-procreative ages in population stability and cyclicity. *Mathematical Population Studies* 3: 79–104

Wachter K W 1991b Elusive cycles: are there dynamically possible Lee–Easterlin models for U.S. births? *Population Studies* 45: 109–35

Wachter K W 1994 The Cohort Feedback model with symmetric net maternity. *Mathematical Population Studies* 5: 25–44

Wachter K W, Lee R D 1989 U.S. births and limit cycle models. *Demography* 26: 99–115

S. Tuljapurkar

Population Dynamics: Classical Applications of Stable Population Theory

1. Introduction

The theory of stable populations is presented in both continuous and discrete form in the article *Population Dynamics: Theory of Stable Populations*. Virtually any real population, whether contemporary or historical, deviates from stability, but the stable model produces a standard for comparison based on simplifying assumptions that allow us to handle data limitations. Some applications of the theory appear in the introductory article just mentioned and in other articles, including *Demographic Techniques: Indirect Estimation*; *Population Dynamics: Momentum of Population Growth*. The present article will review several applications within a structure that specifies the amount of information that is required. The applications are ‘classical’ in the sense that they are relatively well known.

It is helpful first to distinguish between two classes of models that are widely used: survival processes and renewal processes. The prototypical example of a survival process is the single-sex, single-decrement life table, described in the article *Life Table*. A survival process describes a cohort, real or synthetic, from age 0 until some terminal age ω . The main concepts are (a) the survival function, $l(x)$, namely the proportion of the cohort that survives to each age x , and (b) the mean waiting time or remaining life expectancy, e_x^o , at any specific age x .

A renewal process, by contrast, is a survival process with reproduction. With this component, the process generates successive cohorts that are (in general) larger than necessary simply to replace the persons who have died. The cohort interpretation is supplemented with the asymptotic implications of fixed and stable overall rates (of growth, mortality, and fertility), and a fixed and stable age distribution, none of which depends on initial conditions.

If a renewal process is in equilibrium—that is, if reproduction is exactly at replacement value, then the stable population is stationary. It has an intrinsic growth rate of zero, and one does not need explicit

age-specific birth rates to analyze it, for its overall birth rate is identical to the overall death rate. There are several classical applications and generalizations of this scenario, all of which can be regarded as applications of stable population theory because a survival process is a special stable renewal process.

A stable population is typically motivated as the asymptotic consequence of projecting the population indefinitely with fixed age-specific rates. Its overall rates and age distribution converge to functions that do not depend on the initial conditions. However, unless the intrinsic growth rate is zero, there is a fundamental implausibility in this asymptotic interpretation. If the intrinsic growth rate is positive, then the overall population size will increase exponentially, an outcome that cannot be sustained indefinitely in a real population. If the intrinsic growth rate is negative, then the overall population size will *decrease* exponentially until eventual extinction. In probabilistic terms, the fixed-rates assumption implies almost certain ultimate extinction or explosion (see *Population Dynamics: Probabilistic Extinction, Stability, and Explosion Theorems*). It is probably more useful to regard the stable population parameters simply as a synthesis of the underlying age-specific probabilities or rates.

2. Applications Not Requiring Schedules of Rates

Applications of stable population assumptions can be organized according to the amount of information they utilize. Probably the simplest possible applications occur when the only parameter utilized is the intrinsic growth rate, r . If we assume stability between two times t_1 and t_2 , and that the population size at those two times is n_1 and n_2 , then

$$n_2/n_1 = e^{r(t_2-t_1)}$$

Solving for r , we get

$$r = (\ln n_2 - \ln n_1)/(t_2 - t_1)$$

We can then estimate the population size at an intermediate time, $t_1 \leq t \leq t_2$, to be $n_t = n_1 e^{r(t-t_1)}$. This formula, or variations upon it, is often used when the population is not stable, in the absence of better information. Another simple and well-known application is to determine the doubling time associated with a given value of r . Setting $n_2 = 2n_1$ and letting T be the time required for the population to double, we obtain $2n_1 = n_1 e^{rT}$, leading to $T = (\ln 2)/r = 0.6931/r$. For example, if the annual growth rate is 0.01, or 1 percent, a stable population will double in 69.31 years, or approximately seventy years. These and related applications are discussed by Hinde (1998, pp. 154–8), for example.

Suppose that in addition to the intrinsic growth rate, r , we also know the intrinsic birth rate, b (and hence the intrinsic death rate $d = b - r$). Keyfitz (1977, pp. 12–14) uses such information to estimate the

number of persons who have ever lived on Earth, which is the same as the number of persons ever born. His strategy is to piece together long intervals of time within which stability (at different levels) is assumed. If the population is stable between t_1 and t_2 , then the number of person-years lived in that time interval will be

$$PY = \int_{t_1}^{t_2} n_1 e^{r(t-t_1)} dt = (n_2 - n_1)/r$$

PY can be converted to the corresponding number of births by multiplying by b :

$$B = bPY = b(n_2 - n_1)/r$$

(If $r = 0$, then $n_1 = n_2 = n$, $PY = n(t_2 - t_1)$, and $B = bn(t_2 - t_1)$.)

3. Applications that Assume Schedules of Rates

Another class of applications arises if we know r , b , and d , and do not know the survival and fertility schedules but are willing to make some assumptions about them. Coale and Demeny (1983) blended empirical life tables into four sets of so-called regional model life tables (north, south, east, and west) for males and females, keyed to the life expectancy e_0^m at birth and e_{10}^f at age 10 (see *Demographic Models*). They also developed plausible fertility schedules, varying according to the mean age of childbearing but adding up to one child, so that the entire schedule could be multiplied by a hypothetical total fertility rate (TFR). With this information, much of it hypothetical, they were able to generate sets of stable age distributions. It is then possible to estimate a wide range of population characteristics, using only a few pieces of summary information (along with a specification of region and sex). The summary information typically includes certain combinations of the following: a measure of growth, such as r ; a measure of fertility, such as b , the gross reproduction rate (GRR), or the TFR; a measure of mortality, such as d or the average age at death; or a measure of the age distribution, such as the proportion of persons below age 20. Stable age distributions generated in this way can also be used to approximate real age distributions in settings of deficient data. Some methods for the computation of population parameters using the stable population model are described in *Demographic Techniques: Indirect Estimation*.

4. Applications that Require Schedules of Rates

A number of analyses of population dynamics have resulted from examining the sensitivity of population characteristics to disturbances of a stable population. These applications require knowledge of the age-specific schedules for mortality (say m_x for each age x),

of fertility (say f_x for each x), and for any other transitions in the model. For example, Preston (1974) took partial derivatives of the basic stable population formulas to estimate the effect of various age-specific mortality changes upon the intrinsic growth rate, birth rate, and age composition. He found that the effects could be represented as simple functions of the age pattern of the cumulative changes in the mortality rates. Probably the best known of these sensitivity analyses has been on the topic of population momentum, which in its simplest form measures the eventual effect on population growth of an immediate shift to replacement fertility (see *Population Dynamics: Momentum of Population Growth*).

The classical stable population model has no in- or out-migration, but several authors have used the model in innovative ways to evaluate the impact of emigration (Keyfitz 1971) or immigration (Lesthaeghe et al. 1988). Mitra (1990) considered how a population could maintain stationarity through a combination of below-replacement fertility and a steady stream of immigration. With assumptions about the ages and fertility of the immigrants, it is possible to determine the necessary size of the stream of immigration. This model is relevant, at least as a first approximation, to the current mix of below-replacement fertility and immigration in Europe and North America.

The classical formulation of stable population theory only allows individuals to be born, procreate, and die, but later developments permit individuals to move between several statuses or states during their lifetime, as in multistate models (see *Multi-state Transition Models in Demography*). In one obvious application the states represent regions of a territory (Feeney 1970). In another, Keyfitz developed a linkage between aging and career mobility in stable populations (1973; the title of this paper is misleading because he included stable as well as stationary populations). He supposed that advancement in status across a set of discrete categories is essentially based on the steepness of the age distribution. In a setting of high growth, age (as a proxy for experience) tends to produce earlier advancement because there are relatively few people at higher levels. In a setting of low growth, seniority is less valued, because there are relatively many people at higher levels, and promotion will take longer. The approach is again a type of sensitivity analysis, in which the age at reaching a sequence of discrete gateways is expressed as a function of stable population parameters (in particular, the intrinsic growth rate).

5. Applications that Link Individuals to One Another: Kinship

Assume now that we have access to complete schedules of survival and fertility. In continuous form, say that $l(x)$ is the probability that a woman will survive to age

x and $m(x) dx$ is the probability that a woman will have a daughter between ages x and $x+dx$ (with $m(x) > 0$ for x in the reproductive age range $\alpha < x < \beta$ within the full range of ages $0 < x < \omega$).

Single-sex kinship provides a structure by which individuals are linked through lines of descent or ancestry. Two specific models for kinship structures will be described here. In the context of kinship, reference individuals (ego) are linked to their direct descendants, such as daughters; to progenitors, such as mothers; or to their lateral kin, such as sisters. Two-sex approximations are possible. The main outputs are the probability that a specific progenitor is still alive when ego is age a , and the expected number of descendants and lateral kin (in a specific category) who are still alive, or were ever born, when ego is age a . These are obtained from various convolutions of the survival and fertility functions.

The first kinship model to be described was developed by Goodman et al. (1974, 1975). The following exposition will be limited to surviving kin when ego is age a . It can be modified to refer to kin ever born simply by deletion of the outermost survival function (that is, setting it equal to 1).

Each kinship role is expressed with a different formula, but the strategy for generating these formulas is the same for all roles. The steps to link ego with a consanguineal relative are to identify the relevant sequence of births, and the ages of the respective mothers at those dates; to incorporate the fertility rates for those ages, the probability of surviving to those ages, and, for the final link, the probability of surviving to the time when ego is age a ; and integrating over all allowable combinations of ages of the linked kin. Note that the birth of ego herself, and her survival to age a , are assumed.

For example, to link ego with her daughter, the relevant birth is that of the daughter, which occurred when ego was some previous age x . If the daughter is still alive when ego is age a , then the daughter's age is $a-x$. Thus ego experienced the fertility rate for age x , and the daughter survived to age $a-x$. Multiplying in reverse order and integrating over allowable ages x , the expected number of surviving daughters for a woman age a is

$$D(a) = \int_{\alpha}^a l(a-x)m(x) dx$$

(If $a < \alpha$ then $D(a) = 0$ and if $a > \beta$ then $D(a) = D(\beta)$.)

The stable population probability that ego's mother is alive when ego is age a was originally calculated by Lotka (1931), and is useful for estimating adult mortality when some empirical information about the incidence of orphanhood is available. If the mother was age x when ego was born, then the probability that she is still alive a years later is the probability of surviving to age $a+x$, $l(a+x)/l(x)$. Obviously this

depends on the value of x , as well as a , and these survival ratios must be weighted by the probability that the mother was age x to age $x + dx$ when ego was born. In a stable population these weights will be given by $W(x)dx = e^{-rx}l(x)m(x)dx$ (from the characteristic equation, this is the density function for mothers' ages and integrates to 1). Therefore,

$$\begin{aligned} M_1(a) &= \int_{\alpha}^{\beta} [l(a+x)/l(x)]W(x)dx \\ &= \int_{\alpha}^{\beta} e^{-rx}l(a+x)m(x)dx \end{aligned}$$

is the probability that ego's mother is alive when ego is age a .

The expected number of surviving sisters when ego is age a involves a similar logic but will not be given here. Ego is linked to her sisters through their mother. It is necessary to distinguish between older and younger sisters because the survivorship of the mother is problematic after, although not before, the birth of ego. The formulas can be extended to aunts, nieces, first cousins (mother's sister's daughters), and indeed any category of kin.

A noteworthy feature of this application of the basic model of a stable population is that it requires only the basic survival function and maternity function, and a structure to link individuals. As an alternative to linkage by kinship, it should be possible to substitute an organizational structure or some form of vertical and horizontal connections, but we are not aware of any analogous applications.

The formulas are deterministic, rather than stochastic, and will not produce distributions. They assume that the birth of a child is independent of the number of children a woman has already had, that survivorship is independent of biological relationship, and that the population is homogeneous. Although these limitations are severe, it would be difficult to find data corresponding to more realistic assumptions. Calculations using these multiple integrals require some form of quadrature.

In a second application of kinship (Pullum 1982), counts are not conditional upon the survival of ego or any other kin, but are defined as 'eventual expected' numbers of kin. For example, the net reproduction rate

$$N = \int_{\alpha}^{\beta} l(x)m(x)dx$$

is the average number of female births that will be produced, a generation later, from a single female birth. This is equivalent to the eventual expected number of daughters that a woman (ego) will have, defined at the point of ego's birth and the daughters' births.

The eventual expected number of sisters, defined at the point of ego's births and the daughters' births, can be referred to with the symbol S . N is not equivalent to any of the functions defined by Goodman et al. (1974) but S does appear within their scheme as the (expected) number of sisters (older and younger added together) ever born by the time ego has reached age $\beta - \alpha$, the maximum age at which a sister can be born. This formula superfluously conditions on the survival of ego, but allows for all sisters to be born.

It can be shown that N and S are sufficient to obtain the eventual expected numbers of any other types of kin. For example, an average of N^2 granddaughters (daughters of daughters) will be born, SN first cousins (daughters of sisters of the mother), SN nieces (daughters of sisters), and so on. The sex ratio at birth will also be constant in a stable population, and by including it one can easily extend to a two-sex kinship structure. For example, define s to be the ratio of males to females at birth, so that for every female birth there will be $1+s$ births of males and females. This implies that for every female child there will be $1+s$ children, so N is replaced by $(1+s)N$; for every sister there will be $1+s$ siblings, so S is replaced by $(1+s)S$. (Note that instead of one mother there will be exactly two parents.) The number of grandchildren (sons and daughters of sons and daughters) will be $(1+s)^2N^2$; the number of first cousins (sons and daughters of brothers and sisters of the mother or father) will be $2(1+s)^2SN$; the number of nieces/nephews (sons and daughters of brothers and sisters) will be $(1+s)^2SN$; and so on.

Like the applications given in Sect. 3 above, this second model of kinship only requires summary measures of the stable population (in this case S and N), and does not require any schedules of rates.

6. Conclusion

These applications have been presented largely at a conceptual level, without the specific data or numerical conclusions that might help to demonstrate their value (the bibliography and other entries in this encyclopedia describe several further examples). In each case, however, the original goal was to utilize stable population theory to draw approximate conclusions about real populations in situations where data are scarce. The model approximation has certain elegant formal properties and in effect permits the simulation of data that are generally not available for real populations. Indeed, significant improvement in the availability of demographic data is a likely reason why there have been relatively few new applications of the theory in recent decades.

Bibliography

Bourgeois-Pichat J 1971 Stable, semi-stable populations and growth potential. *Population Studies* 25: 235-54

- Cerone P 1987 On stable population theory with immigration. *Demography* 24: 431–8
- Coale A J 1981 Robust estimation of fertility by the use of model stable populations. *Asian and Pacific Census Forum* 8: 5–10
- Coale A J, Demeny P 1983 *Regional Model Life Tables and Stable Populations*, 2nd edn. Academic Press, New York
- Espenshade T J, Bouvier L F, Arthur B 1982 Immigration and the stable population model. *Demography* 19: 125–33
- Feeney G M 1970 Stable age by region distributions. *Demography* 7: 341–8
- Goodman L A, Keyfitz N, Pullum T W 1974 Family formation and the frequency of various kinship relationships. *Theoretical Population Biology* 5: 1–27. Addendum 1975 *Theoretical Population Biology* 8: 376–81
- Hinde A 1998 *Demographic Methods*. Arnold, London
- Keyfitz N 1971 Migration as a means of population control. *Population Studies* 25: 63–72
- Keyfitz N 1973 Individual mobility in a stationary population. *Population Studies* 27: 335–52
- Keyfitz N 1977 *Applied Mathematical Demography*. Wiley, New York
- Keyfitz N 1985 The demography of unfunded pensions. *European Journal of Population* 1: 5–30
- Keyfitz N, Nagnur D, Sharma D 1967 On the interpretation of age distributions. *Journal of the American Statistical Association* 62: 862–74
- Lapkoff S F 1991 A research note on Keyfitz's 'The demographics of unfunded pensions.' *European Journal of Population* 7: 159–69
- Lesthaeghe R, Page H, Surkyn J 1988 Are immigrants substitutes for births? Unpublished paper, European University, Florence. Published in German in 1991 as Sind Einwanderer ein Ersatz für Geburten? *Zeitschrift für Bevölkerungswissenschaft* 3: 281–314
- Lotka A J 1931 Orphanhood in relation to demographic factors: A study in population analysis. *Metron* 9: 37–109
- Mitra S 1983 Generalization of the immigration and the stable population model. *Demography* 20: 111–15
- Mitra S 1990 Immigration, below-replacement fertility, and long-term national population trends. *Demography* 27: 121–30
- Preston S H 1974 Effect of mortality change on stable population parameters. *Demography* 11: 119–30
- Pullum T W 1982 The eventual frequencies of kin in a stable population. *Demography* 19: 549–65
- Schoen R 1986 The two-sex multiethnic stable population model. *Theoretical Population Biology* 29: 343–64

T. W. Pullum

Population Dynamics: Mathematic Models of Population, Development, and Natural Resources

Although the importance of the links between population and environment is widely appreciated, there is little agreement in the literature about the nature of these links.

The dynamic link between population and natural resources—influences are never entirely unidirectional or constant over time—constitutes the main difficulty in assessing formal models and unified theories. To

complicate matters further, this interrelationship depends on the level and type of economic activities, technological development, institutional settings and cultural systems. Even the variables themselves are not uniquely defined, but depend rather on the level and purpose of the analysis.

This article does not survey numerical simulation models of demographic/economic/environmental interactions (see Sanderson 1994 for a review), of which the World 3 Model (Meadows et al. 1972) is the most popular. Instead, it looks at some simple mathematical models that capture stylized facts of the link between population and natural resource dynamics. Such 'mathematical cartoons' help us to explain and understand how various population and environmental characteristics might affect each other, and they can teach us how to respond most effectively to various demographic and environmental developments.

Although it was exhaustible resources that first received theoretical attention, this article concentrates on renewable resources since these are the most interesting and least understood as regards changes in population structures and dynamics. In contrast to nonrenewable resources, the use of which is controlled largely by market prices, renewable ones are very often open-access resources. The overuse of resources if property rights are absent or poorly defined is called the 'first tragedy of the commons' (Hardin 1968) and can be corrected by restricting access to the resource. But as emphasized by Lee (1990), if the population is not fixed—as is presupposed in the models presented in this article—free access through reproduction constitutes the 'second tragedy of the commons.'

1. Malthusian Population Dynamics and Natural Resources

The study of interactions between population and resources has a long history. According to Malthus, population growth reduces material welfare due to diminishing returns to labor on a fixed supply of land. On the other hand the higher the level of material welfare the higher the population growth rate will be. The Malthusian model predicts that 'population will equilibrate with resources at some level mediated by technology and a conventional standard of living' (Lee 1986). Improvements in technology will be offset in the long run by increases in the size of the population, but the standard of living will not be related to the level of technology. As such, the Malthusian model provides a description of a rather primitive society with incomes not too far above the subsistence level and where local renewable resources are an important part of the economic production process.

Renewable resources (agricultural land, forests, lakes, etc.) are not in fixed supply as Malthus assumed. Renewable resources regenerate, but if the rate of

utilization (harvest) exceeds the rate of regeneration, a renewable resource will be depleted or, in the extreme case, irretrievably exhausted. By adding the dynamics of renewable resource growth to the dynamics of population growth, the Malthusian model is capable of explaining patterns of population growth and resource degradation that do not necessarily end up in a single equilibrium (Malthusian trap) (see Brander and Taylor 1998 and Prskawetz et al. 1994).

The structure of these models can best be described in terms of prey–predator dynamics, with the resources, R , being the prey and the population, P , acting as the predator. The dynamics of renewable resources (Eqn. (1)) are commonly described by the standard model of mathematical bioeconomics, where the net growth of the renewable resource, $dR(t)/dt$, is affected by two counteracting factors: indigenous biological growth, $g(R(t))$, and the harvest, $H(R(t), P(t))$, which depends on the stock of resources available to be harvested and the number of people who are harvesting. Indigenous resource growth is modeled by the logistic growth function $g(R(t)) = aR(t)(K - R(t))$, where the coefficient K determines the saturation level (carrying capacity) of the resource stock (i.e., K is the stationary solution of R if the resource is not degraded) and parameter a determines the speed at which the resource regenerates. The functional form of the harvest function is determined by the prevailing economic structure, and it establishes the link between the stock of resources and population. The population growth rate, $(dP(t)/dt)/P(t)$ (Eqn. (2)), is modeled as an increasing function of material welfare $y(t) = y(P(t), H(t))$, which is determined by the level of the harvest and will be reduced by population growth. Whenever material welfare falls below the subsistence level, the population will decline. These Malthusian population dynamics imply that population growth may well adapt to resource constraints in contrast to models with exogenous positive population growth, where the economy collapses if resources do not regenerate quickly enough.

$$\frac{dR(t)}{dt} = g(R(t)) - H(R(t), P(t)) \quad (1)$$

$$\frac{dP(t)}{dt} = n(y(t))P(t). \quad (2)$$

Within this class of models it is possible to investigate how the equilibrium between resources and population stock changes, dependent on the functional relations that govern the dynamics of population and resources. For instance, the efficiency of harvest technology, the degree of substitution between labor and resources, the indigenous rate of resource growth and the carrying capacity of the resource stock will have an effect on material welfare and hence on population growth. In turn, fertility and mortality will affect the resource dynamics via the input of changing

labor stocks in the harvest. An empirical calibration of the model for the case of the small Pacific Easter Island characterized by pronounced population fluctuations was carried out by Brander and Taylor (1998).

Contrary to the Malthusian predictions, increasing population densities might be beneficial, as argued by Boserup (1981), and could well increase the human carrying capacity of the earth. Higher population densities will initiate technological innovations in agriculture, thereby increasing the yields so that the natural environment can support a larger population without reducing the level of welfare. Similar positive feedback mechanisms are captured in a simple mathematical cartoon of the interdependence between the growth in population $P(t)$ and carrying capacity $K(t)$ (expressed in numbers of individuals) by Cohen (1995) as

$$\frac{dP(t)}{dt} = rP(t)[K(t) - P(t)] \quad (3)$$

$$\frac{dK(t)}{dt} = c \left[\frac{dP(t)}{dt} \right] \quad (4)$$

with $r > 0$ and c either negative, zero, or positive. The parameter c , which captures the effect of an increment of population on the carrying capacity, determines the long-term population dynamics, and it can represent technological innovation. When $c = 1$ population size grows exponentially (Euler, eighteenth century); when $c < 1$ population grows logistically (Verhulst, nineteenth century); and when $c > 1$ population grows faster than exponentially (Forester and co-workers, twentieth century).

Malthusian results (in the sense that population will grow to the point where material welfare matches subsistence demand) can be further undermined by allowing population growth to be a choice variable. Zero population growth may well be the optimal choice for individuals in the economy. Since environmental changes are often slow over the course of an individual life span, and since environmental damage may outlive its perpetrators, overlapping generation models (Eckstein et al. (1988)) provide an appropriate demographic structure. Intra- as well as intergenerational conflicts can be modeled in such a framework, taking account of the effects of increases in population and resource exploitation on the future population's quality.

The models presented so far represent traditional societies in which populations derive their living from primary occupations (agriculture, hunting, fishing, etc.) which depend on the availability of resources. But as societies become less bound to the land, more urbanized and more technological, they not only use the environment as a source of natural resources but also as a dump for waste products arising from human activity. Furthermore, an economy's production possibilities are no longer determined by the maximum

sustainable yield of renewable resources. Improvements in technology can increase the sustainable yields or reduce the resource stock required for production, and economic growth will allow for the use of man-made capital in place of natural resources. In open economies with trade, technological change, and economic growth, there is no simple and direct relationship between population growth on the one hand and the environment on the other hand.

Resources might still be at risk, albeit no longer by positive population growth alone. The risk can stem from the environmental costs of consumption and production. It is therefore of great importance to understand how these environmental impacts depend on different population structures.

2. *The Impact of Population on the Environment*

In a seminal paper, Ehrlich and Holdren (1971) express the impact on the environment, I , (e.g., the value of some pollutant) as the product of population, P , and per capita impact, $F(P)$:

$$I = PF(P) \quad (5)$$

Depending on whether diminishing returns or economies of scale are dominant, per capita impact can be increasing or decreasing in the total population. The most familiar form of such a demographic impact model is the so called I-PAT identity, which divides per capita impact $F(P)$ into affluence A , and technology T . Affluence can be measured as GNP per capita, and technology represents the mean impact of an extra dollar of production (impact/GNP). The latter variable encompasses the cleanliness of technology, the effects of institutions, and other factors which can alter the impact efficiency of production.

Demographic impact models (Wexler 1996) constitute multiplicative identities aimed at decomposing the effect on the environment of population on the one hand and that of consumption and technology on the other. The number of variables on the right-hand side of Eqn. (5) can be any number greater than two. Although these models are commonly used, they face several criticisms. They suppress any feedback relation between the variables on the right-hand side; their implications are highly sensitive to the choice of decomposition variables; and the method of decomposition and the role of institutions and culture are ignored. Recent research on the I-PAT identity has dealt with some of these problems. Expressing the identity in terms of variances rather than means (Preston 1996) allows one to incorporate the links between the decomposition variables. Considering households instead of individuals as the unit of consumption (MacKellar et al. 1995) highlights the effect of economies of scale as household size increases. Typical results of such analyses are that the population

has a much stronger impact on the environment in developing countries than in developed countries. In the case of environmental hazards produced by industrial processes, population growth plays a minor role both in developing and in developed countries.

Milik et al. (1996) embed the I-PAT identity into a dynamic model where population P , affluence A , and technology T influence each other and vary over time. Pollution I reduces the regenerative capacity of natural resources, which in turn influences the dynamics of population and affluence. The model highlights the fact that understanding and monitoring the impact of pollution on the environment is important even when outward signs of damage (as a reduction in the resource stock) are not yet visible. Ecosystems often absorb stresses over long periods of time and then eventually reach a disruption level at which the cumulative consequences of stress appear in critical proportions. A policy of waiting until the first signs of environmental deterioration are observed before actions are taken to reduce the impact on the environment may well prove to be disastrous. When action is initiated, the impact on the environment might already be irreversible. Mathematically, these systems can best be described in terms of 'slow-fast' dynamics, where during certain periods some variables vary much more rapidly than others. While population and economic variables can be regarded as slow moving variables, the environment changes much rapidly once the system has absorbed too much stress. For instance, the emission of carbon dioxide occurred almost unobserved for several decades before any attention was paid to the impact of greenhouse gases on global warming.

It becomes difficult to quantify the impact of population on the environment if the demand for local resources is not related to local population growth and if institutional environments are not stable. In fact, 'commercialisation by converting a limited and inelastic subsistence demand to a limitless and elastic export demand (from the standpoint of the region, at least) can lead to much more rapid rates of exploitation than would be implied by population growth alone.' (Repetto and Holmes 1983).

3. *Modeling Environmental Influences on Population Dynamics*

Existing models of environmental influences on population exhibit two shortcomings. First, they generally do not distinguish between the separate effects the environment may have on fertility and mortality. Second, the environmental impact on population growth is commonly modeled to work only through the economic and social variables of the model. There are only two examples in the literature that model the direct effect of environmental stress on fertility and mortality.

The first example (Nerlove 1991) is based on the vicious circle of poverty, population growth, and environmental degradation as evidenced in small rural communities of sub-Saharan Africa. In these economies simple tasks like fetching water and collecting fodder and wood are mostly carried out by children. This means that children are not only valuable to parents for future income but also as a source of current income. Higher population growth leads to more resource depletion, which reduces the marginal productivity of the resource. To offset this effect families have more children, thus depleting the resource even more. In a short parable Nerlove demonstrates that increasing the birthrate when the environment deteriorates will be the optimal choice for a household that maximizes per capita harvest when 'the perceived marginal product of an additional child increases as environment deteriorates.' In addition, in a system in which land rights are acquired through cultivation, a large number of children can imply increased claim to land ownership. This positive feedback between fertility and environmental degradation is in stark contrast to Malthusian dynamics, where population equilibrates with resources. Since the marginal private and social costs of reproduction are not the same, Nerlove then shows how social interventions in the form of taxes or subsidies can be used to induce a specific birthrate corresponding to a socially desirable stationary state.

The underlying mechanism of the above-mentioned vicious circle is neither population growth nor environmental degradation but poverty, which prevents the substitution of alternative fuel sources, and the low status of women and girls, which devalues the large amount of time and effort that they must devote to daily gathering activities. To escape this vicious circle one needs to alleviate poverty and to educate women.

As suggested by Chu (1998), the increasing specialization and level of economic activity typical of advanced societies may also imply a substantial dependency of population dynamics on the state of the environment. Increasing the level of human activities destroys the earth's biodiversity, so that the ability of the environment to absorb negative shocks will decline, while increased specialization makes people less adaptive to environmental shocks. Chu suggests a simple mathematical model of the survival probability $f(\cdot, \cdot)$ to capture both effects:

$$f(m_t, x_t) < f(m'_t, x_t) \quad \text{if } x_t > 0 \quad (6)$$

$$f(m_t, x_t) \geq f(m'_t, x_t) \quad \text{if } x_t \leq 0 \quad (7)$$

where m_t denotes the logarithm of the number of intermediate goods (a measure of the division of labor) with $m_t < m'_t$ and x_t is a realization of the environment value X_t with the critical state $x_t = 0$. The mean and the variance of the random variable X_t will be increasing functions of m_t . By reducing the ability to adapt, increased specialization m'_t will decrease the

survival probability if the environment is already in a disastrous state $x_t < 0$, while the opposite relation will hold if the environmental state is favorable $x_t > 0$.

4. Concluding Assessment of the Literature

Although there is a growing awareness of the population/environment linkage, as evidenced by the increasing number of empirical (case) studies, research has not progressed much in the Malthusian spirit of modeling the complex interaction in simple mathematical cartoons. Even worse, dynamic interactions between population and environment are often neglected—or increasingly complex simulation models are set up where anything can happen.

Further research on this topic is urgently needed. The concern should be shifted away from physical limits to growth towards the ability of social units to respond to environment/population linkages. Although they are more difficult to formalize, such models should also include organizational and institutional issues. The study by Angelsen (1999) provides a first step in this direction.

See also: Development, Economics of; Development: Sustainable Agriculture; Environmental and Resource Management; Environmental Challenges in Organizations; Environmentalism: Preservation and Conservation; Malthus, Thomas Robert (1766–1834); Sustainability Transition: Human–Environment Relationship; Sustainable Development

Bibliography

- Angelsen A 1999 Agricultural expansion and deforestation: modelling the impact of population, market forces and property rights. *Journal of Development Economics* **58**: 185–218
- Boserup E 1981 *Population and Technological Change: A Study of Long-Term Trends*. University of Chicago Press, Chicago
- Brander J A, Taylor M S 1998 The simple economics of Easter Island: a Ricardo–Malthus model of renewable resource use. *The American Economic Review* **88**(1): 119–38
- Chu C Y C 1998 *Population Dynamics: A New Economic Approach*. Oxford University Press, New York
- Cohen J E 1995 Population growth and earth's human carrying capacity. *Science* **269**: 341–6
- Eckstein Z, Stern S, Wolpin K I 1988 Fertility choice, land, and the Malthusian hypothesis. *International Economic Review* **29**(2): 353–61
- Ehrlich P R, Holdren J P 1971 Impact of population growth. *Science* **171**: 1212–7
- Hardin G 1968 The tragedy of the commons. *Science* **162**: 1243–8
- Lee R D 1986 Malthus and Boserup: a dynamic synthesis. In: Coleman D, Schofield R (eds.) *The State of Population Theory: Forward from Malthus*. Basil Blackwell, Oxford, UK pp. 96–130
- Lee R D 1990 Comment: the second tragedy of the commons. *Population and Development Review* **16** (Suppl.): 315–22

- MacKellar F L, Lutz W, Prinz C, Goujon A 1995 Population, households, and CO₂ emissions. *Population and Development Review* 21(4): 849–65
- Meadows D H, Meadows D L, Randers J, Behrens W W 1972 *The Limits to Growth*. Universe Books, New York
- Milik A, Prskawetz A, Feichtinger G, Sanderson W C 1996 Slow-fast dynamics in Wonderland. *Environmental Modeling and Assessment* 1: 3–17
- Nerlove M 1991 Population and the environment: a parable of firewood and other tales. *American Journal of Agricultural Economics* 73: 1334–47
- Preston S H 1996 The effect of population growth on environmental quality. *Population Research and Policy Review* 15: 95–108
- Prskawetz A, Feichtinger G, Wirl F 1994 Endogenous population growth and the exploitation of renewable resources. *Mathematical Population Studies* 5(1): 87–106
- Repetto R, Holmes T 1983 The role of population in resource depletion in developing countries. *Population and Development Review* 9(4): 609–32
- Sanderson W C 1994 Simulation models of demographic, economic, and environmental interactions. In: Lutz W (ed.) *Population–Development–Environment: Understanding their Interactions in Mauritius*. Springer, Berlin, pp. 33–71
- Wexler L 1996 Decomposing models of demographic impact on the environment. *Working Paper WP-96-85*, IIASA, Laxenburg, Austria

A. Fürnkranz-Prskawetz

Population Dynamics: Momentum of Population Growth

The momentum of population growth refers to the tendency of continued growth of a population with a history of high fertility after its fertility has fallen to replacement level. A population with sustained high fertility is growing in size and has a ‘young’ age distribution, that is it has a high proportion of young people. If such a population adopts replacement-level fertility, the population continues to grow before its size stabilizes. The eventual population is a stationary population, i.e., one whose size and age distribution remain constant over time. This ultimate stationary population has an older age distribution than the initial population, and its size is larger. The ratio of the eventual population to the pretransition population size is defined as the momentum of population growth.

1. Momentum Under Specific Assumptions

In a seminal paper, Keyfitz (1971) presented a simple mathematical expression for the population momentum under specific assumptions. Keyfitz assumed that the initial population has a stable age distribution, and that its fertility falls instantaneously to replacement-level fertility. The drop in fertility level is assumed to occur proportionately at all ages

so that the age pattern of fertility remains unchanged. Mortality is assumed to be constant at all times. Under these assumptions, the population momentum M can be expressed as

$$M = \frac{be_0}{r\mu} \frac{R_0 - 1}{R_0} \quad (1)$$

where R_0 is the net reproduction rate (NRR), b is the birth rate, and r is the growth rate, all of the initial stable population, e_0 is the expectation of life at birth and μ is the mean age at childbearing in the stationary population, i.e., the mean of the net maternity function. Equation (1) provides a simple, closed-form expression that permits calculation of an exact population momentum from a few known summary measures of the initial stable population. This avoids the need to project the initial population for a long time until it stabilizes at zero growth as was done by Frejka (1973).

In observed populations, each of the four assumptions is rarely satisfied, and the resultant momentum will be larger, due mainly to delayed fertility transition and to decline in mortality. The Keyfitz momentum M in Eqn. (1) can then be considered as a lower bound of the extent of future population growth. The formula has been used widely, and it has been shown that initial populations with high growth rates produce a substantial population momentum, frequently around 1.6. That is, the initial population grows by 60 percent before its size stabilizes at a constant value even when the most optimistic fertility reduction scenario is assumed.

2. The Anatomy of Population Momentum

The continuation of population growth after an instantaneous drop in fertility to replacement level is due to the large number of women at childbearing ages in the initial population. To understand what ‘causes’ the growth of a population after a fertility drop, the population momentum can be written as a product of two ratios: (a) the ratio of the numbers of births in the ultimate to the initial population, and (b) the ratio of ultimate to initial population size when both have the same number of births. The first factor is the stable-equivalent number of births, and its value is approximately equal to $1/\sqrt{R_0}$, a value less than one for an initially growing population. The second factor is given by be_0 , and is approximately equal to R_0 . Since the product of the two ratios yields the population momentum, the population grows approximately by a factor of $\sqrt{R_0}$ if the pretransition population had an NRR of R_0 (Frauenthal 1975, Kim et al. 1991). Thus the growth of the population is due entirely to the eventual old age distribution relative to the initial young age distribution, but is *not* due to the larger number of births after the fertility drop. The number of births drops when fertility rates drop in spite of the large number of young people in the initial population.

Preston (1986) showed that when fertility drops to replacement level, the population segment below some age near the mean age at childbearing remains virtually constant, so that all of the growth is confined to the population segment above that age, typically around 28. Inspired by this insight, Kim et al. (1991) examined the anatomy of the population momentum in detail. It was shown that the momentum of *any* observed population is given as the ratio of the *proportion* of the observed population under the mean age at childbearing to the *proportion* of its life table population under that age. Thus, the younger the initial population, the larger is the momentum.

The age-specific momentum for broad segments of a population was also presented. Approximately, (a) the momentum for the segment of a population below age 30 is 1, that is the under-30 population size remains constant, confirming Preston's (1986) result; (b) the momentum for age group 30 to 60 is R_0 ; and (c) the momentum for the group above age 60 is R_0^2 . The exact age below which the population size remains constant can be determined exactly (Kim et al. 1991, Kim and Schoen 1993), but age 30 works well for all practical purposes. To illustrate the age-

specific momentum graphically, the trajectory of population segments in five-year age groups after the immediate fertility drop is shown in Fig. 1a (initial $R_0 = 2.9$ in this example). The number of births (here represented by the population size of ages under five) initially drops when fertility falls, then recovers somewhat before it stabilizes at a new level, which is lower than the original level. In this process, the population segment under age 30 (below the lower solid line) remains virtually constant except for a small fluctuation around time 30, which is when the smaller birth cohorts born right after the fertility drop reach that age. The total population growth of some 60 percent after the fertility drop is due entirely to the increase in the population segment above age 30.

Because the total population growth was contributed solely by the population above age 30 while the population under age 30 remains fixed in size, the population momentum and population aging occur simultaneously. The aging process of the population is shown clearly in the percent age distribution of Fig. 1b. The proportion under age 30 makes up 70 percent of the initial population but it decreases to a little over 40 percent in the eventual population, while the population above age 60 increases from a mere 4 percent to almost 20 percent, showing a substantial degree of aging of the population. The results described here hold for any initial population that is not stable, as well as for initially stable populations.

3. Population Momentum and Population Aging

Population momentum and population aging during the transition to stationarity are two aspects of the same phenomenon. The population momentum expresses population aging, and vice versa. The two quantities are separate aspects of the long-term change in size and age structure that results from a reduction in fertility to replacement level.

To make the relationship between the dual consequences of fertility decline—population momentum and population aging—more specific, three measures of aging were considered (Kim and Schoen 1997). They are: (a) the increase in the mean age of population (ΔA), (b) the relative size of the proportion under age 30 (F_{30}), and (c) the relative size of the proportion above age 65 (F_{65+}). The first measure, ΔA , is defined as the increase in mean age of the eventual stationary population from that of the initial population. The second measure, F_{30} , is defined as the ratio of the population under age 30 in the initial population to that in the ultimate stationary population, and the third measure, F_{65+} , is defined as the ratio of the proportion over age 65 in the initial population to that in the ultimate stationary population.

The three measures of aging are highly correlated with each other and with the population momentum, for these four quantities are linear functions of the

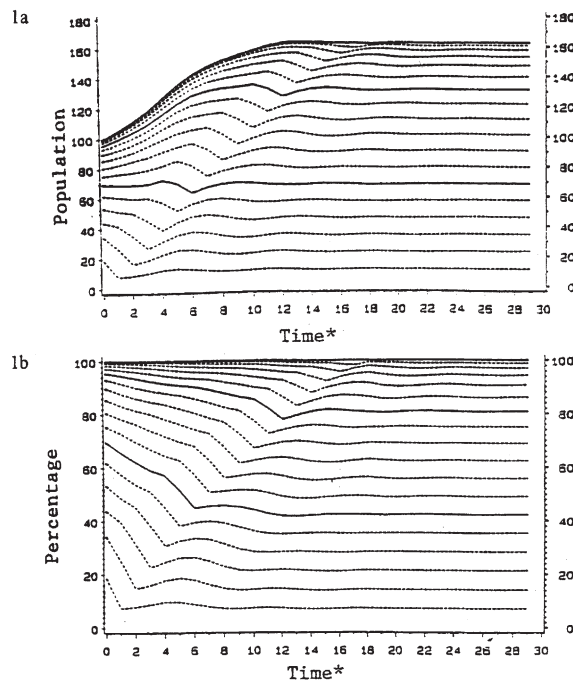


Figure 1

Age distributions and the momentum of the observed population of Singapore (1957): $R_0 = 2.9$. (a) Evolution of population segments from 0 to 90, in five-year age groups. Solid lines represent ages 30, 60, and 90. (b) Population segments from 0 to 90, in five-year age groups as percentages of total population. Solid lines represent ages 30, 60, and 90. *One time unit is five (5) years

intrinsic growth rate, r , of the initial population. The theoretical linear relationships as well as some empirical results are presented in Kim and Schoen (1997). For empirical relationships, two data sets are used. The first is a set of Coale-Demeny West model stable populations (see *Demographic Models*) that includes 39 stable populations with the intrinsic growth rate r ranging from -0.01 to 0.05 in increments of 0.005 at three mortality levels (e_0 values of 60, 70, and 80). The second data set is comprised of 41 observed populations in 1985 (Keyfitz and Flieger 1990).

Figure 2 presents the empirical two-way scatterplots for the momentum, M , and each of the three aging measures generated by the first data set. The first plot (from left to right) shows the relationship between M and ΔA , the second M 's relationship with F_{30} , and the third with F_{65+} . The momentum-mean age linear fit is quite good, and the estimated slope (as well as the theoretical value) suggests that the mean age of the population increases by a little more than two years for every 10 percent increase in population momentum. The middle plot shows that the relative size of the proportions under age 30 is not only linear, but is *identical* to the momentum. This is a direct consequence of the fact that the population size under age 30 remains constant before and after the transition, as discussed in Sect. 2. The third plot presents the relationship between the momentum and the relative size of the proportion above age 65. The fit is not as good as with the other two measures. The theoretical exposition reveals that, for this measure, the contribution from the quadratic term is not negligible, and that the coefficient of the linear term is affected by mortality levels, especially at the high end of the initial growth rate.

Thus, although the proportion over age 65 has great appeal as a conventional index of aging, it performs the poorest among the three measures of aging considered. The increase in the mean age of the population is simple and performs well, both theoretically and empirically. However, the ratio of the proportion under age 30 in the initial population to that in the ultimate population gives the population momentum *directly*.

As Fig. 2 illustrates, these linear relationships hold for *negative* momentum of population decline as well as for positive momentum of population growth. An initial stationary population with $r=0$ bridges the negative and positive momentum seamlessly.

4. Momentum Under Gradual Decline of Fertility

A population's growth potential is significantly underestimated when fertility is assumed to drop instantaneously to replacement level because no population can achieve such an abrupt fertility reduction. If fertility declines gradually over time, the resultant momentum is larger.

If an immediate fertility reduction is adopted by cohorts, the period NRR declines gradually (Frauenthal 1975). Suppose that cohorts born before time 0 maintain the old fertility level at a NRR of R_0 , while all cohorts born after time 0 reproduce at replacement-level fertility. Then the time trajectory of the period NRR resembles a declining logistic curve that begins to decline at time 15 (the beginning age of childbearing), and decreases to 1 at time 45 (the maximum age at childbearing). The *extra* momentum generated by this delayed fertility transition is R_0 . Thus the eventual population momentum under the assumption of a sudden drop in *cohort* fertility is given by $M_c = M \cdot R_0$.

Gradual fertility declines in period fertility have also been examined. Schoen and Kim (1998) consider a linear decline in the growth rate of births to 0 in a given period, during which the number of births grows to a higher level. In this case, NRR reaches replacement level a generation *after* the number of births stabilizes, and the time path of NRR decline is a smooth curve also resembling a declining logistic function. Li and Tuljapourkar (1999) consider a linear path of NRR decline, and derive the exact momentum when the duration of fertility transition is short (e.g., less than 15 years), then show numerically that it is a good approximation for as long as 40 years.

The two formulations give consistent results. The extra momentum increases exponentially as the duration of the fertility transition increases. More specifically, if the *period* NRR declines monotonically to replacement level over a duration, D , the *extra* momentum is larger by a factor $e^{Dr/2} \approx (\sqrt{R_0})^{D/\mu}$ (Schoen and Kim 1998). This is consistent with intuition: it can be interpreted as the growth at half the original growth rate during the given period, or as the growth at the original growth rate for half the duration. Therefore the eventual population

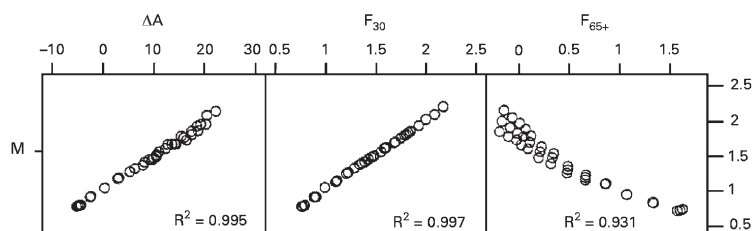


Figure 2

Scatterplots and coefficients of determination (R^2) for M , ΔA , F_{30} , and F_{65+} , coale-demeny female stable population

Thus the momentum generated by a gradual period fertility transition over 30 years is approximately equal to R_0 , equally contributed (multiplicatively) by the age distribution and the delayed fertility transition. The momentum under the sudden *cohort* fertility transition can also be explained in this vein. The extra population momentum, shown to be R_0 , is contributed equally by the growth during the first 15 years at the original growth rate, and during the next 30 years with declining *period* fertility. The growth generated by a *sudden cohort* fertility transition that takes 45 years is equivalent to that of a *gradual period* fertility transition over 60 years.

The extra momentum generated by delayed transition is not sensitive to variations in the time path of fertility decline. A comparison of the momentum generated by a linear decline in the period NRR and that generated by the logistically declining paths (produced by the scenarios of Frauenthal 1975 and of Schoen and Kim 1998) shows that the difference is small (Li and Tuljapurkar 1999). The extra momentum is determined mainly by the number of years a population needs to complete a smooth fertility transition.

5. Future Population Growth

In practice, future population growth is determined by improvements in mortality and the pattern and extent of migration, as well as by the future trends in fertility decline and by the current age distribution. The population growth that results from urbanization of a highly rural population in which the rural sector has higher fertility than in the urban sector is called spatial momentum (Rogers and Willekens 1978, Schoen and Kim 1993). Modern projection/prediction studies show that as the world's populations complete demographic transitions, the momentum inherent in the current young age distribution is the major component of future population growth in most developing countries (Bongaarts and Bulatao 1999). The other components of population growth—the spatial momentum, and the growth due to changes in fertility and mortality—under realistic assumptions is shown to be less than the age momentum (Bongaarts and Bulatao 1999, Brockerhoff 1999). Thus the growth associated with the young current age distribution in developing countries will be the major contributor to the future growth of the world population.

See also: Demographic Transition, Second; Fertility Change: Quantum and Tempo; Population Dynamics: Probabilistic Extinction, Stability, and Explosion Theorems; Population Dynamics: Theory of Non-stable Populations; Population Dynamics: Theory of Stable Populations

Bibliography

- Bongaarts J, Bulatao R A 1999 Completing the demographic transition. *Population and Development Review* **25**: 515–29
- Brockerhoff M 1999 Urban growth in developing countries: A review of projections and predictions. *Population and Development Review* **25**: 757–78
- Frauenthal J 1975 Birth trajectory under changing fertility conditions. *Demography* **12**: 447–54
- Frejka T 1973 *The Future of Population Growth: Alternative Paths to Equilibrium*. Population Council, New York
- Keyfitz N 1971 On the momentum of population growth. *Demography* **8**: 71–80
- Keyfitz N, Flieger W 1990 *World Population Growth and Aging*. University of Chicago Press, Chicago
- Kim Y J, Schoen R 1993 Crossovers that link populations with the same vital rates. *Mathematical Population Studies* **4**: 1–19
- Kim Y J, Schoen R 1997 Population momentum expresses population aging. *Demography* **34**: 421–7
- Kim Y J, Schoen R, Sarma P 1991 Momentum and the growth-free segment of a population. *Demography* **28**: 159
- Li N, Tuljapurkar S 1999 Population momentum for gradual demographic transitions. *Population Studies* **53**: 255–62
- Preston S 1986 The relation between actual and intrinsic growth rates. *Population Studies* **40**: 343–51
- Rogers A, Willekens F 1978 The spatial reproductive value and the spatial momentum of zero population growth. *Environment and Planning, Series A* **10**: 503–18
- Schoen R, Kim Y J 1993 Two-state spatial dynamics in the absence of age. *Theoretical Population Biology* **44**: 67–79
- Schoen R, Kim Y J 1998 Momentum under a gradual approach to zero growth. *Population Studies* **52**: 295–9

Y. J. Kim

Population Dynamics: Probabilistic Extinction, Stability, and Explosion Theorems

1. Extinction, Stable Size, or Explosion?

Historically, probabilistic population theories originate from the *extinction problem*: suppose you know the probabilities of begetting 0, 1, 2, 3,... children. Now determine the probability that a population of a known number of individuals will eventually die out. Already Malthus had noted that a rapid increase of the whole population could occur together with frequent extinction of the separate families constituting the population. Rephrased in the terminology of evolutionary biology, frequent extinction of species does not contradict a rapid growth of the biomass of the world.

The mathematical explanation of this apparent paradox is due to Bienaymé (1845) and independently,

thirty years later and only in parts correctly, to Galton and Watson, who were later to lend their names to the simplest, generation-counting version of the so called branching processes that provide the mathematical structure underlying most stochastic population dynamics. Indeed, if p_k denotes the probability of giving birth to k children, $k = 0, 1, 2, 3, \dots$ then the extinction probability of a population stemming from one ancestor will be the smallest nonnegative solution q of the equation

$$\sum_k p_k q^k = q.$$

If we disregard the trivial case $p_1 = 1$, q will equal 1 if and only if the mean number of children

$$m = \sum_k p_k k \leq 1.$$

If $m > 1$, then the population growth rate per generation will be m , and even though q must be < 1 it can still be pretty large: it is easy to formulate demographically or biologically plausible cases where, say, m exceeds 2 and q is still around 0.8.

If a stochastic population does not die out, it can be proved very generally that it must grow beyond all limits (see Jagers 1991). This is a consequence of stochasticity and the asymmetry between extinction and big sizes: however big the population, there remains an albeit small but still positive risk of extinction, whereas an extinct population cannot be resurrected. This is in contradiction to deterministic population theories which allow population sizes to be stable, cf. Keyfitz 1968.

A finite world can certainly not sustain infinite growth. One can therefore argue that probabilistic population dynamics show that all real populations will die out sooner or later. But before that happens there may be a long period of growth.

2. Population Growth

To describe population growth in real time rather than generation-wise, a more sophisticated description of individual life and reproduction is needed. Probabilistic population dynamics is *individual based* in the sense that it starts from such a description. Deterministic theories, on the other hand, often start directly from an assumed behavior of the population as a whole. Demographic theory is often *pseudoprobabilistic* in that it uses probabilistic notions but makes conclusions only about the expected behavior of the population as a whole. (This is the reason why there exist deterministic populations with stationary sizes: if the underlying stochastic population is critical in the sense that on average each individual is replaced by one child, then the expected total population size will

be stationary, whereas the population itself will die out sooner or later.)

The most general probabilistic formulation of individual life goes as follows: at birth an individual inherits a *type* from her mother. The type determines her survival distribution as well as the distribution of age at childbearing. It also determines the types of the children she will give birth to in her turn. (Think of the type as a genotype, or a genotype combined with the social and biological conditions of life.)

In the *supercritical* case, where there is a positive chance of nonextinction, the probability laws of survival and childbearing determine a positive *Malthusian* parameter, which gives the rate of exponential growth. In the theory of branching processes this rate is usually denoted by α , so that the population will grow like $e^{\alpha t}$, as time t passes. In deterministic theory the conventional notation is often r instead.

Even though exponential growth appears to be a consequence of the independence assumed (tacitly or explicitly) between the reproduction of different individuals, it is interesting to note that such growth has recently been proved also to occur in many cases where individual reproduction depends upon the population as a whole, e.g., its size. For such results in the case of simple Galton-Watson branching processes see Klebaner (1984).

In cases where the reproduction approaches criticality as the population grows, non-Malthusian growth can occur nevertheless, e.g., in the form of linear growth (Klebaner 1994).

3. Stable Population Theory

In contrast to population size, the composition of growing populations tends to stabilize. The description of this composition, or rather the various stable compositions pertaining to different reproduction and survival laws, is the object of stable population theory.

The most well known aspect of stable population composition is the stable age distribution, first described by Euler in 1760 and since then rediscovered repeatedly: if the Malthusian parameter is $\alpha > 0$ and all individuals have the same survival function ℓ , then the age of an individual sampled at random in an old population will have the density

$$\frac{e^{-\alpha a} \ell(a)}{\int_0^\infty e^{-\alpha u} \ell(u) du}$$

at age a , exactly as in deterministic theory. This is, of course, a law-of-large-numbers effect; we are considering ever larger populations.

If, as in the general setup mentioned above, there are several types of individuals, each with its own life span distribution, then the situation will be more

complex. First of all, the type distribution will converge to a stable type law, determined by an eigenmeasure property. Then the survival function ℓ above will have to be replaced by an average of all the different survival functions weighted by the stable type distribution.

But the stable composition describes many other aspects of a balanced growing population besides ages. For example the probability of being first-born can be calculated as well as other aspects of family structure. In modern evolutionary biology the genealogy of individuals and, say, the mutational history plays an important role. In an old population this is described by a stable, typical ancestry, which can be used to determine time back to some event like divergence between species.

See also: Demographic Analysis: Probabilistic Approach; Population Dynamics: Classical Applications of Stable Population Theory; Population Dynamics: Momentum of Population Growth; Population Dynamics: Theory of Stable Populations

Bibliography

- Athreya K B, Ney P 1972 *Branching Processes*. Springer-Verlag, Berlin
- Harris T 1963 *The Theory of Branching Processes*. Springer-Verlag, Berlin
- Jagers P 1975 *Branching Processes with Biological Applications*. Wiley, Chichester
- Jagers P 1991 The growth and stabilization of populations. *Statistical Science* 6: 269–83
- Keyfitz N 1968 *Introduction to the Mathematics of Population*. Addison-Wesley, Reading, MA
- Klebaner F C 1984 Geometric rate of growth in population size dependent branching processes. *Journal of Applied Probability* 21: 40–9
- Klebaner F C 1994 Asymptotic behaviour of Markov population processes with asymptotically linear rate of change. *Journal of Applied Probability* 31: 614–25

P. Jagers

Population Dynamics: Theory of Nonstable Populations

1. Background

To introduce the theory of nonstable population dynamics, we start by presenting a Malthusian model studied by Lee (1974). Consider an overlapping-generation framework in which each individual lives

one or two periods. The first period is childhood and the second is adulthood, and all surviving adults will be in the labor force. The Malthusian model can be characterized by the following two equations: $W_t = f(N_t)$, $b_t = g(W_t)$, where W_t is the wage rate (at time t), N is the size of the adult group, b is the crude birth rate, and $f(\cdot)$ and $g(\cdot)$ are two functions. The $f(\cdot)$ function characterizes the wage/employment relationship in the labor market, and the $g(\cdot)$ function specifies the influence of the labor market reward on fertility.

Suppose the number of children born in period t is B_t , and the child survival rate is l , then $N_t = l \cdot B_{t-1}$. The above two equations can be combined in the following reduced form: $B_t = h(B_{t-1})$, where $h(B_t) \equiv g(f(l \cdot B_{t-1})) \cdot l \cdot B_{t-1}$. The above expression is a recursive equation of B_t , from which the dynamic pattern can be analyzed, and the possible existence of cyclicity can be studied. Here the Malthusian model is interpreted as a model of density dependency, where the density is characterized by the birth size of the previous period. It turns out that the Malthusian model described above may not converge to a ‘dismal’ steady state, as Malthus suggested. Mathematically, the B_t series may converge, diverge, or be cyclic, and the volatility of the birth series B_t crucially hinges on the feedback elasticity of $h(\cdot)$ with respect to B_{t-1} . When this elasticity is sufficiently large, the B_t series can easily have cycles. This is the classical case of nonstable population. Modern versions of nonstable population will be discussed below, after we introduce the general population structure.

2. The General Population Structure

Let $N_t \equiv (N_{t,1}, \dots, N_{t,n})$ be an n -type population vector at period t , where $N_{t,i}$, $i = 1, \dots, n$ indicates the size of type- i population at period t . The type here may refer to any criterion that is used to classify the population, such as age, sex, and occupation. A general formulation of population dynamics can be written as follows:

$$N_{t+1} = Q(N_t)N_t, \quad (1)$$

where $Q(N_t)$ is the transition matrix. The simplest case is when $Q(N_t) = Q$, a time-independent matrix. In this situation, the above equation reduces to a time-invariant Markov chain: $N_{t+1} = QN_t$. It is a well-known implication of the Frobenius–Perron theorem that if Q is positively regular, meaning that there is flexible cross-period mobility across types, then the population vector N_t will converge to a steady state, in which all elements of the N_t vector grow at the same exponential rate, thereby the population composition is also time-invariant. It has been shown by Chu (1998) that, for most population models with time-

independent Q , this positive regularity condition is likely to be satisfied, therefore any nonstable behavior of the population would appear only when $Q(N_t)$ is not a constant matrix, or equivalently when the relationship between N_{t+1} and N_t in Eqn. (1) has some nonlinearity. Below we shall introduce several special cases of the generic setting in Eqn. (1), and discuss their empirical relevance. The reader can see that each model corresponds to a special interpretation of the generic setting in Eqn. (1).

3. Easterlin Cycles

For demographers, the most familiar version of the dynamic models in Eqn. (1) is the Lotka–Leslie age-specific models, where the type criterion is age. Let $N_{t,i}$ be the number of population aged i at period t , B_{t-i} be the birth size at period $t-i$, and l_i be the probability that a person can survive to age i , then the following relationship must hold: $N_{t,i} = B_{t-i} l_i$. The transformed dynamics in terms of B_t is the well-known Lotka's equation:

$$B(t) = \sum_i B_{t-i} \cdot l_i \cdot m_{t,i} = \sum_i B_{t-i} \cdot \phi_{t,i} \quad (2)$$

where $m_{t,i}$ is the average number of births per surviving member aged i at period t , and $\phi_{t,i} \equiv l_i \cdot m_{t,i}$ is the net maternity function. Evidently, the Malthusian model presented in the previous section is a special case of Eqn. (2a) with two age types. As we mentioned, when the fertility function $m_{t,i}$ is independent of t , then the process of B_t follows a time-invariant Markov chain, and the usual convergence result applies. But if $m(t,i)$ is time dependent, then cycles or more volatile behavior may appear. One typical case of such time dependency is when the age-specific fertility rates are functions of previous birth sizes: $m_{t,i} = m_i(B_t)$, where $B_t \equiv (B_{t-1}, B_{t-2}, \dots)$.

Suppose the birth series corresponding to Eqn. (2a) has a steady state. Taking a Taylor expansion around this steady state, Lee (1974) showed that the volatility of the birth series characterized by Eqn. (2a) hinges upon the elasticities of the net maternity function with respect to B_t . When previous birth sizes are larger, usually the density pressure is larger. Intuitively, larger feedback elasticities imply that present reproduction would be more sensitive to previous birth shocks, hence the birth series would be more volatile. Easterlin (1961) argued that the baby-boom/baby-bust cycles observed in the United States are typical examples of a feedback cycle.

How large the feedback elasticities are is an empirical question. In order to estimate such elasticities, two models have been proposed: the first is to assume that $\phi_{t,i}$ is only a function of the cohort size aged i : $\phi_{t,i} = \phi_i(B_{t-i})$, and the second is to assume that $\phi_{t,i}$ is a function of the weighted average of birth sizes of different ages: $\phi_{t,i} = \phi_i(\sum_j w_j \cdot B_{t-j})$. The former is

referred to as the cohort model, and the latter as the period model. However, empirical evidence has so far failed to back up either of these two models in producing persistent birth cycles that fit all characteristics (such as amplitude and period). Moreover, even if there does exist a cyclic solution to Eqn. (2a), the information concerning feedback elasticity is not sufficient for us to tell whether the cyclical solution in question is stable. Related analysis is rather technical, and the reader is referred to Tuljapurkar (1987) for details.

4. Predator–Prey Cycles

The predator–prey model describes a two-type population structure, where the types are prey and predator, respectively, referred to as type-1 and type-2. The simplest formulation of the predator–prey model is characterized by the following difference equations:

$$\begin{aligned} \Delta N_{t+1,1} &\equiv N_{t+1,1} - N_{t,1} \\ &= aN_{t,1} - bN_{t,1}N_{t,2} \end{aligned} \quad (3a)$$

$$\begin{aligned} \Delta N_{t+1,2} &\equiv N_{t+1,2} - N_{t,2} \\ &= -cN_{t,2} + dN_{t,1}N_{t,2}, \\ a, b, c, d &> 0. \end{aligned} \quad (3b)$$

Equation (3) is obviously a special case of Eqn. (1). The prey are considered the only food resource available to the predator. Thus, if $N_{t,1} = 0$, the predator population decreases exponentially at the rate c . With the nonexistence of predators, the prey population grows exponentially at the rate a .

In the steady state of Eqn. (3), $\Delta N_{t,1} = \Delta N_{t,2} = 0$, there is only one nontrivial solution: $(N_1^*, N_2^*) = (c/d, a/b)$. Let the Jacobian matrix of Eqn. (3) around the nontrivial solution be denoted J . It turns out that the determinant of J is

$$J = \begin{pmatrix} a - bN_2^* & -bN_1^* \\ dN_2^* & -c + dN_1^* \end{pmatrix} = \begin{pmatrix} 0 & -bc/d \\ ad/b & 0 \end{pmatrix}$$

$|J| = ac > 0$ and the trace of J is zero.

Thus, we know that the eigenvalues of J are purely imaginary, meaning that besides the equilibrium point $(c/d, a/b)$ itself, the solution trajectories of Eqn. (3) are cyclic and nonstable. But it was also pointed out that when the predator–prey model embodies the concerns of diminishing or increasing returns, then the solution becomes locally stable or divergent.

5. Dynastic Cycles

Based on stylized facts found in Chinese history, Chu and Lee (1994) proposed an occupation-specific population model, in which the population is separated into

three occupations: bandits, peasants, and soldiers. Peasants grow crops and pay taxes, soldiers (rulers) collect taxes and hunt for bandits, and bandits rob food and clash with peasants and soldiers. It is assumed that soldiers are drafted by the government, and all other civilians can choose to become peasants or bandits. Their occupation choice is made depending on which occupation is expected to generate a higher utility. When the bandit/soldier ratio is relatively high, it is likely that the incumbent power regime will be overthrown by the bandit group, and the dynasty switches. It turns out that this pattern of population dynamics characterizes the dynamic evolution of dynasties in Chinese history, and the cycles were called dynastic cycles by historians.

In terms of demography, the occupation switches mentioned above are merely changes in population composition. But there is a unique feature associated with the above-mentioned occupation switching, which we shall explain below. For most animals, it is believed that when density pressure occurs, the environment becomes less favorable, and the reproduction rate of animals is reduced. One might expect that rational human decisions may be able to weaken the outside density pressure through institutional and rational regulation. But the scenario of dynastic cycles seems to be a counterexample. As noted by historians, clashes between soldiers and rebellious peasants in human history were often initiated by famine or density pressure. The higher the bandit ratio, the less likely a peasant can safely keep a good proportion of his crop, and hence the greater attraction for the peasant to join the bandit group and seize other people's crops. Thus, human beings' occupational choice between peasants and bandits has a 'demonstration effect,' which tends to magnify the originally weak density pressure and 'destablize' the dynamics of the population compositional structure.

The cyclical pattern of population composition mentioned above is not the only case we find in human history. Some researchers have considered a neo-classical economic growth model, and focused on the two economic classes in the economy: the laborers and the capitalists. Let u_t be the income share attributed to labor, and v_t be the employment rate of labor. Under some reasonable assumptions, Goodwin (1967) has shown that the dynamics of u_t and v_t mimic that of Eqn. (3). Thus, the conclusion that the capitalists' economy appears to be 'permanently oscillating' was reached. But just as the original predator-prey model is sensitive to variations in institutional specifications, Goodwin's model after some minor modification also generates qualitatively different results. Definite evidence of the nonstability of human population has yet to be found.

See also: Demographic Models; Population Cycles and Demographic Behavior; Population Dynamics:

Classical Applications of Stable Population Theory; Population Dynamics: Theory of Stable Populations

Bibliography

- Cyrus Chu C Y 1998 *Population Dynamics: A New Economic Approach*. Oxford University Press, New York
 Cyrus Chu C Y, Lee R D 1994 Famine, revolt and the dynastic cycle: Population dynamics in historic China. *Journal of Population Economics* 7: 351–78
 Easterlin R A 1961 The American baby boom in historical perspective. *American Economic Review* 51: 860–911
 Goodwin R M 1967 A growth cycle. In: Feinstein C H (ed.) *Socialism, Capitalism and Economic Growth*. Cambridge University Press, Cambridge, UK
 Lee R 1974 The formal dynamics of controlled population and the echo, the boom and the bust. *Demography* 11: 563–85
 Tuljapourkar S 1987 Cycles in nonlinear age-structured models I: Renewal dynamics. *Theoretical Population Biology* 32: 26–41

C. Y. Cyrus Chu

Population Dynamics: Theory of Stable Populations

A population with an age distribution which remains constant over time is referred to as 'stable.' A stable population of constant size is said to be 'stationary.' These concepts are of considerable antiquity. Euler, for example, knew in 1760 the stable age distribution a population needed to have if it were to have a given growth rate. Compilers of some of the early life tables such as Graunt in 1662 appear to have assumed populations which were stationary, and nineteenth-century actuaries certainly made use of the stationary population concept (King 1902).

Subject to age-specific mortality and fertility rates which remain unchanged over time a population will eventually develop a stable age distribution which depends on those mortality and fertility rates but is independent of the initial age distribution of the population. This result, due to Sharpe and Lotka (1911) caused renewed interest in stable populations among demographers. Although demographers usually think of a stable population as one with an age distribution which is unchanging, such a definition is restrictive, because populations can also exhibit stability in respect of other characteristics as well as age and in more general situations.

1. The Population Model of Sharpe and Lotka

Sharpe and Lotka considered the male population alone and assumed that the growth of the female population would be such as to justify assumptions of

constant age-specific fertility and mortality for the males. Subsequent writers have usually applied the model to the female component of the population because of its shorter reproductive age span and the fact that the number of females in the reproductive age range plays a more significant role in determining the number of births than the number of males (Pollard 1973).

Applied to the female population, the model assumes $F(x, t)dx$ females in the age range $(x, x + dx)$ at time t and $B(t)dt$ female births between times t and $t + dt$. Then, if the average number of daughters born per female aged exactly x in time element dt is $\lambda(x)dt$ and the probability that a female survives from age u to age $u + v$ is ${}_v p_u$, the following relationship is immediately apparent:

$$B(t) = \int_0^{\infty} F(x, t) \lambda(x) dx \quad (1)$$

Noting that those alive and aged x at time t , enumerated in $F(x, t)dx$, must either have been born at time $t - x$ ($x < t$) or else have been aged $x - t$ at time 0 ($x > t$), Eqn. (1) may be rewritten as

$$B(t) = \int_0^t B(t-x) {}_x p_0 \lambda(x) dx + G(t) \quad (2)$$

with

$$G(t) = \int_0^{\infty} F(x, 0) {}_t p_x \lambda(x+t) dx \quad (3)$$

The solution to the integral equation (2) reveals that for large values of t , under quite general assumptions

$$B(t) \cong C e^{rt} \quad (4)$$

where the constant r (the *intrinsic growth rate* of the population) is the unique real solution to the integral equation

$$\int_0^{\infty} e^{-rx} {}_x p_0 \lambda(x) dx = 1 \quad (5)$$

and the constant C depends on the initial age distribution of the population as well as the age-specific mortality and fertility rates (Keyfitz 1968, Pollard 1973).

Since the females aged between x and $x + dx$ at time t are the survivors of those born x years earlier between times $t - x$ and $t - x + dx$, it follows from Eqn. (4) that

$$F(x, t) \cong C e^{r(t-x)} {}_x p_0 = (C e^{rt}) (e^{-rx} {}_x p_0) \quad (6)$$

The first factor in the right-hand side of Eqn. (6) indicates that the population grows exponentially asymptotically at rate r , while the second reveals that the age distribution is asymptotically independent of t . In other words, a *stable age distribution* is reached.

2. The Discrete Time Model of Bernardelli, Leslie, and Lewis

Whilst the continuous time model of Sharpe and Lotka is elegant and predicts a stable population asymptotically under conditions of constant age specific mortality and fertility, the discrete time formulation of Bernardelli (1941), Lewis (1942) and Leslie (1945), which makes similar assumptions, is more readily adapted to analyze stability in more complicated situations.

The number of females aged x last birthday at time t is represented by $n_{x,t}$. There is no migration, and no one can live beyond age k last birthday. The average number of daughters born to a female while she is aged x last birthday, those daughters surviving to be enumerated as being aged zero at the next point of time, is F_x , and the proportion of females aged x last birthday surviving to be aged $x + 1$ last birthday a year later is P_x . Under these assumptions, the following matrix recurrence equation must apply to the numbers of females in successive years:

$$\begin{pmatrix} n_{0,t+1} \\ n_{1,t+1} \\ n_{2,t+1} \\ n_{3,t+1} \\ \dots \\ n_{k,t+1} \end{pmatrix} = \begin{pmatrix} F_0 F_1 & F_2 \dots & F_{k-1} F_k \\ P_0 & & \\ & P_1 & \\ & & P_2 \\ & & & \ddots \\ & & & & P_{k-1} \end{pmatrix} \begin{pmatrix} n_{0,t} \\ n_{1,t} \\ n_{2,t} \\ n_{3,t} \\ \dots \\ n_{k,t} \end{pmatrix} \quad (7)$$

This equation may be written more concisely as

$$\mathbf{n}_{t+1} = \mathbf{A} \mathbf{n}_t \quad (8)$$

so that

$$\mathbf{n}_t = \mathbf{A}^t \mathbf{n}_0 \quad (9)$$

Under fairly general conditions, matrix $\mathbf{A}$ is positive regular, in which case it has a dominant eigenvalue α which is positive, of multiplicity one, and greater in absolute size than any other eigenvalue. It follows from Eqn. (9) therefore that for large t ,

$$\mathbf{n}_t \cong C \alpha^t \mathbf{z} \quad (10)$$

where C is a constant which depends on the initial age distribution of the population, and $\mathbf{z}$ is the column eigenvector corresponding to the eigenvalue α . If Π_s is written instead of the product $P_0 P_1 \dots P_s$, the elements of $\mathbf{z}$ are $\Pi_0 \alpha^{-0}, \Pi_1 \alpha^{-1}, \dots, \Pi_k \alpha^{-k}$. It follows, therefore that the number of females aged x last birthday at time t is

$$n_{x,t} \cong (C \alpha^t) (\alpha^{-x} \Pi_x) \quad (11)$$

This is the discrete analogue of the Sharpe and Lotka formula, Eqn. (6). The second factor defines the asymptotic stable age distribution; the first shows the exponential growth of the population (Keyfitz 1968, Pollard 1973).

3. Stability with Migration

The most appropriate model for a population subject to significant migration must depend on the pattern that migration takes. In the case of a country accepting a constant number of immigrants each year with a constant age pattern, for example, the discrete time recurrence of Eqn. (8) needs to be modified as follows:

$$\mathbf{n}_{t+1} = \mathbf{A} \mathbf{n}_t + \mathbf{b} \quad (12)$$

where $\mathbf{b}$ is a constant vector of immigrants. Under this model, immigrants are assumed to adopt immediately the mortality and fertility patterns of their new homeland (see *Immigrant Fertility*). Repeated application of Eqn. (12) for $t = 0, 1, 2, \dots$ reveals that, provided the dominant eigenvalue of $\mathbf{A}$ is not equal to one,

$$\mathbf{n}_t = \mathbf{A}^t \mathbf{n}_0 + (\mathbf{I} - \mathbf{A}^t) (\mathbf{I} - \mathbf{A})^{-1} \mathbf{b} \quad (13)$$

If the rate of natural increase of the population is less than zero, α will be less than one, and for very large t

$$\mathbf{n}_t \cong (\mathbf{I} - \mathbf{A})^{-1} \mathbf{b} \quad (14)$$

The population will become stationary with its age structure determined by the ages of the immigrants (vector $\mathbf{b}$), as well as the mortality and fertility of the host country (matrix $\mathbf{A}$). Where the fertility of the host country is greater than replacement level, α will be greater than one, and the following asymptotic formula follows from Eqn. (13) for very large t :

$$\mathbf{n}_t \cong C' \alpha^t \mathbf{z} \quad (15)$$

In this case, the population grows asymptotically with the same exponential growth rate as it did in the discrete nonimmigration case of Sect. 2. The asymptotic stable age distribution (vector $\mathbf{z}$) is also the same. The constant C' , however, is different, depending on

the initial age structure, the vector of immigrants $\mathbf{b}$, and the mortality and fertility of the population.

In the pathological case where the levels of mortality and fertility in the population correspond exactly to replacement, α is equal to one, and a different algebraic approach is necessary to the solution of Eqn. (12), which yields, for very large t , the following asymptotic result:

$$\mathbf{n}_t \cong C'' t \mathbf{z} \quad (16)$$

The population grows linearly asymptotically and the asymptotic age distribution is the same as in the previous case. The constant C'' depends on the immigration vector, as well as the fertility and mortality rates of the population.

For a population experiencing net emigration with a proportion β_x of females aged x departing each year, the matrix model of Sect. 2 can be utilized with each of the P_x in matrix $\mathbf{A}$ replaced by $P_x - \beta_x$. The mathematical theory can then be applied in the same manner as before, and an asymptotic stable population emerges. The dominant eigenvalue and eigenvector will of course be different from those in Sect. 2. Depending on whether the dominant eigenvalue of the modified matrix $\mathbf{A}$ is less than, equal to, or greater than one, the population will ultimately decline, become stationary or grow in size.

The above models are only two of a very wide range of possible migration models appropriate in different circumstances. They demonstrate how the discrete time model of Sect. 2 may be readily modified to take account of migration and the fact that population stability will usually emerge in the presence of migration when the migration parameters, mortality, and fertility remain unchanged over time.

4. Generalizations

The discrete time models of the earlier sections are readily adapted to study populations partitioned according to other characteristics as well as age. Multiregional demography is the best known example (Rogers 1966, 1975). Females of a population are characterized by age and geographical region in an extended column vector $\mathbf{n}_t$ of dimension mk , where k is the maximum age last birthday attainable and m is the number of regions in the country. As well as surviving and reproducing in the manner of Sect. 2 with rates that may differ from region to region, females can move from one region to another with transition rates which depend on the regions involved and may also depend on age. The matrix recurrence equation is again Eqn. (8), but with $\mathbf{n}_t$ of dimension mk and $\mathbf{A}$ of dimension $mk \times mk$. Conditions for a unique dominant eigenvector are readily determined, and asymp-

totic stability for the population by age and geographic region is predicted.

Another partition characteristic sometimes proposed is the *parity* of woman. Females at time t are enumerated in a vector $\mathbf{n}_t$ with elements $\{n_{x,y,t}\}$ which record the numbers aged x last birthday who already have y children. The only possible transitions for a female aged x and parity y remaining within the system are to age $x+1$ and parity y or to age $x+1$ and parity $y+1$ (ignoring the small possibility of multiple births). Where there is a change in parity, a new female will be introduced into the population aged 0 last birthday and parity 0 if the mother's change in parity corresponds to a female birth. Recurrence matrix $\mathbf{A}$ contains these transition rates. The fundamental recurrence equation is again Eqn. (8). For all realistic populations, the enlarged matrix $\mathbf{A}$ will have a unique dominant eigenvector, and the population enumerated by age and parity will approach stability (Keyfitz 1968).

The discrete-time models of Sects. 2 and 3 have also been applied to workforce planning (Bartholomew 1973, Pollard 1967), membership of learned societies (where Fellows are elected on merit and there is a danger of the society becoming a rapidly aging institution), and membership of other groups including pension schemes (Sherris and Pollard 1980). A learned society, for example, might have a policy of electing five new Fellows each year. The membership of the society at time t can be summarized by a vector $\mathbf{n}_t$ with elements $\{n_{x,t}\}$ representing the numbers at the various ages at that time. The number of new members (five) and their age distribution can be summarized in a vector $\mathbf{b}$. The transition rates in the recurrence matrix are simply the proportions $\{P_x\}$ surviving from age x last birthday to age $x+1$ last birthday a year later, for all relevant ages. Assuming that the ages are listed in their natural increasing order, the transition matrix $\mathbf{A}$ is particularly simple in this case: all its elements are zero, except those immediately below the main diagonal which comprise the survival proportions $\{P_x\}$. With $\mathbf{n}_t$, $\mathbf{b}$, and $\mathbf{A}$ defined in this manner, the recurrence equation for the society is Eqn. (12). All the eigenvalues of $\mathbf{A}$ are zero, and for t greater than the number of ages involved, $\mathbf{A}^t = \mathbf{O}$. The society will approach a stationary state given by Eqn. (14).

All such linear models with realistic assumptions concerning transition between states, entry and reproduction can be shown to predict stable populations asymptotically.

5. Applications

No observed population experiences constant levels of mortality and fertility. Nevertheless the concept of a stable population finds frequent application. Keyfitz's elegant formula for the effect of population momen-

tum (see *Population Dynamics: Momentum of Population Growth*) assumes that the population is stable immediately prior to fertility falling to replacement level. Yet the formula works remarkably well for observed populations which are not stable. Stable population theory may be helpful in deducing results in respect of populations with unreliable and/or limited demographic data. Equation (4), for example, can be used to infer an approximate value for the crude birth rate of a population where contraception is not widely practiced, using only a reliable census of the population and a more or less appropriate life table (Bourgeois-Pichat 1957). Many of the applications of model life tables rely on stable population theory (see also *Population Dynamics: Classical Applications of Stable Population Theory*).

6. Limitations

The discrete time model of Bernardelli, Leslie, and Lewis, like the continuous (female) formulation of Sharpe and Lotka, assumes that the development of the male component of the population is such as to justify the assumption of unchanging female age-specific mortality and fertility. Linear models which include the male component of the population alongside the females have also been proposed, and mathematical analysis analogous to that presented in Sect. 2 predicts a stable age-gender composition for the whole population. Nonlinear models have also been proposed to accommodate the 'two-sex problem' (see *Population Dynamics: Two-sex Demographic Models*). Many of these nonlinear models would also seem to predict populations which are asymptotically stable under conditions of unchanging rates of mortality, fertility and nuptiality.

For many developing countries, for which stable population theory had been an extremely helpful tool, mortality began to decline in the second half of the twentieth century, but fertility remained essentially unchanged from previous levels. To accommodate the improvements in mortality, the concept of a quasi-stable population was introduced: one with improving mortality but constant fertility (Coale 1963).

Bibliography

- Bartholomew D J 1973 *Stochastic Models for Social Processes*, 2nd edn. Wiley, New York
- Bernardelli H 1941 Population waves. *Journal of Burma Research Society* **31**: 1–18
- Bourgeois-Pichat J 1957 Utilisation de la notion de population stable pour mesurer la mortalité et la fécondité des populations des pays sous-développés *Bulletin de l'Institut International de Statistique* (actes de la 30e Session)
- Coale A J 1963 Estimation of various demographic measures through the quasi-stable age distribution. *Emerging Tech-*

- niques in Population Research: Proceedings of the 1962 Annual Conference of the Milbank Memorial Fund
- Keyfitz N 1968 *Introduction to the Mathematics of Population*. Addison-Wesley, Reading, MA
- King G 1902 *Life Contingencies*, 2nd edn. Charles & Edward Layton, London
- Leslie P H 1945 On the use of matrices in certain population mathematics. *Biometrika* **33**: 183–212
- Lewis E G 1942 On the generation and growth of a population. *Sankhya* **6**: 93–6
- Pollard J H 1967 Hierarchical population models with Poisson recruitment. *Journal of Applied Probability* **4**: 209–13
- Pollard J H 1973 *Mathematical Models for the Growth of Human Populations*. Cambridge University Press, Cambridge, UK
- Rogers A 1966 The multiregional matrix growth operator and the stable interregional age structure. *Demography* **3**: 537–44
- Rogers A 1975 *Introduction to Multiregional Mathematical Demography*. Wiley, New York
- Sharpe F R, Lotka A J 1911 A problem in age-distribution. *Philosophical Magazine, Ser. 6* **21**: 435–8
- Sherris M, Pollard J H 1980 Application of matrix methods to pension funds. *Scandinavian Actuarial Journal* 77–95

J. H. Pollard

Population Dynamics: Two-sex Demographic Models

1. Classical Stable Population Theory

A population is in ‘demographic equilibrium’ if the fraction of individuals in each age-sex category remains constant over time and the population is growing (or declining) at a constant rate. Demographers call a population that is in demographic equilibrium a ‘stable population.’ This use of the phrase stable population blurs the distinction between questions involving the existence of a demographic equilibrium and those involving its dynamic stability; hence, I depart from standard demographic terminology and in its place use the phrase ‘demographic equilibrium.’ In deference to traditional usage, however, I refer to demography’s standard model as ‘classical stable population theory.’

Nondemographers are often surprised to learn that the classical stable population theory operates with a one-sex model. The theory postulates an age-specific fertility schedule for women and an age-specific mortality schedule for women. These two schedules, in conjunction with the very strong assumptions that both of these schedules remain constant over time, imply an equilibrium growth rate and an equilibrium age distribution for the female population.

Demography’s two-sex problem is the problem of integrating men into this one-sex, female-based model. The two-sex problem is often motivated by discussing the effect on fertility of a ‘marriage squeeze.’ A marriage squeeze is an imbalance between the female and male populations of marriageable age brought

about, for example, by wars that disproportionately reduce the number of marriageable men. A marriage squeeze can also result from a baby boom: because women typically marry older men, a baby boom increases the number of marriageable women many years before it increases the number of marriageable men. Common sense suggests that a marriage squeeze is likely to reduce age-specific fertility rates for women, but the classical stable population theory, because it assumes that age-specific fertility rates remain constant, cannot accommodate this common sense conclusion.

2. The Birth Matrix-mating Rule Model

Demography’s two-sex problem is solved, in a formal sense, by the ‘birth matrix-mating rule’ (BMMR) model introduced in Pollak (1986). (Feeney (1972), in an unpublished doctoral dissertation, presents a two-sex model with the same basic structure, but without a satisfactory proof of the existence of equilibrium.) The BMMR model has three primitives (i.e., components not derived from more fundamental components):

- (a) a birth matrix, which indicates the number of offspring born to each possible type of union (e.g., the union of a female of age i and a male of age j);
- (b) a mating rule, which indicates the number of unions of each type as a function of the number of females in each age category and the number of males in each age category; and
- (c) age-specific mortality schedules for males and females.

Under relatively weak assumptions, the BMMR model determines equilibrium growth rates and equilibrium age distributions for the female and male populations. It also determines who marries whom or, more precisely, the age pattern of assortative mating. The BMMR model assumes that all fertility is marital fertility and all unions are monogamous, although both of these assumptions are easy to relax. Because the BMMR model allows changes in the age structure of the population (e.g., shortages of marriageable men relative to marriageable women) to change the number of unions, it accommodates our common sense conclusion about the effect of a marriage squeeze. Throughout a marriage squeeze, the birth matrix (i.e., the number of births to each type of union) remains constant, but the number of unions changes, and these changes in the number of unions imply changes in the number of births and, hence, changes in the implied age-specific female fertility rates.

2.1 Linear and Nonlinear Models

Because the classical stable population theory utilizes a linear model, the mathematics is straightforward. The age-specific fertility and mortality schedules for the female population together define a mapping of the female population vector (by age) in period t

into the female population vector in $t + 1$. Because the implied mapping is linear, demographers use the simple and powerful tools of linear algebra to investigate the existence, uniqueness, and dynamic stability of the equilibrium. The matrix corresponding to the mapping of the population vector in period t into the population vector in period $t + 1$ is called the Leslie matrix. The equilibrium age distribution is an eigenvector of the Leslie matrix; the equilibrium growth rate is obtained by subtracting 1 from the corresponding eigenvalue. In linear demographic models, cases in which an equilibrium fails to exist, cases of multiple equilibria, and cases in which the dynamics do not exhibit global stability are generally regarded as pathological.

The BMMR model, like the classical stable population theory, defines a mapping of the population vector in period t into the population vector in period $t + 1$. The initial population vector, together with the mating rule, determines the number of unions of each type; the number of unions of each type, together with the birth matrix, determines the number of newborns in the next period. The initial population vector, together with the mortality schedules, determines the number of individuals in each of the other age-sex categories, just as in the classical stable population theory.

Nonlinearity is a necessary feature of any two-sex demographic model. The two-sex problem arises precisely because the classical stable population theory fails to recognize that the effect on population in period $t + 1$ of adding a female of age i to the population in period t depends on the number of males. In the BMMR model, this dependence is captured by the mating rule. Any mating rule for monogamous unions must satisfy two accounting requirements: a nonnegativity condition ensuring that the number of unions of each type is positive or zero, and an adding-up or balance condition ensuring that the number of mated females of each age does not exceed the total number of females of that age, with a similar requirement holding for males. This adding-up requirement is the fundamental source of nonlinearity in the BMMR model.

Because the BMMR model is inherently nonlinear, establishing the existence of equilibrium and analyzing dynamic behavior in the BMMR model is more complex than in the classical stable population theory. A 'natural' but unsuccessful strategy for establishing the existence of equilibrium is to apply a fixed-point theorem to the mapping of the population vector in period t into a suitably normalized population vector in period $t + 1$; a fixed point of such a mapping—a population vector that maps into itself—is an equilibrium of the BMMR model. The difficulty with this strategy is that the mapping defined by the BMMR model carries some population vectors into the 0 vector and, hence, this strategy fails to establish the existence of a nontrivial equilibrium. An alternative

proof strategy, followed in Pollak (1986), avoids this difficulty by drastically limiting the domain of the mapping and reducing the problem to a single dimension.

2.2 Uniqueness and Dynamic Stability

Multiple nontrivial equilibria are not pathological in nonlinear models, and Pollak (1990a) demonstrates by example that the BMMR model can have multiple nontrivial equilibria. An overly strong apparently innocuous sufficient condition for the BMMR model to have a unique nontrivial equilibrium is that at every population vector an increase in the number of individuals in each age-sex category does not decrease the number of newborns in the next period.

Uniqueness and dynamic stability are intimately related. In a model with multiple equilibria, initial conditions determine which, if any, of several equilibria will be realized, and small differences in initial conditions can lead to large differences in long-run behavior. Pollak (1990a) discusses four sets of demographically meaningful conditions that imply local stability. Local stability rather than global stability is the relevant issue because, *inter alia*, some initial population vectors must lead to the trivial equilibrium. Further work—both theoretical and empirical—is required to determine whether any of the sufficient conditions for uniqueness and dynamic stability are satisfied and whether multiple equilibria or dynamic instability are realistic possibilities in two-sex demographic models.

Simulation is a possible approach to investigating these issues, but simulation techniques are much less informative in models that lack parametric specifications. In linear models, parametric specification is a feature of the model itself, but to simulate the BMMR model, we must first specify a functional form for the mating rule, and then specify the appropriate parameter values. Explicit functional forms for mating rules are discussed in Pollak (1990b).

3. Behavioral Models

Further work is also required to transform the BMMR model from a formal into a substantive model of population age structure and growth. The transformation requires importing behavioral theories from social science into the BMMR model to explain its three primitives—the birth matrix, the mating rule, and the mortality schedules. (In the generalized version of the model in which unions can persist for more than one period, there is a fourth primitive requiring a behavioral explanation: the schedule specifying the probabilities that unions of each type will end in desertion or divorce: see Pollak 1987.) From the

standpoint of social science, however, these three primitives are different kinds of analytical constructs. The mortality schedule is often regarded as a biological datum, although there is ample precedent (from Malthus to recent concern about excess female infant mortality rates in India) for regarding mortality as endogenous. The elements of the birth matrix reflect the decisions of individuals or of families facing economic and biological constraints. The analysis of these decisions is the subject matter of economic and sociological theories of fertility.

Whether we interpret 'unions' narrowly or broadly, the mating rule of the BMMR model is more complex analytically than the mortality schedules or the birth matrix because it reflects not only individual behavior but also the interactions of individuals in the 'marriage market.' The marriage market metaphor is useful because it suggests notions of choice, competition, and equilibrium. That is, the mating rule is not a full-blown theory of marriage or union formation but a reduced form, presumably reflecting the requirements of equilibrium in the marriage market.

There are two approaches to modeling competition and equilibrium in marriage markets, one based on matching and the other based on search. Mortensen (1988) and Weiss (1997) provide accessible introductions to both matching and search approaches; Roth and Sotomayor (1990) provide the definitive treatment of matching models. Both matching and search models begin with the preferences of individuals and analyze their implications for marriage-market equilibrium. They differ, however, in their assumptions about the information available to and the alternatives faced by marriage-market participants. Matching models are nonstochastic and timeless; they implicitly assume that complete and accurate information about potential spouses is instantaneously and costlessly available to all marriage-market participants. Search models, in contrast, are stochastic and dynamic; they recognize explicitly that individuals have less than complete and accurate information about potential spouses, and that acquiring more and better information takes time and, in some search models, requires resources. In a search model, when two individuals meet in the marriage market, they must decide whether to marry or to continue searching, balancing the attractiveness of marriage now to a particular individual against the expected value of remaining unmarried and continuing to search. Because search models do not assume that individuals have complete and accurate information, they may appear to be more 'realistic' than matching models. The price that search models pay for added realism is their sensitivity to specific assumptions about the rate at which potential spouses meet in the marriage market, the impatience or time preference of individuals, and individuals' beliefs about the characteristics and beliefs of other marriage market participants.

Ecologists may find models that allow the possibility of a positive equilibrium rate of population growth unsatisfactory because such models fail to recognize density dependence and other factors that prevent population growth from continuing indefinitely.

4. Summary

To summarize: classical stable population theory is parsimonious both because it allows us to use well-known, powerful mathematical techniques to investigate existence and dynamic stability, and because it allows us to infer a population's equilibrium age structure, growth rate, and dynamic behavior from very little data. The birth matrix-mating rule model, which allows fertility rates to depend on the population's age-sex structure, is more complex analytically and more demanding in its data requirements. In return for these extravagances, the BMMR model solves demography's two-sex problem and provides a framework for addressing the marriage squeeze and other important issues in demography that require a two-sex model.

See also: Gender and Feminist Studies; Gender and Feminist Studies in Economics; Gender and Feminist Studies in History; Gender and Feminist Studies in Political Science; Gender and Feminist Studies in Psychology; Gender and Feminist Studies in Sociology

Bibliography

- Feeney G M 1972 Marriage rates and population growth: The two-sex problem in demography. Ph.D. thesis, University of California, Berkeley, CA
- Mortensen D T 1988 Matching: Finding a partner for life or otherwise. In: Winship C, Rosen S (eds.) *Organizations and Institutions: Sociological and Economic Approaches to the Analysis of Social Structure. American Journal of Sociology, Supplement*. University of Chicago Press, Chicago, pp. S215-40
- Pollak R A 1986 A reformulation of the two-sex problem. *Demography* **23**: 247-59
- Pollak R A 1987 The two-sex problem with persistent unions: A generalization of the birth matrix-mating rule model. *Theoretical Population Biology* **32**: 176-87
- Pollak R A 1990a Two-sex demographic models. *Journal of Political Economy* **98**: 399-420
- Pollak R A 1990b Two-sex population models and classical stable population theory. In: Adams J, Lam D A, Hermalin A I, Smouse P (eds.) *Convergent Issues in Genetics and Demography*. Oxford University Press, New York

- Roth A E, Sotomayor M A O 1990 *Two-sided Matching: A Study in Game-theoretic Modeling and Analysis*. Cambridge University Press, Cambridge, UK
- Weiss Y 1997 The formation and dissolution of families: Why marry? Who marries whom? And what happens upon divorce. In: Rosenzweig M R, Stark O (eds.) *Handbook of Population and Family Economics, Vol. 1A*. Elsevier, Amsterdam

R. A. Pollak

Population Ecology

One part of this research program examines the demography of organizations. It finds that organizational mortality processes depend upon the age and size of the organization, as well as on characteristics of populations and environments. Moreover, there is evidence of an imprinting process, meaning that environmental conditions at certain early phases in an organization's development have long-term consequences. In particular, organizations subject to intense competition have elevated mortality hazards at all ages.

A central theme is structural inertia, the tendency for organizations to respond slowly relative to the speed of environmental change. A central argument holds that inertia derives from the very characteristics that make organizations favored kinds of actors in modern society: reliability and (formal) accountability. It follows that changes in an organization's core features are disruptive and increase mortality hazards, at least in the short-run. Research on this subject tends to support this view.

The concept of niche provides a framework for relating environmental variations and competition to population dynamics and segmentation. Much empirical work examines the niches of organizational populations in terms of dimensions of social, political, and economic environments.

Most research in this field builds on theories of resource partition and of density dependence. Resource-partitioning theory concerns the relationship between increasing market concentration and increasing proliferation of specialists in mature industries. The key implication of this theory concerns the effects of concentration on the viability of specialist organizations (those that seek to exploit a narrow range of resources). The theory of density-dependent organizational evolution synthesizes ecological and institutional processes. It holds that growth in the number of organizations in a population (density) drives processes of social legitimation and competition that, in turn, shape the vital rates. This argument implies certain nonmonotonic relationships between density and rates of founding and mortality. Both theories

have been subject to empirical test in diverse populations of organizations; the preponderance of evidence supports both.

Current research emphasizes formalization (often using tools of formal logic), integration of methods of social network analysis and ecological analysis, and developing implications of organizational ecology for other aspects of social structure.

The population ecology of organizations, commonly called organizational ecology (OE) examines interactions within and between populations of organizations. This approach differs from other sociological research by focusing on the population level, relying on selection logic, and studying the life histories of all organizations in populations studied over their full histories.

This approach built initially on neo-classical population bioecology, which analyzes numerical aspects of population interactions from an evolutionary perspective. In a paraphrase of bioecologist G. E. Hutchinson, Hannan and Freeman (1977) posed the following orienting question of OE: Why are there so many kinds of organizations? They suggested that the social science literature lacked good answers to this question and that seeking answers would clarify the dynamics of the organizational world. Subsequent research has shown that organizational diversity has important consequences for individuals and social structures (Carroll and Hannan 2000).

OE sought initially to resolve a tension in the human ecology concerning what Hawley (1968) termed the principle of isomorphism, 'Units subjected to the same environmental conditions or to environmental conditions as mediated through a given key unit, acquire a similar form of organization.' Hannan and Freeman (1977) argued that Hawley's principle, like other adaptationist premises, does not apply straightforwardly in uncertain, heterogeneous environments. Extending the received wisdom about adaptation requires specification of the underlying dynamic processes and attention to processes of selection and segmentation.

1. The Basic Demography of Organizations

With its concentration on the vital rates of founding and mortality, OE has a strong demographic character. Key structuring processes are specified at the population level. Although mortality histories can always be associated with individual organizations, histories of foundings cannot because the absence of foundings in some period or place must be considered.

Populations must have unit character to serve as useful objects of analysis (Hawley 1968). Researchers generally regard populations as spatial-temporal instantiations of organizational forms, templates for organizing. Several conceptions of the basis of form

have guided these efforts: common resource dependence, structural equivalence in network position, and identities coded in cultural rules.

Constructing organizational demography and ecology requires attention to differences between organizations and organisms. In particular, organizations experience more varied 'birth' and 'death' events, typically lack clear parents, have multilayered, partly decomposable structures, and can transform themselves. The major methodological project has involved dealing with these complexities (Carroll and Hannan 2000).

The resulting strategy examines the full histories of organizational populations, because experience shows that early events can have lasting consequences for population dynamics. It gathers life history data on all organizations in the population(s), including the large and famous as well as the small and insignificant—this is crucial for avoiding problems of selectivity bias. It records detailed information about type of entry (new founding, entry from another industry, merger, etc.) and exit (disbanding, acquisition, transformation, etc.). Finally, it uses event-history methods to estimate the effects on vital rates of characteristics of organization, population, and environment. This research strategy has diffused widely to areas of organizational research that do not necessarily share theoretical commitments.

Ecological research finds that many more organizations have operated in well-studied industries than is usually assumed. Careful archival searches usually find information on hundreds of organizations that get overlooked in published counts. They also reveal that organizational life expectancies are generally an order of magnitude smaller than for individuals.

Much attention has been paid to organizational aging. Stinchcombe (1965) observed that organizations (and organizational populations) experience a liability of newness. Efficient organization requires trust among members, which takes time to build. Creating roles and routines, learning about the environment, and developing relationships with other organizations also take time. Once established, these patterns of relationships enhance survival chances. Much early research found this pattern. Some later research found a variation, a liability of adolescence: the hazard rises during the early lifespan—while initial stocks of endowments get exhausted—before declining. Recent research, using designs that measure (age-varying) organizational size, has produced (mixed) evidence in favor of positive age-dependence (Carroll and Hannan 2000). Two interpretations for this pattern have been proposed. If inertial forces are strong, then the possibilities of adapting to changing environments are limited and older cohorts of organizations have lower fitness—there is a liability of obsolescence. Alternatively, the accumulation of rules and routines impedes adjustment to environmental change—the liability is one of senescence. Current

research seeks to clarify the conditions under which the various patterns arise—see later.

2. *Inertia and Selection*

Stinchcombe's (1965) insight that new organizations get imprinted by their environments provided inspiration for developing the ecological approach. According to Stinchcombe's argument, new organizations get tested against assumptions that are taken for granted but change over time. Hence, the characteristics of organizations that pass such tests and emerge reflect the social structure of the time. Imprinting requires both a mapping of environmental conditions onto the organization and inertia in the imprinted characteristics. Absent the second condition, later adjustments will erode the association between founding conditions and those features.

OE has built on this image of inertia (relative to the speed of environmental change). According to Hannan and Freeman (1989), inertia prevails as an inadvertent by-product of a selection process favoring the properties of reliability (low variance in the quality of performance) and accountability (the ability to construct rational accounts for actions). Achieving these properties depends on structures being reproduced faithfully over time. Yet, high reproducibility means that structures resist transformation. Therefore, this selection process inadvertently favors corporate actors with strong inertial tendencies.

Fundamental change presumably diminishes reliability and accountability. Old learning loses value, and new processes must be learned. Vestiges of the old system conflict with the new one emerging. Thus, even if fundamental change has long-term benefits (the new form might be better aligned with environments and internal processes), change might increase mortality hazards in the short-run. If so, then reorganization-prone organizations have a lower probability of representation in future populations.

Much research has tested the implications of this argument. Most well designed studies reveal that changing core structures does increase the hazard of mortality in the short-run, often substantially (Carroll and Hannan 2000). Moreover, the magnitude of this effect increases with organizational age.

A different way of viewing imprinting yielded one of OE's most robust empirical findings: density delay. Organizations subject to intense competition at founding (as evidenced by a high density of organizations in the population—see later) have elevated mortality hazards at all ages. In other words, density at founding has a (partially) delayed effect on mortality processes. This relationship has been interpreted as reflecting enduring effects either of building structures under scarcity or of positioning at the margins of resource distributions.

3. Organizational Niches

The concept of niche provides a framework for relating environmental variations and competition to population dynamics and segmentation. A population's fundamental niche consists of the combinations of environmental parameters for which its growth rate (fitness) is non-negative in the absence of competition. Much empirical work in OE examines the niches of organizational populations in terms of dimensions of social, political, and economic environments (e.g., Dobrev 1999). Even when such issues do not stand in the forefront, ecologists routinely control statistically for the effects of such dimensions in examining dynamic processes.

Niche theory has also been central to research on competition (Hannan and Freeman 1977, 1989, McPherson 1983). Populations compete if the presence of each reduces the spaces in which the others can sustain themselves (their realized niches). By this reasoning, populations compete if and only if their fundamental niches intersect. The assumed equivalence of niche overlap and competition has played a crucial role in allowing organizational ecologists to relate empirical observations on realized niches to models of population dynamics.

Niche width concerns the variance of resource utilization. Does evolution favor populations of Jacks-of-all-trades and masters of none (generalists) or specialists at a few tasks? Research shows that the organizational world tends toward segmentation into specialist populations (each highly isomorphic) when relevant environmental variations are uncertain, large in magnitude, and long in duration (Hannan and Freeman 1989, Péli 1997).

4. Resource Partitioning

Resource-partitioning theory, OE's best-developed theory of segmentation, concerns the relationship between increasing market concentration and increasing proliferation of specialists in mature industries. The theoretical imagery concerns crowding in a heterogeneous market. Organizations initially target resource segments. Specialists choose narrow, homogeneous targets; generalists choose broad, heterogeneous ones.

The theory assumes an advantage of scale (in production, marketing, or distribution). Because organizations targeting narrow segments cannot grow very large, they have higher (per unit) costs. Organizations with scale advantages expand products to broaden their appeal without losing established domains. Small organizations then face formidable invaders, and many fail. Competition is most intense in resource-dense centers, resulting in dominance by large generalists. Failures of smaller generalists cause some resources to become free. Because larger genera-

lists can best secure these resources, the surviving generalists become larger and more general. However, generalists can rarely secure all of the newly freed resources. As the struggle among generalists proceeds to its monopoly equilibrium, the surviving generalists grow and broaden, but the set of resources they hold collectively actually declines. In other words, more resources become available for specialists as the market concentrates. Therefore, as concentration rises, the viability of specialist organizations increases as well: founding rates rise and mortality rates fall.

The empirical implications of this argument involve an interaction effect between organizational form (specialism-generalism) and concentration in affecting each vital rate. Empirical studies on diverse populations support this hypothesis (Carroll and Hannan 2000). Populations studied include American airlines, brewers, investment banks, microprocessor manufacturers, newspaper publishers, telephone companies, and wineries, as well as British automobile manufacturers and Italian cooperative banks (full citations can be found in Carroll and Hannan 2000).

5. Density Dependence

The theory of density-dependent organizational evolution synthesizes ecological and institutional processes. It holds that growth in the number of organizations in a population (density) drives processes of social legitimation and competition that, in turn, shape the vital rates (Hannan and Carroll 1992). An organizational form gains (constitutive) legitimation when it becomes taken-for-granted (Meyer and Rowan 1977). New organizational forms generally lack such legitimation, which makes organizing difficult. If the form gets promulgated, legitimation increases. Initially, when there are few organizations in a population and the form lacks legitimacy, the returns to legitimation of adding another organization are great. However, when the population contains many organizations, constitutive legitimation increases little—if at all—as density rises further. In other words, a population's legitimation increases with density at a decreasing rate and approaches a ceiling.

Competition refers to the consequences of efforts by multiple agents to acquire the same limited resources. Density-dependence theorists focus on diffuse competition because it has a clear relationship with density as the intensity of diffuse competition within an organizational population rises as a function of the number of potential bilateral competitors. As density increases linearly, the number of possible competitive relations increases geometrically. Thus diffuse competition rises with density at an increasing rate. The theory posits that founding rates are directly proportional to legitimation and inversely proportional to the intensity of diffuse competition and that these relationships are inverted for mortality hazards.

This theory (formalized appropriately) yields two theorems. First, as density increases, founding rates rise initially, reach a peak, and then decline. Second, the relationship for mortality is the mirror image. These theorems, expressed in parametric forms, can be falsified empirically. They have been tested for diverse populations, including automobile manufacturers in Europe and the USA, brewers in Germany and the USA, banks in Denmark, Japan, Singapore, and the USA, and American labor unions, life insurers, newspapers, and social movements of women and ethnic groups. These studies have supported the theory in most cases (full citations can be found in Carroll and Hannan 2000).

6. Emerging Trends

Current research seeks to sort out these varied arguments, learn about conditions of applicability, and (when possible) unify them. One strand begins by reconstructing the fragments in formal languages. In the case of age dependence, a partial integration has been achieved in a formulation in first-order logic (Carroll and Hannan 2000). The next steps require use of nonmonotonic logic, which allows premises to be overridden by more specific ones so that the conclusions of an argument can change as more knowledge about specifics becomes available (Pólos et al. 1999).

Efforts are underway to rebuild the foundations of the research program on an improved formal representation of organizational forms, with forms based on culturally coded and enforced identities (by Pólos et al. 2002). In this construction, density dependence plays a role in the cultural coding process, as do social movements. Processes of enforcement of identities generate inertial mechanisms as well as punishments for fundamental changes (those that violate codes).

Basing forms on identities helps clarify issues in resource partitioning. Specialization based on claims to authentic identity (as in the case of the rise of specialist producers of microbrewed beer in the USA and 'real ale' in the UK) can prevent large generalists from gaining complete control of concentrated markets. Other research seeks better representations of dynamics and segmentation. The formal model of density-dependence serves as the foundation for many of these efforts. One approach rescales the effects of density using a specification of subniches to weight density. Some research weights densities according to an organization's position within a broad technology class or by licensing restrictions, location, and price (Baum and Singh 1994). Alternatively, the weightings are set to reflect the strength of competitors (Barnett 1997). These specifications generate refined understandings of dynamics (at the cost of considerable complexity).

A second variation envisions competition as more localized than legitimation, because political author-

ities can better impede flows of products than ideas. For instance, research on the European automobile industry shows competition operating primarily within national markets but legitimation operating throughout Europe (Carroll and Hannan 2000). The third has shown the effects of density change systematically over a population's history, either directly as a function of population age, reflecting the growing importance of identities, statuses, and reputations in affecting life chances in mature populations or as a function of industry concentration (Barron 1999), in the spirit of resource-partitioning.

The empirical scope of OE also continues to expand. Current research focuses on a diverse set of industries. In addition to those already mentioned, these include: accountancies, bicycle manufacturers, computer manufacturers, disk-drive producers, footwear manufacturers, hotels, law firms, magazine publishers, paper manufacturers, sex shops, and television stations. A major thrust of current work is collecting data on life histories of organizations and on connections among the organizations (including flows of personnel among them). These studies highlight the value of social network analysis for understanding the ecology of organizations.

New processes are also being brought into the picture. Organizational status has been shown to affect life chances, with the impact increasing with the uncertainty in evaluations of quality (Podolny et al. 1996). Social movement research relies increasing on OE as a guide to modeling dynamics (Minkoff 1997); ecologists, in turn, increasingly view social movements as pivotal in the emergence of forms (Wade et al. 1998). Studies of careers find that organizational demography induces considerable job mobility (Haveman and Cohen 1994); at the same time, careers of managers shape strategies and thereby influence life events for organizations (Sørensen 1999).

See also: Competitive Strategies; Organizational; Darwinism; Social; Ecological Economics; Ecological Imperialism; Ecology: Organizations; Ecology, Political; Resource Geography

Bibliography

- Barnett W P 1997 The dynamics of competitive intensity. *Administrative Science Quarterly* **42**: 128–60
- Barron D N 1999 The structuring of organizational populations. *American Sociological Review* **64**: 421–45
- Baum J A C, Singh J V 1994 Organizational niches and the dynamics of organizational mortality. *American Journal of Sociology* **100**: 346–80
- Carroll G R 1985 Concentration and specialization. *American Journal of Sociology* **90**: 1262–1283
- Carroll G R, Hannan M T 2000 *The Demography of Corporations and Industries*. Princeton University Press, Princeton, NJ

- Dobrev S D 1999 The dynamics of the Bulgarian newspaper industry in a period of transition. *Industrial and Corporate Change* **8**: 573–605
- Hannan M T, Carroll G R 1992 *Dynamics of Organizational Populations*. Oxford University Press, New York
- Hannan M T, Freeman J 1977 The population ecology of organizations. *American Journal of Sociology* **82**: 929–64
- Hannan M T, Freeman J 1989 *Organizational Ecology*. Harvard University Press, Cambridge, MA
- Haveman H A, Cohen L E 1994 The ecological dynamics of careers. *American Journal of Sociology* **100**: 104–52
- Hawley A H 1968 Human ecology. In: Sills D (ed.) *International Encyclopedia of Social Sciences*. Macmillan, New York
- McPherson J M 1983 An ecology of affiliation. *American Sociological Review* **48**: 519–35
- Meyer J W, Rowan B 1977 Institutionalized organizations: Formal structure as myth and ceremony. *American Journal of Sociology* **83**: 340–63
- Minkoff D 1997 The sequencing of social movements. *American Sociological Review* **62**: 779–99
- Péli G 1997 The niche hikers guide to population ecology. In: Raftery A (ed.) *Sociological Methodology*. Blackwell, Oxford, UK
- Podolny J P, Stuart T E, Hannan M T 1996 Networks, knowledge and niches. *American Journal of Sociology* **102**: 659–89
- Pólos L, Hannan M T, Carroll G R 2002 Foundations of a theory of social forms. *Industrial and Corporate Change* Forthcoming
- Pólos L, Hannan M T, Kamps J 1999 Aging by default: Building theory from fragmentary knowledge. In: Rott H, Albert C, Brewka G, Witteveen C (eds.) *Proceedings of the Fourth Dutch–German Workshop on Non-Monotonic Reasoning Techniques*. ILCC Scientific Publications, Amsterdam
- Sørensen J B 1999 The ecology of organizational demography. *Industrial and Corporate Change* **8**: 713–44
- Stinchcombe A L 1965 Social structure and organizations. In: March J (ed.) *Handbook of Organizations*. Rand-McNally, Chicago
- Wade J B, Swaminathan A, Saxon M S 1998 Normative and resource flow consequences of local regulations in the American brewing industry. *Administrative Science Quarterly* **43**: 905–35

M. T. Hannan

Population, Economic Development, and Poverty

1. Introduction

Assessing the consequences of population on the pace and process of economic growth is one of the oldest themes in the literature on economics. These assessments have varied enormously over time, spanning the highly pessimistic to the outright optimistic. A systematic review of the major studies in this literature represents a useful way to organize a survey of the

consequences of demographic change. Such an approach places the population debates in perspective, and it infuses a healthy dose of caution in appraising current debates. Specifically, how have the ‘bottom-line’ appraisals of the consequences of demographic change on development changed over time; why have they changed; and what are the most recent contributions to this literature?

2. The Beginnings: Malthus (1798)

The debate began with two propositions by the Reverend Thomas Malthus in 1798 (*First Essay on Population*) that population would grow at a geometric rate (e.g., 1, 2, 4, ...) due mainly to a lack of conscious restraints on fertility, and that food would grow at an arithmetic rate (e.g., 1, 2, 3, ...) due substantially to diminishing returns to increasingly scarce land. The resulting outcome would be food shortages, starvation, and deaths. In the long run, population size would be held in check by food availability and mortality. Population pressures would constrain income per capita to a low level of subsistence—a ‘Malthusian trap,’ as it has been termed. These images caused economics, unfairly, to be dubbed the dismal science.

Fortunately Malthus’ predictions were not sustained by the preponderance of experience over the next two centuries. Couples did not breed without restraint, but rather by consciously managing fertility in response to changing conditions. Food was not unduly constrained by land availability. Instead, technology blossomed and food expanded apace in the very geographic regions where Malthus focused his empirical studies. Ironically, food *surpluses* turned out to be a ‘problem’ confronting many nations, and governments implemented policies designed to curtail farm production.

Clearly the Malthusian ideas regarding population–economic linkages were incomplete, and richer analytical and empirical foundations were needed. The urgency for such a framework was made apparent by demographic events. By the mid-twentieth century, it was recognized that the simultaneous occurrence of declining mortality and exceptionally high and sustained fertility in scores of developing countries was resulting in high population growth rates. A concern emerged that these rates could not be sustained over long periods of time. While, as in the past, fertility would predictably decline (*a la* the Demographic Transition), it was unclear whether such a decline would be soon or rapid enough to avoid potentially deleterious effects on welfare, economic progress, and the environment. Thus, while the ‘Malthusian Problem’ reappeared, approaches to assessing population consequences assumed quite different tacks. It was time for a fresh reassessment.

3. Expanded Elaborations: 1950s, 1960s, 1970s

3.1 The United Nations (1953)

The first major installment that exposed economic-demographic interactions in detail was provided in a seminal study by the United Nations in 1953, *The Determinants and Consequences of Population Trends*. This study was both balanced and comprehensive. Twenty-one economic-demographic linkages were taken up. The impact of population on some factors was judged to be positive (economies due to scale and organization), on some negative (diminishing returns), and on some neutral (technology and social progress). Overall, a net negative impact of undetermined size was considered most likely in much of the developing world, although this conclusion was guarded, and the UN predicted a wide diversity of effects according to country-specific conditions.

3.2 Coale and Hoover (1958)

A second major installment in the evolution of demographic assessments appeared in 1958 in an influential book by Ansley J. Coale and Edgar Hoover (C/H) entitled *Population Growth and Economic Development in Low-Income Countries*. Based on the simulation results of a mathematical model calibrated by Indian data, C/H concluded that India's development would be substantially enhanced by lower rates of population growth. Their analysis rested on two premises. First, household and economy-wide saving would be diminished by large families; it certainly would not be stimulated enough in response to demographic changes to provide the increased capital required by an enlarged population. Second, economy-wide investments would be skewed away from relatively productive activities since funds would be shifted toward so-called unproductive population-sensitive social expenditures (e.g., health and education). This crowding out of productive investment was linked mainly to the youthful age composition of the high-fertility population, and not population size, density, or growth, *per se*.

The C/H study had a profound impact on US population policy. It also attracted the attention of academic scholars given its focus on physical capital (as distinct from the Malthusian focus on land), then considered by many to be a (the?) key to economic development.

The C/H study was not without challenge. Over time its impact waned, in part due to the appearance of empirical studies that failed to unearth empirically strong and consistent impacts of population pressures on Third World saving. This is not surprising. At the household level, such saving impacts are fundamentally based on a life-cycle conceptualization of be-

havior. This requires a lengthy forward looking planning horizon. It also requires an institutional structure (e.g., developed capital markets and/or government social security options) to implement lifetime plans. These conditions do not describe the day-to-day survival mode of most residents of the Third World. Indeed, spending on children represents a form of saving (e.g., parents may expect transfers from their children in old age); and moreover, children can be quite productive both in the household and on the farm.

A second challenge to the C/H framework was the appearance of empirical studies that downplayed the role of country-wide age structures as a determinant of spending on education. Even in the face of population pressures, school enrollments and attainments advanced at historically unprecedented rates in many developing countries. This surprisingly occurred without substantial realignments in social budgets. Nations found ways to economize on education resources, and tradeoffs within education budgets were more pervasive than was assumed in the simple C/H model. In short, the empirical underpinnings of the C/H framework were questioned by an increasing body of evidence. Additionally, over time economic theory itself was in flux, highlighting both the *productivity* of human capital (considered economically unproductive in the C/H model), and the importance of technical change (also absent in C/H). While these limitations do not diminish the importance of C/H's pioneering work, they served to shift the population debates in the ensuing decades.

3.3 National Academy of Sciences (1971)

The most pessimistic assessment of population consequences since Malthus appeared in a study organized by the National Academy of Sciences. The findings, *Rapid Population Growth: Consequences and Policy Implications*, appeared in two parts: Volume I, a short Executive Summary, and Volume II, a large set of papers compiled by a panel of distinguished scholars. The study's findings, as summarized by the Executive Summary, are baffling and difficult to interpret. They must be handled with care.

On the one hand, the Executive Summary conveyed a highly negative, indeed an outright alarmist, assessment of the impact of rapid population growth. Twenty-five separate adverse impacts were listed and assessed; no notable positive effects of demographic change were identified. Moreover, there was a substantial disconnection between the Executive Summary and the findings of the key scholars participating in the study, as documented by the papers found in Volume II. The Executive Summary, whose pessimistic conclusions obtained wide publicity in the 1970s, was demonstrably unfaithful to the scholarly studies in

Volume II. The Summary was not vetted with the scholars who participated in the study, and indeed its authorship is unknown to this day, in spite of extensive attempts to clarify the historical record (an exhaustive historical assessment of this episode is provided in Kelley 2001). Clearly the assessment of the Executive Summary must be strongly discounted. In spite of this, however, a careful reading of the papers and supporting components of the Report do reveal an important insight that assists in illuminating the ebb and flow of population assessments over time.

Specifically, the Executive Summary was by admission based mainly on 'direct,' *shorter-run* impacts of demographic change. ('We have limited ourselves to relative short-term ... issues,' p. vi). In contrast, the major research papers supporting that study in Volume II, which were much less pessimistic, were based on a *longer-run* focus. More than any other factor, the outcomes of the population debates from the 1970s to the present day would turn on this difference in time perspective. *Direct, shorter-run impacts of demographic change are almost always attenuated (and sometimes even offset) by feedbacks that occur only over longer periods of time.*

3.4 United Nations (1973)

In 1973 the United Nations updated its early assessment. This revision arrived at a less eclectic, and a somewhat more pessimistic (but by no means alarmist) evaluation of the various impacts of population growth. This is particularly true of anticipated difficulties of feeding the expanding populations (reverting to traditional Malthusianism), and of pressures on capital formation (reverting to the concerns of Coale and Hoover). In both cases, short-run, direct impacts were highlighted. The report downplays some indirect, longer-run feedbacks that will likely occur due to price responses, and induced innovation and institutional change.

The most important new contribution to the population debate deriving from the 1973 study was a finding by Simon Kuznets that, based on simple correlations, a net negative impact of population growth on per capita output was not obvious in the data. While his work was based on *longer-run* assessments, and while they were appropriately qualified, they were important to conditioning the bottom-line UN assessment. Given the strong priors of demographers and policy makers that the negative impacts of population growth on development were large, the inability to easily confirm this hypothesis through simple, albeit inconclusive, correlations more than any other factor kept the population debate alive during the ensuing decades. The stage was set for a new round of debate.

4. Revisionism: The 1980s

4.1 Julian L. Simon (1981)

This debate was initiated in 1981 with the publication of Julian L. Simon's *The Ultimate Resource*, which challenged the pessimistic views of economic-demographic interactions. The significant impact of Simon's book is explained by several factors. First, it concluded that population growth was likely to exert a *positive* net impact on economic development in many Third World countries *in the intermediate run*—a startling assertion that attracted extensive attention. Second, it illustrated that the outcome of population impacts on the economy are likely to hinge both on the *time* dimension of the assessments, and whether *feedbacks* are included in the analysis. Third, it reintroduced to the debates the roles of alternative demographic impacts, specifically those of density and size, to add to the popular focus on population growth and age structure. Fourth, and importantly, the book elevated the role of *technology* as a (the?) driving force of economic growth. Simon asserted that demographic pressures would favorably influence the nature and pace of technological change. Finally, it was written in a highly accessible, confrontational debating style that goaded and infuriated his skeptics and amused and emboldened his supporters.

These various attributes of Simon's framework are best illustrated by his analysis of the likelihood that demographic pressures would ultimately cause natural resource exhaustion. He illustrated with numerous graphs that over longer periods most natural-resource prices actually declined, even in the face of rising demands stimulated in part by expanding populations. Price-induced substitutions in production and consumption of natural resources, and an expansion of supply through discovery and technical advance, are offered to explain this result. Most importantly, these 'adjustments' or 'feedbacks' were in part *caused by* population pressures, a reality that could be exposed only by extending the analysis over several decades. It is this longer-run perspective, and the *feedbacks* in markets and other institutions, that account for Simon's findings.

Similarly, Simon hypothesized and attempted to document that the pace of technical change, and its bias, were actually stimulated by population pressures, particularly in agriculture where Ester Boserup's (1965) model, first expounded in *The Conditions of Agricultural Growth*, was advanced. She proposed that increasingly productive agricultural technologies are made economically feasible only at higher land densities, and therefore population pressures against increasingly scarce land could have a powerful positive effect on agricultural output growth. Simon extended this notion to observe that major social overhead projects (roads, communications, irrigation) benefited from expanded populations and scale.

While Simon's book was highly influential in its own right, arguably its greatest impact was as a catalyst in setting the stage for reassessing the basic propositions of the population debates. The 1980s witnessed a flurry of major reassessments.

4.2 National Research Council (1986)

The most important and influential reassessment was compiled by a Working Group on Population Growth and Economic Development under the auspices of the National Research Council (NRC). The bottom-line assessment by this group was eclectic, balanced, and nonalarmist:

On balance, we reach the qualitative conclusion that slower population growth would be beneficial to economic development of most developing countries (p. 90).

This carefully worded statement illustrates well the central elements of all of the major population assessments in the 1980s: (a) there are both important positive and negative impacts of population growth (thus, 'on balance'); (b) the actual size of the net impact—and even whether it is strong or weak—cannot be determined given current evidence (thus 'qualitative'); (c) only the direction of the impact from high current growth rates can be discerned (thus 'slower,' and not 'slow'); and (d) the net impact varies from country to country—in most cases it will be negative, in some it will be positive, and in others it will have little impact one way or the other (thus, 'most developing countries').

It is interesting to speculate on factors that account for the guarded and qualified assessment of the NRC group. Three factors can be singled out.

First, the Summary Report highlights *individual and institutional responses* to initial impacts of population change—conservation in response to scarcity, substitution of abundant for scarce factors of production, innovation and adoption of technologies to exploit profitable opportunities, and the like.

Second, the study was compiled almost exclusively by economists, whose faith in the potential for market-induced responses to modify initial direct impacts of population change is greater than that possessed by other social and biological scientists.

Third, research which had been accumulating over the past 10 to 15 years downplayed many of the previous concerns voiced on the impacts of rapid population growth. For example, as noted by Simon and others, the concern that population growth results in the exhaustion of non-renewable natural resources is misplaced. The relationship between population growth and global resource use is not as strong as had been assumed.

Similarly, the concern about a substantial reduction of saving due to rapid population growth is not

confirmed by the data. While some capital shallowing occurs, the impact of this on economic growth is not particularly strong. Finally, the concern that population growth will significantly shift resources from productive physical capital formation into alleged 'less-productive' areas was not sustained by the data. The financing of educational enrollments, which expanded significantly even in the face of population pressures, came from some combination of increases in public (sometimes deficit) spending, reduction in per pupil expenditures, and efficiency gains.

Interestingly, a new concern was elevated in importance by the 1986 Report in the area of renewable resource degradation. Unlike non-renewable resources, property rights are difficult to assign or maintain for renewable resources like fishing areas and rain forests. Overuse can result. The problem is not population growth *per se*, but rather institutional failure. Cutting population growth in half would not solve the problem; it would only postpone the deleterious outcome of resource degradation. The solution requires policies to focus on the *causes* of failure (the need for more effective property rights, market responses, and government policies to correct externalities) rather than a focus on population pressures, which mainly exacerbate the environmental responses.

4.3 Other Studies in the 1980s

The bottom-line assessments of five other studies in the 1980s merit citing. All reflect the balanced, longer-run methodological perspective of the 1986 National Research Council report. Specifically, all highlight direct and indirect responses to population pressures. The resulting bottom-lines, while assigning a negative net impact of rapid population growth are, on average, both qualified and guarded.

The World Bank's *World Development Report* 1984 is concerned most about countries that exhibit exceptionally rapid growth rates: '...population growth—at rates above 2 percent ...—acts as a brake on development' (p. 79), although, 'Up to a point, population growth can be accommodated ...' (p. 79).

An overall appraisal of both the *World Development Report* 1984, and the National Research Council report two years later, is offered by Dr. Nancy Birdsall, who headed the World Bank study and was also a member of the NRC Working Group. She concluded that 'Rapid population growth can slow development, but only under specific circumstances and generally with limited or weak effects' (Birdsall 1988, p. 529).

Allen C. Kelley's 1988 survey for the *Journal of Economic Literature* concludes that '...economic growth...would have been more rapid in an environment of slower population growth, although in a number of countries the impact was probably negligible and in some it may have been positive' (p.

1715). Adverse impacts are most likely to occur where (a) water and arable land are scarce, (b) property rights are poorly defined, and (c) government policies are ineffective and biased against labor.

T. N. Srinivasan's 1988 survey parallels the conclusion of the NRC Report. He emphasizes caution in the formulation of population policies, noting that 'many of the alleged deleterious consequences [of population growth] result more from inappropriate policies and institutions than from rapid population growth. Thus policy reform and institutional change are called for, rather than policy interventions in private fertility decisions to counter these effects' (p. 7).

It appears that in the 1980s a broad consensus was struck on the net assessments of demographic change. Most analysts held that while slower population growth would indeed advance the economic progress of most developing countries, the size of the net impacts would not likely be especially remarkable by comparison with numerous other determinants of economic growth. The shift of the research emphasis during the decade to highlight the importance of (a) *time* in the analysis (modifying shorter-run direct impacts of demography with feedbacks occurring over the longer run), and (b) the need to account for the separate components of demographic change (births, deaths, age, size, density), combined to put the research programs on solid footing. While these revisionist renderings and methodologies were widely embraced, uncertainties still remained, most importantly the need to obtain more precise *quantitative* assessments, especially those that accounted both for the components of demographic change and the timing of their impacts.

5. New Paradigms: The 1990s and Beyond

While most of the 1990s was preoccupied with digesting the revisionist results of the 1980s, population research did advance in several areas.

First, the findings from simple correlations between the rate of population and per capita economic growth appeared to have changed. While a general *lack* of correlation was the widely obtained statistical result for the 1960s and 1970s, in the 1980s the correlation turned negative (see Kelley and Schmidt 1994). On the one hand, most analysts agreed that such simple correlations are difficult to interpret, plagued as they are by failure to adequately account for reverse causation, excessive reliance on cross-section data, sensitivity to the selection of countries, the complexity of demographic linkages that are poorly modeled, spurious correlation, econometric pitfalls, and data of dubious quality. On the other hand, the previous finding of *no* correlation for the 1960s and 1970s in the face of strongly held priors of a negative correlation literally kept the population debate alive. Now, a

change in this relationship from one of no correlation to one of a negative correlation for the 1980s required an explanation. New questions appeared: what accounts for the changed correlations; are the new results robust; are they quantitatively important?

The ability to address these issues coincided with the emergence in the 1990s of empirical 'convergence' models of economic growth. Pioneered by Robert Barro, these empirical paradigms distinguish between factors (economic, political, social, institutional, geographic) that determine each country's *long-run* level of per capita output, and the shorter-to-intermediate-run *transition* of countries to this longer-run state. These models lent themselves to investigating the impacts of demography since they exposed both short- and long-run impacts.

Efforts to model demography using the new convergence models have varied notably. Barro (1997), for example, looked primarily on the longer-run impacts of demography, and found that reductions in the total fertility rate increased the potential for economic growth.

In yet an earlier study, Kelley and Schmidt (1995), building on the Barro core variables, distinguished between several alternative demographic influences on the economy's potential output in the long-run, (e.g., the impacts of population size and density), and timing of demographic impacts (e.g., the timing of reductions in birth and death rates) which influence both the short and long run. These timing specifications highlighted the reality that birth-rate reductions have an immediate positive impact on growth by economizing on child-rearing expenses while, in fifteen or so years, the impacts will be reversed, since there will be fewer persons entering their productive work force years. Death rate reductions, especially infant/child mortality, can have similar timing impacts. Kelley and Schmidt found that these timing/transitional features were empirically important, and the long-run impacts of both density and size added modestly to growth as well.

Bloom and Williamson (1998), also building on Barro's empirical framework (although with different core variables highlighting policy and geography), modified the demographic modeling to break out an accounting reckoning of age-compositional impacts. While explicit modeling of longer-run demographic impacts is absent in their framework, their clean accounting framework clearly exposes the impacts of changing age structures, driven by changes in fertility and mortality. These are quantitatively important impacts on the *transition* to long-run output per capita. Their results focused on East Asia where declines in fertility were rapid and shorter-run transition effects are predictably large.

Finally, Kelley and Schmidt (2001) compared the above (and other) modeling efforts in a single empirical investigation, and came up with a somewhat surprising result: demography accounts for around 20 percent of

changes in output per capita growth from 1960–95 across a wide collection of countries. While for several reasons they consider their findings qualified, it is interesting that these findings are broadly consistent with those of the 1980s. Population *does* matter: its impact is likely adverse over the period 1960–95; this impact varies from decade to decade; components of demographic change exert both positive and negative impacts; these impacts vary notably from place to place; and, as a determining variable of long-run economic prosperity, population's impact is notable, but not remarkable. In the shorter-to-intermediate run, during periods of transition (both demographic, and economic), populations impact can be elevated or diminished, depending on the pace of demographic change and especially on the country's specific institutions (government policy, efficacy of markets, definition of property rights).

See also: Boserup, Ester (1910–99); Demographic Transition, Second; Development: Socioeconomic Aspects; Economic Growth: Theory; Malthus, Thomas Robert (1766–1834); Savings Behavior: Demographic Influences Population Geography

Bibliography

- Barro R J 1997 Determinants of economic growth: A cross-country empirical study. Development Discussion Paper No. 579, Harvard Institute for International Development, April
- Birdsall N 1988 Economic approaches to population growth and development. In: Chenery H B, Srinivasan T N (eds.) *Handbook of Development Economics*. Elsevier Science, Amsterdam
- Birdsall N, Kelley A C, Sinding S (eds.) 2001 *Demography Matters: Population Change, Economic Growth and Poverty in the Developing World*. Oxford University Press, Oxford, UK
- Bloom D E, Williamson J G 1998 Demographic transitions and economic miracles in emerging Asia. *World Bank Economic Review* **12**: 419–55
- Boserup E 1965 *The Conditions of Agricultural Growth*. Aldine, Chicago
- Coale A J, Hoover E M 1958 *Population Growth and Economic Development in Low-Income Countries*. Princeton University Press, Princeton, NJ
- Kelley A C 1988 Economic consequences of population change in the Third World. *Journal of Economic Literature* **26**: 1685–728
- Kelley A C 2001 The population debate in historical perspective: Revisionism revisited. In: Birdsall N, Kelley A C, Sinding S (eds.) *Demography Matters: Population Change, Economic Growth and Poverty in the Developing World*. Oxford University Press, Oxford, UK
- Kelley A C, Schmidt R M 1994 *Population and Income Change: Recent Evidence*. World Bank Discussion Papers 249. The World Bank, Washington, DC
- Kelley A C, Schmidt R M 1995 Aggregate population and economic growth correlations: The role of the components of demographic change. *Demography* **32**: 543–55
- Kelley A C, Schmidt R M 2001 Economic and demographic change: A synthesis of models, findings and perspectives. In: Birdsall N, Kelley A C, Sinding S (eds.) *Demography Matters:*

- Population Change, Economic Growth and Poverty in the Developing World*. Oxford University Press, Oxford, UK
- Malthus T R 1965 *First Essay on Population, 1798, with Notes by James Bonar*. Augustus M. Kelley, Bookseller, New York
- National Academy of Sciences 1971 *Rapid Population Growth: Consequences and Policy Implications*. Johns Hopkins Press for the National Academy of Sciences, Baltimore, MD, 2 Vols.
- National Research Council 1986 *Population Growth and Economic Development: Policy Questions*. National Academy Press, Washington, DC
- Simon J L 1981 *The Ultimate Resource*. Princeton University Press, Princeton, NJ
- Srinivasan T N 1988 Modeling growth and economic development. *Journal of Policy Modeling* **10**: 7–28
- United Nations 1953 *The Determinants and Consequences of Population Trends*. Department of Social Affairs, Population Division, Population Studies No. 17. United Nations, New York
- United Nations 1973 *The Determinants and Consequences of Population Trends*. Department of Economic and Social Affairs, Population Studies No. 50, 2 Vols. United Nations, New York
- World Bank 1984 *World Development Report 1984*. Oxford University Press, New York

A. C. Kelley

Population Forecasts

Population forecasts summarize existing knowledge about the change in the size and structure of populations. They support decision making by providing broad perspectives for future development. In other cases they may provide detailed estimates for specific planning tasks. The methods that are currently in use were mostly developed in the 1920s, 1930s, and 1940s, before modern computers were available. Although the forecasts provide us with important insight into the possible future development of the populations, a major deficiency in those methods is that they provide the forecast users with inadequate assessment of the uncertainty to be expected. The reasons for the unpredictability of demographic processes are only poorly understood. However, the development of new statistical models and the availability of fast computing allows a realistic description of forecast uncertainty in terms that are understandable to nonspecialist users.

1. Development of the Cohort-component Method

1.1 Curve Fitting Approach

Early calculations concerning the renewal of populations were based on simple mathematical models.

Define $V(t)$ as the total population at time t . If the rate of growth is constant, or $V'(t)/V(t) = r$, then the population grows exponentially, $V(t) = V(0)e^{rt}$, from the starting or *jump-off* population $V(0)$. If the rate of growth decreases with increasing population according to $V'(t)/V(t) = r(M - V(t))/M$ where M is a limiting population size, then the growth is *logistic*, $V(t) = Me^{rt}/(1 + e^{rt})$. A discrete time version of the exponential model was used, for example, by T. Malthus in his *Essay on the Principle of Population* in 1798. The logistic model was introduced by F. Verhulst in 1838.

Such models were widely criticized because (a) they do not give age detail; (b) they do not allow a change from increase to decrease; (c) they are too mechanistic for human populations; (d) even though the exponential and logistic models are empirically indistinguishable when $V(t)$ is small relative to M , they give completely different long-term forecasts.

1.2 Incorporation of Age Structure

Methods for describing age-structured populations were developed around the turn of the nineteenth century in many countries, first in England by E. Cannan in 1895. By the end of the 1920s, forecasts by age and sex had also been made for the Soviet Union, The Netherlands, Sweden, Italy, Germany, France, and the United States (de Gans 1999).

Define $V(x, t)$ as the size of the female population at age x at the beginning of year t . Then, for ages $x > 0$, we have the bookkeeping equation $V(x + 1, t + 1) = V(x, t)p(x, t) + N(x + 1, t + 1)$, where $p(x, t)$ is the probability of surviving from age x to age $x + 1$ during year t , and $N(x + 1, t + 1)$ is the net number of migrants during year t who are at age $x + 1$ at the beginning of year $t + 1$. For age $x = 0$ we have $V(0, t + 1) = \sum_x f(x, t)V(x, t) + N(0, t + 1)$, where the summation is over the childbearing ages (often $x = 15, \dots, 49$), and $f(x, t)$ is the expected number of children at age 0 at the beginning of year $t + 1$ that were borne by women at age x one year earlier. Corresponding formulas can be written for the male population, and they can be extended to comprise several regions. This system of calculation is called the *cohort-component* (c-c) method.

1.3 Advantages of the Cohort-component Method

The c-c method can be viewed as a response to the criticisms of Sect. 1.1 in the following sense. (a) Since the c-c method treats the population by age groups, it provides direct estimates of the future population by age and sex. That this is useful for the planning of housing, schools, hospitals etc., was understood from the beginning. (b) The c-c method makes no as-

sumption about the growth rate. It shows that, other things being equal, the growth rate increases with the proportion of women of childbearing ages. In 1925 L. Dublin and A. Lotka introduced the concept of intrinsic rate of growth. For a closed population with unchanging fertility and mortality, it gives the asymptotic rate of growth. As such it provides important insight into the renewal of populations that cannot be detected from current growth rates. Many of the early cohort-component forecasts assumed unchanging fertility and mortality, and no migration, and showed that the size of the childbearing population, and with it the total population, would cease to grow, even though the populations had been growing fast during the early 1900s. This contradicted the forecasts by exponential and logistic models. (c) The reason why the c-c forecasts are less mechanistic than those based on exponential or logistic curves is that they utilize the available information or assumptions about the change of the *vital rates* (of fertility, mortality, and migration). The forecasters of the 1920s knew that fertility and mortality had been declining for decades. How far down would they go and how fast? Examples from other countries or from privileged social groups were used as leading indicators for future mortality. The spreading of the urban lifestyle was seen to lead to a decrease in fertility. Opinion surveys were developed around 1940 to collect data on desired family size (Whelpton et al. 1947). (d) The early forecasters knew that the exponential, logistic, and other curves often gave similar fits to past data, but led to very different forecasts. Since statistical *model choice* depended on judgment but had a drastic effect on the results, why could one not use judgment in other ways as well? The solution by Whelpton to this problem was to set up targets for future vital rates, and to connect the most recently observed rates to the targets by suitable mathematical curves. Dutch forecasters had also used similar methods. This practice has been widely followed, and most national population forecasts use judgmental methods for specifying future vital rates, often via explicit targets (e.g., Andrews and Beekman 1987).

Although the c-c method was an improvement over the curve fitting methods, it was soon understood that it is really a bookkeeping framework, and without an accurate specification of the future change in the vital rates improvements in forecast accuracy would not materialize.

As an example we may consider a forecast of Sweden that was crafted by S. Wicksell in 1926. The forecast had 1920 as the jump-off year with a population of 5.90 million. It had five fertility variants. For 1940 the most likely forecast was 6.53 million inhabitants. The high forecast was 6.87 and the low forecast was 6.28. The actual population size was 6.37 so the error was 2.5 percent. For 1970 the most likely value was given as 6.67 and the high-low interval was [6.26, 8.44]. The true value was 8.08 so the error was -17.5 percent.

Even though many of the errors were large, the high–low intervals were wide and they gave a realistic indication of the uncertainty of the forecast.

2. Foundational Aspects in the Forecasting of the Vital Rates

Most national forecasts of the vital rates are purely demographic, i.e., they do not explicitly incorporate social or economic factors. Only rarely do they take into account the fact that the forecast itself might influence the future development. Partly because of these factors, and partly because of the poor accuracy of past forecasts, there has been a lengthy debate about whether the results of the calculations would be better interpreted as mere ‘projections’ rather than ‘forecasts.’

2.1 The Difficulty with Explanatory Models

Refinement in demographic analysis often refers to the partitioning of the population at risk into more meaningful subgroups. For example, instead of relating births and deaths to the total population (which gives the crude birth and death rates), one may relate births to age of mother and consider deaths by age and sex (which gives the age-specific fertility and mortality rates) (see *Demographic Measurement: General Issues and Measures of Fertility; Demographic Measurement: Nuptiality, Mortality, Migration, and Growth; Demographic Techniques: LEXIS Diagram*). The use of crude rates in exponential or logistic models can imply growth forever, whereas the use of the age-specific rates in a c-c model may imply a future decrease of the same population.

The analysis can be further refined. Age-specific rates may be analyzed for birth cohorts and by parity; marital and nonmarital fertility may be considered separately; or one may consider fertility by race, for example. Empirical evidence of whether this improves accuracy is mixed.

However, the more fundamental social or economic factors that determine the fertility, mortality, and migration of population subgroups are rarely used explicitly in forecasting. This is because future changes in such factors can be harder to forecast than the vital processes themselves. Keyfitz (1982) reviews many of the difficulties.

One attempt at solving the problem has been *conditioning*. By making detailed assumptions about the future course of relevant social and economic factors, forecasters have sought to formulate conditional forecasts, or *scenarios*, in which the evolution of the vital processes would follow assumed relationships with their determinants. A drawback of such an approach is that the manner in which the conditioning is done can have a major effect on the results. A

sensitivity analysis that systematically varies some of the assumptions may then be helpful.

2.2 Self-fulfilling or Self-negating Prophecies

In some countries such as Germany and France the early population forecasts were motivated by the fear of population decline and loss of political power. In The Netherlands overpopulation was the concern. In both cases the forecasts were seen as a means of influencing social policies (de Gans 1999). Inasmuch as the forecasts were influential, they were self-negating.

In other cases such as regional forecasting, a forecast of population growth may cause the local society to develop the infrastructure of an area to allow for future growth. The favorable circumstances created may then turn the forecast into a self-fulfilling prophecy.

Although examples of both types of ‘prophecies’ exist, it has been difficult to demonstrate the causality in a historical context. Deliberate government attempts at lowering (or increasing) fertility, or lowering mortality, have seldom had long-lasting effects that could be attributed to the policies as opposed to other concurrent factors.

2.3 Debate on Terminology

From the very start forecasters have been aware that the future vital rates cannot be known with certainty. Evaluations of past forecasts have vividly demonstrated the unpredictability of demographic change. For example, the post-World War II baby boom was missed by all national forecasters. Mortality has similarly declined faster in the industrialized countries than the forecasters anticipated.

To safeguard their reputation some forecasters have occasionally tried to communicate the uncertainty by using the terms ‘projection,’ ‘illustrative projection,’ or even ‘horoscope’ for the results of their calculations. This is a reasonable practice, e.g., if the calculation consists of a c-c forecast with constant vital rates, when no one expects the rates to remain constant. However, as pointed out by H. Dorn in 1950, the producers of official forecasts typically publish calculations that are based on what they consider as the most likely future vital rates, and users also interpret the results in this way. Indeed, we may define a calculation based on what is regarded as the most likely assumptions as a *forecast*.

3. Statistical Uncertainty in Forecasting

Although probabilistic terminology is used in many branches of demography (e.g., probability of death in

life tables) (see *Life Table*), and experience has shown that forecasts of the future population have been highly uncertain, the forecasting models considered so far have all been deterministic. The recognition of the uncertainties in probabilistic terms opens up new possibilities to enhance the usability of the forecasts.

3.1 Uncertainty in the Events and in the Rates

Consider a child-bearing population of size V . Suppose the number of births during a year is B . We can think of B as a sum of V random experiments with an average probability f of a birth. The expected number of births is $E = fV$. Based on a Poisson approximation one can estimate that in a sequence of B values the relative variability (coefficient of variation) is $E^{-1/2}$. In a small regional population we may have $E = 100$, for example. Then, the births vary by about 10 percent each year due to random variation alone. However, if $E = 10,000$, they only vary by about 1 percent, and if $E = 1,000,000$ they vary by 0.1 percent each year. It was pointed out by J. Pollard in 1973 that much larger relative variations are observed in the numbers of births in different countries, so this type of Poisson variation does not account for the variability of the observed birth processes.

The situation changes if we allow f to change from year to year with a coefficient of variation c (or by 100c percent). Then the relative variation is at least c irrespective of the expectation E . The conclusion is that for forecasting large populations, the uncertainty lies in the randomness of the rates rather than in the Poisson type variation in the events themselves.

3.2 Logical Flaws in the High and Low Variants

Over the years it became customary to produce high and low variants, in addition to the most likely medium variant, to give the forecast users an indication of the plausible range of the future developments. Users have typically ignored the high and low variants. One reason has been the added complexity of having more than one number in a planning context. More fundamentally, an additional reason may have been that the high and low variants have not had a logically coherent interpretation. There are two reasons for this. First, suppose $[B_L, B_H]$ and $[D_L, D_H]$ are high–low intervals for births B and deaths D , respectively. In general, there is no guarantee that the two intervals cover an equal range, because no probability content for the intervals has traditionally been given. Any calculation based on intervals with unequal ranges is difficult to interpret. Second, suppose that the intervals are given a probabilistic interpretation, so that the probability is, e.g., 80 percent that each one of them contains the true value. This was first suggested by L. Törnqvist in 1949.

Suppose there is no migration. Then the increase of the population is $I = B - D$. One can show that the usual high–low interval $[B_L - D_H, B_H - D_L]$ for I has a *higher* probability than 80 percent to cover the increase. Only if the number of births and deaths have correlation -1 does the high–low interval have probability content 80 percent. But the assumption of perfect negative correlation is known empirically to be false.

In a traditional c-c forecast a very large number of vital rates are given high–low variants, and these are combined in complex ways to produce high–low intervals for population sizes at different ages. No uniform interpretation can be given to the latter intervals. For example, a United Nations forecast of the total world population published in 1998 gives a high–low interval of [7.4, 8.8] billions for the year 2030 and [7.3, 10.7] for 2050. Since it is not known what probability contents these intervals are intended to have, it is not clear how they should be interpreted.

3.3 Expressing Uncertainty Probabilistically

Modern probability theory provides a language that can be used to describe the uncertainty of the future vital rates and vital events. This leads to a *predictive distribution* of the future population given what we have observed in the past.

The availability of fast computing allows the use of stochastic simulation to describe numerically the predictive distribution. In practice this is done so that a set of age-specific vital rates is sampled from their respective predictive distributions. A c-c calculation is then made based on the sampled values to produce a possible *sample path* of the population in each age and sex group. This is repeated thousands of times and the results are stored. For any age–sex group, or for any larger population aggregate, one can represent the uncertainty of the forecast in terms of graphics (histograms, boxplots, etc.), or in terms of statistical summaries (means, medians, standard deviations, etc.). Such techniques are available for single region populations, and they are being developed for multi-regional (or multistate) populations, household forecasts, and other functional forecasts.

The key element in such calculations is the formulation of the predictive distributions of the vital rates based on statistical analysis of past data. This can be based on formal time series methods, the analysis of the accuracy of the past forecasts, and judgment.

Predictive distributions provide the forecast users with a richer view of the possible future developments than the traditional population forecasts. For users needing a point forecast only, one would typically offer the median of the predictive distribution. At the other end, a complete predictive distribution is available to a sophisticated user who wants to use formal *loss functions* in *risk analyses*. Stochastic forecasts are

well suited to the analysis of the financial stability of pension funds and other social security systems, for example.

Bibliography

- Ahlburg D A, Land K C 1992 Population forecasting. *International Journal of Forecasting* 8 (Special issue)
- Alho J 1997 Special section on statistical analysis of errors in population forecasting and its implications on policy. *Journal of Official Statistics* 13: 201–304
- Andrews G H, Beekman J A 1987 *Actuarial Projections for the Old-Age, Survivors, and Disability Insurance Program of Social Security in the United States of America*. Actuarial Education and Research fund, Itasca, IL
- de Gans H A 1999 *Population Forecasting 1895–1945: The Transition to Modernity*. Kluwer, Dordrecht, The Netherlands
- Keyfitz N 1982 Can knowledge improve forecasts? *Population and Development Review* 8: 729–51
- Land K C 1986 Methods for national population forecasts: a review. *Journal of the American Statistical Association* 81: 888–901
- Lee R D, Tuljapurkar S 1994 Stochastic population forecasts for the U.S.: Beyond high, medium and low. *Journal of the American Statistical Association*. 89: 1175–89
- Lutz W, Vaupel J W, Ahlburg D A 1998 Frontiers of population forecasting. *Population and Development Review* 24 (Supplement)
- National Research Council 2000 *Beyond Six Billion. Forecasting the World's Population*. National Academy Press, Washington, DC
- Smith D P 1992 *Formal Demography*. Plenum, New York
- Suzman R M, Manton K G, Singer B H (eds.) 1993 *Forecasting the Health of Elderly Populations*. Springer, New York
- United Nations 1998 *World Population Prospects: The 1998 Revision*. United Nations, New York
- van Imhoff E, Keilman N W 1991 *Lipro 2.0: An Application of A Dynamic Demographic Projection Model to Household Structure in the Netherlands*. Swets & Zeitlinger, Amsterdam
- Whelpton P K, Eldridge H T, Siegel J S 1947 *Forecasts of the Population of the United States 1945–1975*. U.S. Government Printing Office, Washington, DC

J. M. Alho

Population Policies, Politics of

Population policy has a long history of engendering political controversy that promises to persist as it is fueled by divergent views of the Good Society. The deeply held beliefs that make population policy an object of intense dispute have their roots in such issues as political and economic ideologies, religion, morality, nationalism, and ethnicity. The impetus for governmental intervention into demographic processes grows out of the realization that births, deaths, and migration are not merely a question of numbers, but

are critical to the social, economic, and political character of a nation. As Kingsley Davis has pointed out, 'Never, apparently, is a demographic goal sufficient by itself to justify a population policy. People judge a population trend to be good or bad only in the light of its presumed social and economic consequences' (Davis 1971, p. 6). Although Western nations have been deeply involved in the effort to curtail fertility in the Third World beginning in the 1960s, the West's experience with population policy has been mainly to encourage higher fertility in their own countries. This article will discuss the difficulties faced by governments that adopt policies to either lower or increase birthrates.

1. The Definition of Population Policy

The controversial nature of population policy extends to the definition of the term itself. There has been a tendency among scholars as well as government officials to abuse and misuse the definition. Population policy has often been defined in terms designed to advance a particular social or demographic agenda rather than to provide a rigorous structure for understanding and analysis. Many writers, including population professionals and students of demography, define population policy as any action of government that directly or indirectly influences population size, distribution, or composition (Eldridge 1954). Other writers go so far as to treat population policy as any government action that directly or indirectly exerts an effect on health, well being, status, or quality of life. These definitions are broad and encompass almost everything a government does. However, this is also the weakness of these definitions as almost every government policy, from the construction of roads to the provision of health services, has demographic implications.

For those interested in the formulation, content, or outcomes of population policy, a more exacting definition is required such as that found in the United Nations volume, *The Determinants and Consequences of Population* (1973). This definition, based on statements by Spengler and Duncan, states that population policy consists of 'a specific set of governmental objectives relative to the population magnitude and/or composition together with the instruments by which it may be possible to achieve them' (United Nations 1973, p. 632). Implicit in this definition is the expectation that government will establish an agency or organization to advance these goals, and will also provide resources to enable that agency to succeed. While the Spengler–Duncan definition provides a rigor that is missing in most definitions of population policy, it may at the same time not allow sufficient latitude for the give and take, negotiation, and compromise that characterize the policy process.

2. Structure Of Population Debate

The structure of the population debate was formed when Karl Marx bitterly attacked and condemned Malthus and his ideas more than a half a century after Malthus published his famous *Essay on Population* in 1798. Although Marx never formulated a 'theory of population,' he adamantly maintained that 'over population' was a consequence of the economic system, and would be remedied once the goal of socialism had been achieved. Marx recognized that if Malthus were correct to say that human poverty is anchored in human nature, then the intellectual edifice of Marxism was threatened. While the Marx–Malthus dispute provides a backdrop to the entire controversy over population policy, neither Marx nor Malthus had a profound affect on policies governments adopted until well into the twentieth century, when governments began to decide that their birth rates were either too high or too low to achieve national goals.

3. Population Policies in the Premodern Era

From earliest times, rulers and governors have attempted to increase the size and growth of their populations, which were periodically ravaged by famine, epidemic disease, and warfare. A large and growing population was universally regarded as a sign of a healthy and vital society while small or decreasing numbers signified weakness and decay. In classical Greece, the most martial city-state, Sparta, enacted laws that made marriage compulsory and imposed special taxes, civil indignities, and political disadvantage on celibates. Procreation was similarly encouraged in Rome when that city was still a city-state intent on conquest. Later, the famous population laws of Augustus, elaborated between 18BC and AD9, tried to address the low fertility of the senatorial class, on which the empire relied for the administration of its colonies (Hutchinson 1967). During this period pro-natalist policy had its origins in nationalism or imperialism, sentiments that would motivate pro-natalism through the mid-twentieth century. Before the modern era, however, the government lacked the means to reach into the hinterland, and it did not command the technologies required to enforce the law. Thus, it is doubtful that these policies were effective.

A second rationale for pro-natalist policy emerged in the Mercantilist states of Europe during the sixteenth to eighteenth centuries. To the Mercantilist rulers, a large and growing population was equated with a rich and prospering nation; therefore, the more people the better. While Mercantilist theorists disagreed over many points, most regarded labor as the most important factor of production. They also observed that high production density fostered industry and commerce and provided the basis for the establishment of colonies that would further enrich the

nation. Mercantilist policies thus encouraged marriage and procreation, severely limited emigration, and encouraged immigration, especially of skilled tradesmen. Once again, effective techniques for implementation were lacking and it is unlikely that the laws had much effect (Spengler 1960).

4. Population Policies in Developed Countries

In the belief that population size and national power were closely related, political leaders in the late 1920s and 1930s saw a low birth rate and population decline as nothing less than a threat to national survival (Tietlebaum and Winter 1985). Concerned by the phenomenon of below replacement fertility, several governments adopted policies to raise their birthrates in an effort to alleviate severe economic depression and assure that the state would not decline in its relative power status in the future.

Mainly in the 1930s, several European governments formulated policy responses to declining fertility, utilizing both repressive and social welfare measures (Glass 1940). The first moves were to re-activate abortion legislation already inscribed in the criminal codes, which imposed Draconian penalties for convicted abortionists and forbade advice on birth control and the advertising and sale of contraceptives—except for condoms which were regarded as prophylactics. On the welfare side, child and family allowances, marriage and housing loans to subsidize maternity leave, benefits to reduce infant mortality, the establishment of nurseries and day care centers, and campaigns of pro-natalist propaganda were among the most significant measures introduced.

While France, Sweden, and other European nations relied mainly on welfare measures to encourage higher fertility, Nazi Germany initiated a vigorous pro-natalist policy for those thought to be of Aryan stock while simultaneously pursuing a policy of sterilizing the 'hereditarily unfit' and 'racially impure.' At the same time that Germany was urging its women to bear more children for the Fatherland, it invaded neighboring countries on the ground that it needed additional space, *Lebensraum*, for its people. Germany also utilized the power of the state and its extensive control apparatus to enforce its prohibition against abortion (Harmsen 1950). However, other than in Nazi Germany the restrictive legislation on abortion in other European nations had only a small effect, as judges and juries were loathe to convict because of the severity of the penalty and their own lack of commitment to the abortion law (Glass 1940).

Despite the anxiety over 'depopulation,' the liberal democracies of Europe could not achieve national consensus over the adoption of coherent pro-natalist policies until the late 1930s. By this time, however, birth rates were already starting to rise as preparations for war contributed to the end of the Depression and

the return of full employment. Pro-natalist measures were introduced in the Soviet Union immediately prior to World War II and spread to Eastern Europe after the war. Moreover, pro-natalism in the Soviet Union was buttressed by Marxist ideology that held that 'overpopulation' was only a problem under capitalism.

Fertility started to fall in Communist bloc states in the mid-1950s, soon after abortion was liberalized in these countries. Eastern Europe responded quickly by placing restrictions on abortion, with the most severe in Romania. Though legal, contraceptives were to all intents and purposes difficult to obtain or unavailable. Child allowances were increased and long-term maternity leave of one to three years was introduced (McIntyre 1975). A benefit was provided that varied between the minimum and average wage and re-employment was assured in the same or a similar job. These policies largely succeeded in maintaining fertility above the replacement level.

In Western Europe, by contrast, fertility fell below replacement level, and governments once again adopted pro-natalist policies. Recognizing that in the contemporary world national power no longer depended on numbers alone, policy makers gave increased emphasis to enhancing the quality of the population, that is, the health, education, and other factors that contribute to what has been called 'the effective population' (Organiski et al. 1972). Equally important as a gauge of national power was economic and technological superiority, as well as a nation's structure of alliances. For those nations intent upon encouraging couples to have at least an additional child, the most effective approach seemed to be a policy that lessened the difficulties of balancing paid work and motherhood (McIntosh 1983). Pro-natalism proved to be less than a successful undertaking for Western democracies as they were constrained by their own liberal democratic politics from implementing the multifaceted measures required for an effective policy aimed at increasing the birth rate.

5. Population Policies in Developing Countries

Although population had been a subject of philosophical speculation and debate since ancient times, it was not until the middle of the twentieth century that it became a cause inciting intense political controversy. There were several developments that moved population from speculation and theory to the policy arena. Of singular importance were demographic trends in less developed and colonial countries that challenged leaders of these nations, especially after they had achieved independence. Earlier, both fertility and mortality rates were high. After World War II the introduction of improved public health measures, especially clean water and improved sanitation, as well as the new availability of antibiotics, led to a decline in

infant and child mortality and an unprecedented increase in the rate of population growth.

Beginning in the 1950s, leaders in developing nations came to regard rapid population growth as an obstacle to their country's economic development, a view that was reinforced by census results which showed that population size and growth exceeded earlier expectations. Four years after gaining its independence from the UK, India adopted a population policy as part of its initial five-year plan in 1951. In roughly the same time period, Pakistan and Sri Lanka instituted population policies. There were also leaders and political elites in many developing countries who, while they recognized the burden of rapid population growth, were nevertheless reluctant to adopt population control policies, partly because of the 'distrust the leaders had of the former metropolitan powers' projection of population as the greatest threat to national development.' (Made with reference to Sub-Saharan Africa in UNECA Population Policy Overview Report 2000).

Meanwhile China, the world's most populous nation, captured the imagination of many Third World countries at the Bucharest Conference in 1974, by proudly proclaiming that socialist China could accommodate increasing numbers of people. Significantly, within five years China implemented the most vigorous population control program in modern history. The sharp changes in China's population policy were not merely a response to demographic conditions but were closely linked to the revolutionary fervor of China's leadership and their economic goals.

The One Child Per Family policy, adopted by the People's Republic of China in 1979, imposed penalties—at times severe—on those who abrogated the law. The policy succeeded and China achieved the sharpest fertility decline that any country had ever experienced. In achieving this goal, many women either sought or were compelled to undergo abortions in order to avoid a second child and avoid the possible legal consequences. A better understanding of demographic processes by Chinese officials combined with resistance to the one child policy within China and criticism from abroad, contributed to a *de facto* modification of the policy (Banister and Harbaugh 1994). Chinese officials realized that their population goals could be achieved, albeit more slowly, and with less public resentment, by implementing the one child policy less strictly.

Population policy came to be recognized internationally as emblematic of a nation's commitment to modernization. This trend was abetted and encouraged by the rich industrial nations, especially the USA, as well as by the United Nations, most particularly the United Nations Population Fund (UNFPA) and the World Bank. Industrial nations of the West felt that high fertility in poor countries would keep them poor forever. Consistent with this attitude, the USA, Canada, European nations, Japan, and

indeed all donor nations began to support efforts to limit population growth.

In 1967, partly as a result of a financial contribution from the USA, the UN began to address the issue of population growth. The Secretary General created a trust fund which eventually became the UN Fund for Population Activities (UNFPA), whose purpose was to encourage countries to identify their population problems and to attempt, with international assistance, to cope with them. Although many branches and organs of the UN would eventually be involved in curtailing rapid population growth, the creation of UNFPA was critical to the entire multilateral effort.

A point that is often neglected is that 'overpopulation' was attractive to both industrial societies as well as to developing countries as an explanation for the poverty of much of the Third World. It was much easier to attribute 'underdevelopment' to excessive population than to confront social and economic inequality, culture, religion, female subordination, or other plausible contributing factors. In the same spirit that western industrial nations found population control an appealing remedy for Third World poverty—for population control did not demand radical changes in the social and economic structure—scholars, intellectuals, and political leaders in developing countries also embraced population control as it did not require them to question fundamental attitudes and beliefs of their society. Thus, poor countries felt they could address their problems of development without first coming to grips with the social costs imposed by established structures of their societies. The discourse on the relationship between development and population has flowed from country to country and in the process has become universal rather than contained by national boundaries. The centrality of population to society assures that population issues will not be confined to any subsector of society but will become an issue that demands the attention of the polity. This has been the case in the past, and there is no reason to question its continuation.

6. The US Influence in Population Policies

The USA played a particularly significant role in shaping world population policy starting in the 1960s and in framing the modern debate over the politics of population. Until the 1960s, the USA was firmly opposed to playing any role, except a negative one, in the incipient population control movement. In words that he later came to publicly regret, President Dwight D. Eisenhower said, 'I cannot imagine anything more emphatically a subject that is not a proper political or governmental activity or function or responsibility' (Critchlow 1999, p. 44). Under Eisenhower's successors the American government became steadily more supportive of family planning at home and abroad.

Once the USA decided to assist poor nations to curtail their population growth, it became an American 'cause.' The USA soon became the world's largest donor to population assistance and used its political influence and financial resources to motivate the United Nations system to support family planning efforts in developing countries. Although the UK, Scandinavia, and other northern European states were important in the spread of the population movement, it was once again the political and economic influence of the USA that played a dominant role in persuading Third World countries that family planning programs were essential to their social and economic development. Equally important, the USA developed a strong professional cadre of population specialists covering every aspect of population and family planning from medicine to demography to management and communications, to assist family planning programs in almost every country in the developing world.

Perhaps of even greater importance in shaping population policy as well as in creating an international community of population experts was the investment made by the USA in training large numbers of men and women from the Third World at American universities. Upon returning home, graduates of these educational programs often filled critical positions in development programs and family planning. The USA also invested heavily in applied research designed to enhance the effectiveness of family planning programs as well as to provide an empirical base for health policy. One early research effort, the surveys of family planning and contraceptive knowledge, attitudes, and practice (KAP), demonstrated in country after country that women were having more children than they desired. The scientific contribution of KAP surveys was limited; however, they were believed to be politically valuable in persuading leaders to adopt national population policies.

As the USA exported its expertise and beliefs in voluntary family planning, so too did it export its profound political problems centering on the population issue. In the USA, the population dispute was entangled with such issues as foreign aid, state intervention in fertility decisions, sex education, contraception as a source of immorality, and, most serious of all, the abortion controversy, the most divisive issue in American society since the Civil War.

Family planners hold that abortion is a necessary backup procedure in the event of contraceptive failure. Other supporters of abortion argued that it is an essential component of any successful effort to limit population growth. A third group, led by but not limited to feminists, held that it is simply a woman's right to choose whether or not to have an abortion. Thus, while family planning and abortion are analytically distinct, there is a general tendency for family planning proponents to be pro-choice. Conversely, anti-abortion or pro-life advocates were prone to broaden their target to include opposition to family

planning. Both pro-life and pro-choice advocacy groups created international linkages whereby they have attempted to replicate the Western abortion controversy in developing countries as well as in the international arena.

7. Population Policies and International Conferences

United Nations conferences reflect the highly political and ideological quality of population policy at the international as well as the national level. Since the Stockholm Conference on Environment in June 1972—a conference generally regarded as highly successful—the United Nations has increasingly relied on the mechanism of international conferences to shape global policy on a variety of social and economic issues ranging from women, children, and human rights to population and development. In contrast to regular UN proceedings, these conferences have been more open to the inclusion and participation of nongovernmental organizations representing a broad spectrum of interest groups. To the extent that the policy process becomes more open and diverse groups acquire access to power, it is almost inevitable that political considerations are likely to play an important, if not a dominant, role in formulating policy at UN population conferences.

United Nations sponsorship of international population conferences actually began with gatherings in Rome in 1954 and Belgrade in 1965. The United Nations jointly sponsored these conferences and the International Union for the Scientific Study of Population (IUSSP), a highly respected international nongovernmental organization (NGO) with a comparatively small but influential worldwide membership of population professionals. The UN and IUSSP invited population experts to participate in the Rome and Belgrade conferences. These meetings, while of some intellectual significance, had only a modest impact on national population policies. Conference participants were invited as 'experts' rather than as representatives of government. Concerned about the lack of policy impact, the United Nations, as well as some of its more powerful members, changed the structure and composition of these early conferences. Their intent was to move the population debate out of the hands of scholars and experts and into the domain of government where population policy is made.

Under the leadership of the United Nations Population Fund (UNFPA), and with the strong support of the USA and other Western donor nations, the UN convened its first of a series of decennial population conferences in which governments rather than individual experts were represented. In contrast to the earlier World Population Conferences in Rome and Belgrade where participants were invited in their individual capacities as experts, delegates to the UN

population conferences in Bucharest in 1974, Mexico City in 1984 and Cairo in 1994, were representatives of their government. Members of national delegations were not exclusively government officials and political leaders; governments also chose to select physicians, family planning activists, social workers, as well as feminists to represent their countries.

Contrary to expectations, these changes did not result in a stronger commitment by governments to population policy. Instead, they opened the door to a variety of interest groups to advance their own agendas and redefine population policy in terms that were more congenial to their interests. Governments also found these intergovernmental conferences a convenient host to their efforts to promote their own ideological and political agendas in collaboration with other friendly governments. Population conferences became far more controversial, interesting, and perhaps even more significant. For example, at Bucharest the Population Plan of Action, the consensus statement of the conference, urged a new and more equitable international economic order. In language that resembles a Marxist critique of development theory, the Background section of the Plan held that population problems are really symptoms of imbalances in the development process and a faulty international economic system. Twenty years later at Cairo, the new definition of population was one that gave primacy to women's reproductive health and the status of women. Nongovernmental organizations were far more prominent at Cairo than at either Bucharest or Mexico City, a move intended to gain the support of the growing universe of nongovernmental organizations. This too resulted in further politicization of the conference proceedings as NGOs at Cairo were articulate, well organized, and embraced an agenda that paid little attention to demographics.

8. The Future of Population Policies

Notwithstanding the near universality of fertility limitation programs in developing nations, world population reached 6.1 billion in mid-2000 and, according to UN estimates, was growing at an annual rate of 1.2 percent or 77 million people per year (United Nations Population Division 2001). According to the United Nations 'medium' projections, by 2050 world population will reach 9.3 billion. This growth is not evenly distributed throughout the world: almost all of it is taking place in developing countries. Clearly, on a global basis, the total fertility rate is above replacement level. Population growth will continue, if for no other reason than the momentum inherent in the age structure, that is, the large numbers of young people who have yet to complete their reproductive years (Bongaarts 1994). Thus, if world population is moving toward stabilization, it is moving at a much slower pace than the population community

predicted or desired. To the extent that societies and political leaders are dissatisfied with demographic conditions in their own countries, there remains little question that they will introduce population policies in an attempt to ameliorate these conditions. Whether they will set clear goals, create an organization to implement these goals, and provide sufficient resources to successfully carry out their policy remains, as in the past, a matter of political commitment.

See also: Environmental Policy; Family Planning Programs: Development and Outcomes; Family Planning Programs: Feminist Perspectives; Fertility Control: Eastern Europe; Fertility Control: Overview; Fertility: Institutional and Political Approaches; Fertility: Political and Political-Economic Perspectives; Malthus, Thomas Robert (1766–1834); Population Policy: International; Reproductive Rights in Developing Nations; United Nations: Political Aspects.

Bibliography

- Banister J, Wu Harbaugh C 1994 *China's Family Planning Program: Inputs and Outcomes*. Center for International Research, Bureau of the Census, Washington, DC
- Basu A M 1997 The 'politicization' of fertility to achieve non-demographic objectives. *Population Studies* 51(1): 5–18
- Bongaarts J 1994 Population policy options in the developing world. *Science* 263: 771–6
- Chamie J 1981 *Religion and Fertility: Arab Christian–Muslim Differentials*. Cambridge University Press, Cambridge, UK
- Critchlow D T 1999 *Intended Consequences. Birth Control, Abortion, and the Federal Government in Modern America*. Oxford University Press, New York
- Coale A J 1986 Population trends and economic development. In: Menken J (ed.) *World Population and US Policy: The Choice Ahead*. W. W. Norton, New York, pp. 96–104
- Davis K 1971 The nature and purpose of population policy. In: Kingsley D, Styles F (eds.) *California's Twenty Million: Research Contributions to Population Policy*. Institute of International Studies, University of California, Berkeley, CA
- Demeny P 1975 Population policy: The role of national governments. *Population and Development Review* 1(1): 147–61
- Dixon-Mueller R 1993 *Population Policy and Women's Rights: Transforming Reproductive Choice*. Praeger, Westport, CT
- Eldridge H 1954 *Population Policies: A Survey of Recent Developments*. International Union for the Scientific Study of Population (IUSPP), Washington, DC
- Finkle J L, McIntosh C A 1994 (eds.) The new politics of population: Conflict and consensus in family planning. *Population and Development Review* 20 (Suppl.)
- Finkle J L, Crane B B 1975 The politics of Bucharest: Population, development, and the new international economic order. *Population and Development Review* 1(1): 87–114
- Glass D V 1940 *Population Policies and Movements in Europe*. Clarendon, Oxford, UK
- Harmen H 1950 Notes on abortion and birth control in Germany. *Population Studies* 3(4): 402–5
- Heidenheimer A, Hecl H, Adams C 1990 *Comparative Public Policy*, 3rd edn. St. Martin's Press, New York
- Hutchinson E P 1967 *The Population Debate*. Houghton Mifflin, Boston

- McIntosh C A 1983 *Population Policy in Western Europe: Responses to Low Fertility in France, Sweden, and West Germany*. M E Sharpe, New York
- McIntyre R J 1975 Pronatalist programs in Eastern Europe. *Soviet Studies* 27(3): 366–80
- McNeill W H 1990 *Population and Politics Since 1750*. University Press of Virginia, Charlottesville, VA
- McNicoll G 1997 Governance of fertility transition: Reflections on the Asian experience. In: Jones G et al. (eds.) *The Continuing Demographic Transition*. Oxford University Press, Oxford, UK
- Miro C A, Potter J E 1980 Population policy: Research priorities in the developing world. *Report of the International Review Group of Social Science Research on Population and Development*. St. Martin's Press, New York
- Organiski A F K, Bueno des Mesquita B, Lamborn A 1972 The effective population in international politics. United States Commission on Population Growth and the American Future. Research Reports, Vol. 4, *Governance and Population*. United States Government Printing Office, Washington, DC, pp. 235–50
- Spengler J J 1960 Mercantilist and physiocratic growth theory. In: Hozelitz B F (ed.) *Theories of Economic Growth*. The Free Press, New York, pp. 3–64, 299–344
- Spengler J J, Duncan O D 1956 *Population Theory and Policy*. The Free Press, Glencoe, NJ, pp. 441, 456
- Teitelbaum M S, Winter J M 1985 *The Fear of Population Decline*. Academic Press, London
- United Nations, Department of Economic and Social Affairs 1973 The determinants and consequences of population trends: New summary of findings on interaction of demographic, economic, and social factors. *Population Studies* 1: 50
- United Nations, Department of Economic and Social Affairs 1999 Below replacement fertility. *Population Bulletin of the United Nations*. Special Issue Nos. 40/41
- United Nations 2000 United Nations economics commission for Africa 2000. *Population Policy Overview Report*. <http://www.uneca.org>
- United Nations Population Division. February 28, 2001. *World Population Prospects: The 2000 Revision (Highlights)*

J. L. Finkle

Population Policy: International

There are two distinct but highly interrelated elements within international population policy. The policy set of goals and activities by states or international entities intended to influence and shape population outcomes in countries beyond their own national boundaries finds expression in both *bilateral* relationships and in *multilateral* activity. Bilateral activities seek to change population outcomes in specific countries. The collectivity of policies of the international community, expressed in multiparty international agreements, charters, covenants, and—most important for the population field—in international conferences and conference resolutions are *multilateral*. This activity is designed to standardize views and practices, and to

form a world-view of desirable goals and outcomes. The two are linked. Countries with active national or bilateral programs also seek to influence the direction of multilateral, international population policy.

1. The Unique Context of the Twentieth Century

Humans have probably always tried to manage fertility at the level of the individual and family. Rites and rituals to encourage fertility are time honored. There are also specific hieroglyphs in Egyptian antiquities citing herbal methods of contraception or abortion. More than a century before Christ, Polybius deplored both the tendency of comfortable urban Greeks to delay marriage and sharply reduce family size, and the resultant decline in population.

National concern and international preoccupation with population issues is very much a phenomenon of the last half of the twentieth century. While anxiety for the health impact on women of large families and unwanted pregnancy has been of long standing concern to health care providers, the major upsurge in interest in population issues originated in concern for the impact of rapidly accelerating global population numbers. Some appreciation of the remarkable quality of twentieth century demographic developments is therefore essential to understanding this policy evolution.

Never before the twentieth century did a human being live through a doubling of population; indeed the global population quadrupled between 1900 and 2000. The population increase in the 1990s alone was the equal of all of those who lived in the seventeenth century. In the preceding millennia, a good century for the human race was one in which the population grew at all. Plagues and normal high death rates in most societies ensured that this was not always the case.

Although everyone dies eventually, human death or mortality at an early age is what keeps population levels low. Absent high levels of mortality, migration, or decreasing fertility, population levels will increase. With the relative decline of famines and epidemics and increases in transport, communication, public health, and control of infectious diseases, nineteenth century Europe grew rapidly. About one-third of this increase was exported to colonies and overseas, a series of deliberate but minimally enunciated international population policies. The same spurt happened in post-W.W. II in the nonindustrialized countries emerging from colonialism, or 'developing world.' The increases in numbers were bigger, and the pressure valve of migration was no longer available. The world was regarded as settled by the major players who controlled immigration flows.

When the global population numbers really began to grow, when the ability to count them improved markedly, and when attitudes regarding sexuality and

family issues began to change in the industrialized world, population issues arrived on the global scene and international population policy began.

Twentieth century population growth changed the shape and weight of international relationships. Over 95 percent of the population increase of the last decades took place in the developing world. At the beginning of the century, almost one person in five was European; at the end, less than one-tenth of the world lived in Europe. The six billionth child was born in the last year of the century. That child will very likely live long enough to see the peak of human population and the stabilization of population numbers. The trends that will cause the decline of high fertility are well begun. Although another huge population increase will take place in the first two decades of the twenty-first century, the end of the high fertility period is in most places now in sight. International population policy in the twenty-first century will likely focus on different issues, already emerging.

2. The Content of Population Policy: Sensitive, Fragmented, Confrontational

There has never been policy consensus on the nature, importance, or remedies in this field. The first stage of policy making revolved around the question of whether rapid population growth and high levels of fertility were in fact a threat, and what to do about them if they were. Many continued to insist that high levels of population growth were a transitory factor that would decrease with development and progress. The answer was seen to be investment in accelerated general development, not specifically targeted programs, i.e., 'development is the best contraceptive.'

Those on the other side of the issue used phrases such as 'population bomb' to describe what Malthus in the late eighteenth century foresaw as predictable crisis deriving from the geometric nature of human growth expansion, not sustainable by the foreseen linear growth of agricultural production. Concerns about the real and also postulated but never provable impacts of population growth expanded beyond food production.

Early advocates for international fertility reduction efforts voiced fears that such growth would lead to mass starvation, perpetual economic underdevelopment, internal conflict, and external wars to expand boundaries. In late 1960s there were confident assertions from certain US thinkers that the battle to feed humanity was over and that widespread famine would shortly begin. In later years, the list of concerns expanded to include the impact on the environment, climate change, and the battle to preserve biodiversity.

On the other side of the ledger, concern for the human rights implications of family planning programs and the emergence of feminist concerns in their

turn significantly reshaped the content of international debate and many population programs. Family planning and contraception services began in the industrialized world with Margaret Sanger and Marie Stopes in the USA and UK, respectively. These programs were delivered through the medical model of a doctor–patient relationship. What developed in the 1960s and 1970s was the novel idea that the state, or a public health service provider, should provide these services on a mass basis in order to bring down the number of births. Thus was born the idea of officially endorsed (or at least tolerated) family planning programs. Begun in Southeast Asia in the 1960s, family planning programs spread throughout the developing world in the subsequent 20–30 years. Internationalization occurred as foreign states and international actors intervened in such domestic programs, sometimes providing the impetus for their creation. The full impact on policy of the multilateralization efforts came somewhat later, with the creation of an international network, influenced by and influencing the existing programs.

The tools of the family planning part of population policy were largely unavailable until the 1960s. Such contraceptives as existed in the first half of the twentieth century were unwieldy, unmentionable, and often illegal. The advent of the contraceptive pill in 1960 provoked a revolution in behavior in the industrialized world. With it grew the conviction that new contraceptives, especially those which could be injected, implanted, inserted, or swallowed as opposed to being applied at the time of sexual relations, had the potential to stop or slow the rising levels of population in the developing world.

Yet another part of debate centered on the degree of importance that should be accorded to these tools and to the provision of contraception services vis-à-vis other fertility-influencing interventions such as girls education. This issue was characterized as the ‘beyond family planning’ debate. Later, with feminism came a strong concern that women’s health issues must override concern for demographic increase in the population agenda. Included were preoccupations related to both the mechanism of program delivery, particularly targets, and several of the recently developed contraceptives. This discernibly altered and shaped population policy programs from the 1990s onward. The emergence of AIDS similarly pulled new elements related to the reproductive health of individuals into the population policy ambit.

These debates and uncertainties and the transitory coalitions that have endorsed particular actions at specific times have given changing goals and forms to international population policy.

While demographic trends can be discussed in statistical and abstract terms, and contraceptive techniques discussed in terms of their mechanisms of action, the reality of population change is grounded in the most personal human relationships, in human

sexuality, and in highly sensitive social issues. These include the nature and power relationships that establish marriage, the role and status of women, the behavior of adolescents, definitions of morality and immorality, the ethical judgments related to sexual activity, and to the very value attached to life at various stages.

For this reason, the myriad of social institutions that define different societies, including organized religions and political forces, have strong interests in most population issues. The major religious have been supportive with the notable exclusion of official Roman Catholic policy and the orthodox elements of many established churches. The moral and ethical issues are complex and the introduction of state activity in these areas therefore often highly contentious. That foreign states might have an interest in the decision of a man and woman in another country to have or not have more children could not but raise very complicated issues indeed.

3. Population Issues Enter Foreign Policy

For most of human history, countries did not have population policies, domestic or international. There was a long tradition of hope for population increase as a sign that prosperity was edging out pestilence and disease. The 1901 Census reports for India expressed the cautious hope that population would increase in the decades ahead. Nongovernmental activity preceded both governmental and intergovernmental activities in the national and international spheres. A major step forward was the formation of the International Planned Parenthood Federation (IPPF) in 1952, which grew by the end of the twentieth century to be a federation of advocacy and service providing organizations in over 150 countries, supported by another 20. The Indian Family planning program began the same year, as did several US nongovernmental organizations (NGOs).

There was a major upsurge in population activities in the late 1960s. Governments became less timorous and more active. Several dozen heads of government agreed to come out in favor of moderating global population growth, and agreed to be featured with photos and quotations in a widely used poster campaign. The World Health Organization (WHO) began population activities. The UN Secretary General established in 1967 a Trust Fund which later became the UN Population Fund or UNFPA. The World Bank adopted a policy to lend for population activities after President Robert McNamara declared that population growth constituted a threat second only to the prospect of nuclear war. (Continuing ambivalence within the Bank about the actual impact of population growth on development has, however, meant that while loans are available in aid of Health Ministry programs, Finance and Planning ministries were never

pressed by bank officials to support or give any special priority to population programs. Many other aspects of national management have been subject to such suasion over the years.)

3.1 The International Community Response: A Wide Spectrum of Policy Perspectives, and Some Domestic Determinants

As foreign policy is the pursuit abroad of domestic policy goals, it could be postulated that even though the desired impact of bilateral population programs was to change outcomes abroad, donor countries would normally not have international population policies until they began to have domestic population policies. This was by and large the case. Most states approached the issue of their domestic population issues with a great deal of diffidence. President Eisenhower is widely quoted as saying that he could imagine 'no subject less fit for government activity' than family planning. The United States became the unquestioned global leader in this field (see Sect. 3.2), while a variety of stances characterized the roles and programs of other countries.

Sweden and the other Nordics countries were early financial supporters of family planning in developing countries, reflecting their relative lack of embarrassment in dealing with these issues domestically. In the later decades, as the influence of women's health advocates grew at home and abroad, Sweden in particular disclaimed linkage between providing contraceptives to women to support them in their own family size decisions and any demographically related outcome.

Successive French governments pursued strong pronatalist policies at home in France and perceived that strong support for family planning programs abroad could well be seen as somewhat racist. There has never been major French support. The UK bilateral programs in many areas included a vigorous population component. Italian governments were widely believed to have exchanged peace with the Vatican on the domestic contraception/abortion front in Italy, in exchange for Italian inactivity abroad. Canada was at one time the third largest supporter of population programs. Interest and support declined rapidly in the 1990s, probably reflecting aid officials' desires to distance themselves from US programs, and some political level wish to avoid exacerbating deemed Quebec sensitivities. Both are longstanding elements of Canadian domestic policy.

The Netherlands from the mid-1990s dedicated a full 4 percent of their aid programs to population and reproductive health. They often played the role of coordinator and host to conferences and meetings on a variety of population-related issues. Their domestic belief in frank discussion and the provision of open explicit information starting at an early age has translated into an active international program.

Japan, under heavy US pressure, dedicated fairly large amounts of funding to population programs, very broadly defined. Once again, the impact of domestic constraints for a long time meant that the nonavailability of hormonal contraceptives to Japanese women precluded their inclusion in the aid package.

The developing world presented a mixed picture. Thinkers in Mexico, Egypt, and India before independence worried about population gains destroying the positive impacts of their own development efforts. India's 1951 five-year plan had population program expenditures. The Southeast Asian countries were early and enthusiastic advocates with national programs established in Thailand, Taiwan, Singapore, and Korea. Over a two-decade period, most southern countries adopted some form of population policy, after a period of initial resistance on the part of many. Virtually all countries eventually formulated population policies, influenced by their own cultural and religious backgrounds, and development aspirations. Until the advent of the international conferences described below, only industrialized countries had international dimensions to their population policies. The advent of the conferences gave developing countries the opportunity to express views on population issues on the world stage.

Bilaterally, the Vatican has acted not as a state but as religious advisor to governments. It has been especially powerful in keeping official support away from these services in Latin America, and to some extent in Africa. This reluctance to endorse any form of contraceptive device continued to include proscriptions against condom use, even as the AIDS epidemic affected millions.

The Marxist countries during the Cold War period had an avowedly antifamily planning stance. The states of the former Soviet Union accorded little or no importance to the provision of contraception with subsequent high reliance on abortion. Population levels dropped sharply in many socialist countries and desperate, coercive measures were adopted by several. The pronatalist policies of Ceausescu in Romania provided the sharpest example. The Marxian view was that difficulties caused by population growth really reflected maladjustment in wealth distribution. Although they did not proselytize to any major extent, this view was very strongly echoed by many developing countries in the early years. Particularly since it fit well with other themes of the New International Economic Order, the 'development is the best contraceptive' theme became the most commonly used rhetorical position in opposition to the new enthusiasm.

3.2 Global Leadership—the United States Role

Unquestionably, the global leader in this field has been the United States. American activity began in the 1960s and the US quickly became the most enthusiastic

proponent of population programs. A bipartisan, centrist coalition in Congress maintained support for international family planning programs, even in the Reagan years. Domestic NGOs appeared in the US for the express purpose of lobbying for greater global attention to this issue. The early name of one such group—ZPG or Zero Population Growth—indicates the intensity of public policy concern that underlay US policy aimed stabilization of global population levels.

Later with the resurgence of the religious right and the 1990s shift to political conservatism, the abortion issue re-emerged to cause violent schisms in US society. It defined presidential campaigns, filled the courts, and provoked violence against domestic abortion providers. The bipartisan consensus that had supported the family planning programs shattered. Republicans' public policy stance became the equation of family planning with abortion, and therefore opposition to all forms of public involvement in it. Democrats remained the champions of such programs.

The international policy reflection of this schism caused sharp reductions and severe restrictions on US support. Migration for economic improvement and to escape unfavorable political and social situations around the world will continue to be a growing population issue.

Specific measures of censure for China's one-child policy were included in funding conditions. Multilateral organizations were cut off from funding. A so-called gag rule was enacted. It prohibited the release of US family planning funds to international or foreign NGOs which use their own funds to provide abortion services or participate in policy debates on abortion outside the United States. This measure, which would be unconstitutional in the USA itself, was put into effect once by administrative decree in the Reagan years, and again by law in the final Clinton years.

The US flag has flown higher over population assistance in most countries than anyone else's. One of the advantages was the creation of unparalleled expertise in American government, nongovernmental, academic, and program delivery circles. The visibility was also a liability, especially in the early years when sensitivities were high about western enthusiasm for reducing southern births. Because of the intensity of US internal debate, and the rapidity of policy shifts, the US has often been at odds with the rest of the world. It was far out in front on global targets and Malthusian rhetoric in 1974, then pulled back a decade later when just about everyone else was finally coming to some policy consensus in favor of attenuating population growth.

3.3 The Tools of Population Policy: Aid Programs

The tools of bilateral population programming have included policy statements, speeches, persuasion,

advertising, and funding. Assistance has included policy-influencing computer simulations, assistance to demographic and census departments, soap opera presentations, and research on a wide variety of reproductive health elements. Much program activity focused very directly on the provision of family planning services. Funds have been provided for contraceptive development, dissemination, and for the creation of national family programs.

Family planning programs had no real counterpart in developed countries, and were often 'vertical' or stand-alone programs, like some immunization campaigns. They occasionally became the focus of religious, traditional, and xenophobic criticism. These clinics often provided the only services available or available to poor women. In terms of impact on fertility, family planning programs were deemed to have succeeded; they probably accounted for about 40 percent of the decline in otherwise anticipated births that characterized the world as the demographic transition progressed in the countries of Asia and Latin America.

Because of the sensitivity of population and reproductive health issues, many donor countries directed a substantial percentage of their assistance through multilateral and international delivery mechanisms. Almost US\$500 million flowed through multilateral agencies two years after the 1994 Cairo Conference. The programs of the primary multilateral agency, the UN Population Fund, reached \$320 million in its high water years, usually providing a channel for about 25 percent of available donor funding. World Bank lending at one point reached \$500 but declined at the end of the century. The regional banks have not been major players, with some exception for the Asian Development Bank. An increasingly important element of support was the assistance provided by (largely US) foundations to the population field, reaching as high as \$150 million just after the 1994 Cairo Conference.

Aid or official development assistance to population/family planning was never large in relation to overall expenditures in the field or to overall levels of overseas development assistance. By the end of the twentieth century, developing countries were paying three-quarters of the costs of their own reproductive health and population programs.

Nor have population programs dominated the overall aid programs. Even within the US program, population assistance only ever represented about 7 percent of all US aid. Australia, Denmark, Finland, The Netherlands, Norway, and UK the population comprised about 3 percent and in France and Italy only 1 percent of their respective ODA programs. In dollar terms, when the totality of all countries' official development assistance was running around US \$60 billion per year, population assistance (mostly family planning) never got above \$2.0 billion in total. Eight countries almost always supplied 90 percent of all

population assistance. The US gave the biggest amount, usually about half of all bilateral aid to population. Denmark, Norway, Sweden, and The Netherlands gave higher percentages, relative to their own economic weight.

If not significant as a percentage of overall aid, or as a proportion of each country's aid program, foreign aid to population has often been *very* significant in relation to the total health budget of many developing countries. It has had a catalytic impact in determining the scope, content, impact, and in some places existence of programs across the world.

4. Multilateral Policy Making

The second zone of international population policy activity is multilateral or many sided in nature. It is important for two reasons. As just noted, multilateral organizations have been important sources of technical and capital assistance, and vital to program delivery. The second key multilateral activity relates to international population policy making within the international community. Here, all countries have roles; the multilateral organizations provide the venue and organization for this policy-making activity.

4.1 The Population Conferences—Policy Tracers Par Excellence

It is through the phenomenon of the international population conferences that the changes and shifts of international policy in this area can be traced. Margaret Sanger organized an international population conference in Geneva in 1920. It resulted in the formation of the first international nongovernmental population organization, the International Union for the Scientific Study of Population (IUSSP), mostly but not exclusively demographic in purview.

Modern population conferences are not like those of the World International Trade Organization where treaties are negotiated and monitored. Unlike conferences such as the 1948 Universal Declaration of Human Rights, they do not require countries to take legal steps to bring domestic activity into line with new policy norms. They are rather international conclaves of policy expression, where undertakings of good faith are made by all, and statements are made and resolutions drafted which together give a comprehensive picture of how the world's thinking has evolved on an issue. They also give the world the opportunity to comment on individual programs: the one-child policy of China and the specific programs of Indonesia and India have from time to time drawn attention.

The preparatory work preceding such conferences is key, often including the collation of detailed country statements which can include reviews and revisions of laws, budgets, policies, and program descriptions.

Most important, the preconference activity occasions consultation and negotiation on the domestic front in most countries. As the issue has internationalized, comparable activity takes place on a regional basis, and globally among NGOs. As NGOs and women's groups have become more important players, both the openness of the preparatory process and the difficulty in reaching consensus have grown. At the conferences themselves, NGO representatives have come to take part in the delegations and to play a role in formulating the consensus documents. Generally, countries reflect at these conferences the positions they have adopted domestically. The Vatican chooses on this issue to behave as a State in terms of attendance at international conferences. It works through lobbying, diplomatic demarches, and public diplomacy.

Population conferences as international policy making occasions therefore provide highly visible venues for intense debate about sensitive reproductive health issues of safe abortion, genital cutting, violence against women, and services for adolescents. These continue to provoke confrontation and contention.

4.1.1 Bucharest—1974. By the end of the 1960s, there was considerable enthusiasm—in the industrialized world—for serious attention to be given to population levels in the developing world. Serious attention in this era really meant a focus on family planning. In Bucharest, the developing countries both contested the view that population growth was deleterious, and asked to be paid more for going along with the industrialized country view. The socialist states and developing countries successfully insisted on bringing development, broadly and loosely defined, into the equation. The industrialized world did not disagree—John D. Rockefeller stated that ‘... the only viable course is to place population policy solidly within the context of general economic and social development in such a manner that it will be accepted at the highest levels of government and adequately supported.’

Post-Bucharest, program content did not change appreciably, concentrating primarily on family planning with some maternal and child health elements.

4.1.2 Mexico—1984. Many countries in the developing world, notwithstanding enunciated public positions, implemented population-limiting policies and programs in the largest growth period in family planning program activity, from 1974 to 1994. The last region to adopt population policies was Africa, following the landmark African regional conference in 1984, part of the lead-up to the Mexico City Conference. The 1984 world, except for the Vatican and allies, was reasonably united behind these programs, and was demanding more support for them. The US administration was by that time in full retreat from

former enthusiasm. 'Population growth is neither positive nor negative; it is a neutral phenomenon' became the leitmotif of the American administration position. Despite this policy level coolness, US budgetary and program support for continued growth in these programs were maintained.

4.1.3 Cairo—1994: a near universal consensus on a 20-year program of action. The preparatory sessions for Cairo were extremely contentious with severe criticisms of the existing approach to population programs coming from both conservative forces and from the new strong feminist voices. The pre-conference work was arduous, and the result was a complex, far reaching program of action that scarcely touched demographic forecasts and forces, but acknowledged at great length the complex context within which decisions about childbearing are made. The conference and action plan focused on health services, education, economic development, and livelihoods for women.

There was general acceptance that family planning must be voluntary and available, something never before achieved in conference documentation. The programs were to be centered on the needs of the individual and provided in a context which would take the client's entire reproductive health into account, particularly exposure to sexually transmitted diseases and AIDS. Targeted development interventions such as female education, credit for women, gender equity in inheritance and land ownership, infant and maternal mortality reduction, became explicit components of population policy.

Ninety percent of the 180 governments representing about 95 percent of the world's population approved the action plan in totality; a few conservative countries and the Vatican approved it in part. A price tag of \$17 billion was established for achieving the program.

There is fairly general agreement that the Cairo Conference made an appreciable difference to the field. Five years after Cairo, progress under the Action Plan was reviewed. Although the same issues provoked the same dissent from many of the same sources, there was consensus that considerable progress has been made in policy, program redesign, increased partnership, and collaboration directed toward implementation of the action plan. International aid fell away from rather than toward the \$7.5 billion target collectively undertaken in Cairo, but many developing countries were found to have made policy legislative and/or institutional changes in the area of population and development and reproductive health and rights. Improved transparency in governance, expanded activity of voluntary associations, improvements in communications, and legal and policy modifications were found to have advanced the prospects for the participatory approach seen to be at the center of successful implementation of reproductive health and

rights programs. Much of course remained to be done: enormous problems of maternal mortality, violence against women, and unmet need for contraceptive services all persisted, and at the end of century, two-thirds of the world's illiterates were girls and women.

5. *The Future of International Population Policy*

Is international attention to population growth finished? One of the results of the Cairo Conference was a sharp refocusing of debate away from demographic concerns toward reproductive health issues. Analysis has continued but with a sharp diminution on discussion of several former key elements. In terms of program content, many now realized to be most important in addressing high levels of fertility, especially girls education, are being addressed for their own importance. As the world continues to urbanize, and as the situation of women improves in many areas, fertility levels will continue to decline, and with them, demographic growth. It is indeed possible that countries and groups will begin to discuss policy measures that have succeeded in *raising* family fertility intentions.

Judging from declining flows of development assistance, and of international attention, there will probably be relatively less attention paid to assisting with service provision. Some continuing attention will be given to the high levels of maternal mortality. Violence against women will continue to be addressed both nationally and internationally. The same levels of international assistance and interest do not, however, seem to be available in the service of reproductive rights, as could be mobilized to slow population growth, before it became politically incorrect to speak of doing so.

This is the case despite very high levels of measurable unmet need for contraceptive services and even though an additional two billion will be added to the global population before stability is reached, with most of that two billion added in areas where food and water are critical now. These pressing global issues will be addressed primarily by applying resources to freshwater availability, food production and availability, etc., and rather less to addressing the numbers of people in need of them.

There are three main issues that will likely form the core of international population policy in the future: reproductive health issues touching infectious disease, and the linked issues of migration and aging.

Reproductive health issues will occupy an expanding place in the global agenda, particularly those related to the spread of AIDS and other sexually transmitted diseases. The realization of the security threat posed by these diseases began to accelerate with the turn of the century.

Migration for economic improvement and to escape unfavorable political and social situations around the world will continue to be a growing population issue.

A different set of international institutions and government departments from those concerned with fertility are the central actors on this issue.

While the dramatic nature of fertility increase in the twentieth century is widely known, there is considerably less appreciation of the fact that the equally dramatic decline in fertility has radically changed the number of people within age groups in many countries. Aging may well be the dominant policy issue of the twenty-first century. The spread of aging populations is uneven. While there are more Europeans moving into the over-65 year old category than the under 16 year old group, at the turn of the century there was still a 26:1 ratio in the other direction in the totality of the developing world. In developing countries that initiated rapid declines of fertility in the 1960s, such as Taiwan, Korea, China, Thailand, Mexico, and Brazil, issues connected with aging are now rising on the agenda. This issue will internationalize primarily because of the impact on migration created by declining labor force size, and the worries that this creates vis-à-vis social security pressures and health system financing.

See also: Corporate Governance; Fertility Change; Quantum and Tempo; Fertility Control: Overview; Food Security; Greening of Technology and Eco-technology; Population Ecology; Population Pressure, Resources, and the Environment: Industrialized World

Bibliography

- Bongaarts J 1994 Population policy options in the developing world. *Science* **263**: 771–76
- Cohen J E 1995 *How Many People can the Earth Support?* W W Norton, New York
- Brown L R 1995 *Who Will Feed China: Wake-up Call for a Small Planet*. W W Norton, New York
- Erlich P R 1968 *The Population Bomb*. Ballantine, New York
- Gulhati K, Bates L M 1994 Developing countries and the international population debate: Politics and pragmatism. In: Cassens (ed.) *Population and Development: Old Debates, New Conclusions*. Overseas Development Council, Washington, DC, and Transaction Publishers, New Brunswick, NJ
- Jain A (ed.) 1998 *Do Population Policies Matter? (Fertility and Politics in Egypt, India, Kenya and Mexico)*. Population Council, New York
- Jejeebhoy S J 1995 *Women's Education, Autonomy, and Reproductive Behaviour*. Clarendon Press, Oxford, UK
- Malthus T R 1798 *An Essay on the Principle of Population*. J. Johnson, London
- Myers N 1995 *Environmental Exodus: An Emergent Crisis in the Global Arena*. The Climate Institute, Washington, DC
- Sen G, Germain A, Chen L C (eds.) 1994 *Population Policies Reconsidered: Health, Empowerment, and Rights*. Harvard University Press, Boston
- Simon J L 1977 *The Economics of Population Growth*. Princeton University Press, Princeton, NJ

Sinding S W, Ross J A, Rosenfield A 1994 Seeking common ground: Unmet need and demographic goals. *International Family Planning Perspectives* **20**: 23–7, 32. Reprinted in Mazur LA (ed.) 1994 *Beyond the Numbers: A Reader on Population, Consumption and the Environment*. Island Press, Washington, DC

United Nations Flows of Financial Resources in International Assistance for Population

M. Catley-Carson

Population Pressure, Resources, and the Environment: Industrialized World

1. Introduction

Few issues have attracted as much ink, and as much intellectual passion, as have debates about population, resources, and the environment. It is important to understand that many discussions of 'environment' have foundered due to at least four different uses of the term:

(a) As in 'the global environment' or the 'global ecosystem.' Explanations at this high level of abstraction often are so general as to defy contradiction or verification.

(b) As a shorthand reference for all natural resources, especially all sources of raw materials, energy, and food. Less general than the first use, but still combines very different settings and industries under one rubric.

(c) To connote the purity of air and water, especially as affected by urbanization and industrialization and their impacts on rural or 'natural' landscapes.

(d) As an antonym of 'heredity.' This is a quite different and often confusing meaning in which 'environmental' effects are those determined by social rather than biological processes, along with interaction between the two.

The disputes, which go back at least two centuries, touch upon theories and evidence central to much of modern intellectual history. Moreover, the participants in the debates have included many of the giants of Western scholarly and political thought, from Smith, Ricardo, and Malthus, to Marx and Engels, to Keynes and the Myrdals.

2. Two Fundamental Phases of Western Political Thought, 1750–1940

We can identify two fundamental phases of the developing debate on population, resources, and environment in the West. The first extends from 1750–

1880, the central period of industrialization in Western Europe and North America. The second covers the period 1880–1940, a period when industrial growth rates in Western Europe first peaked and then plateaued, Western imperial power peaked and then declined, and Western political instability and collective violence reached unprecedented levels. These two phases of ascending European political and economic power, followed by its slow decline, are reflected in the attendant debates about the balance among demographic growth, resources, and the environment.

During the expansionary phase, optimists from Godwin to Condorcet to Marx foresaw essentially unlimited expansion of material resources to support a growing population. What had to be changed was not the resource base, but rather the power of landowner and/or capitalist classes to place their demands above the needs of the majority. While Marx and his followers demanded revolutionary transformation of urban, industrial society, utopians sought to escape from its corruptions and create a new kind of society remote both physically and morally from the old. With the 20/20 vision of hindsight, we now can say that these utopian and Marxian views have not proven to be sustainable in the real world.

Meanwhile, pessimists of the period such as Malthus and his followers saw utopians as innocents unaware of the impossibility of human perfectibility. Malthus held that the human sexual instinct was a constant that would produce population growth ultimately leading to disaster. Idealistic efforts to reduce poverty (via ‘transfer payments’ in modern terminology) would prove self-defeating, he argued, because higher incomes would in turn lead to higher fertility and lower infant and child mortality. Hence, policies intended to relieve poverty would worsen the environment by putting additional pressure on the food supply through encouraging demographic increase. For Malthus, the only morally acceptable escape was by individuals exercising ‘prudential restraint’ of their fertility, by which he meant avoidance of marriage unless and until the couple had sufficient resources to support a family, accompanied by abstinence outside of marriage. Failing this, balance between population and resources would be imposed via the dreadful ‘positive checks’ of famine, plague, and war.

In similar hindsight we can say that Malthus’s forecasts were mistaken too. He underestimated how fast food production and economic activity would increase, and dismissed as infeasible the conscious control of fertility within marriage that later came to be the norm in urban/industrial societies. His provocative arguments did, however, stimulate a passionate debate on population and the environment. In particular, his views were assaulted vigorously by Marx, thereby establishing a destructive ‘right-left’ dimension to more than a century of population/resource/environment debates that ensued.

Other conservative writers of the nineteenth century focused less on population issues and more on the ‘destruction’ of the ‘natural environment’ produced by capitalist industrial development and urbanization. In response to such claims, Marx and his followers acknowledged the validity of their criticisms, but argued that these developments were a function of capitalism, and hence transitional.

The second phase of the population/environment debate began toward the end of the nineteenth century, when the use of ‘environmental’ as an antonym to ‘genetic’ emerged in discussions of human decadence and perfectibility. Such arguments derived from enthusiasm about Darwinian arguments, though many of their ‘eugenicist’ proponents were far from well informed of the scientific evidence involved.

In general, they argued that the human gene pool was deteriorating, with powerful detrimental impacts on society and its betterment. The urban environment was itself genetically damaging, producing a cycle of residents who bore large numbers of stunted offspring who in turn raised large families of equally unhealthy children. Meanwhile, social elites were controlling their own family size, resulting in a collective ‘enfeeblement.’ For these proponents, this posed profound dangers to state and empire by reducing military and international power, while the increasing fertility differentials between social classes foretold the prosperous being swamped by a flood of urban disability. What was essential, they argued, was intervention to reverse the rot.

While many eugenicists were politically conservative, others were liberals or socialists. Though relatively few were fascists, the eugenics movement ultimately was compromised fatally when its notions of biological decadence were co-opted by the Nazis.

3. Complex Phenomena, Simpler Analyses

Any intersection among three forces as multifaceted as population, resources, and the environment—each of which is itself a subject of convoluted debate—is bound to be one of challenging complexity, required balanced and nuanced analyses. Yet, unhappily, there is an enduring tendency for arguments that anoint one of the three forces as dominant. Proponents on all sides tend toward single-cause explanations of complex social processes. Some emphasize the biological, others the economic, still others the ecological. Each group embraces one unifying theme to the exclusion of insights from other perspectives, and produce pronounced polemics of denunciation and retort.

Malthus’s celebrated essay on the principle of population was itself such a response to earlier arguments from utopians (including his father). It in turn provoked passionate denunciations by utopian socialists and Marxists, some of whom argued that there were no natural limits on population growth.

Similarly, eugenicists of the early twentieth century exaggerated the importance of biological and genetic elements for social issues such as fertility, poverty, and illiteracy, and such arguments in turn led social scientists to exclude biological elements entirely from their discussion of the 'demographic transition.'

In short, there has been a mentality of the adversary underlying much of the discussion about population, resources, and the environment, redolent of the style of political, religious, and legal advocates who ignore the similarities of their approaches to those they condemn.

Another source of dissonance during the nineteenth and twentieth centuries was the tendency to search for or assert 'scientific laws' of population, resources, and/or environment. Marx and Engels claimed the mantle of science for their analyses of capitalism while attacking utopians seeing the world not as it was but as they wished it to be. Malthus too criticized utopians such as Godwin and Condorcet in the same way. The fact that Marx's arguments did not explain social conditions within his own lifetime, and that Malthus's distinction between geometric and arithmetic progressions of population and food was contradicted by trends underway as he wrote, failed to dissuade either from his claims of scientific validity.

Similarly, the near-appropriation of Darwinian theory by political proponents of social Darwinism, and later of then-modern genetics by those promoting eugenics as social policies, illustrates the deep mystique of science in political arguments of this period. Invoking 'scientific' evidence of biological limits to human progress, eugenicists derided moderate and revolutionary reformers' assurances that environmental or political change could ameliorate the lot of the poor. Such excessive arguments from eugenics then led social scientists from the 1930s onward to formulate a theory of the demographic transition that excluded biology.

Meanwhile, in the Soviet Union Lysenko and his associates developed a Soviet 'scientific' response to eugenics that asserted that environmental factors trumped genetic forces, and more specifically that hereditary characteristics could be acquired rapidly from the environment and subsequently transmitted reliably to later generations. This 'scientific' perspective offered Soviet officials the opportunity to use 'environmental' (i.e., social, economic, political) policies to manipulate the gene pool—an opportunity that produced devastating famines as a consequence.

4. Economists and Demographers

Economists and demographers of the twentieth century also tended to unduly limit the horizons of their analyses and data. Neo-classical economics flowered at a time when economic growth was so rapid that it was easy to ignore resource and environmental issues.

Natural resources must have appeared unlimited to those at the center of expanding global empires, while the bulk transport allowed by steam power, railroads, and refrigeration rendered moot early concerns about limited agricultural land.

The emerging field of demography had multidisciplinary roots in economics, biology, mathematics, statistics, and epidemiology. Yet leaders of the nascent field were offended by the excesses of some biologists and others who sought to explain Western fertility declines in exclusively biological terms, and by the conservative political views that increasingly were associated with such explanations. The emergence of fascism exacerbated their concerns, leading them to relocate their 1936 scientific meeting from Rome to London and to boycott another meeting in Berlin.

At this moment in the intellectual history of demography a social and economic explanation of Western fertility coalesced into the 'theory of the demographic transition.' Though not really a theory in the sense of offering testable predictive hypotheses, the demographic transition provided a grand historical generalization that fertility declines could be wholly attributed to economic and social forces of the industrial revolution. In so doing, it ignored biological explanations then in vogue of the same phenomenon, as well as elements of culture, language, and religion unless these were driven by socioeconomic change (e.g., the effect of urban and industrial development in weakening the hold of the extended family on fertility behavior.)

The thinking of both economists and demographers was affected deeply by the Great Depression of the 1930s. To some degree they interpreted it as a consequence of 'underutilization,' both of capital and human resources in the economy and of the ability to bear children. Gunnar and Alva Myrdal in Sweden and Enid Charles in England produced gloomy, sometimes apocalyptic writings anticipating the terrible consequences of continuing low fertility. The Myrdals' prescription for dealing with 'the population crisis' was social legislation that evolved into the Swedish welfare state.

World War II focused concerns in different directions, but the war was followed by a long period of buoyant economic growth accompanied by much higher fertility levels. Expectations of unlimited economic growth replaced the gloom of the 1930s. Fertility trends ranged from powerful and sustained baby booms in the United States, Canada, Australia, and New Zealand to smaller and shorter-term 'boomlets' in most of Western Europe.

By the early 1960s, concerns about the negative environmental consequences of economic growth began, accompanied by worries about accelerating demographic increase both due to the Western baby booms and to rapid mortality declines with sustained high fertility rates in formerly colonial developing countries. Both concerns contributed to growing

awareness and activism about environmental problems, enhanced further during the 1970s by the oil crisis, the accident at the Three Mile Island nuclear plant, and the appearance of economic 'stagflation' that had no explanation in the canon of Keynesian economics. These events suggested there might, after all, be limits to natural resources, technological advance, and economic prosperity. In the United States and Western Europe, a new kind of political environmentalism emerged, including a direct electoral form in the German 'Green' party.

During the 1980s, and partially in reaction to environmental politics, there emerged in the United States an odd amalgam of earlier utopian, Marxian, and *laissez-faire* thought, deploying apparently scientific data and economic analysis, in what came to be called libertarianism or 'cornucopianism.' In its most enthusiastic variants, cornucopians argue that there are no natural limits to economic and population growth, and that the availability of natural resources is rising rather than declining. Some asserted that the 'ultimate resource' is human intellectual power to innovate technologically and create new resources, and hence the more humans on earth, the more resources there would be, *ad infinitum*.

Many of those who embraced such arguments did not seem to appreciate their intellectual origins, especially their Marxian pedigree. They were promoted strongly by opinion organs of American new right groups (the Heritage Foundation, Cato Institute, and the editorial board of the *Wall Street Journal*).

5. *Environmental and Ecological Approaches*

While many biologists, economists, demographers, and politicians have tended toward simple arguments about population, resource, and environmental issues, environmental scientists have emphasized the place of the human species in a complex ecological web of biological systems in constant interaction with the physical environment. In so doing, they have pointed appropriately to its vulnerability to the external insults of human technologies. However, there has also been a tendency to minimize or ignore the potentially positive environmental impacts that can also be produced by human technologies.

Consider for example the 'epidemiological transition' (see *Mortality, Epidemiological, and Health Transitions*). The triumph of human knowledge over food shortages and many endemic infectious diseases has produced dramatic improvement in human mortality since the eighteenth century. This was, of course, the primary reason for the acceleration of rapid human demographic growth since 1750, but it may also be seen as a fundamental technological transformation of the ecological systems involving humans, their food supply, and their 'predators.' Though environmental writers generally do not describe these particular

technological achievements in negative terms, some prominent writers have expressed skeptical views of their long-term implications for the global ecosystem.

In general, environmental and ecological writers frequently have emphasized the negative effects of other elements of science and technology, such as those surrounding energy consumption and production and modern agriculture, and paid rather less attention to the potential of new technologies for expanding natural resources or ameliorating environmental damage from other human activities. Of course sophisticated ecologists take such potentials fully into account, but there remains among some elements of the environmental movement a tendency to de-emphasize the potential environmental positives of technological innovation—in some sense a polar extreme from the technological enthusiasms of some economists. To many ecologists, the technological optimism of much modern economics, and of both Marxists and the new right, ignores the laws and complexities of nature.

To some degree we have come full circle. When environmentalists denounce the rape of the rain forest or the despoliation of freshwater streams and lakes, they echo older and eloquent laments such as those of D. H. Lawrence at the ruination of the English countryside or Dickens's portrayal of the industrial wasteland in *Hard Times*. Concerns about the environment, like those about population and resources, have deep intellectual roots in Western intellectual traditions.

6. *Conclusion*

Given this rich but often contentious intellectual history, what might be expected of future discussions of population, resources, and environment? During the nineteenth century, most of the major political thinkers were forced by the powerful changes all around them to analyze population change and its interaction with resources and the environment. Typically they sought single (or at least simple) explanations of the complex transformations they could see, leading inevitably to an intensely ideological debate. Instead of building upon the work of earlier writers, they sought to refute their adversaries' errors. First the utopians, Malthus, Marx, and their allies clashed over the potential for demographic and economic growth with prosperity. Later biologists, eugenicists, and social scientists contested about the biological vs. environmental causes of declining fertility rates and increasingly 'enfeebled' urban populations.

These classical debates about population/resources/environment are a polemicists' dream: sweeping generalizations, exaggerated rhetoric, and emotional accusations. Those in the Marxian, socialist, and liberal-Christian traditions were certain

that when the majority took power from the reactionary elite, they would build a more humane environment unconstrained by demographic issues. Those following Malthus, and later the eugenicists, argued to the contrary that human progress is constrained by the laws of biology and demography.

The political and economic crises following World War I tended to mute these earlier debates about population growth, resource scarcity, and environmental degradation. After World War II, it was relief at the end of global war and the reappearance of prosperity that further eclipsed environmental and demographic questions. Only since the 1960s have these issues re-emerged, after decades of inattention.

There is good reason to fear that future debates may echo those of the past: ideological blinders, harsh rhetoric, narrow analytic perspectives. But there are also signs of hope. Resource economists have shown that it is possible to combine the power of economic theory and analysis with those of ecology and environmental perspectives. Demographers too have broadened their dependency upon purely socio-economic analyses to embrace insights from biology, ecology, political science, and other relevant disciplines. The hope is that these more ecumenical trends will persist and flower, so that our collective understanding of population, resources, and the environment may someday be as subtle and complex as the phenomena themselves.

See also: Development: Sustainable Agriculture; Development Theory in Geography; Ecological Economics; Environmental and Resource Management; Environmental Challenges in Organizations; Environmental Economics; Environmental Vulnerability; Environmentalism: Preservation and Conservation; Population Policies, Politics of; Resource Geography; Resource Institutions; Sustainable Development

Bibliography

- Daly H E 1996 *Beyond Growth: The Economics of Sustainable Development*. Beacon Press, Boston
- Homer-Dixon T F 1999 *Environment, Scarcity, and Violence*. Princeton University Press, Princeton, NJ
- National Research Council, Board on Sustainable Development 2000 *Our Common Journey: A Transition Toward Sustainability*. National Academy Press, Washington, DC
- Ness G D, Drake W D, Brechin S R (eds.) 1993 *Population–Environment Dynamics: Ideas and Observations*. University of Michigan Press, Ann Arbor, MI
- Polunin N (ed.) 1998 *Population and Global Security*. Cambridge University Press, Cambridge, UK
- Ramphal S S, Sinding S W (eds.) 1996 *Population Growth and Environmental Issues*. Praeger, Westport, CN

M. S. Teitelbaum

Population Size Estimation and Capture–Recapture Methods

How many people were missed by the US census? How many injecting drug users are there in Glasgow? Homeless in Scotland? Industrial accidents in Maryland? Bugs in a computer program? How many words did Shakespeare know?

How do we estimate the sizes of such populations? If the population includes a subgroup of identifiable individuals and if their number is known, a simple approach is available. If a random sample can be taken from the whole population and the proportion of the subgroup that occur in the sample is calculated, that proportion can be used to scale up the sample size to provide an estimate of the unknown size of the population. Three important ‘ifs’ occur in the last two sentences that cause problems for this dual-system estimation (DSE) procedure, also known as capture–(or mark–) recapture (CR) from its widespread use in the study of wildlife populations by means of animals caught, given marks, released and subsequently recaptured.

Both in this simplest form, and in the many extensions proposed to overcome the problems posed by these ‘ifs’ in the real world, this CR method has many relationships with other statistical methods. The problem is one of sampling from a finite population. If an indicator variable y is defined to be 1 for an individual in that population, 0 otherwise, the unknown size of that unique population is the population total Y of y . The CR estimate is the classical ratio estimate based on the auxiliary indicator variable x which is 1 for an individual in the subgroup, 0 otherwise. Laplace’s (1786) estimation of the population of France, with subgroup those registered as born in the previous three years and a cluster sample of over two million, can be regarded as the first use of CR as well as the first use of a superpopulation model (Pearson 1928, Cochran 1978). Two books discuss CR as sampling theory—by Thompson (1992) and Skalski and Robson (1992).

1. Statistical Theory

Early discussion, summarized in e.g. Seber (1982) or Cormack (1968), focused on distributions for the observations—hypergeometric, multinomial etc.—under different rules for defining the subgroup and stopping the sampling. Practical differences are minor. The distribution of the estimate, especially for small subgroups or samples, is more important. If the population size is N , the size of the known subgroup X , and the sample size n contains x from the subgroup, then the intuitive, also the maximum likelihood,

estimate of N is $[nX/x]$, known as the Petersen estimate after the Danish fisheries biologist who first proposed its use in biology. This is the observed number of different individuals $[X+n-x]$ plus the estimate $[(X-x)(n-x)/x]$ of the number U of unobserved individuals. As a ratio estimate this is biased: the recommended correction (Chapman 1951) adds 1 to the denominator of the estimate of U . The distribution of the estimate of U is skew, and if U is small, or based on small observed numbers, then a confidence interval based on assumed normality can cause embarrassment by including negative values for U . Likelihood can be used directly (Cormack 1992) to avoid this. Other approaches have been proposed, with different loss functions and/or Bayesian methods. The latter are discussed below.

The known subgroup often comprises individuals identifiable because they occur in another list or sample. In wildlife applications, in which the next developments took place, the two samples were often obtained by the same protocol at different occasions, with a long enough period between for thorough mixing of the mobile animals to be (questionably) assumed. Generalizations of this two-sample case were made, to more samples (Darroch 1958) and, as time periods extended, to allow for the population to change through birth, death, or migration between samples. In open populations, survival estimation from recapture data on marked animals is more reliable than estimation of population size. Open population models were applied to historical lists by James and Price (1976): the willingness of residents of a medieval town to contribute to the town's defense, and their ability to avoid paying a tax to the king, are clearly displayed, together with evidence of unsuspected epidemics.

The natural statistical way of representing counts in a number of categories, formed by the presence in or absence from a number s of lists, is as the cells in a 2^s contingency table. Models and analytical techniques are summarized in books such as Bishop et al. (1975). In CR one cell is not observed, a structural zero, and the aim is, by formulating a series of models for the observed categories in the table, to estimate the missing number by extension of these models. When additional information on individuals is available in the form of quantitative covariates, such as age or severity of disease, the use of logistic models at the individual level was suggested independently by Alho (1990) and by Huggins (1989). The Horvitz–Thompson estimator is used, with inclusion probabilities, of being in each list, estimated from the model, a procedure developed differently by Pollock and Otto (1983). With no covariate information, or with categorical information only—e.g., gender, ethnicity, geographical region—log linear models form the current standard approach. Maximum likelihood estimates (MLE), perhaps bias corrected, are then derived. In this framework it is natural to examine the data for interactions between

categories, evidence that the samples were not random. Birth and death, unequal availability of different individuals, and behavioral change consequent on physical capture of animals, have all been expressed as interactions within the log linear framework.

Even thus far, numerous quirks arise whereby the corresponding statistical theory does not quite apply to the CR problem. Estimation from a log linear model does not maximize the complete likelihood: nuisance parameters of sampling probabilities are estimated from the observed cells, and N estimated conditionally on them. Normally such multinomial models for complete tables are fitted assuming a Poisson model, giving the estimates various desirable properties of the exponential family. With N unknown, the distributions of the estimate of N are different under multinomial and Poisson models though the point estimate is the same and estimates of nuisance parameters have the same properties. Regularity conditions for desirable properties of MLE do not hold in the multinomial model, but do in the Poisson. The Horvitz–Thompson correspondence shows that the estimate of N is calculated via estimates of $1/(\text{inclusion probabilities})$. Why then seek unbiased estimates of these probabilities? It is their inverses that are of interest.

2. Health and Social Studies

One widespread application in the social sciences is to ascertain how many individuals have some health or social problem, numbers needed for policy and budgetary decisions. Data are pre-existing lists from different agencies, each attempting to cover the whole population. Such lists have no or little random element in their formation. They will not, in general, be statistically independent. Individuals are referred from one agency to another. Conversely, two agencies may overlap, or be perceived to overlap, in their function so that an individual attending one may choose not to attend the other. Agencies primarily serving different sections of the population, geographically, by age, severity of problem, or whatever, will also be negatively dependent. Such correlation bias results in the standard estimate being an underestimate if lists are positively dependent, an overestimate if negatively dependent.

With two lists dependence is not identifiable: independence must be taken on trust. With more than two, log linear models can be used to elucidate the pattern of dependence. With three lists, eight models with different combinations of pairwise dependences are available. Extrapolation to the missing cell to estimate U and thus N requires the three-factor interaction to be zero or else expressible, from a model at the individual level, as a function of lower-order interactions. Most studies use some form of model selection to determine which dependencies must be

allowed for, usually based on log likelihood ratio statistics from the analysis of deviance—the residual deviance and the difference between deviances from nested models. The complexity of a model is balanced against its badness of fit. Many authors have advocated parsimony: others argue that a parsimonious model with independence between lists is difficult to believe *a priori*. With the latter attitude, it might seem best to always fit the saturated model, including all identifiable interactions. There are two objections to this. The first is that if all lower-order interactions are nonzero, considerable doubt is cast on the necessary act of faith that the highest-order interaction is exactly zero. The other is that the estimate is very sensitive to any small cell count, and gives either a zero or infinite estimate for the missing cell if any cell count is zero, an event which becomes more likely the more lists are included in the study.

Certain patterns of cells with small numbers cause problems for all analyses, in that they reveal that the data contain very little information about some dependencies, either having few overlaps between lists or, perhaps surprisingly, when one list has a very high coverage. Asymptotic distributions of criteria used for model selection no longer hold, and should be replaced by bootstrap resampling of data predicted by different models.

More than three lists can be analyzed similarly except that, with more than five or six, the small number problem usually dominates and routine application of a program to analyze log linear models becomes impossible. One approach to model selection is then to start with the model including all pairwise dependencies, perhaps use backward elimination to simplify, and add higher-order terms by forward selection. However, making inferences conditional on the truth of a model selected from the data has been recognized as the Achilles heel of statistical inference. The alternative of averaging over all models has been applied to CR by several authors (see Sect. 5).

In the study design a list should be sought which is likely to be independent of the others. In studies on drug abuse, lists from police arrests have sometimes appeared to fulfil this condition relative to lists from social and health agencies. Questions of commonality of definition do arise. Moreover, patterns of interactions have to allow for individual heterogeneity as well as direct dependence between lists. The resulting confounding is discussed below.

3. Heterogeneity

Animals, particularly humans, are not the exchangeable colored balls required by ideal mathematical theory. There are reasons why individual i is not observed in list j . To estimate the numbers missing from all lists requires exploration of these reasons.

Exploration and thought are crucial to give even a qualitative impression of the effect of such individual heterogeneity on the estimate. To quantify it needs modeling of p^{ij} , the probability that individual i is observed in list j . An early attempt by Cormack (1966) considered $p^{ij} = \theta_i * \beta_j$, but this is superseded by the logit model: $\log \{p^{ij}/(1-p^{ij})\} = \theta_i + \beta_j$ introduced by Rasch (1960) for educational test scores. In the context of repeated samples with the same protocol and same effort, so that β_j may be assumed constant, Burnham and Overton (1978) introduced the idea of the individual effects θ_i being a random sample from some distribution. Some workers have developed the consequences of assuming a beta distribution, but, if no specific form is assumed, the set of capture frequencies—numbers of individuals seen k times—are jointly sufficient statistics for N after integration over the distribution.

Without θ_i these models represent independent lists: differences between individuals necessarily cause dependence between lists. The Rasch model for individual behavior corresponds to a log linear model for cell counts with quasisymmetry, all interaction terms of the same order having the same value (Fienberg et al. 1999). Assuming a random model for an individual contribution's to his probability of being in a list, and integrating over that distribution, creates a superpopulation model for a sample survey. With nonrandom lists, any inference must be model-based, and estimates from superpopulation models are standard in survey sampling. However, while it is easy to accept that an estimate of, say, mean income in a superpopulation gives a reasonable inference for the one finite population of interest, doubts surface when the parameter of interest is the size of that one finite population. Unfortunately, allowing direct dependence between lists in a Rasch model for an individual is not the same as a log linear model with only that one extra dependence term. CR has also been expressed as a latent class problem by Agresti (1994).

Heterogeneity can be reduced by analyzing subpopulations as separate strata. Differences in models selected and in estimated inclusion probabilities can be studied. Samples from some strata may however be too small to permit effective model selection.

4. Coverage

Closely related to CR is the classical species problem. How many are there in the world? As a specific example, how many words did Shakespeare know? Concordances list all occurrences of any word in each play. Some words are used in only one play. There must be many he knew, but never used, just as there are many undiscovered species of beetle. Some words occur much more frequently than others, stark evidence of heterogeneity between individuals (words).

Apart from elegant models for specific applications, debate about sampling distributions and behavior of estimates has paralleled, and often predated, that about CR. A good review has been provided by Bunge and Fitzpatrick (1993). A key concept is that of coverage, the proportion of individuals of species represented in existing lists, perhaps easier to estimate than N , the unknown number of species. Chao and colleagues (e.g., Chao and Lee 1992) have developed various nonparametric estimates of N via coverage. These can be linked to martingale approaches to CR by Yip (1991).

5. Bayesian Methods

Bayesian methods have a long history in mark-recapture, even if Laplace's inverse probability argument is not considered pure Bayes. Early formulations make the point that independence of population size N and capture probabilities may not be appropriate *a priori*. Not only prior beliefs about N but also model uncertainty can be built smoothly into the analysis. Madigan and York (1997) develop models, with some analytical tractability, for decomposable graphical models with hyper-Dirichlet priors for cell probabilities, respecting the model's pattern of conditional independence, and allowing informative priors for N . Bayesian analysis of hierarchical log linear models has also been developed, an approach linked to the Rasch model by Fienberg et al. (1999).

6. Census Undercount

Recent national censuses show nontrivial undercounts. The possibility of adjustment has been widely explored and debated, in the US courts as well as the scientific and social literature. Because of the scale and expense, there are at least four major differences from applications discussed above. One is that matching and other nonsampling errors, present in other applications, gain prominence. Another is that the second list, from follow-up field survey, is necessarily performed only on a sample of the population. A third is that only two lists are the norm. The fourth is that estimates are required not only for the population as a whole, but for particular minority groups in relatively small geographical areas. Statistical summaries of the US debate have appeared in special sections or issues of the *Journal of the American Statistical Association* (1993) and of *Statistical Science* (1994), and in Anderson and Fienberg (1999).

Matching errors tend to identify two records of the same individual as different rather than vice versa. With two lists this leads to overestimation of N . Attempts to correct this, by algorithms which declare individuals identical, even when differing in some

identifiers, can give false positive matches, and hence underestimation. Probabilistic record linkage is discussed by Jaro (1995) and models for its effect on CR estimates by Ding and Fienberg (1996).

In recent usage, the second list comes from a postenumeration survey (PES), an intensive field survey of small geographical areas, defined slightly differently in different countries, in a complex randomized sample plan, typically with strata and clusters at two or more levels. The relationship between CR and ratio estimators allows CR models to be combined formally with the sample design, though sampling theory suggests other ways of using auxiliary variables.

With two lists, the required independence must be built into the design. Sampling units must be defined, small enough, to minimize heterogeneity. Small numbers cause instability, and attempts are made to correct this by spatial smoothing over neighboring areas. The multiplicity of ways in which sample-based inference can be integrated with model-based inference from multilevel data with CR models is a topic of much current debate.

7. Summary

The first difficulty with the use of CR is the definition of the population. If in a study of diabetes all observed individuals can be stratified by the medical care prescribed, then the missing numbers are not of undiagnosed diabetics, but of diabetics who have been diagnosed, but whose records have been lost. If different agencies have different definitions of homelessness, or cover different geographical areas, then which population is it hoped to estimate? No statistical analysis of the observed data resolves these problems. Second, there is the problem of matching individuals, a problem which grows with the size of the study, and the separation of the analyst from the data collection. The more sensitive the subject of the study, the more prevalent will be deliberate misinformation from an individual. Matching problems will vary between countries that have differing levels of confidentiality and anonymity. Third, there are the analytic problems caused by the unreality, in many cases, of the assumption of random sampling. Lists are not independent, individuals are not behavioral clones, and the two aspects of dependence and heterogeneity may be at least partially confounded unless care is taken to obtain data which allows disentanglement. Finally, there is the problem of the variety of analytical approaches—which to believe? With several lists, some models can be rejected because of lack of fit to the observed counts. However, failure to reject a model is not positive proof of its truth. Estimation of population size still requires extrapolation to the unobserved, the validity of which is untestable without randomness in the sampling.

Although absolute estimates of population size are often unjustifiable, it may be possible, by repeating the same protocol for list-formation at different times or places, to obtain reliable assessments of population change or geographical variation in incidence of a disease or social phenomenon.

See also: Censuses: History and Methods

Bibliography

- Agresti A 1994 Simple capture-recapture models permitting unequal catchability and variable sampling effort. *Biometrics* **50**: 494–500
- Alho J M 1990 Logistic regression in capture-recapture models. *Biometrics* **46**: 623–5
- Anderson M J, Fienberg S E 1999 *Who Counts? The Politics of Census-taking in Contemporary America*. Russell Sage Foundation, New York
- Bishop Y M M, Fienberg S E, Holland P W 1975 *Discrete Multivariate Analysis*. MIT Press, Cambridge, MA
- Bunge J, Fitzpatrick M 1993 Estimating the number of species: a review. *Journal of the American Statistical Association* **88**: 364–73
- Burnham K P, Overton W S 1978 Estimation of the size of a closed population when capture probabilities vary among animals. *Biometrika* **65**: 625–33
- Chao A, Lee S-M 1992 Estimating the number of classes via sample coverage. *Journal of the American Statistical Association* **87**: 210–17
- Chapman D G 1951 Some properties of the hypergeometric distribution with applications to zoological censuses. *University of California Publications in Statistics* **1**: 131–60
- Cochran W G 1978 Laplace's ratio estimator. In: David H A (ed.) *Contributions to Survey Sampling and Applied Statistics*. Academic Press, New York, pp. 3–10
- Cormack R M 1966 A test of equal catchability. *Biometrics* **22**: 330–42
- Cormack R M 1968 The statistics of capture-recapture methods. *Annual Review of Oceanography and Marine Biology* **6**: 455–506
- Cormack R M 1992 Interval estimation for mark-recapture studies of closed populations. *Biometrics* **48**: 567–76
- Darroch J N 1958 The multiple-recapture census I: estimation of a closed population. *Biometrika* **45**: 343–59
- Ding Y, Fienberg S E 1996 Multiple sample estimation of population and census undercount in the presence of matching errors. *Survey Methodology* **22**: 55–64
- Fienberg S E, Johnson M S, Junker B W 1999 Classical multi-level and Bayesian approaches to population size estimation using multiple lists. *Journal of the Royal Statistical Society A* **162**: 383–405
- Huggins R M 1989 On the statistical analysis of capture experiments. *Biometrika* **76**: 133–40
- James T B, Price N A 1976 Measurement of the change in populations through time: capture-recapture analysis of population for St. Lawrence parish, Southampton, 1454 to 1610. *Journal of European Economic History* **5**: 719–36
- Jaro M A 1995 Probabilistic linkage of large public health data files. *Statistics in Medicine* **14**: 491–8
- Laplace P-S 1786. Sur les naissances, les mariages et les morts a Paris depuis 1771 jusqu'en 1784; et dans toute l'etendue de la France, pendant les annees 1781 et 1782. *Memoires de l'Academie des Sciences (1783)* 693–702
- Madigan D, York J C 1997 Bayesian methods for estimation of the size of a closed population. *Biometrika* **84**: 19–31
- Pearson K 1928 On a method of ascertaining limits to the actual number of marked members in a population of given size from a sample. *Biometrika* **20A**: 149–74
- Pollock K H, Otto M C 1983 Robust estimation of population size in a closed animal population from capture-recapture experiments. *Biometrics* **39**: 1035–49
- Rasch G 1960 *Probabilistic Models for Some Intelligence and Attainment Tests*. University of Chicago Press, Chicago
- Seber G A F 1982 *The Estimation of Animal Abundance and Related Parameters*. Griffin, London
- Skalski J R, Robson D S 1992 *Techniques for Wildlife Investigations: Design and Analysis of Capture Data*. Academic Press, San Diego, CA
- Thompson S K 1992 *Sampling*. Wiley-Interscience, New York
- Yip P 1991 A martingale estimating equation for a capture-recapture experiment in discrete time. *Biometrics* **47**: 1081–8

R. Cormack

Populism

Populism is a form of mass politics. The core idea, as the etymology would suggest, rests with the claim to represent or act in the name of the people, understood as ordinary or common people, the majority, or the masses, as opposed to elites, privileged or special-interest groups, the establishment, or the power bloc. It is generally distinguished from two other forms of mass politics: class politics (based on class organizations of unions and social democratic parties) and pluralism (based on multiple, shifting, overlapping interest groups and a rejection of the populist dualism that stresses the representation of the general will of the people vs. some minority). 'Populism' has been used as a term of approbation by movements themselves and as a term of opprobrium by critics. Analysts have also used the term, applying it to diverse phenomena.

1. Empirical and Conceptual Variation

With origins in ancient Rome, where the *Populares* opposed the aristocratic *Optimates*, the first modern usage is probably the Russian *Narodnik* movement of the 1860s–70s, which never achieved mass adherents but represented a radical articulation of the idea that wisdom resides with the common people and that socialism could be based on the traditional peasant commune. American populism, beginning in the 1880s

and continuing particularly in the South through the twentieth century, has received much attention, as has Latin American populism of the mid-twentieth century. Currently there is a resurgence throughout the world of movements and leaders that have been dubbed populist: in Latin America (Collor, Fujimori, Chavez, Menem), in Western Europe (Haider, Le Pen), in the United States (Buchanan), and in post-Communist states (Meciar, Milosevic, Lukashenko). Many other movements have also been considered populist by at least some analysts, including those led by Nasser, Nyerere, Mao, Hitler, Father Coughlin, Gandhi, Chartists, and even Jimmy Carter.

The populist label has been attached to such a wide variety of political movements (right and left, from above and from below) that it is hard to stabilize any core meaning that can work rigorously as an analytical concept. Some analysts propose conceptualizations that are compilations of traits of specific movements, providing no basis for abstracting out a more parsimonious conceptualization that serves for broader, comparative analysis (Pipes 1964, Goodwyn 1976, Hicks 1955). Others, engaged in a more comparative quest, have highlighted a particular trait or subset of traits (Germani 1962, Di Tella 1965, Ionescu and Gellner 1969, Laclau 1977, Conniff 1999, Taggart 1995).

These approaches typically emphasize some components of the following.

(a) *Mobilization* or collective action from below, with deep roots in local communities.

(b) A *reform* or anti-status quo movement.

(c) A *support base* constructed by a discourse that attempts to promote an identity as the 'people' rather than a class, and sometimes aimed at neutralizing class identities. The political group constructed in this way is embedded in a particular adversarial juxtaposition to a vague 'other': elite (economic and/or political), alien, special interest, or minority group, often corrupt or conspiratorial. At the same time that a mass discourse seeks to construct an inclusionary base, populist movements in fact may be quite specifically supported by a particular class or classes.

(d) A multiclass or nonclass-based *coalition*.

(e) A *heteronomous* or politically subordinated, masssupport base, lacking autonomous power and subordinated to the leader and/or the interests of another class in the coalition.

(f) A particular *leadership style*, specifically strong, personalized, sometimes charismatic leadership.

(g) A leader-mass *linkage* that is direct, rather than mediated through organizations.

(h) Use of *rhetoric* and *oratory* that is anti-theoretical and anti-intellectual rather than coolly analytic, expository, explanatory, or abstractly elaborated in theory or ideology.

(i) An '*irrational*' form of politics in the sense that it attracts support on an emotive or moralistic rather than an interest or programmatic basis: sometimes

identified as a politics of status-anxiety rather than economic demands.

(j) A set of *economic policies* or an economic model that emphasizes growth, distribution, and demand-side stimulation, and deemphasizes fiscal and market constraints. In this way it is seen as an economic policy that is undisciplined, inefficient, and distorted or contaminated by the goal of winning political support.

Confronted with the great number and diversity of both movements and traits, Tindall (1972) proclaimed a 'semantic identity crisis.' Some analysts have responded with an attempt to construct subtypes. Canovan adopted a 'phenomenological' and 'naturalist' approach—an admittedly atheoretical exercise that yields seven types. These accommodate diversity but 'do not really look like seven varieties of the same thing; on the contrary, some of them seem quite unconnected with others' (Canovan 1982, p. 551). Roberts (1995) attempted to retain a conceptual core or commonality while accommodating diversity by proposing a 'radial' conceptualization. Here the aim is to delineate a set of traits that defines a 'pure' populism, and to recognize that subtypes (or actually existing populisms) may not in fact have all the traits in the definition. The diversity of populist movements is thus accommodated by constructing a typology of 'diminished' subtypes—types differentiated in terms of traits in the core definition which they lack.

2. Family Resemblance and Analytical Dimensions

The important point about 'populism' as it has been used is not just that types of populism lack the full set of traits. Taken as a whole, conceptualizations of populism share a family resemblance. However, different combinations of traits describe a diverse set of cases (or types) that may share few, if any, traits. Thus, 'populism' can encompass quite incongruous, even 'opposite' or inverse movements.

In the above list of traits, each seems to follow the other unproblematically. However, as one moves from the beginning toward the end of the list, one travels from authentic reform movements from below toward demagoguery—from programmatic movements based in grass-roots collective action and/or organization to instrumental top-down mobilizations of support by political leaders to facilitate their own political ambitions, or, in a Marxist framework, to promote the interests of a superordinate class. In between, populism has been associated with quite standard electoral politics and the pragmatic strategies of political leaders, particularly in catch-all parties, in order to achieve electoral victory, strategies which can be presented positively, as rational politicians instrumentally adjusting positions to conform to the median voter, and negatively, as pandering or the marketing of

candidates. (Canovan (1982) includes politicians in catch-all parties as populists.)

More important than trying to define some common core of populism is delineating some dimensions for comparing very different forms of mass politics. The different traits printed in italics in the above list point to distinct 'domains' which can be treated as dimensions on which movements analyzed as populist vary. The range of variation is evident from a description of three of the most successful sets of populist mass movements: the historical, 'classical' cases of populism in the United States and Latin America, which have received the most attention in the literature, and the recent cases emerging in diverse settings throughout the world. These cases differ dramatically with respect to, for example, leader-follower linkage, composition and heteronomy of support base, class/mass discourse, programmatic/interest vs. 'irrational' orientation, and economic policy.

3. Concrete Examples

American populism of the 1880s and 1890s arose out of the economic distress of farmers and was based in a web of thousands of independent organizations formed to address grievances at a local level. In 1892 the two regional alliances which these organizations had formed founded the Populist Party to contest national elections. The Populist Party thus emerged from the autonomous mobilization and organization of farmers. Its platform is interesting for its programmatic nature and its class discourse. Though free silver was one of the most important and well-known demands, the platform included many other economic and political planks: government distribution of credit; nationalization of railroads; graduated income tax; direct election of president, vice-president, and senators; initiatives and referenda; and a number of demands of interest to urban labor and unions. The Populists saw themselves as embodying the will of all the 'toiler classes' against the 'criminal classes.'

After a poor showing, the party retreated and in 1896 backed the candidate of the Democratic Party, William Jennings Bryan, who adopted many Populist planks and whose campaign reached oratorical heights. Those (Hicks 1955, Goodwyn 1976) who see populism as a grass-roots agrarian movement identify 1892 as its peak, after which it was coopted by the Democrats under Bryan. To the extent they see an ongoing populist thread, the analysts who identify populism with agrarian radicalism trace it forward to other progressive movements.

The conceptual confusion about populism is reflected in the fact that those who identify populism with an unmediated, oratorical leadership style see Bryan not as the coopter but as the prototype (Hofstadter 1955). Emphasizing majoritarian and crusading traits that imply intolerance and express the

irrationality of status anxiety, they trace American populism to the mobilization of an unorganized support base by paternalistic, demagogic Southern governors (e.g., Huey Long and George Wallace) and to McCarthyism, in which economic interests were even more distant, with the primary targeting of a foreign enemy and conspiracy against the nation. This alternative conception obviously involves very different sets of attributes.

'Classical' Latin American populism (of Cardenas in Mexico, Peron in Argentina, Betancourt in Venezuela, Haya de la Torre in Peru) emerged in the 1930s-40s as a strategy to win political support among the unionized proletariat in the context of the struggle by rising urban elites to attain or consolidate power. These movements consisted of urban, multiclass coalitions. It was a coalition between unequal partners, with the autonomy of workers limited to varying degrees, but generally workers had some bargaining power and political weight. The discourse was notably classist, and support mobilization occurred on the basis of real material and organizational concessions, not simply rhetorical appeal or charismatic leadership. Far from involving direct, unmediated leader-follower linkages, these were organizing periods when unions were strengthened, political parties were founded, and unions became affiliated to these parties. Union and party organizations played a central role in mobilizing support (Collier and Collier 1991).

If the first phase of early US populism is the autonomous expression and demands of farmer organizations and classical Latin American populism is the strategy of political leaders to construct an alliance with the unionized proletariat, contemporary populism is quite different. Like some US strands, contemporary populism ('neopopulism' in Latin America) generally consists of electoral support mobilization by leaders who make strong direct, rhetorical appeals to the 'people' unmediated by organizations. Yet unlike the 'classical' cases, contemporary populism occurs in the context not of a state interventionist economic model consistent with demand-side (re)distributive policies, but of the global move to marketizing policies, which are often unpopular. These policies have led to social-institutional change involving the growth of the informal sector and contingent workforce and the defensive position of organized labor. Politicians find fertile ground in the actual or threatened grievances and social dislocation resulting from new economic models, but once in office often adopt marketizing policies. Nonprogrammatic, nonclassist, noninterest-based appeals for support are thus a rational approach for sustaining governments pursuing marketizing policies.

The contrast of populisms is particularly evident in Latin America. There, classical populism represented a certain kind of inclusion and controlled mobilization of the new proletariat; neopopulism represents exclusion and demobilization. The one organizes classes;

the other disorganizes them or disarticulates their organizations. One is an alliance and bargains with the organized working class with broad payoffs (e.g., wage policy, food subsidies); the other is a rhetorical appeal to the unorganized masses (and particularly targets the growing informal sector) with selective, paternalistic distributions (e.g., local infrastructure). Postcommunist populism is a parallel form of 'post-organized' mass appeal when class solidarities and organizations have weakened or disappeared. In advanced countries, populist appeals have similarly targeted those threatened by the prospect or reality of the global relocation of production and labor, economic insecurity, and the pressure to flexibilize or liberalize labor regimes.

These examples suggest the diversity of 'populist' movements. The minimal commonality they share—that they seek to 'represent' or appeal to ordinary people in one way or another—is not very interesting; their differences are more important for elucidating the who, what, why, or how questions one might ask about them, and fitting them into larger theoretical and explanatory frameworks. If one is to use the term as more than political discourse or as a broad label, if the term is to do analytical work, it will be most helpful to delineate a common set of dimensions for scoring cases. Analysis of the resulting subtypes, with their distinct logics and dynamics, may be more useful than an ultimately elusive effort to achieve a common overall conceptualization of populism.

See also: Mass Media, Political Economy of; Political Culture; Political Discourse; Political History (History of Politics); Political Sociology; Reform: Political; Social Movements, Sociology of; Voting: Class

Bibliography

- Canovan M 1982 Two strategies for the study of populism. *Political Studies* 30: 544–652
- Collier R B, Collier D 1991 *Shaping the Political Arena*. Princeton University Press, Princeton, NJ
- Conniff M (ed.) 1999 *Populism in Latin America*. University of New Mexico Press, Tuscaloosa, NM
- Di Tella T 1965 Populism and reform in Latin America. In: Veliz C (ed.) *Obstacles to Change in Latin America*. Oxford University Press, Oxford
- Germani G 1962 *Política y sociedad en una época de transición*. Paidós, Buenos Aires
- Goodwyn L 1976 *Democratic Promise: The Populist Movement in America*. Oxford University Press, New York
- Hicks J 1955 *The Populist Revolt. A History of the Farmer's Alliance and the People's Party*. University of Nebraska, Lincoln, NE
- Hofstadter R 1955 *The Age of Reform. From Bryan to FDR*. Knopf, New York
- Ionescu G, Gellner E (eds.) 1969 *Populism. Its Meaning and National Characteristics*. Macmillan, London
- Laclau E 1977 *Politics and Ideology in Marxist Theory. Capitalism, Fascism and Populism*. NLB, London

- Ostiguy P 1997 Peronism y antiperonismo: Bases socioculturales de la identidad política en la Argentina. *Revista de Ciencias Sociales* 6: 133–213
- Pipes R 1964 Narodnichestvo: A semantic inquiry. *Slavic Review* 24: 441–58
- Roberts K H 1995 Neoliberalism and the transformation of populism in Latin America. The Peruvian case. *World Politics* 48: 82–116
- Taggart P 1995 New populist parties in Western Europe. *West European Politics* 18: 34–51
- Tindall G 1972 Populism: A semantic identity crisis. *Virginia Quarterly Review* 48: 501–18
- Weyland K 1996 Neopopulism and neoliberalism in Latin America: Unexpected affinities. *Studies in Comparative International Development* 31: 3–31

R. B. Collier

Pornography

The topic of pornography has been discussed from many vantage points and often hotly debated by scholars and researchers. Some have focused on the functions and effects of pornographic mass media as they pertain to understanding the impact of pornography on individual consumers. For example, some theorists have argued that the particular type of pornography an individual is attracted to may serve as a 'window to the mind,' revealing core developmental conflicts of the psyche (Stoller 1976). Others have considered the beneficial and/or negative effects of pornography on individuals in the realms of fantasy, sex education, esthetic/artistic expression, sexual addiction, etc. (Gagnon 1977, Gordon 1980). Yet other scholars have approached the topic at a broader level, exploring such topics as the relevance of pornography to group relations, social institutions, and the role of government regulation (e.g., Strossen 1995, Weinstein 1999). Assertions have been made that pornography use contributes to unequal power relations between dominant and subordinate classes, particularly male subjugation of women (Itzin 1992). Others have claimed that pornography is a degrading threat to social institutions and values, such as the traditional family structure (Wojtyla 1981); and that it has functioned as an agent of social change (Sorokin 1956), even as a contributing factor to major social revolutions (Hunt 1993).

It is beyond the scope of this article to consider all of these diverse approaches to the topic of pornography. Instead, the focus here may be described as a 'social science-centered' approach revolving around two central questions: First, what are the findings of social scientific studies on the effects of pornography on men (who are much more likely than women to be the consumers of such media) in the areas of aggressive attitudes and behavior? Second, how can we under-

stand this major asymmetry in male vs. female consumption of pornography?

An extremely large media industry has existed for many years that sells sexually explicit material, including movies, magazines, books, videos, and, increasingly, Internet-based media. This 'pornography industry' was recently described by *Forbes* magazine as a \$56 billion global industry that has become much more mainstream in recent years. Some companies marketing very explicit portrayals on the Internet are now listed on the Nasdaq stock exchange (Morais 1999).

1. Defining Pornography

Pornography refers to sexually explicit media that are primarily intended to sexually arouse the audience. Typical content consists of nudity and/or various sexual acts (e.g., intercourse, fellatio, etc.). Hunt (1993) notes that the production of media for this purpose is a relatively modern phenomenon, beginning toward the latter part of the nineteenth century. O'Toole (1999, p. 1) similarly argues that earlier in European history the use of sexually explicit depictions was primarily for political purposes, '... to satirise, criticise, to tilt at the Church, the state, the monarchy. ... Porn was controlled during this period not because it was obscene but because it was seditious, blasphemous, or defamatory.'

Pornography does not include sexual content that is 'embedded' or interwoven with much nonsexual content (e.g., a movie that primarily includes nonsexual content but also has one or a few sexual scenes). The research literature in that area has been recently summarized by Malamuth and Impett (2000). Sometimes the distinction between 'embedded' and explicit sexual content is not so clear. For example, *Playboy* magazine regularly includes considerable nonsexual content (e.g., interviews with politicians and jazz artists) as well as nude portrayals; the Starr Report, an official state document focusing on President Bill Clinton, included much sexually explicit content. Nevertheless, the focus here is on content that is primarily designed to sexually arouse the consumer.

In terms of legal definitions, the terms 'pornography' and 'obscenity' need to be distinguished. Obscenity refers to pornographic content that has been judged by the legal system to be illegal. In the USA, most states use the 1973 Supreme Court decision in *Miller vs. California* as the legal definition of 'obscene.' To be judged obscene under this case the material must satisfy three conditions: (a) The average person, applying contemporary community standards, would find that the work, taken as a whole, appeals to the prurient interest; (b) the work depicts sexual conduct in a patently offensive way; and (c) the work, taken as a whole, lacks serious literary, artistic, political, or scientific value.

With increasing use of new types of media such as the Internet, there may be changes in the content of pornography and the feasibility that governments can restrict access to content considered by some as offensive or harmful. For example, Barron and Kimmel (2000) recently measured the levels of sexually violent content in representative samples of sexual magazine, video, and Internet sex stories sites. They found that the Internet content was significantly higher in violence than the videos or magazines. This difference was particularly high when considering 'extreme' forms of violence (e.g., use of weapon on a victim, torture, mutilation) where 17 percent of the Internet sites contained such material as contrasted with only 2 percent in the other media. Furthermore, the Internet content depicted men in dominant positions, as victimizer and not victim, in far greater proportion than magazines and videos. The authors note that the Internet is more 'democratic' than the other media, which are constrained by dependence on commercial advertisers. The researchers suggest the Internet content is '... as close as one can get to men's direct expressions of their own fantasies, unconstrained by the demands of the marketplace or the high costs of producing and distributing those fantasies to others' (p. 166). It should be noted, however, that the men who currently use such Internet sex sites may not be representative of the larger population and may have particular attraction to more aggressive content.

2. Research on the Effects of Sexually Explicit Media

2.1 Differences Among Studies

Varied methodologies and dependent measures have been used to study the effects of pornography. Much research has centered on the potential link between pornography exposure and aggression, particularly sexual aggression. Most of the studies conducted within North America, may be described along two orthogonal dimensions based on whether they used experimental (random assignment to conditions) or correlational methodology, and whether the dependent variable they assessed was a response presumably affecting sexual aggression (e.g., attitudes supporting such aggression) or some measure of actual aggression. Such studies may therefore be organized into four categories: (a) experiments manipulating exposure to pornography and examining the effect upon attitudes as the dependent measure; (b) experiments manipulating exposure to pornography and assessing the impact on aggressive behavior as the dependent measure (in laboratory settings); (c) studies of the correlations between individual differences in pornography exposure in naturalistic settings and men's attitudes supporting aggression; and (d) two

types of studies examining the correlations between pornography exposure and aggressive behavior in naturalistic settings, those comparing criminals vs. noncriminals on pornography use, and within samples of noncriminals, those comparing different levels of reported past sexual aggression and individual variability in use of pornography.

Several meta-analyses of pornography systematically synthesize the research literature in three of the four categories and partly in the fourth category of study. Research correlating pornography use and sexual aggression within the noncriminal population has not yet been synthesized by a metanalytic review. Here, fortunately, a recent publication using a national representative sample of all men in some form of posthigh school education is available to help fill this void

Below is a summary of the findings of the various meta-analyses and the findings of the national study.

2.1.1 Experiments on the effects of pornography on attitudes. Allen et al. (1995a, 1995b) conducted a meta-analysis of experiments, both laboratory and field studies, on the effects of pornography exposure on attitudes. The meta-analysis revealed that across the various studies conducted, there was a significant though modest effect, with violent pornography resulting in significantly greater increase in attitudes supporting aggression than exposure to nonviolent pornography (although there was also some effect for both types of pornography exposure). Therefore, the overall data showed that exposure to pornography, under controlled conditions, does cause an increase in attitudes supporting sexual aggression.

2.1.2 Experiments on the effects of pornography on laboratory aggression. The meta-analysis of Allen et al. (1995a) concluded that men exposed to non-violent or violent pornography portraying sexual acts were more aggressive than those exposed to neutral content, in the laboratory research reviewed. In contrast, exposure to portrayals of nudity alone (without any sexual acts) resulted in reduced aggression in comparison to exposure to neutral content.

2.1.3 Pornography and attitudes favoring sexual aggression in naturalistic settings. Although correlational studies do not enable cause and effect conclusions, they often have the advantage of assessing responses occurring in naturalistic settings. Allen et al.'s (1995b) meta-analysis examined these studies. They found that there was a weak positive correlation that was judged not to be statistically significant between amount of exposure to pornography and atti-

tudes favoring sexual aggression. Therefore, the correlational data for attitudes in naturalistic settings did not yield supportive data for the conclusions emerging from the experimental studies conducted in controlled settings.

2.1.4 Pornography and aggression in naturalistic settings. Allen et al. (2000) conducted a meta-analysis focusing on the pornography use of convicted sex offenders as compared to those of men from the non-criminal general population. They examined several types of dependent measures: (a) frequency of pornography use; (b) age of first exposure; (c) the degree to which pornography was a direct prelude to some sexual act (masturbation, consensual sex, or forced sex); (d) and degree of sexual arousal, measured by direct genital measures of such arousal. The findings across all of the studies and measures combined did not show that criminality was associated with frequency of pornography exposure nor with age of first exposure. After viewing pornography, criminals were more likely than noncriminals to engage in some sexual acts such as masturbation, consensual, or criminal sex.

Finally, these investigators conducted a related meta-analysis examining the degree to which different types of stimuli affect people in laboratory settings. They focused on studies of physiological sexual arousal (assessed by direct genital measures) to various types of content. The researchers concluded that sexual criminals were generally more sexually aroused than noncriminals in laboratory studies examining physiological arousal to sexual stimuli. However, studies which separated portrayals of consenting and nonconsenting sex found that in comparison to non-criminals, sex criminals were more aroused by violent sex. By contrast, the difference was in the opposite direction with consenting sexual portrayals, with nonrapists showing relatively more sexual arousal. It should be noted that although criminals may be less aroused sexually by consenting depictions than non-criminals, the data on 'sexual acting out' (see above) suggest that they may still be more likely than noncriminals to engage in some sexual activity following exposure to either types of pornography exposure.

Although a meta-analysis has not been conducted on the few available studies examining among non-criminals the potential relationship in naturalistic settings between use of pornography and individual differences in self-reported sexual aggression, these few studies do point to a significant association (for a review see Malamuth et al. 2000). However, these few studies have some methodological limitations, which were rectified in a national probability study recently reported by Malamuth et al. (2000). This study found that for the majority of men, high

pornography use is not indicative of high risk for sexual aggression: Among those classified as being at relatively low risk for committing sexually aggressive acts (based on various measures previously found to predict such risk), there was only a small difference in sexually aggressive behavior between those who reported different levels of pornography use.

In contrast, pornography use was indeed a very good 'marker' of higher sexual aggression levels when these researchers considered a particular group of men: those who, based on the other measures of risk for sexual aggression, were previously determined to be at high risk. It was found that among these high-risk individuals, those who additionally were very frequent users of pornography were much more likely to have engaged in sexual aggression than their counterparts who consumed pornography less frequently. The researchers cautioned, however, that such data alone cannot be used to infer cause and effect conclusions, even if they are useful for risk assessment. They did speculate that:

associations between pornography consumption and aggressiveness toward women could be explained by a circular relationship between high coercive tendencies and interest in certain content in pornography, whereby aggressive men are drawn to the images in pornography that reinforce and thereby increase the likelihood of their controlling, impersonal, and hostile orientation to sexuality. The way relatively aggressive men interpret and react to the same pornography may differ from that of nonaggressive men. (Malamuth et al. 2000, p. 85)

2.2 Cross-cultural Comparisons

One of the most glaring apparent contradictions in the literature on pornography results from research conducted in different cultures (as well as using different methodologies). In research conducted primarily in Denmark (Kutchinsky 1970, 1991) and in Japan (Diamond and Uchiyama 1999) there has not been evidence of increased criminal sexual acts as a function of the wider availability of pornography. In commenting on such findings, Malamuth and Donnerstein (1984, p. 141) suggested some time ago that '... there may be considerable variations among individuals within a culture in susceptibility to media influences. Similarly, cultural factors may create major individual differences in the role and impact of media stimuli on members of differing societies.'

A similar point was also aptly made by Giglio (1985, pp. 289–90):

The cultural environment in each country is a factor to consider in understanding the prevailing public attitudes towards pornography. The Danes appear to enjoy a more natural approach to sex in general. Public nudity, for example, is more acceptable in Denmark than in the United States ... In a society where it is possible on a warm summer

Sunday to visit the Rosenborg Palace in the heart of Copenhagen and see dozens of partially nude women in the surrounding gardens, usually accompanied by family or friends, without the slightest public disturbance, lends credibility to Kutchinsky's theory that the Danes lack the sustaining desire for pornography to make it a profitable domestic business.

In keeping with the position that aggressive-sexual predispositions may moderate the impact of exposure to certain types of pornography, it would be expected that within countries such as Denmark, fewer men than in the USA are at high risk for sexual aggression. If this is correct, then the lack of an association between pornography use and sexual aggression would map nicely on to the findings we have obtained. Although we do not have direct evidence bearing on this issue, there are some potentially relevant data. Zak and Knack (in press) compared the levels of trust that people had for each other in various countries throughout the world. They found strong differences, with countries such as Denmark being at the very highest levels of trust between people, and these levels were considerably higher than in the USA. To the extent that such general levels of trust may also be associated with trusting between men and women (an important component of the measures used to assess risk for sexual aggression), men may be at relatively low levels of risk for sexual aggression in a society such as Denmark. Therefore, when considering the possible impact of pornography on individuals within a cultural context, it may be essential to take into account the larger context within which such exposure is embedded.

3. Gender Differences in Pornography Consumption

What is clear is that across all of the cultures that have been studied, substantial differences exist in the type of sexually explicit media men and women are more likely to be attracted to (for a review, see Malamuth 1996, Oliver 2000). Many studies provide evidence that large gender differences exist in the consumption of and gratification derived from sexually explicit media. For example, in comparison to women, men consume more regularly, are more sexually aroused by, have more favorable attitudes to, and react with less negative affect to, portrayals featuring nudity of the opposite sex and/or sexual acts devoid of relationship context (e.g., Bryant and Brown 1989, Laumann et al. 1994, Mosher and Maclan 1994, Stauffer and Frost 1976). In contrast, men are less likely than women to consume sexually explicit media that emphasize 'sexual communion' (e.g., sex accompanied by emotional closeness) and romance (e.g., Faust 1980, Krentz 1992, Lawrence and Herold 1988, Perse 1994). Various other related gender differences found in pornography research are summarized by Malamuth (1996).

3.1 Adaptive Problems and Sexual Media

These differences can be discussed using the theoretical framework of evolutionary psychology, which predicts when gender differences are or are not expected, their direction, and why they occur (Buss 1999). Generally, males and females are expected to have the same psychological mechanisms in domains where natural selection has favored a universal human adaptive solution. However, different mechanisms are expected to exist in domains in which males and females have recurrently encountered different evolutionary problems and/or where the 'optimal solutions' differed for the genders. Evolutionary theory predicts that sexuality is one of these domains. Somewhat different mating strategies probably evolved for males and females due to differences in maximum reproductive capability and parental investment. The consumption of sexually explicit media might be the result of differing evolved psychological mechanisms interacting with environmental factors.

As Malamuth (1996) discusses in some detail, there appears to be a high correspondence between the adaptive problems that evolutionary psychologists have argued underlies a 'short-term' mating strategy and the recurrent content of sexually explicit media primarily consumed by men. According to evolutionary psychologists, males who engaged in ancestral environment in a short-term sexual strategy faced problems about how to get access to as many fertile females as possible while minimizing the commitment and investment to any one woman. Pornography marketed towards males primarily portrays casual sex with numerous, accessible women who display fertility cues through age and body shape. There is similarly a high level of correspondence between female-oriented sexual media (e.g., romance novels, portrayals of sex within a relationship context) and the adaptive problems that led females to evolve a long-term sexual strategy. These problems included identifying and securing commitment from a man who had the ability (e.g., high status) and willingness to invest heavily in her and her offspring. Additionally, due to the increased vulnerability associated with pregnancy and childbirth, females preferred to mate with men who could offer physical protection. Finally, women attempted to identify men who were kind and sensitive, responses that may be strong cues of parenting skills. Romance novels, one of the most popular forms of sexually explicit media for women, frequently portray a woman securing a relationship with a high status, physically protective male who also possesses parenting skills and the willingness to invest time and resources in a single female and her offspring.

In contrast, some researchers have argued that the gender differences described above can be explained exclusively as a product of the differing socialization of males' and females' cultural roles. While this process of socialization is probably relevant to these and many

other gender differences, such a model needs to consider why such norms and rules of socialization emerged in the first place as well as to account for a large number of other related gender differences in the area of sexuality (Buss and Schmitt 1993).

4. Concluding Comments

Pornography research has often been influenced by ideological/political perspectives with vested interests in particular conclusions. This may have led to framing of research questions and design of some studies in ways that encourages simple conclusions, while not readily accommodating more nuanced conclusions (Linz and Malamuth 1993). Nevertheless, research in this area has progressed considerably. In future work, it is essential not to use an 'either-or' lens in which research is cast simply into questions such as whether pornography exposure is generally harmful or not. As suggested in the research summarized here, depending on such factors as the cultural milieu, the individual's background, the particular content of the stimuli, the type of responses focused on, the content of exposure, the consumer's environmental circumstances, and the way 'harm' is defined, differing conclusions may result. Some may wish for a 'one handed' set of conclusions, but the research more accurately justifies a 'multihanded' perspective that reveals the richness and complexity of the issues related to the study of pornography.

See also: Aggression in Adulthood, Psychology of; Freedom of the Press; Internet: Psychological Perspectives; Media Effects on Children; Media Ethics; Rape and Sexual Coercion; Regulation: Sexual Behavior; Sexual Attitudes and Behavior; Sexual Behavior: Sociological Perspective; Violence and Media

Bibliography

- Allen M, D'Alessio D, Brezgel K 1995a A meta-analysis summarizing the effects of pornography II: Aggression after exposure. *Human Communication Research* **22**: 258-83
- Allen M, D'Alessio D, Emmers-Sommer T M 2000 Reactions of criminal sexual offenders to pornography: A meta-analytic summary. In: Roloff M (ed.) *Communication yearbook* **22**. Sage, Thousand Oaks, CA, pp. 139-69
- Allen M, Emmers-Sommer T M, Gebhardt L, Giery M 1995b Pornography and rape myth acceptance. *Journal of Communication* **45**: 5-26
- Barron M, Kimmel M 2000 Sexual violence in three pornographic media: Toward a sociological explanation. *The Journal of Sex Research* **37**: 161-8
- Bryant J, Brown D 1989 Uses of pornography. In: Zillmann D, Bryant J (eds.) *Pornography: Research Advances and Policy Considerations*. Erlbaum, Hillsdale, NJ, pp. 25-55
- Buss D M 1999 *Evolutionary Psychology: The New Science of the Mind*. Allyn & Bacon, Boston

- Buss D M, Schmitt D P 1993 Sexual strategies theory: An evolutionary perspective on human mating. *Psychological Review* **100**: 204–32
- Diamond M, Uchiyama A 1999 Pornography, rape, and sex crimes in Japan. *International Journal of Law & Psychiatry* **22**: 1–22
- Faust B 1980 *Women, Sex, and Pornography: A Controversial and Unique Study*. Macmillan, New York
- Gagnon J H 1977 *Human Sexualities*. Scott, Foresman, Glenview, IL
- Giglio E D 1985 Pornography in Denmark: A public policy model for the United States? *Comparative Social Research* **8**: 281–300
- Gordon G 1980 *Erotic Communications: Studies in Sex, Sin and Censorship*. Hastings, New York
- Hunt L 1993 Introduction: Obscenity and the origins of modernity, 1500–1800. In: Hunt L (ed.) *The Invention of Pornography: Obscenity and the Origins of Modernity, 1500–1800*. Zone, New York, pp. 9–45
- Itzin C (ed.) 1992 *Pornography: Women, Violence and Civil Liberties*. Oxford University Press, Oxford, UK
- Krentz J A 1992 (ed.) *Dangerous Men and Adventurous Women*. University of Pennsylvania Press, Philadelphia, PA
- Kutchinsky B 1970 Pornography in Denmark: Pieces of a jigsaw puzzle collected around New Year 1970. In: *Technical Reports of the Commission on Obscenity and Pornography*, US Government Printing Office, Washington, DC, Vol. 4
- Kutchinsky B 1991 Pornography and rape: Theory and practice? Evidence from crime data in four countries where pornography is easily available. *International Journal of Law and Psychiatry* **14**: 47–64
- Laumann E O, Gagnon J H, Michael R T, Michaels S 1994 *The Social Organization of Sexuality: Sexual Practices in the United States*. University of Chicago Press, Chicago
- Lawrence K, Herold E S 1988 Women's attitudes toward and experience with sexually-explicit materials. *Journal of Sex Research* **24**: 161–9
- Linz D, Malamuth N M 1993 *Pornography*. Sage, Newbury Park, CA
- Malamuth N M 1996 Sexually explicit media, gender differences and evolutionary theory. *Journal of Communication* **46**: 8–31
- Malamuth N M, Addison T, Koss M 2000 Pornography and sexual aggression: Are there reliable effects and how can we understand them? *Annual Review of Sex Research* **11**: 26–91
- Malamuth N M, Donnerstein E 1984 Correlational and cross-cultural studies on pornography and aggression. In: Malamuth N M, Donnerstein E (eds.) *Pornography and Sexual Aggression*. Academic Press, New York, pp. 139–41
- Malamuth N, Impett E A 2000 Research on sex in the media: What do we know about effects on children and adolescents. In: Singer D, Singer J (eds.) *Handbook of Children and the Media*. Sage, Newbury Park, CA, pp. 269–87
- Morais R C 1999 Porn goes public. *Forbes Magazine*, 14 June, p. 214
- Mosher D L, MacLan P 1994 College men and women respond to X-rated videos intended for male or female audiences: Gender and sexual scripts. *Journal of Sex Research* **31**: 99–113
- Oliver M B 2000 The respondent gender gap. In: Zillmann D, Vorderer P (eds.) *Media Entertainment: The Psychology of its Appeal*. Erlbaum, Mahwah, NJ, pp. 215–34
- O'Toole L 1999 *Pornocopia: Porn, Sex, Technology and Desire*. Serpent's Tail, London
- Perse E M 1994 Uses of erotica and acceptance of rape myths. *Communication Research* **21**: 488–515
- Sorokin P 1956 *The American Sex Revolution*. Sargent, Boston
- Stauffer J, Frost R 1976 Male and female interest in sexually-oriented magazines. *Journal of Communication* **26**: 25–30
- Stoller R 1976 Sexual excitement. *Archives of General Psychiatry* **33**: 899–909
- Strossen N 1995 *Defending Pornography: Free Speech, Sex, and the Fight for Women's Rights*. Scribner, New York
- Weinstein J 1999 *Hate Speech, Pornography, and the Radical Attack on Free Speech Doctrine*. Greenwood Press, Westport, CT
- Wojtyla K (Pope John Paul II) 1981 *Love and Responsibility* [Willetts H trans.]. Farrar, Straus & Giroux, New York
- Zak P J, Knack S in press Trust and growth. *Economic Journal*

N. M. Malamuth

Positivism, History of

Positivism is the name of a social and intellectual movement that tried to learn from the mistakes of the Enlightenment project that eventuated, first, in the Reign of Terror following the French Revolution of 1789, and second, in the irrationalism of the Weimar Republic following Germany's defeat in World War I. While it has been customary to distinguish between the quasipolitical movement called 'positivism' originated by Auguste Comte in the 1830s and the more strictly philosophical movement called 'logical positivism' associated with the Vienna Circle of the 1930s, both shared a common sensibility, namely, that the unchecked exercise of reason can have disastrous practical consequences. Thus, both held that reason needs 'foundations' to structure its subsequent development so as not to fall prey to a self-destructive scepticism. In this respect, positivism incorporates a heretofore absent empiricist dimension to the risk-averse orientation to the world historically associated with Platonism.

1. Positivism's Platonic Political Roots

In terms of Western intellectual history, positivism inflects Plato's original philosophical motivation through a secularized version of the Christian salvation story, in which Newton functions as the Christ figure. This captures both the spirit of Auguste Comte's original project and its residual effects in twentieth century logical positivism, which dropped the overt historicism of Comte's project, while retaining the fixation on Newton as the model for what it means to express oneself scientifically and a vague belief that greater scientific knowledge will deliver salvation. Indeed, positivism's core conceptual problem has been to define a scientific vanguard capable of both offering guidance to the unenlightened and itself

changing in light of further evidence and reflection. Its Platonic roots are worth recalling.

Plato was affected deeply by the volatile political environment created by the public dialectical encounters of Socrates and the other Sophists who mentored the leaders of his Athenian youth. Thus, after the fall of Athens, Plato was keen to restrict the dialectic to a cloistered setting, only to be unleashed to the larger society once initiates were sufficiently secure in their own knowledge and judgement that they would not be unduly swayed by popular demands. Where Plato had hoped to produce implacable philosopher-kings who would rule as absolute monarchs, positivists typically have envisaged a more differentiated but no less authoritative (authoritarian?) rule by experts, each an oligarch over their domain of knowledge. In this respect, positivism is bureaucracy's philosophical fellow traveler.

Not surprisingly, positivism's relationship to democracy has been chequered. Like Plato, positivists have feared protracted public disagreement most of all and hence have tended to demonize it as 'irrational' and 'noncognitive.' Their image of 'plural' authority presumes nonoverlapping competences, such that legislative questions are reduced to judicial ones concerning the expertise to which one should defer. This is in sharp contrast to most republican conceptions of democracy, which base representation on interest groups, each invested with a competence appropriate to the promotion of its interests but typically about matters in which other interest groups might also have competence. Since matters of public policy typically affect several such groups at once, collective decisions do not turn on identifying the right group whose judgement should hold sway. Rather, it is expected that public debate will issue in a solution that transcends all the groups' starting points but nevertheless manages to serve their respective interests.

Thus, there is a fundamental ambiguity in positivism's appeal to organized reason, or 'science,' in the public sphere. Sometimes this ambiguity is finessed by saying that positivists regard science as the main source of political unity. At the very least, this implies that it is in the interest of all members of society to pursue their ends by scientific means, as that may enable them to economize on effort and hence allow more time for the fruits of their labour to be enjoyed. Ernst Mach (discussed in Sect. 3) comes closest to defending this position in its pure form. It fits comfortably with the libertarian idea that democratic regimes should enable maximum self-empowerment.

However, many positivists have drawn a further conclusion that can thwart this libertarian impulse. From Comte onward, it has been common to argue that science can unify the polity by resolving, containing, or circumventing social conflict. Here a well-established procedure or a decisive set of facts is supposed to replace more 'primitive' and volatile forms of conflict resolution such as warfare and

sometimes even open debate—all of which supposedly compromise the integrity of opposing viewpoints in the spirit of expedience. Accordingly, a scientific politics should not merely satisfy the parties concerned: it should arrive at the 'correct' solution.

To be sure, even this mentality admits of a democratic interpretation, as positivist social researchers have been in the forefront of presenting 'data' from parties whose voices are unlikely to be heard in an open assembly. Typically, this has occurred in surveys designed to represent the full empirical range of a target population. Nevertheless, the question remains of exactly who reaps the political benefits of these newly articulated voices: the people under investigation, the investigators themselves, or the investigators' clients? Moreover, once a target population has been registered empirically, do its members remain 'objects of inquiry' or are they promoted to full-fledged inquirers capable of challenging the original investigators' findings and methods? Probably the most sophisticated treatment of these questions in the context of positivistically inspired US social policy research is to be found in Campbell (1988).

These delicate questions arise because ultimately positivism turns Plato on his side by converting a static hierarchy into a temporal order. Where Plato imagined that authority flowed downward from the philosopher-king in a caste-based social structure, positivists have envisaged that all of humanity may pass (at a variable rate) through a sequence of stages that retrace the socioepistemic journey from captivity to autonomy. In the positivist utopia, it is possible for everyone to be an expert over their own domain. Moreover, there is a recipe for the conversion of Platonism to positivism. It proceeds by isolating a domain of inquiry from the contingencies surrounding its manifestations so that its essential nature may be fathomed. Whereas Plato reserved such inquiry to philosopher-kings, positivists have more often turned to state-licensed professional bodies. And instead of Plato's intellectual intuition (*nous*), positivists attempt to gain epistemic access by comparative historical and experimental methods.

This recipe can be illustrated in the work of Otto Neurath, an organizer of the Vienna Circle. He wanted to isolate the essence of the 'war economy' so that its efficient central planning mechanism can be transferred to environments where it would have more socially salutary consequences. Here Neurath anticipated what Alvin Gouldner would call the Janus-faced character of the 'welfare-warfare state,' whereby the same organizational structure (in this case, a concentration of resources in the nation-state) can have radically different consequences, depending on the supporting political environment. Nevertheless, as Neurath's many critics pointed out, positivism seems to have inherited Platonism's political naivete, which confuses the fact that, say, the 'war economy' can be identified analytically as a feature of many societies

and the analyst's ability to transfer it to new social environments—bar ring the imposition of sufficient force to hold all other environmental factors constant. If anything deserves the name of the 'positivist fallacy' it is this too easy assimilation of the forum to the laboratory. The history of positivism can be neatly captured as a Hegelian dialectic, the three moments of which are epitomized by the work of Auguste Comte (thesis), Ernst Mach (antithesis), and the Vienna Circle (synthesis). However, these moments have historically overlapped, occasionally coming together in figures such as Neurath. The career trajectories of positivism's standard bearers help explain the direction taken by their thought.

Comte was an early graduate of the *Ecole Polytechnique* who believed that its Napoleonic mission of rendering research a vehicle for societal transformation had been betrayed, once he himself failed to achieve a permanent academic post. Mach was a politically active physicist on the losing side of so many of the leading scientific debates of his day that his famous chair in Vienna, from which the logical positivists sprang, was awarded on the strength of his critical-historical studies, not his experimental work. Finally, the intellectual leader of the Vienna Circle, Rudolf Carnap, had to abandon physics for philosophy because his doctoral dissertation topic was seen as too 'metatheoretical' for a properly empirical discipline. For Carnap and others who came of age in World War I, physics had devolved into another specialized field of study, rather than—as it had still been for Einstein—natural philosophy pursued by more exact means.

2. Thesis: Comtean Positivism

Auguste Comte coined the word 'positivism' to refer to the positive side of the Enlightenment's essentially destructive approach to traditional social and intellectual orders. Comte agreed with the Enlightenment wits that scientific rationality is the model of social rationality, the exemplar of which is Newtonian mechanics. However, according to Comte, the French Revolution had refuted the wits' belief that people's 'natural reason' would come to light once they were freed from their captivity to religion and other traditional forms of authority. Comte thus designed positivism to replace both the church and the sovereign with scientific experts who together would function as a secular priesthood for modern society. 'Sociology,' another of Comte's coinages, referred at once to this political vision and the scientific project that would render it a legitimate form of authority.

Positivism in this basic sense has remained the philosophy underwriting science as a specialized profession of 'scientists' with an identifiable order of merit and career structure. Ironically, none of the movement's first phase protagonists—including Comte

himself, John Stuart Mill, and Herbert Spencer—ever called a university home. Nevertheless, they all earned their living from producing knowledge (some would call it 'ideology'), unlike those for whom the pursuit of knowledge was a refined way of consuming wealth. In this important respect, positivism was aligned with the ascendant bourgeoisie and hostile to such aristocratic forms of learning as the gentleman-scientist and the academic humanist. Comte arrived at this position from his mentorship under Count Henri de Saint-Simon, a spiritual godfather of today's management gurus, who at the dawn of the industrial age anticipated that organized knowledge would be the principal factor of economic production and political control in what Daniel Bell would call, 150 years later, 'post-industrial society.'

Given its explicitly anticlerical posture, positivism was more often found in medical schools and polytechnic institutes than traditional universities. (French universities did not become fully secular institutions until the Third Republic in 1870.) The medical interest is noteworthy because, in the nineteenth century, positivism was used to legitimize everything from phrenological and psychosomatic research to the experimental approach to medicine championed by Claude Bernard. What they all shared was the recognition that physical constitution plays a significant role in defining psychic, or spiritual, life. This basic assumption was sufficient to challenge religion's ministerial monopoly.

More generally, positivism's anticlericalism has been among its most powerful export values, especially in Latin America (where Comte's motto 'Order and Progress' remains emblazoned on the Brazilian flag). Without a thoroughgoing Protestant Reformation, Latin Americans found Comte's secularized version of Catholicism much less utopian than did their European counterparts. To be sure, background cultural allegiances divided positivist loyalties, with Francophile Latin Americans (e.g., Brazil) preferring Comte and Anglophile ones (e.g., Argentina) Spencer. In either case, though, positivism was equally likely to justify authoritarian and democratic regimes. Perhaps the most lasting intellectual residue of positivism's Comtean origins has been the default association that social scientists make between reliable knowledge and statistically significant empirical correlations. Hempel (1942), the canonical positivist account of historical explanation, placed the Vienna Circle's distinctive stamp on this point. Nowadays, its import is easily lost in methodological discussions that trivialize correlations as deficient accounts of causation, proof that positivism owes too much to Francis Bacon. However, following Karl Pearson, perhaps the first self-declared positivist to hold a professorship in Britain, the appeal to correlations served at least four distinct functions.

First, correlations enabled an economy of thought by summarizing observations already made. This would decrease the likelihood of 're-inventing the

wheel,' thereby expediting overall human progress. Second, correlations forced investigators to look beyond the singular striking success to more generalizable, albeit mundane, tendencies. Here the positivism helped moderate the hype surrounding new inventions and discoveries. Third, the focus on correlations implied that scientific work could be mechanized by treating individual experiences as instances of more general types. This helped dispel the idea that scientists must have a special 'genius' for finding hidden causes beneath the welter of data. Fourth and probably most significant in Pearson's own turn-of-the-century context, the appeal to correlations echoed the 'uniformitarian' approach to natural history championed by Darwin, whose theory of evolution presupposed that all biological change has occurred by principles that remain in effect today. Thus, there was no need for either a divine creator behind the scenes whose actions transcend observed correlations or a mystification of statistically extraordinary events as 'miracles.'

3. *Antithesis: Machian Positivism*

Ernst Mach symbolized the second moment of positivism's dialectic: he gave the movement a dose of its own medicine, namely, a scientific sense of science's own limitations. Mach's counsel emerged in the midst of the ironic, perhaps perversely self-defeating, character of *fin de siècle* Viennese culture, whereby Karl Kraus's quip that psychoanalysis is the disease for which its cure was eventually incorporated into positivistic lore as Ludwig Wittgenstein's dark saying that philosophy is the ladder that must be climbed in order to be discarded (Janik and Toulmin 1973).

In the long term, this sense of scientific self-restraint has made its strongest impression in the dual doctrines of 'value neutrality' and 'academic freedom,' which are most closely associated in the social sciences with Max Weber. Although these doctrines normally are treated as separate, it is difficult to motivate either without examining their historical interconnection. The missing link is the idea that, absent further empirical demonstration, the value of theoretical discourse—the most obvious epistemic marker distinguishing experts from the lay public—is purely heuristic. Theory enables those who already know something to know much more; however, it obscures the vision of those who have yet to know anything.

Thus, the various Mach-inspired positivist projects of 'reducing' theoretical discourse to its empirical bases have been motivated by a keen sense of the contingency of when and where new scientific insights arise. The distinction that Hans Reichenbach and Karl Popper drew between the contexts of discovery and justification canonized this point for contemporary philosophy of science. The unique means by which a discovery is made is neither necessary nor sufficient for demonstrating its validity. Elementary logic textbooks

demonize the failure to abide by this point as the 'genetic fallacy.' For Mach, the contingent nature of discovery places a special burden on scientists to render their insights as widely as possible, without denying themselves the right to follow it up in their own way. The Vienna Circle positivists may be seen as having compromised this legacy by limiting the context of justification to the received canons of inductive and deductive reasoning. But perhaps they had good reason.

Mach may have been too keen to let the lay public draw whatever conclusions they wished from their studies of science. The second moment of positivism enabled the conversion of scientific knowledge into a pure instrument detached from any prior theoretical or normative commitment. In practice, this 'neutral' stance made science available to the highest bidder, who in turn could hold a monopoly over its use (or nonuse). Consider the fate of the law of diminishing marginal utility (LDMU), which states that if there is enough of a good to satisfy a basic level of need, then each additional increment of the good satisfies the need less than the previous increment. LDMU first appeared in John Stuart Mill's normative political economy, the source of modern welfare economics. The German translation of Mill's *System of Logic* in 1849 had introduced the first phase of positivism to Mach and such contemporaries as Wilhelm Wundt and Wilhelm Dilthey, initiating the long-standing debates over the reducibility of the methods of the *Geisteswissenschaften* (Dilthey's rendering of Mill's 'moral sciences') to those of the *Naturwissenschaften*. While Mill's later work, *On Liberty*, would inspire Mach's own views linking academic freedom to the right of dissent, Mach refused to follow Mill in drawing normative conclusions from putatively scientific facts, even though Mill's aim of removing the accumulated advantage of the rich was close to Mach's own heart.

Mill treated LDMU as a scientific basis for redistributionist social policies. He inferred that each additional increment of income earned by the rich would improve their lives less than the same increment transferred to the poor. Mill presumed a common standard of interpersonal utility and the idea that different interpersonal endowments could be explained as reversible historical accidents. Moreover, he modeled LDMU on a certain interpretation of Newton's laws, which holds them to be true in an ideal physical medium but not in actual physical reality. Thus, Mill made the ideality of the scientific laboratory stand for the normative basis of social reform. What is now called the 'marginalist revolution' in economics reinterpreted the Newtonian precedent for LDMU in the value-neutral terms that captured the social situation of Mach and other second generation positivists who called a university home.

In the 1870s, Mill's academic nemesis, William Stanley Jevons, had begun interpreting LDMU as the

emergent product of interactions between agents whose respective utilities are in principle incalculable, that is, an empirically stable resolution of unknowable constitutive forces. Today we would say that LDMU is a macrolevel market effect, in which the parameters of the relevant market are left unspecified. Thus, cases that fail to conform to LDMU do not force a difficult normative judgement about whether the law or the case is at fault; rather, one casts around for an appropriately defined market to which LDMU might apply, since one may not have considered all the effects over a large enough scale, over a long enough period.

Second generation positivists came to regard LDMU as holding no clear implications for social policy. This opened the door to the curious twentieth century 'noninterventionist' policies that Milton Friedman and others have used to oppose scientific justifications of the welfare state. Here positivism's antitheoretical and antinormative stance is explicitly identified with the inductively self-organizing order of the so-called invisible hand. The shift from Jevons to Friedman is indicative of what logicians call the 'modal fallacy': The denial that X implies Y is interpreted to mean that X implies not Y , where X = LDMU; Y = State-managed redistribution of income. Thus, an economic principle that was seen originally as not necessarily licensing a particular policy intervention came to be seen as an outcome that obviates the need for any policy intervention whatsoever.

Nevertheless, the transition between the first two moments of positivism was neither smooth nor clearly separated from the third. Free market liberals continued to invoke Comte's original sense of positivism to stigmatize large-scale, usually Marx-inspired, social planning of the economy (e.g., Hayek 1952). A target of these studies was Vienna Circle organizer, Otto Neurath, who supposed that a specialist understanding of the persistent underlying structures of political economy provides a better handle on social policy than the spontaneously aggregated experiences of individual producers and consumers. While Neurath's Marxist orientation to social planning was close to Comte's privileging of expert over lay judgement, he also accepted the Machian view that useful knowledge should be spread as widely as possible. In that sense, Neurath's interest in designing an ideographic language ('Isotype') to enable the populace to understand capitalism's inequities was born of the same impulse as his debates with Rudolf Carnap over the construction of the 'protocol statements' that constitute the fundamental language of science. Here the second moment of positivism naturally shades into the third.

4. *Synthesis: The Vienna Circle*

The Vienna Circle was started by Moritz Schlick, another physics-trained philosopher who succeeded Mach to the chair of the history of inductive sciences

at the University of Vienna. Much has been made of the Vienna Circle's preoccupations with the conceptual and empirical foundations of science, the demarcation of science from nonscience, and the unification of the special sciences under a single metaphysical and/or methodological rubric. Less has been said to explain the Circle's partial return to positivism's communitarian and authoritarian origins, especially after Mach had made positivism virtually synonymous with the liberalization of epistemic authority. Nevertheless, there were few true liberals in the Vienna Circle: Schlick and Wittgenstein were conservatives, Carnap and Neurath socialists. Popper, the Circle associate closest to Mach's own politics, was not easily accepted into the fold.

This antiliberal turn can be explained by changes in science's social relations in the German-speaking world over the first quarter of the twentieth century. Seen against this background alone, the Vienna Circle appears utopian and even prone to nostalgia; but with Nazism's ascendancy in the 1930s, the 'logical positivist' movement quickly acquired a defiant and progressive cast. At first, the Circle recoiled from Machian positivism because of the ease with which Mach's neutral instrumentalism allowed scientists, especially chemists, to harness their knowledge for aggressive military-industrial ends. Consequently, Germany's humiliating defeat in World War I paved the way for the antiscientific backlash of Weimar culture that formed the backdrop to the Circle's meetings. In recognizably Platonic fashion, the Circle forged a philosophy that protected science from this backlash, while divesting it of any potentially corrupting links to technology. The key was the idea that a genuine science must satisfy certain logical and empirical criteria relating to the public testability of knowledge claims. The model for this requirement was ongoing debates over the foundations of mathematics, which Carnap and Reichenbach had already transferred to debates over the presuppositions of relativity and quantum theory.

However, one decidedly anti-Platonic insight that the Vienna Circle retained from Mach, which increased in importance over the years, was that esoteric forms of knowledge must be 'reducible' to their core 'cognitive' components so as to facilitate both their evaluation and utilization. Specifically, the Vienna Circle aimed to demystify 'metaphysical' jargon that encoded prejudice without extending insight. (Carnap famously cited Heidegger in this respect.) At the peak of its popularity in the 1940s, logical positivism was linked with such American best sellers of the 'general semantics movement' as Hayakawa (1978). Here the foe had shifted from Heidegger to Hitler, with the positivists arguing against the irrationalism of mass propaganda. Perhaps the most famous single work of the positivist movement, Ayer (1936), written by the British observer at the Vienna Circle, can be seen as a transitional work between these two phases.

Given this history, logical positivism appears in a rather ironic light in the highly publicized *Positivismusstreit* that transpired in 1961–8 between Karl Popper and Frankfurt School doyen, Theodor Adorno. It turns out that both Popper and Adorno demonized ‘positivism’ in social science for the exact same ills as the logical positivists had demonized Heidegger and Hitler a quarter-century earlier, namely, a ‘jargon of authenticity’ and its contribution to an ‘authoritarian personality’ (two relevant Adornian phrases). Moreover, the two antagonists presupposed the logical positivists’ ideal of unified inquiry and the inadequacy of defining society as simply the totality of ‘social facts.’ Where Popper and Adorno clearly differed from the positivists was over the latter’s promotion of a sharp fact-value distinction, whereby the presumed unity of factual knowledge offsets the irreducible plurality of value orientations. Here positivism courted a ‘relativism’ that both antagonists strongly opposed. Nevertheless, the ascendancy of relativism, in the guise of ‘postmodernism,’ would render Popper and Adorno’s wranglings over ‘critical reason’ beside the point by the 1980s.

In detecting positivism’s latent relativism, Popper and Adorno not only anticipated the future; they also recalled the ‘value neutrality’ that had originally turned Machian positivism into a vehicle of power politics. In the social sciences, Max Weber’s distinction between charisma and bureaucracy as modes of legitimation enabled both politicians and scientists to turn a blind eye to each other’s activities, as the two groups were supposedly driven by radically different normative orientations: the politician acts, while the scientist provides reasons. Similar criticisms have been made of other movements in the social sciences affiliated with positivism, most notably Hans Kelsen’s legal positivism and B. F. Skinner’s radical behaviorism. Yet, even Neurath was happy to call his project ‘neutral Marxism.’

These cases are united by a refusal to blur the boundary between science and politics, which in practice has meant a rejection of social democratic, welfare-state policies in favor of more extreme political stances: either complete withdrawal or outright authoritarianism. Once again, positivism would seem to have fallen victim to the modal fallacy: because politics and science can contaminate each other, therefore they inevitably will, and hence each must be left to its own devices.

5. Final Assessment: Does Positivism Have a Future?

After the leading members of the Vienna Circle migrated to the US in the 1930s, logical positivism seeded that country’s analytic philosophy establishment for the second half of the twentieth century. However, this is the only context in which positivism

possibly dominated an established discipline. For the most part, positivism has been embraced by disciplines that have yet to achieve academic respectability, even in the natural sciences, where Mach found his strongest support among chemists, biomedical scientists, and psychologists—not physicists. (It is often overlooked that positivism’s reliance on Newtonian mechanics as the model for all science generally was not appreciated by a physics community jealous to guard its guild privileges.) Unsurprisingly, positivism’s most ardent supporters have been social scientists, not for the Comtean reason that sociology is the pinnacle of all science but for the more mundane reason that positivism seemed to offer a strategy for rendering one’s activities ‘scientific.’

These matters came to a head with the publication of Kuhn (1970), which Carnap enthusiastically endorsed as the final instalment of the logical positivists’ *International Encyclopedia of Unified Science*. It provided an account of scientific change based largely on the history of physics that was quickly embraced by social scientists, as Kuhn stressed science’s self-organization over its larger societal impact. However, unlike previous positivist accounts, Kuhn’s explicitly fixated on particular scientific disciplines (or ‘paradigms’) without presuming that science as a whole is heading toward a unified understanding of reality. By implication, then, all scientific standards are discipline-relative. Kuhn’s approach suited what is now called the ‘postmodern condition,’ which the *Positivismusstreit* had desperately tried to stave off.

In retrospect, the popularity of Kuhn (1970) is better understood as signifying positivism’s decadent phase than, as it was originally seen, a fundamental challenge to positivism. The decadent state of contemporary positivism can be gleaned from the ongoing debates over the scientific status of ‘sociobiology.’ Each side of the debate captures part of the positivist legacy: proponents like E. O. Wilson hark back to positivism’s Comtean vision of unified knowledge in aid of societal transformation, while opponents like Richard Lewontin draw on positivism’s more recent restriction of the bounds of science to discipline-specific criteria.

If positivism has a future, it lies in rekindling a sense of ‘Science’ that transcends the boundaries of particular scientific disciplines. This is how Comte originally thought about the discipline he called ‘sociology.’ He claimed that it was the last to develop, not simply because of the complexity of its human subject matter, but more importantly, sociology had to reconstitute the (natural) sciences that historically preceded it. Too often the history of positivism’s quest for unified science has been interpreted as exclusively a matter of applying the methods of physics to less developed sciences. The reciprocal movement is really more important, namely, the application of sociological findings to the future direction of science as a whole. Such interdisciplinary projects as ‘social epis-

temology' and the 'science of science' have tried to fulfill this side of the Comtean promise, which may be redeemed once the present intellectual climate changes.

See also: Behaviorism; Comte, Auguste (1798–1857); Critical Theory: Frankfurt School; Durkheim, Emile (1858–1917); Empiricism, History of; Enlightenment; Kuhn, Thomas S (1922–96); Logical Positivism and Logical Empiricism; Mill, John Stuart (1806–73); Popper, Karl Raimund (1902–94); Positivism: Sociological; Wittgenstein, Ludwig (1889–1951)

Bibliography

- Adorno T (ed.) 1976 *The Positivist Dispute in German Sociology* (trans. Adey G, Frisby D). Heinemann, London
- Ayer A J 1936 *Language, Truth and Logic*. Gollancz, London
- Campbell D T 1988 *Methodology and Epistemology for Social Science: Selected Papers*. University of Chicago Press, Chicago
- Cartwright N, Cat J, Fleck L, Uebel T 1996 *Otto Neurath: Philosophy between Science and Politics*. Cambridge University Press, Cambridge, UK
- Fuller S 1993 *Philosophy of Science and Its Discontents*, 2nd edn. (orig. 1989). Guilford Press, New York
- Fuller S 2000 *Thomas Kuhn: A Philosophical History for our Times*. University of Chicago Press, Chicago
- Hayakawa S I 1978 *Language in Thought and Action*, 4th edn. (orig. 1939). Harcourt Brace and Jovanovich, New York
- Hayek F 1952 *The Counter-Revolution of Science: Studies on the Abuse of Reason*. University of Chicago Press, Chicago
- Hempel C G 1942 The function of general laws in history. *Journal of Philosophy* 39: 35–48
- Janik A, Toulmin S 1973 *Wittgenstein's Vienna*. Simon and Schuster, New York
- Kolakowski L 1972 *Positivist Philosophy From Hume to the Vienna Circle* (trans Guterman N). Penguin, Harmondsworth, UK
- Kuhn T S 1970 *The Structure of Scientific Revolutions*, 2nd edn. (orig. 1962). University of Chicago Press, Chicago
- Proctor R 1991 *Value-Free Science? Purity and Power in Modern Knowledge*. Harvard University Press, Cambridge, MA
- Turner S P 1986 *The Search for a Methodology of Social Science*. Kluwer, Dordrecht, Netherlands
- Zea L 1963 *The Latin American Mind* (trans. Abbott J A, Dunham L). University of Oklahoma Press, Norman, OK

S. Fuller

Positivism: Sociological

Positivism arose in the social sciences in an effort to legitimate sociology as a scientific discipline. In essence, positivism asserts that the social sciences, and sociology in particular, are natural sciences in

which abstract laws of human organization can be formulated to explain the operative dynamics of the social universe. The plausibility of these laws are then to be assessed against systematically collected empirical data.

1. The Rise of Positivism

Advocacy for a positivistic social science began with Auguste Comte's (1830–42) *Course of Positive Philosophy*. Comte was, no doubt, influenced by Montesquieu's ([1748] 1900) *The Spirit of Laws*, where the notion of formulating laws about the operation of human society was first suggested, and by both Turgot and Condorcet, from whom Comte received, via his mentor, Saint-Simon, the notion of the three stages in the evolution of idea systems. For Comte, the various sciences developed at different rates, with sociology (and by implication the other social sciences as well) emerging last. As each science develops it goes through three stages—theological, metaphysical, and eventually, positivistic. By the early 1800s, the idea that sociology and other social sciences could take their place at the table of science was widespread, but it was Comte who forged the various lines of argument into a forceful advocacy for a science of society. Comte's preferred name for sociology was *social physics*, and this name demonstrates his desire to develop a sociology organized around basic laws of the social universe.

In *Course of Positive Philosophy*, Comte posits a view of sociology as a theory-driven enterprise. As a descendant of the French Enlightenment, and as one who admired greatly what Newton had done for astrophysics, Comte argued for a sociology in which abstract laws of human organization would explain the dynamics of the social universe. His exemplar was Newton's law of gravity, and yet he never developed any such law—although he saw the 'law of the three stages' as something like Newton's law of gravity. Still, from the opening pages of *Positive Philosophy*, a clear message is sent: 'The first characteristic of Positive Philosophy is that it regards all phenomena as subject to invariable natural Laws' (1830, pp. 5–6, emphasis in original). In this same passage, Comte argues against a search for 'Causes, whether first or final'; rather the goal of positivism is the 'discovery of laws, with a view to reducing them to the smallest possible number.' In later portions of *Positive Philosophy*, Comte (1830–42) buttressed this emphasis on abstract laws by criticizing raw empiricism, for 'the next great hindrance to the use of observation is the empiricism which is introduced by those who, in the name of impartiality, would interdict the use of any theory whatsoever.' Comte concludes this passage with an even stronger statement: 'No real observation of any kind of phenomena is possible, except in as far

as it is first directed, and finally, interpreted, by some theory.'

Thus, in Comte's view, all observation is mediated through a theory, and the goal of positivism is to make the theories explicit and formal, while employing systematic methods to gather data on the plausibility of a theory. He advocated four basic methods: observation, experimentation, comparison, and historical analysis in order to collect the broadest array of data possible to assess the validity of laws that are presumed to explain the operation of the social universe in all times and places.

By 1830, then, Comte had laid out in bold and incisive fashion the basic tenets of positivism. As Comte's mental instability caused his star to fade, positivism was carried forward by Herbert Spencer in England and Emile Durkheim in France. The titles of almost all of Spencer's important works indicate the positivistic slant of his work in both the natural and social sciences: *First Principles*, and *The Principles of* (sequentially) *Psychology*, *Biology*, and *Sociology*. But nowhere was Spencer's advocacy more powerful than in his first published sociological work, *The Study of Sociology* (1873), where he not only reaffirms his commitment to seeing the universe as guided by forces whose dynamics can be captured by abstract principles, but also articulates in a systematic manner the sources of bias that can distort the formulation of these principles and their systematic assessment by data. Twenty-two years later, Emile Durkheim ([1895] 1938) made a similar plea in his *The Rules of Sociological Method* for a sociology based upon laws about 'social facts.' Both Spencer and Durkheim personified how positivistic sociology could be conducted, mixing sophisticated functional theory with presentations of data. Indeed, from Comte's early analogies to organisms, positivism was associated initially with functional sociology, a theoretical point of view that was to flower in the middle of the twentieth century, and then to die on the vine in the latter part of the century. And as functionalism faded, so did positivism as a coherent point of advocacy. Yet, much of the fate of positivism in the twentieth century is associated with intellectual ferment in Europe around what was to become known as the Vienna Circle.

2. The Vienna Circle and the Transformation of Positivism

During the first half of the twentieth century, a group of intellectuals located mostly in and around Vienna debated issues of thought and reasoning as well as connections between the empirical world and modes of representing this world (Johnson 1983). Two outside and somewhat marginal figures of the Vienna Circle—Ernst Mach (1893) and Karl Popper (1959, 1969)—are critical to the debate that raged within the

Circle for the first half of the twentieth century. Mach had proposed a radical empiricism in which all conceptions of the universe as guided by natural laws were rejected in favor of mathematical descriptions of relations among empirically observable variables. This extreme position called for a radical empiricism, emphasizing that formal thought must always be tied to, and inducted from, observable empirical facts. At the other pole were those concerned with the structure and logic of formal thinking to either the exclusion or subordination of empirical observations to rules of logic. Karl Popper's famous solution to this impasse was to emphasize that theories can never be proven by empirical facts, only disproven. The goal of science is, therefore, to assess the empirical plausibility of hypotheses deduced formally from abstract principles by repeated assaults from empirical observations. For many decades, this argument became the new standard way to assert the viability of positivism; and it led to the notion of 'logical positivism' in which emphasis is on the logic of deduction from abstract principles to testable hypotheses. Comte had never conceptualized how one makes abstract laws available for empirical assessment, and so this point of emphasis was perhaps required in a more complete formulation of positivistic science.

3. Vienna Circle Debates and US Sociology

The Vienna Circle had much more influence on US social science than on disciplines in Europe where scholars were, in general, more philosophically sophisticated than their US counterparts. But much more was involved; US social science adopted Popper's argument because these Vienna-inspired arguments were compatible with the new social science emerging in the USA.

If we look at US sociology of the 1920s and 1930s, there was a vacuum of theory. Spencer's star had faded, and other figures in sociology's European pantheon—Weber, Durkheim, Simmel, and Pareto—remained largely untranslated and, hence, unknown to many US sociologists. And yet, US sociology had been very much committed to Comte's view of science, and many early texts refer to 'principles' of various subjects in a manner reminiscent of Spencer. At the same time, quantitative statistical methods were being adopted by US sociologists under the influence of Franklin Giddings (1901), and later his students and followers. Sociologists were thus very much interested in finding a legitimating mantra for statistical analysis of data that could make claims to being scientific (Turner and Turner 1990). One mantra was quantification, *per se*, but many early sociologists had a more Comtean–Spencerian view of science as a search for laws. Popper's solution provided the reconciliation of the debate within not

only the Vienna Circle but also within conflicting camps of US sociology as well: theories were to be tested against quantitatively arrayed data using new statistical techniques borrowed from the European agronomist, Karl Pearson. Through much of the mid-century, the theory part of this reconciliation took a back seat to quantification. And, not surprisingly, it is from this period that the misperception of positivism as 'raw empiricism' becomes prominent, despite the fact that this label is wholly inappropriate to positivist epistemology.

The radical empiricism of the Vienna Circle found a comfortable home in American quantitative research in the last quarter of the century, and for a brief time, the logical part of positivism was trumpeted through the effort to 'build' and 'construct' theories. A host of books on 'theory building' and 'theory construction' emerged in the late 1960s, beginning with Hans Zetterberg's (1965) *On Theory and Verification in Sociology*, and continuing through the 1960s and 1970s (e.g., Blalock 1969, Dubin 1969, Gibbs 1972, Hage 1972, Reynolds 1971). These books attracted great interest in an effort to overcome the radical empiricism of quantitative research, since they tried to reconnect methods and theory along the lines suggested by Popper. The problem was that these books had a cookbook and mechanical character, in essence trying to do for theorizing what standardized quantitative protocols had done for quantitative methods. As a result, they soon fell into disuse—at about the same time that functionalism, which had been allied with positivism since its origins, was also going out of style. The result was that as functionalism faded and as the rather mechanical and restrictive character of theory construction books was exposed, positivism as Comte had conceived of it also declined. Indeed, the very term 'positivism' is now seen as anachronistic or as an epithet for raw empiricism.

4. The Critics of Positivism

Beyond inviting the label of 'raw empiricist,' proclaiming oneself a positivist in the contemporary era is to invite an additional series of attacks on scientific sociology in general and, more specifically, on the prospects for formulating universal laws on the dynamics of the social universe. Some of the most common lines of criticism are enumerated below (Turner 1992).

4.1 The Phenomenological Critique

The world 'out there' can only be known through human consciousness, and so the nature of reality is ultimately filtered through the human mind. Hence, laws of human organization cannot be formulated without a prior understanding of the dynamics of the human mind.

4.2 Hermeneutics

The fundamental nature of the social world can be changed by human agency. Hence, since the nature of social organization is subject to change in its most fundamental character, universal laws about generic processes are, at best, time bound and, at worst, the illusions of social theorists. Moreover, even if a law of social organization is articulated, it soon becomes a tool for changing the fundamental processes described by the law, thereby obviating the validity of the law.

4.3 Critical Discourse

Any set of laws describing the social universe operates as a legitimating umbrella for current social arrangements, many of which are oppressive. Thus, science legitimates the status quo and closes off discourse on less oppressive alternatives.

4.4 Historical Particularism

Social reality is historically specific, reflecting particular confluences of events at a particular moment in time. Hence, any effort to generalize beyond a given historical epoch is doomed to failure, since history constantly remakes the very nature of social reality. The most recent version of this kind of argument can be found in postmodernism, where it is postulated that a fundamentally new reality has emerged, obviating the laws of modernist social organization.

4.5 Scientific Politics

Since science is an organized activity, the knowledge produced by scientists reflects organizational, professional, and political processes as much as reality 'out there.' Thomas Kuhn's (1970) famous work reinvigorated what had previously been known as 'the sociology of knowledge'—an approach revolving around analyses of the social context of knowledge production. Scientific knowledge is, therefore, distorted by the very social processes that organize scientists.

5. The Defense of Positivism

The five lines of criticism listed above make interesting points, and then take them to the extreme. But they represent an *ad hoc* coalition of positions that oppose positivism, seeing positivism as a 'failed epistemology' in the social sciences. Others (e.g., Turner 1985) have sought to defend positivism, addressing the specifics of these points of critique, while trying to re-establish the search of universal and abstract laws about generic social process.

The phenomenological and scientific politics critiques are undone by the fact that the same problems exist for the physical and biological sciences; and yet scientists in these areas still manage to generate verifiable knowledge. The hermeneutical critique poses a unique challenge to scientific positivism in the social sciences, but it is merely assumed by those making this critique that the universe is constantly changing its fundamental character. To make their point stick, critics have to demonstrate that the fundamental forces governing the social world have changed; to assert this to be the case without demonstration with data makes the critique less convincing. The historical particularism critique encounters similar problems, for while it is true that specific events and processes are the product of a particular configuration of events, the underlying dynamics of specific historical conditions may still be generic processes of the social universe. Every specific event is the product of a causally unique historical chain of causes—such is the case in both social and physical universes—but this fact does not mean that the forces that move this chain of causes along are unique historically. Indeed, positivists would argue that these forces are universal, operating in all times and places.

The critical discourse critique is perhaps the most important because, to the extent that generalizations from current empirical reality are viewed as ‘laws’ of how things are, these statements summarizing current empirical reality will assert implicitly that the status quo is the way things *must* be. But the critical theorists run into the opposite problem of postulating ideal states that must be pursued, even if they violate basic laws of human organization and, as a consequence, positively harm people. If ideology were informed by scientific laws, then it is the theory more than the ideology that is useful in reconstructing the social universe in humane ways.

6. The Prospects of Positivism

Positivism will perhaps never be accepted uncritically in sociology, and perhaps not in anthropology and even some branches of psychology and political science. In other social sciences, especially economics, the basic arguments of positivists are not even subject to debate. Positivism was a unique advocacy created by Comte to legitimate the scientific study of human organization by a discipline reluctantly labeled ‘sociology,’ but his advocacy was never accepted by all in the nineteenth or twentieth centuries, and it is unlikely to be accepted in the twenty-first century. Indeed, the term positivism might be best phased out, since it appears to raise such a bright red flag to an *ad hoc* coalition of positions critical of the scientific pretensions of some social sciences. In a sense, theoretical

and empirical work dedicated to discovering and assessing general laws has been relegated to the status of yet another specialty in some social sciences, although one still can find positivistic elements within virtually all specialties of social science. Thus, there is some hope that science will persist in those disciplines, particularly sociology and anthropology, where doubts about science are greatest, but positivism in any guise will never be the only epistemology within sociology and anthropology.

See also: Comte, Auguste (1798–1857); Durkheim, Emile (1858–1917); Empiricism, History of; Hempel, Carl Gustav (1905–97); Hermeneutics, History of; Logical Positivism and Logical Empiricism; Popper, Karl Raimund (1902–94); Positivism, History of; Sociology, Epistemology of; Sociology: Overview; Spencer, Herbert (1820–1903); Theory: Sociological; Wittgenstein, Ludwig (1889–1951)

Bibliography

- Blalock H M Jr. 1969 *Theory Construction: From Verbal to Mathematical Formulations*. Prentice Hall, Englewood Cliffs, NJ
- Comte A 1830–42 *Cours de philosophie positive: Les préliminaires généraux et la philosophie mathématique*. Bachelier, Paris
- Dubin R 1969 *Theory Building*. Free Press, New York
- Durkheim E [1895] 1938 *The Rules of Sociological Method*. Free Press, New York
- Gibbs J 1972 *Sociological Theory Construction*. Dryden Press, Hinsdale, IL
- Giddings F 1901 *Inductive Sociology*. Macmillan, New York
- Hage J 1972 *Techniques and Problems of Theory Construction in Sociology*. Wiley, New York
- Johnson W M 1983 *The Austrian Mind: An Intellectual and Social History, 1848–1938*. University of California Press, Berkeley, CA
- Kuhn T 1970 *The Structure of Scientific Revolutions*, 2nd edn. University of Chicago Press, Chicago
- Mach E 1893 *The Science of Mechanics* [trans. McCormack T J]. Open Court, La Salle, IL
- Montesquieu C [1748] 1900 *The Spirit of Laws*. Colonial, London, 2 Vols.
- Popper K 1959 *The Logic of Scientific Discovery*. Hutchinson, London
- Popper K 1969 *Conjectures and Refutations*. Routledge & Kegan Paul, London
- Reynolds P D 1971 *A Primer in Theory Construction*. Bobbs-Merrill, Indianapolis, IN
- Spencer H 1873 *The Study of Sociology*. Routledge & Kegan Paul, Boston
- Turner J H 1985 In defense of positivism. *Sociological Theory* 3: 24–30
- Turner J H 1992 The promise of positivism. In: Seidman S, Wagner D G (eds.) *Postmodernism and Social Theory*. Blackwell, Oxford, UK, pp. 75–94
- Turner S P, Turner J H 1990 *The Impossible Science: An Institutional Analysis of American Sociology*. Sage, Newbury Park, CA

Zetterberg H L 1965 *On Theory and Verification in Sociology*, 3rd edn. Bedminster Press, New York

J. H. Turner

Possession (Linguistic)

In linguistics, possession is conceived as a relationship of appurtenance between a possessor (P'or) and a possessed (P'ed, possessum, possessee) for which there are specific paradigms of expression in languages of the world. On the conceptual side, this relationship of appurtenance ranges from material possession via kinship and body-part relationships to the domain of abstract relations (someone's ideas, someone's activities, etc.). On the formal side, one distinguishes mainly between attributive possession (e.g., 'my house,' 'the man's house') and predicative possession ('the house is mine,' 'the house belongs to me,' 'I have a house'). However, it is not always easy to distinguish the conceptual domain of possession from other cognitive domains, and also on the formal side there is a continuum of expression rather than some strictly defined set of constructions. For instance, in many languages (e.g., in English or Icelandic) constructions like 'I have ploughed an acre' (expressing an event in time) historically go back to possessive expressions whose literal meaning was 'I have an acre as something ploughed,' and there are also numerous ways of crossing the boundary between attributive and predicative sentential constructions. For instance, many languages many alternate between attributive constructions like German *der Sohn des Mannes* (ART:NOM son:NOM ART:GEN man:GEN, 'the man's son'; where ART = article, GEN = genitive, NOM = nominative) and *dem Mann sein Sohn* (ART:DAT man:DAT his:NOM son:NOM, lit. '(to) the main his son'), where the dative (DAT) was originally some indirect object in predications like *sie entführten dem Mann seinen Sohn* (they:NOM kidnapped ART:DAT man:DAT his:ACC son:ACC, lit. 'they kidnapped (to) the man his son'; where ACC = accusative). These latter constructions (which are typically associated with some psychological involvement on the part of the P'or) are known as *dativus sympathicus*-constructions, or are referred to in the context of possessor raising or external possession. Since psychological involvement also correlates with the relative 'closeness' and inalienability of the possessive relation, it is customary to distinguish between inalienable and alienable possession. Alienable possession normally refers to material belongings as opposed to kinship and body-part relationships, and there are many languages which have standardized ways of distinguishing these domains. In addition to expressing a concept which is not just relevant for linguistics, possessive categories also have

specific functions in the discourse, which suggests that to a certain extent linguistic possession must be differentiated from non-linguistic concepts of possession.

1. Approaches to Possession

1.1 The Literature on Possession

The linguistic literature on possession concentrates on different aspects. Seiler's monograph (Seiler 1983) is still the most comprehensive attempt at dealing with the cross-linguistic evidence for the continuum of constructions in the domain of possession (see Sect. 1.2 below). Heine (1997) focuses mainly on predicative possession and on typical ways in which such constructions develop in the process of grammaticalization. Attributive possession is usually extensively described in traditional grammars under the headings of possessive pronouns and genitival constructions. McGregor and Chappell (1996) is a standard reference to the typology of inalienable possession, and Payne and Barshi (1999) focus on the typology of external possession. Other publications include discussions on the grammatical status of specific items from the domain of possession. For instance, it is sometimes unclear whether possessive pronouns behave like genitival arguments or like adjectives. The discussions in this area are rather technical and must be seen in the context of DP-analysis (for reference see Abney 1987, Löbel 1995). In what follows, this article will focus on some important issues from the typology of possessive constructions.

1.2 Details on Typological Research on Possession

The most difficult problem for the typological description of possession is the great variety of expression and the fact that there is no simple correspondence between a conceptual relationship of possession and traditional linguistic categories like possessive pronouns or possessive verbs (in fact, possessive verbs are cross-linguistically rather rare). For instance, a Latin construction like *mihi est liber* (me:DAT is book:NOM, lit. 'to me is (a) book') contains no 'classical' possessive category, but conceptually it figures as a possessive construction. Conversely, a construction like 'my train' in 'I've got to catch my train' does not really refer to the train which I own, though it does involve a 'classical' possessive pronoun. Finally, there are instances of so-called 'possessor deletion' as in the German *er hob die Hand* (lit. 'he raised the hand') vs. the English 'he raised his hand,' which unlike the German example contains an explicit possessive pronoun.

As a solution to these problems, Seiler (1983) proposes that the expression of possession constitutes an operational program ranging from minimal

explicitness of the relationship (indicativity) to maximal explicitness (predicativity). Indicativity literally means to 'point' at what one may take for granted, while predicativity establishes a relationship by overt means. The 'indicative' and less explicit techniques are mainly characteristic of the attributive pole (with constructions like 'my train'), while the most predicative type of construction is found in the area of possessive verbs (with constructions like 'I own a house'). For 'my train' the relationship is still very vague, while 'I own a house' is very specific about the relationship in question.

In this context, it is no coincidence that inalienable, inherent possessive relationships usually need less 'effort' to be encoded than alienable relationships. For example, in the Amerindian language Cahuilla the expression 'his eye' would just involve the direct juxtaposition of a personal pronoun and the lexical stem for 'eye,' while the expression 'his dog' needs the addition of a 'possessive classifier' meaning 'pet': compare *he-puc* (3.SG-eye, 'his eye'; where SG = singular) vs. *he-ʔas ʔawal* (3.SG-pet dog, lit. 'his pet dog'). In other words, the (inalienably conceived) body-part term 'eye' is inherently relational, such that the possessive relation between 'he' and 'eye' is more or less 'self-understood'; for 'dog' the situation is different, and the relationship between 'he' and 'dog' must be established by some relational item like 'pet.' By way of contrast, one would need far more effort in Cahuilla to use a word like 'eye' in a 'derationalized' sense. Thus, while 'the dog' is simply *ʔawal*, 'the eye' (in some abstract sense) would be *puc-il* (eye-DER; where DER = derationalizer). Seiler lists a number of techniques commonly employed for the expression of possessive relationships, which can all be placed on a continuum between maximal implicitness and maximal explicitness of the relation. These techniques include plain juxtaposition of P'or and P'ed ('N-N-constructions,' see above), connectives (CONN), possessive classifiers (possessive) cases, existential copula constructions, and possessive verbs. The Persian *iza:fa*-construction would be an example for connectives (cf. *ha:ne-je pedar* (house-CONN father, 'the father's house')), and for possessive classifiers see Cahuilla. Latin *domus patris* (house:NOM father:GEN) is an instance of a case construction. A case construction is a more 'explicit' technique than the connective technique in the sense that so-called 'cases' may allow far more oppositions than all-purpose connectives, etc.; thus, apart from using a possessive genitive, many languages also employ locatives in attributive possession (cf. Modern Icelandic *höfundur á mér* (lit. 'the head on me,' 'my head') which is fairly explicit way of specifying how the body-part relates to the P'or). Case constructions also include expressions like the German *der Mann mit dem Hut* ('the man with the hat'). Here the relationship of possession is expressed from a different perspective than *der Hut von dem Mann* (lit. 'the hat from the

man,' 'the man's hat'); in the former construction, the nucleus is the possessor, while in the latter the nucleus is the possessed. (For the change of perspective also see Sect. 1.3 below.)

With the help of existential copulas, the case constructions can often be turned into literally predicative expressions of possession (cf. the German *der Hut ist von dem Mann* (lit. 'the hat is from the man,' 'the hat belongs to the man') and Icelandic *ég er með bíl* (lit. 'I am with (a) car,' 'I have a car with me'). Finally, languages may also employ actual verbs of possession as in 'I have or own a book' and 'the book belongs to me,' etc. This technique allows the greatest number of oppositions (the number of lexical verbs is virtually limitless), and since this technique is the most predicative one in the sense of Seiler, it is hardly 'natural' in the context of inalienable and 'self-understood' possession. For example, clauses like 'I have a father' or 'I have a head' are considered deviant, unless modifications are added (cf. 'I have a nice father,' 'I have a swollen head'). Nevertheless, though in principle inalienable possession tends towards the 'indicative' pole, and alienable possession tends towards the 'predicative' pole, there are always possibilities of making the implicit explicit and the explicit rather implicit. For instance, Indo-European possessive pronouns may go back to locative constructions of much the same type as the ones which are now commonly used for attributive body-part possession in Icelandic (see above), and the Icelandic attributive constructions in question like *höfundur á mér* (lit. 'the head on me') originate in turn from predicative *dativus sympathicus*-constructions like 'he hit (on) me on the head.' Therefore, there appears to be some shift from the more explicit and predicative to the less explicit in the process of grammaticalization. At the same time, as a reaction to a decrease of explicitness, new explicit constructions may be coined. In this context, Seiler claims that indicativity and predicativity represent complementary principles which are in competition with each other in the operation of giving expression to conceptual possession.

Heine (1997), too, is interested in grammaticalization channels, but he concentrates on the lexical sources of (mostly non-attributive) possession and on the ultimate destiny of possessive items in other domains of grammar. As for non-attributive possession ('predicative possession' in a more restricted sense than in Seiler's framework), Heine (1997, p. 47) presents the following typology based on various conceptual 'schemata': 'X takes Y' ('action schema'), 'Y is located at X' ('location'), 'X is with Y' ('companion'), 'X's Y exists' ('genitive'), 'Y exists for/to X' ('goal'), 'Y exists from X' ('source'), 'as for X, Y exists' ('topic'), 'Y is X's (property)' ('equation'). As an example for the action schema Heine mentions Portuguese, while Turkish has a location schema. Somali has a companion schema, while the genitive schema is found again in Turkish. The goal schema is

manifested in Tamil and Hungarian, the source schema manifests itself in Ik (Nilo-Saharan), Cahuilla has a topic schema, and Russian has an equation schema. In many of these languages different schemata coexist. Again, there is no clear separation between possession and very many other kinds of conceptual domains.

For the different schematic types mentioned it is possible to identify very common paths of development and transitions to other categories. Especially postural verbs (e.g., 'stand,' etc.) often develop into locative copulas, and the resulting location schema ('X is located at Y') may develop into a companion schema ('X is with Y'); both the location schema as well as the companion schema then tend to be re-employed as expressions of possession (for companion cases ('be with') cf. Icelandic *ég er með bil* (see above), and for location cf. Finnish *Liisalla on mies* (lit. 'at Liisa is (a) husband,' 'Liisa has a husband')). The same construction types also often grammaticalize to expressions of aspect (Swahili has a 'be with'-progressive, and Icelandic has a 'be at'-progressive). Aspectual categories also commonly develop from possessively interpreted action schemata like 'to take X' (> 'to have X' > 'to have gone'), and aspectual categories may ultimately be interpreted as tense categories ('past tense').

The shortcomings of traditional descriptions of possessive constructions are most noticeable in the area of inalienable possession. The *dativus sympathicus*-constructions are extremely common, and the cases used in possessive constructions for the P'or or noun phrase include not only genitives, datives, locative constructions, and the plain juxtaposition mentioned above, but occasionally even accusatives (e.g., Faroese kinship relations). As a consequence, it is necessary to give special attention to alternatives to traditional grammatical encodings of possession. The book edited by McGregor and Chappell (1996) presents a rich cross-linguistic survey of inalienable possession; common strategies associated with inalienables apart from 'he hit me on the head'-constructions are 'incorporating' strategies like 'he is bigheaded' (vs. ?'he is bigcarred'). Quite typically, the P'or may not be mentioned at all on the possessed noun phrase (cf. the phenomenon of 'possessor deletion' (see above)). But conversely, inherent relationality may sometimes have to be obligatorily marked (lit. 'his hands-his'). As a variant of possessor raising giving prominence to psychologically involved possessors, also different honorifics may be used in the context of body-part relations (cf. Japanese, see Tsunoda 1996).

Payne and Barshi's definition (1999, p. 3) of 'external possession' covers more than the constructions used in possessor raising: here external possession denotes any construction in which the P'or figures as a core argument of the verb and in a constituent separate from that which contains the P'ed. This definition includes applicatives (e.g., Mixean, see Zavala 1999),

'double subject constructions' (Japanese, see Shibatani 1994) and expressions where the P'or is the subject of a verb with an incorporated P'ed (Mohawk, see Baker 1999). What is common to all these constructions is some difference of 'empathy,' 'discourse-prominence,' and 'fore-grounding' in the relationship between P'or and P'ed as compared to the 'standard' possessive constructions. In fact, all kinds of possessive constructions can be ordered in terms of the comparative discourse prominence for P'or and P'ed. With this in mind, it is necessary to be aware of some specifically linguistic functions of possessive categories.

1.3 The Linguistic Side of Possession

Different degrees of discourse prominence can be tested, among other things, with criteria like accessibility for relativization, etc. Languages differ considerably in the extent they 'normally' give prominence to possessors and possessed items, but there is usually variation between various options in the same language, too. Compare English 'her dog *who (= she)' vs. 'this dog of hers, who (= she),' where the ordinary possessive pronoun cannot be taken up in a relative clause, but the phrase 'of hers' can. Normally, there is some continuum of possessor prominence from possessive pronouns via personal pronouns in the genitive to actual arguments of the verb (in the dative or accusative, etc., as in 'possessor raising' constructions). And while the P'ed is totally inaccessible for relativization in 'the blue-eyed man *which (= the eyes) ...,' the P'ed can be relativized in 'the man with the blue eyes which (= the eyes) ...,' etc.

In this context, possessive constructions do not simply express some extra-linguistic concept, i.e., the constructions in question must also be judged with respect to their function in the discourse. This is particularly conspicuous in connection with genitive constructions. It is true that 'normally' the genitive phrase 'expresses' the possessor (cf. 'the man's car'), but sometimes the situation is reversed. Compare German *das Auto des Mannes ist alt* (lit. 'the car (P'ed: NOM) of the man (P'or: GEN) is old') vs. *der Besitzer des Autos ist alt* (lit. 'the owner (P'or: NOM) of the car (P'ed: GEN) is old'). The answer why it is the P'or who is in the genitive in the first example, and why is it the P'ed which is in the genitive in the second ultimately has to do with discourse pragmatics. The noun phrase in the genitive typically encodes a known referent which can be used for the identification of something which is less inherently identifiable. 'Normally,' it will be human P'ors who are 'better known' than inanimate referents or dependent referents like body-parts, etc., but in examples like German *der Besitzer des Autos mit dem Kennzeichen XY wird gebeten, sich mit der Polizei in Verbindung zu setzen* ('the owner of the car with the registration number XY is requested to contact the police') it is the inanimate

referent which happens to be better identifiable. In this case, the P'or will have to be expressed by a special relational noun (*Besitzer* 'owner'), but the genitival phrase is still basically used for identification in the discourse, even though it now encodes the P'ed. These considerations may also explain why many languages have paradigms like 'the man's car,' 'John's car,' 'his car,' 'my car,' while there is often not a corresponding construction like '*I with the car,' '*John with the car' (which in principle just looks at the relationship from another perspective just like 'the man with the car'): 'Normally,' referents like 'I' and 'John' are identifiable without further specification, while dependent items are not default referents in the discourse. Therefore it is quite 'natural' for possessed items to carry some additional marking for making them identifiable in the discourse (e.g., with reference to an animate, 'ego-like' P'or).

Also the distinction between so-called *eius*- and *suus*-possession ('of somebody else' vs. 'of his own') reflects differences of discourse prominence. In Modern Icelandic, the *suus*-possessive *sinn* only refers to prominent participants (usually a subject) within the clause. Conversely, *hans* (*eius*) is used mainly when the pronoun refers to a possessor mentioned outside of the clause. Interestingly, *hans* distinguishes masculine, feminine, and neuter with reference to the P'or, while *sinn* distinguishes masculine, feminine, and neuter only with reference to the P'ed. Hence, the 'distant' P'or will be far better retrievable by means of *hans* than by means of *sinn*, which prefers 'near' and very salient noun phrase referents. Since discourse-relevance is a direct correlate of high animacy and conceptual independence, it is clear why the linguistic treatment of possessive relations will tend to be associated with distinctions like [+/-animate] and [+/-alienable]. As animacy also correlates with control just like any relationship of possession, it is not surprising that there are conceptual shifts from possession to distinctions of agency (cf. 'I have beaten' (agentive) vs. 'I am beaten' (non-agentive)) and vice versa. Agent (AG) and patient (PAT) are also differentiated in nominalized possessor-of-the-act constructions (cf. 'Peter's (AG) beating (action) of the dog (PAT),' or Tongan *ko 'eku* (alienable) *taa'i 'o* (inalienable) *e kuli'i* (lit. 'it is my (active) beating of the (passive) dog,' I am currently beating the dog')). Apart from that, agent and patient relations have also frequently been observed to coincide with differences of imperfective vs. perfective aspect. From this point of view, the typical correspondences between possession, voice, and aspect are not surprising. Like the voice distinction, possession also implies some shift of perspective from an 'ordinary' discourse referent (who is animate and independent) to some referent which is usually inanimate and dependent. It is these shifts of perspective and the many alternative forms of encoding which make possession a highly interesting domain of study for linguistics and cognitive science.

Bibliography

- Abney S 1987 *The English Noun Phrase in its Sentential Aspect*. MIT Press, Cambridge, MA
- Baker M 1999 External possession in Mohawk: Body parts, incorporation, and argument structure. In: Payne D, Barshi I (eds.) *External Possession*. Benjamins, Amsterdam, pp. 293–324
- Heine B 1997 *Possession. Cognitive Sources, Forces, and Grammaticalization*. Cambridge University Press, Cambridge, UK
- Löbel E 1995 Zur kategorialen Bestimmung der Possessiva in der NP/DP: Possessiva als Adjektive. In: Tappe T, Löbel E (eds.) *Struktur der Nominalphrase*. Bergische Universität-Gesamthochschule, Wuppertal, Germany, pp. 54–86
- McGregor W, Chappell H (eds.) 1996 *The Grammar of Inalienability. A Typological Perspective on Body Part Terms and the Part-Whole Relation*. Mouton de Gruyter, Berlin
- Payne D, Barshi I (eds.) 1999 *External Possession*. Benjamins, Amsterdam
- Seiler H 1983 *Possession as an Operational Dimension of Language*. Narr, Tübingen, Germany
- Tsunoda T 1996 The possession cline in Japanese and other languages. In: McGregor W, Chappell H (eds.) *The Grammar of Inalienability. A Typological Perspective on Body Part Terms and the Part-Whole Relation*. Mouton de Gruyter, Berlin, pp. 565–630
- Zavala R 1999 External possessor in Oluta Popoluca (Mixean): Applicatives and incorporation of relational terms. In: Payne D, Barshi I (eds.) *External Possession*. Benjamins, Amsterdam, pp. 339–72

J. Broschart

Postal Systems

Postal systems are administratively coordinated communication networks that originated in antiquity, yet remain important today. During their 4,000-year history, they have changed significantly, making it hard to generalize about their scale, scope, and accessibility. Until fairly recently, for example, postal systems often conveyed people in addition to information and goods.

1. Overview

Most surveys of the world's postal systems are institutional genealogies. Postal systems beget one another in a grand procession of names, places, and dates. Such an approach has the merit of bringing together a wealth of anecdotal material not easily found elsewhere (Scheele 1970). Yet it exaggerates continuity, discourages systematic comparison, and obscures the role of the postal system as an agent of change.

This article takes a comparative institutional approach. The world's postal systems are divided into three groups: imperial, corporate, and national. The section on national postal systems is subdivided into three parts: origins, postal reform, and twentieth- and twenty-first-century developments. The article concludes with a brief discussion of postal systems and social science, with suggestions for future research.

2. Imperial Postal Systems

Imperial postal systems enabled political leaders to conduct diplomacy, plan military campaigns, and oversee routine administration. The most highly developed were those of the Persians and the Romans. The Persian postal system under Emperor Darius (521–486 BC) employed teams of mounted couriers to transmit clay tablets at 14-mile intervals along a 1,600-mile route that stretched from the Tigris River to the Aegean Sea. The Greek historian Herodotus praised its speed and regularity: 'Neither snow, nor rain, nor heat, nor gloom of night stays these couriers from the swift completion of their appointed rounds.'

The Roman postal system, the *cursus publicus*, was established by Emperor Augustus (29 BC–AD 14) and lasted 500 years. Relays of couriers with horse-drawn carts transmitted official dispatches along with government officials and their baggage. Routes extended throughout the empire and, during military campaigns, to generals in the field. Messages were written on papyrus, parchment, or wax-coated tablets, paper remaining unknown in Europe until after AD 1000. The relays explain the etymology of 'postal'; it was derived from *positus*, the past participle of *ponere*, 'to place.' Couriers traveled 50 miles per day—a speed rarely exceeded until the nineteenth century.

Access to the *cursus publicus* was restricted largely to government officials. For this reason, some scholars term it a government communication network, and reserve the term postal system for networks open to the general population. Merchants relied on private messengers, seafarers, or friends. Maritime links were particularly important for Christian apostles such as Paul, who used them to send epistles throughout the Mediterranean world.

3. Corporate Postal Systems

Long after the fall of Rome, imperial postal systems continued to flourish in Asia, South America, and the Middle East. In Europe, however, no comparable institution existed for centuries. When postal systems revived in the 1200s, their organizers were no longer imperial regimes, but corporate bodies, including the University of Paris, the monks of Cluny, and the German butchers' guild.

Corporate postal systems rarely maintained relays of mounted couriers and often lacked the trappings of power. Yet in several respects they marked a distinct advance. Many crossed political boundaries, establishing links *between* states as well as *within* them. Few were restricted, like the Roman *cursus publicus*, to government officials. Open access was facilitated by the recent introduction of paper, a light-weight writing medium that was easy to transport. The policy also helped defray costs, since travel was expensive and many corporate postal systems were self-supporting. Henceforth, the primary constraint for potential postal patrons was financial: postal rates were typically fairly high, limiting access to the well-connected and the well-to-do.

Among the most important corporate postal systems were those established by the Italian city-states. By the end of the thirteenth century, couriers linked Venice with Constantinople, and Florence, Genoa, and Siena with northern France. The most enduring was the *Corrieri Bergamaschie* that Amadeo Tasso established in Bergamo, near Milan, in 1290. This institution much impressed the Hapsburgs, who, in the 1400s, enlisted the Tasso and Delle Torre families to extend it throughout the Holy Roman Empire.

To oblige the Hapsburgs, the families Germanicized their names to Thurn and Taxis. Shortly thereafter, in 1489, Johann von Taxis became the Hapsburgs' first postmaster. Within the next few years, he and his brother Franz von Taxis established postal links between the Hapsburgs' far-flung domains and Rome, France, and Spain, and opened the service to the general population (Dallmeier 1977). Couriers carried a distinctive post horn, an instrument that remains to this day a symbol of the German Deutsche Post. Although the rise of the nation-state greatly restricted its scope, the Thurn and Taxis postal system survived until 1867, when it was bought out by the Prussian government (Behringer 1990).

4. National Postal Systems

4.1 Origins

Corporate postal systems stymied the rulers of the emergent nation-states in several ways. Most obviously, they created an unregulated communication channel that threatened political stability. In addition, they deprived rulers of a useful tool with which to monitor the flow of information and ideas. In response, nation-states throughout Europe, led by France and England, established postal systems under tight government control (Coase 1955).

The French and English postal systems originated, like the imperial postal systems of antiquity, as government communication networks. In France, the founder was Louis XI (1461–83), and in England,

Henry VIII (1509–47). In both countries, the early 1600s marked a decisive turning point. Within a few decades, both governments granted access to the general population, instituted a monopoly, established a rate structure and delivery schedule, and organized a surveillance mechanism to monitor their subjects' correspondence: the *cabinet noir* in France; the secrets office in England (Vaille 1946–55, Robinson 1948). Shortly thereafter, the English government mandated a major rate increase, transforming its postal system into a revenue source, which it would remain until the twentieth century (Dauntton 1985).

Beginning in the 1600s, national postal systems did much to hasten the emergence of the newspaper—at the time, a new form of political communication. This relationship was particularly close in England, British North America, The Netherlands, and Prussia. By permitting the regular transmission of information on current events, postal systems became, in effect, an early broadcast media. Having gained control of the means of communication, governments used it to foster a distinctive world view. In so doing, they helped forge the national identities that are such a fundamental feature of the modern world (Mueller 1986).

Nowhere was the relationship of the postal system and the newspaper closer than in the USA. In the 1790s, following the adoption of the federal Constitution, the American government invested its postal system with an educational rationale far more expansive than the fiscal rationale that had previously prevailed. The new mandate, codified in the Post Office Act of 1792, hastened the circulation of newspapers by conveying them at extremely low rates while enormously expanding the number of localities to which they could be sent (John 1995). A landmark in postal history, this policy may well have marked the first time a government deliberately subsidized the transmission of a particular kind of information.

By 1828, the American postal system, with almost 8,000 post offices, was larger and more geographically extended than any postal system in Europe. At this time, the USA boasted 74 post offices per 100,000 inhabitants, Britain 17, and France four (John 1995). Prior to the coming of the railroad and telegraph, the American government had established the preconditions for the disembodied public sphere that French traveler Alexis de Tocqueville (1835, 1840) would explore so perceptively in *Democracy in America*.

4.2 Postal Reform

Victorian commentators often hailed the postal system of their day as a fundamentally *new* institution that was little older than the railroad and telegraph. For the majority of the population, this was often true. Prior to the 1840s, postal rates remained so high that letter writing was, in practice, confined to the few.

Even in the USA, if a farmer or artisan received anything in the mail, it was more likely to be a newspaper than a letter.

In both Britain and the USA, the high cost of letter postage hastened the establishment of clandestine letter-delivery services in direct violation of the postal monopoly. Following the advent of the railroad in the 1830s, these operations expanded greatly, hastening the major rate reduction known as postal reform.

Postal reform began in Britain, where it was championed by educator Rowland Hill. Hill recognized that high letter rates encouraged illicit mail carriage, and posited that a major reduction would hasten a massive expansion in letter writing. To buttress his argument, he demonstrated in a famous pamphlet, *Post Office Reform: Its Importance and Practicality* (1837), that rates bore no relation to expenses. All but a tiny fraction of the cost of delivering a letter, Hill calculated, derived from its *routing* through the system rather than from its *conveyance* from sender to recipient.

Hill prevailed and, in 1840, the British government simplified the rate schedule and reduced the basic letter rate to one penny. Hill styled his reform uniform postage; it was commonly called penny postage. Hill's reform greatly increased mail volume, just as he had predicted. In addition, it hastened the advent of several familiar postal practices, including mandatory prepayment, mail boxes, and the postage stamp (Dauntton 1985).

Many governments followed Hill's lead. The USA lowered its letter rates in 1847 and 1851; France in 1848. Japan, eager to emulate the best practices in the West, established in 1870 a postal system modeled closely on Britain's (Westney 1987). Increased mail volume, in conjunction with the advent of the ocean-going steamship, hastened the proliferation of transnational postal treaties. These diplomatic initiatives were regularized in 1878 with the establishment of the Universal Postal Union (UPU), today a branch of the United Nations. Closely related to postal reform was the substitution of the railroad for the stagecoach as the preferred mode of conveyance. By the 1830s, stagecoach service had attained a high level of development, especially in Britain and France. Interestingly, Rowland Hill based his famous cost calculations on the stagecoach, rather than the railroad, which had yet to be widely adopted. Stagecoach speeds often averaged 10 miles per hour, a standard rarely matched since the demise of the *cursus publicus* of imperial Rome. The use of the railroad to convey mail posed various scheduling dilemmas and could increase postal costs; in the USA, for example, it quickly ran the postal system into debt. Yet it proved indispensable, greatly increasing carrying capacity and revolutionizing postal routing. In Britain, the USA, and most industrialized nations, it was long customary to sort the mail on-board moving trains, a practice known as railway mail.

Postal reform ushered in a golden age of letter writing. In many countries, the exchange of holiday greeting cards became a popular custom. Soldiers' letters emerged, especially in wartime, as a notable mode of literary expression. Although letter writing would, in the twentieth century, face competition from telephony, the habit persists, and, since the popularization of E-mail, may well be enjoying a revival.

In the decades preceding World War I, each of the world's leading postal systems—Britain, France, Germany, and the USA—undertook major new responsibilities. Each established a parcel post and instituted a rudimentary banking service known as postal savings. And each, with the exception of the USA, took over the telegraph and telephone, establishing a ministry of Posts, Telegraphs, and Telephones, known in France and elsewhere as the PTT (Bertho 1981, Perry 1992). This expansive era in postal history reached its apogee around 1900, when it was common for journalists to hail the postal system as the greatest business in the world (Fuller 1972).

4.3 Twentieth and Twenty-first Century Developments

In both the developed and the developing world, postal systems remain indispensable elements of the information infrastructure (Sinclair 1984). In the latter part of the twentieth century, however, much changed. Beginning around 1970, rapid changes in communications technology led many governments to re-evaluate the union of the postal system, the telegraph, the telephone. Deregulation followed. In Britain, the break-up came in 1981, and in France, in 1991.

The Internet poses a further challenge, yet, in a sense, it is nothing new. Since the mid-nineteenth century, postal systems have co-existed with earlier modes of electric communications. The USA is the originator not only of the Internet, but also of 40 percent of all the world's mail. In the early twentieth century, pundits routinely predicted that telegraphy would soon render mail delivery obsolete, yet telegraphy has been everywhere supplanted, while postal systems thrive.

Additional challenges stem from the emergence of new competitors, e.g., DHL in Europe, Federal Express and UPS in the USA. Express firms rely on improvements in air travel to compete head-on with the national postal systems in key high-value markets. In response, many governments have re-examined—and, in several cases, narrowed—the scope of the postal monopoly (Campbell 1994).

Postal systems have met these challenges in various ways. Many have reinvented themselves as public corporations, a transformation sometimes called privatization. If this trend continues, postal systems in the twenty-first century may share less with the national postal systems out of which they evolved than the corporate postal systems of medieval Europe. The

French postal system, known today as La Poste, has been a publicly traded corporation since the 1970s. The British postal system, the Royal Mail, has aggressively sought new markets. So, too, has the Deutsche Post, which issued its first securities in 2000, and the US Postal Service, which has not been a government department since 1970 and no longer receives a government subsidy.

5. Postal Systems and Social Science

Scholars agree that the printing press was an agent of change. Only recently, however, have postal systems been studied in a similar way. The last edition of the *International Encyclopedia of the Social Sciences* (1968) contained no entry on the postal system, and no references to it in the index.

The rise of the new institutionalism in history, sociology, and political science bodes well for postal studies. So too does the vogue for French literary theory, with its focus on the complexity of the communication process (Siegert 1993). Developments in information technology promise further insights, offering new perspectives on modes of communication that are seemingly being displaced. For postal scholars, Hegel's dictum holds true: the owl of Minerva flies at dusk.

Research opportunities abound. Comparative data remain scarce, especially for the period prior to 1840. Without it, it is hard to frame credible generalizations about what postal systems did and how they worked. Topics that invite exploration include the development and significance of scheduling; censorship and surveillance; postal privacy; communication networks and the rise of nationalism; challenges to the postal monopoly; changing rationales for postal systems; and the shifting meaning of the equal access ideal.

See also: Communication and Democracy; Communication and Transportation, History of; Communication: Electronic Networks and Publications; Communication: Geographic Aspects; Internet: Psychological Perspectives; Mass Communication: Empirical Research; Mass Communication: Normative Frameworks; Mass Communication: Technology; Telegraph; Telephone

Bibliography

- Behringer W 1990 *Thurn und Taxis: Die Geschichte ihrer Post und ihrer Unternehmen*. Piper, Munich
- Bertho C 1981 *Télégraphes et Téléphones de Valmy au Microprocesseur*. Librairie Générale Française, Paris
- Campbell R M 1994 *The Politics of the Post: Canada's Postal System from Public Service to Privatization*. Broadview Press, Peterborough, Canada
- Coase R H 1955 The postal monopoly in Great Britain: An historical survey. In: Eastham J K (ed.) *Economic Essays in*

- Commemoration of the Dundee School of Economics, 1931–55. William Culross, Coupar Angus, UK, pp. 25–37
- Dallmeier M 1977 *Quellen zur Geschichte des europäischen Postwesens, 1501–1806*. 3 parts. Lassleben, Kallmünz, Germany
- Daunton MJ 1985 *Royal Mail: The Post Office since 1840*. Athlone, London
- Fuller WE 1972 *The American Mail: Enlarger of the Common Life*. University of Chicago Press, Chicago
- John R R 1995 *Spreading the News: The American Postal System from Franklin to Morse*. Harvard University Press, Cambridge, MA
- Mueller M 1986 The currency of the word: War, revolution, and the temporal coordination of literate media in England, 1608–1655. M.A. thesis, University of Pennsylvania, Philadelphia, PA
- Perry CR 1992 *The Victorian Post Office: The Growth of a Bureaucracy*. Royal Historical Society, London
- Robinson H 1948 *The British Post Office: A History*. Princeton University Press, Princeton, NJ
- Scheele CH 1970 *A Short History of the Mail Service*. Smithsonian Institution Press, Washington, DC
- Siebert B 1993 *Relais: Geschichte der Literatur als Epoche der Post, 1751–1913*. Verlag Brinkmann & Bose, Berlin; 1999 *Literature as an Epoch of the Postal System*. Stanford University Press, Stanford, CA
- Sinclair J 1984 *Uniting a Nation: The Postal and Telecommunication Services of Papua New Guinea*. Oxford University Press, Melbourne, Australia
- Vaillat E 1946–55 (1947) *Histoire Générale des Postes Françaises*. 7 Vols. Presses Universitaires de France, Paris
- Westney DE 1987 *Imitation and Innovation: The Transfer of Western Organizational Patterns to Meiji Japan*. Harvard University Press, Cambridge, MA

R. R. John

Postcolonial Geography

Historically speaking, a special relationship existed between geography and the making of empires. Pragmatic geographies of orienteering, surveying, mapping and description were the frontier technologies of the operations that brought distant territories under the control of metropolitan centers. As historian Edward Said famously noted in his book *Culture and Imperialism* (1993, p. 271), imperialism is ‘an act of geographical violence through which virtually every space in the world is explored, charted, and finally brought under control.’ At no time was the relationship between geography and colonialism more intimate than during the span of modern European expansion. Such worldly colonizations might be understood simply as metropolitan appropriations of the resources of other peoples and lands. Yet this material self-interest was motorized by a complex set of values that made these often audacious and violent appropriations make perfect moral sense to those whose interests were being served. Not least, deeply

embedded assumptions about racial superiority and apportionings of civility, gave legitimacy to these territorializations. The end of the nineteenth century marked the close of an era in which the discipline of geography was intoxicatingly enlivened.

The modern discipline of geography lives with this past. For a dwindling few it remains a source of nostalgic pride. For many it is source of considerable ambivalence. It is, in part, out of this ambivalence that some of the work that might claim the title ‘postcolonial geography’ emerges. The global authority assumed by imperial geographical knowledges was laid to rest finally by the self-reflexive shock wave that passed through the discipline in the mid-1980s. Geographers who had traditionally marauded across class, gender, race, and other divides began to interrogate the politics of representation associated with such reportage (see special issue of *Environment and Planning D: Society and Space* 1992). This was a symptom of a discipline that was hurriedly retreating from its own past.

To understand what postcolonial geographies are it is essential to understand what is meant by the term postcolonial. As the prefix suggests, postcolonialism refers to the state that comes after colonialism. Most simply, this is the political condition that arrives when a colony is granted independence from its colonial rulers. Many have questioned the ‘post’ of such postcolonialism. The movement from colony to independent nation–state is not a sharp break with the past. Postcolonial states remain complexly entangled in their colonial pasts, just as they are constantly negotiating various kinds of neocolonialism, be it in the form of geopolitical influence, cultural effect, or economic leverage. The persistence of various kinds of colonialism has led various scholars to note that the term postcolonialism is prematurely celebratory. Others are similarly skeptical, suggesting that the term does nothing more than describe a condition of a newly diasporic intelligentsia.

Those geographies to claim the term ‘postcolonial’ reflect a use common to the social sciences and humanities more generally. Here the term refers to an ongoing set of relationships and processes inaugurated by the experience of colonization. Postcolonialism is not then a state that arrives once colonialism ends, but a struggle that comes into being because of, and alongside, the fact of colonization. In this sense, colonial and postcolonial formations are cohabitants. For example, colonial authorities were always dealing with postcolonial effects produced by this cohabitation. This might have included certain subtle, but nonetheless unsettling hybridizations. More starkly, it might have included counter-colonial formations through which the colonized sought to reinstate their authority.

Not surprisingly, many different kinds of formations, periods, and circumstances have come to be described as postcolonial. Such elasticity in the term

postcolonial compromises both its descriptive and analytical worth. Yet postcolonialism is still a useful term to apply to a particular intellectual (and political) stance. Indeed, postcolonial studies are defined as much by their theoretical field of influence as they are their empirical concerns. Nowhere is this more evident than in the discipline of geography. Postcolonial geographies are varied in their subject matter, but may be drawn together as a loosely unified set of geographies intent on bringing into view a set of phenomena that work against colonial authority. This might include evident ruptures in the apparent coherence of colonial occupation, it might include the hybrid outcomes of cohabitation, the so-called 'minor' knowledges and alternative terrains of the colonized, and postcolonial formations that aim to wrest back control of territories.

Much of the current work within postcolonial geography, as with postcolonial studies more generally, is based on revisiting the site of colonialism in order that the silences that exist in the colonial record may be uncovered. This recovery project, it is assumed, will lead to a radical rethinking of the forms of knowledge and social identities authorized under colonialism. Historical geography has been much enlivened by this critical return to the imaginaries, logics, and practices of colonialism. In 1993 Livingstone (1993) published his elegant account of the geographical tradition, within which he paid particular attention to the relationship between European geography, race, and empire. In the same year Driver (1992) published an important article that 'unmasked' modern geography's relationship to European colonization. By demonstrating the way geographical knowledges framed other worlds, his paper drew the analytic terms set by Edward Said's landmark contribution, *Orientalism* (1978), into the heart of a rapidly postcolonizing discipline.

Critical appraisals of geographers and the knowledges they produced have continued as an important theme within the discipline. Harley's (1988) work on the relationship between cartographic knowledges and power, is exemplary in this regard. His re-appraisal of cartographic practices and colonialism contributed to an on-going scholarship (both within the discipline and outside) on cartographies and colonial power. The spatial emphasis in geography's critical engagement with the colonial archive is also evident in accounts of other, more populist, visual representational fields. Photography, for example, played an important role in bringing home the experiences, peoples and lands of those faraway new worlds (Ryan 1997).

Geographical practitioners, institutions, knowledge systems, and pedagogies of various kinds have succumbed to the scrutiny of this autocritique. Particular attention has been given to the systems of knowledge used to understand, order and benefit from the often exotic nature found in the new worlds. These knowl-

edge systems made meaningful peoples and lands that, from a eurocentric worldview, were so radically different they could at first only be understood as 'savages' or 'wastelands.' Two showcase collections of studies detailing diverse examples of geography's relationship to various imperialisms (old and new, European, and other) appeared in the mid-1990s (Godlewska and Smith 1994, Bell et al. 1995).

Delivering a gender sensitive perspective to the geographical critique of imperialism has shown that this is seemingly masculinist project, is one in which women (colonial and colonized) are complexly embedded, both as fellow adventurers and as objects of desire. The letters and journals of women who traveled to colonized lands have been an important source of information for rich scholarship on the women in the colonial project.

The global scale expansions of imperialism ground themselves in micro-geographies of occupation and domestication. As Jacobs (1996, p. 19) notes, the pure idea of imperialism 'tumbled into the fractured and erratic everyday practices of the personalities ... who were forced together in the making of colonies.' One of the key collective contributions of the now vast body of critical colonial geographies has been to point to the contingencies of colonialism as a located, place-specific practice. For some time a key concern of this work has been the way relations of differentiation and domination have been enacted in and through space, and how colonized subjects subvert or negotiate those spatial controls.

A significant subset of critical colonial and postcolonial geographies concerns the city. Although not a geographer, the work of Anthony King has been highly influential in shaping the direction of this inquiry within the discipline. His theory of colonial urban development outlined the economic, political, and cultural processes that gave rise to and defined the form of the cities of colonial territories. Colonial cities were important sites in the transfer of modern capitalist values to new worlds and functioned as important centers of power for administering indigenous people and resources. In these cities, town-planning ordinances about cleanliness and sanitation became the mechanism through which colonial assumptions about racial difference were translated into on-the-ground technologies of segregation and exclusion. Jacobs (1996) extends this urban postcolonial scholarship by documenting the ways in which imperial pasts and postcolonial struggles continue to shape contemporary cities. More recent scholarship has turned its attention to the way imperial geographies impacted upon the townscapes and spatial arrangements of the metropolitan cities that stood at the heart of empires. Driver and Gilbert (1999) offer an exemplary collection on the history of cross-mappings between empire and modern European cities.

The geographies accounted for thus far largely confine their attention to the archive or to historic

landscapes and territorializations. The intellectual project of making a critical return to the scene of colonialism is undertaken in the belief that it contributes to a decolonization of the discipline. Such a return can and does have its postcolonizing value, but there remains a certain ambiguity around just how postcolonial these modern adventures in critical colonial geography might be. Not least there is a tendency in these geographies for the colonized subject (the 'other') to be brought into view simply by way of the colonial discourses that framed them. Can attending to the colonial scene with a postcolonial lens dislodge the effects of that history? Or does it simply re-inscribe the colonial moment as the master referent of all that follows? Colonial constructs, practices, and landscape transformations have significant after effects, they flow into present times both sustaining vectors of domination and exploitation and providing the site against which alternate identities, places, and ways of being in the world are asserted. It is to the geographies that attend to the contemporary scene of postcolonialism that we now turn.

Racism and racialized difference rightly dominate the critical analysis of imperialism and the making of colonies, for it was on the basis of this register of difference that many of the most audacious imperial appropriations have been made. Despite this, there is an intellectual distance between the radical geographical tradition of studies of race and racism and geographies that describe themselves as 'postcolonial.' Some recent work has begun to converge around contemporary notions of whiteness, the postcolonial politics of antiracism and the emergence of new forms of postcolonial racism.

With geographical knowledges and technologies playing such a central role in the making of colonial territories, they have come to play an important part in the unmaking of these territories. There is, for example, a growing literature within geography on the use of cartographic technologies, place renaming, and alternative geographical knowledges, in various forms of postcolonial nationalism and indigenous rights movements, as well as newly commodified forms of conquest. Similarly, geographers have begun to show how modern environmentalisms are complexly entangled with colonialist assumptions about nature and race: bringing forth new kinds of (romantic) racisms and exclusions alongside of new possibilities for political alliance.

An exemplary body of self-consciously decolonizing geographical writing is the scholarship that has emerged around the racist spatial logic of settlement in Southern Africa. That lingering colonial construct of apartheid gave rise to stark geographies of differentiation and disadvantage based on racially defined 'group areas.' These racist spatialities called into being many passionately political geographies of reform. Similarly postcolonial geographies are emerging in the local scholarship of other former colonial territories and

settler states. Some of these geographies continue to remind us of the tenacity of colonial and neocolonial constructs and the way in which they are embedded in the physical and ideological landscapes of places and nations. This might include the ways in which the landscapes and identities of those once colonized are now being absorbed into new touristic circuits or commodity markets. Other postcolonial geographies genuinely reorient the concerns and subject matter of the discipline. They may install radically alternative evaluations of land and resources, that are directed away from dominant models of use and in so doing give legitimacy to previously marginalized ways of being in the world. They may chart hopeful examples of resistance to domination, be it in the form of minor subversions or stark oppositions. They may even be so optimistic as to speculate about new models of co-existence. What all of these emerging geographies of postcolonialism reveal is how specific historical and geographical contexts have delivered a peculiar political and moral imperative to the project of decolonizing the disciplinary gaze.

See also: Colonialism, Anthropology of; Colonialism: Political Aspects; Geography; Globalization and World Culture; Imperialism, History of; Imperialism: Political Aspects; Racism, History of; Subaltern History

Bibliography

- Bell M, Butlin R, Heffernan M (eds.) 1995 *Geography and Imperialism, 1820–1940*. Manchester University Press, Manchester, UK and New York
- Blunt A, Rose G (eds.) 1994 *Writing Women and Space: Colonial and Postcolonial Geographies*. Guilford Press, London and New York
- Bonnett A 2000 *White Identities: Historical and International Perspectives*. Prentice Hall, Harlow, UK
- Crush J 1994 Post-colonialism, de-colonization, and geography. In: Godlewska A, Smith N (eds.) *Geography and Empire*. Blackwell, Oxford, UK, pp. 333–50
- Driver F (ed.) 1992 Geography's empire: Histories of Geographical knowledge. *Environment and Planning D: Society and Space* 10: 23–40
- Driver F, Gilbert D 1999 *Imperial Cities: Landscape, Display and Identity*. Manchester University Press, Manchester, UK and New York
- Godlewska A, Smith N (eds.) 1994 *Geography and Empire*. Blackwell, Oxford, UK
- Harley J B 1988 Maps, knowledge and power. In: Cosgrove D, Daniels S (eds.) *The Iconography of Landscape*. Cambridge University Press, Cambridge, UK, pp. 277–312
- Jacobs J M 1996 *Edge of Empire: Postcolonialism and the City*. Routledge, London and New York
- Livingstone D N 1993 *The Geographical Tradition*. Blackwell, Oxford, UK
- Marcus G E (ed.) 1992 *Environment and Planning D: Society and Space* 10(5) (whole issue)
- Phillips R 1997 *Mapping Men and Empire: A Geography of Adventure*. Routledge, London and New York

- Robinson J 1995 *The Power of Apartheid: State, Power and Space in South African Cities*. Butterworth-Heinemann, Oxford, UK and Boston
- Ryan J R 1997 *Picturing Empire: Photography and the Visualization of the British Empire*. The University of Chicago Press, Chicago
- Smith D M (ed.) 1992 *The Apartheid City and Beyond*. Routledge, London and New York
- Sparke M 1994 White mythologies and anemic geographies—a review. *Environment and Planning D: Society and Space* 12: 105–23
- Yeoh B S A 1997 *Contesting Space: Power relations and the Urban Built Environment in Colonial Singapore*. Oxford University Press, Oxford, UK

J. M. Jacobs

Postcolonial Law

Postcolonial theory is largely associated with South Asian scholarship, subaltern and literary studies, analyses of resistance, and a general rethinking of a dominant European historiography that places the 'West' at the center of the world. Postcolonial theory plays a particularly important function in its acknowledging and recovering the ongoing significance of colonized peoples in shaping the epistemologies, philosophies, practices, and shifting identities of what is taken for granted as dominant Western subjects and subjectivities. Most importantly in the context of law, postcolonial theory, and the various colonial sites through which it is inscribed and substantiated, punctuates the degree of violence and the technologies of power involved in understanding concepts such as modernity and capitalism as well as contemporary state and trans-state nationalisms.

That being said, postcolonial theory rarely considers the importance of law and imperial governance (however, see Baxi 2000, Spivak 1999). Likewise, legal scholarship has largely ignored the significance of a postcolonial perspective (see Darian-Smith and Fitzpatrick 1999). Certainly sociolegal scholarship has gone some way to bring to the fore some of the implications of gender, ethnicity, race, class, age, religion, and so on, in the formulating of the rule of law and the maintenance of its legitimizing practices. However, the particularly uncomfortable position offered by a postcolonial perspective, which places side by side the oppressed and the oppressor, and their enduring psychological and material interdependence, is generally treated as a radically different line of inquiry from that which seeks to analyse economic and political inequalities promulgated through a particular legal system. What, after all, do cultural and literary studies that focus on the colonial situation have to do with an empirical study of law in places such as the United States? In an attempt to dismantle this prevailing attitude of mutual exclusivity between

postcolonial and legal studies, discussed below are some of the insights postcolonial theory brings to bear on an understanding of contemporary legal processes in order to identify, analyse, and assess what is meant by the term 'postcolonial law.'

1. Defining a 'Postcolonial' Perspective

Throughout the late 1980s and 1990s, a considerable body of scholarship devoted to issues of postcoloniality emerged. And with this proliferation of research in Anglo/European universities, there was, and is still, much debate over the meaning and scope of postcolonial terminology (see Schwarz and Ray 2000, Loomba 1998, Williams and Chrisman 1994). Despite these ongoing deliberations, it is helpful to turn to the most significant implications of a postcolonial perspective which mark it as distinctly different from, yet complementary to, other contemporary theoretical investigations. This process will help to flesh out the contours of postcolonial studies and, at the same time, underscore its intellectual lineage.

Postcolonial studies emerged initially among Indian and British scholars in the wake of World War II and in an era of general demise of colonial regimes. Spurred on by radical Western thinkers such as Marx, Nietzsche, Althusser, Fanon, Jameson, Foucault, and Derrida, Edward Said's important publication *Orientalism* (1978), these scholars of historiography critiqued Eurocentric and Indocentric interpretations alike for their assumptions of superiority in European countries vis-à-vis the past and future of Indian people and a modernist concept of the Indian nation (Prakash 1990, 1992, O'Hanlon and Washbrook 1992, Chakrabarty 1992).

Building on Marxist, poststructuralist, and feminist theories, postcolonial scholars have played a major role in bringing to the fore the chronological and psychological relations between the former colonized and colonizers, whom, they argue, cannot be understood except in conjunction with each other (see Gandhi 1998). Postcolonial theorists do not claim that colonialism was experienced in the same way under different regimes, just as they recognize that today colonialism operates in very different ways from its earlier configurations. Nonetheless, while paying attention to the details of specific contexts, postcolonial scholars agree that, in order to understand all contemporary histories of peoples and places, irrespective of whether there historically existed in any given site an explicit colonial regime, it is crucially important to remain aware of the enduring presence of former colonial discourses that posit 'civilized,' 'progressive,' and 'just' Europeans against 'barbaric,' 'static,' and 'unethical' indigenous populations.

According to these scholars, colonial assumptions of Western superiority endure across time and across attempts to blur racial and ethical boundaries. While racism is not equivalent to colonial

regimes, categories of racial difference were used by colonial governments to varying degrees as strategies of power and control over indigenous peoples (Stoler 1989, Dirks 1986). Today, despite claims of increasing acceptance of multiculturalism and cultural diversity, racial categories and differences continue to exist, though often in less overt manifestations. And, according to many postcolonial theorists, these boundaries of difference are insurmountable given the psychological intersubjectivity between former colonizers and colonized which constantly necessitates a desire for re-representations of difference (see Fitzpatrick 1999). Drawing on a variety of theoretical perspectives—including an Hegelian master/slave dialectic, phenomenological essentialism, and psycho-analytical insights gleaned from Fanon and Lacan—postcolonial theorists such as Homi Bhabha, Gayatri Spivak and Benita Parry bring to the fore complex understandings of how oppressed peoples resist and seek empowerment (see Gates 1991).

In recognizing processes of mutual desire and negation between the 'master' and 'slave,' postcolonial scholars are forced to recognize the profound irony in attempting to present alternative historical narratives and identities other than those conventionally supplied by the West (Trivedi and Mukherjee 1996). The very moment of self-determination by peoples formerly colonized requires the adoption of European knowledge such as notions of progress and development, concepts of self and personhood, as well as forms of government, state-building, and, perhaps most importantly of all, law. Thus, according to a postcolonial perspective, all assertions of freedom and self-awareness require elements of mimicry and voyeurism. As Douglas Robinson has noted:

Postcolonial or subaltern scholars claim it is at once essential and impossible to forge a 'new' postcolonial identity: *essential*, because those colonial constructs were at once alien and negative, because they came from the outside and destroyed much of value in the indigenous cultures, and because an effective post-colonial politics requires the development of more positive indigenous visions: but also *impossible*, because colonial discourse continues to inform even these postcolonial attempts to break free of it, and tends to condition even the imagination of a 'new' (postcolonial) identity along 'old' (colonial) lines. (Robinson 1997, pp. 19–20)

A prominent postcolonial theorist, Dipesh Chakrabarty, has responded to the ironies of alterity by calling for the 'provincializing,' or decentering, of Europe and European epistemological knowledge. Chakrabarty (and others) argue that non-Western knowledge has been historically ignored and precluded from historiographical accounts of humanist understanding and intellectual endeavor (Chakrabarty 1992, 2000a). In an effort to critique 'the "Europe" that modern imperialism and (third world) nationalism have, by their collaborative venture and violence,

made universal,' Chakrabarty urges scholars 'to write into the history of modernity the ambivalences, contradictions, the use of force, and the tragedies and the ironies that attend it' (Chakrabarty 1992, pp. 20–21).

2. Postcolonial Law

Nowhere in the histories of modernity has the 'use of force, and the tragedies and the ironies that attend it' been so obvious as they are in the context of law. Law, in a variety of ways, was the formal mechanism and institutional frame through which colonial governments oppressed and controlled indigenous peoples. Of course, there are many varieties of imposition and reception of law, a few even categorized as voluntary. In some cases, such as in Africa (Mamdani 1996), this involved coopting native chiefs and traditional procedures of arbitration and dispute resolution. In other cases, such as Australia, the British declared the continent *terra nullius* or vacant, and on the basis that native populations were thought to be less than human they were deemed to have no law at all. As a result, traditional methods of peacekeeping were overlooked or deliberately obliterated. Law, emblematic of European rationalism, universal objectivity, individual property rights, and sovereign state authority, provided the justification and the discourse of capitalism, and its necessary corollary of social inequality, be it based on the grounds of gender, class, or ethnicity.

It is impossible to separate colonial from postcolonial law since each developed in collaboration and in conjunction against and through each other. The metropole has always been connected with its colonial peripheries, be these geographically distanced outposts beyond state boundaries, or socially and politically isolated enclaves (such as native reservations) within it. The process of decolonization and the emergence of new states throughout the twentieth century clearly demonstrates the extent to which independent nations have had to adopt modern forms of state building, and the institutions, bureaucracies, constitutions, and legal ideologies of their former oppressors in order to participate in national and international legal communities. The rule of law, the handmaiden of Western capitalist enterprise, exemplifies the postcolonial dilemma that requires new states and non-Western states to mimic and adopt, at least to a certain extent, the legal structure of so-called democratic society.

3. Postcolonial Law and Globalization

Prevailing discussions about law and globalization epitomize the ironies presented by postcolonial law. For instance, The World Bank, the United Nations, the European Union, GATT, NAFTA, and other international legal arrangements all require that, in

order for a country to participate in a global political economy, it must demonstrate commitment to, and adherence with, the foundational values of modern Western law. This does not mean that Western law, or more specifically Anglo/American law, remains impervious to the challenges and collaborations of interacting with different legal values and processes. In a postcolonial world, just as in the colonial situation, there are always ongoing modifications and appropriations between legal systems and the various cultural values that legal processes represent. One only has to think of human rights to appreciate the extent to which its apparently universal actual application has to be constantly modified according to prevailing social, political, and economic values existing within particular peoples and places. That being said, however, the dominance of Anglo/American law is overpowering and not easily destabilized.

An assumption of Western superiority is evidenced by a majority of United States and European socio-legal scholars who continue to treat the rule of law as a discrete entity or thing, much like the nation-state (which law has historically served), and religion (which law has in effect replaced), and not as a dynamic product of historically contested cultural (and colonial) interactions. Thus up to the end of the twentieth century, in arguments both for and against the enduring significance of the nation-state in the increasing context of globalization, the meaning of law is taken as a given, and it is only its spheres of centralized legal activity and influence that are being brought into question. As a result, legal analysts tend to look only at the privileged domains of legal interaction amongst lawyers, judges, business people, and entrepreneurs, and not from the perspective of ordinary people (e.g., Dezalay and Garth 1996). An important point is that what happens in the offices of international and state law firms, as well as bureaucratic and regulatory institutions, may have very little to do with the operation of law on the ground in the barrio, reservation, schoolyard, parking lot, hospital, jail, home-cleaning service, bank, and employment office. These realms of legal interaction, while not so overtly important in macroeconomic and political terms, are, nonetheless, where the majority of ordinary people interface with the law and understand its meanings. And it is in these realms of everyday legal interaction where challenges to static legal ideologies and definitions coagulate, solidify, and, in some cases (such as the riots against the World Trade Organization in Seattle in 2000), become articulated.

4. Challenges and Limits of Postcolonial Law

To really accept the challenges and resistances presented by postcolonial law, be it from within or from without the boundaries of the nation-state, requires that legal scholars accept that there is no universal legal code and legal objectivity. This in turn would

mean that the relative dominance of one set of legal values over another would be a constant topic of review and analysis. The naturalized centrality and superiority of an American/European legal perspective would be dislodged and, borrowing Chakrabarty's terminology discussed above, necessarily 'provincialized.' Given the relations between law, capitalism, and a global political economy, it is no surprise that Western legal scholarship has largely ignored, some would argue deliberately, the long-term implications of the presence of postcolonial law.

That being said, the ironies and limits of a postcolonial perspective applies to the context of law as it does elsewhere. Desires to elevate the singularity, independence, and significance of peoples formerly colonized, and their particular legal ideologies and practices, must also recognize the extent to which postcoloniality is a product—culturally, politically and economically—of the colonial enterprise. There is no obvious solution to how to move beyond the inherent limitations whereby postcolonial knowledge is posited as part of the enduring problem of Western oppression and domination of others. Still one thing seems sure: in order to reveal the intersections between legal processes and new forms of colonialism and imperialism that are emerging within the technological and economic innovations of the twenty-first century (such as the Internet), what is first required is a basic rethinking of the currently prevailing Eurocentric, modernist, state-bound understanding of what constitutes law.

See also: Colonialism: Political Aspects; Law and Development; Law: Imposition, Reception, and Colonial; Modernity; Modernity: History of the Concept; Nationalism: Contemporary Issues; Nationalism: General; Nationalism, Historical Aspects of: Africa; Nationalism, Historical Aspects of: South Asia; Nationalism, Historical Aspects of: The West; Postcolonial Geography; Postcoloniality

Bibliography

- Baxi U 2000 Postcolonial legality. In: Schwarz H, Ray S (eds.) *A Companion to Postcolonial Studies*. Blackwell, Oxford, pp. 540–55
- Chakrabarty D 1992 Postcoloniality and the artifact of history: who speaks for 'Indian' pasts? *Representations* 37 (Winter): 1–26
- Chakrabarty D 2000a *Provincializing Europe: Postcolonial Thought and Historical Difference*. Princeton University Press, Princeton, NJ
- Chakrabarty D 2000b A small history of subaltern studies. In: Schwarz H, Ray S (eds.) *A Companion to Postcolonial Studies*. Blackwell, Oxford, pp. 467–85
- Darian-Smith E 1996 Postcolonialism: a brief introduction. *Social and Legal Studies* 5(3): 291–99
- Darian-Smith E, Fitzpatrick P (eds.) 1999 *Laws of the Postcolonial*. University of Michigan Press, Ann Arbor, MI
- Dezalay Y, Garth B 1996 *Dealing in Virtue: International*

- Commercial Arbitration and the Construction of a Transnational Legal Order*. University of Chicago Press, Chicago and London
- Dirks N B 1986 From little king to landlord: property law, and the gift under the Madras Permanent Settlement. *Contemporary Studies in Society and History* 28: 307–33
- Fitzpatrick P 1999 Passions out of place: law, incommensurability, and resistance. In: Darian-Smith E, Fitzpatrick P (eds.) *Laws of the Postcolonial*. University of Michigan Press, Ann Arbor, MI, pp. 39–60
- Gandhi L 1998 *Postcolonial Theory: A Critical Introduction*. Colombia University Press, New York
- Gates H Jr 1991 Critical Fanonism. *Critical Inquiry* 17 (Spring): 457–70
- Loomba A 1998 *Colonialism–Postcolonialism*. Routledge, London and New York
- Mamdani M 1996 *Citizen and Subject: Contemporary Africa and the Legacy of Late Colonialism*. Princeton University Press, Princeton, NJ
- Ngugi wa T 1986 *Decolonising the Mind: The Politics of Language in African Literature*. James Currey, London
- O'Hanlon R, Washbrook D 1992 After orientalism: Culture, criticism, and politics in the Third World. *Comparative Studies in Society and History* 34 (1): January, pp. 141–67
- Prakash G 1990 Writing post-orientalist histories of the Third World: Perspectives from Indian historiography. *Comparative Studies in Society and History* 32 (2): April, pp. 383–408
- Prakash G 1992 Can the 'Subaltern' Ride? A reply to O'Hanlon and Washbrook. *Comparative Studies in Society and History* 34 (1): January, pp. 168–84
- Robinson D 1997 *Translation and Empire: Postcolonial Theories Explained*. St. Jerome Publishing, Manchester
- Said E 1978 *Orientalism*. Vintage, New York
- San Juan E 1998 *Beyond Postcolonial Theory*, 1st edn. St. Martin's Press, New York
- Schwarz H, Ray S (eds.) 2000 *A Companion to Postcolonial Studies*. Blackwell, Oxford, UK
- Spivak G C 1999 *A Critique of Postcolonial Reason*. Harvard University Press, Cambridge, MA
- Stoler A 1989 Rethinking colonial categories: European communities and the boundaries of rule. *Contemporary Studies in Society and History* 31(1): 134–61
- Trivedi H, Mukherjee M (eds.) 1996 *Interrogating Post-Colonialism: Theory, Text and Context*. Indian Institute of Advanced Study, Rashtrapati Nivas, Shimla, India
- Williams P, Chrisman L (eds.) 1994 *Colonial Discourse and Post-Colonial Theory: A Reader*. Colombia University Press, New York

E. Darian-Smith

Postcoloniality

Postcoloniality is the unprecedented, complex predicament faced by individuals, societies, and their governments after the end of European colonial rule. An increasing number of social scientists, especially in the 1980s and 1990s, have found conceptions of postcoloniality useful for characterizing many contemporary societies. Theories of postcoloniality have

emerged especially as more established alternatives—theories of modernization, of underdevelopment, of the Third World, and of revolution—have lost their power to explain the present and predict or guide the future. The premise of most postcoloniality theories is that 'The history of the West and the history of the non-West are by now irrevocably different and irrevocably shared' (Sangari 1995, p. 147).

Writings on postcoloniality range widely in discipline and scholarly methods, and might be typifiable first of all by an attitude, even a mood: disillusionment coupled with commitment, disillusionment with the conspicuous failure of decolonization to produce a world of peace, prosperity, and real liberation, coupled with commitment to diagnose that failure and seek improvements or alternatives. In a British colonial prison during World War II, Jawaharlal Nehru wrote that 'Out of this crucible of war, wherein so much is melting, we hope that something finer will emerge for the West as well as the East, something that will retain all the great achievements of humanity and add to them what they lacked' (Nehru 1982 [1946], p. 81). Thirty-seven years later, as he helped launch the 'Subaltern Studies' school of history, Ranajit Guha declared the 'historic failure of the nation to come to its own' in postcolonial India, its failure as either a bourgeois democracy on the European model or a socialist 'new democracy.' He declared the study of this failure 'the central problematic' for the historiography of India (Guha 1982, p. 7). Similarly, Frantz Fanon wrote in the late 1950s, in the midst of the Algerian anticolonial war, that 'Decolonization is the veritable creation of new men,' that 'the European game has finally ended; we must find something different,' and that 'It is a question of the Third World starting a new history of Man' (Fanon 1968, pp. 36, 312, 315). Thirty-five years later, in the Afterword added to his classic philological study *Orientalism*, Edward Said specifically abjured the idea of 'fixed identities battling across a permanent divide' of Europeans and others (Said 1994, p. 335), in favor of more nuanced multicultural critique of power. While some theorists voice a continuing optimism that, especially in Southern Africa, Eastern Europe, and Southeast Asia, the end of the twentieth century has inaugurated a new age of revolutions (e.g., Comaroff 1996), the dominant tendency in most studies of postcoloniality is critical dissection of utopian anticipations.

To date, discussion of postcoloniality has been more extensive in the Humanities than in the Social Sciences. Critical, often reflexive, theorizing about postcolonial identity and experience has been more extensive than social scientific research into political, economic, and cultural dynamics of postcoloniality. However, controversies in postcolonial theory have provoked fruitful research in anthropology, history, geography, sociology, political science, and other social sciences, characteristically entailing profound

political as well as analytic questions. Good available readers on postcolonial studies include Williams and Chrisman (1994) and Ashcroft et al. (1995). This article centers on efforts to address three powerful antinomies in postcolonial theory, concerning (a) the agency of dominated groups, whether they determine their own fate, (b) the power of the West, whether Western paradigms are inevitable, and (c) the goal of liberation, and when and how postcoloniality might ever end. The article concludes with the tensions between postcoloniality theory and postmodernism, and with efforts to relate postcoloniality and globalization, arguing against portrayals of globalization as the end of postcoloniality, in favor of depiction of the two as intrinsically related.

1. *Can the Subaltern Speak?*

In 1988, literary critic Gayatri Spivak posed the above question as a challenge to the 'Subaltern Studies' school of historians (Spivak 1988). The Subalternists, following Gramsci's theory of class politics in culture, have sought to uncover hidden realities of dominated people's experience and agency in South Asia. The 'Subaltern Studies' project, and Spivak's question, have led to debates and research globally on class, status, culture, and power especially in colonial and postcolonial societies.

The premises of the Subalternists have been, first, that prior histories of colonialism, decolonization, and the postcolonial nation-state have focused on elites, and second, that especially under colonial conditions, elite culture has not been hegemonic, i.e., has not established rituals and routines for social order unchallenged by 'subalterns' or lower class and status, dominated people. In order to explain why a unified postcolonial nation has failed to take shape, Subalternists seek to understand the social experience, collectivities and initiatives of 'subaltern' groups. The original South Asian 'Subaltern Studies' of peasant insurgencies and labor struggles have joined a much broader stream of scholarship on colonial and postcolonial politics, broader not only geographically but also topically, including especially in the 1990s many studies of sex and gender relations, institutions, and imaginaries. However, the issue of the failure of decolonization and the postcolonial nation-states remains central. Efforts have begun to identify positive, existing alternatives to 'the nation' as a form of social organization, not only in South Asia but also in East Asia (notably Duara 1995), the Americas and elsewhere, and also to understand the social legacies of specifically colonial and postcolonial state structures (e.g., in Africa (Mamdani 1996), in Indonesia (Pemberton 1994), and in India (Pandey 1990)).

In the context of research seeking to isolate colonial pathologies and celebrate successful resistances to them, Spivak's question attracted enormous attention because it challenged a second tier of scholarly

premises. Phrased with a literary critic's focus on voice, 'Can the Subaltern speak?' quickly metamorphosed into a question about coherent, autonomous agency, and initiative. Is it possible for 'subaltern' people to shape their own society and culture? Precisely as a radical among philologists, Spivak had insisted that representation in a literary or semiotic sense must be reconsidered in connection with representation in politics, representation in the sense of any capacity for a person to be the agent of, to stand for, the will of other people. Do subalterns have any real power to represent themselves, or is that precisely the lack that constitutes them as subaltern?

Social scientists have come to recognize empirical as well as theoretical issues here. In research on postcolonial societies, the question of subaltern political agency has intersected with studies of class and ethnic structure of anticolonial movements and postcolonial governments, and with studies of democracy in postcolonial conditions. While one tendency in contemporary scholarship is to discover unnoticed ways and means by which subaltern groups preserve themselves, maintain autonomous self-control, and even augment their life chances (e.g., J. C. Scott 1985), another trend is to seek aspects of political structure, global and local, that might account for the reproduction of extremities of domination and violence in postcolonial societies (e.g., Tambiah 1996), and a third trend is to seek to reintegrate the questions about colonial and postcolonial subalterns with studies of domination in metropolitan societies (e.g., Stoler and Cooper 1997). Are postcolonial governments especially likely to be bound to the interests of urban, elite, and ethnically marked subpopulations? Are 'subaltern groups' in such societies less likely to exercise routine rights as citizens, and more likely to sustain whole institutional apparatuses of avoidance and resistance, 'informal' economies, even labor migrations kept deliberately invisible to the postcolonial state? No consensus has emerged. However, it does grow increasingly clear at least that, systematically, the nation-state does not 'come into its own' in postcolonial societies, and 'subaltern' classes in them are neither powerless nor as powerful as citizens in the West.

2. *Provincializing Europe*

Is the postcolonial world destined (or doomed) to follow the West's lead? To fit Western molds? Is the history of the West the central axis of world history? If it isn't, then how can sense be made of colonial history? Research assessing the successes and failures of postcolonial societies that sought alternatively to emulate and to refuse Western models (China, Singapore, Japan, Malaysia, Indonesia, India, and many sites in Africa), especially concerning successes of non-Western capitalisms, and the failures of self-proclaimed socialist revolutions, suggest that efforts to understand the West either as 'exceptional,' successful

for particular cultural or material reasons, or as 'universal,' a distillation of necessities, both fail to grasp the more complex dynamics of the actual history of capitalism, colonialism, and the rise of the nation-state. Anticipations kindred to those of Fanon and Nehru, hoping more or less radically for a new order to supplant the civilization of Europe, were common among anticolonial intellectuals of the mid-twentieth century. These anticipations frequently entwined in a variety of ways with socialist and communist projects for social reconstruction and/or revolution, an end to a capitalist era. For this and other reasons, decolonization became deeply enmeshed in the narrow frame of the Cold War, and was dramatically reconceived within it. The question of the future relation of excolonies and excolonizers became, temporarily, reduced to matters of allegiance in a grand global dualism. Social scientists operated under the premise that all 'Third World' new nations were destined as they 'modernized' to join either the 'free' or 'Communist' worlds (see Pletsch 1981). However, by the end of the century much that was taken for granted in this Cold War common sense was in doubt, was under scrutiny, or was repudiated outright, not only concerning the viability of a revolution against capitalism but also concerning modernization and the nation-state.

In the late-twentieth century, great popularity accrued among scholars for Benedict Anderson's vaguely Marxist theory that the nation as an 'imagined community' was central to the century's global culture (Anderson 1983), but at the same time, disquiet about actual postcolonial nations and states led to new critical questions. Partha Chatterjee mounted the most influential challenge to the premise of the nation-state for excolonies. Chatterjee amplified and reconsidered the implications of parts of Anderson's argument, that 'the nation' was a form for politics linked intrinsically to Enlightenment social theories, a political format pioneered in the Americas and Europe, and then 'pirated' in already 'modular' format as colonies were refashioned and new nations built. Insofar as this is what happened, then, decolonization was not actually a break from Western hegemony. 'Europe and the Americas, the only true subjects of history, have thought out on our behalf not only the script of colonial enlightenment and exploitation, but also that of our anticolonial resistance and postcolonial misery' (Chatterjee 1993, p. 5). Chatterjee doubts that 'our imaginations must remain forever colonized' (Chatterjee 1993, p. 5). However, the question of how to decolonize political imagination, and whether or where it has already been decolonized, has been a provocative one for postcolonial studies. While another influential critic has argued that postcolonial critique of Western universalisms has opened a 'third space' outside both local, cultural particularisms, and modernist universalism (Bhabha 1994, pp. 36-9), others have defined postcolonialism instead as the

dream that this could be done, 'the need, in nations or groups which have been victims of imperialism, to achieve an identity uncontaminated by universalist or Eurocentric concepts and images' (During 1987, p. 3), or the quest for a new relationship with 'the intimate enemy' (Nandy 1983). In this sense, as Lenin once hoped that a United Front politics in colonies could use bourgeois nationalism as a stepping stone to a proletarian revolution, a postcolonial perspective is proffered by some as the key to a new political future. With or without such a dream, the matter remains of how to understand the impact of 'the West' and its (possibly parochial) universals on one-time colonies in both the past and the present. Touchstones for controversy have included 'world system' models of a globally coercive Western-centered economy, and models of globally hegemonic Western culture. World system economic models have largely been encompassed by approaches emphasizing the irreducibility of both global and local economic forces, and the necessity to understand both. While support increases for models of economic 'globalization' that see global economic forces and relations increasing in their influence over time, the causes of the rise of the West as the pre-eminent capitalist center, and the place of colonialism in that history, are still much debated. Similar controversies have followed models of globally hegemonic Western culture, notably Edward Said's landmark study of the kind of discourse he named 'Orientalism,' discourse that constitutes the East and West comparatively, the East as the inferior antithetical other. Said's study has been criticized, ironically, for itself overriding the possibility of non-Western political discourses, for itself undermining the possibility of a world history centered anywhere but the West. Studies have also begun of processes of cultural globalization and of the crucial process of the localizing intermixture of global and local cultural elements that clearly accompanies the development of global cultural connections, the process Hannerz has named 'creolization' (Hannerz 1992).

At their extremes, globalization models can be as simple in their symmetries as conceptions of postcoloniality can be stark in their asymmetries. Recharacterizations of global culture in terms of flows, links, and hybridities, and inquiry into 'modernity at large' (Appadurai 1996), have added an important complexity to studies of postcoloniality, undermining the starkness and simplicity of the colonizer-colonized dualism, and challenging the premise of continuing cultural superiority of the West. However, the situation of postcoloniality requires attention to asymmetries. Reconsidering the past, present, and future place of the West, postcolonial theorists have grown as skeptical of triumphalist visions of reversal, of a non-Western overcoming of Western culture, as they are critical of intrinsically colonial visions of Western superiority over the rest. In *Provincializing Europe*, Dipesh Chakrabarty suggests a way out of the anti-

nomies generated by efforts to reckon with Western domination (Chakrabarty 2000). The paradox in all postcolonialisms that seek a triumphant postcolonial succession to Europe at history's center, Chakrabarty shows, is that even in such a success, the historical structure would still be 'First in Europe, and then elsewhere.' Europe indeed pioneered the form of historicist explanation that gives time a single processual center, giving capital and its culture of Enlightenment a unified place, time, and pattern of development in Europe, while occluding all inassimilable complexities of its earlier and later history elsewhere. However, Chakrabarty argues for abandoning such historicism, in favor of a method which never confines universal history, and provincializes Europe not from a new center, but instead by understanding Western cultural history, also, as one of creolizations (to use Hannerz's term here). Following Chakrabarty (or the similar, 'anti-Hegelian,' argument of Duara (1995)), one can provincialize Europe, confront and unmake colonial asymmetries, without a triumphalism that will inevitably, fatally, unmask as mimicry of European colonialism's own naive historicist egocentrism. Reflection on how, in recent but past centuries, Europe encountered and came in particular, powerful, but not all-powerful, ways to embody and develop universals of capital, democracy, science, etc., might point toward alternative developments in the past and the present, and possibilities not yet recognized for the future, rather than renewing an imperative to overcome empires already dead. In any case, postcolonial theorists have grown weary of battle with colonial ghosts.

3. *When, and How, Can Postcoloniality End?*

It will now be useful, and easy, to clarify the distinction between postcolonialism and postcoloniality. If postcolonialism is the aspiration for a profound political perspective to emerge from the experience of decolonization, postcoloniality is the unprecedented political predicament that decolonization has delivered many societies. The antinomies that fuel the struggle to constitute a postcolonialist political vision are, in less ideological form, equally inescapable for those without postcolonialist aspirations. Postcolonialist quests for political mission via subaltern standpoints struggle over what to expect from and for subaltern political agency. However, vast, dominated, and politically active populations no doubt exist in subaltern positions in postcolonial societies. Any coherent postcolonialism would seem to require a history both centered on and decentering of Europe, but, no doubt, Europe did once dominate, in various and varying ways, the societies it colonized. The question of how, and whether, that domination can be (or has been) ended is crucial to the third antinomy for postcolonialism, the third aspect of predicament in postcoloniality.

Anticolonial theorist and activist Franz Fanon insisted on the goal of liberation, not mere decolonization, for European colonies. Decolonization was something done for, and to, one-time colonies, where the goal should be actual liberation. In fact, as Fanon perceived, in the aftermath of World War II, the European empires disintegrated with extraordinary rapidity, and (except for the French and Portuguese) with little violence: Fanon perceived why. In the atmosphere of the Cold War, Western powers and especially the newly pre-eminent United States greatly feared radical new states emerging from wars of liberation, and preferred to shape successor regimes and constitutions through a pre-emptive orderly process. As Fanon put it, the message flashing was 'for God's sake let's decolonize quick,' as the pieces of European empires were reconfigured into the nation-state form of 'self-determination' required by the new, US-fashioned United Nations (Fanon 1968, pp. 70, 79).

By many diagnoses, decolonization and not genuine liberation was what the new 'Third World' received. But the problem of what would constitute genuine liberation, and which groups merit it, has become increasingly complex. As Firth (2000) has demonstrated, rhetorics of decolonization and liberation have expanded and proliferated after the breakup of most of the European empires has already been accomplished, expanding to articulate aspirations both of groups within nation-states (Biafrans in Nigeria, Kurds in Iraq) and new kinds of groups (women, indigenous peoples, even environments). If the premise of the United Nations has been that the nation-state as a political form constitutes the ultimate form of self-determination for all peoples, that simple, symmetric premise has been called into question. While most reckonings of postcoloniality focus upon the need to overcome asymmetries past, another approach entirely is to reckon it the onset, rather than the aftermath, of something, the predicament constituted for capital-poor excolonies by their re-establishment within the formal symmetries, and limited liabilities, of the United Nations model (Kelly and Kaplan 2001). From this perspective, the goal of liberation itself has become a very specific trap, as formally independent and autonomous political standing comes systematically to entail perpetual clientage, expanding debts and ongoing substantive subservience to outside agencies and agendas. Like the 'free laborer' perceptively critiqued by Marx as free of all access to means of production, these liberated states are, by their very freedom and formal equality, radically limiting of the life chances of their citizens, especially their subalterns. In an interesting recent controversy, Tambiah (1996) rejects the arguments of Pandey (1990) explaining postcolonial communalist violence in South Asia as a legacy of colonial policies that emphasized and institutionalized communal political differences. Tambiah argues that endemic

political violence is a phenomenon of democracy especially in postcolonial conditions, conditions that engender a spiral of destructive, effervescent leveling crowds. The crowds seek not liberation but leveling, destruction of privileges perceived as unmerited, especially across ethnic lines.

Boundary issues have arisen concerning which societies are postcolonial, especially concerning East Asian societies never formally colonized, and the Americas where the legacy of earlier, Iberian domination and earlier independence movements must be calibrated with the history of European colonies in Africa, Asia, and elsewhere. Sociologist Haejoang Cho has argued forcefully that Korea and other East Asian societies were colonized by the West, with becoming colonized defined as coming to require foreign means, methods, and tools simply to live and to think (Cho 1992–94). For the Americas, Klor de Alva (1992) has argued that few or none of the societies outside the Caribbean are postcolonial in fact, that regardless of mestizaje ideologies, Latin American societies are descended not from colonized peoples but largely from the colonizers. As with most of Latin America, whether the United States is postcolonial depends upon the definition adopted. The question is most interesting when connected to different boundary issues concerning postcoloniality: its relations with postmodernism, and with globalization.

Is the post- in postcolonial the post- in postmodern? (So asked another literary critic, Appiah (1992).) Can either end? Dramatic, relativistic postmodernist proclamations that the Enlightenment universals are dead have been countered, Appiah argued, by the fact of misery in postcolonial societies, leading postcolonial political movements to ground their calls for new reforms in ethical universals. While postmodernism offered a first alternative vision to the stalled modernization logic of the Cold War, at present conceptions of both postcoloniality and globalization offer more dynamic alternative formulations. How, then, are postcoloniality and globalization connected? Some scholars envision globalization as the successor, the late capitalist liberator from postcoloniality, as the globe moves from Western modernity to modernity at large or to alternative modernities, negating all vestiges of asymmetric colonial relations. The counter argument (Kelly 1998) is that postcoloniality and globalization are two sides of the same coin, the coin of American power, phenomena not necessarily of a 'late' capitalism but certainly of the new form capitalism has taken under US leadership in the twentieth century. If this is the case, if decolonization and globalization are the fruit of US-led quests for political self-determination and economic open doors, then Europe has already been provincialized by an explicitly anti-imperial agenda, and the predicaments of postcoloniality are the consequences, precisely, of a postcolonialism already instituted. David Scott has perceptively observed, in the first book with 'After

Postcoloniality' in its title, that in the late twentieth century, in the wake of the global resurgence of liberal democracy and the failures of socialist states both Soviet and non-aligned, 'it is no longer so clear what "overcoming" Western power actually means' (D. Scott 1999, p. 14). Perhaps, then, the time has come for dividing 'the West' into its colonial European and twentieth-century American versions. In that case, postcoloniality could end only when American power is as thoroughly confronted as European power has been, and the limitations intrinsic to the formal symmetries of the political present are as fully overcome as have been the formal asymmetries of the colonial past.

See also: Colonialism, Anthropology of; Colonialism: Political Aspects; Colonization and Colonialism, History of; Fourth World; Globalization and World Culture; Hegemony: Anthropological Aspects; Imperialism, History of; Imperialism: Political Aspects; Nationalism: Contemporary Issues; Postcolonial Law; Subaltern History; Subaltern Studies: Cultural Concerns; Third World

Bibliography

- Anderson B 1983 *Imagined Communities: Reflections on the Origin and Spread of Nationalism*. Verso, London
- Appadurai A 1996 *Modernity at Large: Cultural Dimensions of Globalization*. University of Minnesota Press, Minneapolis, MN
- Appiah K A 1992 The postcolonial and the postmodern. In: Appiah K A (ed.) *In My Father's House: Africa in the Philosophy of Culture*. Oxford University Press, New York, chap. 7, pp. 137–57
- Asad T 1993 *Genealogies of Religion: Discipline and Reasons of Power in Christianity and Islam*. The Johns Hopkins University Press, Baltimore, MD
- Ashcroft W, Griffiths G, Tiffin H (eds.) 1995 *The Post-colonial Studies Reader*. Routledge, London
- Bhabha H 1994 *The Location of Culture*. Routledge, New York
- Chakrabarty D 2000 *Provincializing Europe: Postcolonial Thought and Historical Difference*. Princeton University Press, Princeton, NJ
- Chatterjee P 1993 *The Nation and its Fragments: Colonial and Postcolonial Histories*. Princeton University Press, Princeton, NJ
- Cho H 1992–94 *T'alsikminji Chisikin ui Kullikki wa Samilkki* [Reading Texts, Reading Life: The Point of View of a Post-Colonial Intellectual], 3 vols. Ttohaniumunhwa Press, Seoul
- Comaroff J 1996 Ethnicity, nationalism, and the politics of difference in an age of revolution. In: Wilmsen E N, McAllister P (eds.) *The Politics of Difference: Ethnic Premises in a World of Power*. University of Chicago Press, Chicago, pp. 162–83
- Duara P 1995 *Rescuing History from the Nation: Questioning Narratives of Modern China*. University of Chicago Press, Chicago
- During S 1987 Postmodernism or post-colonialism today. *Textual Practice* 1: 32–47
- Fanon F 1968 *The Wretched of the Earth*. Grove Press, New York

- Firth S 2000 Decolonization. In: Borofsky R (ed.) *Remembrance of Pacific Pasts: An Invitation to Remake History*. University of Hawaii Press, Honolulu, HI
- Guha R 1982 On some aspects of the historiography of colonial India. In: Guha R (ed.) *Subaltern Studies I*. Oxford University Press, Delhi, pp. 1–8
- Hannerz U 1992 *Cultural Complexity: Studies in the Social Organization of Meaning*. Columbia University Press, New York
- Kelly J D 1998 Time and the global: against the homogeneous, empty communities in contemporary social theory. *Development and Change* 29: 839–71
- Kelly J D, Kaplan M 2000 *Nation and Decolonization*. University of Chicago Press, Chicago, IL
- Klor de Alva J 1992 Colonialism and postcoloniality as (Latin) American mirages. *Colonial Latin American Review* 1: 3–23
- Mamdani M 1996 *Citizen and Subject: Contemporary Africa and the Legacy of Late Colonialism*. Princeton University Press, Princeton, NJ
- Nandy A 1983 *The Intimate Enemy: Loss and Recovery of Self Under Colonialism*. Oxford University Press, New York
- Nehru J 1982[1946] *The Discovery of India*. Oxford University Press, New Delhi
- Pandey G 1990 *The Construction of Communalism in Colonial North India*. Oxford University Press, New Delhi
- Pemberton J 1994 *On the Subject of 'Java.'* Cornell University Press, Ithaca, NY
- Pletsch K E 1981 The three worlds, or the division of social scientific labor, circa 1950–1975. *Comparative Studies in Society and History* 23: 565–90
- Said E 1994 *Orientalism*. Vintage Books, New York
- Sangari K 1995 The politics of the possible cultural critique 7. In: Ashcroft W, Griffiths G, Tiffin H (eds.) *The Post-colonial Studies Reader*. Routledge, London, pp. 143–7
- Scott D 1999 *Refashioning Futures: Criticism After Postcoloniality*. Princeton University Press, Princeton, NJ
- Scott J C 1985 *Weapons of the Weak: Everyday Forms of Peasant Resistance*. Yale University Press, New Haven, CT
- Spivak G C 1988 Can the subaltern speak? In: Nelson C, Grossberg L (eds.) *Marxism and the Interpretation of Culture*. Macmillan Education, Basingstoke, UK, 271–313
- Stoler A L, Cooper F 1997 Between metropole and colony: Rethinking a research agenda. In: Cooper F, Stoler A L (eds.) *Tensions of Empire: Colonial Cultures in a Bourgeois World*. University of California Press, Berkeley, CA, pp. 1–56
- Tambiah S J 1996 *Leveling Crowds: Ethnonationalist Conflicts and Collective Violence in South Asia*. University of California Press, Berkeley, CA
- Williams P, Chrisman L (eds.) 1994 *Colonial Discourse and Postcolonial Theory: A Reader*. Columbia University Press, New York

J. D. Kelly

Post-Keynesian Thought

1. Introduction

For Joan Robinson (1978; *CEP*, in reprinted Vol. 5, 1979, p. 210), ‘the expression *post-Keynesian*... [applied] to an economic theory or method of analysis

which takes account of the difference between the future and the past’: a helpful point from which to start but it must be remembered that the approaches to political economy reflecting Post-Keynesian thought are there for historical reasons and because of logical associations. Post-Keynesianism is a broad church. Overlaps at each end of a long spectrum of views are marginal, reflecting little more than hostility towards mainstream neoclassical economics and methodology, and conventional forms of Keynesian economics as opposed to the economics of Keynes. Some Post-Keynesians attempt to synthesize the principal strands. Others regard this as misguided. Post-Keynesianism should use a situation—and issue-specific method, a ‘horses for courses’ approach, itself an all-embracing structure at the methodological level.

In this essay historical origins are discussed first; it then deals with approaches to fundamentals issues political economists have tackled since the Physiocrats and Adam Smith theories of value, distribution, price-setting, accumulation, growth and the cycle, and the role of money and financial intermediaries, especially whether money is endogenous or exogenous. Post-Keynesian thought belongs to the tradition that the *raison d'être* of economics is its bearing on policy. Stress is put on the role of institutions in determining economic and political processes, the political philosophies of Post-Keynesians and the policies following from their theories.

All these issues, the methodologies implied in them, the particular ‘vision’ of the economy associated with each strand, are intertwined. This article concentrates on origins, pioneers, and issues. For a full cast list see, for example, Eichner and Kregel (1975), Hamouda and Harcourt 1988; reprinted in Sardoni 1992, Arestis 1992, Lavoie 1992.

2. Classical Roots

Keynes’s new theory reached back over neoclassical economics to join up with the pre-occupations and approaches of our classical forebears. Much that was integral to Keynes’s system, especially its Marshallian offshoots, was jettisoned, at least by Post-Keynesians sympathetic to the viewpoints of the classical economists and Marx. This rules out two prominent American Post-Keynesians, Sidney Weintraub and Paul Davidson, whose structures are amalgams of Marshall and Keynes and who are hostile to Marx, Michal Kalecki, and Piero Sraffa.

An important aspect of classical economics for Post-Keynesian thought is the organizing concept of the surplus—its creation, extraction, distribution, and use. It is the core of their system and its modern counterpart, dovetailing logically with a view of economic processes as dynamic, evolving, progressing organic systems, not the more static allocative process characterizing neoclassical economics until neoclassical growth theory emerged in the 1950s. Linked to the

concept is the notion of a class society, sometimes people, sometimes income, sometimes overlapping. Classes perform different functions. Each is treated as homogenous, with no need to start from the isolated individual economic agent of modern analysis.

Associated with the surplus concept is that of the long-period position, containing the centers of gravitation of the system (Smith's natural prices, Marx's prices of production, Marshall's long-period normal equilibrium prices (and quantities)). With them comes the view that *general* theory is analysis of interrelationships of dominant, persistent forces in society. Modern counterparts are Sraffa's work on value and distribution and Pierangelo Garegnani's (and others') long-period interpretation of Keynes.

3. The Marshallian-Keynesian Roots

Turning to the Marshall-Keynes connection, Marshall's system, as Keynes understood it and within which he perceived himself as operating up to *A Treatise on Money* (1930; *C.W.*, Vols. V, VI, 1971) is noted. Basic ingredients were emphasis on the long period, the real and money dichotomy, the logically prior need for a Say's Law long-period equilibrium position for the Quantity Theory of Money to explain the general price level. Keynes liberated himself from this to write the theory of a monetary production economy—*The General Theory* (1936; *C.W.*, Vol VII, 1973).

Effective demand occurs at the intersection of the aggregate demand and supply functions, usually implying involuntary unemployment. The money rate of interest (which clears the money, not the saving and investment market) rules the roost. Uncertainty is a permanent environment for important economic decisions, especially those concerning investment and holding financial assets for speculative and other purposes. The consumption function is stable, the investment function, volatile; they are connected by the Kahn-Meade multiplier. Finally (an aberration for most Post-Keynesians), the general price level is explained by short-period Marshallian analysis at the level of economy as a whole. These are the jumping-off points for American Post-Keynesians especially.

4. Keynes's Theory of Accumulation

Accumulation theories under our rubric start from, or incorporate, the ingredients of Keynes's theory. The details are usually stringently criticized. The recipes emerging mix the same ingredients in more acceptable ways. Within its own framework Keynes's theory was unacceptable. Keynes confused accumulation's stock and flow aspects by naming his principal concept the marginal efficiency of capital (*MEC*). Capital is a stock; he was mainly concerned with investment, a

flow, and so the marginal efficiency of investment (*MEI*).

Keynes's arguments why higher rates of accumulation are associated with lower *MEI* values, with short-period flow equilibrium achieved when $MEI = i$, (i = the rate of interest), were far-fetched or non-clinching. Keynes assumed rising marginal costs and (usually) marginal-cost pricing. At higher values of accumulation in a given situation, there are higher supply prices and lower *MEI* values. Individual free competitors have, in effect, rational expectations; it is as if they use the eventual equilibrium, not the market price of capital goods in their calculations. Otherwise, their combined actions could not bring about that overall rate of production of capital goods which establishes the appropriate supply price and *MEI* value equal to i .

Second, Keynes argues that because we do not know the future, we project the present into the future unless there are good reasons for not doing so. Higher investment levels are associated with higher sales, prices, profits, output, and employment. Keynes's investment analysis assumes that expected longer-period demand curves for the products concerned remain stable while future short-period supply curves associated with possible levels of investment move to the right, implying lower expected prices, quasi-rents and *MEI* values for investment projects now. If Keynes were true to himself, expected demand curves would move too. Lower values of prices, etc. would not necessarily follow. Keynes used a static analysis to analyze a dynamic process, a criticism first made by Kalecki reviewing in Polish *The General Theory* (1936, *C.W.*, Vol. 1, 1990, pp. 223–32), subsequently by Joan Robinson 1951–1980 and Tom Asimakopulos 1971.

They reacted by setting out the two-sided relationship between accumulation and profitability—actual investment is a major determinant of current profits; current profits, a major determinant of expected profits, which determine planned accumulation in situations of given financial conditions and long-term expectations. This is the basis of Kalecki's theory of distribution, accumulation, growth and the cycle, and Joan Robinson's famous banana diagram, Robinson (1962, p. 48).

Pasinetti (1997) remains true to Keynes, pointing out that the *MEI* soon became identified with the marginal product of capital, that investment was treated as a 'deepening' process responding to different i values. Pasinetti argues Keynes made no such claim—he argued that in a given situation, lower i values make more of the existing investment projects seem profitable *regardless of capital-intensity orderings*. Davidson (1972), too, accepted Keynes's analysis, using Keynes's alternative presentation to describe investment projects by their *NPVs* at ruling i (s). He used Keynes's spot, forward, and futures analysis to go from investment possibilities—prices exceed supply prices—to determine equilibrium flows per period.

5. *Post-Keynesian Theories of Value, Distributions, and Price Setting*

Post-Keynesian theories of value, distribution, and price-setting by, usually, an oligopolistic price-leader firm are now examined. Both sets of theories arise from dissatisfaction with supply and demand theories. The theories criticized and proposed as alternatives are macroeconomic. They arose from the empirical observation *cum* theoretical deduction of classical political economy that, with free competition, there is a tendency to a uniform rate of profit in all activities. Two questions arose: what is the origin?, and what determines the size of the system-wide rate of profits (Sraffa's usage) (Sraffa 1960) to which individual rates tend to be equal? The most explicit, refined answers were Marx's. He emphasized that the ultimate source of value is the difficulty of reproduction in a system of production of commodities by means of commodities (as opposed to utility and scarcity in the alternative tradition). Reconciliation is needed between the embodied labor values of commodities in the sphere of production and prices of production in the sphere of distribution and exchange. The latter contain the overall rate of profits and are the centers of gravitation of the system.

The other major alternatives to the neoclassical marginal productivity explanation of returns to capital and labor and shares of wages and profits are those Kaldor (1955–56) called Keynesian. He was not the first to set forth such a theory. Kalecki (1933, reprinted in 1971, pp. 78–9) had long before: 'capitalists may decide to...invest more in a given period than in the preceding one, but they cannot decide to earn more...therefore, their investment...decisions...determine profits,...not vice versa.' Kaldor located their origins in the widow's cruse story of *A Treatise on Money* (1930; *C.W.*, Vol. V, 1971, p. 125) Joan Robinson (1977; *C.E.P.*, Vol. V, 1979, pp. 184–96) provided a lucid account of Kalecki's theory. Consider the case of no consumption by rentiers, no saving by wage-earners and no government or overseas sectors. Profits (*II*) exactly equal investment (*I*), causation runs $I \rightarrow II$. Kaldor's version is the best known. Common features are different saving behavior at the margin as between wages and profits (and/or wage-earners and profit-receivers) and the Keynesian/Kaleckian proposition that investment leads and saving responds. Kaldor's theory is, paradoxically, long-period and relates to a fully employed economy. Kaldor argued, never coherently, that a growing economy must be fully employed in the long-period sense. Planned saving and investment (and their share of full employment income) are equalized when the income distribution is such that the matching amount of saving is forthcoming.

The process bringing equality depends upon Kaldor's (claimed) empirical observations that while short-term money prices and money-wages are sticky,

so that changes in income equalize saving with investment, longer-term prices are more flexible than wages. These differential changes continue until equality is secured, implying that modern wage-earners have real wage levels so far above the classical-Marxian subsistence levels that they passively accept what remains after profit-receivers have first bite of the cherry: the opposite of the older scenarios in which wages had to be met and the surplus available for profits (and thus accumulation) is what is left.

The other theories under this heading are short-period and not constrained to be at full employment. Saving is equalized with investment by changes in activity and income distribution as profit margins change relative to wage costs, a rest state usually exhibiting involuntary unemployment. (In the most developed models there are roles for the government and overseas sectors.)

The most rigorous presentation of modern classical theory is Sraffa's *Production of Commodities by Means of Commodities* (1960). Sraffa's model is formally similar to Leontief's input-output model; the *conceptual* interpretation is different. With production with a surplus and explicit wage rate (*w*) and rate of profits (*r*), Sraffa shows we must know from outside the value of one of the distributive variables. (The classical economists and Marx chose *w*, Sraffa chose *r*.) The prices of production and the value of the other distributive variable are simultaneously determined. The prices of production are centers of gravitation in the particular circumstances of the time, around which market prices fluctuate—or perhaps on which they converge. (Modern work suggests they do not necessarily do *either*.)

Sraffa starts with a circulating, single-use commodity model, then takes us through non-renewable and/or addable to resources (land and rent) and joint production (to analyze durable capital goods, not wool and mutton). He ends with the choice of techniques in the investment decision, showing the possibility of reswitching, destroying the conceptual foundation of a well-behaved demand curve for 'capital' in the alternative tradition.

6. *The Size of the Mark-up*

Alfred Eichner's version (1976) is the most widely known variant of the theory of mark-up size (apart from Kalecki's 'degree of monopoly' theory). Eichner related its size to the investment plans and financial requirements of oligopolistic firms. As his investment function is flawed, we present here as representative Adrian Wood's (1975) long-period steady-state model of an oligopolistic firm, steady-state because expectations of the values of relevant variables are realized. It is a partial analysis—with a given level of aggregate demand, the firm strives for its optimum share of

demand for the industry's products at the expense of other firms in the industry.

The firm wants its sales revenue to grow as fast as possible, subject to two constraints formalized as an opportunity and a finance frontier. The former reflects that there is eventually a trade-off between profit margin size on turnover and sales revenue growth. Increasing growth requires the margin to be squeezed by a price cut, a rise in sales costs, or both. Assuming one best-practice technique, and given external financial conditions, each growth rate of sales requires a particular margin to finance the capacity to make it possible. The two frontiers' intersection defines the margin associated with highest growth rate. Introducing choice of techniques defines a family of interacting frontiers. Because opportunity frontiers move out at a decreasing rate—greater investment per unit of sales lowers costs at a decreasing rate—while finance frontiers fan out proportionately—the margins, growth rates locus forms a concave curve containing a maximum margin, growth rate combination.

Wood's analysis is in Golden Age logical time—expectations realised, no incompatibilities present. Harcourt and Kenyon (1976, reprinted in Sardoni 1992, pp. 48–66) set the problem in historical time—disappointed expectations determine what happens next—using Salter's vintage analysis. The logical, historical time distinction is due to Joan Robinson (1962, pp. 23–9). Setting the price which finances the accumulation program simultaneously determines the accumulation needed to satisfy expected future sales and still profitable existing capacity (in the sense of covering variable costs) able to contribute to future sales. Frederic Lee (1998) tells the Post-Keynesian price theory story from the earliest times to the present.

7. *Post-Keynesian Accumulation: Moses and the Profits*

Post-Keynesian theories of accumulation were stimulated by Harrod's immediate pre- and post-second world war contributions towards a dynamic economics 1939; 1948. Harrod in effect rediscovered the systemic instabilities of capitalism Marx brought out in his reproduction schemas analysis. Marx asked: what conditions must be satisfied, period by period, in order that aggregate demand and supply, and their compositions, match? This threw up the impossibility of these being achieved by capitalists doing their own things with regard to production, employment, and accumulation.

Harrod obtained the same result by bringing out the long-period destabilizing effects of a momentary gap between planned saving and investment when short period stabilizing effects are abstracted from by assuming the immediate achievement of investment planned for each short period. The growth rate of income allowing planned saving to equal investment

might be such that if business people had known it, they would not have planned those investment levels. The saving and investment functions are so related that excess demand (supply) situations are to the right (left) of the equilibrium position; the system gives the wrong signals as to what to do next. Unless the system is growing at the warranted rate (g_w) (the growth rate allowing expectations and plans to be realized), it moves away from it—all regardless of whether g_w coincided with the natural rate of growth (g_n), the growth rate implied by the economy's supply-side characteristics.

Solow and Swan interpreted Harrod as assuming a fixed capital-output ratio, and wondered whether this was responsible for his disturbing results. As pioneering Keynesians, they assumed an all-wise government removed effective demand puzzles. They asked whether in a long-period sense, Marshall's 'dynamical principle of substitution seen ever at work' allowed a competitive price mechanism to give appropriate price signals, causing the production methods chosen by accumulators to lead g_w towards g_n . The answer is 'yes' in the simplest one, all-purpose commodity models. These results attracted the Cambridge, UK growth theorists' attention. Already suspicious of smooth continuous substitution possibilities, they wished to bring in the 'book of blue prints' approach (long 'at home' at MIT and elsewhere within a different context). The Cambridge UK theorists were loathe to leave out aggregate and effective demand. Their writings mostly remained within the confines of steady-state growth models, learning to walk before running, except for Kaldor who used his distribution mechanism to equalize g_w with g_n .

Kalecki, Josef Steindl, and Richard Goodwin followed a different path, one eventually congenial to Kaldor and Joan Robinson. Kalecki initially developed trendless cycles; in his last paper (1968, reprinted in Kalecki 1971, pp. 165–83) on these themes, trend and cycle were indissolubly mixed: 'the long-run trend [is] but a slowly changing component of a chain of short-period situations ... [not an] independent entity' [Kalecki 1968, reprinted in 1971, p. 165].

This view was developed independently by Goodwin, reaching maturity in 'A Growth Cycle,' Goodwin (1967). He used the Volterra prey-predator model; the analogy of 'the symbiosis of two populations—partly complementary, partly hostile—is helpful in ... understanding ... the dynamical contradictions of capitalism, especially when stated in a ... Marxian form' (Goodwin 1967, reprinted in 1982, p. 167). He analyzed the wages, profits fight, and feedbacks, spawning an expanding literature.

Goodwin integrated effective demand and production interdependence in Goodwin and Punzo (1987). This impressive eclectic book reflects influences from Marx, Schumpeter, Keynes, von Neumann, Joan Robinson, Sraffa, and Kalecki, the developments of catastrophe theory, the concept of 'bifurcation,' and the

Volterra prey-predator biological analogy. Goodwin concentrated on the nature of evolutionary structures experiencing from time to time large jumps and breaks, his key to the cyclical development of economies characterized by production interdependencies.

Joan Robinson (1962) synthesized these ideas in her banana diagram. From Kalecki she derived a relationship between actual accumulation and achieved profitability; from Keynes she derived the 'animal spirits' function relating desired accumulation to expected profitability, itself a function of achieved profitability, in a situation of given financial conditions and long-term expectations. With plausible conjectures about the shapes and positions of the two curves, iteration takes the economy to a stable intersection point where the expected and desired are achieved, her version of Harrod's g_w . It is not a solution of Harrod's (or Domar's) problem; there is no reason why g_w coincides with g_n . Nor are there any mechanisms to take it to g_n . Even if the economy attains g_w , nothing ensures that it will not be driven away in the future—neither relationship can be expected to remain stable.

Discussing *The General Theory* twenty five years on, Joan Robinson signaled the change of method lying behind her analysis in 1956 and 1962.

'The short period is here and now, with concrete stocks of the means of production in existence. Incompatibilities in the situation...determine what happens next. Long-period equilibrium is not at some date in the future: it is an imaginary state of affairs in which there are no incompatibilities in the existing situation.' Joan Robinson (1962, reprinted in *C.E.P.*, Vol. 3, 1965, p. 101).

These developments constitute an attack on the procedure of setting up theoretical issues as searches for existence and uniqueness of equilibria, themselves determined by factors independent of those responsible for stability, local, and global. Similarly, the applied procedure of separating trend from cycle with independent factors responsible for each is also discarded.

8. Post-Keynesian Method

There are two views on method within historically defined Post-Keynesian thought. Increasingly, commentators put neo-Ricardians in a class of their own; The author does not agree. Sraffa's critique of neo-classical economics and positive contributions are integral parts of the historical and logical developments, especially when his underlying Marxism is recognized. The neo-Ricardian stance (on method) is identified as common to all strands of political economy until the 1920s when neoclassical economists, including Hayek, sensed the incoherence of the

concept of the marginal product of capital and its relationship to the long-period method. They reacted by changing the questions and method to the temporary equilibrium analysis of, for example, J.R. Hicks's *Value and Capital*. Neo-Ricardians argue that general theory can only be written about the characteristics of long-period positions reflecting the ultimate outcomes of relationships associated with persistent forces. The economist's role is to identify these, making rigorously explicit their interrelationships. This, they argue, was common to Smith, Marx, and also Marshall who fitted the 'new' supply and demand theories into the old method, making long-period equilibrium prices and quantities the heart of the *Principles*. Therefore, Keynes's theory must be long-period to be a true revolution Garegnani, Eatwell, Milgate, Rogers, and Krishna Bharadwaj refer to passages in *The General Theory* supporting this. Short-period theory has no role in its own right, as opposed to situation-specific analysis at different levels of abstraction.

For Kalecki, Kaldor, and Joan Robinson and their followers, and Paul Davidson and his (the theories differ considerably), the long-period method is a non-starter for descriptive analysis. (For doctrinal debates even Joan Robinson granted it a role. If conjectures or concepts are shown to be incoherent under ideal conditions, this is a legitimate critique of approaches, conceptions, intuitions, and theories.) For descriptive analysis, we must start from a given short period with inherited historical circumstances and accompanying expectations. Long-term expectations and factors are relevant for some current decisions, e.g. investment, pricing; short-term expectations and factors are relevant for others, e.g. production, employment. The story unfolds from one short period to another. (We have simplified: some current activity results from past decisions; some current decisions are relevant for future activity. This led Keynes to abandon period analysis, despairing of finding a determinate time unit into which all relevant interrelationships could be fitted.) There is an element of convergence in the discipline recently because this philosophy underlies work on path-dependence even when the *theory* is neoclassical.

These new developments link back to old approaches to method. Keynes, following Marshall, argued that which variables were regarded as endogenous (determined) and which were exogenous (determining) were arbitrary on any absolute criteria. The crucial decider was the issue in hand.

'The object of our analysis is, not to provide a machine, or method of blind manipulation, [furnishing] an infallible answer, but to provide...an organized and orderly method of thinking out particular problems; and, after we have reached a provisional conclusion by isolating the complicating factors one by one, we...go back on ourselves and allow...for the probable interactions of the factors amongst

themselves...the nature of economic thinking,' Keynes (1936; as in *C.W.*, Vol VI, 1973, p. 297).

Marshall hoped we would get definite, if partial, results and, if we went right round the economy, would be able to bring all results together in a full, overall picture. He forgot that the procedure was inconsistent with his deeper vision that economic processes were akin to systemic interrelated biological processes. (Jan Kregel argues that the *problem* is intractable.)

Sheila Dow (1991) stresses the unifying methodological aspect of Post-Keynesian thought, justifying diverse methods. She recalls to mind Keynes on method and theorizing: there is a continuum of languages which runs from poetry and intuition through lawyer-akin arguments to mathematics and formal logic. Each has a role depending upon the issues, or aspects of issues, discussed. Keynes's emphasis was championed by George Shackle, his favourite example was Keynes's use of 'sentiment.' It could not be defined mathematically, but it gave insights, conjured up atmosphere, invoked the environment of situations analyzed. Kregel (1976) identified in *The General Theory* three models of economic reality of differing degrees of complexity, the most complex being that of shifting equilibrium. The methods were common to Keynes and Post-Keynesians when modeling decision-making in uncertain environments.

Post-Keynesians following Goodwin, Kalecki, and Joan Robinson eschew the Marshallian (or even Keynesian) method; others combine a similar 'vision' of society with explicit use of the Marshallian/Keynesian method or, even earlier, the classical economists' implicit method. Outstanding examples include Pasinetti whose many years work on a multi-sectoral growth model came to fruition in 1981 (with a simple exposition in 1993). Pasinetti integrates Keynes's macroeconomic aggregates with the production interdependence of the classicals, Marx and Sraffa. He sets out the consistency conditions required to maintain full employment growth in a multi-sector model in which the demands for commodities over time follow their respective Engels curves, technical progress occurs and, in later stages of the analysis, the economy is open, trading with and borrowing from or lending to, the rest of the world. The best-known exposition of his insights is Pasinetti (1962); he obtained the 'remarkable result,' $r = g/s_c$. It has proved extremely robust as complications have been added—international trade, government, financial intermediaries, endogenously determined saving ratios.

Another major contribution is Marglin's contribution to the conflict inflation literature (1984). (Rowthorn (1977) derived the same results.) Marglin has a Keynesian-Marxian 'vision' but is Marshallian in method. His aim is to establish long-period 'equilibrium' relationships—short-period stations approach long-period crosses which stay 'put' until and

after they are reached. His principal result is that sustained rates of inflation allow disappointed aspirations of wage-earners and capitalists to be shared in uneasy but sustainable truces.

Post-Keynesian thought starts from real world observations, for example, Kaldor's famous 'stylized facts': broad empirical generalizations holding, in a rough and ready way, for long runs of historical time, requiring situation-specific theories to explain them. Kaldor pioneered the theory of cumulative causation, a deep idea learnt from Allyn Young (and Gunnar Myrdal) Young alerted Kaldor to the pervasive existence of dynamic increasing returns in economic life. Kaldor exploited these illuminations, and inter-regional relationships, to analyze specific periods of economic history.

A wolf pack analogy illustrates the contrast between mainstream axiomatic, equilibrium theory and Post-Keynesian cumulative causation theory. The former thinks of a wolf pack running along. If some wolves stray, powerful forces return them to the pack. The latter identifies powerful forces making those leaving the pack get further ahead or fall further behind, at least for long periods of time—hence the rejection of the equilibrium concept in descriptive analysis.

9. Endogenous Money

Keynes, before and after *The General Theory*, was an endogenous money person, arguing that the demand for money, especially for credit, helped determine the interest rate and the supply of money. Was his stance in *The General Theory*, where he is interpreted as having an exogenous money supply with the interest rate, the price allowing it to be voluntarily held, an aberration? The answer (given most recently by Sheila Dow (1997) is that the money supply is *given*, not *exogenous*. By 1937 he reverted to type when discussing the finance motive for holding money (read, demanding credit)—he had banks creating credit in response to demands for finance for investment projects, a base from which start Post-Keynesian debates on money and credit demand and supply.

Within these discussions, there are claims that the money supply is entirely demand (not supply) determined, summarized in the phrase. Horizontalists versus Verticalists, see Moore (1988). Moore's position is extreme. Most reject the verticalist interpretation, a stance attributed to much mainstream theory, but allow important roles for credit rationing and rising interest rates in determining outcomes in the financial assets market. It is almost universally argued that meeting demand for funds for investment and now, with credit for all, consumption too, creates deposits. This contrasts with the mainstream view that deposits of present and past saving initiate potential lending by financial intermediaries. The former view was fervently expressed by Keynes: 'The investment

market can become congested through shortage of cash, never...through shortage of saving.' (C.W., Vol. XIV, 1973, p. 222). Kalecki independently arrived at the same view. Post-Keynesians start from here, responding to new institutional developments and refining, extending the analysis. They reflect the major change James Meade attributed to Keynes: Keynes changed us from looking at the world as a saving dog wagging an investment tail to one in which an investment dog wagged a saving tail.

Within this set of discussions, some stress the stock aspects of money and financial assets, others, the flow aspect of credit as finance for important expenditures. Associated with this distinction are Post-Keynesians concentrating on one period only in their analysis of money—accommodationists. They contrast with structuralists, stressing period by period analysis in the manner of the mature Kalecki and John Hicks. The upshot is to preserve Keynes's liquidity preference theory when explaining demand for money and financial assets, levels and patterns of interest rates and banks extending credit guided by their liquidity preference, so that not all demand is necessarily accommodated, see Cottrell (1994).

10. Policy

The connection between Post-Keynesian analysis and the real world includes a healthy preoccupation with policy. The mean Post-Keynesian's policy proposals match the mean Bastard Keynesian's—a slightly left of center package deal of monetary, fiscal and incomes policies, combined with a longer-term commitment to freer trade while preserving protective positions until 'the time is ripe' and preference for fixed exchange rates (with institutional means allowing orderly realignments now and then). Revamping Bretton Woods's structures, to remove built-in contractionary biases, guides suggestions on the international stage. Deregulation of financial markets and 'freeing up' of exchange rates unleashed an era of instability following the tranquil experience of 'the Golden Age of Capitalism.' This has led to attempts, e.g. the Tobin tax on foreign exchange transactions, to curb the excesses of speculative movements in the foreign exchange markets, stock exchanges, and property markets.

Where Post-Keynesians feel they have the upper hand is in their argument that cumulative causation processes characterize the workings of markets and economic systems. If such processes operate it is harder to argue for beneficial systemic effects from speculation—that it reduces fluctuations and gets the market (system) to underlying long-period equilibrium positions more quickly than otherwise would be the case. If they are not there, speculation makes matters worse. We are led to middle way programs between ruthless

laissez faire (but not that competitive) capitalism and authoritarian, inefficient, centrally planned economies. Heinrich Bortis (1997) profoundly analyzed the philosophical, political and economic aspects of these developments. He synthesized three major stands of Post-Keynesianism: the neo-Ricardian approach to capture the longer term aspects of economies' developments, the Kaleckian-Robinsonian, their cyclical aspects, and the American Post-Keynesians, short-period behavior in product and money markets under uncertainty. To this he allied a democratic socialist platform, with philosophical foundations and a comprehensive outline of policy, nationally and internationally. Another stalwart is John Cornwall. Over the years he developed his original theoretical critique of Harrod for supposing that g_w and g_n were independent of each other and supplemented his analysis of the implications for growth of systemic supply and demand interdependence, with a steady stream of down-to-earth, humane recommendations for institutional reforms, see Harcourt and Monadjemi (1999).

Bortis and Cornwall were influenced by Kaldor. Kaldor's last book (1996), contains his mature views on how the world works: policy proposals are in the last chapter. He has long emphasized the importance of industrial increasing returns and different pricing behavior as between the (export) products of primary producing countries and countries producing industrial goods. He combined policies for relative stability in exchange rates worldwide with international buffer stock schemes for primary commodities. The latter serve the dual purpose of reducing fluctuations in prices and incomes which damage the economies producing these commodities (and those buying them) while simultaneously providing worldwide liquidity for trade and capital movements, a provision free of disadvantages arising when one major country's currency performs this role, as happens now.

See also: Keynes, John Maynard (1883–1946); Keynesianism and State Interventionism; Marxian Economic Thought; Political Economy, History of; Robinson, Joan (1903–83)

Bibliography

- Arestis P 1992 *The Post-Keynesian Approach to Economics*. E. Elgar, Aldershot, UK
- Asimakopulos A 1971 The determination of investment in Keynes' model. *Canadian Journal of Economics* 4: 382–8
- Bortis H 1997 *Institutions, Behavior and Economic Theory. A Contribution to Classical-Keynesian Political Economy*. Cambridge University Press, Cambridge, UK
- Cottrell A 1994 Post-Keynesian monetary economics. *Cambridge Journal of Economics* 18: 587–605
- Davidson P 1972 *Money and the Real World*. Macmillan, London

- Dow S 1991 Keynes's epistemology and economic methodology. In: O'Donnell R M, Keynes J M (eds.) *Philosophy, Economics and Politics*. St Martin's Press, New York, pp. 144–167
- Dow S 1997 Endogenous money. In: Harcourt C C, Riach P A (eds.) *A 'Second Edition' of The General Theory*. Routledge, London, Vol. 2, pp. 61–78
- Eichner A S 1976 *The Megacorp and Oligopoly: Micro Foundations of Macro Dynamics*. Cambridge University Press, Cambridge, UK
- Eichner A S, Kregel J A 1975 An essay on Post-Keynesian theory: a new paradigm in economics. *Journal of Economic Literature* **XIII**: 293–314
- Goodwin R M 1982 *Essays in Dynamic Economics*. Macmillan, London
- Goodwin R M, Punzo L F 1987 *The Dynamics of a Capitalist Economy: A Multi-sectoral approach*. Basil Blackwell, Polity Press, Oxford
- Harcourt G C, Monadjemi M S 1999 The vital contributions of John Cornwall to economic theory and policy. In: Setterfield M (ed.) 1999. *Growth, Employment and Inflation. Essays in Honour of John Cornwall*. St Martin's Press, New York
- Harcourt G C, Riach P A (eds.) 1997 *A 'Second Edition' of The General Theory*, 2 Vols. Routledge, London
- Harrod R F 1939 An essay in dynamic theory. *Economic Journal* **49**: 14–33
- Harrod R F 1948 *Towards a Dynamic Economics. Some Recent Developments of Economic Theory and their Application to Policy*. Macmillan, London
- Kaldor N 1955–56 Alternative theories of distribution. *Review of Economic Studies*. **23**: 83–100
- Kaldor N 1996 *Causes of Growth and Stagnation in the World Economy*. Cambridge University Press, Cambridge, UK
- Kalecki M 1936 Some remarks on Keynes' theory. In: Osiatynski J (ed.) 1990 *Collected Works of Michal Kalecki*, Vol I. Oxford University Press, Oxford, UK, pp. 223–32
- Kalecki M 1971 *Selected Essays on the Dynamics of Capitalist Economies 1933–71*. Cambridge University Press, Cambridge, UK
- Keynes J M 1971–89 *The Collected Writings of John Maynard Keynes*, 30 Vols. for the Royal Economic Society. Macmillan, London
- Kregel J A 1976 Economic methodology in the face of uncertainty: the modeling methods of Keynes and the Post Keynesians. *Economic Journal* **86**: 209–25
- Lavoie M 1992 *Foundations of Post-Keynesian Economic Analysis*. E. Elgar, Aldershot, UK
- Lee F S 1998 *Post Keynesian Price Theory*. Cambridge University Press, Cambridge, UK
- Marglin S A 1984 *Growth, Distribution and Prices*. Harvard University Press, Cambridge, MA
- Moore B J 1988 *Horizontalists and Verticalists: The Macroeconomics of Credit Money*. Cambridge University Press, Cambridge, UK
- O'Donnell R M (ed.) 1991 *Keynes as Philosopher-Economist*. St Martin's Press, New York
- Pasinetti L L 1962 'Rate of profit and income distribution in relation to the rate of economic growth. *Review of Economic Studies* **29**: 267–79
- Pasinetti L L 1981 *Structural Change and Economic Growth: A Theoretical Essay on the Dynamics of the Wealth of Nations*. Cambridge University Press, Cambridge, UK
- Pasinetti L L 1993 *Structural Economic Dynamics. A Theory of the Economic Consequences of Human Learning*. Cambridge University Press, Cambridge, UK
- Pasinetti L L 1997 The Marginal Efficiency of Investment. In:

- Harcourt G C, Riach P A (eds.) *Capital Controversy: Post-Keynesian Economics and the History of Economic Thought*. Routledge, London, I: pp. 198–218
- Robinson J 1956 *The Accumulation of Capital*. Macmillan, London
- Robinson J 1962 *Essays in the Theory of Economic Growth*. St Martin's Press, London
- Robinson J 1978 Keynes and Ricardo. *Journal of Post Keynesian Economics* **1**: 12–18
- Robinson J 1951–1980 *Collected Economic Papers*, 6 Vols. Blackwell, Oxford, UK
- Rowthorn R E 1977 Conflict, inflation and money. *Cambridge Journal of Economics* **1**: 215–39
- Sardoni C (ed.) 1992 *On Political Economists and Modern Political Economy. Selected Essays of G C Harcourt*. Routledge, London
- Sraffa P 1960 *Production of Commodities by Means of Commodities. Prelude to a Critique of Economic Theory*. Cambridge University Press, Cambridge, UK
- Wood A 1975 *A Theory of Profits*. Cambridge University Press, Cambridge, UK

G. C. Harcourt

Postmodern Urbanism

1. Postmodernism in Urban Studies

In the field of urban studies, three principal references are discernable in postmodern thought:

- (a) a series of distinctive cultural and stylistic practices that are in and of themselves intrinsically interesting;
- (b) the totality of such practices, viewed as a cultural ensemble characteristic of the contemporary epoch of capitalist society; and
- (c) a philosophical and methodological discourse antagonistic to the precepts of Enlightenment thought, most particularly the hegemony of any single intellectual persuasion.

Implicit in all these approaches is the vision of a 'radical break,' i.e., a discontinuity between past and future trends (whether cultural, political-economic, or philosophical).

The pivotal point of departure in a postmodern urbanism is that just as the tenets of modernist thought have been undermined and a multiplicity of new ways of knowing put in their place, so, analogously, in postmodern cities the logics of previous urbanisms have evaporated; and in the absence of a single new imperative, multiple forms of (ir)rationality clamor to fill the vacuum. The localization, sometimes literally the concretization, of these multiple logics is creating the emerging time-space fabric of what may be called a postmodern society. The contradictions between a modernist and a postmodern urbanism are encoun-

tered most directly in the differences between the 'Chicago School' and the 'Los Angeles School' of urban structure.

1.1 *The Chicago School*

The most enduring of the Chicago School models was the zonal or concentric ring theory (Burgess 1925). Based on assumptions that included a uniform land surface, universal access to a single-centered city, free competition for space, and the notion that development would take place outward from a central core, Burgess concluded that the city would tend to form a series of concentric zones. At the core of Burgess's schema was the Central Business District (CBD), which was surrounded by a transitional zone, where older private houses were being converted to offices and light industry, or subdivided to form smaller dwelling units. The transitional zone was succeeded by a zone of workingmen's homes, which included some of the oldest residential buildings in the city, and stable social groups. Beyond this, newer and larger dwellings were to be found, occupied by the middle classes. Finally, the commuters' zone extended beyond the continuous built-up area of the city (where a considerable portion of the zone's population was employed).

Other urbanists noted the tendency for cities to grow in star-shaped rather than concentric form, along highways that radiate from a center with contrasting land uses in the interstices. This gave rise to a sector theory of urban structure, advanced in the late 1930s by Homer Hoyt (1939), who observed that once variations arose in land uses near the city center, they tended to persist as the city grew. Distinctive sectors thus expanded out from the CBD, often organized along major highways. The complexities of real-world urbanism were further taken up in the multiple nuclei theory of C. D. Harris and E. L. Ullman (1945). They proposed that cities have a cellular structure in which land uses develop around multiple growth-nuclei within the metropolis—a consequence of accessibility-induced variations in the land-rent surface and agglomeration (dis)economies.

Much of the urban research agenda of the twentieth century was predicated on the percepts of the concentric zone, sector, and multiple nuclei theories of urban structure. However, the specific and persistent popularity of the Chicago concentric ring model is probably related to its beguiling simplicity and the enormous volume of publications produced by adherents of the Chicago School.

1.2 *The Los Angeles School*

During the 1980s, a group of loosely associated scholars began to examine the notion that what was happening in the Los Angeles (LA) region was

somehow symptomatic of a broader socio-geographic transformation taking place within the US. Their project was based on certain shared theoretical assumptions, as well as the view that LA was emblematic of some more general urban dynamic. One concern common to all adherents of the LA School was a focus on restructuring, which included deindustrialization, the birth of the information economy, the decline of nation-states, and the rise of the Pacific Rim. Such proliferating logics often involved multiple theoretical frameworks that overlap and coexist in their explanations of the burgeoning global/local order—a heterodoxy consistent with the project of postmodernism.

1.3 *From Chicago to LA*

Based on the differences between Chicago and LA, the distinction between a modernist and postmodern urbanism may be summarized as follows: whereas traditional concepts of urban form imagine a city based on a single central core that organized its hinterland (Chicago-style), in postmodern urbanism it is the hinterlands that organize what is left of the center (as in Los Angeles).

2. *Postmodern Urbanism: Urban Form and Urban Process*

Contemporary urban development in southern California provides a wealth of evidence that is symptomatic of a burgeoning postmodern urbanism.

2.1 *Edge Cities*

Joel Garreau (1991, p. 3) noted the central significance of Los Angeles in understanding contemporary metropolitan growth in the US. He asserts that: 'Every single American city that is growing, is growing in the fashion of Los Angeles,' and claims there are 26 edge cities within a five-county area in southern California. The classic location for contemporary edge cities is at the intersection of a suburban beltway and a hub-and-spoke lateral road. The central conditions that have propelled such development are the dominance of the automobile and the associated need for parking, the communications revolution, and the entry of women in large numbers into the labor market. One essential characteristic of the edge city is that politics is not yet established there. Into the political vacuum moves a 'shadow government'—a privatized proto-government that is essentially a plutocratic alternative to normal politics. Shadow governments can tax, legislate for, and police their communities; but they are rarely accountable, are responsive primarily to wealth (as opposed to numbers of voters), and are subject to few constitutional constraints (Garreau 1991, p. 187).

2.2 *Privatopia*

'Privatopia,' perhaps the quintessential edge-city residential form, is a private housing development based in common-interest developments (CIDs) and administered by homeowners' associations (McKenzie 1994, p. 11). Sustained by an expanding catalogue of covenants, conditions, and restrictions (or CC&Rs), privatopia has been fueled by ideology of 'hostile privatism' that has provoked a culture of non-participation. McKenzie notes how this 'secession of the successful' has altered concepts of citizenship, in which 'one's duties consist of satisfying one's obligations to private property' (1994, p. 196).

2.3 *Cultures of Heteropolis*

Provoked to comprehend the causes and implications of the 1992 civil disturbances in Los Angeles, Charles Jencks (1993, p. 32) zeroes in on the city's diversity as the key to LA's emergent urbanism. The consequent built environment is characterized by transience, energy, and unplanned vulgarity. Jencks views this improvisational quality as a hopeful sign, since it accepts the different voices that create a city, suppresses none of them, and makes from their interaction some kind of greater dialogue (1993, p. 75). This is especially important in a city where 'minoritization' is 'the typical postmodern phenomenon where most of the population forms the "other"' (Jencks 1993, p. 84).

2.4 *City as Theme Park*

California in general and Los Angeles in particular, have often been promoted as places where the American (suburban) dream is most easily realized. Many writers have used the 'theme park' metaphor to describe the emergence of its associated cityscapes. For instance, Michael Sorkin (1992) describes theme parks as places of simulation without end, characterized by a-spatiality, plus technological and physical surveillance and control. What is missing in this new cybernetic suburbia is not a particular building or place, but the spaces between; that is, the connections that make sense of forms (p. xii). What is missing, Sorkin maintains, is connectivity and community.

2.5 *Fortified City*

The downside of the southern Californian dream has been the subject of countless dystopian visions. In one powerful account, Mike Davis noted how southern Californians' obsession with security has transformed the region into a fortress. This shift is manifested in the

physical form of the city, which is divided into fortified cells of affluence and places of terror where police battle the criminalized poor. These urban phenomena, according to Davis, have placed Los Angeles 'on the hard edge of postmodernity' (Davis 1992a, p. 155).

2.6 *Interdictory Spaces*

Elaborating upon the notion of fortress urbanism, Steven Flusty observed how various types of fortification have extended a canopy of suppression and surveillance across the entire city. His taxonomy of interdictory spaces (1994, p. 16–17) stresses how spaces are designed to exclude by a combination of their function and cognitive sensibilities. Some spaces are passively aggressive. Other configurations are more assertively confrontational. Flusty notes how combinations of interdictory spaces are being introduced 'into every facet of the urban environment, generating distinctly unfriendly mutant typologies' (1994, pp. 21–33). Some are indicative of the pervasive infiltration of fear into the home; others betray a fear of the public realm.

2.7 *Historical Geographies of Restructuring*

In his history of Los Angeles between 1965 and 1992, Ed Soja (1996) attempts to link the emergent patterns of urban form with underlying social processes. He identified six kinds of restructuring, which together define the region's contemporary urban process. In addition to 'Exopolis,' the city without (a massive simulation of what a city should be), Soja lists: 'Flexcities,' associated with the transition to post-Fordism, especially deindustrialization and the rise of the information economy; and 'Cosmopolis,' referring to the globalization of Los Angeles, both in terms of its emergent world-city status and its internal multicultural diversification. Three specific geographies are consequent upon these dynamics:

- (a) 'Splintered Labyrinth,' which describes the extreme forms of social, economic, and political polarization characteristic of the postmodern city;
- (b) 'Carceral' city, referring to the 'incendiary urban geography' brought about by the amalgam of violence and police surveillance; and
- (c) 'Simcities,' the term Soja uses to describe the new ways of seeing the city—an epistemological restructuring that foregrounds the postmodern perspective.

2.8 *Fordist and Post-Fordist Regimes of Accumulation and Regulation*

Many observers agree that one of the most important underlying processes in the contemporary political

economy is represented by the shift from a Fordist to a post-Fordist industrial organization. Allen Scott's basic argument is that there have been two major phases of urbanization in the US. The first related to an era of Fordist mass production, during which the paradigmatic cities of industrial capitalism (Detroit, Chicago, Pittsburgh, etc.) coalesced around industries that were themselves based upon ideas of mass production. The second phase is associated with the decline of the Fordist era and the rise of a post-Fordist 'flexible production.' This is a form of industrial activity based on small-size, small-batch units of (typically subcontracted) production that are nevertheless integrated into clusters of economic activity (Scott 1993). Post-Fordist regimes of accumulation are associated with analogous regimes of regulation, or social control. Perhaps the most prominent manifestation of changes in the regime of regulation has been the retreat from the welfare state, which has led many people to the brink of poverty just at the time when the social welfare 'safety net' is being withdrawn.

2.9 Globalization

The global/local dialectic is another important (if somewhat imprecise) *leitmotif* of contemporary urban process. In his reference to the global context of LA's localisms, Mike Davis (1992b) claims that if LA is in any sense paradigmatic, it is because the city condenses the intended and unintended spatial consequences of post-Fordism. He insists that there is no simple master logic of restructuring, focusing instead on two key localized macro-processes: the over-accumulation in southern California of bank and real estate capital, principally from the East Asian trade surplus; and the reflux of low-wage manufacturing and labor-intensive service industries, following upon immigration from Mexico and Central America. Through such connections, what happens today in Asia and Central America will tomorrow have an effect in Los Angeles.

2.10 Politics of Nature

The natural environment of southern California has been under constant assault since the first colonial settlements. Human habitation on a metropolitan scale has only been possible through a widespread manipulation of nature, especially the control of water resources. On one hand, southern Californians tend to hold a grudging respect for nature, living as they do adjacent to one of the earth's major geological hazards, and in a desert environment that is prone to flood, landslide, and fire. On the other hand, its inhabitants have been energetically, ceaselessly, and sometimes carelessly unrolling the carpet of urbanization over the natural landscape for more than a

century. This uninhibited occupation has engendered its own range of environmental problems, most notoriously air pollution; but it also brings forth habitat loss and dangerous encounters between humans and other animals. In this sense, a postmodern urbanism is axiomatically an environmental issue (Wolch and Emel 1998).

In sum, these diverse dynamics suggest an urban process:

(a) that is driven by a global restructuring and permeated and balkanized by a series of interdictory networks;

(b) whose populations are socially and culturally heterogeneous, but politically and economically polarized;

(c) whose residents are educated and persuaded to the consumption of dreamscapes even as the poorest are consigned to 'carceral' cities;

(d) whose built environment, reflective of these processes, consists of edge cities, 'privatopias,' and the like; and

(e) whose natural environment is being erased to the point of unlivability.

3. The Postmodern Urban Condition

In their model of postmodern urbanism, Dear and Flusty (1998) synthesize the diverse spatial manifestations of the postmodern urban process into a vision of 'keno capitalism.' According to their characterization, urbanization is occurring on a quasi-random field of opportunities. Capital touches down as if by chance on a parcel of land, ignoring the opportunities on intervening lots, thus sparking the development process. The relationship between development of one parcel and nondevelopment of another is a disjointed, seemingly unrelated affair. While not truly a random process, it is evident that the traditional, center-driven agglomeration economies that guided urban development in the past no longer apply. Conventional city form, Chicago-style, is sacrificed in favor of a noncontiguous collage of parceled, consumption-oriented landscapes devoid of conventional centers, yet wired into electronic propinquity and nominally unified by the mythologies of the (dis)information superhighway. The consequent urban aggregate is characterized by acute fragmentation and specialization—a partitioned gaming board subject to perverse laws and peculiarly discrete, disjointed urban outcomes.

The validity and full potential of the concept of postmodern urbanism can be judged only by a thorough comparative urban analysis that goes beyond Los Angeles (Dear 2000). Four contemporary emphases seem especially promising:

(a) world cities, or the emergence of a very few 'megacities' involved in key command and comfort functions in the globalizing political economy;

(b) dual cities, or the burgeoning polarizations between rich and poor, powerful and powerless, computer literate and illiterate, etc.;

(c) hybrid cities, or the fragmentation of existing social alliances and the rise of unprecedented hybrid categories and places; and

(d) 'cybercities,' or the radical undermining of conventional geographies by the telecommunications revolution.

See also: Critical Realism in Geography; Humanistic Geography; Postmodernism in Geography; Postmodernism in Sociology; Urban Anthropology; Urban Studies: Overview

Bibliography

- Burgess E W 1925 The growth of the city. In: Park R E, Burgess E W, McKenzie R D (eds.) *The City: Suggestions of Investigation of Human Behavior in the Urban Environment*. University of Chicago Press, Chicago, IL, pp. 47–62
- Davis M L 1992a Fortress Los Angeles: The militarization of urban space. In: Sorkin M (ed.) *Variations on a Theme Park*. Noonday Press, New York, pp. 154–80
- Davis M L 1992b Chinatown revisited? The internationalization of downtown Los Angeles. In: Reid D (ed.) *Sex, Death and God in L.A.* Pantheon Books, New York, pp. 54–71
- Dear M 2000 *The Postmodern Urban Condition*. Blackwell, Oxford, UK
- Dear M, Flusty S 1998. Postmodern urbanism. *Annals, Association of American Geographers* **88**: 50–72
- Flusty S 1994 *Building Paranoia: The Proliferation of Interdictory Space and the Erosion of Spatial Justice*. Los Angeles Forum for Architecture and Urban Design, West Hollywood, CA
- Garreau J 1991 *Edge City: Life on the New Frontier*. Doubleday, New York
- Harris C D, Ullman E L 1945 The nature of cities. *Annals of the American Academy of Political and Social Science* **242**: 7–17
- Hoyt H 1939 *The Structure and Growth of Residential Neighborhoods in American Cities*. US Federal Housing Administration, Washington, DC
- Jencks C 1993 *Heteropolis: Los Angeles, the Riots and the Strange Beauty of Hetero-architecture*. St. Martin's Press, New York
- McKenzie E 1994 *Privatopia: Homeowner Associations and the Rise of Residential Private Government*. Yale University Press, New Haven, CT
- Scott A J 1993 *Technopolis: High-technology Industry and Regional Development in Southern California*. University of California Press, Berkeley, CA
- Soja E 1996 Los Angeles 1965–1992: The six geographies of urban restructuring. In: Scott A J, Soja E (eds.) *The City: Los Angeles & Urban Theory at the End of the Twentieth Century*. University of California Press, Berkeley, CA, pp. 426–62
- Sorkin M (ed.) 1992 *Variations on a Theme Park: The New American City and the End of Public Space*. Hill and Wang, New York
- Wolch J, Emel J (eds.) 1998 *Animal Geographies: Place, Politics and Identity in the Nature–Culture Borderlands*. Verso, London

M. J. Dear

Postmodernism in Geography

The impact of postmodernism on the field of geography has been particularly intense and far-reaching, perhaps more so than in any other social science. Although still controversial, with few geographers specifically identifying themselves as postmodernists, postmodern and related *poststructuralist* critiques of traditional modes of interpreting and explaining geographical phenomena have influenced scholarly research and teaching in nearly all branches of the discipline. Especially influential has been postmodernism's role in shaping the development of what has come to be called *Critical Human Geography* and in the rethinking of such core concepts as space, place, region, and environment (see also *Postmodern Urbanism*). At the same time, this postmodernization of geography has contributed significantly to expanding geography's impact on other disciplines, both in the social sciences and the humanities. Geographical writings on postmodernism and postmodernity have been an integral part of the recent resurgence of interest in concepts of space and place across a wider spectrum of academic disciplines than ever before in the twentieth century.

1. The Postmodern Turn in Geography

Postmodernism entered geography as an offshoot and extension of critical responses to the positivist 'spatial science' paradigm that began to dominate the discipline in the 1960s. These critical responses took two major forms. A radical or *Marxist Geography* consolidated in the 1970s as a critique of prevailing modes of explanation in spatial science and as a means of making geography more socially and politically relevant. Rather than focusing primarily on statistically measurable surface appearances and searching for empirical regularities based on the frictions of distance and spatial covariation, how one geographical pattern or distribution correlates with another, Marxist geographers analyzed and explained what they called *specific geographies* as socially constructed outcomes of underlying processes such as class formation and capitalist accumulation. This carried with it a strong emphasis on structural explanation of geographical phenomena (see *Structuralism*) and gave rise to a specifically geographical (urban, regional, international, ecological) political economy perspective that would influence the discipline significantly in subsequent years.

The second stream of critical response, developing initially as *Humanistic Geography*, also concentrated on alternative modes of explanation, but emphasized the role of human agency and subjectivity much more than the effects of powerful structuring forces emanating either from society or the environment (Ley and Samuels 1978). Drawing on phenomenology, exis-

tentialism, and hermeneutics as well as traditional forms of cultural geography and landscape analysis, these geographical humanists were less concerned with space and political economy than with *place, nature-society relations*, and the interpretation of *cultural landscapes*. Continuing long-established links with anthropology, this reinvigorated form of *Cultural Geography* would also influence the development of postpositivist critical human geography over the next several decades as an important countercurrent to the Marxist critique.

By the early 1980s, the spatial science paradigm had passed its peak of influence and many new approaches were entering geography, each competing to define the leading edge of postpositivist research. At the same time, significant critiques were developing both within and between Marxist and humanistic geography (and within spatial science as well), adding further to the proliferation of new perspectives. Perhaps the most challenging critiques came from feminist geographers, who saw an exploitative and demeaning *masculinism* inherent in every branch of what would be labeled modern geography. This patriarchal order was not only expressed in the workplace (e.g., in university privileges, hiring, salaries) but also in the very practice of geography, whether theoretical or empirical, positivist, Marxist, or Humanist. The feminist critique echoed on a broader platform many other concerns over what was perceived as an excessive narrowing of perspectives affecting and constraining the development of (modern) critical human geography. It was in this atmosphere of increasing epistemological fragmentation and escalating internal critique of modern geographical practices that postmodernism first entered the disciplinary discourse.

The deepest divisions within postpositivist geography revolved around traditional philosophical dichotomies or binary oppositions: Subject–Object, Idealism–Materialism, Agency–Structure. Their particular expression, however, took the form of an opposition between approaches that emphasized spatial political economy vs. those that stressed place-based culture. In the late 1970s, there had been some efforts to bring these opposing positions together, most notably in the writings of Derek Gregory. Influenced by the social theorist Anthony Giddens, also then at Cambridge University, Gregory (1978) argued for a critical synthesis of ‘structural’ and ‘reflexive’ approaches to geographical explanation that would go beyond their traditional antagonism to create a new, socially and politically ‘committed’ critical human geography. Gregory was one of the first geographers to recognize the potential relevance of postmodern and closely associated poststructuralist debates to achieving this goal, and for the next two decades would be a leading voice in adding a postmodern, poststructuralist, and, later, postcolonial perspective to geographical research, theory, and teaching (Gregory and Walford 1989, Gregory 1994).

In addition to the influence of Gregory’s work and the growing impact on geography of Giddens’s structuration theory, there were other developments in the early 1980s that attempted to overcome the divisions and fragmentation of postpositivist geography by at least in part introducing some aspects of the postmodern and poststructuralist critiques (Olsson 1980). In retrospect, however, the key turning point in the postmodernization of geography came from outside the discipline, with the publication (in 1984) of Fredric Jameson’s *Postmodernism, or the Cultural Logic of Late Capitalism*. Jameson, a widely recognized Marxist literary critic and cultural studies scholar, had recently become interested in continental European theories of space, place, and architecture, and especially in the work of the French philosopher and urbanist Henri Lefebvre. Then teaching in the History of Consciousness program at the University of California-Santa Cruz, Jameson had organized a lecture tour by Lefebvre in the previous year and hosted his stay as visiting professor at Santa Cruz in 1984.

While he saw architecture as the ‘privileged’ language of postmodernism, Jameson was aware of the new developments in critical human geography and gathered a group of geographers, urbanists, and architects, mainly from California, at a conference to celebrate Lefebvre’s creative spatial theories. When the article appeared, presenting a postmodern and explicitly spatial critique that effectively bridged Marxist political economy and critical cultural studies, innovatively interpreting emerging forms of contemporary urban geography, and calling for a new ‘political esthetic’ based on what Jameson termed a ‘cognitive mapping’ of ‘postmodern hyperspace,’ its effect on geography was particularly powerful.

The impact was especially intense among geographers working in California and, in particular, in Los Angeles, from whose geography Jameson drew some of his most imaginative examples and where there already existed one of the largest clusters of critical human geographers in the US. In the ensuing years, the geography and the geographers of Los Angeles would play a leading role in defining postmodernism and extending its impact on the discipline.

2. Postmodernism Seen from Los Angeles

By the mid-1980s, after decades of relative neglect by urban studies scholars, Los Angeles had become the research focus for a group of geographers and urban planners interested in making theoretical and practical sense of the new urbanization processes that were transforming the contemporary city, not just locally but elsewhere as well. The earliest work on this *urban restructuring* emphasized the radical political economy perspective that was so central in Marxist geography, but there was an openness to alternative perspectives

and practices that arose in large part from the close connections that had developed between geography and urban planning at local universities such as UCLA and USC. These connections linked theorizing the political economy of urban restructuring to the practical demands of public policy, urban politics, and local community activism.

Developing from these early studies was a conceptual framework that saw in the restructuring of Los Angeles the emergence of a new era of urban and regional development affecting cities all over the world in the aftermath of the tumultuous urban crises of the 1960s. This new kind of urbanism and urban geography was interpreted as the product of deep changes in the local, regional, and national economy associated with the rise of *post-Fordism*, increasingly *flexible* methods of industrial production, and the powerful local effects of *globalization* and new information technologies (Scott 1988; see also *Economic Geography*; *Industrial Geography*; *Technology Districts*; *Economic Restructuring: Geographic Aspects*; *Information Society, Geography of*). Built in to this primarily economic conceptualization of urban restructuring was also a critical postmodern interpretation that saw in the dramatic changes taking place in the century's last three decades the rise of a distinctively postmodern as well as post-Fordist urbanism.

In 1986, a special issue devoted to recent research on Los Angeles appeared in the journal *Environment and Planning D: Society and Space*, with an editorial introduction by Scott and Soja describing Los Angeles as the most representative, if not paradigmatic, city of the late twentieth century. The journal, with its founding editor Michael Dear and close editorial ties to British geographers such as Derek Gregory and Nigel Thrift, had quickly become the major journal for current debates relating geography to social theory and cultural studies since its origins four years earlier. In this special issue, two articles appeared, one by Dear and the other by Soja, that were the first explicit statements arguing for a postmodern perspective in geography and planning. Drawing on the changing geography of Los Angeles and the innovative urban scholarship of Jameson and Lefebvre, the authors described postmodernism in three related ways: as a *method* of textual and visual representation, as a contemporary *condition* of everyday life, and as a comprehensive epistemological *critique*. In each of these approaches to postmodernism in geography, emphasis was placed on the distinctiveness of the present era, on what is new and different in the contemporary world, an emphasis that in the broadest sense defines a postmodern perspective.

Defined as a method of textual and visual representation, postmodernism aimed at the intentional *deconstruction* of what were seen as outmoded modernist practices of writing geography, with the intent of opening up creative new alternatives. Rather than

accepting the impress of powerful continuities with the past, for example, there would be playful and disruptively sequenced historical references, accentuating the synchronic and the simultaneous rather than a sequential diachronic narrative. The orderliness of visual form and textual narrative was similarly deconstructed through collage, pastiche, interjected verse, and other unruly modes of literary and visual representation. Explanation and interpretation thus became increasingly involved in *discourse analysis*, with geographies seen as complexly constructed 'texts' to be discursively 'read' as simultaneously real and imagined, a spatial version of the language games and deconstructive practices of such leading postmodernist philosophers as Lyotard, Derrida, and Foucault.

Postmodernism defined as an empirical condition of contemporary life was much easier to accept in geography than what appeared to many to be merely a fashionable and unduly confusing style of writing. That at least part of the world had entered what could be described as a postmodern era and that there were *postmodern geographies* and a *postmodern urbanism* worthy of being studied as such posed relatively little challenge to most geographers. The growing role of economic and industrial geographers in making practical and theoretical sense of the restructuring processes affecting the contemporary world reinforced this widening impact, and helped to create a close association between the geographical condition of postmodernity and the impact of post-Fordist economies of flexible production, globalization processes, and the information society. Just how extensive this empirical postmodernization was, whether it was positive or negative or both, and, especially, how it should affect the practice of geography beyond the description of places like Los Angeles remained much more controversial.

The greatest resistance arose, not unexpectedly, over postmodernism as a deep epistemological critique, especially after subsequent writings by Dear (1988) and Soja (1989) pushed this critique into the very core of geography as a discipline. They argued that all forms of explanation in (modern) geography were fundamentally flawed, for, among other reasons, they rested on epistemological assumptions that excluded too many important issues and voices, especially related to gender, race, class, and what postmodern critics call the marginalized 'Other'; were based on all-encompassing or totalizing metanarratives that too confidently proclaimed the possibility of obtaining complete and accurate knowledge about human geographies; and were excessively bound up in oppositional binaries that left little room for alternative forms of knowledge production. Without abandoning modernism and modern geography entirely, Dear and Soja contended that the restructuring of empirical geographies that had been taking place in the last third of the twentieth century had so changed the contemporary world that established methods of analysis

and explanation in geography, postpositivist or otherwise, could no longer be so confidently relied upon. New and different approaches, an assertively postmodernist geography rather than simply a descriptive geography of the postmodern condition, needed to be opened up and explored.

3. *Postmodernism in Geography: 1989–Present*

Two books published in 1989 were particularly influential in shaping the directions that would be taken by postmodernism in geography in the 1990s: Soja's *Postmodern Geographies: The Reassertion of Space in Critical Social Theory* and Harvey's *The Condition of Postmodernity: An Enquiry into the Origins of Cultural Change*. Both authors were rooted in Marxist geography and political economy, and shared a deep commitment to the creation of a historical materialism that was simultaneously profoundly geographical and spatial. Where they differed most was in the degree to which this construction of a historical and geographical materialism required a critical deconstruction and reconstitution of the ontology and epistemology of Marxist analysis itself.

Harvey adaptively recognized the challenge to radical geographical analysis posed by the postmodern condition but remained confident in the persistent power of historical (and geographical) materialism as a critical methodology for understanding and changing the contemporary world. There were many aspects of the condition of postmodernity that were new and different from the past, but they could best be understood through a more geographical Marxism that viewed postmodernity as just the most recent phase of capitalist development rather than as a disruptive or epochal break with the past.

Soja was more avowedly a postmodernist, insisting that a much deeper restructuring of Marxism was necessary to engage effectively with contemporary geographical realities. He called for a radical rethinking of two fundamental or ontological premises embedded in Marxism, one relating the social and the spatial dimensions of capitalist societies and the other linking time and space, history and geography. Following Lefebvre, Soja argued that there was a more balanced and mutually causal relation between society and space, a sociospatial dialectic that had been submerged in Marxism's overemphasis on social relations of production and aspatial class analysis. Similarly, he saw a persistent privileging of history over geography, the temporal over the spatial, in a particular form of *historicism* that severely constrained the development of a balanced and mutually interactive historical and geographical materialism. He called instead for a more balanced and three-sided ontology and epistemology that dynamically related the spatial, social, and historical dimensions of human

existence (spatiality, sociality, and historicity), with no one of the three inherently privileged over the others (Soja 1996).

The contrasting positions taken by Harvey and Soja on the implications of postmodernism were virtually ignored in the immediate reaction in geography to the two books. Instead, their publication by two prominent male geographers authoritatively writing about postmodernity provoked a concerted response from feminists to what was perceived as another round of masculinist appropriation of the discipline's leading edge of research and theory. Both were accused of 'flexible sexism' (Massey 1991), of living in 'Boy's Town' hopelessly 'lost in space,' of imposing their authoritative power too forcefully, no matter how different their positions on postmodernism. Many of the harshest feminist critics shared an interest in Marxist geography and in the postmodern critiques of modernism, but were suspicious of the particular directions suggested by Harvey and Soja for the future of critical human geography.

The end result of this confrontation and critique was both a deflection and a *refraction* of the impact of postmodernism. A new generation of geographers, especially in Great Britain, set the confrontation aside to work in many different ways at applying postmodernist and poststructuralist perspectives to geographical analysis and interpretation, avoiding both the narrowed historical materialism of Harvey's approach to the postmodern condition and the overly enthusiastic celebrations of postmodernism associated with the work of Soja and Dear.

To an unusual degree, some version of the postmodern epistemological critique (often subsumed under the safer sounding label of poststructuralism) was absorbed into an increasingly eclectic critical human geography in the 1990s, but more as a justification for theoretical and methodological pluralism than as an encompassing or integrative philosophy. Continuing debates about explanation in geography were significantly expanded in scope and increasingly focused on such issues as the construction of *difference* and *identity*; the relations between *space*, *knowledge*, and *power*; the pervasive problems of *representation*; and the increasing confusion of the real and the imagined in *cyberspace* and *hyper-reality*. Growing attention was given to different scales of geographical analysis, ranging from studies of the human body, the most intimate geography, to the new landscapes being produced by the globalization of capital, labor, culture, and information flows. Nearly every subfield became opened up to a multiplicity of interpretive perspectives, including cartography (see Harley in Barnes and Gregory 1997) and geographical information systems (Pickles 1994), with little concern for establishing an overarching theorization or disciplinary paradigm.

Feminist geographers such as Doreen Massey (1994), Gillian Rose (1993), and Linda McDowell

(1993) were particularly influential in shaping the continuing and open-ended postmodernization of geography, cautiously promoting a gender-based postmodern critical epistemology as part of the widening pluralism of interpretive methodologies. Similarly, Kathy Gibson and Julie Graham (writing as Gibson-Graham 1996) called for new directions in Marxist geography based on their reinterpretations of contemporary feminist, postmodernist, and poststructuralist theory. The new cultural geography, building on the early work of Cosgrove and Daniels (1988), Duncan (1990), and Barnes and Duncan (1992), moved increasingly closer to critical cultural studies and deconstructive discourse analysis, with perceptive readings of cultural landscapes as texts and 'signifying practices.' Cultural and political economy perspectives were partially reconnected through a growing interest in the *postcolonial critiques* of scholars such as Edward Said, Gayatri Spivak, Arjun Appadurai, and Homi Bhabha (Gregory 1994, Soja 1996), while geographical studies of the body, sexuality, the unconscious, and the formation of human subjectivity, drawing on the psychoanalytical theories of Lacan as well as the ideas of Foucault and Lefebvre, created other kinds of bridges between subjective place-based culture and spatial political economy (Pile and Thrift 1995, Pile 1996).

There has also developed a vigorous antipostmodernism in the 1990s, even among those whose work most geographers would associate with a postmodern perspective. Many Marxist geographers, for example, continuing to follow David Harvey's lead (see Harvey 1996), have narrowed postmodernism almost entirely to being a politically diverting and divisive aspect of the contemporary *condition* of global capitalism. Efforts to continue to press a postmodern epistemological critique are therefore seen as politically retrogressive and antagonistic to the radical project of historical and geographical materialism. At the other end of the political spectrum, more conservative geographers frequently reduce postmodernism to superficial and often foolish playfulness, blunting any challenge to their established convictions. But very few geographers today are without an opinion on postmodernism or unaware of its impact on the discipline.

As the new century begins, it can be argued that postmodernism and the interpretive approaches associated with it (poststructuralism, postcolonial critiques, post-Marxism, postmodern feminism, methods of deconstruction and discourse analysis) have been absorbed into geography as an integral part of the contemporary discipline (see such recent overviews of modern and postmodern geography as Barnes and Gregory 1997, Benko and Strohmayer 1997, Peet 1998, Massey et al. 1999; see also Jones et al. 1994 and Watson and Gibson 1995). Postmodernism has been blunted and deflected in its boldest claims, continues to be actively resisted by many, but remains part of

the consciousness and imagination of most human geographers.

The achieved and continuing impact of postmodernism can be summarized in several ways. First, it has helped to bring geography into closer contact with social theory and philosophy, and nearly every branch of the social sciences and humanities, from film studies and literary criticism to anthropology and economics. This has contributed in turn to what might be called a 'de-disciplining' of geography: an opening up of its traditional borders with other fields; a wider diffusion of its core concepts and ways of studying space, place, region, and environment; and the emergence of an extraordinary pluralism within geography with respect to theory, epistemology, and methods of empirical analysis. Intense frictions and disagreements remain, to be sure, but there may be no other time in its existence as a distinct discipline when geography has been so paradigm-free and yet so critically engaged with the major issues and events of our times. While these developments are not entirely due to the impact of postmodernism, it has certainly played a stimulating role.

Bibliography

- Barnes T J, Duncan J (eds.) 1992 *Writing Worlds: Discourse, Text, and Metaphor in the Representation of Landscape*. Routledge, New York
- Barnes T J, Gregory D (eds.) 1997 *Reading Human Geography: The Poetics and Politics of Inquiry*. Arnold, London
- Benko G, Strohmayer U (eds.) 1997 *Space and Social Theory: Interpreting Modernity and Postmodernity*. Blackwell, Oxford, UK
- Cosgrove D, Daniels S 1988 *The Iconography of Landscape*. Cambridge University Press, Cambridge, UK
- Dear M 1988 The postmodern challenge: Reconstructing human geography. *Transactions I B. Geography* 13: 262-74
- Dear M 1986 Postmodernism and planning. *Environment and Planning D: Society and Space* 4: 367-84
- Duncan J S 1990 *The City as Text: The Politics of Landscape Interpretation in the Kandyan Kingdom*. Cambridge University Press, Cambridge, UK
- Gibson-Graham J K 1996 *The End of Capitalism (As We Knew It): A Feminist Critique of Political Economy*. Blackwell, Oxford, UK
- Gregory D 1978 *Ideology, Science, and Human Geography*. Hutchinson, London
- Gregory D 1994 *Geographical Imaginations*. Blackwell, Oxford, UK
- Gregory D, Walford R (eds.) 1989 *Horizons in Human Geography*. Macmillan, London
- Harvey D 1989 *The Condition of Postmodernity: An Enquiry into the Origins of Cultural Change*. Blackwell, Oxford, UK
- Harvey D 1996 *Justice, Nature and the Geography of Difference*. Blackwell, Oxford, UK
- Ley D, Samuels M (eds.) 1978 *Humanistic Geography: Prospects and Problems*. Maaroufa Press, Chicago
- Jameson F 1984 Postmodernism, or the cultural logic of late capitalism. *New Left Review* 146: 53-92

- Jones J P, Natter W, Schatzki T R (eds.) 1994 *Postmodern Contentions: Epoch, Politics, Space*. Guilford Press, New York
- Massey D 1991 Flexible sexism. *Environment and Planning D: Society and Space* 9: 31–57
- Massey D 1994 *Space, Place and Gender*. University of Minnesota Press, Minneapolis, MN
- Massey D, Allen J, Sarre P (eds.) 1999 *Human Geography Today*. Polity Press, Cambridge, UK
- McDowell L 1993 Space, place and gender relations: Part II: Identity, difference, feminist geometries and geographies. *Progress in Human Geography* 17: 305–18
- Olsson G 1980 *Birds in Egg*, 2nd edn. Pion, London
- Peet R 1998 *Modern Geographical Thought*. Blackwell, Oxford, UK
- Pickles J (ed.) 1994 *Representations in an Electronic Age: Geography, G.I.S., and Democracy*. Guilford Press, New York
- Pile S 1996 *The Body and the City: Psychoanalysis, Space and Subjectivity*. Routledge, London
- Pile S, Thrift N 1995 *Mapping the Subject: Geographies of Cultural Transformation*. Routledge, London
- Rose G 1993 *Feminism in Geography: The Limits of Geographical Knowledge*. University of Minnesota Press, Minneapolis, MN
- Scott A J 1988 *Metropolis: From Division of Labor to Urban Form*. University of California Press, Berkeley, CA
- Soja E W 1986 Taking Los Angeles apart: Some fragments of a critical human geography. *Environment and Planning D: Society and Space* 4: 255–72
- Soja E W 1989 *Postmodern Geographies: The Reassertion of Space in Critical Social Theory*. Verso, London
- Soja E W 1996 *Thirdspace: Journeys to Los Angeles and other Real-and-imagined Places*. Blackwell, Oxford, UK
- Watson S, Gibson K (eds.) 1995 *Postmodern Cities and Spaces*. Blackwell, Oxford, UK

E. W. Soja

Postmodernism in Sociology

Prominent fields of postmodernist research have been work and organizations, political action, science and technology, commodification, consumption, gender, media, and popular culture. Classical sociological topics like symbolic consumption have been further explored and enriched. Sociological postmodernism has also pioneered domains like computers and the Internet. It is to be expected that in the future postmodernist research will evolve in the direction of concrete field studies, with an emphasis on the cultural forms induced by computers and the Internet.

In sociology, postmodernism designates (a) a cluster of theoretical and meta-theoretical approaches; (b) an analysis of postmodernity, understood as encompassing the social and cultural features of late capitalism; (c) an extension of sociological inquiry to new domains; and (d) new forms of sociological expression. Sociological postmodernism is thus a form of sociological analysis, a kind of sociological sensibility, and a sociologists' social and intellectual condition at the same time. Its varieties share a series of theoretical

and methodological premises, but differ in their conclusions and research programs. Core common elements are: (a) avoiding recourse to a set of universally valid assumptions as theoretical and methodological foundations, together with (b) the key role ascribed to notions like subject, identity, text, and symbol in the analysis of society.

1. Origins of Postmodernism

The term 'postmodernism' has been used in Latin-American literary criticism since the 1930s, and in Anglo-American debates since the 1940s, in order to designate new forms of expression in their relationship with the aesthetic of modernism. In history, the term was used by the British historian Arnold Toynbee in *A Study of History* in 1947 and designated the latest phase of Western civilization. In sociology, it was introduced by Amitai Etzioni's book *The Active Society* in 1968. Starting with the early 1970s, the term appeared more and more frequently in sociological texts in France, North America, and the UK. At that time, it was already common stock in literary theory and criticism. An essay written by the French philosopher Jean-François Lyotard, *The Postmodern Condition: A Report on Knowledge* (1979) played a seminal role in making this term widely known and used in the social sciences. Therefore, postmodernism was partially a conceptual import; its diffusion in the 1970s was part of a wider innovative movement in the discipline, marked, among others, by the rise of social constructivism and feminism. At the same time, it continued and reformulated a series of topics already present in sociology.

The intellectual roots of sociological postmodernism can be identified in the works of some key nineteenth and early twentieth century philosophers, sociologists, and linguists. Karl Marx's critique of capitalism, Georg Simmel's analysis of modernization processes, Friedrich Nietzsche's critique of epistemology, and Ferdinand de Saussure's theory of language play here a prominent role. The French structuralist movement (Claude Lévi-Strauss's anthropology, Roland Barthes's semiotic theory, Jacques Lacan's psychoanalysis), and the poststructuralist one (Michel Foucault, Jacques Derrida), Ludwig Wittgenstein's later works, and Clifford Geertz's anthropology have also played a considerable role. Among other influences there are the Frankfurt School (Walter Benjamin, Theodor Adorno, Max Horkheimer, and Herbert Marcuse), symbolic interactionism, ethnomethodology, and phenomenological sociology.

2. The Tenets of Postmodernism

There are two main tenets, mirroring the distinction between postmodernism and post-modernity: while the latter designates the social and cultural features of

late capitalism, distinguishing them from earlier phases (modernity), the former designates a theoretical and methodological program, as distinct from earlier (modernist) ones.

2.1 The Analysis of Postmodernity

The analysis of postmodernity developed at the beginning of the 1970s, continuing a preoccupation already found in the works of Georg Simmel, Thorstein Veblen, and the Frankfurt School. Postmodernity is a late phase in the evolution of the capitalist system, seen either as a continuation and intensification of modernity, or as a substantially new stage. It is characterized by the transition from manufacture to a decentralized services economy; science and complex technological systems play a key role here. Decentralization, together with the pervasiveness of expert and technological systems in all domains of everyday life lead to a fragmentation and diversification of social identities. The electronic media (television, computers, video) are prominent in the constitution and presentation of social reality; visual and linguistic symbols become pervasive, while popular cultural forms diversify and expand. Society is therefore fragmented and heterogeneous; social order is always local and contextual, and works as an assemblage of symbolic codes. Production and consumption, social stratification, as well as personal and collective identities become dependent on such codes. They are understood as a support, a medium for, and an outcome of social action. This notion becomes a general explanatory paradigm: every domain of societal organization can be seen as a symbolic system and analyzed accordingly.

Consequently, sociology (an outcome of modernity itself) should take fragmentation, decentralization, relativization, and the mediated character of social reality into account; it should adapt its own tenets, claims, as well as its own modes of analysis and representation to these changes. Starting from these premises, Zygmunt Bauman (1992) has argued that notions like sociality, habitat, self-constitution, and body are better suited for the analysis of contemporary social processes than the traditional concepts of society and order.

2.2 The Critique of Modernist Sociology

During the 1980s, a second theme made its way from philosophy and literary theory into sociological debates. It was a critique of the status of sociological theory as a modernist product, and it drafted a new methodological program. This debate raised a series of questions about the validity claims of scientific theories, the nature of scientific truth, and the status of the knowing subject, arguing that: (a) all knowledge is contextual and local; (b) the validity claims of any

scientific theory are not to be found in some abstract, universal criteria, but are rather the results of either negotiated consensus or power struggles; and (c) as a consequence, the knowing subject does not dispose of universal criteria to ascertain the validity and truth of his or her knowledge. The subject should not be seen as autonomous with respect to an objective, given world, but rather as the product of social circumstances. Since knowledge has a contextual and local character, it follows that scientific theories are re-interpreted according to local conditions: they are open-ended and cannot be controlled by their authors. Scientific theories are not determined by external validity and truth criteria; they work as systems of symbols enabling certain courses of social action and promoting a certain worldview. On these grounds, there should be a structural similarity between texts and social order: both are locally determined, network-like, open-ended systems of symbols requiring certain kinds of action from the actors embedded in them.

Notions like text and performance should occupy a prominent place in the analysis of social phenomena. Sociology should focus on investigating and describing how symbol systems work in particular social contexts, instead of producing all-encompassing theories. This claim, together with the one that the knowing subject is the product of social circumstances, have sometimes been described in a more radical form as 'the end of sociological theory' and the 'death of the author,' respectively. A corollary of these arguments is that sociological theories should acknowledge the fragmentation and diversity inherent in the local character of knowledge; they should explore their consequences for the world we live in, and should experiment with new modes of representation—like poetry, or dialogical forms. To some extent, these epistemological claims overlap with those of the first strand of postmodernist sociology. However, we encounter an important difference: while the first strand derives the fragmentation, relativization, and diversification of knowledge from social changes, the second one takes them as stemming out of a number of epistemological, universally valid first principles.

At the end of the 1980s and in the early 1990s, such meta-theoretical topics were intensely debated in sociology and, for a short time, even eclipsed the first strand of postmodernism. However, the acceleration of globalization processes brought social analysis to the forefront again.

3. The Methodology of Postmodernist Sociology

One important methodological aspect is the anti-positivist and antiverificationist stance: there is no totalizing explanatory model and no universally valid verification procedure. Postmodernist sociology stresses the role of qualitative methods in the analysis of contemporary social phenomena, as a consequence

of the prominence of visual and linguistic symbols. While some authors (Richardson 1997) emphasize interpretation and subjectivity, others (Lemert 1997) argue that methodology must preserve objectivity and logical consistency. There has been an increased sociological interest for discourse analysis, understood as a cluster of techniques for grasping the emergence, functioning, and context-boundedness of symbolic codes. Ethnographic methods, earlier reserved for the study of 'exotic' societies or of marginal social phenomena, also play an important role. This has increased the interest for texts and writing processes as sources of conceptual enrichment in sociology. Poetry, drama, or dialog are promoted as legitimate complements to the established, academic expository modes. Hybrid forms of sociological expression have appeared too—for example, in works combining academically written texts, poetry, and artwork (Haraway 1997).

4. *Fields of Postmodernist Research*

Postmodernist analyses have been prominent in domains like work and organizations, political action, science and technology, commodification and consumption, gender, media, and popular culture. George Ritzer (1993), for example, has examined the new forms of work which emerge in the knowledge-based, global economy: they are characterized by fragmentation, local specialization, flexibility, and mobility, on the one hand, and by global dispersion and typification, on the other. Under these circumstances, the temporal and spatial conditions of work change substantially: while the lifespan of products shortens, they need to be replaced faster, and worktime expands. Tensions arise between requirements for flexibility and mobility, on the one hand, and typification, on the other. New directions in commodification and consumption complete this picture. Commodification is understood as the unrelentless expansion of market exchange, which reshapes social relationships according to its own logic; the pendant of this process is the increased social significance of the act of consumption. Jean Baudrillard (1972), for example, has argued that symbolic consumption becomes one of the main social mechanisms through which individual and collective identities, as well as social relationships are defined and reproduced.

In political sociology, the analysis has focused on political discourses and symbolisms, on the spectacular form taken by political events, and on the role of television as a medium for political action. Another direction has been to examine the strains put on democracy by the media-induced forms of political action, and by the globalized economy. Authors like Chantal Mouffe (1992) have asked whether the new economic and political conditions do not require devising new forms of democratic participation and of political action.

With knowledge lying at the core of the postmodern society, science has been a domain of sustained investigation. The focus has been on the changes brought about by the new information and biomedical technologies in social life, as well as on the role and effects of complex technological systems. Bruno Latour (1991) and Donna Haraway (1997), among others, have argued that such systems are so pervasive and deep-reaching, and that social life is dependent on them to such an extent that new forms of social symbiosis have emerged; these are described with the help of notions like the cyborg society, hybrids, or networks of human and nonhuman actors. These authors have called for a revision of the conceptual apparatus of sociology on the grounds of the relationships between human actors and technical systems.

A further domain of investigation, situated at the crossroads between feminist and postmodernist sociology, has been that of gender and sexuality. Postmodern diversification brings forth a fragmentation and multiplicity of gender and sexual identities. The distinctions between genders become blurred; some forms of sexual identity formerly considered marginal gain a new significance, and new sexual identities emerge (e.g., transsexuality). Owing to the local and contextual character of knowledge, the gender specificity of knowledge and experience has to be taken into account. Several authors (e.g., Smith 1990) have explored the ways and the implications of gender-specific knowledge in the workplace, the field of politics, popular culture, and the public sphere.

One of the key domains in which the impact of postmodernist sociology has been felt is that of the electronic media and communication. Since the social changes brought about by computers and television have been considerable, questions arise about new, media-induced forms of social life and organization. According to Jean Baudrillard, among others, large portions of the social world are now media generated, and whole series of political and social events have a media-supported existence: they exist as simulacra, as symbolic codes which can be reproduced endlessly, without making recourse to an original. Other domains of investigation have been computers in the workplace, or the media-induced forms of social life, like those arising in and around the Internet (Turkle 1995). Finally, postmodernist sociology has also considerably influenced the subfields of cultural and urban studies, reinvigorating a research tradition which goes back at least to the Chicago School of sociology.

5. *The State of Postmodernist Research*

Postmodernist sociology has been around for approximately three decades; while controversial and contested for some of its more radical theoretical stances, many postmodernist ideas have made inroads into mainstream sociology. At the same time, it has

continued and enriched a series of themes already present in classical sociology—like the sociology of consumption, or the study of popular cultures. It has also pioneered some new domains of research—the study of Internet cultures, of the media, or (together with ethnomethodology) that of the impact of computers. Together with feminist sociology, it has had a considerable influence on the sociology of gender and sexuality. Its empirical, domain-oriented impact has been more significant than that of its pure theoretical tenets.

6. New Research Directions

The cultural forms induced by the electronic media (e.g., the culture of software programmers, of Internet users) have not been studied in depth until now. The changes in consumption patterns and market relationships brought about by the Internet, the rapid rise of popular electronic exchanges, the Internet's role in the globalization processes are topics which await deeper exploration too. It is to be expected that postmodernist sociology will evolve more and more in the direction of empirical, concrete field studies, focusing on these phenomena.

See also: Culture; Consumption, Sociology of; Cultural Relativism, Anthropology of; Feminist Epistemology; Feminist Theory: Postmodern; Globalization, Anthropology of; Identity: Social; Modernity; Political Discourse; Postmodern Urbanism; Postmodernism in Geography; Postmodernism: Methodology; Postmodernism: Philosophical Aspects; Relativism: Philosophical Aspects; Sociology, Epistemology of; Structuralism, Theories of

Bibliography

- Baudrillard J 1972 *Pour Une Critique de l'Économie Politique du Signe*. [For a Critique of the Political Economy of the Sign]. Gallimard, Paris
- Bauman Z 1992 *Intimations of Postmodernity*. Routledge, London
- Etzioni A 1968 *The Active Society: A Theory of Societal and Political Processes*. Collier-Macmillan, London
- Haraway D 1997 *Modest Witness @ Second Millenium: FemaleMan® Meets OncoMouse™; Feminism and Technology*. Routledge, New York
- Lash S (ed.) 1991 *Poststructuralist and Postmodernist Sociology*. Elgar, Aldershot, UK
- Latour B 1991 *Nous n'Avons Jamais Été Modernes*. [We Have Never Been Modern]. La Découverte, Paris
- Lemert C 1997 *Postmodernism Is Not What You Think*. Blackwell, Cambridge, MA
- Lyotard JF 1979 *La Condition Postmoderne: Rapport sur le Savoir*. [The Postmodern Condition: A Report on Knowledge]. Editions de Minuit, Paris
- Mouffe C 1992 *The Return of the Political*. Verso, London
- Richardson L 1997 *Fields of Play*. Rutgers University Press, New Brunswick, NJ

- Ritzer G 1993 *The McDonaldization of Society: An Investigation into the Changing Character of Contemporary Social Life*. Pine Forge Press, Newbury Park, CA
- Smith D E 1990 *Texts, Facts, and Femininity. Exploring the Relations of Ruling*. Routledge, London
- Toynbee A 1947 *A Study of History*. Oxford University Press, London
- Turkle S 1995 *Life on the Screen: Identity in the Age of the Internet*. Simon and Schuster, New York

A. Preda

Postmodernism: Methodology

If methodology is defined as referring to the foundations of knowledge, as how we perceive and understand reality, as well as how we study it, then there is not only a single implicit postmodern methodology. There are several. But all challenge the methodological assumptions associated with rigorous, modern social science inquiry, be it qualitative or quantitative (Dickens and Fontana 1994). Indeed, there are serious philosophical differences between postmodernism and the enlightenment norms of traditional science. Some postmodernists are uninterested in generalization, definitive explanation, replication, validity, reliability, etc. Neither prediction nor theory building are a major concern. They look to deconstruction as method. Deconstruction tears a text apart, revealing its contradictions and assumptions. Other postmodernists criticize modern methodology and seek to revise it, to build upon it. They look for alternatives to what they call the old, past, closed conventional social science methods inherited from the natural sciences. But they do so without entirely rejecting reason, consistency and coherence of argument. Although they abandon foundations, they retain what they call preferences.

Many postmodernists see the need for new standards for evaluating knowledge. Their criteria are likely to be subjective in nature, including, for example, flexibility, sensitivity, and interactivity. They look for beauty, strength, or force in a text. Some value elegance of expression and style, and seductiveness of content while others say that such criteria are community specific and so cannot be enumerated. Still other postmodernists willingly employ more direct valuations, such as negation of oppression, exploitation, and domination, or the affirmation of liberation, freedom, insubordination, and resistance as key standards (Rosenau 1992).

1. A Postmodern Methodological Focus

Postmodern methodology focuses on the margins and postmodern social scientists highlight the unique and the unusual. They concentrate on the enigmatic

and appreciate the unrepeatable. They seek to multiply paradox, to invent ever more elaborate repertoires of questions as much as to determine answers. They look to 'differance' (Derrida 1972) rather than sameness, regularity, and synthesis. 'Differance' is a structuring principle that tells us that definition rests not on some inherent characteristics of a thing but rather in terms of its negative and positive references to other texts. It also suggests that in the end, definition is never final, always postponed, deferred forever (Derrida 1981).

Postmodern methodology seems driven by substantive concerns quite different from those modern social scientists focus upon. Postmodernists direct their attention to everyday life and they gravitate toward mininarratives or stories. Postmodern social science is a 'systematic commentary on the knowledge of daily life' (Bauman 1990). This means examining what is assumed to be quite ordinary and attributing a new importance to it. Postmodern social science makes room for the unexpected, the mystical, and the metaphysical. Concepts such as fate and spirituality are, once again, legitimate (Griffin 1988).

1.1 *Linking Substance and Method*

If there is a close association of substance and method, then conventional social science methods are not helpful given the postmodern focus on the margins, the unexpected, the paradoxical, and the unrepeatable. Empirical methods require repeatability so that careful measurement can be carried out and statistical norms met. But the subject matter of interest to postmodernists does not recur necessarily and its very interest may be in the fact that it is transient, constantly in flux, and difficult to capture. Because of their substantive foci, these postmodernists require methods that are open and flexible, rather than highly structured, codified, and technically detailed. They must be adaptable to a great variety of situations because of the range of potential application. There is no easy methodological recipe for doing this type of research.

1.2 *Going Beyond Qualitative Methods*

Postmodernists would reject any argument that their methodology is simply qualitative social science in a different guise. The gamut of modern qualitative approaches developed over the years (from observation to participant observation) retains a preoccupation with generalization and replication as much as an appreciation of complexity. Those employing conventional qualitative methods in social science have invested enormous energy in separating out what is procedurally acceptable from the unqualified. All this is abandoned in the postmodern methodology

that embodies an attitude of humility and respect. It does not try to convince others that its own methodological perspective is right, superior, truthful, objective, the best assessment, or the source of the only correct answer. It does not settle on solutions or give privileged voice to any particular, singular point of view. It does not attribute authority to authorship. It queries and interrogates constantly. It follows its own advice: no new metanarrative. A metanarrative is a global worldview, a 'mastercode' that assumes the validity of its own truth claims.

2. *An Upbeat Postmodern Methodology*

What methodological alternatives does postmodernism have to offer vis-à-vis modern social science methodology? It emphasizes personal wisdom, lived experience, and knowledge as rich sources of material that inform writing and research. We are counseled to look to music, art, poetry, and literature instead of census data and public opinion polls. A postmodern social scientist would have more confidence in feelings, faith, empathy, emotion, insight, sensitivity, intuition, imagination, subjective judgment, and introspection as method. Postmodern social scientists are obviously and openly more comfortable with these than with an exclusive confidence in reason and logic. But by praxis, if not discourse, some postmodernists communicate that their alternative methodologies are of most value if disciplined by reason and logic. This amounts to coming to terms with the best of the Enlightenment heritage but without being intimidated by it. It suggests a refusal to accept the authority of modern science but also a desire to salvage from it whatever is of interest, such as the possibility of an interactive and reciprocal relationship between, say, reliable knowledge and meaningful participation in the public sphere, the creation of social life and political action, the enhancement of personal development, self-expression, self-determination and collective praxis (Jennings and Graham 1996). Doing social science might also have to be fun, enjoyable, entertaining rather than somber, listless, and mundane (Bakhtin 1973, Todorov 1984). It would not seek to demonstrate the obvious so much as to highlight the overlooked and the unrecognized.

2.1 *Methodology as Art and Literature*

Postmodern social science is often viewed as vision and art more than trained observation. It disperses any clear distinction between self and other; fact and values; author, text, and reader. It is inspired by the cultural and contextual preoccupations of anthropology where the distinction between literature and

science was never clearly established. Anthropology had to be content with differing 'evocative devices of metaphor and imagery' because it was so difficult to arbitrate between different narratives from the past (Richardson 1988). Postmodern social science methodology centers on 'listening to,' and 'talking with' the other (Tyler 1986). The reader 'writes' the postmodern text and in so doing postmodern social science gives voice to those previously silent. It empowers through the methods of interpretation and criticism.

Postmodern methodology employs a literary capacity and it re-invigorates the mininarrative as a story that is worthy of notice even if it is not necessarily true in the modern sense of science. Storytelling is both a method and a means to emancipatory power (Newman and Holzman 1999). The need to reconcile various stories, narratives or research findings may or may not be urgent. The postmodernism tolerance for diversity and its indulgence of discrepancy reflect the fact that there is no rush to create, to construct, to accumulate, and to impose new knowledge in the modern sense.

2.2 Methodology as Interpretation

Postmodern methodology is interpretive but postmodern interpretation bears but slight resemblance to the more conventional social science forms of interpretation. Modern social science understands interpretation to be the careful consideration of data with the goal of locating patterns. Meaning is not arbitrary and some interpretations are assumed to be better than others. But neither does postmodernism resemble hermeneutic interpretation that is more subjective and less ambitious than modern social science's interpretation. Hermeneutic interpretation, unlike postmodern interpretation, seeks to probe the 'silences,' to uncover a deeper meaning, masked and hidden perhaps, but waiting to be discovered (Dreyfus and Rabinow 1983, Madison 1988). There is a current within postmodernism that says there simply is no 'meaning' to uncover. In many ways hermeneutic interpretation accepts truth, universality, and a shared discourse (a community of understanding).

Postmodern interpretation is *introspective*, *intersubjective*, and *anti-objectivist*, a form of individualized understanding. Postmodern theories of interpretation no longer 'ground,' they rather 'mediate' (Kugler 1988). They range from 'interpretative reason' (Bauman 1990) to prayer and chanting as 'interpretative communication.'

2.3 Methodology for a Complex Intertextual World

Postmodernists see the world as so complex that any direct causal statements relating to reality would

necessarily be simplistic. They argue that it is impossible to establish temporal priority in a simple fashion. They similarly are suspicious of modern versions of time that clearly delineate the past from the present, and include a strict version of linear time or geography. Their preference is for a present-oriented social science that erases linear time and ignores strictly bounded forms of geography. Their analysis emphasizes postmodern intertextuality, including, for example, chaos theory with its view of nonlinear social dynamics, and a definitely more complex version of reality (Young 1992).

If the social universe for these postmodernists is intertextual, this has an impact on their methodology. Relationships, interconnections, and interactions are areas of substantial interest. Conventional research tools are hardly appropriate. Quantitative methods isolate and simplify what is being studied, holding as many extraneous variables as possible constant so as to tease out causal, predictive associations. These postmodernists seek methodologies that integrate what is being studied with all that is related to it even if this complicates the research task, because in the end explanation in any absolute sense may be impossible.

2.4 Relativism and Questioning Postmodernists

Postmodernism is plagued, because of its European philosophical precursors, with the charge of relativism (Gellner 1992) but many postmodernists defy this heritage and reject relativism. This is especially true for the postmodernists who come from a philosophically committed heritage: Marxism, feminism, environmentalism, peace movements, ecology groups, and religious studies (Griffin 1988, 1990, Kumar 1995, Leonard 1997). This leads them to argue that while deconstruction is an interesting first step to a postmodern methodology, it is limited because its goal is to undo all constructions. Deconstruction's intent is not to improve, revise, or offer a better version of the text, so they reject it. Postmodernism, in this view, must ultimately move to reconstruction and by extension, methodology remains essential.

3. Methodology Defined Narrowly

If methodology is defined narrowly as the rules and regulations of inquiry, then postmodernists must question the whole undertaking, arguing that there is no such thing as a post-modern methodology. 'Rules,' as such, are modern, not postmodern in character because they imply hierarchy, a series of good-bad evaluative criteria, an assessment on the basis of a set of agreed upon standards. Postmodernists reject rule-bound approaches as 'metanarratives.' Critics of

postmodernism, however, respond that the absence of rules is itself a rule, the rule that 'anything goes' (Feyerabend 1975, Heller 1987) and that there is, thus, a postmodern methodology.

4. Rejecting Methodology Altogether

Postmodernists, inspired by Heidegger and Nietzsche, are antiscience and anti-Enlightenment, and this leads to an absolute relativism and to methodological nihilism. This postmodern orientation emphasizes the hopelessness and meaninglessness of life; theirs is the pessimistic, dark side of postmodernism (Rosenau 1992).

Some such postmodernists practice epistemological impossibilism. They do not intend to provide 'clear reasoned defense.' They tell us their texts offer 'no consistent argument' and that 'not all points made are "necessary."' They boast that by design 'some essential matters are forgotten' (Corlett 1989). To insist that these postmodernists develop a self-conscious methodology is to 'do violence' to their views, to impose one's own perspective, and thus diminish them by forcing them to acknowledge your views (Ashley and Walker 1990a, 1990b, Lyotard 1988). In the end, from this postmodernist methodological perspective, we can never really know anything, not even our own feelings and emotions. These postmodernists speak of the death of the author (as agent) and of the absence of the subject (as a real person) with a specific identity. This means that they escape agency, accountability, and responsibility. They deny causality, question efforts to represent reality (through social science theory), and they contend that all interpretations are equally interesting (Miller 1981).

All textual meaning, all interpretation, is undecidable within this postmodern perspective, because there is no final sense of a text and everything is a text in this postmodern world. No interpretation can be regarded as superior to any other (Latour 1988). There is a multiplicity of readings and these postmodernists refuse to privilege one statement over another. 'Every understanding is a misunderstanding' (Culler 1982, Norris 1988). Within this intellectual framework there is no need to bother about methodology at all.

5. Conclusion

Postmodern methodologies will be judged in the long term by their product. In the arts and literature, postmodern methods can be fascinating, challenging, and interesting. They stimulate creativity and innovation as exemplified by the writings of Umberto Eco, Italo Calvino, Milorad Pavic, and Salman Rushdie. They dissolve the boundary between fiction and social critique as the works of Jean Baudrillard and Michel

Foucault so well illustrate. The fruitfulness of postmodern methodologies in the applied social sciences is yet to be demonstrated. Many feel them to be seriously flawed and epistemologically dubious. Post-modern methods contributed little toward the solution of the greatest social problems of the twentieth century. But then, that is not what they set out to do; it was never the intent.

See also: Deconstruction: Cultural Concerns; Feminist Theory: Postmodern; Globalization, Subsuming Pluralism, Transnational Organizations, Diaspora, and Postmodernity; Postmodernism in Geography; Postmodernism in Sociology; Postmodernism: Philosophical Aspects; Symbolic Interaction: Methodology

Bibliography

- Ashley R K, Walker R B J 1990a Reading dissidence/writing the discipline: Crisis and the question of sovereignty in international studies. *International Studies Quarterly* **34**: 367-416
- Ashley R, Walker R B J 1990b Speaking the language of exile: dissident thought in international studies. *International Studies Quarterly* **34**: 259-68
- Bakhtin M 1973 *Problems of Dostoevsky's Poetics*. Ardis, Ann Arbor, MI
- Bauman Z 1990 Philosophical affinities of postmodern sociology. *The Sociological Review* **38**: 411-44
- Corlett W 1989 *Community without Unity: A Politics of Derridian Extravagance*. Duke University Press, Durham, NC
- Culler J 1982 *On Deconstruction: Theory and Criticism after Structuralism*. Cornell University Press, Ithaca, NY
- Derrida J 1972 *Marges de la philosophie*. Editions de Minuit, Paris
- Derrida J 1981 *Positions*. University of Chicago Press, Chicago
- Dickens D R, Fontana A (eds.) 1994 *Postmodernism and Social Inquiry*. Guilford Press, New York
- Dreyfus H L, Rabinow P 1983 *Michel Foucault, Beyond Structuralism and Hermeneutics*, 2nd edn. University of Chicago Press, Chicago
- Feyerabend P 1975 *Against Method: Outline of an Anarchistic Theory of Knowledge*. New Left Books, London
- Gellner E 1992 *Postmodernism, Reason, and Religion*. Routledge, London
- Griffin D R (ed.) 1988 *The Reenchantment of Science: Postmodern Proposals*. State University of New York Press, Albany, NY
- Griffin D R (ed.) 1990 *Sacred Interconnections: Postmodern Spirituality, Political Economy, and Art*. State University of New York Press, Albany, NY
- Heller A 1987 Mouvements culturels et changements des modes de vie quotidienne depuis la deuxième guerre. In: Corten A, Tahon M B (eds.) *La Radicalité du quotidien*. VLB éditeur, Montreal, PQ
- Jennings L E, Graham A P 1996 Postmodern perspectives and action research: reflecting on the possibilities. *Educational Action Research* **2**
- Kugler P 1988 From modernism to post-modernism: Some implications for a depth psychology of dreams. *Psychiatric Journal of the University of Ottawa* **13**: 60-5

- Kumar K 1995 *From Post-Industrial to Post-Modern Society: New Theories of the Contemporary World*. Blackwell, Oxford, UK
- Latour B 1988 *The Pasteurization of France*. Harvard University Press, Cambridge, MA
- Leonard P 1997 *Postmodern Welfare*. Sage Publications, Thousand Oaks, CA
- Lyotard J F 1988 *L'inhumaine Causerie sur le Temps*. Galilee, Paris
- Madison G B 1988 *The Hermeneutics of Postmodernity: Figures and Themes*. Indiana University Press, Bloomington, IN
- Miller J H 1981 The disarticulation of the self in Nietzsche. *The Monist* 64: 247–61
- Newman F, Holzman L 1999 Beyond narrative to performed conversation. *Journal of Constructivist Psychology* 12: 23–40
- Norris C 1988 *Deconstruction and the Interests of Theory*. Pinter, London
- Richardson L 1988 The collective story: Postmodernism and the writing of sociology. *Sociological Focus* 21: 199–208
- Rosenau P M 1992 *Post-Modernism and the Social Sciences*. Princeton University Press, Princeton, NJ
- Todorov T 1984 *Mikhail Bakhtin: The Dialogical Principle*. University of Minneapolis Press, Minneapolis, PA
- Tyler S 1986 Post-modern ethnography: from document of the occult to occult document. In: Clifford J, Marcus G E (eds.) *Writing Culture: The Poetics and Politics of Ethnography*. University of California Press, Berkeley, CA
- Young T R 1992 Reinventing sociology. In: *Mission and Methods in Postmodern Phenomenology*. The Red Feather Institute, Weidman, Michigan

P. V. Rosenau

Postmodernism: Philosophical Aspects

'Postmodernism' is a term that defies simple definition. Although earlier uses are recorded, the term has been widely used since the 1960s to refer to a range of antimodernist attitudes and strategies in art criticism, architecture, and literary criticism. References to postmodernism in philosophy and social theory did not become common before they appeared in the work of Lyotard and Rorty during the early 1980s. Philosophical postmodernism quickly became associated with the idea of a 'crisis' in representation, abandonment of the belief in a universal subject of reason and value, and the rejection of unitary schemas of progress inherited from the Enlightenment. In place of these foundational motifs of modern thought, postmodernism emphasized difference, diversity and irreducible conflict in human affairs rather than unity and the possibility of harmony.

There are several reasons why postmodernism should be 'a particularly unstable concept' (Bertens 1995, p. 12). First, the concept of postmodernism stands in a variety of complex relations to other concepts such as modernism, poststructuralism, deconstructionism, modernity and postmodernity. Since

the meaning of these concepts is equally uncertain, so our understanding of 'postmodernism' will be affected by the particular meaning attached to these related concepts. Second, to the extent that postmodernism is considered a cultural movement alongside others such as romanticism, positivism, or structuralism, then it too will encompass a variety of distinct aesthetic and intellectual strategies. Bertens' *The Idea of Postmodernism: a History* begins with illustrations of the sometimes contradictory ways in which the term has been employed. For example, by 1970 modern painting had long abandoned the idea that representation or narrative were essential to its art. In explicit reaction against modernism understood as the self-reflexive exploration of the formal possibilities of painting, postmodern painting involved a reprise of representation and narrative. In a similar fashion, postmodern architecture rejected the formalism of the postwar International style and embraced historical and vernacular styles. By contrast, at least in some of its manifestations, literary postmodernism involved a move in the opposite direction away from narrative and representation towards a form of radical self-reflexivity. Bertens concludes that, depending on the artistic discipline, 'postmodernism is either a radicalization of the self-reflexive moment within modernism, a turning away from narrative and representation, or an explicit return to narrative and representation. And sometimes it is both' (Bertens, 1995, p. 5). In view of the many different phenomena to which it has been applied, it is common for discussions of postmodernism, to begin, as this one has, with the claim that the term has no clear meaning (Eco 1985, p. 65, Rorty 1999, p. 262).

1. Modernity/Modernism

Just as modernism is often regarded as a response to the process of modernization which transformed the material conditions of European society during the nineteenth century, so postmodernism is often presented as a consequence of the restructuring of capitalism at the end of the twentieth century. The historical period of modernity which extends from the middle of the eighteenth through to the middle of the twentieth century saw the development of industrial capitalist infrastructure and means of production such as railways, automobiles, air transport, electricity, and telegraphic and telephonic communication. It also saw the growth and modernization of cities, the spread of public schools, hospitals, and prisons, and the development of bureaucratic and rational procedures for the government of social life. 'Modernism' is commonly used to refer to an artistic sensibility which embraced innovation and change. In its early stages, modernism involved a positive sense of living at the dawn of a new era. To be modern implied an heroic commitment to action and to revolutionizing oneself

as well as the means of artistic production, hence the dynamic of successive avant-gardes in the particular arts. During the latter part of the nineteenth and the early years of the twentieth century, modernism meant experimentation with the style and form of the various arts.

Theorists of postmodernity assert a parallel connection between the changes associated with the restructuring and globalization of late capitalism and the emergence of a postmodern culture. Two of the leading theorists, Fredric Jameson and David Harvey, argue that changes in technology, urban space, and forms of consumption have created the conditions of a new mode of experience of self and others (Jameson 1991, Harvey 1989). In particular, they argue that the 'compression' of social space and time which results from new electronic and communications technologies is reflected in a new cultural sensibility. Moreover, just as the modernist commitment to the destruction of traditional ways of life and thought was often linked to a tragic sense of loss, so some theorists of postmodernity argue that a similar kind of homelessness is generated by the late twentieth century mutations in our experience of space and time. Jameson is among those who take a negative view of the cultural experience of postmodernity on the grounds that it does not sustain a sufficiently robust sense of self and history. He argues that the rapid evolution of postmodern urban space has overtaken 'the capacities of the individual human body to locate itself, to organize its immediate surroundings perceptually, and cognitively to map its position in a mappable external world' (Jameson 1991, p. 83). On his account, postmodernity is experienced as incapacity in the face of complex and discontinuous social spaces and as disorientation in the absence of a rich sense of the past or future. The result is a widely documented fragmentary and schizophrenic experience of subjectivity, along with a tendency towards affectless repetition or 'blank parody' of past forms of artistic and cultural production.

For others, modernist and postmodernist alike, the appropriate affective response to change is much less one of nostalgia for older ways of being at home and more a question of orientation towards the future. Nietzsche, who is in many respects the source of much postmodern philosophy, took the condition of 'homelessness' as a metaphor for the condition of those, like himself, who looked forward to 'the strengthening and enhancement of the human type': 'We children of the future, how *could* we be at home in this today?' (Nietzsche 1974, p. 338). In their efforts to make positive sense of the experience of being estranged from the social world in which they must live, postmodernists often have recourse to the concept of irony. Richard Rorty calls ironists those who are conscious of the contingency and mutability of the 'final vocabularies' employed to make sense of their lives, yet nonetheless committed to the values which

those vocabularies embody (Rorty 1989, p. 73). Umberto Eco argues that the postmodern attitude is encapsulated in the irony which acts and speaks in full awareness of the degree to which all our ways of acting and speaking involve repetition of the past: 'the postmodern reply to the modern consists of recognizing that the past, since it cannot really be destroyed, because its destruction leads to silence, must be revisited, but with irony, not innocently' (Eco 1985, p. 67).

While postmodernism is supposed to involve rejection of the presuppositions of modern art, philosophy, and social theory, commentators and users of the term hesitate over whether postmodernism is primarily a temporal or a typological concept. The prefix and the manner in which the term has been employed in the arts suggest a temporal reference, even though estimates of the beginning of the postmodern era range from the 1860s to the 1980s. But some theorists of postmodernism in the arts and culture treat it as simply the radicalization of tendencies already present in the modern. In philosophy, Rorty argues that, to the extent that postmodernism stands for anything distinctive, it refers to a pluralism shared by a range of modern thinkers such as Nietzsche and William James (Rorty 1999). In similar fashion, Jean-Francois Lyotard argues that postmodernism in painting stands for an aesthetic of the sublime which embraces the ever-renewed attempt to 'present the unrepresentable' (Lyotard 1983). For these reasons, Eco suggests that postmodernism should not be chronologically defined but rather treated as 'an ideal category—or better still, a *Kunstsollen*, a way of operating. We could say that every period has its own postmodernism, just as every period would have its own mannerism' (Eco 1985, p. 66).

2. Unity and Plurality

A feature of many cultural responses to the transient, fleeting and ephemeral character of modern life is their attachment to ideas of unity, whether in the object to be represented or the artefact to be produced. Thus, for example, modernists believed that a machine or a city could be a unified, smoothly functional object, and a novel or work of art could be a complete, self-contained meaningful entity. Where there is experimentation with diverse modes of representation, this is undertaken in the attempt to convey a single complex and multifaceted reality. And even where there is a dizzying succession of conceptions of the nature of a particular art such as painting, there is nevertheless a conviction at each stage that at last the truth has been found.

Even though 'modernism' does not have the same currency as a classificatory term in philosophy and social theory as it does in the arts, there are epistemological parallels to this modernist attitude towards

unity. Modernist thought would be that which tolerated perspectivism only on the assumption that there is a single, intelligible reality to be presented. Even when there is an acceptance of surface incoherence, as in Marx's account of the contradictions of capitalism or Freud's account of conscious behavior, this is balanced by the discovery of an underlying structure which produces such contradictory effects. In this sense, Harvey argues that modernism 'took on multiple perspectivism and relativism as its epistemology for revealing what it still took to be the true nature of a unified, though complex, underlying reality' (Harvey 1989, p. 30). By contrast, postmodernism is associated with the acceptance of plurality in both ways of being and ways of knowing. In opposition to the ideal of unified objects of knowledge, whether texts or artefacts, postmodernism refers instead to networks, open systems, or the dissemination of meaning. In opposition to the modernist ideal of unified science, postmodernism upholds the idea of irreducible difference and even incommensurability among the varieties of knowledge.

3. Poststructuralism and Postmodernism

Postmodernism coincides with the work of a number of French thinkers who came into prominence at the end of the 1960s, notably Baudrillard, Deleuze, Derrida, Foucault, and Lyotard. By different intellectual paths, these thinkers rejected the idea that theories and concepts simply represent a pre-existing reality. For some, such as Derrida or Baudrillard, this conclusion was reached via a radicalization of Saussurian theses about the arbitrariness of signs and the manner in which their sense is derived from differential relations to other signs. For others, such as Foucault, similar conclusions were reached via Kantian views of the role played by language and thought in the construction of its own objects. A consequence of Foucault's historicization of the Kantian idea that objects of knowledge are determined in part by the underlying rules governing discourse is that 'truth' acquires a history and can no longer be measured solely against an independent reality.

While it is true that the work of Foucault and others does imply a form of historical relativism with regard to truth, it is overstating the case to suggest that the resultant 'crisis in representation' amounts to 'a deeply felt loss of faith in our ability to represent the real, in the widest sense' (Bertens 1995, p. 11). It is not our ability to represent reality but rather what Deleuze calls the representational 'image of thought' that is impugned by poststructuralist epistemology (Deleuze 1994, pp. 129–67). Foremost among the influential philosophical assumptions which make up the representational image of thought is the idea that thought naturally seeks truth. For Rorty, too, what lies at the heart of postmodernism in philosophy is the abandonment of the correspondence theory of truth and

the 'theologicometaphysical belief that Reality and Truth are One—that there is One True Account of How Things Really Are' (Rorty 1999, p. 262). A less negative characterization might say that postmodernism does not so much reject the idea that thought represents reality as embrace more complex views of what representation entails. In this manner, Foucault's genealogies of punishment and sexuality drew attention to the ways in which what passes for truth in the social sciences is bound up with the exercise of power over individuals and populations. More generally, postmodernism has drawn attention to new aspects of social reality, such as the systems of thought and language which condition public discourse. It has also given rise to new kinds of analysis of cultural and social phenomena: genealogical, deconstructive, narratological, and so on. In these ways, postmodernism has given rise to new research methods and new objects of analysis in anthropology, sociology, and other social sciences (Clifford and Marcus 1986, Dickens and Fontana 1994). In addition, postmodernism has been widely taken up in jurisprudence and the philosophy of law, where it has been described as the most influential trend at the end of the twentieth century (Litowitz 1997, p. 1).

4. Universalism and Freedom

One of the most influential formulations of postmodernism in philosophy was that provided in Lyotard's *The Postmodern Condition* (Lyotard 1984), first published in 1979. Here, Lyotard uses 'modern' to refer to any form of knowledge which seeks legitimation by appeal to some 'grand narrative, such as the dialectics of the spirit, the hermeneutics of meaning, the emancipation of the rational or working subject, or the creation of wealth.' He then defines the postmodern attitude as one of 'incredulity towards metanarratives' (Lyotard 1984, pp. 23–4). Lyotard is not concerned in this essay with postmodernity understood as an historical epoch, since his inquiry is limited to 'the condition of knowledge in the most highly developed societies' (Lyotard 1984, pp. 13). Rather, his aim is to show that the supposed transcendent status of scientific knowledge is an illusion sustained by a certain narrative language game. He argues that the privileged status of scientific knowledge is destined to disappear once it is accepted that this status depends upon a narrative which privileges science over other kinds of language game. This narrative of legitimation has either followed the French Enlightenment link between the pursuit of knowledge and human emancipation, or the German Idealist construal of history as the process of the realization of reason through the sciences. Lyotard asserts not only that these 'metanarratives' have lost their appeal but that the very idea of transcendent legitimation has given way to forms of legitimation which are local and immanent in particular social

practices and forms of life. The postmodern condition therefore is the state in which we must accept this plurality of micronarratives and address the conflicts that arise other than through the imposition of a single master narrative.

In *Modernity—An Incomplete Project*, Habermas (1983) provided an influential response to Lyotard and other poststructuralist thinkers. He characterized the philosophical project of modernity as the desire for unification of the three spheres of cultural value distinguished by Kant—the theoretical, the practical, and the aesthetic—and for the application of the results of these distinct kinds of judgment to the improvement of everyday life. According to this view, the project of modernity consisted in the effort of Enlightenment thinkers ‘to develop objective science, universal morality and law, and autonomous art, according to their inner logic. At the same time, this project intended to release the cognitive potentials of each of these domains to set them free from their esoteric forms. The Enlightenment philosophers wanted to utilize this accumulation of specialized culture for the enrichment of everyday life—that is to say for the rational organization of everyday social life’ (Habermas 1983, p. 9). The major obstacle preventing the realization of this project, according to Habermas, lay in the manner in which these spheres have become separated from one another, both institutionally and in their internal logics and forms of argumentation, and in the manner in which they have become increasingly separated from a lifeworld dominated by the instrumental rationality of the theoretical sphere.

Habermas’s solution to this problem appealed to the procedural rationality implicit in language use, or what he calls ‘communicative reason,’ as a means to restoring the possibility of intersubjective consensus. Both his diagnosis and his solution to the problem of modernity are founded upon the Enlightenment belief that universal standards of judgment are necessary for the development of both reason and moral progress. By contrast, postmodern thinkers reject the idea of progress in human affairs in favor of the idea that there are many ways in which the present differs from the past. In common with Lyotard, they reject the idea of universal standards of judgment, value or ‘legitimation.’ Although Habermas characterizes French thinkers such as Foucault, Derrida, and Lyotard as ‘young conservatives’ (Habermas 1983, p. 14), most of them do not wish to abandon the ideals of freedom and emancipation but only the approach which links these to universality, unity, and consensus and does so in the name of the Enlightenment.

In response to Habermas’s criticisms, Foucault presented his practice of genealogy as serving the cause of freedom, while suggesting that it represented a critical ethos or attitude towards the present which had always belonged to the Enlightenment. Genealogical criticism does not seek to identify the universal

structures of human knowledge and moral action, but rather takes the form of an historical investigation into the means by which we have been constructed as certain kinds of subject. Against the presumption that human reason is an unchanging and undivided faculty, Foucault views reason as divided among the distinct rationalities embodied in particular systems of thought and practice such as ‘medicine,’ ‘discipline,’ or ‘sexuality.’ The aim of this type of historical investigation, Foucault suggests, is to determine ‘in what is given to us as universal, necessary, obligatory, what place is occupied by whatever is singular, contingent, and the product of arbitrary constraints’ (Foucault 1984, p. 45). Thus, in contrast to the idea that progress in human freedom depends upon the extension of the sphere in which universal rationality governs social life, Foucault presents progress in human freedom in negative terms as the escape from particular forms of domination. For many poststructuralist theorists, freedom is not linked to the achievement of universal standards of value or to the hope that dissensus will eventually give way to consensus. Rather, it is dissensus and the experience of irreducible difference which are the conditions of freedom and creativity or change in human affairs.

5. *Difference and Multiplicity*

In his later work, Lyotard introduced the concept of the differend in order to characterize the postmodern condition with respect to political and moral judgment as the situation in which conflict between two parties cannot be resolved for lack of a rule of judgment common to both (Lyotard 1988, p. 9). Foucault also described his critical attitude towards the present as one which endorsed differentiation from the past rather than movement towards some future identity (Foucault 1984). However, the recent origins of this concern with difference must be sought in the philosophical work of Deleuze and Derrida published during the 1960s. Both sought to elaborate a non-dialectical concept of difference, which would see it neither as the simple contrary of identity, nor as ‘dialectically’ identical with identity (Descombes 1980, p. 136). Hegel was a focus of their criticism because he represented the culmination of a metaphysical tradition which treated identity as primary and difference as the derivative or secondary term. In *Difference and Repetition*, Deleuze (1994) explicitly rejects the link which Hegel forged between difference and contradiction, arguing that contradiction is not the condition or ground of difference but on the contrary: ‘It is not difference which presupposes opposition but opposition which presupposes difference, and far from resolving difference by tracing it back to a foundation, opposition betrays and distorts it’ (Deleuze 1994, p. 51). Limitation or opposition is a distortion of difference, according to Deleuze, because difference in

itself implies 'a swarm of differences, a pluralism of free, wild or untamed differences' (Deleuze 1994, p. 50). Similarly, in *Of Grammatology*, Derrida (1976) drew upon Heidegger's concept of the ontological difference between being and beings and Saussure's concept of the differentiation which gives rise to units of signification in a sign system to draw attention to the movement of deferral and differentiation (*différance*) which underlies all production of meaning (Derrida 1976, pp. 56–65).

Both Deleuze and Derrida were concerned with the problem of how to conceive of the identity of objects such as meanings, texts or signs which are subject to continuous variation, or how to envisage a form of unity which is not identical with itself. To this end, Derrida relied upon a series of concepts such as *différance*, the trace and iterability, while Deleuze relied upon a concept of open or variable multiplicity derived from Bergson. The concept of multiplicity, he and Guattari suggest, 'was created precisely in order to escape the abstract opposition between the multiple and the one, to escape dialectics, to succeed in conceiving the multiple in the pure state, to cease treating it as a numerical fragment of a lost Unity or Totality or as the organic element of a Unity or Totality yet to come, and instead distinguish between different types of multiplicity' (Deleuze and Guattari 1987, p. 32). In his collaborative work with Guattari (Deleuze and Guattari 1987), the concept of abstract multiplicities provided the formal structure for the elaboration of a series of philosophical concepts of such things as desire, language use, and forms of social organization. Deleuze and Guattari defend their idiosyncratic practice of philosophy as the invention of new concepts in order to bring about new ways of speaking and acting.

A distinction is often drawn between 'postmodern' theorists, such as Deleuze–Guattari, Derrida, and Foucault, and theorists of postmodernity such as Jameson and Harvey. The theorists of postmodernity tend to adopt a broadly Marxist approach to society and culture, and to the cultural phenomena they see as representative of the social totality, in order to pass judgment. Their 'modernist' social theory presupposes that a total picture is possible and that cultural phenomena can be understood as effects of a certain economic 'logic.' They proceed on the assumption that critical distance is possible and that an appropriate interpretative framework will provide such distance. By contrast, the 'postmodern' theorists tend to abandon economic or other forms of historical determinism, along with the idea that a totalizing theory of society and history is possible or desirable, in favor of the idea of multiple, overlapping, and divergent histories written from the distinctive perspective of particular social groups, peoples, or interpretative frameworks (power, desire, the body, etc.). This distinction is not absolute, since there are those such as Baudrillard, Foucault, and Lyotard who might be

included in both camps although at different stages of their careers. However, it is characteristic of post-modern theorists that their work involves a self-conscious attempt to theorize in a different manner, while also searching for critical strategies which do not rely upon a normative standpoint outside or apart from the object of criticism.

6. *Simulacra and Simulation*

Philosophies which seek to make difference an object of affirmation must do more than reverse the traditional hierarchy between identity and difference. The mere inversion of hierarchy does not change the fundamental relation between the elements involved, nor does it change the nature of those elements. For this reason, Derrida's practice of deconstructive philosophy always envisaged a further stage after the initial hierarchy has been overturned which would involve 'the irruptive emergence of a new "concept," a concept that can no longer be, and never could be, included in the previous regime' (Derrida 1981, p. 42). In the course of his essays he produced or drew attention to a series of such paradoxical concepts, including the trace, writing-in-general, metaphoricity, and Plato's *pharmakon*, all of which served to disrupt conceptual oppositions and hierarchies in philosophy. Similarly in *Plato and The Simulacrum* (Deleuze 1990, pp. 253–66), Deleuze argued that the concept of simulacra served to undermine or 'overturn' the primacy of identity as this was embodied in Plato's conception of a model or form of which reality is only an imperfect copy.

However, it was the sociologist Jean Baudrillard who took up the concept of simulacra and applied it to aspects of everyday life. In his earlier work, Baudrillard had argued that industrial society at the end of twentieth century had entered a new phase in which consumption rather than production served to structure social relations. He argued that the Marxian view of commodities as useful items produced in order to realize exchange value must be replaced by a semiotically informed view of commodities as signs, produced in order to realize their 'sign value' as differential elements of a social code of signification. In a series of essays published during the early 1980s, Baudrillard built upon this analysis and other accounts of the power of mass media to argue that society had entered a new era in which simulation had replaced representation as the dominant principle of signification. The essays translated in *Simulations* (Baudrillard 1983) combined a Nietzschean inspired critique of metaphysics with a descriptive account of the present stage of social life as a regime of simulacra in which the distinction between reality and the forms of its representation has broken down.

According to Baudrillard, there are many areas of social life, such as the media, where signs or other

representations do not faithfully represent a prior reality but function in its stead. For practical purposes we live in a hyper-real social world of representations without referents, images without originals. The proliferation of media transmitted information and the speed with which it circulates creates a perpetual cultural present in which contemporary society has begun to lose its capacity to retain its own past.

7. Criticism Without Foundations

While critics object to the nihilistic, impressionistic and highly selective nature of Baudrillard's descriptions, his analyses of technological, media, and political events as simulations served to draw attention to this pathological feature of the experience of social reality at the end of the twentieth century. His essays have been criticized for their failure to provide any stable foundation for criticism. However, Baudrillard openly acknowledges their ironic and hyperbolic style, describing them as 'theory-fictions' intended to provoke and to challenge contemporary common sense rather than convince by evidence or argumentation. Such a provocative function is often invoked in defense of postmodern theory. At best, it is argued, postmodern approaches serve to problematize established concepts and to draw attention to ways in which traditional universalist and humanist assumptions function to exclude or marginalize differences.

Postmodern theory is also criticized for its limited and negative approach to normative issues. Critics assert that the anti-universalism of Foucault, Lyotard, and others is incapable of providing grounds for moral or political judgment much less comprehensive programs for social reform. Defenders argue that their aim is both more limited and more modest, namely to work on the limits of what it is possible to think or to say and thereby to contribute to the emergence of new norms which may also become bases for judgment. To those who argue that the task of problematizing existing ways of thinking or challenging existing norms is ultimately of limited importance, postmodernists reply that not to do so is to limit the possibilities for thinking and acting differently. For this reason, they argue, the ethos or *Kunstwollen* of postmodernism should be seen as a necessary element of every genuinely critical approach.

See also: Foucault, Michel (1926–84); Phenomenology: Philosophical Aspects

Bibliography

- Baudrillard J 1983 *Simulations* [trans. Foss P, Patton P, Beitchmann P]. Semiotext, New York
 Bertens H 1995 *The Idea of Postmodernism: a History*. Routledge, London
 Best S, Kellner D 1991 *Postmodern Theory: Critical Interrogations*. Macmillan, Basingstoke, UK

- Clifford J, Marcus G E (eds.) 1986 *Writing Culture: the Poetics and Politics of Ethnography*. University of California Press, Berkeley, CA
 Deleuze G 1990 *The Logic of Sense* [trans. Lester M, Stivale C, Boundas C, ed.]. Columbia University Press, New York and Athlone Press, London
 Deleuze G 1994 *Difference and Repetition* [trans. Patton P]. Athlone Press, London and Columbia University Press, New York
 Deleuze G, Guattari F 1987 *A Thousand Plateaus: Capitalism and Schizophrenia*. [trans. Massumi B]. University of Minnesota Press, Minneapolis, MN and Athlone Press, London
 Derrida J 1976 *Of Grammatology* [trans. Spivak G C]. Johns Hopkins University Press, Baltimore, MD
 Derrida J 1981 *Positions* [trans. Bass A]. University of Chicago Press, Chicago
 Derrida J 1982 *Margins of Philosophy* [trans. Bass A]. University of Chicago Press, Chicago
 Descombes V 1980 *Modern French Philosophy* [trans. Scott-Fox L, Harding J M]. Cambridge University Press, Cambridge, UK
 Dickens D R, Fontana A 1994 *Postmodernism and Social Inquiry*. Guildford Press, New York
 Eco U 1985 *Reflections on the Name of the Rose*. Secker and Warburg, London
 Foucault M 1984 What is enlightenment? In: Rabinow P (ed.) *The Foucault Reader*, 1st edn. Pantheon, New York, pp. 32–50
 Habermas J 1983 Modernity—an incomplete project? In: Foster H (ed.) *The Anti-aesthetic: Essays in Postmodern Culture*. Bay Press, Port Townsend, WA, pp. 3–15 [Reprinted 1981 version of Modernity versus postmodernity. *New German Critique* 22: 3–14]
 Habermas J 1987 *The Philosophical Discourse of Modernity* [trans. Lawrence F G]. MIT Press, Cambridge, MA
 Harvey D 1989 *The Condition of Postmodernity: an Enquiry into the Origins of Cultural Change*. Blackwell, Oxford, UK
 Jameson F 1991 *Postmodernism, or, the Cultural Logic of Late Capitalism*. Duke University Press, Durham, NC
 Litowitz D E 1997 *Postmodern Philosophy and Law*. University Press of Kansas, Lawrence, KS
 Lyotard J-F 1983 Answering the question: what is postmodernism? In: Hassan I, Hassan S (eds.) *Innovation/Renovation: New Perspectives on the Humanities*. University of Wisconsin Press, Madison, WI, pp. 329–41
 Lyotard J-F 1984 *The Postmodern Condition: a Report on Knowledge*. University of Minnesota Press, Minneapolis, MN
 Lyotard J-F 1988 *The Differend: Phrases in Dispute*. University of Minnesota Press, Minneapolis, MN
 Nietzsche F 1974 *The Gay Science* [trans. Kaufmann W]. Vintage Books, New York
 Rorty R 1989 *Contingency, Irony, and Solidarity*. Cambridge University Press, Cambridge, UK
 Rorty R 1999 Afterword: pragmatism, pluralism and postmodernism. In: *Philosophy and Social Hope*. Penguin, London, pp. 262–77

P. Patton

Postsocialist Societies

'Postsocialist societies' is the general term for the former socialist countries which are in the course of full or partial restructuring of social life. The problem

of their definition (in many cases as 'postcommunist societies'), which ignores substantial differences in their past, is examined, along with certain common and significant problems that have acquired a specific form in the 'process of transition.'

1. Definition

The term 'postsocialist societies' itself raises a serious problem: societies which were considered (call themselves or identified as) socialist, and which have renounced the 'socialist course of development' since the 1989/90 changes in government, have practically no content of their own. These societies are classified in one and the same group, defined in one and the same way from the perspective of their past—their existence as 'socialist societies'—which they have renounced. On the other hand, this definition from the perspective of their past suggests that they are presumed to be societies in the course of self-identification—seen as transition from totalitarianism to democracy, from centrally planned to market economy—which have embarked on the arduous course towards building civil society.

This raises a serious question of the possible criteria by which a concrete national society would shed its 'postsocialist' label, that is, would have ultimately defined itself and be defined from the perspective of the present rather than of the past. At what stage does the present existence of the respective society provide sufficient grounds for regarding it as a society with sustained development based on values asserted in it (and shared by its citizens)?

Changes in the form of government are a necessary but not a sufficient prerequisite for radical restructuring of social life, for significant and sustainable changes in the concepts of the world, of human relations, and individual actions. That is precisely why the term is 'postsocialist societies' rather than 'postsocialist states'—the form of state government may or may not be socialist, whereas the declared renunciation of the 'socialist course of development' in itself does not free a society of the values, ideas, and expectations accumulated over the years, nor of the established relations and groups, and the very attitudes to and mechanisms of their establishment. At the same time, the course of self-identification of the 'postsocialist societies' is indefinite—their point of departure is known, but one cannot know for sure what they will ultimately become.

Of course, this is a generalization only, since the majority of 'postsocialist societies,' those in Europe, are expected to embrace precisely democratic values, a market economy, and civil society as the pillars of self-identification—most of them have not merely declared this but are making real progress in this respect. They are apparently aware that they will no longer be

labeled as 'postsocialist societies' once they complete the process (already under way with the majority) of accession to the European Union (EU). In other words, for those national societies the criteria of recognition of a completed process of self-identification are the same as the criteria of fully-fledged EU membership, which in itself means that they have already chosen a course of development, that their development is predetermined—all they have to do is attain the ultimate objective. This does not apply to the non-European 'postsocialist societies' such as China, or to many of the ex-Soviet republics (some of which are European in purely geographical terms); nor does it apply to the special case of 'rump' Yugoslavia.

2. Differences from the Past

The term 'postsocialist' creates a false impression of uniformity of the societies in question (even though this definition is indefinite itself) and, respectively, of uniformity in the process of departure from their previous development as socialist. This idea is largely based on the previous term 'socialist societies' which, along with everything else, has the ideological function of hiding or belittling the internal distinctions of the respective 'socialist societies,' as well as of rendering as entirely insignificant, from the perspective of their existence as 'socialist,' the differences in their previous development and the concrete process of their 'transformation' to socialism. They are 'socialist societies,' and all details of the history of their transformation to socialism and of the national specificity of their existence are insignificant. Yet it is precisely those formerly 'insignificant' differences (declared and considered insignificant, but to a large extent also virtually insignificant from the perspective of 'the general line' of development, of the rallying of the respective societies into the 'socialist camp'), which also largely determine the form of renunciation of the 'socialist course of development,' that have become significant accelerators or impediments—generally speaking, significant determiners of the national specificity of the existence of the concrete societies as 'postsocialist,' as well as of the pace and consistency of the transition period, of the ways of reaching national consensus on the parameters of the change, and of the actual concept and definition of 'national consensus.'

Forgetting the history of the formation of a particular nation-state and the way in which it 'embarked on the road to socialist development,' and/or failing to take into consideration the influence of national specificity on the existence of the respective society as socialist, usually results in overemphasizing the impact of international institutions (economic and financial, as well as political) whose actions are guided by economic and geopolitical interests, as well as of the

role of the personality of the respective political leaders (in the concrete 'postsocialist society'). Thus the changes and their pace in 'postsocialist societies' (especially in Europe) prove directly and crucially dependent on 'foreign' interests or the individual abilities of politicians. Consequently, individual/group responsible action proves impossible or, rather, pointless. Even if it were possible, its ineffectiveness is readily justified by counteracting 'external forces' or inactive 'internal factors.'

Social scientists, among others, are accountable for the formation of explanatory models centered on the two above-mentioned variables. In an effort to explain the transition period and its problems, they deviate from their scientific position under pressure from their partisan allegiance; concentrate on the present, 'forgetting' history; resort to biased selection of historical facts to legitimate their interpretation of events; build a macroexplanatory model of transition in general, ignoring national differences; or cite separate facts or data from opinion polls without the historical-theoretical framework requisite for their interpretation.

The production and public presence of such 'explanatory models' of the problems of the transition period, actually generate further tensions in the post-socialist society. First, they are explanatory and deliberately created for the purpose of introducing ultimate clarity among ordinary citizens about current developments; second, they are even more convincing on the strength of their authors' scientific prestige; third, they are qualitatively heterogeneous—rather than just different—interpretations, and at the same time each one of them claims to be the only truthful interpretation. This lopsided way of 'understanding' developments, history, the origin of problems in social life, of events and political actions, proves to be yet another factor (along with political allegiance and social position) that antagonizes individuals and groups.

3. *Present Similarities or Key Problems of Transition*

Two ranges of problems may be identified: problems associated with identity and problems associated with social interaction. Although they are not confined to postsocialist societies only, these problems have acquired special significance and a specific form. They may also be generalized as 'political,' 'economic,' or 'social' from a macrosociological perspective on social reality, on the processes occurring within the latter, and their specific intertwining. The macrosociological perspective proposed here is based on the understanding that the restructuring of a society which has renounced the socialist principles and norms (or, generally speaking, any society in the course of self-identification as something that is qualitatively dif-

ferent) proceeds from the presumption of the major significance of the changes in the ways of thinking and action and is seriously impeded by the continuing effect of previous models of thinking of the world and the others, patterns of behavior, and mutual relations of individuals and groups.

3.1 *Problems of Identity*

These problems concern apparently quite heterogeneous phenomena which, however, are experienced by the individual or group in a similar way as problems of self-identification: from 'nationalism,' which has become indicative of postsocialist societies, to 'unemployment' and 'poverty,' which are rarely treated as problems of identity (defined foremost as social and economic problems).

The preoccupation with national/ethnic identity, which has apparently divided (albeit in quite different ways) the postsocialist societies, stems from two specific features of the ideology upheld in socialist society: national/ethnic specificities within the boundaries of a separate state and between the states from the 'socialist camp' are insignificant—the significant self-identification is 'socialist citizens' (which does not mean that ethnocultural differences have been forgotten within the separate states, or that they never were an undisclosed reason for actual social inequality); the significant difference is directly associated with significant self-identification—the localization of everything 'strange,' 'there,' beyond the 'iron curtain.'

The disintegration of 'the camp' and the fall of 'the wall' dividing the two worlds, inevitably rejecting the validity of significant self-identification and difference, generates the need of 'national' self-identification. Whereas the question 'Who are we?' focuses the attention inwards, on the 'national community' and back in history, on the national community's existence and formation before its 'depersonalization,' the attempt at defining 'us' (who are united within a state) has exposed differences and has led to intolerance and discrimination against different ethnic and/or religious minorities. This process, common to all post-socialist societies, has acquired qualitative specificity (largely determined according to the constitution of the respective 'socialist societies'): from violent disintegration of multinational states (the Soviet Union and Yugoslavia), to the peaceful division of Czechoslovakia, to the more or less conflict-ridden confrontations between 'national majority' and 'ethnic/national minorities' in Bulgaria, Romania, Hungary, and elsewhere. Yet the common problem of emphasizing internal differences (by origin, language, religion) has remained beyond the qualitative specificity: internal differences which are thought to be significantly distinctive of people to the point of their

existence within more or less closed groups, and perception of respective individuals as significantly predetermined by their 'belonging' to those groups. Regardless of the self-perception of the respective individuals, 'the others' ascribe identity to them and treat them (engage in or avoid relations with them) from the perspective of this identity.

The problematic process of identification with European (Western) values unfolds against the background of these 'internal' problems of self-identification and ascription of identity. This process is problematic mainly because of the complex and contradictory combinations of three commonplace ideas: 'We are European,' 'We have had a tragic destiny (not without the involvement of the West), now the West should help us become like them,' 'Once we are admitted to the European Union, we shall become Europeans.' Although the images and expectations underlying the apparently unambiguous images of 'Europe,' 'the West,' 'we are European,' 'we shall become Europeans' are quite different, their focus is arguably on 'become like them,' whereas the principal dimension of the process of 'becoming' is the expected well-being.

This is not accidental. And it is directly associated with another aspect of the identity crisis in post-socialist society—unemployment and poverty. These two related but different problems are doubtless indicative of the scope of the economic crisis, but their personal aspect concerns the deepest layers of individual identity, largely because the status of being unemployed is unfamiliar, virtually absent in the experience of the 'socialist individual,' whereas poverty is a process which individuals who have found themselves in a postsocialist society do not associate with democracy and a market economy.

3.2 Problems of Social Interaction

The change (or, rather, the orientation towards change) in the principles of organization of social life triggers tensions at all levels of social interaction. At the level of the individual, these tensions take the form of a large-scale crisis in human relations. At the level of institutionalized interactions, they are manifested in 'corruption' of significant proportions.

The 'crisis in human relations' could be reworded too—as a moral and a political crisis, as well as a problem of interethnic relations. For the everyday person, however (insofar as we can talk about an everyday person in a world where 'what is taken for granted'—if anything at all—is always questioned), this problem appears as tensions in relations, as a failure to understand the 'other' (or vice versa), avoidance of certain others, isolation of and from strangers, and so forth. That problem covers the whole

range from the closest social environment—family, relatives, friends—to neighbors and colleagues, to functional relationships with anonymous others.

The problem generally defined as a 'crisis in human relations' actually results from the intertwining and superimposing of diverse manifestations and echoes of political, economic and cultural values and other changes on the macrosocial level. The realized opportunity of being different in our political preferences and allegiance, in our economic status, of choosing between different cultural models, of adhering to different values, has inevitably destroyed established models of family, friendly, neighborly, business, and interethnic relations. Yet 'new' models may appear only as a result of a long and complex process of construction, mutual adjustment, and assertion—they are not something that can be borrowed (imported, grafted), or something that becomes valid simply because it is (or appears to be) 'better.'

Similar to nationalism, corruption has become emblematic of postsocialist societies. And just like nationalism, corruption is not an unfamiliar phenomenon in Western societies. Moreover, corruption is an affliction of democratic society insofar as the latter has sufficiently large opportunities for responsible decision making, instead of the strictly prescribed rules for any action which are typical of any totalitarian society and relieve its members of responsibility. Yet the opportunity for responsible decision making is also an opportunity for irresponsible action. In the post-socialist society, the complex and contradictory intertwining of still effective old schemes of control and decision-making, and the deficiency of new norms (both on the institutional level of laws, regulations, ordinances, etc., as well as on the level of individual and group consciousness—role models, values, notions, etc.) turn corruption into an epidemic. The affliction acquires huge proportions due to, *inter alia*, the internally contradictory perception of corruption: on the one hand, it is seen as an inevitable relation in a situation of transition, of reformulating and re-regulating of institutions, social positions, rules, and so forth, which are invisibly present in the social sphere and are manifested in status, security, possessions, and money of those who have effected the replacement; on the other, corruption is embodied in the bribe-taker, insofar as bribe-takers are those who have rare resources, by rule of power, which many other persons would like to use. However, the bribe-giver is simply one of many: bribe-givers do not control the 'rules' of replacement, the way of negotiations, the size of the bribe to the bribe-taker, they are simply 'this person' who has happened to be in this position and, in this sense, anyone may end up in this position. Thus the bribe-taker is regarded as an irresponsible power-holder and the bribe-giver as 'any one of us' forced by the circumstances, and therefore not responsible for their involvement in relations of corruption.

See also: Democratic Transitions; Eastern European Studies: Culture; Eastern European Studies: Economics; Eastern European Studies: Politics; Ethnic Conflicts; Identity: Social; Minorities; Nationalism, Sociology of; Poverty, Sociology of; Prejudice in Society; Social Change: Types; Socialist Societies: Anthropological Aspects; Soviet Studies: Culture; Soviet Studies: Economics; Soviet Studies: Politics

Bibliography

- Andrusz G, Harloe M, Szelenyi I (eds.) 1996 *Cities After Socialism: Urban and Regional Change and Conflict in Post-socialist Societies*. Blackwell, Cambridge, UK
- Bukowski C G, Racz B (eds.) 1999 *The Return of the Left in Post-communist States: Current Trends and Future Prospects*. Edward Elgar, Cheltenham, UK
- Elster J, Offe C, Preuss, Boenker F, Goetting U, Ruel F W 1998 *Institutional Design in Post-communist Societies: Rebuilding the Ship at Sea*. Cambridge University Press, Cambridge, UK
- Fowkes B 1999 *The Post-communist Era: Change and Continuity in Eastern Europe*. St Martin's Press, New York
- Jawłowska A, Kempny M (eds.) 1994 *Cultural Dilemmas of Post-communist Societies*. IFiS, Warsaw, Poland
- Kornai J 1995 *Highway and Byways: Studies on Reform and Post-communist Transition*. MIT Press, Cambridge, MA
- Kornai J 1997 *Struggle and Hope: Essays on Stabilization and Reform in a Post-socialist Economy*. Edward Elgar, Cheltenham, UK
- Michta A A, Prizel I (eds.) 1992 *Post-communist Eastern Europe: Crisis and Reform*. St Martin's Press, New York
- Nagle J D, Mahr A 1999 *Democracy and Democratization: Post-communist Europe in Comparative Perspective*. Sage, London
- Opalski M (ed.) 1998 *Managing Diversity in Plural Societies: Minorities, Migration and Nation-building in Post-communist Europe*. Nepean Forum Eastern Europe, Nepean, ON, Canada
- Rose R, Mishler W, Haerpfer C 1998 *Democracy and its Alternatives: Understanding Post-communist Societies*. Johns Hopkins University Press, Baltimore, MD
- Rupesinghe K, King P, Vorkunova O (eds.) 1992 *Ethnicity and Conflict in a Post-communist World: The Soviet Union, Eastern Europe, and China*. St Martin's Press, New York
- Tismaneanu V 1998 *Fantasies of Salvation: Democracy, Nationalism, and the Myth in Post-communist Europe*. Princeton University Press, Princeton, NJ

M. Grekova

Post-traumatic Stress Disorder

1. Introduction

Psychological responses to the experience of traumatic events have long been recognized, although the meaning attached to and the terms used to describe such

reactions have varied over the years, and continue to do so, influenced by the prevailing beliefs and interests of the time and depending on the culture studied. 'Post-traumatic stress disorder' (PTSD) is the term now given, at least in Western psychiatry, to a constellation of symptoms which may follow exposure to a traumatic event, and which cause clinically significant distress or impairment in social, occupational, or other important areas of functioning. The word 'trauma' is derived from ancient Greek surgical terminology, meaning an injury stemming from penetration of the body defense (intactness) such as skin. Exposure to major trauma does not cause psychiatric disorder in all those who experience it, possibly due, at least in part, to differences in vulnerability and resilience. The concept of PTSD has been extended from adults to include children, although it is still uncertain to what extent the developmental status of the child influences the symptoms (Scheeringa et al. 1995). The child's perception of its family's reaction to a disaster may also be relevant.

Unfortunately, the recent burgeoning of interest in PTSD has perhaps given some the idea that PTSD is the only response after trauma. This is not the case (see Udwin 1993 for a discussion regarding children). It should be emphasized at the outset that PTSD is only one of a number of possible psychological sequelae of traumatic experiences, and the concept is not without its critics.

2. History

Early descriptions of the reactions of individuals and societies during the pre-Christian period, including the Sumerian, Hebrew, and Ancient Greek eras, are described by Boehnlein and Kinzie (1992). More recent examples which have been described include the response of Samuel Pepys, as described in his diary after the Great Fire of London in 1666. Pepys wrote of his 'dreams of the fire and falling down of houses,' and even six months later described how he was unable to sleep 'without great terrors of fire' (Daly 1983). Charles Dickens, the English novelist, was involved in a railway accident in June 1865, and also helped in rescuing others. He subsequently became anxious about travelling, and wrote, 'I am not quite right within, but believe it to be an effect of the railway shaking' (Trimble 1984). Two years after the accident, he was still experiencing 'sudden vague rushes of terror' (Ackroyd 1999).

The development of the railways stimulated interest in the psychological responses to traumatic events. Doctors at that time noted that there were many victims of railway accidents who appeared to suffer no physical injury and yet who complained of a variety of symptoms. Indeed, the term 'compensation neurosis' was introduced in 1879 following the increase in

invalidism reported after railway accidents. Some argued that post-traumatic symptoms were the result of physical damage to the nervous system following relatively mild trauma. John Eric Erichsen, a professor of surgery at University College Hospital, London, used the term 'railway spine' to describe such injuries. A contrary view was taken by others. Herbert Page, a London surgeon, proposed in 1885 that accident victims were suffering from 'nervous shock,' attributable to a functional (psychological) disturbance rather than physical injury. Jean-Martin Charcot, a Parisian neurologist, supported the psychological model, arguing that the syndrome of 'railway spine' was a form of hysteria. He believed that, in the immediate aftermath of an accident, the victims were in a suggestible state and that their own idea of physical injuries influenced the development of symptoms. Charcot, in turn, influenced Sigmund Freud, who later suggested that the physical symptoms of hysteria were due to the repressed memories of traumatic events. The development (in many countries) of workmen's compensation acts, which provided for financial compensation to those injured during the course of work, also led to increased interest in the psychological effects of traumatic experiences (Trimble 1984).

A major influence on the development of the concept of PTSD has been war. The belief that the psychological responses to traumatic experiences were the result of physical injury or dysfunction of the central nervous system was strongly held and represented in the concept of 'shell-shock' which prevailed during the First World War in Europe. During the First World War, the British Army reported that seven to ten percent of the officers and three to four percent of the other ranks suffered 'mental breakdowns' (Gersons and Carlier 1992). Some soldiers were treated as cowards and deserters, and punished. The number of death sentences passed was 3,500, of which approximately 282 were executed prior to 1917. However, many more were considered ill and in need of treatment (Deahl 1997). Some, such as the Scottish doctor, William Rivers, saw the problem as a psychological one, and tried psychological techniques, such as abreaction and catharsis, as well as simply letting the person talk about their traumas. Rivers held that repression explained how an individual suppressed painful and terrifying experiences, and that these later became converted into physical and mental symptoms. World War II also produced large numbers of psychiatric casualties, despite more stringent recruit selection. Mental disorder accounted for 31 percent of medical discharges from the British Army.

After wars, increased rates of mental illness have been reported in service veterans who have seen combat. For example, between 1964 and 1975, 3.14 million US servicemen and women served in the war in Vietnam. Approximately 58,000 died and 300,000 were wounded. Of combatants, 20 percent witnessed

atrocities or abusive violence. It has since been estimated that, of all US Vietnam veterans, approximately 15 percent (450,000) currently suffer from PTSD, whilst an additional 11.1 percent of male and 7.8 percent of female veterans suffer at least some symptoms of traumatic stress related to the experience in Vietnam (see Deahl 1997 for a review).

Psychological sequelae in both adults and children, including PTSD, have now been reported after a wide variety of traumatic events, such as a fire in a nightclub in Boston (Lindemann 1944), and, more recently, after a variety of both human-made and natural disasters, kidnapping, torture, and road traffic accidents, as well as in rape victims, after abuse, in refugees, after domestic violence, and even after childbirth. Epidemiological studies suggest that PTSD or subthreshold PTSD may be present in a significant proportion of a population, although it may not be recognized as such. Comorbidity may be present.

3. Classification

Attempts have been made to categorize responses to traumatic experiences. Publication of the first edition of the *Diagnostic and Statistical Manual of Mental Disorders (DSM-I)* (American Psychiatric Association 1952) coincided with the Korean war, and included the categories of 'gross stress reactions' and 'anxiety neurosis/transient situational disturbance' with reference to 'response to overwhelming environmental stress.' The second edition (*DSM-II*) (American Psychiatric Association 1968), which developed during a period of relative global peace, eliminated the diagnostic category of gross stress reaction.

In 1980, diagnostic criteria for PTSD were described in the third edition of the *Diagnostic and Statistical Manual of Mental Disorders (DSM-III)* (American Psychiatric Association 1980). These criteria were based on information obtained from Vietnam veterans, and probably also reflected the experience of Holocaust concentration-camp survivors from World War II. The diagnostic criteria were modified in the revised edition (*DSM-III-R*) published in 1987. There was increasing emphasis on avoidance phenomena. It was acknowledged that children may be affected, that 'survivor guilt' was not a primary feature, and that symbolic reminders of the trauma may be important.

Meanwhile, in Europe, the ninth revision of the *International Classification of Diseases (ICD-9)* (World Health Organization [WHO] 1978) incorporated the subcategories of 'acute reaction to stress' and 'adjustment reaction' following exposure to acute stress. Acute reaction to stress described transient responses to exceptional physical or mental stress, lasting only a few hours to a few days. Adjustment reactions were longer lasting, but usually lasted only a few months.

The diagnostic criteria are presently defined in the *Diagnostic and Statistical Manual of Mental Disorders*

(*DSM-IV*), published by the American Psychiatric Association in 1994, and *The ICD-10 Classification of Mental and Behavioural Disorders*, published by the World Health Organization in 1992. There is also the publication *The ICD-10 Diagnostic Criteria for Research* (World Health Organization 1993). According to *ICD-10*, PTSD is a delayed and/or protracted response to a stressful event or situation (either short- or long-lasting) of an exceptionally threatening or catastrophic nature (including serious accident). Typical symptoms include episodes of repeated reliving of the trauma in intrusive memories or dreams, a sense of numbness and emotional blunting, with feelings of detachment from other people, inability to enjoy activities, and avoidance of activities and situations reminiscent of the trauma. Other symptoms, such as hyperarousal of the autonomic nervous system, hypervigilance, an enhanced startle reaction, difficulty sleeping, anxiety, and depression may also be present. The *ICD-10 Diagnostic Criteria for Research* are similar, but specify that there must be persistent remembering or 'reliving' of the stressor in intrusive 'flashbacks,' vivid memories, or recurring dreams, or in experiencing distress when exposed to circumstances resembling or associated with the stressor. In addition, for the diagnosis of PTSD to be made using these criteria, there should be an actual or preferred avoidance of circumstances which resemble or are associated with the stressor, which was not present before exposure to the stressor, and either an inability to recall, either partially or completely, some important aspects of the period of exposure to the stressor, or persistent symptoms of increased psychological sensitivity and arousal (not present before exposure to the stressor), as shown by any two of the following—difficulty sleeping, difficulty concentrating, irritability or outbursts of anger, hypervigilance, or an exaggerated startle reaction. These symptoms should have started within six months of the stressor.

According to *DSM-IV*, PTSD occurs after a traumatic event in which the person experienced, witnessed, or was confronted with an event or events that involved actual or threatened death or serious injury, or a threat to the physical integrity of self or others. The person's response should have involved intense fear, helplessness, or horror. For the diagnosis of PTSD, the individual should be persistently re-experiencing the traumatic event, such as experiencing recurrent and intrusive distressing recollections of the event (which, in children, may be manifest in repetitive play in which themes or aspects of the trauma are expressed), having recurrent distressing dreams of the event (in children, the content of the dreams may be less recognizable), acting or feeling as if the traumatic event were recurring, experiencing intense psychological distress at exposure to internal or external cues that symbolize or resemble an aspect of the traumatic event, or experiencing physiological reactivity on exposure to internal or external cues that symbolize or

resemble an aspect of the traumatic event. There should be evidence of persistent avoidance of stimuli associated with the trauma and numbing of general responsiveness (not present before the trauma), as indicated by three (or more) of the following: efforts to avoid thoughts, feelings, or conversations associated with the trauma, efforts to avoid activities, places, or people that arouse recollections of the trauma, an inability to recall an important aspect of the trauma, markedly diminished interest or participation in significant activities, a feeling of detachment or estrangement from others, a restricted range of emotions, a sense of a foreshortened future. There should also be at least two persistent symptoms of increased arousal (not present before the trauma), such as difficulty falling or staying asleep, irritability or outbursts of anger, difficulty concentrating, hypervigilance, an exaggerated startle response. These symptoms should have lasted more than one month and be causing clinically significant distress or impairment in social, occupational, or other important areas of functioning. 'Delayed-onset' PTSD may be seen where a person initially appears to have dealt with a traumatic experience but, perhaps years later, symptoms of PTSD develop, possibly after a reminder of the original trauma or after a relatively trivial incident.

While diagnostic classification systems are useful, they are not a checklist, and clinical judgment is considered essential in their use. Criteria are not 'written in stone' but evolve in the light of clinical experience and information from research. For example, there is ongoing discussion about the validity of the 'stressor criterion,' since PTSD may occur after seemingly minor accidents. It seems likely that, in the future, there will be attempts to eliminate as many differences as possible between the ICD and DSM systems of classification (see *Nosology in Psychiatry*).

4. Theoretical Models

The concept of PTSD has been accompanied by a variety of theoretical models (neuropsychiatric, psychoanalytical, cognitive-behavioral) to explain its origins and maintenance in the traumatized individual. For example, a number of psychobiological findings in PTSD, such as altered autonomic function and alterations in the hypothalamic-pituitary-adrenal axis, and studies of endogenous opioids (such as enkephalins) may be relevant (see *Endocrinology and Psychiatry*). Psychoanalytic models include those which regard the re-experiencing phenomena as primary, leading to defensive functioning, and those which regard the reaction to the stressor as a failure of adaptation. Common to all dynamic models is a concern with the meaning a person gives to a traumatic event and/or its aftermath. Behavioral models of PTSD have used the principles of conditioning theory to explain the development and maintenance of PTSD,

while cognitive models of PTSD have included the concepts of learned helplessness, the attributional style of individuals, cognitive appraisal, and, more recently cognitive-processing theory. However, no single model offers a comprehensive view of PTSD. It has also been suggested that an understanding of memory may assist in the understanding of the phenomenology of responses seen after traumatic events. It is to be hoped that a clearer understanding of the origins of PTSD will help in the development of effective therapeutic techniques.

5. Assessment

There are a number of questionnaires or scales which may be used to help assess the psychological responses of traumatic events. Davidson et al. (1997) describe a 17-item self-rating scale for PTSD, which measures each *DSM-IV* symptom on five-point frequency and severity scales. Costello et al. (1998) have described a new measure of life events and PTSD for children and adolescents aged nine to seventeen years, for use in both epidemiological and clinical studies. One of the most widely used questionnaires in either research into, or clinical work on, the psychological sequelae of traumatic events, is the Horowitz Impact of Event Scale (Horowitz et al. 1979). This is a 15-item self-rating scale with two subscales: intrusion and avoidance. It is normally used to assess the response after a single traumatic event, rather than after exposure to a chronic stressful situation. It gives a subjective estimate of the frequency of intrusive recall of the traumatic event and of attempts to avoid such recall, and thus a measure of psychological disturbance when compared to normative data. It does not include questions about physiological reactivity and cannot, therefore, be used to make any diagnosis. It was not originally designed for use with children but has been shown to be useful in those aged eight years and over. A score of 20 or above is often considered to be a cause for concern for civilians in peacetime, although others have used a score of 30 or above to indicate significant post-traumatic stress.

6. Treatment

A variety of approaches are presently used, both in adults and in children. In adults, behavioral and cognitive approaches have been increasingly used over the past 15 years or so, usually involving exposure (either real-life or imagined), cognitive therapy, or anxiety management training. Clinically, exposure and cognitive restructuring are often included in an anxiety management program. Psychotherapy may be individual, group, or with a family. A further technique, 'eye movement desensitization and reprocessing' (EMDR), may be a useful therapeutic technique, but its value and the mechanism of any effects that it has

require further elucidation. There is little evidence that drug treatments have a central role in the treatment of PTSD.

The emphasis when treating children has been on cognitive behavioral approaches that aim to help survivors of a traumatic event make sense of what has happened and to master their feelings of anxiety and helplessness. Sometimes support may be appropriately offered in groups. Since children live with adults, whose reactions to events are part of the context of that event for the child, most treatment plans for children should involve both families and schools. This may be especially relevant when the adults have themselves experienced the traumatic event. Distress in the adults may mean that they are less able to recognize distress in the children for whom they have responsibility, or that they are relatively unavailable to respond to the children's needs (see *Child and Adolescent Psychiatry, Principles of*).

Further research is needed to evaluate the effectiveness of the differing therapies and also to determine the effectiveness of screening populations at high risk of PTSD (for example, after a major disaster) to find out whether any interventions may prevent the development of PTSD or other psychiatric morbidity in those identified as being especially at risk. However, there is no clear evidence that offering counseling routinely to all survivors of traumatic events is helpful—indeed, there is some suggestion that it may be harmful (Wessely et al. 2000).

7. Areas of Controversy

PTSD is an evolving concept. There has been continuing debate on whether individual, as opposed to stressor characteristics, are decisive in the development of PTSD. It may be inappropriate to use a single medical concept to describe the responses seen after different traumatic experiences. Distress after the experience of traumatic events has been variously interpreted in the past. The current emphasis on the 'medicalization of distress' in Western cultures may, perhaps, reflect the prevailing expectations of that society and culture. It is still unclear whether it is appropriate to use a concept which has developed in Western psychiatry to other cultures, and how appropriate it is to extend the concept to describe the responses seen after a wide range of traumatic events. The present classifications assume a dichotomous rule—disorder is either present or absent, with no opportunity to consider severity.

An important consequence of the diagnostic category of PTSD is that it has facilitated, and perhaps encouraged, claims for compensation after traumatic events. Clinicians undertaking forensic assessments are advised that diagnostic criteria are clinical guidelines, and that there may be an imperfect fit between questions of ultimate concern to the law and information contained in a clinical diagnosis.

8. Conclusions

The concept of PTSD is an evolving one, of uncertain value, and it seems likely that there will be further developments and refinements in our understanding of individual responses to traumatic events, and how best to describe and classify them. The importance of developmental influences, the significance of symptoms in differing cultures, and the risk and protective mechanisms involved require further attention.

See also: Mental Illness, Etiology of

Bibliography

- Ackroyd P 1999 *Dickens*. Vintage, London
- American Psychiatric Association 1952 *Diagnostic and Statistical Manual of Mental Disorders*, 1st edn. American Psychiatric Association, Washington, DC
- American Psychiatric Association 1968 *Diagnostic and Statistical Manual of Mental Disorders*, 2nd edn. American Psychiatric Association, Washington, DC
- American Psychiatric Association 1980 *Diagnostic and Statistical Manual of Mental Disorders*, 3rd edn. American Psychiatric Association, Washington, DC
- American Psychiatric Association 1987 *Diagnostic and Statistical Manual of Mental Disorders*, 3rd rev. edn. American Psychiatric Association, Washington, DC
- American Psychiatric Association 1994 *Diagnostic and Statistical Manual of Mental Disorders*, 4th edn. American Psychiatric Association, Washington, DC
- Boehnlein J K, Kinzie J D 1992 Commentary. *DSM* diagnosis of post-traumatic disorder and cultural sensitivity: A response. *Journal of Nervous and Mental Disease* **180**: 597–9
- Costello J, Angold A, March J, Fairbank J 1998 Life events and post-traumatic stress: The development of a new measure for children and adolescents. *Psychological Medicine* **28**: 1275–88
- Daly R J 1983 Samuel Pepys and post-traumatic stress disorder. *British Journal of Psychiatry* **143**: 64–8
- Davidson J R T, Book S W, Colket J T, Tupler L A, Roth S, David D, Hertzberg M, Mellman T, Beckham J C, Smith R D, Davison R M, Katz R, Feldman M E 1997 Assessment of a new self-rating scale for post-traumatic stress disorder. *Psychological Medicine* **27**: 153–60
- Deahl M 1997 Effect of conflict on combatants. In: Black D, Newman M, Harris-Hendriks J, Mezey G (eds.) *Psychological Trauma: A Developmental Approach*. Gaskell, London, pp. 134–47
- Gersons B P R, Carlier I E 1992 Post-traumatic stress disorder: The history of a recent concept. *British Journal of Psychiatry* **161**: 742–8
- Horowitz M, Wilner N, Alvarez W 1979 Impact of Event Scale: A measure of subjective stress. *Psychosomatic Medicine* **41**: 209–18
- Lindemann E 1944 Symptomatology and management of acute grief. *American Journal of Psychiatry* **101**: 141–8
- Scheeringa M S, Zeanah C H, Drell M J, Larrieu J A 1995 Two approaches to the diagnosis of post-traumatic stress disorder in infancy and early childhood. *Journal of the American Academy of Child and Adolescent Psychiatry* **34**: 191–200
- Trimble M R 1984 Post-traumatic stress disorder: The history of a concept. *Psychiatry in Practice* **May**: 5–7
- Udwin O 1993 Annotation—Children's reactions to traumatic events. *Journal of Child Psychology and Psychiatry* **34**: 115–27

- Wessely S, Rose S, Bisson J 2000 Brief psychological interventions ('debriefing') for treating immediate trauma-related symptoms and preventing post-traumatic stress disorder. *Cochrane Database of Systematic Reviews* **2**: CD000560 [Medline]
- World Health Organization 1978 *Mental Disorders: Glossary and Guide to the Classification in Accordance with the Ninth Revision of the International Classification of Diseases*. World Health Organization, Geneva, Switzerland
- World Health Organization 1992 *The ICD-10 Classification of Mental and Behavioural Disorders: Clinical Descriptions and Diagnostic Guidelines*. World Health Organization, Geneva, Switzerland
- World Health Organization 1993 *The ICD-10 Classification of Mental and Behavioural Disorders: Diagnostic Criteria for Research*. World Health Organization, Geneva, Switzerland

M. Newman

Potlatch in Anthropology

'Potlatch' refers to a ceremonial exchange complex of aboriginal peoples of the Northwest Coast of North America. Originating in precontact practices, the potlatch evolved in the nineteenth century into a quasicommercial system of acquiring prestige through large gifts—and occasionally destruction—of native and trade goods. Among the Kwakiutl of Vancouver Island and neighboring groups, the potlatch came to include a supplemental economy of loans and interest, allowing for prestations so large that they beggared the host. This extreme form, appearing between roughly 1870 and 1930, earned it a central place in the ethnographic imagination.

1. History and Origins

1.1 Precontact Trade and Ceremonial Exchange

'Potlatch' derives from the Chinook jargon word *patshatl*, to give. Chinook jargon was a trade language centered in the Columbia River basin, but extending north to the central Northwest Coast. It predates European contact, and served as a *lingua franca* among groups engaged in aboriginal trade. Archeological evidence suggests the existence of intergroup ceremonial exchange alongside precontact trading relations. However, it is clear that such exchange was both quantitatively and qualitatively different from the historic potlatch. It involved modest amounts of animal pelts and a few types of prestige goods, and was known under different labels in different areas. It reaffirmed but likely did not alter rank relations within and between groups. Thus, it is fair to say that the potlatch, as a translocal institution, was a product of the fur-trade era.

1.2 The Fur Trade: Influx of Wealth

The early fur trade of the 1770s and 1780s involved Russians, Spanish, British, and Americans. Trade focused on the sea otter and was conducted primarily from ships. This had a limited effect on aboriginal groups. However, by the late 1780s the sheer number of trading vessels, and the shift to land-based trading, had begun to bring changes. The transfer of the Spanish colony on Juan de Fuca Strait to the British in 1793 marked the beginning of a process that would see Hudson's Bay Company forts established all along the coast, and eventually a substantial colony at Victoria. To the north, the Russians established a fort at Sitka, in Tlingit territory, and to the south, Americans established posts on the Columbia in the 1810s. The shift to land-based trade brought with it an influx of wealth and trade goods, as well as epidemic disease, and a broad range of new cultural influences.

Despite the negative effects of contact, the early nineteenth century was characterized by an efflorescence of art and expressive culture on the Northwest Coast. Nowhere is this seen more clearly than in the plastic arts. The introduction of metal woodworking tools and, somewhat later, commercial paint, allowed for an explosion of art in the form of masks, architectural elements, storage vessels, and free-standing monumental sculpture. Such objects were commissioned with new commercial wealth; these objects, in turn, were vehicles by which this wealth was displayed and validated. Wealth had always been closely associated with rank. The problem arose in that rank was ideologically a closed system: a limited number of ranked positions, associated with titles, privileges, and symbolic property ('crests') were believed to exist, and these positions were filled in accordance with regular principles of descent and succession. However, during the fur trade period *nouveaux riches* rose to prominence. New crests were created for them by artists who were experimenting with both themes and techniques. Simultaneously, such persons also staked claim to unused traditional crests or those in dispute. The potlatch was the forum in which such claims were made public, and were either validated or rejected by the larger community.

1.3 The Dominion Period: Response to Traumatic Change

After the 1860s, which saw both Canadian Dominion and the US purchase of Alaska, aboriginal groups were increasingly encapsulated within Euroamerican political, social, and economic institutions. Simultaneously, the full effects of disease mortality were being felt, producing upheavals in aboriginal society. The potlatch, an effective means of dealing with demographic and social change, was mobilized to address issues such as the amalgamation of groups, the lack of personnel to fill important rank positions,

changing relations with Euroamericans, and changes in the relative status of local groups. In the process, potlatches became larger and more expansive. A wide range of aboriginal practices, such as shamanic winter dances, became incorporated within the potlatch. Simultaneously, the potlatch replaced intergroup violence, which was suppressed, as a means for gaining prestige at neighbors' expense.

During this period, the potlatch was suppressed by Canadian authorities, who passed a 'Potlatch Law' in 1885. Although unenforceable for most of its existence, it did result in famous prosecutions and convictions of Kwakiutl potlatchers in 1923. In addition to criminal punishment, the state seized 'potlatch goods,' including ritual objects. These provided the basis of ethnological collections at museums in Ottawa and Toronto. In the 1970s and 1980s, these items were returned to Kwakiutl communities at Cape Mudge and Alert Bay, which have built museums to house them.

2. Geographical Range and Variation

2.1 The Northern Potlatch

The northern region, consisting of matrilineal societies, used the potlatch primarily in the context of funerals and succession. For the Tlingit and Haida, potlatches were between moieties, and stressed co-operation and solidarity between hosts and guests. A minor theme of competition was introduced, with two separate guest groups attempting to outdo each other in the degree of 'comfort' provided to the host. Tsimshian potlatches were similar, but, lacking moieties, were somewhat more agonistic. They were associated with marriages, as well as other life-crisis rituals; marriage was seen in oppositional terms. In addition, potlatches were used to make territorial claims.

2.2 The Southern Potlatch

South of the Tsimshian, social organization was cognatic, with a matrilineal orientation in the north becoming more patrilineal in the south. Among the Nootka, potlatches were essentially private entertainments sponsored by a chief for his people, and thus stressed social harmony. Elsewhere, however, potlatches were primarily agonistic. Marriage potlatches, including those held to repay bridewealth, were couched in military terms. As group membership and rank were less fixed in cognatic societies, they were more the object of political manipulation, and thus of representation in the potlatch.

2.3 The Kwakiutl Potlatch

The extreme example of agonistic prestation was provided by the Kwakiutl, who live in northeastern

Vancouver Island and the adjacent islands and mainland. During the 60 years after Dominion, the potlatch intensified to the point that Benedict (1934), in her flawed analysis of Boas' material (Boas 1966), described the Kwakiutl as 'megalomaniacal' and 'paranoid.' Fierce competition among brothers, cousins, and affines for succession to titles, the existence of numerous latent titles, the rise of a *nouveaux riche* class called 'Eagles,' and a system of intergroup ranking, all contributed to the intensity of the potlatch, which participants viewed as warfare by other means. Potlatches opened with guests arriving in war regalia, and a mock battle on the beach, reminiscent of the murderous raids that were endemic to this area.

Kwakiutl potlatches were predicated on rivalries between chiefs and, by extension, their groups. In addition to the usual means of amassing wealth, would-be hosts engaged in a system of loans with 100 percent interest, which allowed lenders to build fortunes quickly, and borrowers to obtain last-minute infusions. The currency was Hudson's Bay Company blankets, supplemented and eventually replaced by cash. Copper shields were the ultimate valuable; bid up in previous transactions, 'coppers' reached cash values of several thousand dollars. In addition, trade goods played an increasing role. By the start of the twentieth century, manufactured items such as shotguns, washing machines, and pool tables were given away, along with quantities of flour, rice, cloth, and utensils. In some of the most extreme potlatches, goods including coppers were destroyed.

3. Functional Accounts

3.1 The Potlatch as Total Social Phenomenon

The French sociologist Marcel Mauss (1990) noted that the potlatch is an integral element of the rank system of Northwest Coast societies. It is a mechanism by which claims to rank status are validated or challenged, and changes in social order are communicated to the larger community. The potlatch rests on opposition between host and guest. The host—a chief or high-ranking person and, by extension, his social group—was concerned with conducting certain types of business: naming heirs, contracting marriages, conducting funerals, proclaiming new titles and privileges. This business was witnessed, in a juridical sense, by the guests, who were paid for their service. Potlatch gifts went beyond mere payment for services, however, since they were also taken as *prima facie* evidence for the validity of the claim.

Wealth was a necessary concomitant of high rank. High-rank positions implied a certain degree of extraordinary power, similar to the *mana* of Polynesian chiefs, which necessarily resulted in wealth. Inability to provide an abundant potlatch would seriously undermine the legitimacy of the claimant's assertions. By the same token, an extravagant potlatch could

'fasten on' a title, merging the holder's social identity with the permanent qualities of the title. Such acts were seen as a sort of reincarnation, by which the living title holder became incorporated in the title, which encompassed the holder's predecessors and mythical founder. For chiefly positions, the ascendance of a new holder was often marked by events symbolic of the (re)creation of the social world, including human sacrifice, the razing of old chiefly houses and poles, and their replacement by new ones.

In chiefly succession potlatches, a broad range of political, legal, and economic issues were implicated. In all groups, chiefly titles involved territorial claims of some sort. These titles were based on mythical narratives, set locally, and taking place at the beginning of human history, thus staking aboriginal claim to land. In some groups, such as the Nootka, such territorial claims were very specific and resembled Western notions of ownership. In all Northwest Coast cultures, chiefs were minimally seen as stewards of the land and its inhabitants, and were responsible for regulating relations between human and nonhuman populations (which were believed to include more than the flora and fauna recognized by biological science). This was accomplished by ritual means, including prayers, attraction magic, and first fruit rituals. The latter were used to regulate the seasons on fish and game. Although the informing ideology was moral, not ecological, such regulation had the effect of conserving resources.

The potlatch thus involved relations between the social group and the nonhuman outside world, the prehuman myth world, and the group's ancestors. In this way, the group was situated within a larger universe, and was thus defined and delimited. Potlatches likewise served to delineate the internal organization of society, and the functional relations obtaining between its components. In a general sense, one potlatches to one's affines, who constitute the friendly other. In northern matrilineal groups, this opposition was absolutely clear. Especially in the moiety system of the Tlingit and Haida, one's affines were a separate corporate group; the cooperation of the two sides presented an image of social totality cosmological in scope.

3.2 The Potlatch and Marriage Exchange

Potlatches were associated with marriage in most areas; even where they were not, the potlatch was transacted between affines. Politically, potlatches were vital in renewing alliances among groups. Among the northern groups, whose aristocrats practiced preferential cross-cousin marriage, long-term stable exchange relations involving two or more local groups were established. Such groups provided political, military, and economic support, as well as external validation of chiefly claims. Among cognatic groups, no such long-term structural relations existed, but

were rather produced and reproduced in a series of exchanges tied to specific marriages (payment and repayment of bridewealth, naming of children, funeral of a spouse). The connection between marriage and potlatching was so strong that in groups such as the Kwakiutl, Bella Coola, and Heiltsuk, aristocrats would engage in serial monogamy and even fictive marriage in order to obtain prestige from potlatching. Potlatching varied along several dimensions (size, number, occasion, agonism, and others) depending on the precise configuration of kinship, descent, and marriage rules.

3.3 *The Potlatch and Ecological Variation*

Since the 1960s, materialist scholars have viewed the potlatch as a virtual homeostatic mechanism, distributing personnel and food efficiently through space and time. The earlier conception of the Northwest Coast as free from food pressures is justly rejected. The potlatch is seen as having four distinct, but inter-related functions: (a) spurring overproduction, (b) encouraging storage of food, (c) allowing for the exchange of food for prestige goods, and (d) regulating group membership and territorial boundaries, matching resources with personnel. This approach assumes local variation in resources, a condition that does not exist everywhere, but is more pronounced in the north. Item (d) was effected much more often by warfare than by potlatching. However, these remain at least potential functions of the potlatch.

3.4 *Rank, Class, and Political Power*

Northwest Coast society was stratified by both rank and class. Potlatching was the domain of chiefs and nobles; however, both commoners and slaves played important roles in amassing resources. By definition, only persons with titles were in a position to potlatch. Potlatching thus reinforced vertical, as well as horizontal, social divisions. Simultaneously, potlatching was the mechanism by which social mobility was integrated into the social system.

The potlatch both manifested and represented political power. By drawing together a large number of guests, especially if they were from far away or their chiefs were especially high ranking, a host was demonstrating his prestige and political network. The material goods and food distributed were an index of his land holdings and their productivity, which ultimately depended upon the chief's legitimacy. Potlatches were fora for authoritative political discourse, defined technically as monologic, multichannel, redundant communication. A significant aspect of this communication was an illocutionary speech act. This was supplemented with nonverbal media, including plastic arts and dance. The political success of a potlatch was usually measured in terms of the degree to which these utterances remained unchallenged.

They may be challenged either directly, in rivalry gestures, or more subtly through various sorts of corrosive discourse, including humor. However, in some groups, such as the Kwakiutl, such elements were institutionalized within the structure of the potlatch. This had the effect of heightening tensions within the potlatch, and giving it a more game-like structure, in which a wide range of outcomes was possible.

4. *Problems of Reference*

4.1 *Reification*

Given the degree of variation over space and time, it is unclear that the potlatch, *as such*, exists. Rather, our understanding of it is conditioned by reifications imposed by government agents and missionaries, anthropologists, and aboriginal people themselves. The Potlatch Law of 1885 first tried to specify 'the Indian festival known as the "Potlatch"'. Problems of reference significantly delayed successful prosecution of the law. During this intervening time, the symbolic labor of many people served to bring descriptions and practice closer together.

4.2 *Invention of Tradition*

Aboriginal people themselves reified the potlatch by defining their ceremonial praxis as potlatch, a foreign word and concept mediated through Euroamerican society. By the time Franz Boas recorded ethnographic data among the Kwakiutl in the 1890s (Boas 1966), the concept of potlatch was present. This consciousness of 'the potlatch,' on the part of both Boas and the Kwakiutl, was shaped largely by the existence of the Potlatch Law. Clearly, aboriginal practice and self-awareness of practice were conditioned by exogenous expectations and categories. It is not surprising that during a period of rapid change such an invention of tradition would occur. It allowed aboriginals to retain symbolic control of their destiny, by redefining novel structures and practices as traditional. Because of this useful quality, potlatching remains a common practice today.

4.3 *The Potlatch as Anthropological Totem*

The potlatch remains one of the most durable anthropological concepts, despite the reanalyses and deconstructions to which it has been subjected. Beginning with Boas, anthropologists defined potlatch as both self and other, transparent and exotic. This contradictory and mediational quality has made potlatch an especially useful metaphorical vehicle. Potlatch has been viewed as, *inter alia*, metaphorical banking, warfare, and mental illness. At the same time, various aspects of Western societies, including

politics and consumer capitalism, have been compared to potlatch. Its hybrid nature, and its origins in the desire on the part of Euroamericans to describe aboriginals, and aboriginals to represent change and continuity, make it an attractive concept, but one that resists critical usage.

See also: Boas, Franz (1858–1942); Colonialism, Anthropology of; Elites, Anthropology of; Exchange in Anthropology; Exchange: Social; Mauss, Marcel (1872–1950); Political Anthropology; Totemism

Bibliography

- Adams J W 1973 *Gitksan Potlatch: Population Flux, Resource Ownership, and Reciprocity*. Holt, Rhinehart, and Winston, Toronto, Canada
- Benedict R 1934 *Patterns of Culture*. Houghton Mifflin, Boston
- Boas F 1966 *Kwakiutl Ethnography*. University of Chicago Press, Chicago
- Bracken C 1997 *The Potlatch Papers: a Colonial Case History*. University of Chicago Press, Chicago
- Codere H 1950 *Fighting with Property: a Study of Kwakiutl Potlatching and Warfare, 1792–1930*. Augustin, New York
- Cole D, Chaikin I 1990 *An Iron Hand Upon the People: the Law Against the Potlatch on the Northwest Coast*. University of Washington Press, Seattle, WA
- Drucker P, Heizer R F 1967 *To Make My Name Good: a Reexamination of the Southern Kwakiutl Potlatch*. University of California Press, Berkeley, CA
- Harkin M E 1997 *The Heiltsuks: Dialogues of Culture and History on the Northwest Coast*. University of Nebraska Press, Lincoln, NE
- Kan S 1989 *Symbolic Immortality: the Tlingit Potlatch of the Nineteenth Century*. Smithsonian Institution, Washington, DC
- Lévi-Strauss C 1982 *The Way of the Masks*. University of Washington Press, Seattle, WA
- Mauss M 1990 *The Gift: the Form and Reason for Exchange in Primitive Societies*. Hall W D (trans.) Routledge, London
- Rosman A, Rubel P G 1971 *Feasting with Mine Enemy: Rank and Exchange Among Northwest Coast Societies*. Columbia University Press, New York
- Schulte-Tenckhoff I 1986 *Potlatch: Conquête et Invention: Réflexion sur un Concept Anthropologique*. Lausanne: Éditions d'en Bas
- Seguin M 1985 *Interpretive Contexts for Traditional and Current Coast Tsimshian Feasts*. National Museums of Canada, Ottawa, Canada
- Walens S 1981 *Feasting with Cannibals: Metaphor and Morality in Nineteenth Century Kwakiutl Culture*. Princeton University Press, Princeton, NJ

M. E. Harkin

Poverty and Child Development

Income poverty is when a family's income fails to meet a federally established threshold; this threshold differs across countries. In general, living at or below a

poverty threshold signifies that a family lacks basic financial resources, and thus children may lack access to necessary resources including food, shelter, and health care. Historically, the measurement of poverty as well as research linking poverty to child development has been an Anglo-Saxon endeavor (Smeeding et al. 2000). However, international organizations, such as the United Nations Children's Fund (UNICEF), have fostered global concern. A growing body of research suggests that poverty has consequences for children's well-being. Internationally, the prevalence of poverty among families with children has profound implications for researchers, policy makers, and practitioners concerned with children's development.

This article, divided into three sections, reviews the literature on the role of income poverty in the lives of children and families. The first section considers different measures of poverty as well as international trends in child poverty. Second, a review of recent studies examining the association between poverty and children's physical, cognitive, and emotional development is provided. The last section reviews the potential pathways through which poverty may influence children's health and well-being. The conclusion briefly highlights broad policy implications.

1. Measuring Poverty and International Trends

1.1 Income and Poverty

In most countries, income poverty typically is defined by official poverty thresholds. Those families whose income falls above the poverty line are considered 'not poor,' whereas, families whose incomes are at or below the threshold are considered 'poor.' Two general types of thresholds exist, absolute and relative. *Absolute poverty thresholds* are based on an absolute level of resources; as such, a family's poverty status is not influenced by other families' incomes. For instance, the USA has an absolute poverty threshold based on anticipated food expenditures (thrifty food basket), taking into account household size, and multiplied by three; the thresholds are adjusted annually for changes in cost of living. In contrast, Canada and most Western European nations employ relative poverty thresholds. A *relative poverty threshold* is calculated by comparing a family's resources to that of the average family. These thresholds are, generally, defined by half of the median or mean income of all households in the respective country. Thus, other families' incomes can change a family's poverty status, even if their income is relatively stable over time.

1.2 International Trends in Child Poverty

Using the poverty measures described, international comparisons of child poverty rates are considered. Comparisons of the proportion of children living in

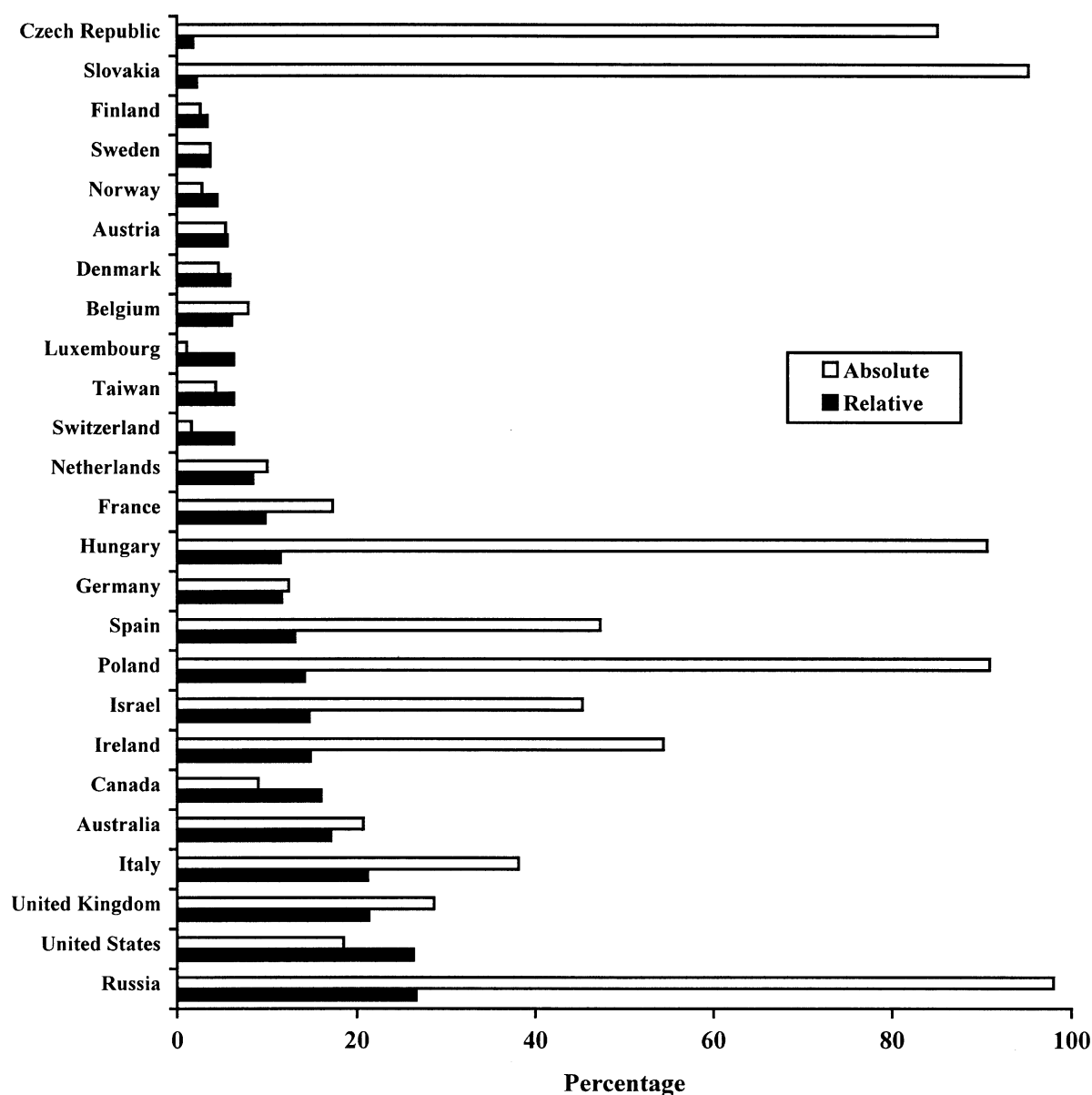


Figure 1ⁱ

Percentage of children living in poverty in industrialized nations using relative and absolute measuresⁱⁱ

ⁱAdapted from: Bradbury and Jantti 1999

ⁱⁱRelative poverty rate is 50 percent of overall national median income for survey year. Absolute rate is \$15,299, based on 1995 US official poverty line for a two-parent family with two children

poverty in industrialized countries vary depending upon the type of poverty measure utilized (see Fig. 1, Bradbury and Jantti 1999). Using a relative measure, where families living at or below 50 percent of the national median income are considered to be poor, the USA ranks 24th out of 25 in terms of low child poverty

rates, despite general economic prosperity. With the exception of the USA and UK, countries with higher levels of national income tend to have lower relative child poverty rates than countries in economic transition (Bradbury and Jantti 1999, Smeeding 1997). However, if an absolute measure of poverty is applied

using the 1995 US criteria, this pattern shifts. Although the child poverty rate in the USA declines only slightly, the country's ranking jumps to 12th place, while Russia and Hungary, countries in economic transition, have child poverty rates of 90 percent or higher. In contrast, child poverty rates remain low in the Nordic countries and Germany and are not appreciably altered by changing poverty measures (Bradbury and Jantti 1999).

The overall level of economic development of a country is not the only factor influencing child poverty trends, as over time, countries with similar levels of development fared differently. Diverse demographic differences between countries, such as the prevalence of single-parent versus two-parent families, account for some of the discrepancies, as well as different roles assumed by governments for fighting child poverty. International poverty rates also are affected by the inclusion or exclusion of social transfers including unemployment benefits, illness or disability assistance, income assistance, and housing benefits (Eurostat 1999). Within the European Union (EU) countries, the USA, and Canada, poverty rates drop from approximately 20 percent to 25 percent when transfers are included as part of a family's annual income (Eurostat 1999, Smeeding 1997). Countries generally considered 'welfare leaders,' namely the Nordic countries, have, on average, lower poverty rates than other countries with fewer social programs.

2. Poverty and Child Health and Well-being

This section provides an overview of the effects of poverty on child health and well-being. The findings are discussed for five developmental age groups: prenatal to two years, early childhood (3 to 6 years), late childhood (7 to 10 years), early adolescence (11 to 15 years), and late adolescence (16 to 19 years). Each age group incorporates at least one significant transition in a child's life, such as school entrance or exit or biological maturation. The developmental challenges confronted during each period are relatively universal and may entail new modes of adaptation to biological, psychological, or social changes. The domains of well-being addressed include: physical health; school readiness and achievement; and behavior, social, and emotional problems. Each age period is associated with somewhat different indicators.

The review presented draws heavily from several recent endeavors to investigate the influence of poverty on child development. The first initiative is a cross-national comparison of key child indicators reported by UNICEF (2000), which provides a macrolevel view of how countries' overall levels of economic development affect child well-being. The second effort is a comprehensive evaluation of research on poverty and children's physical, emotional, and cognitive devel-

opment. This analysis by Brooks-Gunn and Duncan (1997) provides a 'risk assessment' of the effects of poverty on child health and well-being at a single point in time by examining outcomes for poor and nonpoor children drawn from large, national cross-sectional studies, primarily in the USA. Additional figures are reported from Canadian national data (Ross et al. 1989, Ross and Roberts 1999). In contrast, the final undertaking, an interdisciplinary initiative to conduct parallel evaluations across multiple data sets with income data measured over the life course, centers on understanding the role of the dynamics of poverty (depth, timing, and persistence) on a range of outcomes, after accounting for child and family background characteristics. These results on the long-term effects of poverty on children and youth were described in a volume, *Consequences of Growing Up Poor* (Duncan and Brooks-Gunn 1997); statistical criteria were used to determine the magnitude of income effects reported.

2.1 Physical Health

A country's level of economic development is associated with key indicators of child health; in developing countries (primarily those in sub-Saharan Africa and South Asia), over half of all deaths among children five years of age or younger are associated with malnutrition (UNICEF 2000). In many of these countries, AIDS/HIV is also contributing to rising infant mortality rates.

Research in the USA and Canada indicates that before children are born, poverty takes a toll on their health; children from poor families are approximately twice as likely to be low birth weight (less than 2,500 grams) than children who escape poverty (Brooks-Gunn and Duncan 1997, Ross and Roberts 1999, Ross et al. 1989). Other threats to poor children's health include lead poisoning and experiences with hunger as well as increased risk of mortality, hospitalization, and stunting (small for height). Poor children display more health problems and disabilities than do children from affluent families (Newacheck et al. 1993). When the longer term effects of poverty are considered, living in poverty during the childhood years has been found to be associated with stunting later in life (Duncan and Brooks-Gunn 1997).

2.2 School Readiness and Achievement

Access to basic education is associated with a country's level of economic development, such that in many developing countries, at least one-third of children are not enrolled in primary school, resulting in lower literacy rates compared with industrialized countries (UNICEF 2000). Among children of all ages, living in

poverty is also associated with deficits in cognition and achievement (Brooks-Gunn and Duncan 1997, Ross and Roberts 1999, Ross et al. 1989). Findings from US studies indicate that poor children 3 to 17 years of age are about one and a half times more likely than nonpoor children to display learning disabilities. Among school-age children, poverty increases the risk that children repeat a grade, perform poorly in school, are expelled or suspended from school, and obtain low scores on tests of school readiness and achievement. By late adolescence, poor youth are more likely than their nonpoor peers to drop out of school and to be unemployed. Taken together, these studies point to the disadvantage poverty bestows upon children's academic development; however, they do not shed light on the long-term effects of poverty on achievement.

A series of studies by Brooks-Gunn, Duncan, and colleagues have sought to understand the longer term effects of poverty on young children's cognitive and school outcomes. Findings have been highly consistent; after accounting for family and child background characteristics, income has large effects on young children's cognition, verbal ability, and achievement with a large effect defined as about five to six points on standardized tests. The depth of poverty is also important; children who were very poor (deep poverty defined as incomes 50 percent or less than the poverty threshold) scored particularly low on measures of cognitive and school achievement. The number of years a child lives in poverty matters as well for children's school readiness, particularly living persistently (all years) in poverty.

Longitudinal studies of young adolescents in the USA and Canada, generally, report small to modest income effects on indicators of school performance including grade placement and grade point average (Duncan and Brooks-Gunn 1997). Among studies of late adolescence, income was also found to have small to moderate associations with measures of completed schooling (Haveman and Wolfe 1994). The timing of poverty is important for late adolescent achievement; experiences with poverty during early childhood, as opposed to other developmental periods, are most salient for educational attainment (Duncan and Brooks-Gunn 1997).

2.3 Behavior, Social, and Emotional Problems

Similar to health and achievement, the threats to poor children's social and emotional health are great. A country's level of economic development places children at risk for exposure to hazardous and exploitative labor as well as conflict, war, and violence (UNICEF 2000). Even in the USA, poor children are twice as likely to be the victim of a violent crime and live in unsafe neighborhoods than nonpoor children (Brooks-Gunn and Duncan 1997). Among older children in Canada, poor children exhibit higher levels

of emotional problems than nonpoor children, and are approximately twice as likely to have psychiatric and conduct disorders than their nonpoor peers (Ross et al. 1989). Finally, among adolescents in the USA, the rate of teenage unmarried parenthood is three times higher among poor female adolescents than nonpoor youth. These risk figures suggest that poor children and youth are faring at less than an optimal level.

In longitudinal studies, family income has small to moderate associations with the prevalence of young and school-age children's behavior problems as reported by parents and teachers (Brooks-Gunn and Duncan 1997). The depth, timing, and persistence of experiences with poverty during early childhood also are linked to children's socioemotional well-being. Finally, small to moderate income effects on adolescent anxiety and hyperactivity have been reported (Duncan and Brooks-Gunn 1997).

3. The Pathways between Poverty and Child Development

In the previous section, the direct effects of poverty on children's and youth's health and well-being were reviewed. Together, the research presents a consistent picture of the harmful effects of poverty, particularly during early childhood. To increase our knowledge about the influence of poverty on children's development, it is necessary to identify the pathways between poverty and child development. This exercise is particularly constructive for drawing policy implications. Several possible processes have been hypothesized to transmit the effects of income to child development; each is reviewed in turn.

3.1 Home Environment

The quality of the home environment is a salient mechanism through which income may alter child development. The home environment entails: emotional warmth displayed by parents while interacting with their children; provision of stimulating and learning experiences in the home; and physical surroundings, such as safety of play areas and cleanliness. Poor children are likely to be raised in less emotionally, cognitively, and physically enriching environments than children from more affluent families. Several studies of young children have found that quality of the home learning environment accounts for a substantial share of the effect of poverty on children's achievement and behavioral outcomes.

3.2 Child Care and Schools

Child care and schools are important contexts for children that may transmit the effects of low income to developmental outcomes. During young childhood,

high quality child care and early intervention programs have demonstrated short-term and long-term benefits for poor children's academic, social, and behavioral outcomes (Yoshikawa 1994). Moreover, these programs have the potential to enhance parenting outcomes as well, such as mental health and parenting skills. Among older children and youth, income effects on achievement outcomes may operate in large part through school characteristics, such as quality, climate, and demographics (Jencks and Mayer 1990).

3.3 Parenting Behavior

Economic hardship resulting from income loss, unemployment, job instability, and economic insecurity rather than low income *per se* is another potential mediator of income effects on child well-being. According to this 'family stress model,' economic hardship is associated with parental stress and depression which leads to more inconsistent, unsupportive, harsh, and punitive parenting practices. These reduced quality parent-child interactions, in turn, are associated with emotional and school problems in children (Conger et al. 1994, McLoyd et al. 1994).

3.4 Parental Mental Health

Maternal emotional well-being may be a path through which poverty affects child development; it may operate directly on child well-being or through parenting, as discussed under the family stress models. Substantial evidence exists that poverty has negative consequences on adult mental health (Kessler 1997). Additional studies reveal links between maternal emotional well-being and children's behavior (Velez et al. 1989). One study of young children found that income might have an indirect effect on children's behavior and cognition by means of maternal depressive symptoms and coping skills (Duncan et al. 1994).

3.5 Neighborhoods

An additional mechanism through which family income may impart its effects to child outcomes is neighborhood of residence. Inadequate income constrains the choice of where a family chooses to live, and as such, low income families may reside in very poor neighborhoods characterized by crime, unemployment, lack of resources, few role models, and absence of adult supervision of children and youth. A growing body of research indicates that neighborhood conditions affect children's development above and beyond family income, rather than indirectly (Leventhal and Brooks-Gunn 2000). Affluent (compared with

middle income) or high SES neighbors appear to have beneficial effects for children's school readiness and achievement outcomes, while low income (compared with middle income) or low SES neighborhoods have negative consequences for children's and youths' emotional and social well-being. Neighborhood residence, notably income/SES, is also associated with quality of home environment, parenting practices, quality of child care and schools, and peer group attributes.

3.6 Health Care

Accessibility to health care is another pathway through which poverty may operate on child health and well-being. Although children from economically disadvantaged families are in poorer health, they are less likely to receive health care services or receive services of lower quality than children from more affluent families (Newacheck et al. 1993). Several studies suggest that access to quality care may mitigate health-related problems.

3.7 Exposure to Violence

The final process by which poverty may exert an influence on child well-being is exposure to violence inside as well as outside of the home. Children's exposure to violence as either a witness or a victim may account in part for the link between income and children's health. Both family and neighborhood income are associated with exposure to violence (Martinez and Richters 1993, Sampson et al. 1997). Exposure to violence, in turn, is associated with psychiatric symptoms.

4. Conclusion and Future Directions

In conclusion, large international discrepancies are seen in child poverty rates. Poverty rates, however, depend upon the type of measure used, demographic differences across countries, and government policies. The research indicates that poverty has negative short-term and long-term influences on a host of child outcomes. In addition, the timing, depth, and duration of experiences with poverty are also important; deep and persistent poverty during early childhood are especially detrimental to child well-being. Increasingly, researchers are beginning to explore the potential pathways through which poverty transmits its deleterious effects to children's development. There is growing evidence that the home environment and parental behavior are primary mechanisms of influence. Such knowledge is crucial for policies seeking to counter the negative consequences of poverty for poor children and their families. These policies can target

family resources directly or the mechanisms through which poverty operates on child development. In addition, policies can impact children directly through, for example, assistance with child care, or indirectly through a policy's impact on the family unit and neighborhood, for instance labor policies. Policies that coordinate the needs of both children and their parents are likely to be more efficacious than those that focus solely on one or the other. An important task for researchers will be to evaluate the effects of poverty-related policies on children's development.

See also: Child Care and Child Development; Early Childhood: Socioemotional Risks; Infancy and Childhood: Emotional Development; Infant and Child Development, Theories of; Parenting in Ethnic Minority Families: United States; Population, Economic Development, and Poverty; Poverty and Gender in Affluent Nations; Poverty and Gender in Developing Nations; Poverty, Culture of; Poverty in History; Poverty: Measurement and Analysis; Poverty Policy; Poverty, Sociology of; Urban Poverty in Neighborhoods

Bibliography

- Bradbury B, Jantti M 1999 *Child Poverty Across Industrialized Nations*. Innocenti Occasional Papers, Economic and Social Policy Series No. 71. UNICEF, Sydney, Australia
- Brooks-Gunn J, Duncan G J 1997 The effects of poverty on children. *The Future of Children* 7(2): 55–71
- Conger R D, Ge X, Elder G, Lorenz F, Simons R 1994 Economic stress, coercive family process and developmental problems of adolescents. *Child Development* 65: 541–61
- Duncan G J, Brooks-Gunn J (eds.) 1997 *Consequences of Growing up Poor*. Russell Sage Foundation Press, New York
- Duncan G J, Brooks-Gunn J, Klebanov P K 1994 Economic deprivation and early-childhood development. *Child Development* 65: 296–318
- Eurostat 1999 Social benefits and their redistributive effect on the EU. *Statistics in Focus*, Population and Social Conditions, no. 13/99. Eurostat, Luxembourg
- Haveman R, Wolfe B 1994 *Succeeding Generations: On the Effects of Investments in Children*. Russell Sage Foundation, New York
- Jencks C, Mayer S E 1990 The social consequences of growing up in a poor neighborhood. In: Geary L L (ed.) *Inner-city Poverty in the United States*. National Academy Press, Washington, DC, pp. 111–86
- Kessler R C 1997 The effects of stressful life events on depression. *Annual Review of Psychology* 48: 191–214
- Leventhal T, Brooks-Gunn J 2000 The neighborhoods they live in: The effects of neighborhood residence upon child and adolescent outcomes. *Psychological Bulletin* 126: 309–37
- Martinez J E, Richters P 1993 The NIMH Community Violence Project: II. Children's distress symptoms associated with violence exposure. *Psychiatry* 56: 22–35
- McLoyd V C, Jayaratne-Epstein T, Ceballo R, Borquez J 1994 Unemployment and work interruption among African American single mothers: Effects on parenting and adolescent socioemotional functioning. *Child Development* 65: 562–89

- Newacheck P W, Stoddard J J, McManus M 1993 Ethnocultural variations in the prevalence and impact of childhood chronic conditions. *Pediatrics* 91: 1031–8
- Ross D P, Roberts P 1999 *Income and Child Well-being: A New Perspective on the Poverty Debate*. Canadian Council on Social Development, Ottawa, ON
- Ross D P, Shillington R, Lochhead C 1989 *The Canadian Fact Book on Poverty*. Canadian Council on Social Development, Ottawa, ON
- Sampson R J, Raudenbush S W, Earls F 1997 Neighborhoods and violent crime: A multilevel study of collective efficacy. *Science* 277: 918–24
- Smeeding T M 1997 *Financial Poverty in Developed Countries: The Evidence from LIS*. Working Paper No. 155. Syracuse University, New York
- Smeeding T M, Rainwater L, Burtless G 2000 *United States Poverty in a Cross-national Context*. Syracuse University, New York
- United Nations Children's Fund (UNICEF) 2000 *The State of the World's Children 2000*. UNICEF, New York
- Velez C N, Johnson J, Cohen P 1989 A longitudinal analysis of selected risk factors for childhood psychopathology. *Journal of the American Academy of Child and Adolescent Psychiatry* 28: 861–4
- Yoshikawa H 1994 Prevention as cumulative protection: Effects of early family support and education on chronic delinquency and its risks. *Psychological Bulletin* 115: 28–54

T. Leventhal and J. Brooks-Gunn

Poverty and Gender in Affluent Nations

Women are more likely to be poor than men in nearly all Western nations, with the greatest disparity found in the USA. The underlying causes for women's poverty vary across nations but generally fall into one of three main categories—demographic composition, economic conditions, and government policy. In this article, poverty is defined, and the advantages and disadvantages of different poverty measures are noted. Then, the recent scholarship on gender and poverty is briefly reviewed, and women's poverty rates in various Western, industrialized countries are compared. Finally, the three primary causes of women's poverty are discussed.

1. Definitions of Poverty

The concept of 'poverty' is relatively new, having only entered the public lexicon in the latter half of the twentieth century. Poverty can be defined in either absolute or relative terms. It is typically measured with respect to families—not individuals—and is adjusted for the number of persons in a family (because larger families are assumed to need more income but also to benefit from economies of scale). *Absolute* measures of poverty (typically used in Anglo countries such as the

USA and the UK) define poverty in relation to the amount of money necessary to meet basic needs such as food, clothing, and shelter. *Relative* poverty measures (typically used in European countries and by academic researchers) define poverty in relation to the economic status of other members of the society. One half of the median family income is often chosen as the poverty threshold. Absolute and relative measures of poverty reflect divergent ideas about the meaning of poverty—whether it represents the inability to afford basic necessities, or comparative economic disadvantage—and both have advantages and disadvantages. The advantage of an absolute poverty measure is that it clearly indicates whether or not a family's income is sufficient to sustain a minimal level of subsistence.

However, if an absolute threshold is adjusted only for inflation, during periods of economic growth that same threshold will increasingly reflect a greater level of relative destitution. Relative measures avoid this problem because they are linked to the average standard of living: when median income goes up, the poverty threshold goes up. Yet, because of this practice, 'poverty' in one country may imply a vastly different economic status than 'poverty' in another country. Relative measures are particularly affected by the income distribution: in countries with a more unequal distribution, poverty rates will be high; in countries with a more equal distribution, they will be low—regardless of the actual standard of living that 'poverty' represents.

2. Women's Poverty in Comparative Perspective

The salience of gender for understanding poverty was first noted by Diana Pearce (1978) who coined the term 'the feminization of poverty.' Pearce observed that women and children living in female-headed

families were over-represented among the poor, and that the proportion of adults in poverty who were women had been increasing. Since that time, various scholars have used data to examine trends in men's and women's poverty rates—and the ratio between them—in order to explore how economic status may be affected by gender (see, for example, McLanahan et al. 1989, Casper et al. 1994). Feminist scholars have raised important issues about the gendered nature of welfare state policies (Folbre 1984, Gordon 1990, 1994, Orloff 1993, O'Connor et al. 1999) and the inequitable distribution of the costs of children (Folbre 1994a, 1994b).

Because poverty is measured at the family level, for married couples, poverty is by definition the same for women as for men: if a married-couple family is classified as poor, the husband, wife, and any children are all considered to be poor. Only for single individuals (defined by most researchers as neither married nor cohabiting) can the rates of poverty differ by gender. Thus, differences in the proportion of the population that is single, and differences in the poverty rates of single men and single women, underlie the overall gender differences in poverty both within and across nations.

Table 1 presents poverty rates in eight Western, industrialized nations. A relative measure of poverty is used, and an individual is counted as poor if he or she lives in a household with disposable money income that is less than half the median for households in the nation, adjusted for size. The table presents poverty rates for all women and men, for single women and men, and for two groups of single women that are particularly disadvantaged—single mothers and elderly women living alone.

Overall, the table highlights three fundamental observations about gender and poverty in highly industrialized nations. First, within each of the demo-

Table 1
Poverty rates for select demographic groups in eight industrialized nations^a

Country (year) ^b	All women	All men	Single women	Single men	Single mothers ^c	Elderly women living alone
Australia (1994)	10.7	8.2	24.7	14.8	38.6	62.1
Canada (1994)	13.1	11.6	29.1	17.4	40.5	16.2
France (1989)	10.7	9.6	16.4	13.2	24.7	NA
Germany (1994)	9.8	8.3	22.6	11.8	39.1	12.7
Netherlands (1991)	5.6	4.9	13.2	8.1	25.7	3.4
Sweden (1992)	2.9	4.0	4.8	9.3	3.4	14.7
United Kingdom (1995)	10.9	9.1	21.4	13.6	31.9	50.1
United States (1994)	15.1	11.0	32.0	16.6	47.1	43.1

Source: Figures for the first five columns are from Christopher et al. (2001); data for the last column are from Smeeding (1998).

NA: Not available.

^a Poverty is defined as less than 50 percent of a country's median income. All columns except the last are for nonelderly adults ages 25 to 54; the final column refers to elderly women ages 65 and older. 'Single' indicates that an individual is neither married nor part of a cohabiting male/female couple.

^b Years of data for the first five columns are shown in parentheses. For the figures on elderly single women, years of data are as follows: Australia 1989, Canada 1991, Germany 1989, Netherlands 1991, Sweden 1992, U.K. 1991 and US 1994.

^c Single mothers are women living with a child under the age of 18 and without an adult male.

graphic groups shown, poverty levels vary across the nations observed. For example, Sweden has very low poverty rates for most categories, while the USA has notably higher poverty rates. Second, in all but one country shown, women are more likely to be poor than men; the exception is Sweden, where men are slightly more likely to be poor than women. In the USA, the gap between women's and men's poverty rates is the highest—about four percentage points. As would be expected—since the single population drives the gender differences in poverty—the gender gap in poverty rates is larger for single individuals than for the adult population overall. Again, Sweden runs counter to the trend, with a lower poverty rate for single women than for single men. For all other countries, the difference in poverty rates between single men and women ranges from about three percentage points in France to more than 15 percentage points in the USA.

The third observation is that among single women, single mothers and elderly women living alone are two groups that face particular economic disadvantage. For single mothers, poverty rates are 25 percent or higher for all countries shown except Sweden; in Australia, Canada, and Germany, about 40 percent of single mothers are poor; and in the USA, nearly half of all single mothers live in poverty. For elderly women living alone, poverty rates range from 3 percent in The Netherlands to 62 percent in Australia, with rates of more than 40 percent observed in both the USA and the UK.

3. Why are Women Poor?

There is no single explanation for why women are poor—and why they have higher poverty rates than men—particularly as one looks across countries that have vastly different histories, cultures, and social and economic contexts. Three primary factors have emerged, and each will be discussed in turn.

3.1 Demographic Composition

Since differences in poverty rates by gender are fundamentally determined by the size and characteristics of the population that is single (neither married nor cohabiting), the demographic distribution in various countries has an important effect on women's poverty rates. Several major demographic trends have led to an increasing proportion of adult women (many of whom have children) living separately from men. First, differential mortality rates by gender imply that women live longer than men, and women are, thus, more likely to live alone in their later years. Since single elderly women have higher poverty rates than either men living alone or women living

Table 2
Single mothers in eight industrialized nations

Country (year)	Single mothers as a percent of all women ages 25 to 54
Australia (1989)	7.8
Canada (1994)	8.5
France (1989)	5.8
Germany (1994)	7.0
Netherlands (1991)	6.0
Sweden (1992)	9.9
United Kingdom (1995)	9.8
United States (1994)	12.7

Source: Data are from Smeeding et al. (1999) using the Luxembourg Income Study (LIS).

with a spouse (Smeeding 1997), in countries with a greater proportion of elderly women living alone, poverty rates for women will be higher. Yet, mortality rates among Western, industrialized nations are not highly variant; therefore, poverty among elderly women living alone has only a modest impact on overall gender differences in poverty compared to the second major demographic factor, the rise in single-parent families.

Increases in divorce and nonmarital childbearing have led to notable growth in the number of single-mother families. Although the extent of single parenthood varies across countries, the trend has been remarkably similar—single parenthood has increased in nearly all Western, industrialized nations since the 1960s (McLanahan and Casper 1995). Table 2 shows single mothers as a percentage of all women ages 25 to 54 for eight industrialized nations. Of the countries shown, the proportion of single mothers ranges from 5.8 percent in France to 12.7 percent in the USA (Smeeding et al. 1999).

If the economic costs of rearing children were evenly distributed between mothers and fathers, then even if the parents were not married, the sex-poverty ratio would not be affected. However, in reality, single parenthood has a much greater affect on women's poverty rates than men's (McLanahan and Kelly 1999). Women are more likely to have custody of a child following divorce or nonmarital childbirth, and many fathers who live away from their children do not provide adequate financial support for them. Having co-resident children decreases a household's economic status by definition, because children require economic resources but do not contribute any additional income. Unless the government provides generous financial support when the noncustodial father fails to pay (as is the case in Sweden), in countries with a high proportion of single mothers, poverty rates for women will be high. For this reason, the prevalence of single motherhood represents the major demographic factor underlying the gender disparity in poverty rates.

Table 3

Economic characteristics of women and single mothers ages 25 to 54 in eight industrialized nations

Country (year)	Labor force participation rates		Mean earnings ^a	
	All women	Single mothers	All women	Single mothers
Australia (94)	63.0	49.3	\$15,522	\$11,904
Canada (94)	73.1	63.2	\$17,391	\$16,922
France (89)	57.2	70.3	\$13,671	\$14,083
Germany (94)	68.0	67.4	\$15,033	\$14,745
Netherlands (91)	50.0	38.1	\$13,650	\$10,946
Sweden (92)	88.0	83.8	\$13,075	\$12,362
United Kingdom (95)	60.5	37.7	\$15,196	\$13,597
United States (94)	74.2	71.8	\$21,186	\$18,547

Sources: Data are from special tabulations of the Luxembourg Income Study (LIS) data by Karen Christopher at the University of Arizona.

^a For those who have non-zero earnings. Figures are in 1994 US dollars, adjusted for purchasing power parity of each country.

3.2 Economic Conditions

The rise in female labor force participation, especially among married women with children, represents one of the most dramatic socioeconomic changes in the West in the latter half of the twentieth century. Table 3 shows that labor force participation rates for all women (in eight countries) range from 50 percent in The Netherlands to 88 percent in Sweden; employment rates for single mothers vary between 38 percent in The Netherlands to 84 percent in Sweden. While women are increasingly entering the labor force, their employment rates remain lower than those of men. Further, women have average earnings that are significantly lower than those of men. As shown in Table 3, women's average earnings (among those who have any earnings) range from approximately \$13,000 in Sweden to about \$21,000 in the USA; earnings for single mothers vary between about \$11,000 in The Netherlands to \$18,500 in the USA.

Because, relative to men, women have less income from earnings to support themselves and their children, women are more likely to be poor. (This, however, is not a factor affecting elderly women's poverty as the elderly are no longer of working age.) Yet, earnings *per se* are not the only reason why women have higher poverty rates. Rather, the fact that women disproportionately bear the responsibility of rearing children has important indirect effects on their economic status. In particular, caring for children limits the amount of time that women are able to work (as well as the energy they may have available to invest in their work), and it also increases the costs of working because women must pay for out-of-home child care (not counted in the poverty measure). In addition, women who reduce their work hours or leave the labor force in order to care for children lose vital work experience which affects both their earnings and their seniority in the workplace.

3.3 Government Policy

All individuals and families receive income from two primary sources—the market and the state—in addition to any private transfers they may receive. Government policy has a significant effect on women's economic well-being because it can either obviate or intensify the inequalities that result from the operation of the market economy. Among single individuals living without children, government policy has little impact on women's poverty because men and women are largely treated the same. Although there may be some differences between elderly men and women in how pensions are allocated in various countries, differences in government policies primarily affect the economic status of single mothers and their children.

Welfare states vary dramatically in the extent to which they 'socialize' the costs of children. Feminist scholars have noted that many members of society benefit from children being brought up well (Folbre 1994a, 1994b), and yet market mechanisms fail to equitably apportion costs for this 'public good' (Christopher et al. 2001). In countries where more of the costs of childrearing are borne by governments instead of parents, women's poverty rates tend to be lower. In general, single mothers tend to benefit if a greater proportion of welfare assistance is universal rather than means-tested, because they are not faced with high marginal tax rates on their earnings that result from decreased welfare benefits. The overall level of income assistance provided to single mothers varies notably across industrialized nations; mean transfers for single mothers as a percentage of median equivalent income range from 22 percent in the USA to 71 percent in The Netherlands (unpublished tabulations of Luxembourg Income Study data by Timothy Smeeding and Lee Rainwater).

Numerous scholars have developed typologies that classify welfare states along particular dimensions.

Table 4

Poverty rates for children in single-mother families^a before and after government programs^b

Country (year)	Pretransfer poverty rate	Post-transfer, post-tax poverty rate	Percent decline
Australia (1989)	73.2	56.2	23.2
Belgium (1992)	50.7	10.0	80.3
Canada (1991)	68.2	50.2	26.4
Denmark (1992)	45.0	7.3	83.8
Finland (1991)	36.3	7.5	79.3
France (1984)	56.4	22.6	59.9
Germany (1989)	43.9	4.2	90.4
Ireland (1987)	72.6	40.5	44.2
Italy (1991)	31.7	13.9	56.2
Luxembourg (1985)	55.7	10.0	82.0
Netherlands (1991)	79.7	39.5	50.4
Norway (1991)	57.4	18.4	67.9
Sweden (1992)	54.9	5.2	90.5
Switzerland (1982)	33.7	25.6	24.0
United Kingdom (1986)	76.2	18.7	75.5
United States (1991)	69.9	59.5	14.9

Source: Rainwater and Smeeding 1995.

a Single-mother families are families where one female adult resides in the household; cohabiting-parent families are not included here. Poverty is defined as 50 percent of a country's median adjusted income.

b Government programs include income and payroll taxes and all types of government cash and near-cash transfers. Child care and other non-cash benefits are not included; this is particularly important for countries such as France which provide generous child care subsidies that are not reflected in these calculations.

Probably the best known is by Esping-Anderson (1990) which classifies welfare states into three categories according to decommodification (the ability of individuals to subsist apart from labor and capital markets) and stratification (differences between classes).

(a) *Social democratic states*, typified by the Scandinavian countries, promote gender equality and provide the most generous support to single-parent families. These nations emphasize full employment, and work and welfare are intricately intertwined, with some part-time public jobs provided for mothers. Many benefits and allowances are universal, supplemented by specific programs targeted to needy families. Women's employment is encouraged by the provision of high-quality childcare services and extensive parental leave. Also, the government guarantees that support is provided to children who do not live with both parents. Because of these policies, mothers are able to combine work and parental responsibilities and to 'package' income from both market and government sources.

(b) *Conservative-corporatist nations* also have generous transfer systems, but they are more traditional with respect to family values and expectations (largely due to the significant influence of the Church). Benefits are generally targeted toward families (rather than individuals), and inequalities across families and households may be sustained. An important component of welfare policy is social insurance, which is linked to labor force participation and may differ by

occupation. Austria, France, Germany, and Italy are representative conservative-corporatist states.

(c) In the *liberal welfare states* (notably the Anglo countries), assistance is primarily means-tested with modest universal transfers, primarily paid to the elderly or disabled. Emphasis is on the market and individuals' labor force activity as the primary means for resource allocation. Benefit levels are typically meager when compared to average wage levels, and recipients are often stigmatized by nonrecipients. Supports for working mothers, including public childcare, are limited, with little or no paid parental leave. Parents who live away from their children are expected to pay child support, although the government does not provide any assistance for children of noncustodial parents who fail to pay. Because welfare benefits are income-tested, women have a difficult time combining work and welfare in order to escape the 'poverty trap.'

These variations in welfare state policy lead to dramatic differences in the economic status of women and children. Table 4 presents poverty rates for children living in single-mother families both before and after government taxes and transfers are included in income (while childcare and other non-cash benefits are not included). The table shows that, with several exceptions, the percentage decline in the poverty rates roughly clusters according to the three categories of nations in Esping-Anderson's typology. Government tax and transfer policies reduce the poverty rates in each of the four social democratic nations shown in the table (Denmark, Finland, Norway, and Sweden)

by 68 percent or more. In the corporatist nations, poverty reductions fall in the middle of the spectrum with a 60 percent decline in France and a 56 percent decline in Italy. Germany is an important exception, with its government policies reducing poverty for children in single-mother families by fully 90 percent. The lowest reductions in poverty rates are noted in the liberal states, such as 26 percent for Canada, 23 percent for Australia, and 15 percent for the USA. The UK represents another exception because its policies reduce poverty by 76 percent—a much greater decline than in the other liberal states.

Some feminist scholars have criticized Esping-Anderson's framework because it emphasizes workers' dependence on employers while ignoring women's dependence on men. Orloff (1993) argues, for example, that welfare regimes should be judged by the extent to which they allow women to establish independent households. In contrast, other feminist scholars have raised questions about women's dependence on the welfare state (Gordon 1990, 1994) and gender biases in the welfare state (Nelson 1990).

4. Conclusion

Ultimately, reducing women's poverty in industrialized countries requires the convergence of all three factors—demography, the market economy, and government policy—to better support women and their children. This is because women's economic status depends on whatever resources they can conjointly obtain from the family, the market, and the state (Christopher et al. 2001). First, women's poverty is closely tied to whether the responsibility for children is shared equally between mothers and fathers. Private transfers, regulated by the government, may be just as important as public transfers in compensating for the demographic changes that have resulted in women bearing a disproportionate share of the cost of children. Second, women benefit from market economies that pay men and women equally. Despite the recent gains in wages relative to men, women (especially mothers) earn less than men in all Western countries. Countries that promote gender equity in wages and that maintain high minimum wages reduce the poverty rate of all women, especially single mothers. Third, government policy can be assessed both in terms of the level of support provided to single mothers and the extent to which public support encourages or discourages labor force participation. In particular, childcare subsidies and parental leave policies enable women to balance their responsibilities as parents and workers, which ultimately reduces poverty and increases economic independence.

See also: Divorce and Gender; Domestic Violence: Sociological Perspectives; Economic Development

and Women; Economic Globalization and Gender; Family and Gender; Family Bargaining; Gender, Economics of; Inequality; Labor Movements and Gender; Lone Mothers in Affluent Nations; Motherhood: Economic Aspects; Poverty and Gender in Developing Nations; Poverty Policy; Poverty, Sociology of; Power in Society; Race and Gender Intersections; Sex Differences in Pay; Sex Segregation at Work; Social Welfare Policies and Gender; Underclass; Urban Poverty in Neighborhoods; Welfare Programs, Economics of; Welfare State

Bibliography

- Casper L M, McLanahan S S, Garfinkel I 1994 The gender-poverty gap: what we can learn from other countries. *American Sociological Review* **59**: 594–605
- Christopher K, England P, McLanahan S, Ross K, Smeeding T M 2001 Gender inequality in poverty in affluent nations: the role of single motherhood and the state. In: Vleminkx K, Smeeding T M (eds.) *Child Well-Being, Child Poverty and Child Policy in Modern Nations*. The Policy Press, Bristol, UK, pp. 199–219
- England P 1997 Gender and access to money: what do trends in earnings and household poverty tell us? *8th International Meeting. Comparative Project on Class Structure and Class Consciousness*. Canberra, Australia
- Esping-Anderson G 1990 *The Three Worlds of Welfare Capitalism*. Polity Press, Cambridge, UK
- Folbre N 1984 The pauperization of motherhood: patriarchy and public policy in the United States. *Review of Radical Political Economics* **16**(4): 72–88
- Folbre N 1994a Children as public goods. *American Economic Review* **84**: 86–90
- Folbre N 1994b *Who Pays for the Kids?: Gender and the Structures of Constraint*. Routledge, New York
- Gordon L (ed.) 1990 *Women, the State, and Welfare 1890–1935*. University of Wisconsin Press, Madison, WI
- Gordon L 1994 *Pitied But Not Entitled: Single Mothers and the History of Welfare*. Harvard University Press, Cambridge, MA
- McLanahan S, Casper L 1995 Growing diversity and inequality in the American family. In: Farley R (ed.) *State of the Union: America in the 1990s, Vol. 2 Social Trends*. Russell Sage Foundation, New York, pp. 1–45
- McLanahan S S, Kelly E L 1999 The feminization of poverty: past and future. In: Chafetz J (ed.) *Handbook of the Sociology of Gender*. Kluwer, Dordrecht, The Netherlands, pp. 127–45
- McLanahan S S, Sorenson A, Watson D 1989 Sex differences in poverty, 1950–1980. *Signs: Journal of Women in Culture and Society* **15**: 102–22
- Nelson B J 1990 The origins of the two-channel welfare state: workmen's compensation and mother's aid. In: Gordon L (ed.) *Women, the State, and Welfare*. University of Wisconsin Press, Madison, WI
- O'Connor J S, Orloff A S, Shaver S 1999 *States, Markets, Families: Gender, Liberalism, and Social Policy in Australia, Canada, Great Britain, and the United States*. Cambridge University Press, New York
- Orloff A S 1993 Gender and the social rights of citizenship. *American Sociological Review* **58**: 303–28

- Pearce D M 1978 The feminization of poverty: women, work and welfare. *Urban and Social Change Review* 11: 28–36
- Rainwater L, Smeeding T M 1995 Doing poorly: the real income of American children in a comparative perspective. In: Skolnick J H, Currie E (eds.) *Crisis in American Institutions*. Allyn and Bacon, Boston, pp. 118–25
- Smeeding T M 1998 Reshuffling responsibilities in old age: the United States in a comparative perspective. In: Gonyea J (ed.) *Re-Securing Social Security and Medicare: Understanding Privatization and Risk*. Gerontological Society of America, Washington, DC, pp. 24–36
- Smeeding T M, Ross K, England P, Christopher K, McLanahan S S 1999 Poverty and parenthood across modern nations: findings from the Luxembourg Income Study. Luxembourg Income Study, Working Paper No. 194

S. S. McLanahan and M. J. Carlson

Poverty and Gender in Developing Nations

Poverty as used in contemporary literature is an economic measure of income or consumption. A person or household is considered below the 'poverty line' if such income or consumption falls below a minimum standard needed to provide basic human needs. Economic development in developing countries is considered successful when the gross national product (GNP) increases and raises the per capita income. Rapid population increase dilutes the impact of GNP and led to international family planning programs as one approach to poverty reduction. This direct connection between economic development and population decrease is often questioned. The insensitive implementation of family planning programs has been widely criticized by women who initiated the emphasis on women's education and health to reduce both poverty and family size at the 1994 United Nations International Conference on Population and Development in Cairo (see *Reproductive Rights in Developing Nations*).

The poverty line is linked to a minimum standard within each country which varies by country, region, or culture. Attempts to compare income poverty across societies has led the World Bank to utilize the Purchasing Power Parity (PPP) that accounts for currency fluctuation and local consumption habits and is commonly referred to as the market-basket approach.

Although economic measures dominate the literature on poverty and on development, critics have objected to such sterile methods of accounting. Additional indicators of wellbeing have been combined into more textured indices of progress or poverty. The social aspects of poverty do not lend themselves to such aggregation. To understand the multidimen-

sional nature of poverty, scholars are today investigating a broad range of social and cultural practices as well as governmental policies. Only with this comprehension of the causes of poverty can more effective programs for poverty alleviation be designed. This effort has taken on increased urgency in light of current economic trends which have exacerbated income disparity within countries and pushed more people into poverty (Lundberg and Milanovic 2000).

Differential patterns of poverty are experienced by women and men due largely to their social roles and responsibilities. The increase in the number of women-headed households throughout the world, caused by rapid economic transition, migration, and urbanization, has led to the dissolution of many traditional support systems. As a result, women must both nurture and provide a livelihood for their children in an economy that privileges full-time work away from home. This double burden limits both a woman's choice of jobs and her mobility, thus reducing her income causing the feminization of poverty.

Even within needy families, women may be poorer than their husbands because they have little control over how income or food are allocated. Food aid programs now deliver food directly to poor women who show more equity in feeding their children Pinstrup-Andersen et al. 1999. Similarly, development agencies have focused more of their income projects on women (see *Economic Development and Women*). While some critics argue that loading additional work responsibilities on women has created a crisis of time, others demonstrate that income increases women's household bargaining power (see *Family Bargaining*). Conservatives object to programs that provide women with jobs because they believe that upsetting the gender division of labor encourages further disintegration of families. In contrast, feminists argue that the focus on welfare programs to alleviate the worst aspects of poverty will do nothing to reduce the subordination of women. Recent approaches to alleviating poverty emphasize asset creation and housing rights to empower women.

The first section explores the broadening of definitions, measures, and social indicators of poverty. The second section presents arguments for a gendered approach to poverty alleviation and discusses studies that elaborate on, or question, earlier assumptions. The last section describes significant programs designed to assist women out of poverty and reviews salient criticisms of them.

1. Definitions and Causes of Poverty

Defining poverty in economic terms alone raises issues of the accuracy of statistics in developing countries and encounters the range of moral objections to valuing life only in material terms. Development

agencies are aware of the unreliability of data from developing countries yet cognizant of the demands by policy makers for easily replicated measures. To compare poverty globally, the World Bank uses estimates of local income expressed as US\$1 and US\$2 per day to identify levels of poverty, estimating that in 1998, 1.2 billion people worldwide had consumption levels below \$1 a day and 24 percent of the population of the developing world and 2.8 billion lived on less than \$2 a day (World Bank 2000).

1.1 Indicators Beyond Income

As the GNP became the common method of measuring progress, critiques appeared. Alternative measures, such as the Physical Quality of Life Index (PQLI) promoted by the Overseas Development Council in Washington DC in the 1970s, used indicators for life expectancy, infant mortality, and literacy for an index that often reversed ideas of improved life through increased income. For example, Saudi Arabia rated high for income and low for PQLI while Sri Lanka had a low GNP and a high PQLI. These indicators actually measure women's status, though that was not the original intent, since poor health for women results in high maternal mortality that often also means high infant mortality while women's literacy reflects educational opportunities and cultural openness.

Following this trend, the Human Development Index (HDI) refines these indicators into three indices based on life expectancy, education, and gross domestic product (GDP). Utilizing these and other indices, the Human Development Report makes explicit the need for gender analysis in the Gender Development Index (GDI) and the Gender Empowerment Measure (GEM). Saudi Arabia is placed in the medium human development list but ranks low in GDI and has no entry for GEM. Three developing countries, Sri Lanka, Thailand, and Jamaica, have a GDI rank that is more than 10 points higher than the HDI rank (UNDP 1998).

In 1997 the Human Development Report introduced the Human Poverty Index (HPI) to emphasize that poverty is much more than a lack of material wellbeing, but also a denial of opportunities to live a healthy and creative life and enjoy both self-esteem and the respect of others. While the HDI measures national progress, the HPI measure the deprivations of those who are left out of progress (UNDP 1998 p. 25). Significant is the finding that the US has highest human poverty among the 17 developed countries listed although it has the highest per capita income measured in PPP.

1.2 Social Aspects of Poverty

As world poverty increases, development agencies are searching for more accurate social indicators for edu-

cation, health, access to services and infrastructure. Other nonincome dimensions of poverty are being studied to tease out and evaluate illusive patterns of vulnerability, social exclusion, and gender inequities. Analyses of poverty in terms of access to social capital, entitlements to food, income, land, and houses are being conducted. These indicators do not lend themselves to aggregation into a single index. Rather this broadening of research reflects the complexities of addressing social aspects of persistent poverty (Chambers 1995). Reports on urban poverty in cities around the world and programs that improve the lives of the poor have been published in three issues of *Environment & Development*. In preparation for the World Development Report 2000/1 on Poverty and Development, the World Bank commissioned a wide range of working papers (World Bank 2000). Gendered poverty is the concern of studies produced by the United Nations Research Institute for Social Development (UNRISD 1999). The Inter-American Development Bank on causes of intergenerational transmission of poverty has resulted in an initiative on early childhood programs.

1.3 Structural Influences on Poverty

Economic and political forces external to culture and community also contribute to the current trend of increased poverty and income disparity around the world. The rapidity of the present economic transitions toward global trade and largely unfettered markets is exacerbating income disparity and contributing to the increase in poverty. Economic prescriptions to improve creditworthiness of countries, such as the Structural Adjustment Policies (SAP) of the International Monetary Fund and the World Bank, are widely blamed for the reduction of health and education budgets in those countries. Subsequently, these financial institutions have instituted changes in conditionality to require governments to balance cuts in the social safety net with other measures to reduce expenditures by trimming bloated bureaucracies or eliminating inefficient industrial projects.

Governments privilege different groups through macroeconomic choices regulating trade, investments, and food policy, among others (UNRISD 1999). Urban poverty is much influenced by the actions of city and municipal governments. Inefficient and corrupt bureaucracies contribute to income inequality and limit the effectiveness of poverty alleviation programs. In 1999, research teams in 23 countries held *Consultations with the Poor* (Narayan et al. 1999). A large majority of the poor said they were worse off, had fewer opportunities, and were more insecure than in the past. Most had little good to say about government personnel or services and turned to their own networks and institutions for support. The presence of non-governmental organizations (NGOs) was uneven; their programs received mixed reviews, often seen as

irrelevant because, despite the goals of participation espoused by NGOs, the poor do not feel that their voices are heard.

2. Poverty as Gendered

The gender division of labor traditionally allocated domestic production to women, activities that included providing subsistence as well as care to their families. The shift from labor exchange to a monetized economy tended to leave women with tasks and responsibilities that are not paid and therefore are excluded from national accounts. Opportunities for men to earn money are more plentiful but may require migration for short or long periods of time, weakening family cohesion and kin networks while changes in traditional status hierarchies as well as income paths increases insecurity. Once men earn money they often feel entitled to spend it on drinking, gambling, or women rather than on their families. These events, drinking, insecurity, and lack of kin supervision, have led to a worldwide epidemic of domestic violence and AIDS (see *Domestic Violence: Sociological Perspectives*).

2.1 Women-headed Households

The same transitions encourage the increase of households headed by women (see *Economic Development and Women*; *Economic Globalization and Gender*). Definitions of and data on women-headed households are contested, as are most statistics in developing countries. Estimates of 30 percent have been found valid since 1975, but studies have refined categories to note the numbers of widows, or of women with children returning to parental homes and becoming invisible. Women-headed households are perceived as the poorest, though regional and class variations are noted. The International Food Policy Research Institute (IFPRI) reported that women-headed households in more than half of the countries examined in Asia and sub-Saharan Africa are poorer than households containing adult men. Recent household studies link urban poverty not only to gender but to life cycle, age, and housing options. Food security is of prime concern in poor households; women are more likely than men to direct more resources to food and children (United Nations 1995, p. 129). These findings have important implications for poverty alleviation programs.

2.2 Time Pressures of Women's Double Day

In order to purchase needed commodities or pay school fees, women are pushed to earn income as well

as carry out domestic production. The dual responsibilities of women tend to limit their income opportunities and have generally resulted in women earning less than men in most of the world. Statistics on women's work, along with a discussion of problems of the definition of work and the imprecision of much data, are frequently published by the UN's Statistical Division (United Nations 1995). Time-use studies underscore the fact that women work longer hours than men in a large number of countries (see *Time-use and Gender*). This has led feminist scholars to promote appropriate technology to reduce women's work burdens (Tinker 1987). Others deplore the labor-intensive poverty programs since heavy physical work, called for in such poverty alleviation projects as road building, reduce the already limited nutritional reserves of poor women (UNRISD 1999). Government cost reduction programs that cut social spending generally add to women's time pressures as they increase hours of care-giving to children and relatives and often sacrifice income activities in the process (Moser 1996).

3. Poverty Alleviation Programming

Poverty reduction has been a major goal of economic development since the thrust toward basic human needs and redistribution with growth identified in the Second Development Decade, 1970–80 documents. The focus on basic needs encouraged development programs to address the concerns of people not just countries and promoted the women in development movement. Pressures from the growing global women's movement forced the debates over structural adjustment and free trade to consider their impact on women (see *Economic Development and Women*). At the end of the twentieth century, a New Policy Agenda emerged which considered market-led growth as the primary engine for reducing poverty. In developing countries, agriculture and labor-intensive activities were expected to provide employment. The role of the state was limited to service delivery, assisted by NGOs and other civil society organizations. The fact that the unfettered global market economy has led to more poverty has forced rethinking of poverty and its alleviation as noted above.

3.1 Programs to Alleviate Poverty

In the 1970s poverty alleviation programming was focused on rural areas. Appropriate technology was introduced to reduce women's energy and time required for subsistence activities such as collecting water and fuel, and processing food. Improved cookstoves and water-seal toilets also reduced the needs for water and fuel. Time saved became available for

literacy programs or income activities which at first were ill-adapted to poor women's skills (see *Economic Development and Women*).

3.2 Microcredit and Microfinance

Most current poverty alleviation programs focus on savings, credit for microenterprise, and the organizing home-based workers through NGOs. In combination these programs assist the poor women and men to accumulate assets and expand ties within the community which in turn encourages investments in human capital and creates social capital. The impact is more dramatic for women than men in those traditional societies where women and men are segregated socially, where women's mobility is severely limited, and where women's work is largely unremunerated. Not surprisingly, two of the most successful microcredit programs are in Bangladesh, the Grameen Bank, and India, SEWA (Self-Employed Women's Association). Savings are a prerequisite for credit which is utilized to start or improve tiny entrepreneurial activity from raising a goat to selling street foods to marketing fish. All programs require frequent attendance at group meetings which function both to break down caste, ethnic, and kin barriers and provide mechanisms for wide-scale organizing. SEWA has spearheaded the rapid expansion of Homenet, a worldwide organization for home-based workers which advocates for the extension of pensions, health benefits, and childcare to these workers. Meetings of these organizations promote the sharing of information on health, contraception, nutrition, and dealing with domestic violence as well as on business techniques. Women learn the benefits of investing in the health and education of their children, especially daughters. The Grameen Bank is proud of the number of children of members now attending secondary schools and colleges.

Profits from microenterprise and the savings programs have allowed women to purchase more durable assets such as a house or agricultural land. Many urban NGOs support programs by which the poor obtain housing through sweat-equity and credit. Women as the primary family caretakers need the security of shelter. Most laws or customary practices in developing countries still privilege male ownership of house and land (see *Land Rights and Gender*). Registration for these assets remains an obstacle though laws and traditions are changing, often under pressure from NGOs (Tinker and Summerfield 1999). The surplus of widows from war and AIDS has recently convinced both Uganda and Rwanda to allow women to inherit land. Globally, the rise of women-headed households and absentee partners, now exacerbated in postconflict regions, and their critical need for shelter has prompted the Gender Unit of the United

Nations Centre for Human Settlements to mount a campaign for security of house tenure for women (UNCHS 1999).

Programs to provide income and housing to women empower them and enhance their household bargaining position (see *Family Bargaining*). The social capital gained through membership in organizations enhances women's community presence and provides them with a voice in elections. Such power addresses gender inequalities and provides ways for women to rise out of poverty.

See also: Boserup, Ester (1910–99); Development, Economics of; East Asian Studies: Gender; Economic Development and Women; Economic Globalization and Gender; Family Planning Programs: Development and Outcomes; Family Planning Programs: Feminist Perspectives; Fertility Transition: China; Fertility Transition: East Asia; Fertility Transition: Economic Explanations; Fertility Transition: Latin America and the Caribbean; Fertility Transition: Southeast Asia; Gender and Technology; Latin American Studies: Gender; Near Middle East/North African Studies: Gender; Poverty and Gender in Affluent Nations; Poverty Policy; Poverty, Sociology of; Reproductive Rights in Developing Nations; Rural Industrialization in Developing Nations and Gender; Southeast Asian Studies: Gender

Bibliography

- Chambers R 1995 Poverty and livelihoods: whose reality counts? *Environment and Urbanization* 7(1): 173–204
- Lundberg M, Milanovic B 2000 Globalization and inequality: are they linked and how? Working paper. World Bank, Washington, DC
- Moser C O N 1996 *Confronting Crisis: A Comparative Study of Household Responses to Poverty and Vulnerability in Four Poor Urban Communities*. Environmentally Sustainable Development Studies and Monographs Series No. 8. World Bank, Washington, DC
- Narayan D, Chambers R, Shah M, Petesch P 1999 Global Synthesis: Consultations with the Poor. Discussion Draft, Poverty Group, World Bank, Washington, DC
- Pinstrup-Andersen P, Pandya-Lorch R, Rosegrant M W 1999 *World Food Prospects: Critical Issues for the Early Twenty-First Century*. International Food Policy Research Institute, Washington, DC
- Renner M, Sheehan M O 2000 *Vital Signs 2000: The Environmental Trends That Are Shaping Our Future*. Worldwatch Institute, Washington, DC
- Tinker I 1987 The real rural energy crisis: Women's time. *The Energy Journal* 8(87): 125–46
- Tinker I, Summerfield G (eds.) 1999 *Women's Rights to House and Land: China, Laos, Vietnam*. Lynne Rienner, Boulder, CO
- United Nations 1995 *The World's Women 1995: Trends and Statistics*. Prepared by the Statistical Division, ECOSOC, United Nations, New York

United Nations Centre for Human Settlements 1999 *Women's Rights to Land, Housing and Property in Postconflict Situations and During Reconstruction: A Global Overview*. Land Management Series #9. UNCHS, Nairobi, Kenya

United Nations Development Programme 1998 *Human Development Report 1998*. Published annually since 1990 with different emphases. Oxford University Press, New York

United Nations Research Institute for Social Development (UNRISD), United Nations Development Programme (UNDP), Centre for Development Studies, Kerala 1999 *Gender, Poverty and Well-Being: Indicators and Strategies*. Report of the UNRISD, UNDP and CDS International Workshop, November 1997. UNRISD, Geneva

World Bank 2000 *World Development Report 2000/1: Attacking Poverty*. Oxford University Press, New York

I. Tinker

Poverty, Culture of

The culture of poverty concept was developed in the USA during the 1960s primarily through the best-selling ethnographic realist publications of the cultural anthropologist Oscar Lewis, who tape-recorded eloquent life histories of the urban poor. He reprinted numerous versions of his definition of the term 'culture of poverty' in short journal articles and also in the introductions to his books on family life among Mexicans, Puerto Ricans, and Cubans living in shanty towns and ghettos (Lewis 1961, 1966a, 1966b, 1967). Lewis's culture of poverty struck an academic identity politics nerve, and at the turn of the millennium the concept remained enmired in a bitter polemic over how to analyze and engage politically the persistence of poverty in the midst of postindustrial plenty.

1. Ideological Backdrop to the Culture of Poverty

In the USA, irrespective of the theoretical orientation of researchers, most discussions on poverty polarize around value judgments concerning individual self-worth or around racial/ethnic stereotypes. US attitudes towards poverty are rooted in the country's colonial Calvinist/Puritanical heritage and are exacerbated by the historical importance of racialized hierarchies that have legitimized genocide, slavery, colonization, and immigration control. This helps explain why the culture of poverty concept continues to generate so much emotional heat while shedding so little conceptual light. The uses and misuses of the concept offer a fascinating case study in the sociology of knowledge illustrating the political interfaces between theory, empiricism, art, and ethnocentric moralizing in the social sciences.

Poverty research throughout history has been more successful at reflecting the biases of an investigator's society than at analyzing the experience of poverty. The state of poverty research in any given country emerges almost as a litmus for gauging contemporary social attitudes toward inequality and marginalization. For example, while Lewis's books are read by a US public as an individualistic interpretation of the persistence of poverty that blames victims, in France his work is interpreted as a critique of society's failure to remedy the injuries of class-based inequality under free market capitalism.

2. Defining the Culture of Poverty

The socialist sociologist Michael Harrington was the first prominent academic to use the phrase 'culture of poverty' in a major publication. His book, *The Other America*, documented rural poverty in Appalachia and represented a moral call to action that anticipated the War on Poverty initiated by President Johnson in 1964 (Harrington 1962). As a first-generation son of impoverished Jewish immigrants who was influenced by Marxism in his youth, Lewis shared Harrington's social democratic commitment to combating poverty (Rigdon 1988, Harvey and Reed 1996). Ironically, however, Lewis's popularization of the culture of poverty concept is said to have tolled an intellectual death knell to the optimistic idealism of the mid-1960s that advocated eradicating poverty through direct state intervention (Katz 1989). This is because Lewis's definition of the culture of poverty stressed that a significant minority of the poor (approximately 20 percent) were trapped in self-perpetuating cycles of dysfunctional behaviors and attitudes: 'By the time slum children are age six or seven they have usually absorbed the basic values and attitudes of their subculture and are not psychologically geared to take full advantage of changing conditions or increased opportunities' (Lewis 1965a, p. xlv).

This kind of psychological reductionist and individualistic interpretation of the persistence of poverty resonated with US popular blame-the-victim discourse. Ironically, in the same articles or book introductions in which he defined the culture of poverty, Lewis also included radical political statements contradictory to the implication that poverty is caused by self-perpetuating deficient value systems. For example, in his *Scientific American* version of 'The Culture of Poverty' he quotes Frantz Fanon, praises Castro's Cuba, and criticizes 'free-enterprise, pre-welfare-state stage capitalism' for spawning the culture of poverty (Lewis 1966b). At the same time he states that 'it is much more difficult to undo the culture of poverty than to cure poverty itself,' and advocates 'psychiatric treatment' for poverty in the USA (Lewis 1966b, p. 25).

In other words, the culture of poverty concept was confused theoretically at its inception. Unfortunately, Lewis never managed to clarify what he intended to mean. His published correspondence reveals that he was profoundly disturbed by the blame-the-victim interpretation of the causes of poverty that he triggered in the USA: 'There is nothing in the concept that puts the onus of poverty on the character of the poor' (Lewis 1967, p. 499).

The notion of a culture of poverty, consequently, should not be treated as a full-blown theory. As presented by Lewis, it was merely a bundle of some 70 traits which he did not link to a particular processual or dynamic logic. In fact, he never even listed all 70 of the traits that he claimed existed. The theoretical sloppiness of the culture of poverty concept may well be a product of the McCarthyist anticommunism in US academia that impinged on Lewis during his formative years in the 1950s. Four decades after their inception, his lists of culture of poverty character traits appear embarrassingly arbitrary, ethnocentric, and psychologically reductionist (Lewis 1966a, p. xlviii):

... a high incidence of maternal deprivation, of orality, of weak ego structure, confusion of sexual identification, a lack of impulse control, a strong present-time orientation with relatively little ability to defer gratification and to plan for the future, a sense of resignation and fatalism ... male superiority, and a high tolerance for psychological pathology of all sorts.

3. *The Policy Implications of the Culture of Poverty*

In the late 1960s and 1970s Lewis's culture of poverty concept produced an outpouring of political and academic reactions, primarily in the USA but also in Mexico and Puerto Rico (Leacock 1971, *Alteridades* 1994). Despite conceiving of his work as a call for the expansion of public-sector intervention on behalf of the poor, his concept took the popular intellectual spotlight off the need for structural economic reform, and glossed over the social power dynamics revolving around class, ethnic, gender, and colonial inequalities. Policy makers, if they paid any attention to the culture of poverty concept, interpreted it as advocating the need to rehabilitate the deficient cultural value systems of poor children through the agency of psychiatric social workers. For example, in the applied policy realm Lewis consulted in the development of the Head Start Program in the USA (Rigdon 1988), which has been criticized retrospectively as an attempt to 'take inner-city preschoolers who live in lead-painted, rat-infested tenements without steady heat or hot water, and metamorphose them into bright-eyed, upper-middle-class overachievers' (Bourgeois 1995, p. 325). Significantly, in the year 2001, Head Start was still

identified by both liberals and conservatives as one of the only successful antipoverty programs of the 1960s.

4. *The Theoretical Implications of the Culture of Poverty*

Unfortunately, most of the hostile academic responses to Lewis's culture of poverty concept have limited themselves to contradicting Lewis's empirical assertions, rather than to critiquing theoretically his psychological reductionism, his sloppy use of the culture concept, and his failure to link in a dynamic manner macrostructural political and economic forces—including gender power relations—to ideology, culture, and individual values (Valentine 1968, Stack 1974; for a political economy exception, see Katz 1989; for a feminist literary criticism exception, see Franco 1989). The bulk of the negative reaction hinges on a political concern for replacing the negative imagery of Lewis's painful but expressive ethnographic portraits of the everyday suffering of urbanized families, with positive images of the worthy poor, struggling for upward mobility against all odds. A late 1990s rehabilitating of the culture of poverty concept from a Marxist perspective dismissed the virulence of the US progressive reaction against the culture of poverty concept as a sectarian 'ultra Bolshevism' that swept the New Left when the general public was drifting ideologically to the Right following the War on Poverty. This precipitated a 'fruitless game of radical one-upmanship' among frustrated intellectuals, who were completely marginal to public political discourse, and who chose instead to devote their energies to proving their dedication to protecting the image of the poor (Harvey and Reed 1996). More importantly, the urgent righteousness of the anti-culture of poverty social science literature is comparable to the polemics against Moynihan's 1967 patriarchal attribution of the 'tangle of pathology' in the black family as being the central cause for the persistence of poverty among urban African-Americans (Rainwater and Yancey 1967).

The angry denial by academics of the existence of the types of violence and self-destructive behaviors described ethnographically by Lewis among the vulnerable families that he tape-recorded and described reveals how far removed intellectuals can be from the inner-city street. Although Lewis's writing deserves criticism for presenting his subjects in a decontextualized pornography of violence, sexuality, and emotional brutality, none of the behaviors or personalities described by Lewis should shock anyone who is familiar with everyday life in the US inner city or Latin-American shanty towns. On the contrary, Lewis's ethnographic realist descriptions, unfortunately, still ring true four decades after they were written. His disturbing material, however, demands

theoretical explanation and political contextualization, and that is where both Lewis and his critics and admirers have largely failed. By confining the debate to a worthy vs. unworthy poor dichotomy, the internecine squabbles between leftist, liberal, and conservative academics mimetically reproduce the right-wing hegemony in popular US culture that equates poverty with sinfulness.

Arguably, the polemics of righteousness that the culture of poverty prompted scared a generation of social scientists away from ethnographic analyses of inner-city poverty in the USA and, to a lesser extent, around the world (Wilson 1987, pp. 13–16). Indeed, accusations of supporting a ‘culture of poverty interpretation’ are still frequently invoked in polemical identity politics attacks between academics over any representation of poverty that is not flattering to poverty’s victims (cf. Lassalle and O’Dougherty 1997). Hence the virulence of the ‘underclass debate’ in sociology spawned by William Julius Wilson’s book *The Truly Disadvantaged* during the late 1980s through the 1990s (Katz 1989, Wacquant 1997).

From a theoretical perspective, the legacy of the culture of poverty debate has impoverished research in the social sciences on the phenomenon of social suffering, everyday violence, and the intimate experience of structural oppression in industrialized nations. Most importantly, by remaining mired in debates driven by identity politics, researchers have minimized the painful experience of day-to-day survival among the persistently poor. Epidemiological data on the associations between social class interpersonal violence, domestic violence, health outcomes, education outcomes, substance abuse, etc. are simply ignored by most poverty researchers in the USA.

The vacuum of critical intellectual engagement with the phenomenological experience of poverty has allowed right-wing academics subscribing to facile neoliberal blame-the-victim interpretations to capture popular imagination and policy debates—especially in the USA. Social or cultural reproduction theory, which emerged out of studies of poor youth at the intersection of the disciplines of education, sociology, and anthropology during the 1980s and early 1990s, offered a critical theoretical alternative. By focusing on the power dynamics of the interface between culture and social inequality, social/cultural reproduction theorists address the empirical reality of the existence of patterns of interpersonal self-destruction without obscuring structural political forces. Although vulnerable to critique for being overly functionalist, these theories allow for the reinscription of agency among the poor, as well as an autonomous role for culture in political economy (Bourdieu and Passeron 1977, Willis 1981).

Few serious social science researchers, if any, at the dawn of the twenty-first century would dare utilize the term ‘culture of poverty’ in their analysis, despite the fact that their empirical and theoretical work addresses

cultural expressions of social suffering due to the entrenchment of urban poverty in industrialized and postindustrialized societies. The problematic analytical and political utility of the culture of poverty concept demonstrates how dangerously essentializing the phrase ‘culture of ...’ can become with respect to any concept. Indeed, anthropologists cannot agree upon a useful definition for culture; nor do they understand how it operates without turning it into a black box of totalizing essences (Gupta and Ferguson 1997).

The culture of poverty furore reminds us that academics fight so hard over so little especially when marginalized political perspectives are at stake. At the turn of the millennium, much of the world’s population survives precariously in shanty towns, housing projects, tenements, and homeless encampments where mind-numbing, bone-crushing experiences of poverty engulf the socially vulnerable. Meanwhile, concerned academics continue to fiddle in their ivory towers, arguing over how to talk correctly about the structural violence of poverty.

See also: Class; Social; Conflict Sociology; Income Distribution; Inequality; Poverty: Measurement and Analysis; Poverty Policy; Poverty, Sociology of; Racism, History of; Racism, Sociology of; Underclass; Urban Poverty in Neighborhoods; Wealth Distribution

Bibliography

- Alteridades 1994 Oscar Lewis Revisitado. *Alteridades* 4 Special Issue
- Bourdieu P, Passeron J-C 1977 *Reproduction in Education, Society, and Culture*. Sage Publications, London
- Bourgeois P I 1995 *In Search of Respect: Selling Crack in El Barrio*. Cambridge University Press, New York
- Franco J 1989 *Plotting Women: Gender and Representation in Mexico*. Columbia University Press, New York
- Gupta A, Ferguson J (eds.) 1997 *Culture, Power, Place: Explorations in Critical Anthropology*. Duke University Press, Durham, NC
- Harrington M 1962 *The Other Poverty: Poverty in the United States*. Macmillan, New York
- Harvey D L, Reed M H 1996 The culture of poverty: an ideological analysis. *Sociological Perspectives* 39: 465–95
- Katz M B 1989 *The Undeserving Poor: From the War on Poverty to the War on Welfare*. Pantheon Books, New York
- Lassalle Y, O’Dougherty M 1997 Death without weeping: The violence of everyday life in Brazil. *Radical History Review* 69: 243–60
- Leacock E 1971 *The Culture of Poverty: A Critique*. Simon and Schuster, New York
- Lewis O 1961 *The Children of Sanchez: Autobiography of a Mexican Family*. Random House, New York
- Lewis O 1966a *La Vida: A Puerto Rican Family in the Culture of Poverty—San Juan and New York*. Random House, New York
- Lewis O 1966b The culture of poverty. *Scientific American* 215: 19–25

- Lewis O 1967 The children of Sanchez—Autobiography of a Mexican family. *Current Anthropology* 8: 480–500
- Rainwater L, Yancey W L (eds.) 1967 *The Moynihan Report and the Politics of Controversy*. MIT Press, Cambridge, MA
- Rigdon S M 1988 *The Culture Façade: Art, Science, and Politics in the Work of Oscar Lewis*. University of Illinois Press, Urbana, IL
- Stack C B 1974 *All Our Kin: Strategies for Survival in a Black Community*. Harper and Row, New York
- Valentine C A 1968 *Culture and Poverty: Critique and Counter-proposals*. University of Chicago Press, Chicago
- Wacquant J D 1997 Three pernicious premises in the study of the American ghetto. *International Journal of Urban and Regional Research* 21(2): 341–53
- Willis P 1981 *Learning to Labor*. Columbia University Press, New York
- Wilson W J 1987 *The Truly Disadvantaged: The Inner City, the Underclass, and Public Policy*. University of Chicago Press, Chicago

P. Bourgois

Poverty in History

1. The Concept of Poverty

There is no universal concept of poverty which can be applied to all cultures, societies, or times. Nowadays, poverty is usually defined in relation to the medium-income level of a society, and the poverty line is drawn where a family has less than 50 percent of this medium income available. Such seemingly 'exact' definitions are not possible before general income statistics were generated by statistical offices in the twentieth century (with forerunners in some of the most developed countries in the nineteenth century, and even earlier in some cities).

Some researchers have sought to define an absolute level of poverty by drawing the borderline at the minimum input of calories which a person needs to survive. But even this definition cannot be used as a universal measure since this minimum differs with age, sex, climate, the individual body, health, and the kind of work a person does. Moreover, poverty certainly begins before a person dies of hunger. Therefore, 'poverty can be defined objectively and applied consistently only in terms of the concept of relative deprivation' (Townsend 1979, p. 31). But what is relative deprivation?

2. The Size of Poverty

If we were to take present standards of living in advanced societies as a yardstick, most people, probably more than 90 percent, in all cultures and ages, would have to be defined as paupers. Only rulers, their

immediate entourage, and some large landholders and merchants were free of poverty. Most people working in handicrafts or on the land whom we would term today as the middle classes lived at least at the borderline of poverty. When the harvest failed, epidemics arrived, or war devastated a region—and this occurred frequently even in Europe up to the seventeenth and eighteenth centuries—they could be driven into misery. In other regions of the world, where periods of drought or torrential rains recur regularly, it is even today difficult for the majority of people to work themselves out of poverty.

For Western Europe we have sporadic estimates of the size of poverty for single cities and smaller regions since the later middle ages, and by Gregory King for a whole country, England, in 1688. They vary considerably, partly because the criteria used often remained vague and depended on the prejudice and the purpose of the author or institution taking the numbers. In some relatively well-to-do cities like Toledo, Lyons, Verona, Augsburg, Amsterdam, and Norwich in the sixteenth and seventeenth centuries, the poor numbered between 5 and 10 percent of the population in 'normal' times, rising to between 20 and 30 percent in times of wars, epidemics, or harvest failures. Similar figures have been found for some rural hinterlands of such cities, e.g., around Nuremberg or Zurich. (Jütte 1994, p. 53).

According to Gregory King, however, paupers constituted the majority of the population of England at the end of the seventeenth century. While half a million families (comprising about 2.7 million persons) increased the wealth of the country, about 850,000 families (comprising more than 2.8 million persons) decreased it by depending on others for their livelihood (Mathias 1983, p. 24). Other contemporaries give, however, different figures.

In eighteenth century France estimates of the size of poverty vary according to the definition between one-third and half of the population. When, during the French Revolution, the definition of poverty was sharpened by excluding beggars, vagabonds, and others, 'unwilling to work' the Comité de Mendicité stated that 39 percent of the working men were not able to support a family of five persons (Hufton 1974, p. 22).

One can distinguish at least three different measures used by officials in Western Europe in early modern times to count their paupers. The first was to count the recipients of welfare in one way or the other. This depended, of course, on the welfare policy applied at a particular location. Known figures vary from 2 percent for the city of Berlin in 1665 (but 7.2 percent in 1799) to 14.8 percent in Trier in 1623. But it is also known that in nearly all cities, the majority of people who applied for welfare were refuted. (In Amsterdam in 1799 37,500 received help, while 81,080 were refuted.) The second measure were censuses of the poor. These were rare, and the few available figures vary between

9.5 percent in Valladolid, Spain, in 1561 and 20–25 percent in Huddersfield, England, in 1622. The third method is to count the ‘fiscal poor’: the people (mostly heads of households) exempted from taxes because of poverty. Here we find a much higher level. The lowest available figure is 4.7 percent for Cologne in 1582 while the highest was 77 percent for Verona in 1635. Most other cities record shares of the tax-exempt households between 35 and 50 percent (Jütte 1994, pp. 46–56).

Taking into account that most of these figures stem from cities in the wealthier parts of Europe, it is easy to imagine the degree of poverty in other regions, particular the notoriously poorer mountainous regions, or the countryside in Eastern Europe, not to speak of the less developed continents.

3. *Who Are the Paupers?*

Certain groups of paupers can be found in all cultures and centuries, for instance, the bodily or mentally handicapped who are not able to work, and father- or motherless children (except those who live in extended families). Very often mothers with fatherless children are being reported amongst the poor. Finally, elderly people, particularly widows, frequently fell into poverty, at least before old age pensions were introduced in the late nineteenth and twentieth centuries. Even then they rarely received pensions high enough to support them. These people can be called ‘structural poor.’ They probably formed the 5–10 percent of paupers registered even in well-to-do cities and regions. In addition are those who have become temporarily unable to work because of sickness or injury, particularly during wars or natural disasters. One could term them ‘cyclical poor.’ Their numbers could swell easily to more than half of the population if events like epidemics, crop failure, or war devastated a region. At such times usually all those who lived on the margin of subsistence—until very recently the majority of the population—fell into deprivation.

All these groups have been seen since the Middle Ages by both worldly and religious authorities as ‘deserving poor,’ to help whom was a task for church, local or central governments, and private charity. They are to be distinguished from the ‘undeserving poor’: able-bodied idlers, beggars, vagrants, vagabonds, and prostitutes, who had to be educated or penalized, being expelled or—since the sixteenth century in Europe—put into workhouses or prisons (often one and the same).

For modern observers the biggest group are, however, the ‘working poor’: small peasants or handicraftsmen day laborers in cities as well as in the countryside, and the unemployed. Nor should slaves be forgotten. Though not all slaves in ancient Europe or seventeenth to nineteenth century America were

poor—some had learned crafts or worked as privileged servants, even as artists or managers—the great majority certainly were, and slavery existed also in other cultures and in some European countries as late as the seventeenth and eighteenth centuries.

4. *The Causes of Poverty*

For analytical purposes one can distinguish between different causes of poverty but one has to be aware that in reality they often were intertwined. Jütte (1994) distinguishes between accidental, cyclical, and structural causes. Accidental causes are, for instance, sickness, which counted according to the English Poor Law Commission in 1909 for ‘at least one-half of the total cost of pauperism’ (p. 21). In early modern Europe plague, influenza, typhus, typhoid fever, or smallpox did not only kill large segments of the population but also impoverished many of those who had survived. Warfare is another ‘accidental’ cause. Bad harvests, as well as periods of accelerated population growth and structural changes in the economy, Jütte counts as ‘cyclical,’ while he sees ‘structural’ reasons mainly in demographic patterns: persistent large families with many children and a non-secured old age. One would have to add seasonal causes to this typology. In northern countries each winter produced unemployment and poverty.

Another analytic scheme could be to distinguish between natural, demographic, sociopolitical, and economic causes of poverty. In all pre-industrial societies man was subject to natural conditions and disasters: drought or flooding, fire, unbearable heat in summer, or ice-cold storms in the winter time could stop his activities, and the natural fertility of his land limited his productivity. Richer or poorer soil, the availability of transport routes on rivers, lakes, and oceans, the availability of raw materials, like wood, copper, or iron ores determined the results of his labor. Demographic factors played an important role—as long as productivity could not be raised markedly, population growth could lead to increased poverty. Sociopolitical factors influenced the degree of and the distribution of wealth and poverty in all cultures, societies, and ages. Centralized or decentralized power, the strictness or looseness of social and political hierarchies, the degree and distribution of freedom, serfdom, or slavery amongst the population influenced also the living standards. Where geographic mobility was restricted, as in the case of slaves and serfs, people could rarely evade poverty by moving to better places. Where upward social mobility was restricted even hard work could help little to improve the lot of the lower strata of society. Here India with her caste system is probably the worst case, but Chinese and Eastern European peasants also fared worse than Western European ones, or Europeans in their new

settlements overseas. Economic conditions are closely connected to the political and social order. Where market conditions could develop a growing minority at least had the chance to escape poverty. But economic development did not occur harmoniously and evenly. The rise of one sector or region could cause impoverishment in another. When Dutch and English woolen products flourished in northwestern Europe in the sixteenth and seventeenth centuries, Italian cities suffered from the new competition. When Indian cotton products appeared on the European markets, the European woolen and linen producers had a difficult time, and when the British and Western European textile industries mechanized their production in the late eighteenth and nineteenth centuries Indian calico producers lost out.

Since the Industrial Revolution numerous such displacements took place and led to regional and branch- or industry-wide pauperization. Most, but not all, were temporary and could be stopped by adjustments to new technologies or other activities. Social and geographical mobility helped to absorb such shocks. There was, however, no overall impoverishment connected with industrialization. On the contrary, rather, the growth of productivity connected with mechanization and rationalization improved even the lot of the poorer segments of the population in the long run. The social problems at the beginning were caused rather by the population growth, which accelerated noticeably from the later eighteenth century. Without industrialization and the growth in productivity in agriculture too, Malthus' law that population grows faster than productivity would not have become obsolete.

There has been a long discussion amongst economic historians of Britain about whether the standard of living of the working classes during the Industrial Revolution improved or deteriorated. There is agreement that it improved from the 1830s when productivity grew recognizably and regularly and the growth of industry created more better-paid jobs. The pessimistic school maintained, however, that between about 1780 and 1830 the living standard of the workers deteriorated. This discussion suffers from a lack of adequate statistics, particularly of consumer prices and regional differences. As always when structural changes occur, there were winners and losers. The winners were skilled laborers and those who could move to better-paid jobs in growing regions; the losers were the immobile, unskilled people in areas not enjoying growth in industry, trade, and services (Fischer and Bajor 1967, and an ongoing discussion in the *Economic History Review*).

The discussion was first aired amongst British social reformers and Members of Parliament in the 1820s and 1830s, culminating in F. Engels' famous book on the situation of the working classes in England (1845) in which he painted an unrealistic picture of the idyllic life of the workers in the putting-out industry and

contrasted it with that of the industrial workforce. It was also influenced by the last grave subsistence crisis which occurred in parts of Western Europe during the second half of the 1840s. Ireland in particular, but also regions on the continent like Silesia, was hit by bad harvests including a nearly total loss of potatoes, the poor man's main food. Millions died or emigrated. The not-yet-mechanized linen industry could not compete with mechanized cotton industry. Pauperism seemed on the rise again, but this was a cyclical rather than a structural phenomenon. Since the middle of the nineteenth century living standards of workers improved in most countries of Western Europe and North America without, however, erasing poverty. In many countries serious empirical work on the size and causes of poverty began only now, culminating (amongst others) in the seventeen volumes Charles Booth published between 1889 and 1897 about *Life and Labour of the People of London*.

5. Welfare Policies

There is agreement amongst students of poverty that (modern) welfare policy began only in the nineteenth century, with the onset of social security schemes. But to care for the poor was seen as a moral obligation in all major religions and it was institutionalized at very early stages of history. Compassion was, however, only one reason for this: the other was fear, for paupers were regarded as dangerous. This twin motivation runs right through the centuries. The second motive, fear, seems to have dominated public policy in the Roman Empire, but early Christians insisted that *caritas* was essential. The Early Fathers of the Church, like St. Augustine, preached that poverty was a virtue and some founders of Christian orders, as in the case of St. Francis of Assisi, obliged their brethren to adhere to it. The Church and its monasteries were thus the main bodies responding to the fate of the poor during most of the Middle Ages, though not the only ones. Self-help played a role not only amongst families and neighborhoods but also in the craftsmen's guilds, which cared for their sick people and sometimes for the widows and elderly masters. Brotherhoods existed not only for journeymen but sometimes also for unskilled laborers, and in the later Middle Ages in some cities such as Frankfurt, Strasbourg, or Zurich even for 'the blind and the lame' (Fischer 1982, p. 29). In other countries, such as Spain, confraternities which enrolled men and, in some cases, women from different occupations, offered mutual aid alongside spiritual welfare to their members. Christian belief convinced wealthier citizens to secure their salvation by giving alms, establishing foundations for the poor or leaving parts of their fortunes to institutions such as hospitals.

Hospital pauperum already constituted small centers of welfare in the early Middle Ages. Often they cared for twelve persons, symbolizing the twelve apostles. In

the twelfth and thirteenth centuries many such hospitals were founded, not only by bishops and monasteries but also by kings, princes, merchants, guilds, and cities. About 600 of such hospitals existed in England in the middle of the fourteenth century. Some of them specialized, for example, in the care of children or of old people, or of those with certain maladies like leprosy. Often they supplied only minimal help, not even food was necessarily provided, and their inhabitants received permission to beg. Medical care was an exception since very few physicians were available. Most of the hospitals in the smaller towns remained all-round institutions for helpless people, but a few housed several hundred people.

In the later Middle Ages cities took over some such hospitals or founded new ones. They installed poor relief administrators, who began investigations of individual cases of poverty and made surveys for their towns or districts. Poor relief was to be made more efficient and rational; pure almsgiving was no longer enough. The 'deserving poor' were to be separated from the 'undeserving' ones. Orderly procedures were being introduced in administering poor relief. This was particularly marked in northern and central Europe while in France, Spain, and Italy worldly authorities remained more reluctant to take poor relief out of the hands of the Church.

This rationalization and bureaucratization began well before the Reformation, and afterwards Catholic and Protestant regimes did not necessarily follow different policies—both, for instance, tried to educate the poor, particularly the children, though with meager results. But there were differences as to the centralization or decentralization of such policies. Mainly Protestant cities centralized their institutions into a 'common chest' to exercise better control. In the German Empire the Imperial Diet of 1497 declared that each community should take measures to care for its own poor, and this principle was enacted in the Imperial Police Ordinance of 1530 which remained in force until the end of the Empire in 1806. Switzerland, The Netherlands, Britain, and the Scandinavian countries followed similar paths, while in the Mediterranean countries the many decentralized institutions remained more or less undisturbed.

While Britain is seen as the pioneer of a general poor tax in 1572, the municipality of Paris had introduced such a tax already in 1551 after the French King had given it the authority 'to supervise and conduct all things required for the treatment and the relief of the poor' (Jütte 1994, p. 119). But Britain pioneered early modern poor regulations in several respects. Already in the middle of the fourteenth century Parliament ordered that the working poor should be distinguished from beggars, and in 1388 tried to restrict the mobility of beggars. From the 1530s to around 1600 numerous statutes were enacted, of which those of 1572 and 1596 became known as the 'Old Poor Laws,' which remained in force until 1834. The Judges of the Peace

had to select four men of substance to act as overseers of the poor. These men had to raise the poor tax, distribute help, and take care of the education of orphans. They were supposed to punish beggars and vagabonds, and could establish houses for correction. As in other countries, the result of their endeavors was very uneven. The poor taxes always fell short of the needs. It has been estimated that only 7 percent of the expenditure for the poor originated from taxes—by far the greater part came from private donations and charitable trusts which were established by wealthy citizens. One rich Londoner alone set up institutions for the poor in 216 communities (Fischer 1982, p. 43f.). Such trusts also ran hospitals and schools.

In the minds of the early modern poor administrators and the wider public, education to work became more important than poor relief. Therefore, workhouses were established in the major cities and some smaller towns. The hope was that they would both accustom the poor to work and produce some profit. Usually they failed in both respects. Particularly in The Netherlands, the Scandinavian countries, and in Germany they became rather a combination of workhouse and house of correction, the famous example being the Amsterdam *tuchthuis* founded in 1595. In smaller towns such houses often tried to perform four functions: educating orphans, getting adults to work, penalizing beggars and criminals, and taking care of handicapped or sick people.

The English Poor Law of 1834 was supposed to remedy the ages-old deficiencies in welfare policies. After a Parliamentary Commission had found that the 'Old Poor Law,' by prescribing assistance to every needy person, had actually provided a disincentive to work, the new law stipulated that only such paupers who were willing to live (and work) in poorhouses should get relief. It turned out that there were not enough such institutions to house everybody who did not find work on his own or did not earn enough to survive, particularly with a larger family. The new law did not remove the ancient causes of poverty, neither could it take care of the temporary unemployment of industrial workers, but it revived the discussion about social problems in the country. A whole set of measures had to be taken to improve the lot of the poorer classes: while more employment was to be created by industry, transport business, and services. Parliament regulated child and female labor, and began to enforce safety measures at the workplace, while local communities improved the housing and sanitary conditions. In 1876 about 19 percent of the population were insured by voluntary Friendly Societies, relatively more than in Germany's obligatory health insurance in 1885, but the most needy could not afford the contributions. Private charity remained, therefore, important in Britain. In 1897 Britain introduced obligatory accident insurance, in 1908 old age pensions, and in 1911 a combined health- and unemployment insurance.

Other European countries were faster in creating obligatory accident-, health-, and old age insurance. But unemployment, one of the most important reasons for poverty, was being regarded as non-insurable even at the end of the nineteenth century. In Germany, unemployment insurance came into force only in 1927. Most of these insurances were quite inadequate at the beginning. They covered industrial workers only in some branches, not day-, farm-, and home-workers, thus excluding the poorest part of the labor force. Bismarck's Old Age Insurance provided means only at the age of seventy when most people had died already, and widows received only 20 percent of the earning of their husbands in accident insurance. But the principles of a modern welfare state were established, and the basic institutions for it were created before or shortly after World War I.

See also: Income Distribution; Inequality; Inequality: Comparative Aspects; Poverty, Culture of; Poverty: Measurement and Analysis; Poverty Policy; Poverty, Sociology of; Social Inequality in History (Stratification and Classes); Social Security; Social Stratification; Social Welfare Policy: Comparisons; Wealth Distribution; Welfare; Welfare Programs, Economics of; Welfare State; Welfare State, History of

Bibliography

- Abel W 1974 *Massenarmut und Hungerkrisen im vorindustriellen Europa. Versuch einer Synopsis* [Mass Poverty and Subsistence Crisis in Pre-industrial Europe]. Paul Parcy, Hamburg and Berlin
- Aldcroft D H, Catterall R (eds.) 1996 *Rich Nations—Poor Nations. The Long Run Perspective*. Edward Elgar, Cheltenham, UK
- Fischer W 1982 *Armut in der Geschichte. Erscheinungsformen und Lösungsversuche der 'Sozialen Frage' in Europa seit dem Mittelalter* [Poverty in History]. Vandenhoeck and Ruprecht, Göttingen, Germany
- Fischer W (ed.) 1995 *Lebensstandard und Wirtschaftssysteme* [Living Standards and Economic Systems]. Fritz Knapp, Frankfurt, Germany
- Fischer W, Bajor G (eds.) 1967 *Die Soziale Frage. Neuere Studien zur Lage der Fabrikarbeiter in den Frühphasen der Industrialisierung* [The Social Question, Recent Studies about the Situation of Industrial Workers in the Early Phases of Industrialization]. Koehler, Stuttgart, Germany
- Fischer W, van Houtte J A, Kellenbenz H, Mieck I, Vittinghoff F (eds.) 1980–1993 *Handbuch der Europäischen Wirtschafts- und Sozialgeschichte* [Handbook of European Economic and Social History]. Klett-Cotta, Stuttgart, Germany, 6 Vols
- Flynn M 1989 *Sacred Charity: Confraternities and Social Welfare in Spain, 1400–1700*. London.
- Geremek B 1988 *Geschichte der Armut* [History of Poverty]. Artemis, Zürich, Switzerland
- Gutton J P 1974 *La Société et les pauvres en Europe, 16e–18e siècle* [Society and the Poor in Europe, 16th–18th Centuries]. Presses universitaires, Paris
- Himmelfarb G 1984 *The Idea of Poverty: England in the Industrial Age* London
- Hippel W von 1995 *Armut, Unterschichten, Randgruppen in der frühen Neuzeit* [Poverty, Lower Strata, Marginal Groups in Early Modern Times]. Oldenbourg, Munich, Germany
- Huften O W 1974 *The Poor of Eighteenth-Century France 1750–1789*. Clarendon, Oxford, UK
- Illife J 1987 *The African Poor. A History*. Cambridge University Press, Cambridge, UK
- Jütte R 1994 *Poverty and Deviance in Early Modern Europe*. Cambridge University Press, UK
- Landes D S 1998 *The Wealth and Poverty of Nations. Why Some Are So Rich and Some So Poor*. Norton, New York
- Lindenmeyr A 1996 *Poverty is Not a Vice: Charity, Society and the State in Imperial Russia*. Princeton University Press, Princeton, NJ
- Lis C, Soly H 1979 *Poverty and Capitalism in Pre-industrial Europe*. Harvester, Hassocks, UK
- Mathias P 1983 *The First Industrial Nation. An Economic History of Britain 1700–1914*, 2nd edn. Methuen, London
- Mollat M (ed.) 1974 *Etudes sur l'histoire de la pauvreté (Moyen Age–XVI siècle)* [Studies in The History of Poverty: Middle Ages to 16th Century]. Sorbonne, Paris, 2 Vols
- Pullan B 1994 *Poverty and Charity, 1400–1700. Europe, Italy and Venice*. Variorum, Aldershot, UK
- Riis T (ed.) 1981–1990 *Aspects of Poverty in Early Modern Europe*. Klett-Cotta, Stuttgart, Germany, 3 Vols
- Rimlinger G 1971 *Welfare Policy and Industrialisation in Europe, America and Russia*. Wiley, New York
- Salama P 1994 *Pauvreté et inégalités dans le tiers monde* [Poverty and Inequality in the Third World]. Paris
- Thorp R 1998 *Progress, Poverty and Exclusion: An Economic History of Latin America in the 20th Century ...* Washington, DC
- Tomaskovic-Devey D 1988 *Poverty and Social Welfare in the United States*. University of Colorado Press, Boulder, CO
- Townsend P 1979 *Poverty in the United Kingdom. A Survey of Household Resources and Standards of Living*. Penguin, Harmondsworth, UK
- Williamson J G 1991 *Inequality, Poverty and History*. Blackwell, Cambridge, UK
- Woolf S 1986 *The Poor in Western Europe in the Eighteenth and Nineteenth Centuries*. Methuen, London

W. Fischer

Poverty Law

'Poverty law' refers to two distinct but related topics. One has to do with the welfare state, or more particularly, that part of the welfare state that seeks to address the needs of the poor. In the United States, as is true of all advanced industrial economies, many programs fit this definition. In this article, the definition will be restricted to income support for single mothers and their children—the program formerly called 'AFDC' (Aid to Families with Dependent Children) now replaced by 'TANF' (Temporary As-

sistance to Needy Families). Although not the largest program, by far, it is the most controversial. This is the program that is referred to as 'welfare,' which has become the code word for the stereotype inner-city, African-American woman, often a substance-abuser, having children to stay on welfare, and breeding a criminal class. 'Welfare' is the word used to describe the 'underclass' in America.

The second and related meaning of 'poverty law' refers to the legal rights revolution that blossomed in the 1960s that focused on welfare and other poverty issues. Although Americans have always resorted to the judicial system to try to further political goals, the modern era can probably be dated from the court victories of the National Association for the Advancement of Colored People (NAACP) and its offshoot, the Legal Defense Fund (LDF, also known as the 'Inc. Fund'). The LDF victory in *Brown vs. Board of Education* (1954), where the US Supreme Court declared school desegregation unconstitutional, ushered in an incredibly active period of law reform litigation brought by a wide variety of legal service, public interest, and private lawyers. In addition to civil rights (for women, other racial and ethnic minorities), consumers, and environmental protection, rights were won for the mentally ill, the disabled, children with learning disabilities, health care patients, criminal defendants, prisoners, gays and lesbians, and so forth. The particular concern of this article is with lawyers who worked on behalf of the poor.

The two parts of 'poverty law'—'welfare' and poverty lawyers—will be discussed together, because their developments were intertwined.

1. Aid to Single Mothers Prior to the 1950s and 1960s

Programs to provide income support for poor single mothers were enacted by the states between 1910–20. They were part of the Progressive Era Child Saving Movement. Social reformers, alarmed at the number of immigrant children growing up in the urban slums, expanded the definition of child neglect and removed children from what they considered to be baleful living conditions and sent them to either farms in the mid-West or reformatories. Juvenile courts were established at the turn of the century with jurisdiction over 'delinquent, dependent, and neglected' children. The latter two were considered 'pre-delinquent.' Then, the Child Savers decided that if the home was just poor, but otherwise 'fit and proper,' perhaps the mother could be supported with a grant as an alternative to removing the children. The programs were called 'Aid to Dependent Children,' (ADC) although popularly known as 'Mothers' Pensions.' In most jurisdictions, they were administered by the juvenile or county courts, assisted by local charities, to make sure that the homes and the mothers were 'fit and proper.'

This early history set the pattern for the next 40 years. The programs were locally administered, small, and basically restricted to white widows. Other single mothers—divorced, separated, never married, and of color—were not considered 'fit and proper.' The consequences of exclusion were severe. Those who were excluded were lumped together with the rest of the 'undeserving poor'—childless adults, the long-term unemployed, others who were excluded from other assistance programs. The 'undeserving poor,' including most single mothers, were thus required to support themselves as best they could in the paid labor market. The establishment of the American welfare state, in 1935, changed the pattern for single mothers only slightly. ADC became a grant-in-aid. The Federal Government assumed about half of the costs and set a few eligibility and procedural conditions, but substantive and administrative control remained with the states and the counties.

AFDC (Aid to Families with Dependent Children, 1962) changed radically starting the late 1950s and early 1960s. The previous system of exclusion collapsed and the rolls and costs exploded. In 1950, two million people were on the rolls; in 1960, three million; then 13 million by the end of the 1980s. Costs rose from about \$500 million in 1950 to about \$23 billion. Most dramatically, however, was the change in the demographic characteristics of the rolls. In streamed the previously excluded—divorced, single, never married, with increasing proportions of nonwhites. There were various reasons for the changes—the massive migration of African Americans out of the rural south to the northern cities; the entry of women into the paid labor force; the severe unemployment in the inner cities, with the accompanying increase in poverty; rising social unrest, disorder, crime, and riots (during the early 1960s, there were riots in 24 cities); the pressure exerted on local governments by the Federal Government's War on Poverty; the welfare rights movement, and the legal rights movement; the courting of innercity voters by the Democratic Party; and the rise of single mother families. This was the era of the Great Society. A variety of federally sponsored local initiatives and activist groups attacked local institutions and practices, including welfare. The federal courts, War on Poverty lawyers, and welfare rights organizations, such as the National Welfare Rights Organization (NWRO), forced open the state AFDC gates. Welfare became a 'right'; social movement groups actively recruited potential eligibles, reaching more than 90 percent in the early 1970s. Welfare was now in 'crisis.'

2. The War on Poverty; Poverty Lawyers

The first legal aid society was established in 1876 in New York City to ease the transition of German immigrants. Thereafter, with the prodding of the

American Bar Association, organizations spread serving the various needs of the poor. By the early 1960s, there were 249 offices. However, the individual offices were small and underfunded. The caseload was mostly landlord-tenant, consumer, and family matters, but few divorces, adoption, bankruptcy, and juvenile proceedings. Legal aid offices did not want to get involved in morally controversial issues. The work was individual service cases. There was no attempt at law reform, test case litigation. The first public defender system was established in 1910. The public defender system grew much more slowly than legal aid. In contrast to legal aid which was based on private charity, the public defender depended on public funding.

The big change in legal services for poor came with Kennedy Administration (1961–1963). Legal rights for the unrepresented was an integral part of the New Frontier and then President Johnson's War on Poverty. The overarching strategy was 'community action,' whereby the poor would help themselves (with lots of government support). If the poor were to be empowered, they would need legal help. There was a struggle, however, over the model and control of legal services. Local bar associations pressed for an expansion of their established individual service offices. The alternative model came from the more aggressive, test-case, law-reform activities, primarily, but not exclusively the NAACP and particularly, the LDF. It would be hard to overestimate the influence of the LDF class action, law-reform strategy during this period. Aided by an activist federal judiciary, at the stroke of the judicial pen, so it seemed, legal rights and legitimacy were given to disadvantaged groups. This prospect of significant social change via the legal system had an enormous appeal to activist, young, elite lawyers. The US Supreme Court, led by Chief Justice Earl Warren, and the NAACP litigation seemed the perfect example of what law and lawyers could do. This was a period of enormous social activism, particularly with civil rights. Although there were the usual compromises with the bar associations as to the organization and funding of legal services, the aggressive, law-reform approach became the most important influence in the development of the War on Poverty's Office of Economic Opportunity (OEO) Legal Services program (1965), as well as in foundation support for consumer, environmental, and public interest law. As we shall see, it also became the focus of the conservative attack on Legal Services.

The specific model for Legal Services was the Ford Foundation's Grey Areas program, where community-based legal services was visualized as an integral component of the tools necessary for the militant self-assertion of the poor and the disenfranchised. Legal services were offered for a range of activities—education, employment, welfare, health and mental health, individual service, and community development. The dominant approach was using test

case litigation to establish new rights and the rule of law in dealing with government administrative agencies. The most important legal needs of the poor were thought to be those that concerned their relations with public service programs such as welfare and housing.

While many ideas and organizations contributed to the development of Legal Services, three stand out. One was the influential article by Charles Reich (1964), then a professor at Yale Law School, who argued that relationships with government should receive the same status and protection as traditional property rights. He coined the term, the 'new property.' The other was the articles published by Jean and Edgar Cahn (1964, 1970) which presented the case for neighborhood, activist legal services offices. The third was the support for Legal Services by the American Bar Association which became crucial not only in the contests with local bar associations, but also against the rising conservatism specifically directed at Legal Services.

By 1966, OEO was giving US \$20 million to 130 local legal services organizations throughout America. The local offices, in turn, were supported by a half dozen backup centers which provided research support for the litigation activities. From the mid-1960s to mid-1970s, Legal Service lawyers and their allies won a significant number of cases in a variety of areas. The US Supreme Court held that welfare could not be denied simply because the mother cohabited with an unrelated male (the 'man-in-the-house' rule); that states could not withhold welfare or provide lower benefits to the poor who migrated from other states. The most significant welfare case was *Goldberg vs. Kelly* (1970) which held that welfare was a statutory entitlement and that before welfare could be terminated, the client had a due process right to a trial-type hearing at the administrative level with the right of judicial review. During this period store-front offices would advise welfare clients of their rights, both statutory and administrative. Aided by lawyers, clients were emboldened to confront the line staff, and, in certain parts of the country, organize mass demonstrations and sit-ins at local offices. Significant victories were won in both federal and state courts—for example, persons were entitled to a due process hearing to contest the repossession of property or eviction from public housing; the mentally disabled had the right to be represented by an attorney in commitment hearings and if confined to an institution, a 'right to treatment.' In addition, significant legislation was also passed—e.g. the Voting Rights Act, the Americans with Disabilities Act, the Education for All Handicapped Children Act, which, among other things, sought to guarantee parent participation in school special education decisions.

Goldberg vs. Kelly was the high-water mark. Legal Services lawyers tried to establish that the right to a basic income was a fundamental right that deserved the same protection as other fundamental rights, such

as speech, religion, and the right to be free from racial discrimination. But the Supreme Court almost immediately took back what it gave in *Goldberg*. The Court held that welfare regulation was similar to other economic and social matters and could be regulated according to a less exacting standard of reasonableness. Subsequent cases applied the 'reasonable regulation' standard in upholding maximum grants regardless of family size, percentage reductions of AFDC grants as compared to grants for the aged, blind, and disabled, lower amounts of school funding in poorer districts, and unannounced caseworker visits as not violating the Constitutional prohibition of unreasonable searches and seizure.

By this time, the reaction to the activism of the 1960s began to take hold. Not surprisingly, Legal Services' successes against government agencies aroused the hostility of state and local political leaders. Legal Services earned the undying enmity of Ronald Reagan, as governor of California (1967–69) when one of the offices successfully blocked his attempts to cut back on welfare and health care for the poor. From time to time, governors attempted to prevent Legal Services offices from bringing suits against government agencies or on behalf of particular clients (e.g., migrant workers).

3. *Change in Welfare*

Reflecting the liberal moment, the first response of the Kennedy Administration to the 'welfare crisis' was social services to strengthen families and help them towards self-support. The 'social service amendments,' however, never provided much money or services. Rolls continued to grow, and the social service approach was abandoned in 1967. Thereafter, the conservative approach took over. The perception was that AFDC, especially at the local level, was out of control. The program was overly complex, administratively impossible, and hence, subject to rampant 'waste, fraud, and abuse.' The first reaction of the conservatives was a major overhaul of administration. Many of the complex rules were abolished (most notably, those providing for special grants to meet special needs).

Caseworkers were downgraded to eligibility technicians (often with little more than a high school education) whose main task was to monitor rules and reduce errors. The Federal Government imposed strict error-rate standards on the states, and the states, in turn, imposed strict standards on local county administration. The various state and local programs became increasingly rationalized and computerized, all designed to check eligibility and financial payments. Grants to families were either reduced or, more commonly, not allowed to keep pace with inflation. Average grants declined by 45 percent between 1970 and 1996.

The second response was to return to first principles. Those previously excluded from AFDC were required to work in the paid labor market; they were part of the 'undeserving' poor. The conservatives argued that they did not lose their negative moral ascription merely because misguided liberals allowed them into the AFDC program. Beginning in 1967, the first federally-imposed work test for able-bodied AFDC recipients was enacted (WIN, the Work-Incentive Program). Liberals vigorously fought WIN, arguing that if nonwelfare mothers were not required to work, then it was unfair and punitive to require welfare mothers to work. The conservatives said that these historically undeserving women were morally different from mothers who were either man-dependent or self-sufficient. Over the next 20 years, there was a standoff. The work programs, both state and federal, never succeeded in the sense of either reducing rolls and costs through employment, or setting the poor to work, or imposing sanctions for failure to comply. It was administratively easier and less costly to simply defer the vast majority of work-eligible welfare recipients.

Both sides remained unhappy. The conservatives continued to attack the 'entitlement state' on the grounds that there are responsibilities as well as rights in the social contract. Then, in the late 1980s, the liberals changed. They now agreed that AFDC mothers should be expected to work. They gave two reasons. Since the majority of nonwelfare mothers were in the paid labor force, it is only reasonable to expect welfare mothers to work. Second, families are better off, both materially and socially, when the adults are gainfully employed. The Family Support Act of 1988 attempted to strengthen the work requirements. In the meantime, states began to experiment with welfare-to-work programs under waiver from the Federal Government.

The welfare 'crisis' continued to boil. President (then candidate) Clinton promised to 'end welfare as we know it.' The 1994 mid-term elections produced a strongly conservative Republican Congress, which enacted the Personal Responsibility and Work Opportunity Act of 1996. AFDC was abolished, replaced with Temporary Assistance to Needy Families (TANF). State authority over welfare, already considerable, was increased by converting the federal entitlement into block grants. The work requirements were strengthened significantly, now enforced by two sets of time limits: recipients can no longer receive aid for more than two consecutive years (states have the option of shorter limits); and there is a cumulative 5-year lifetime limit. States are required to move an increasing percentage of recipients into work or work-related activities; states are required to reduce grants for recipients who refuse to engage in work or work activities. In addition, there are many 'family values' conditions—for example, requiring mothers under 18 to live with their parents or in adult-supervised

settings; requiring cooperation in the establishment of paternity; states can refuse to pay for children conceived while the mother was on welfare; or deny cash assistance for life to persons convicted of a felony (which, in most states, could be possession of a small amount of marijuana). The Act also modifies several other programs which can impact on the wellbeing of the poor—health care, food stamps, disability, and so forth.

The consensus behind the work provisions (as well as the family values) is that welfare encourages dependency with all its attendant ills for both the adults and the children. Tough work requirements, enforced by time limits, will not only reduce welfare costs, but, more importantly, replace the entitlement status and permissiveness of the current system with values of responsibility and self-sufficiency. TANF explicitly rejects education and training in favor of the ‘work first’ strategy. The argument for this strategy is: (a) there are plenty of jobs for those who want to work; (b) by taking any job, even an entry-level job, and sticking with that job, a person will move up the employment ladder; (c) the problem with welfare recipients is that they do not have the motivation or the incentives to leave welfare and enter the paid labor market; and (d) state programs have shown that recipients can be moved from welfare to work.

Many states began implementing their programs between 1993 and 1995. In the meantime, starting in 1994, AFDC, nationwide, began a sharp decline—from 5 million families to 2.8 million, in 1998. State and Federal politicians, of course, are claiming ‘success’ for the welfare reforms. Aside from the fact that the declines in the rolls started well before the initiation of the welfare-to-work programs, the evidence from the state programs is that comparatively few recipients entered the labor market as a result of the programs and they earned very little more than the controls. For example, in the most ‘successful’ program—the standard-bearer for the ‘work first’ strategy—Riverside, California—at the end of the demonstration, two-thirds of the experimentals were not working and over half never worked during the entire 3-year period.

The reasons for the lack of success of the state programs are that the assumptions behind the ‘work first’ strategy are totally misconceived as to who welfare recipients are, why they are on welfare, and the characteristics of the low-wage labor market. Despite the impressive job growth, there has been a stagnation, if not decline, in the wages and benefits for the less skilled, less educated worker. Jobs are increasingly contingent or short-term. Most Americans may be better off, but the bottom fifth of the population is worse off; the average annual income of this group is US \$8800 (1999), down from US \$10,000 in 1997. Despite the stereotype, most welfare recipients are adults with small families (1.9 children on average), and are on welfare for relatively short periods—

between 2 and 4 years. They have extensive labor market connections and many combine welfare with work. But work is uncertain. Jobs disappear, child care breaks down, there are family and transportation problems. They return to welfare because they are usually ineligible for unemployment benefits (leaving a job for family reasons is a ‘voluntary quit’).

The welfare reforms do little, if anything, to meet these problems. Despite the job growth, there is a shortage of jobs for the low-skilled in many inner cities. Welfare recipients are in tough competition. Most lack a high school degree; they have young children; and they often have transportation problems. They have to rely on informal child care, which is often uncertain and still costs between a fifth and a third of monthly income. They do not have health insurance. And many have additional barriers—lack of skills, and networks, poor health, depression, abuse and violence. Although ex-recipients much prefer work to welfare, they are rarely better off financially and still remain in poverty. As the rolls decline, more of the hard-to-employ are left. In the meantime, sanctions are increasing and the two-year time limits on welfare are due. While there is very little data on those who have been turned away from welfare but are not in regular employment, in general, poverty and hardship is increasing for families, with potentially very serious impacts on the children. In 2000, nationwide more than a third of the homeless are families, an increase of more than ten percent since 1985. In the past, AFDC provided at least some financial stability.

4. *A Crippled Legal Services*

When President Reagan took office in 1981, he proposed no funding for Legal Services for seven successive budgets. The program survived, but funds were sharply cut, forcing significant reductions in staff. Consequently, groups of cases (e.g., bankruptcies, disability-benefits) had to be dropped by many offices. The Nixon Administration (1969–74) had also tried to abolish the program and distribute the funds to the states. When that plan failed, the compromise was to create an independent corporation, the Legal Services Corporation (LSC) (1974) and, for the first time, restrictions were imposed on the kinds of cases that Legal Services could handle—school desegregation, the draft (the Vietnam War), abortion, post-conviction civil challenges. The back-up centers were eliminated. On the other hand, the Corporation’s budget was increased dramatically, from US \$71 million in 1975 to US \$321.3 million in 1981. In that year, the LSC made annual grants to 323 programs which employed 6218 lawyers in 1425 offices providing legal assistance in over 1.2 million matters. That was the last increase. The 1994 Republican Congress made deep funding cuts (from US \$400 million to US \$278) as part of a three-year plan to abolish Legal Services altogether,

and added the most severe restrictions to date. Congress prohibited all class actions; eliminated the collection of attorneys' fees; prohibited participation in administrative rulemaking, lobbying, litigation on behalf of prisoners, representing drug-related public housing evictions, representing certain categories of immigrants, and challenging state, welfare-reform laws under any circumstances—even if such laws violate the US Constitution or federal laws.

The restrictions went into effect in August 1996, and Legal Services had to withdraw from more than 600 class-action lawsuits across the country. In 1997, Legal Service programs employed about 3500 attorneys. They were assisted by about 59,000 private attorneys, mostly on a pro bono basis. The program closed 1.2 million cases and provided services for 1.9 million clients. In 1998, the budget for Legal Services was actually US \$21 million less than in 1980. The American Bar Association has estimated that only about 20 percent of the legal needs of the poor are being met. Thus, at a time when there is growing inequality and poverty, when more than 40 million Americans lack health care, and when the poor are facing one of the harshest welfare reforms in recent history, they are also being denied effective legal assistance.

In 2001 the US Supreme Court (*Legal Services Corporation vs. Velazquez* (2001)), held that the restriction on legal services lawyers from challenging welfare reform statutes or regulations in the course of representing individual clients seeking welfare benefits was unconstitutional; perhaps a new chapter on representing the poor will open.

5. The Welfare State in Western Europe

In the US, the current welfare reform can be looked at as a continuation of American welfare policy. From the earliest days of the colonies, the central mission was to separate out the 'deserving' poor from the 'undeserving.' The latter were defined primarily in terms of the labor market—are they able-bodied, could they work? If so, then no welfare. The mission of welfare policy was to exclude, rather than include. Excluded were the deviants—the moral deviants—and this was carried out at the local level. In the USA there is no concept of 'social citizenship.'

Europe could not have been more different. In the decades following World War II, until the oil shocks of the 1970s, despite the differences between countries, the basic approach was universalism, inclusion. There were significant protections and supports for workers, and generous benefits for those who were not in the labor market. The basis for the European welfare state was full-employment prosperity, that is, the system was successful as long as most people were not drawing significant benefits, but rather, were working and contributing to the welfare state. It was this prosperity

that allowed the nationalist, peak, corporatism bargaining—concertation—between labor and government. Labor would agree to wage restraints in return for jobs and benefits.

By 2000, much had changed. The demographics of the welfare state have changed to an aging population, greatly increasing pension and health care costs. In addition, for most of the Continent, the economies have been sluggish with persistent long-term employment. Increasing proportions of the long-term unemployed are the vulnerable—youth, immigrants, lone mothers. The current dilemma is that with the high rates of persistent, long-term employment, demands are being made on the welfare state. But under the European Union monetary restrictions, neither tax increases nor budget deficits can be used to finance the increased demands.

Thus far, efforts to develop social protections at the European level have not been successful. There is now pretty much of a consensus that negative integration—the removal of trade barriers, the free flow of labor—is far different than positive integration which requires fundamental changes in government structures, decision making processes, claims to legitimacy—and, most important, solidarities across borders. In fact, the present differences among European welfare states are so significant, that most commentators believe that national institutions are too embedded to converge on a common approach.

When the conservatives were in office, the argument was made that the reason for the sluggish economies is that the generous welfare state has made labor too costly, thus preventing job growth. In the global economy, the foundations of the welfare state—a society of steady, well-paying jobs—are being replaced by major changes in both economic structures (low paying, part time work) and demographics (declining proportion of workers as compared to retirees), the disabled, an increasing proportion of women in the paid labor force, the rise in lone parent families, and immigrants. The foundations of the present welfare state—the intact family, male breadwinner—are no longer valid. The 'passive' solutions of unemployment—early retirement and long-term unemployment and disability—are no longer feasible. Pointing to the US and UK experience, conservative politicians and EU officials argue that labor has to become more 'flexible' and labor market protections have to become more 'active'—that is, encourage people to re-enter the labor market by shortening the period of unemployment benefits, requiring training, education, and employment, raising the retirement age, and tightening disability.

The voters rejected these arguments and the Social Democrats have replaced most conservative governments. While promising to protect the welfare state, they are acutely aware of the constraints on budget deficits or raising taxes. Some countries—Sweden, Denmark, and the Netherlands—have been able to

emerge from the recession in the early 1990s, maintain or re-establish concertation between major employers and labor, and protect their systems of social protection. In other countries, the Social Democrats are struggling. There have been some changes—the proportion of unemployed who qualify for benefits has been declining, there is more use of means-tested benefits, the reduction of benefits in some jobs, the abandonment of centralized wage bargaining in some countries. Most countries are trying to replace ‘passive’ benefits with an ‘active’ labor market policy. But so far, these changes have been incremental, and the basic welfare states are still intact. It has proven easier to change markets than social welfare states.

While there is still strong social support for basic welfare state programs, the effective, political resistance, it is argued, basically applies to the existing workforce, not the excluded workers and the vulnerable groups. In the meantime, even though transfers may not have been significantly reduced, we see the quiet undermining of the welfare state—the break-up of national collective bargaining, the growing disparity in wages, the spread of low-wage work, and hostility towards immigrants. The deterioration of the labor market will undermine the social welfare provisions. As pointed out, it was the combination of earnings and benefits that accounted for the comparative equality in Europe. In the final analysis, the foundation of families rests on the economic viability of the adults. Two things are happening in the labor market. As noted, large numbers of the vulnerable population are increasingly facing more-or-less permanent exclusion from the labor market. Others will be increasingly forced back into the labor market, but good, permanent jobs will not be there. An effective active labor policy requires both a large public investment and an integrated infrastructure of labor market institutions, a commitment to some basic income equality, together with a strong safety net. As a growing number of people become dependent, the political pressure increases to restrict income support for the able-bodied. Although there is an increase in need-based and conditional benefits, and there seems to be still strong support for retirees and for children, the crucial question is how solid is the support for the able-bodied? At the present time, welfare policy in Western Europe is at an impasse.

See also: Justice, Access to; Legal Representation of the Poor; Lone Mothers in Affluent Nations; Lone Mothers in Nations of the South; Poverty and Child Development; Poverty in History; Poverty: Measurement and Analysis; Poverty Policy; Poverty, Sociology of; Social Welfare Policies and Gender; Social Welfare Policy: Comparisons; Urban Poverty in Neighborhoods; Welfare; Welfare Programs, Economics of; Welfare State; Welfare State, History of

Bibliography

- Abramovitz M 1988 *Regulating the Lives of Women: Social Welfare Policy from Colonial Times to the Present*. South End, Boston
- Bell W 1965 *Aid to Dependent Children*. Columbia University Press, New York
- Davis M 1993 *Brutal Need: Lawyers and the Welfare Rights Movement 1960–1973*. Yale University Press, New Haven, CT
- Cahn E, Cahn J 1964 The war on poverty: A civilian perspective. *Yale Law Journal* 73: 1317–52
- Cahn E, Cahn J 1970 Power to the people or the profession? The public interest in public law. *Yale Law Journal* 79: 1005–48
- Ellwood D 1988 *Poor Support: Poverty in the American Family*. Basic Books, New York
- Esping-Andersen G 1999 *Social Foundations of Postindustrial Economics*
- Garfinkel I, McLanahan S 1986 *Single Mothers and their Children: A New American Dilemma*. Urban Institute Press, Washington, DC
- Gordon L 1994 *Pitied but not Entitled: Single Mothers and the History of Welfare, 1890–1935*. Free Press, New York
- Handler J, Hasenfeld Y 1997 *We the Poor People: Work, Poverty, and Welfare*. Yale University Press, New Haven, CT
- Handler J, Hollingsworth E, Erlanger H 1978 *Lawyers and the Pursuit of Legal Rights*. Academic Press, New York
- Katz M 1986 *In the Shadow of the Poorhouse: A Social History of Welfare in America*. Basic Books, New York
- Mink G (ed.) 1999 *Whose Welfare?* Cornell University Press, Ithaca, NY
- Patterson J 1981 *America's Struggle Against Poverty, 1900–1980*. Harvard University Press, Cambridge, MA
- Piven F, Cloward R 1977 *Poor People's Movements: Why they Succeed, How they Fail*. Pantheon, New York
- Reich C 1964 The New Property. *Yale Law Journal* 73: 733–87
- Skocpol T 1992 *Protecting Soldiers and Mothers: The Political Origins of Social Policy in the United States*. Harvard University Press, Cambridge, MA
- Trattner W 1989 *From Poor Law to Welfare State: A History of Social Welfare in America*, 6th edn. Free Press, New York

J. F. Handler

Poverty: Measurement and Analysis

The issue of low income and poverty has concerned social scientists for centuries. In 1776 Adam Smith linked economic poverty directly with the lack of ‘not only the commodities which are indispensably necessary for the support of life, but whatever the custom of the country renders it indecent for creditable people, even of the lowest order, to be without’ (Smith 1776, Book V, Chap. II, Pt. II, Article 4th). Interest in this problem was a natural accompaniment of efforts to understand why nations experience economic growth. If the economy of a nation grows, it is only natural to wonder how the fruits of this growth will be distributed among people and why some people are excluded from

the benefits of economic prosperity. A primary issue is to define poverty precisely, and then to document its extent consistent with this definition, that is, to count the poor. Having identified the poor population, questions regarding the nature, causes, and cures of poverty follow naturally. The process of measuring poverty and analyzing its causes and consequences has advanced social science research methods and techniques.

1. The Measurement of Poverty

The most prominent early effort to measure poverty was undertaken by B. S. Rowntree in 1902, in his attempt to assess the extent of poverty in York, England (Rowntree 1902). Since that time, there have been numerous attempts to measure poverty in a variety of countries or more narrowly defined areas.

Although reducing poverty is a nearly universal goal among both nations and scholars, there is no commonly accepted way of identifying who is poor. Some argue for a multidimensional poverty concept, such as being 'housing poor' or 'health poor' or lacking in social contacts, thought of as social deprivation. Economists tend to prefer a concept of hardship that reflects economic position or economic well-being, somehow measured; that concept has come to dominate the measurement of poverty.

The measurement of economic poverty seeks to identify those families whose economic position, defined in terms of *command over resources*, falls below some minimally acceptable level. In addition to requiring a definition of economic well-being, the measure must specify a *minimum level of well-being* (or 'needs') in terms that are commensurate with 'resources.' Sen (1983) considered the needs standard (or poverty line) to have 'some absolute justification of its own,' a level below which 'one cannot participate adequately in communal activities, or be free of public shame from failure to satisfy conventions' (p. 167). An ideal poverty measure would allow for differentiation according to household size and composition and would enable people to compare the level of access with resources over time.

Having determined the measure of household resources, the needs standard—or poverty line—appropriate for various households allows one to compare the level of household resources with its needs, and thereby to indicate if the family is poor or not poor. When this process is followed for all households in a nation (or some other jurisdiction), one obtains the *poverty rate*, or the percentage of the people (or households) in the jurisdiction who are living in poverty.

Within this economic perspective, a number of economic well-being indicators have some claim to identifying the economic position of households. For

example, the official US poverty measure relies on the *annual cash income* of a family, and compares this with some minimum income standard or 'poverty line.' Alternative—and equally legitimate—measures suggest that the level of *annual consumption* better reflects a family's level of living, or that some measure of a family's *capability to be self-reliant* or a family's *own assessment of economic well-being* are superior concepts for measuring the level and composition of poverty. No matter which perspective is adopted, a large number of judgments and choices are necessary for developing a concrete poverty measure. For example, if income is taken to be the best indicator of economic status, is annual, multiyear, or lifetime income the appropriate measure? Should we examine pretax, pretransfer income or income after accounting for taxes and/or transfers? Should in-kind income be counted or not? Each decision implies both a different poverty rate and poverty population and a different set of antipoverty policies. Finally, irrespective of the decisions made, the resulting poverty measure can be either absolute or relative. The indicator is absolute if the definition of needs is fixed, so that the poverty line does not vary with the standard of living. A relative poverty measure reflects changes in the overall standard of living of the society.

1.1 Absolute Income Poverty Measures

The most prominent absolute income poverty measure is that of the USA, which is the only industrialized country to have an official poverty measure. See Ruggles (1990) and Citro and Michael (1995) for further discussion of the US definition and other definitions. Since the late 1960s, the nation has reported on the level of national poverty and the composition of the poor population. This indicator seeks to identify those families that lack sufficient annual cash income to meet what is judged to be their annual needs. The needs indicator is a fixed income threshold, constructed originally as a multiple of the food needs of households of different sizes and compositions. This threshold changes only as the price level changes, but not as average incomes in society change.

This and other income poverty indicators have a particular philosophical basis. The implied social objective is that, together, the community's efforts and those of the individual should ensure that some minimal level of living is attained. Although people may experience hardship in many dimensions—education, housing, food, social contacts, security, environmental amenities—only a sufficiently low level of money income matters in determining who is poor. Such a measure also presumes that money income is a good proxy for welfare (or utility), and that a particular year's income is an acceptable indicator of longer-run income.

Aside from taking exception to the social objective that underlies the official measure, most other criticisms of it focus on the adequacy of the annual cash income measure of 'command over resources.' While the current cash income numerator of the poverty ratio may reflect the extent to which the family has cash income available to meet its immediate needs, it indicates little about the level of consumption spending potentially available to the family. For many families, annual income fluctuates substantially over time owing to unemployment, decisions to change jobs, or changes in health. Even as an indicator of a family's ability to meet its immediate needs, the current income measure is flawed—it reflects neither the recipient value of in-kind transfers (e.g., food or medical care subsidies) which support the economic well-being of low-income families, nor the taxes for which the family is liable. Similarly, while current cash income—and hence the official poverty measure—reflects financial flows in the form of interest and dividends from the assets held by individuals, the assets themselves are not counted, nor is the value of leisure (or voluntary nonwork) time reflected in the measure. The arbitrary nature of the denominator of the poverty ratio—the minimum-income-needs indicator—has also been criticized. Given the weak conceptual basis and the crude empirical evidence on which the dollar cut-offs rest, including the equivalence scales designed to account for the income needs of different family sizes and structures, the official US poverty lines are viewed generally as arbitrary.

These weaknesses have stimulated efforts in the US to improve the official poverty measure. In 1995, a panel appointed by the National Research Council issued its report suggesting revisions in the official measure. The poverty thresholds were modified so as to reflect a broader commodity bundle of needs (relative to the food need basis of the current thresholds) and a new set of equivalence scales. Moreover, the measure of resources was revised to include the value of in-kind transfers received by a household and to exclude estimates of the labor and income taxes for which the household is liable. (see Citro and Michael 1995). While alternative versions of this new measure are being published, the new measure has not yet been adopted formally.

1.2 Relative Income Poverty Measures

Use of a relative income poverty measure takes into consideration changes in the overall economy (in terms of wages and prices); it stresses the relationship of the resources of a living unit to those available to other living units in the society.

One prominent relative poverty definition, used extensively by the European Community, considers those with incomes less than one-half of the median income to be in poverty, hence reflecting the view that

poverty is only meaningful when compared with overall income or spending levels. Often, a one-half of median income standard is used in the analysis of poverty in the European Community, but more recently, a 60 percent of the median standard has been adopted (see Bradshaw 2001).

In addition to the arbitrariness of the 50—or 60—percent threshold, this measure has other weaknesses. Critics point out the weakness of a relative measure in assessing the efficacy of antipoverty efforts. Moreover, the relative income standard ensures that the poverty threshold will rise most rapidly in periods of economic growth, during which time those at the bottom of the distribution also experience real growth in both earnings and consumption. Hence, even though poor families may perceive themselves as better off during a prosperous period, the poverty rate may not fall, thus overstating the poverty problem. As Ruggles (1990) has stated: 'poverty cannot decline under a relative poverty measure without some change in the shape of the income distribution as a whole.'

1.3 Consumption-based Poverty Measures

A primary criticism of income poverty measures is that the annual income concept on which they rest is a poor indicator of the permanent income (or lifetime resources) of the family unit. Using such a measure, a wealthy family with a well-educated head and substantial assets, but a year of low income, would be classified as 'poor.'

One proposal designed to avoid this problem involves use of measured family consumption to determine poverty status, based on the argument that family consumption is a superior proxy for the family's permanent income, or family command over resources. Slesnick (1993) argues that income measures of poverty are 'severely biased' (p. 2) in that *temporarily* low-income families are regarded as *permanently* needy. The consumption-based poverty indicator proposed by Slesnick uses household real consumption expenditure per equivalent adult (taken to be the quotient of real household consumption and a household-specific cost-of-living index and an equivalence scale) as the indicator of resources.

Slesnick's consumption poverty measure has been criticized on several grounds. One particularly salient criticism concerns the arbitrary and potentially misleading equivalence scales developed by Jorgenson and Slesnick (1987) that are employed. (see Triest 1998). While other consumption-based indicators (see Cutler and Katz 1991) have shown a growing poverty rate over the last two decades, the Slesnick measure suggests that poverty in the USA has been decreasing secularly over that period.

In general, the largest impediment to utilizing a consumption-based index is the difficulty of obtaining complete and accurate family expenditure data, at

least relative to measuring a family's income. Furthermore, consumption may not fully reflect a family's true well-being; it is possible that simple frugality may be mistaken for poverty.

1.4 Subjective Measures of Poverty

An alternative view of poverty is the 'Leyden School' approach, which relies on the subjective responses of individuals to questions that seek their impressions of their own economic position or well-being, relative to some norm. (see Hagenaars 1986, van Praag et al. 1982). As with income and consumption poverty measures, subjective poverty indicators are based on an 'access to resources' concept, but because subjective thresholds change over time, these measures are relative poverty indicators.

Subjective poverty measures are based on survey responses to questions regarding the minimum level of income or consumption that respondents consider to be 'just sufficient' to allow them to live a minimally adequate lifestyle. If families respond that their own level of living exceeds that minimum, it is possible by observing their actual income to infer both a poverty line and a poverty rate. Because subjective measures are based on individual opinions of what constitutes 'minimally adequate' or 'enough to get by,' they require that individual perceptions of these notions reflect the same level of real welfare for all respondents.

A number of subjective measures have been developed and tested, mainly in Europe. Despite rather minor differences in terminology and question phrasing among these measures, they have yielded highly diverse results. The effectiveness of subjective measures is also limited by the nature of the data collection method. Most estimates are based on small sample sizes, yielding large standard errors.

1.5 Inability to be Self-reliant as a Poverty Measure

One additional poverty indicator should be mentioned—one that views those who are without the capability to be self-reliant as the ones with the lowest economic position. This view rests on two considerations, one conceptual and the other practical. The *conceptual reason* is the more basic. This position has its foundations in the writings of Sen (1992, 1997), who argued 'that the basic failure that poverty implies is one of having minimally adequate capabilities,' (1992, p. 111) and hence that 'poverty is better seen in terms of capability failure than in terms of the failure to meet the "basic needs" of specified commodities' (1992, p. 109).

There is also a *policy-related reason* for developing a capability measure of poverty. In recent years there has been renewed civic debate regarding appropriate norms and standards for individual responsibility and

behavior, and hence the appropriate role of the state. A prominent viewpoint in this debate has emphasized the merits of individual independence (relative to reliance on government programs), the negative effects of government programs on individual behavior, and the desirability of a smaller economic and social policy role for government. If society is to base policy on the central social goal of 'economic independence,' then it would seem important to identify the size, composition, and growth of the population of citizens who do not have the capability to be independent in a market economy.

An example of such a capability-based poverty measure is based on the concept of a family's net earnings capacity (the amount that the adults in a family could earn if they worked full time, full year, adjusted for the constraints on working at capacity due to health problems or disability and the expenses required if all of a family's working-age adults did work at capacity). If earnings capacity is less than a poverty threshold, the family is defined as 'unable to be self-reliant,' and in poverty (see Haveman and Bershadker 2001). A related measure of family capability is Becker's (1965) concept of 'full income,' which includes both income realized through market work and the value of leisure time.

Measuring 'self-reliant poverty' requires several implicit conventions, norms, and assumptions. For example, such a measure is reliable only if the 'norm' of full-time, full-year work is an accepted, socially determined norm representing the full use of human capital. Of necessity, attribution of poverty status to any particular family requires prediction from statistical estimates rather than values measured in survey data (such as income), and hence may be viewed as less objective than, say, income-based measures.

2. The Analysis of Poverty

Since the mid-1960s and the US declaration of a War on Poverty, social science researchers have addressed a number of issues concerning the nature, causes, and cures of poverty. In the process, substantive additions to economic knowledge and to research methods have been developed.

2.1 Understanding the Determinants of Poverty: Demographics, Discrimination, Human Capital, Immigration, and Macroeconomic Performance

A major objective among social science researchers has been to understand the processes which combine to result in some people failing to attain minimal standards of economic well-being. Although the number of such processes and institutions is very large, and the interactions among them complex, research has been concentrated in several areas.

The underlying *demographic structure* of the population—age, gender, family arrangements, fertility choices—plays a major role in determining whether families or individuals are poor. For example, the share of the population that is elderly, and hence unable to be self-reliant through work and earnings, influences the prevalence of poverty in a society. Research has tracked the poverty status of the older population and evaluated the ability of older people to escape poverty given their own resources and social support. Similarly, changes in the prevalence of single-parent families caused by rising divorce and non-marital childbearing rates are also fundamental determinants of poverty. Families with children but only a single adult tend to be restricted in their ability to be self-reliant from own earnings. Social scientists have studied the trends and determinants of these choices in an effort to better understand the causes of poverty.

If the attitudes of the majority of the population toward minority groups are such as to lead to lower payments for equivalent work or to inferior schooling, housing, or mobility opportunities, poverty rates among minority groups would tend to be high. Given the prevalence of racial minorities among the poor population, the possible causal role of *labor market, education, and housing discrimination* has been an important topic of study, especially by economists. While the existence and extent of racial income gaps had been studied prior to the 1960s, later research delved deeper than previous work in measuring and explaining the sources of these income gaps. See Cain (1986) for a review of this line of analysis. Given that competition would be likely to erode the effects of discriminatory behavior by employers (Becker 1957), research turned to explanations involving ‘statistical discrimination’ and ‘market signaling’ as the basis for continued discrimination as a cause of poverty.

The role of *education and training*—human capital—in enabling people to attain self-sufficiency also became a focus of poverty research and analysis. The question of the effect of increases in schooling on individual earnings occupied the efforts of numerous researchers, and education came to be analyzed as a process that created human capital. Estimates of the rates of return on schooling seemed to justify diverting resources from private sector investments in physical capital to public investments in education. Rosen (1987) surveys the literature in this area. While economists focused on the earnings returns from schooling, other social scientists sought to document the effect of schooling on student attainment (such as test scores), social mobility, and economic status. This research documented the complexity of the processes by which some children succeed while others fail, and led to a search for improved data and statistical methods capable of disentangling the effects of education on attainments. This ‘human capital’ approach to schooling carried over to numerous studies of the

effectiveness of governmental programs designed to provide skills and training to adult workers. See Heckman and Lochner (2000) for a review and interpretation of the evidence on the economic effects of education and training.

Concern with the problem of poverty in the USA coincided with a rapid increase in *immigration*, often of individuals with low levels of schooling or language competence. The potential impact of immigration in eroding wage rates for unskilled workers in urban labor markets containing large numbers of immigrants became an important policy and research issue.

The effect of the overall performance of the economy on the poverty rate was also the subject of substantial research and analysis, beginning in the 1960s. During the 1970s, economic growth and low unemployment seemed to be powerful instruments in reducing unemployment. However, during the 1980s and early 1990s, economic growth seemed to have lost its antipoverty effectiveness. Evidence suggests that during the 1990s, the level of poverty was again responding to macroeconomic performance. See Haveman and Schwabish (2000) for a review of this research and for recent evidence.

2.2 Understanding Social Mobility, Status Attainment, and Income Dynamics

Research on the causes of poverty have tended to focus on static issues—factors that appear to be causally related to low economic position. Simultaneously, a complementary line of research addressed questions of the dynamic processes by which people either succeed or fail to attain economic position—processes that indicate the openness or fairness of society. While research on these social mobility processes had been undertaken prior to the 1960s, those studies were limited by the absence of both a conceptual framework and data on individuals over time—longitudinal data—necessary to trace these dynamic patterns. The early 1960s witnessed a number of important research contributions designed to increase understanding of the process of social mobility. One of the most central of these presented a causal model of the socioeconomic life cycle, distinguishing the interactions among three primary life phases: family, schooling, and work (see Duncan and Hodge 1963). The socioeconomic status of the job was taken to be the ultimate indicator of success, with family back-ground and schooling (as an intervening variable) viewed as the determinants of that status. This framework stimulated several empirical studies of these relationships, yielding insights into changing intergenerational correlations involving occupations and income, changes in the degree of social mobility, changes in the role of education in fostering mobility, and changes in mobility patterns between whites and nonwhites. A variant of this approach sought to

identify the factors that contribute to social or economic status, distinguishing those that relate to exogenous family and background factors from those amenable to policy influence. Two prominent examples of this approach are Jencks et al. (1972, 1979). Family characteristics and choices and intellectual inheritance were seen as playing important roles, but school quality and the racial mix of neighbors and peers were found to be less important. Haveman (1987) reviews and critiques this research.

While these analyses were revealing, they demonstrated the difficulty of making progress in this area without detailed data on individuals and families that tracked their attributes and performance over time. The Panel Study of Income Dynamics, the National Longitudinal Surveys, and the Survey of Income and Program Participation are among the most prominent and high quality of these data collection efforts. Extensive use of them has enabled major advances in understanding the nature and causes of poverty, including the lengths (or persistence) of poverty and welfare stays among the poor, trends in income trajectories among families with various characteristics, the identification of the long-term or permanently poor population, and estimates of the relative roles of family background, schooling choices, neighborhood characteristics, and labor market discrimination in determining the probability of being poor. A complementary field of study has used these data to explore the determinants of these family characteristics (e.g., poverty-level incomes, single-parent status, low parental education, welfare reciprocity), schooling quality, neighborhood characteristics, and own schooling and fertility choices on a variety of children's attainments (see Haveman and Wolfe 1995, Duncan and Brooks-Gunn 1997).

2.3 Antipoverty Effects of Policy Interventions

Poverty researchers have also sought to describe the various policy measures in place to assist poor families and to identify changes in or alternatives to these policies that might be effective in reducing poverty. Probably because the main antipoverty efforts were in the area of income support measures, analysts concentrated their assessments on such policies. A voluminous literature describing the nature of income support policies and critiquing their incentive (efficiency) and equity characteristics has influenced policy changes over the period since the 1960s. These analyses revealed low 'take-up' rates among people eligible for support, horizontally uneven support patterns among equally poor people, the uncertain rationale for and implications of providing support via in-kind benefits vs. tax subsidies vs. cash transfers, the relative effectiveness of alternative programs in reducing poverty rates or the intensity of poverty, and the presence of incentives for undesirable behaviors (e.g., reductions in work, increases in family breakups, and

nonmarital childbearing) implicit in the structure or administration of the programs.

These descriptive and policy analysis studies played an important role in guiding numerous antipoverty policy reform efforts that have occurred over the past several decades, including proposals for 'negative income tax' plans, work-related subsidies, the substitution of in-kind for cash benefits, and 'welfare-to-work' initiatives. They also provided a catalog of important questions for which answers would be relevant in designing effective policy: how large are work effort reductions in response to income transfers? Has public income support reduced family stability, increased unemployment, reduced savings, or increased nonmarital childbearing? (see Haveman 1987).

2.4 Labor Supply, Family, and Fertility Effects of Income Support Policy

Concerns with the work incentive effects of income support policy have dominated debates over this antipoverty strategy. Numerous analysts responded to this concern with empirical studies documenting the extent of these behavioral responses to public income support programs. These studies, employing a variety of survey data and empirical methods, found that public benefits tend to reduce employment probabilities, lower work hours per year if working, and possibly reduce wage rates among both recipients and potential recipients of benefits.

Most of the early studies employed cross-sectional survey data in order to measure the responses of individuals to the increased income from transfer benefits and the reduced effective wage rate associated with working. Later studies relied on the results from carefully designed 'social experimental' studies of these potential responses (see below). The essential characteristic of these studies is the random assignment of families to alternative forms of transfer program designs (in terms of the generosity of income guarantees and the extent of 'benefit reduction' rates), and the observation of the behavior and performance of these families over subsequent years. Both forms of studies confirmed that, as expected, work effort falls as nonlabor income rises and as the effective wage rate falls. While the estimates varied widely, nearly all of them raised serious questions about the efficiency and effectiveness of this income support approach to helping poor people.

While these studies focused on the work effort responses of people to the availability of income support, other analyses (including randomized approaches) tested the family structure, childbearing, health care utilization, housing quality, and welfare utilization choices of low-income families in response to both welfare-type policies and reform efforts in the educational, housing, health insurance, and public employment areas. Danziger et al. (1981) and Moffitt (1992) have surveyed this research.

3. Poverty Research and Social Science Methods

In addition to adding to the body of knowledge regarding the nature, causes, and cures of poverty, the growth of poverty-related research since the 1960s has also contributed to the development of new research methods of general applicability in the social sciences (Haveman 1987). Four contributions deserve mention.

The first contribution consists of a new field of evaluation analysis and research. As analysts, in response to the demands of policy makers, focused on the poverty problem, an emerging analytical area, evaluation research, was also being spawned. Interest in this applied research approach extended far beyond the poverty issue, but research on social and antipoverty policy issues provided this then-nascent subdiscipline with some of its basic methods and case studies. Without the emphasis of policy makers on the need to evaluate the effectiveness of the numerous new antipoverty initiatives, it is unlikely that this field of analysis, with its journals and professional organizations, would have developed as rapidly as it has, and in the same directions.

Second, as noted above, randomized social experimental research was first applied to proposed policy interventions in the poverty area—the negative income tax proposals, school and housing vouchers, employment and training measures, and so on. Without the emphasis on the effect of antipoverty policy interventions on family economic status during the 1960s and 1970s, it is unlikely that this important social science research method would have attained the status it now holds among social science researchers.

One of the problems that analysts face in assessing the effects of policy and other changes on the behavior of people is the need to rely on samples of observations that may be composed of individuals who are selectively included in the sample, and hence unrepresentative of the larger population for which inferences are desired. This concern dominated the study of labor markets well before demands to assess the impacts of antipoverty measures, especially in the employment and training area, where factors that determined the presence of a person in a research sample were also determinants of the effect that was being studied (see Ashenfelter 1978). Several approaches to this problem were developed by econometricians (see especially Heckman 1979), and these statistical tools became a standard part of applied statistical analyses.

Finally, in part because of the development of large representative samples of national populations, analysts began developing computer models designed to simulate the effects on the income, well-being, and behavior of each observation in a data set of actual or proposed policy measures. The application of such microdata simulation models found its most welcomed audience among analysts concerned with the effects of changes in taxation and income transfer policy on low-

income families. Without the impetus provided by the need to assess the impacts on poor people of proposed policies, this advance to predictive model building would not have occurred. This research methodology is assessed in Citro and Hanushek (1991).

See also: Education: Skill Training; Human Capital: Educational Aspects; Income Distribution: Demographic Aspects; Male Dominance; Mobility: Social; Policy History: State and Economy; Poverty and Gender in Affluent Nations; Poverty, Culture of; Poverty in History; Poverty Policy; Poverty, Sociology of; Social Inequality in History (Stratification and Classes); Social Mobility, History of; Social Stratification; Wealth Distribution

Bibliography

- Ashenfelter O 1978 Estimating the effect of training programs on earnings. *Review of Economics and Statistics* **60**: 47–57
- Becker G S 1957 *The Economics of Discrimination*. University of Chicago Press, Chicago
- Becker G S 1965 A theory of the allocation of time. *Economic Journal* **75**: 493–517
- Bradshaw J 2001 The measurement of absolute poverty. In: Schokkaert E (ed.) *Ethics, Poverty and Inequality and Reform in Social Security*. Ashgate, Aldershot, UK
- Cain G C 1986 The economic analysis of labor market discrimination: A survey. In: Ashenfelter O, Layard R (eds.) *Handbook of Labor Economics*. North Holland, Amsterdam
- Citro C, Hanushek E 1991 *Improving Information for Social Policy Decisions: The Uses of Microsimulation Modeling*. National Academy Press, Washington, DC
- Citro C, Michael R 1995 *Measuring Poverty: A New Approach*. National Academy Press, Washington, DC
- Cutler D M, Katz L F 1991 Macroeconomic performance and the disadvantaged. *Brookings Papers on Economic Activity* **2**: 1–61
- Danziger S, Haveman R, Plotnick R 1981 How income transfer programs affect work, savings, and the income distribution: A critical review. *Journal of Economic Literature* **19**: 975–1028
- Duncan G J, Brooks-Gunn J (eds.) 1997 *Consequences of Growing up Poor*. Russell Sage Foundation, New York
- Duncan O D, Hodge R W 1963 Education and occupational mobility. *American Journal of Sociology* **68**: 629–44
- Hagenaars A J M 1986 *The Perception of Poverty*. North Holland, Amsterdam
- Haveman R 1987 *Poverty Policy and Poverty Research*. University of Wisconsin Press, Madison, WI
- Haveman R, Bershadker A 2001 The ‘inability to be self-reliant’ as an indicator of poverty: Trends for the US 1975–1997. *Review of Income and Wealth* **47**(3)
- Haveman R, Schwabish J 2000 Has macroeconomic performance regained its antipoverty bite? *Contemporary Economic Policy* **18**(4): 415–27
- Haveman R, Wolfe B 1995 The determinants of children’s attainments: A review of methods and findings. *Journal of Economic Literature* **33**(4): 1829–78
- Heckman J 1979 Sample selection bias as a specification error. *Econometrica* **147**: 153–61
- Heckman J, Lochner L 2000 Rethinking education and training policy: Understanding the sources of skill formation in a

- modern economy. In: Danziger S, Waldfogel J (eds.) *Securing the Future: Investing in Children from Birth to College*. Russell Sage Foundation, New York
- Jencks C, Smith M, Acland H, Bane M J, Cohen D, Gintis H, Heyns B, Michelson S 1972 *Inequality: A Reassessment of the Effect of Family and Schooling in America*. Basic Books, New York
- Jencks C, Bartlett S, Corcoran M, Crouse J, Eaglesfield D, Jackson G, McClelland K, Mueser P, Olneck M, Schwartz J, Ward S, Williams J 1979 *Who Gets Ahead? The Determinants of Economic Success in America*. Basic Books, New York
- Jorgenson D W, Slesnick D T 1987 Aggregate consumer behavior and household equivalence scales. *Journal of Business and Economic Statistics* 5: 219–32
- Moffitt R 1992 Incentive effects of the US welfare system: A review. *Journal of Economic Literature* 30(1): 1–61
- Rosen S 1987 Human capital. In: Eatwell J, Milgate M, Newman P (eds.) *The New Palgrave: A Dictionary of Economics*. Palgrave Publishers, Hampshire, UK, Vol. 2, pp. 681–90
- Rowntree B S 1902 *Poverty, A Study of Town Life*. Macmillan, London
- Ruggles P 1990 *Drawing the Line: Alternative Poverty Measures and Their Implications for Public Policy*. Urban Institute Press, Washington, DC
- Sen A 1983 Poor, relatively speaking. *Oxford Economic Papers* 35: 153–69
- Sen A 1992 *Inequality Reexamined*. Russell Sage Foundation and Harvard University Press, Cambridge, MA
- Sen A 1997 *On Economic Inequality*. Clarendon Press, Oxford, UK
- Slesnick, D T 1993 Gaining ground: Poverty in the postwar United States. *Journal of Political Economy* 101(1): 1–38
- Smith A 1776 *Wealth of Nations*. Modern Library Edition, 1993. Random House, New York
- Triest R K 1998 Has poverty gotten worse? *Journal of Economic Perspectives* 12(1): 97–114
- van Praag B M S, Hagenaars A J M, van Weeren H 1982 Poverty in Europe. *Review of Income and Wealth* 28(3): 345–59

R. H. Haveman

Poverty Policy

Poverty policy or the politics of poverty is a constituent element of social policy. It covers public action that aims to deal with poverty when and if poverty is seen as a public issue. While various forms of poverty or pauperism have always invited some action of close communities, the term poverty policy is reserved for interventions in which the modern state has played a role.

1. The Political and Intellectual Context

Poverty policy is informed by public policy in general or social welfare policy in particular. These policies are by their very nature ideologically loaded. The political labels to be applied may be debated. For the present purpose the conventional political categories of conservatism, liberalism, and social democracy appear

useful. The ideology based on the supremacy of the market will be termed neoliberalism. Conservatism and neoliberalism may be situated more to the right, liberalism and social democracy more to the left of the political spectrum. By means of these categories an involved pattern will be presented in an oversimplified way to offer an intellectual context for poverty policy.

The political right perceives poverty as an individual rather than a social problem, attributing its causes to the variations in individuals' merits and achievements. One conservative position adds as a main cause of individual failure genetic inheritance that explains for instance the poverty of the intellectually 'inferior' races in America (Herrnstein and Murray 1994). The variation between individuals being natural, poverty itself is seen as a natural and inevitable phenomenon. Its alleviation for altruistic reasons may be approved of, but attempts at its eradication are rejected as futile. Conservatives may prefer voluntary charity as a means of poverty alleviation. Neoliberals may accept state intervention in the form of a well-targeted safety net for the truly needy. The political right usually regards poverty policy as the only acceptable type of social welfare policy.

In a left perspective, poverty is considered as the outcome of social inequalities generated by the economic, political, and social power structure, by the unequal distribution of all the forms of capital. Poverty policy is seen as a must to alleviate immediately the lot of the poor by means of public redistribution. The prevention of the continuous reproduction of poverty is also a major concern, though. This objective requires broader public policies. Public intervention should promote a fairer access to all socially relevant material and symbolic resources, which in turn requires curbing the domination of the market over society. The success of such policies depends on, among other things, the empowerment of people and strong economic and social rights. In this broader policy framework, poverty policy becomes residual.

2. History

Poverty policy followed closely the economic and social changes connected with unfolding capitalism in Europe from about the fifteenth century. Four periods may be distinguished, although their contours are inevitably vague.

2.1 The Beginnings of a Public Policy Concerning the Poor

Central intervention focusing on the poor was prompted by several factors. The closely knit primary communities—the family, the neighborhood, the parish, the feudal manor, later the guild—endorsed for a long time the care of the weakest members unable to survive without this protection, who would have

otherwise endangered the cohesion or the stable reproduction of the group. The ability of self-help of the small communities was slowly eroded by wars, epidemics, the laicization of society, the loosening of feudal bonds, new forms of mobility, and an increasing number of marginalized vagrants. Most of these processes were connected to urbanization and industrialization. Following these developments, the number of the unsupported poor multiplied while the traditional forms of spontaneous or organized charity were waning. The direct encounter with the misery and the physical suffering of the poor affected uncomfortably their betters, whose sensibility had increased with the Enlightenment (Mollat 1978). More importantly, the poor have become dangerous. They may have caused direct personal danger through unrest, banditism, or subsistence crime. The more indirect danger was that they represented or exuded physical and moral 'evils' (Swaan 1988). The two aspects of increasing poverty—physical suffering and threats to public safety—invited Janus-faced interventions. The politics of poverty combined from the start the functions of policing (control and punishment) and of helping (poverty alleviation) (Geremek 1987). Legal regulations at the local or central level combining these two functions appeared with increasing frequency from the fifteenth and sixteenth centuries onwards (Castel 1995) in most Western countries and even in Eastern Europe. Until the early seventeenth century, the policing element was probably stronger (under the reign of Henry VIII between 1509 and 1547, 72,000 vagrants were executed in England). The supportive function became more explicit in the first major Poor Law of Europe adopted in England in 1601. This was the Elizabeth 43 Act for the Relief of the Poor. In its Section 2 it declared: 'It is agreed and ordered by the present Assembly that each town shall provide carefully for the relief of the poor, to maintain the impotent, and to employ the able...' The French ordinance of Moulin of the same year applies almost identical wording. At this time even if regulation was central, its implementation was the duty of the local authorities in collaboration with the churches and voluntary organizations. Despite the more explicitly prescribed obligation to give assistance, the policing function never disappeared.

Poverty policy was beleaguered from the earliest times by several dilemmas. One of them was the separation of the genuinely helpless from the able-bodied poor. Those belonging to the first group—the sick, the invalid, the frail elderly, sometimes orphans or widows—were usually considered as deserving help. The members of the second category were mostly seen as undeserving of support. It was assumed that they could maintain themselves were they not work-shy or lazy. It was acknowledged only very late—towards the end of the nineteenth century or even later—that the majority of the able-bodied poor would have been willing to work had they had access to a job. Another

lasting problem was the separation of the local poor from strangers. The locality was willing to endorse the responsibility for its own poor but was keen to exclude strangers; hence the various acts concerning the duties of provision and the rights to expulsion of the settlements, and also the practice of visible stigmatization of those getting assistance or of those allowed to beg in the locality. These regulations were ubiquitous from Pennsylvania in North America to the town of Győr in Hungary. A third problem centered on the level of assistance. This level was haphazard at least until the wide acceptance of the 'less eligibility' principle defined by the (English) New Poor Law of 1834. The principle implied that 'the pauper's condition should be rendered less eligible, or more unpleasant than that of the poorest paid independent laborer.' The workhouse was then conceived of so as to assure maximum feasible less-eligibility. The principle has also been applied to define the level and the modalities of access to other types of assistance. A further dilemma that lost its edge by now was the choice between so-called indoor and outdoor relief. Indoor relief meant the poorhouse or the workhouse not only in England but all over Europe and America (termed, for instance, *Hôtels de Dieux* and *hospitaux pour renfermement des mendiants* in France) from the sixteenth century onwards. The regulation of poor relief in Elberfeld, Germany, became the model for organized outdoor relief only in the nineteenth century.

2.2 *The Handling of Industrial Poverty*

With the unfolding of capitalism, the character of poverty changed. The number of the urban poor multiplied, and urban destitution acquired tragic dimensions. Early nineteenth century legislation dealing with urban hygiene, or the Factory Acts protecting young children and mothers, were concerned less with utter poverty than with the survival of the whole community. Public assistance continued to play its role often with a strong punitive edge, but it also acquired new dimensions. The workers were affected by new insecurities as compared with the 'pre-industrial' poor. The risks of grave accidents, of unprovided old age, or of job losses all increased, and were slowly recognized as new causes of poverty. Moreover, the workers started to organize and mobilize themselves. The expansion of political rights forced the states to be more responsive to the pressures of vocal political thinkers and of popular movements. From the last third of the nineteenth century progressive taxation, various branches of social insurance, and in a few countries fully or nearly universal flat-rate benefits were added to the instruments of public redistribution. The monopolistic role of poverty policy within social welfare policy was thus ended. From the early twentieth century some countries—Sweden in the first

place—formulated more ambitious projects about the improvement of the working and living conditions of the whole community, and about the abolition of poverty as well as of a separate and autonomous poverty policy.

2.3 *The Welfare State and Poverty Policy*

During and after World War II, the extension of public responsibility for all-encompassing social security became an accepted, albeit never uncontested, principle in most developed capitalist states. The seminal document founding the idea of the Welfare State was the Beveridge Report, a proposal prepared for the British government published in 1942 under the unassuming title 'Social Insurance and Allied Services.' It proposed to fight by means of universal benefits and services the five 'giant evils,' Want, Disease, Ignorance, Squalor, and Idleness—that is, to end poverty and its root causes. The experiences of World War II, a widespread appeal of left ideals, and the enforced competition with the 'communist' world in Central and Eastern Europe triggered all over Western Europe the development of social protection, increasingly underpinned by labor law and social law.

The welfare state project was never and nowhere fully successful. Yet in its three ascending decades it considerably reduced social inequalities, deep and even shallow poverty, and improved the working, living, and housing conditions and the existential security of the majority. Poverty policy did not disappear, but in most countries its role—and the role of social assistance—have become marginal (Hatzfeld 1971, Esping-Andersen 1990). The situation was different in the USA where universal schemes were never accepted and their function was fulfilled by the targeted schemes devised within the framework of poverty policy. However, from the 1960s on, the War on Poverty and the project of a Great Society improved the welfare programs. For about 15 years they became more generous and better adjusted to needs. The results are contested by many on the right. Yet statistical evidence shows that the poverty rate fell by about half, some inequalities were reduced, and housing, health, and nutrition indicators rose also among those on the assistance rolls.

In Central-Eastern Europe, pre-war poverty policy had many elements in common with the most conservative, more punitive than generous systems of Western Europe. After World War II, the modern systemic differentiations were supplanted by a monolithic system in which a single party's politics dictated the rules of all subsystems such as the economy, law, or social policy. Welfare policy was absorbed by the economy, which was in its turn dominated by politics. Poverty policy did not exist for two more reasons. On the one hand, price policy, full employment policy, broad and almost all-encompassing health, pension, etc., schemes were constructed so as to produce a not

too unequal income distribution, and to prevent relatively successfully the emergence of deep poverty. On the other hand, the dominant ideology denied the existence of the poverty that still existed, and therefore did not make any provision to handle it. It had no need and no space for a specific poverty policy, leaving many social ills unattended (Ferge 1986).

2.4 *The Withdrawal of the Welfare State and the Revival of Poverty Policy*

Following the oil shock and the advent to power of the Reagan and Thatcher administrations, the tide turned. A neoliberal agenda has unfolded requesting less state intervention in the operation of the free market and the retrenchment of the social state. The collapse of totalitarian state socialism bolstered this process.

The new trends affect the traditional institutions of social policy. Universal solutions with access tied to citizenship rights are losing ground. In the case of social insurance, there are efforts to privatize it, to weaken its solidary (nonactuarial) components, and to lower its standards. The efforts to deregulate the labor market have weakened labor rights and increased the number of working poor. The consequences are increasing poverty and unmet needs. The countries in 'transition' are inordinately hit by these changes. The solution offered everywhere is the increased role of poverty policy, particularly of means-tested benefits. The political support of programs targeted only to the poor may be so weak, however, as to threaten subsistence rights seriously.

The new ideology has affected social policy everywhere, albeit not to the same extent. The European welfare states having integrative programs with strong historical roots and wide public support were less affected than other parts of the world. In the USA even the welfare programs seem to have suffered severe cuts (Piven and Cloward 1982, Katz 1993). In many Latin American and 'postsocialist' countries, proposals figuring on the neoliberal agenda are replacing the former, more comprehensive solutions. In other parts of the world where welfare policies hardly existed before, they seem to appear mainly as variants of poverty policy. In East Asia this is explained by the Confucian ideology of self-reliance and familism (Goodman et al. 1998). In desperately poor third-world countries poverty policy seems to be the only alternative to improve the situation, albeit the mobilization and empowerment of the poor seem to acquire greater weight than in earlier poverty programs (World Bank 2000, United Nations 1998).

3. *Current Theory and Research*

Theory and research on poverty and on poverty policy are strongly interwoven. An attempt is made here to focus on the policy aspects. One school of thought

deals with the theoretical and political possibilities of poverty reduction on a world scale. It explores the possibilities and instruments of global redistribution. Supranational agencies, particularly the United Nations, the International Labor Office, and more recently the World Bank as well as some civil organizations, are stimulating this research. Many elements of these endeavors are summarized in the Report of the UN World Summit on Social Development held in 1995. A second trend deals with the role of poverty policy within public policy. It explores the respective roles of the state, of civil society, and of the market (economic growth) in decreasing poverty and social exclusion. The main question is whether the 'trickling down' of economic growth is a sufficient condition to lift the poor or whether additional public and civil efforts are needed. If the latter is the case (as research seems to show), the question is whether focusing on the poor is sufficient, or a broader approach is warranted. A third large field explores the instruments of poverty policy and of their evaluation. These are partly methodological issues mentioned in Sect. 4. Some of them have acquired momentous theoretical importance. The age-old problem of poor relief, the kernel of the 'social question,' was always how to separate those worthy of help and compassion from the undeserving poor. The growing number of able-bodied poor because of lasting unemployment and the increase in single parenthood and low-paid jobs put the issue again forcefully on the agenda. One direction of research—more to the left—looks for 'third way' solutions including investment in human capital and the creation of acceptable jobs. The conservative research agenda is involved in proving the harms of welfare programs that stifle incentives, destroy stable families and lead to a dependency culture.

4. Distributional Issues

If help is to be given, it has to be known at least who to give it to, what to give and how, and how to assess the results of the policy.

4.1 Who Are to Be Helped?

The poor and the needy may be defined in different ways in poverty research and in poverty politics. In research the definition may be based on objective or subjective, or on absolute or relative poverty criteria; it will depend on whether the research relates to one country or is comparative, whether it describes one point in time or covers processes—let alone the idiosyncratic outlook of the researcher. Poverty policy usually applies definitions which relate to one country and one time-point, albeit they may change with time. The main instrument of targeting help is the minimum income or assistance level, which becomes *de facto* a poverty line. Those under the line may get help. The poverty line (unlike in research) has to be unam-

biguous. It is usually defined by central or local governments on the basis of pragmatic considerations. It may be lower than a scientifically based adequate minimum income standard (Eardley et al. 1996).

The poverty line may relate to the individual, the family, or the household. In the last two cases different weights may be assigned to family or household members, that is, different equivalence scales may be applied (OECD 1997). In addition, poverty policy may target help also to those who are above the poverty line but whose resources are insufficient to cover some special needs. Examples are assistance offered to cover unusually high housing costs or health costs, or in the case of a crisis situation.

4.2 Who to Give to, in What Form, and How Much?

Help may not be offered to all those under the poverty line. The selection of the beneficiaries is in part a political issue discussed above. Methodologically it has to be decided whether to apply only an income test, or a means (resource) test comprising an income test and an asset test, or whether to add also various behavior tests (a work test, a cohabitation test, and such like). The form of the provision also matters. There are arguments for and against assistance in cash to be used freely, or tied in some way, given in kind or in intermediary forms such as vouchers or food stamps. When a poverty line is established, the policy may be to offer enough assistance to cover the whole poverty gap (the difference between the set line and the actual income) or to alleviate only the situation of the poor without eliminating the poverty gap. The adequacy of the provisions is a further major issue. The European Union has strongly and repeatedly recommended a 'right to sufficient resources' or a minimum income for all as a measure to reduce poverty and to promote social integration (Schulte 1997). In other parts of the world governments may accept poverty lines well below the adequacy level either because of insufficient resources or because of ideological reasons (Veit-Wilson 1998). The provisions may then still assure subsistence, or only mere survival, or not even that (Gans 1995). The 'less eligibility' principle appears still to survive.

4.3 The Assessment of the Policies

The methods of evaluation of poverty policy were developed partly in connection with the war on poverty in the USA, and partly by independent researchers in Britain and elsewhere in Europe. Evaluation deals with such issues as the efficiency and effectivity of the various programs (to what extent did they reach the target groups and did they solve the problems); the measurement of the impact of policies on horizontal and vertical equity; the costs of relief; the public acceptance of different programs; the reasons of varying take-up rates; the merits and

demerits of social assistance based on right or on discretion; the dangers of emergence of poverty traps, unemployment traps, or savings traps—all hampering efficient help. The assessment of the negative incentives mentioned above as one of the most important theoretical issues requires involved research methods, too (Haveman 1987).

5. The Future

The future of poverty and of the orientation of poverty policy depends to a large extent on political forces. Supranational agencies, the United Nations in the first place, are committed to poverty reduction on a global scale, with emphasis on the poorest countries. The attainment of this objective would require a stronger international state with economic mandates and a stronger international civil society to curb the strengthening market—conditions different to fulfil.

The future of individual countries will depend on the global trends, but also on their own policies. One may expect the continuation of the trends described in Sect. 2.4 with some variations and innovations (McFate et al. 1995). In the European Union and in some left-liberal circles there is a search for new instruments to combat poverty and exclusion such as the citizen's basic income (Parijs 1992), or a new integration of economic and social policy (Atkinson 1995). This orientation implies a reduction of the role of means testing in favor of more encompassing public policies. These proposals usually meet with the resistance of economically and politically strong groups.

All in all, at the start of the new millenium there are genuine threats of widening within- and between-country inequalities, of growing poverty and of the increasingly harsh treatment of the poor whose right to subsistence is questioned. The problem is not so much the scarcity of resources as their distribution. Vigorous public and poverty policies would be needed to achieve more equity in order to prevent major tensions.

See also: Feminist Economics; Industrialization; Inequality; Inequality: Comparative Aspects; Poverty, Culture of; Poverty: Measurement and Analysis; Poverty, Sociology of; Social Security; Underclass; Welfare; Welfare Programs, Economics of; Welfare State, History of

Bibliography

- Atkinson A B 1995 *Incomes and the Welfare State*. Cambridge University Press, Cambridge, UK
 Castel R 1995 *Les Métamorphoses de la Question Sociale: Une Chronique du Salarial* [The Transformations of the 'Social Problem': A Chronicle of Wage-earners]. Fayard, Paris

- Eardley T, Bradshaw J, Ditch J, Gough I, Whiteford P 1996 *Social Assistance in OECD Countries*. HMSO, London
 Esping-Andersen G 1990 *The Three Worlds of Welfare Capitalism*. Polity Press, Cambridge, UK
 Ferge Z 1986 *Fejezetek a Magyar Szegénypolitika Történetéből* [The History of the Politics of Poverty in Hungary]. Magvető, Budapest, Hungary
 Gans H J 1995 *The War Against the Poor. The Underclass and Antipoverty Policy*. Basic Books, New York
 Geremek B 1987 *La Potence et la Pitié. L'Europe et les Pauvres du Moyen Age à Nos Jours* [Gallows and Compassion. Europe and the Poor from the Middle Ages to the Present]. Gallimard, Paris
 Goodman R, White G, Kwon H J (eds.) 1998 *The East Asian Welfare Model: Welfare Orientalism and the State*. Routledge, London
 Hatzfeld H 1971 *Du Paupérisme à la Sécurité Sociale* [From Pauperism to Social Security]. Librairie Armand Colin, Paris
 Haveman Robert H 1987 *Poverty Policy and Poverty Research: The Great Society and the Social Sciences*. University of Wisconsin Press, Madison, WI
 Herrnstein R J, Murray C 1994 *The Bell Curve: Intelligence and Class Structure in American Life*. Free Press, New York
 Katz M B (eds.) 1993 *The 'Underclass' Debate. Views from History*. Princeton University Press, Princeton, NJ
 McFate K, Lawson R, Wilson W J (eds.) 1995 *Poverty, Inequality, and the Future of Social Policy: Western States in the New World Order*. Russel Sage Foundation, New York
 Mollat M 1978 *Les Pauvres au Moyen Age*. Hachette, Paris
 OECD 1997 *Family, Market and Community. Equity and Efficiency in Social Policy*. OECD, Paris
 Parijs van P. 1992 *Arguing for Basic Income: Ethical Foundations for a Radical Reform*. Verso, London
 Piven F F, Cloward R A 1982 *The New Class War: Reagan's Attack on the Welfare State and its Consequences*, 1st edn. Pantheon Books, New York
 Schulte B 1997 Juridical instruments of the European Union and the European Communities. In: Beck W, van der Maesen L, Walker A (eds.) *The Social Quality of Europe*. Kluwer Law International, The Hague, pp. 51–76
 Swaan de A 1988 *In Care of the State: Health Care, Education and Welfare in Europe and the USA in the Modern Era*. Oxford University Press, Oxford, UK and Polity Press, New York
 United Nations 1998 *Poverty Reduction Strategies. A Review*. United Nations, New York
 Veit-Wilson J 1998 *Setting Adequacy Standards. How Governments Define Minimum Incomes*. The Policy Press, Bristol, UK
 World Bank 2000 *World Development Report 2000/2001: Attacking Poverty*. Oxford University Press, New York

Z. Ferge

Poverty, Sociology of

Sociologists have long written about the topic of poverty, as have economists and other social scientists. Indeed, this article is complemented by several others with similar themes. Readers may therefore want to also consult other entries as well (e.g., *Poverty:*

Measurement and Analysis; Poverty and Gender in Affluent Nations). Here we begin with a basic description of the nature and causes of poverty as seen by sociologists, and second with the techniques used to measure poverty by sociologists and other social scientists in both national and cross-national contexts. Our study also focuses on cross-national comparisons with rich countries.

1. Sociology and Poverty

Because the fundamental concept of poverty concerns itself with having too little resources or capabilities to participate fully in a society, the study of poverty has found a wide following in sociology. Sociologists have been more interested in the explanation of poverty than in its measurement, though almost all sociologists believe that poverty statistics are meaningful social indicators of basic needs (Townsend 1993, Piachaud 1987, Ringen 1985). Sociological interests in poverty center around the ideas of the 'culture' of poverty and the effects of 'place' on poverty. The urban—central city—ghetto aspects of this issue are well treated in the work of Wilson (1987), Massey and Denton (1993), and Massey (1996); Harrington's classic work, *The Other America*, covers rural deprivation as well (1962). Roles of culture, power, social structure, and other factors largely out of control of the individual are the main forces that sociologists use to explain poverty. The basic working hypothesis is that individuals are strongly influenced by the physical and cultural context in which they live. 'Neighborhood' exerts a strong influence on behavior and concentrated poverty in central city ghettos therefore has a strong effect on future life chances and long-run deprivation. Indeed, in the United States, this has spawned interest in the so-called underclass, which goes beyond poverty alone to include all persons with dysfunctional behaviors living in 'bad' neighborhoods (Mincy et al. 1990). The European term for a similarly disaffected population is 'social exclusion' which is also beyond poverty measurement *per se* (Room 1999, Hills 1999, Glennerster et al. 1999).

Gender and household structure also frequently play a role in sociological theories of poverty (e.g., Miller 1996, Ringen 1985). As women's roles in society often leave them with care-giving responsibilities, with low-paying careers and jobs, and therefore in poor economic circumstances, poverty is often high among single female parents and elderly widows. Thus, we find a preponderance of women in poverty in most rich nations (Casper et al. 1994, McLanahan et al. 1989, Christopher et al. 2001).

In general, sociologists are critical of the economist's perspective of free choice models whereby individuals control their own destiny and are therefore the cause of their own poverty (Piachaud 1987). Some sociolo-

gists even feel that poverty has a functional role to play in capitalist society (e.g., Gans 1972). Ultimately, however, sociologists, like all other social scientists, need to first establish the breadth and depth of this social phenomenon called 'poverty' before they can meaningfully analyze it and explore its ultimate causes and remedies. Thus, we turn to measures and comparisons of poverty employed by sociologists and other social scientists within and across nations.

2. Poverty in Rich Nations: Concepts and Measures

We begin by noting that poverty measurement began as an Anglo-American social indicator. In fact, official measures of poverty (or measures of low income) exist in very few nations. Only the United States (US Bureau of the Census 1999) and the United Kingdom (Department of Social Security 1993) have official poverty series. Statistics Canada publishes the number of households with incomes below a low income cutoff on an irregular basis, as does the Australian government. In Northern Europe and Scandinavia, the debate centers instead on the level of income at which minimum benefits for social programs should be set. In other words, their concept of insufficient 'low income' is directly fed into programmatic responses to social needs (Björklund and Freeman 1997).

While poverty measurement is an exercise that is particularly popular in the English-speaking countries, most rich nations share the Anglo-Saxon concern over distributional outcomes and the well-being of the low-income population. There is no international consensus on guidelines for measuring poverty, but international bodies such as the European Statistical Office (Eurostat 1998), the United Nations Development Programme (UNDP 1999), the United Nations Children's Fund (UNICEF 2000), and the Organization for Economic Cooperation and Development (OECD 2000), have published several cross-national studies of the incidence of poverty in rich countries in recent years. The large majority of these studies are based on the Luxembourg Income Study (LIS) database which can be accessed remotely via Internet at www.lis.ceps.lu. Some examples of these studies include Förster (1993), Smeeding et al. (1990), Kenworthy (1998), Jäntti and Danziger (2000), and Smeeding et al. (2000), many of which are directly available online from the LIS web page.

The measure of poverty in rich nations involves the comparison of some index of household well-being with household needs. When command over economic resources falls short of needs, a household (or person or family) is classified as poor. Well-being refers to the material resources available to a household. The concern with these resources among most social scientists, including sociologists, is not with material

consumption, *per se*, but rather with the capabilities such resources give to household members so that they can participate fully in society (Sen 1983, 1992). These capabilities are inputs to social activities, and participation in social activities gives rise to a particular level of well-being (Coleman and Rainwater 1978, Rainwater 1990). Methods for measuring a person's or household's capabilities differ according to the context in which one assesses them, either over time or across nations or among subpopulations within a nation.

Most sociologists recognize that all advanced societies are highly stratified (e.g., Townsend 1979). Hence, some individuals have more resources than others. The opportunities for social participation are affected by the resources that a household disposes, particularly in nations like the United States where there is heavy reliance on the market to provide such essential services as health care, post-secondary education, and child care. Money income is therefore a crucial resource. Of course, there are other important kinds of resources, such as social capital, noncash benefits, primary education, and access to basic health care, all of which add to human capabilities (Coleman 1988). These resources may be available more or less equally to all people in some societies, regardless of their money incomes.

Sociologists believe that there are many forces in rich societies that reduce well-being by limiting capabilities for full participation in society, including inadequacies in neighborhoods where people live. Sociologists are therefore interested in racial discrimination, neighborhood violence, low quality public schools, lack of good jobs, and job instability, all of which increase economic insecurity, reduce human capabilities, and increase poverty (e.g., Massey 1996).

2.1 *Absolute vs. Relative Poverty*

An absolute poverty standard is defined in terms of a level of purchasing power that is sufficient to buy a fixed bundle of basic necessities. A relative standard, on the other hand, is defined relative to the typical income or consumption level in the wider society. The purchasing power of a relative poverty standard will change over time as society-wide income or consumption levels change; an absolute poverty standard will change only with the prices of commodities. Most sociologists inherently favor the relative definition of poverty.

And, in fact, all measures of poverty or economic need are relative, because context is important to the definition of needs. The World Bank uses poverty measures of \$1 to \$2 per person per day—or \$1,095 to \$2,190 per year for a family of three—for the developing nations of Africa or Latin America (Ravallion 1996). In contrast, the 1998 United States, 'absolute' poverty threshold was \$13,003 for a family of three—about 6 to 12 times the World Bank's poverty

line—while one-half of median income, the preferred relative poverty standard, was another 25 percent above this poverty line or 8–15 times the poverty standard in poor countries. Moreover, as economic inequality has increased in most rich societies over the past 20 years, the study of relative deprivation and poverty has taken on a new life (Gottschalk and Smeeding 2000). The link between marginalization, relative poverty, and the globalization of trade is just beginning to emerge as a topic of prime interest among sociologists (e.g., Alber and Standing 2000).

In fact, cross-national comparisons of poverty in rich countries rely almost exclusively on a relative concept of poverty. Cross-national studies typically compare the percentage of persons living with income below some fraction, normally half, of the national median income. This kind of comparison is consistent with a well-established theoretical perspective on poverty (Townsend 1979, Sen 1983, 1992). However, some feel that such a standard is too low. For instance, the European Statistical Office Working Group on Poverty Measurement recently recommended using 60 percent of the national median income as a common poverty threshold for European Community poverty studies in the new millennium (Eurostat 2000).

2.2 *Alternative Measures of Resource Poverty*

Because there is no single commonly accepted way to measure poverty among sociologists, there is a desire to go beyond the popularly used income poverty definition. As a result, there exists a wide variety of additional poverty measures that substitute for or complement the preponderance of income-based measures used by quantitative sociologists and economists (e.g., Ruggles 1990, Haveman and Mullikin 1999, Boltvinik 2000). In principle, poverty is a multi-dimensional concept and should reflect several aspects of personal well-being. Forms of deprivation other than economic hardship can certainly be relevant to poverty measurement and to anti-poverty policy-making.

Even restricting consideration to economic poverty, measures other than annual income might be used to estimate household resources. For households with fluctuating incomes, annual income can provide a poor approximation for permanent or long-run average income. Some social scientists suggest that annual consumption represents a better proxy for permanent income and hence should be preferred over annual income for measuring household resources (Haveman and Mullikin 1999). However, broad-based consumption data are more difficult to come by than are income data. A number of authors have suggested that separate measures of needs ought to be developed for different goods and services (Aaron 1986). Housing and health care are often mentioned in this context, while the former is of much greater interest in the United Kingdom (DSS 1993).

2.3 Social Indicators

Social indicators present an entirely different method of poverty measurement. Researchers have a long tradition of using social indicators to measure both economic hardship and affluence. Analysts often define direct measures of deprivation, such as lack of food, heat, or access to health care (Bauman 1998, Mayer and Jencks 1993). Because each household's need is imperfectly measured by an income or consumption threshold, many households with incomes above the thresholds may still experience hardships, such as hunger, or face high rent payments or heavy medical expenses that prevent its members from buying enough food or fuel. Further, individual members of a household may suffer hardship because of unequal sharing of incomes within the household. In these cases, direct measures of material hardship might offer a useful indication of poverty, especially in a cross-national context, and many sociologists are therefore attracted to such measures (Mayer and Jencks 1993).

2.4 Time Poor

Finally, resources other than access to material goods might be important determinants of poverty. The literature in the 'time poor' has grown in recent years after a nascent period. The argument is that time is often the scarcest commodity in market work-oriented societies. Further, time spent on home production is often undervalued or not valued at all. The gender bias in time use is therefore at the heart of these types of analyses (Vickery 1997, Douthitt 1994). In particular, recent research has focused on the time crunch on single parents who must both earn a living in the labor market and care for their children (Folbre 1994, England and Folbre 1999).

3. Summary

Sociologists and other social scientists have used a variety of methods to define and measure poverty. Each approach has its strengths and weaknesses. Over the past few decades, income-based poverty measures using annual assessments of household income have provided the main statistics for tracking poverty over time, across nations, and across sub-populations within nations for sociologists and for other social scientists. They have also provided important tools for testing causal hypotheses concerning the explanations, causes, and cures for poverty, including place, space, race, gender, and other popular theories of poverty in sociology. They are complemented by social indicator-based measures of inadequacy, and by the adequacy of time for market work, home work, and the nurturing and care of family members. Sociological studies of

poverty have both a rich heritage and have become an important area of research in increasingly unequal rich societies.

See also: Equality and Inequality: Legal Aspects; Income Distribution; Inequality; Inequality: Comparative Aspects; Poverty and Gender in Affluent Nations; Poverty and Gender in Developing Nations; Poverty, Culture of; Poverty in History; Poverty Law; Poverty Policy; Social Stratification; Wealth Distribution

Bibliography

- Aaron H 1986 Comments on methods of measuring noncash benefits. Reported. In: *Noncash Benefits: An Evaluation of the Census Bureau's Measurement Conference*. US General Accounting Office, Washington, DC
- Alber J, Standing G 2000 Social dumping, catch-up or convergence? Europe in a comparative global context. *Journal of European Social Policy* 10: 99–119
- Bauman K J 1998 *Direct Measures of Poverty as Indicators of Economic Need: Evidence from the Survey of Income and Program Participation*. Population Division Technical Working Paper No. 30. US Bureau of the Census, Washington, DC
- Björklund A, Freeman R 1997 Generating equality and eliminating poverty—the Swedish way. In: Freeman R B, Topel R, Swedenborg B (eds.) *The Welfare State in Transition: Reforming the Swedish Model*. University of Chicago Press, Chicago
- Boltvinik J 2000 *Poverty Measurement Methods: An Overview*. United Nations Development Programme, SEPED Series on Poverty Reduction, United Nations, New York
- Casper L M, McLanahan S S, Garfinkel I 1994 The gender-poverty gap: What we can learn from other countries. *American Sociological Review* 59: 594–605
- Christopher K, England P, McLanahan S, Ross K, Smeeding T 2001 Gender inequality in poverty in affluent nations: The role of single motherhood and the state. In: Vleminckx K, Smeeding T (eds.) *Child Well-being, Child Poverty and Child Policy in Modern Nations: What Do We Know?* Policy Press, Bristol, UK, pp. 199–220
- Coleman J S 1988 Social capital in the creation of human capital. *American Journal of Sociology* 94 (Suppl. 5): S95–S120
- Coleman J, Rainwater L 1978 *Social Standing in America: New Dimensions of Class*. Basic Books, New York
- Department of Social Security 1993 *Households Below Average Income, A Statistical Analysis, 1979–1993/94*. Government Statistical Service, London, UK
- Douthitt R A 1994 *Time to do Chores? Factoring Home-production Needs into Measures of Poverty*. Discussion Paper No. 1030–94, Institute for Research on Poverty, University of Wisconsin, Madison, WI
- England P, Folbre N 1999 Who should pay for the kids? *Annals of the American Academy of Political and Social Science* 563: 194–207
- Eurostat 1998 *Recommendations of the Task Force on Statistics on Social Exclusion and Poverty*. European Statistical Office, Luxembourg
- Eurostat 2000 *Report of the Working Group: Statistics on Income, Social Exclusion and Poverty*. European Statistical Office, Luxembourg

- Folbre N 1994 *Who Pays for the Kids? Gender and the Structures of Constraint*. Routledge, New York
- Förster M 1993 Comparing poverty in 13 OECD countries: Traditional and synthetic approaches. *Studies in Social Policy* #10. OECD, Paris
- Gans H 1972 The positive functions of poverty. *American Journal of Sociology* 78(2): 275–89
- Glennester H, Lupton R, Noden P, Power A 1999 *Poverty, Social Exclusion and Neighborhood: Studying The Area Bases of Social Exclusion*. CASE paper 22, Neighbourhood Research Centre for the Analysis of Social Exclusion, London School of Economics, London UK
- Gottschalk P, Smeeding T M 2000 Empirical evidence on income inequality in industrialized countries. In: Atkinson A B, Bourguignon F (eds.) *Handbook of Income Distribution*. North Holland, New York, pp. 261–308
- Harrington M 1962 *The Other America: Poverty in the United States*. Macmillan, New York
- Haveman R, Mullikin M 1999 *Alternatives to the Official Poverty Measure*. University of Wisconsin, Madison, WI. Presented at a Conference on 'Poverty: Improving the definition after 30 years'
- Hills J 1999 *Social Exclusion, Income Dynamics and Public Policy*. Annual Sir Charles Carter Lecture, Northern Ireland Economic Council, Report No. 129, Belfast, UK
- Jäntti M, Danziger S 2000 Income poverty in advanced countries. In: Atkinson A B, Bourguignon F (eds.) *Handbook of Income Distribution*. North Holland, Amsterdam, pp. 309–378
- Kenworthy L 1998 *Do Social-welfare Policies Reduce Poverty? A Cross-national Assessment*. Luxembourg Income Study Working Paper #188, Center for Policy Research, Syracuse University, Syracuse, NY
- Massey D 1996 The age of extremes: Concentrated poverty and affluence in the 21st century. *Demography* 33: 395–412
- Massey D S, Denton N A 1993 *American Apartheid: Segregation and the Making of the Underclass*. Harvard University Press, Cambridge, MA
- Mayer S E, Jenks C 1993 Recent trends in economic inequality in the United States: Income versus expenditures versus material well-being. In: Papadimitriou D B, Wolff E N (eds.) *Poverty and Prosperity in the USA in the Late Twentieth Century*. St. Martins Press, New York
- McLanahan S, Sorensen A, Watson D 1989 Sex differences in poverty 1950–1980. *Signs: Journal of Women in Culture and Society* 15: 102–22
- Miller S M 1996 The great chain of poverty explanations. In: Øyen E, Miller S M, Samrod S A (eds.) *Poverty: A Global View*. Scandinavian University Press, Oslo, Norway, pp. 569–586
- Mincy R, Sawhill I, Wolf D 1990 The underclass: Definition and measurement. *Science* 248: 450–3
- Organization for Economic Cooperation and Development 2000 *Trends and Driving Factors in Income Distribution and Poverty in the OECD Area*. Social Policies Studies Division Occasional Paper #42, OECD, Paris
- Piachaud D 1987 The distribution of income and work. *Oxford Review of Economic Policy* 3: 41–61
- Rainwater L 1990 *Poverty and Equivalence as Social Constructions*. Luxembourg Income Study Working Paper #55, Center for Policy Research, Syracuse University, Syracuse, NY
- Ravallion M 1996 *Economic Journal* 106: 1328–44
- Ringen S 1985 Toward a third stage in the measurement of poverty. *Acta Sociologica* 28: 99–113
- Room G J 1999 Social exclusion, solidarity and the challenge of globalization. *International Journal of Social Welfare* 8: 166–74
- Ruggles P 1990 *Drawing the Line: Alternative Poverty Measures and their Implications for Public Policy*. Urban Institute Press, Washington, DC
- Sen A 1983 Poor, relatively speaking. *Oxford Economic Papers* 25: 153–169
- Sen A 1992 *Inequality Reexamined*. Harvard University Press, Cambridge, MA
- Smeeding T M, O'Higgins M, Rainwater L 1990 *Poverty, Inequality and the Distribution of Income in a Comparative Context: The Luxembourg Income Study (LIS)*. Wheatshaeaf Books, London, UK and Urban Institute Press, Washington, DC
- Smeeding T, Rainwater L, Burtless G 2000 *United States Poverty in a Cross-national Context*. Center for Policy Research, Luxembourg Income Study Working Paper #244, Syracuse University, Syracuse, NY
- Townsend P 1979 *Poverty in the United Kingdom: A Survey of Household Resources and Standards of Living*. University of California Press, Berkeley, CA
- Townsend P 1993 *The International Analysis of Poverty*. Harvester Wheatsheaf, London, UK
- UNICEF Innocenti Research Centre 2000 *A League Table of Child Poverty in Rich Nations*. Innocenti Report Card No. 1, UNICEF, Florence, Italy
- United Nations Development Programme 1999 *Human Development Report 1999: Globalization With a Human Face*. United Nations, New York
- US Bureau of the Census 1999 Poverty in the United States: 1998. *Current Population Reports* P60–207, US Government Printing Office, Washington, DC
- Vickery C 1997 The time-poor: A new look at poverty. *Journal of Human Resources* 12: 27–48
- Wilson J 1987 *The Truly Disadvantaged: The Inner City, the Underclass and Public Policy*. University of Chicago Press, Chicago

T. M. Smeeding

Power in Society

Power has been seen as an 'essentially contested' concept (Lukes 1974). Some of the major recent contours of this contestation will be addressed in this entry.

1. Weber: Ideal Types of Rule and Power

At the core of modernism, according to Weber, lies the rationalization of various institutional arenas such as the market, technology, law, state, the growing pervasiveness of rational calculation in all spheres of life, and the disenchantment of more spiritual or superstitious forms of calculation. Disenchantment means exposure to a world in which meaningfulness is never

Table 1
Power/authority dynamics

	Versus tradition	Versus charisma	Versus rational-legality
Tradition	Rival dynasties	Dynasty confronting a revolutionary hero	Privilege ranged against rights according to law
Charisma	Princess Diana versus the British Royal Family	Hitler against Stalin during World War II	Nelson Mandela against De Klerk's Apartheid Government of South Africa
Rational-legality	Oliver Cromwell leading the Parliament against the monarch, Charles I	The Marcos Government in the Phillipines against Cory Aquino	Two governments disputing each others claim to the same territory, such as India and Pakistan over Kashmir

'given.' Instead, meanings clash, struggle against the resistance of others and are never wholly secure. Rational action consists precisely in the capacity to respond to the uncertainties of a world without meaning. Analytically, he conceived of three main types of rule around which rational action might revolve: principles of charisma, tradition, or rational legality. Each of these three types is a source of a particular type of *Herrschaft* (authority, or in better translation, domination).

In Weber's view, rationality had only a limited number of sources. It might derive from those few individuals who choose to impose value and meaning upon the world by confronting the world as something other than resolutely unyieldingly, absolutely and unbreakably cast in a mold. There were precious few such people in Weber's vision, other than charismatic leaders. Charisma is often identified with historical figures regarded as ordained with exceptional grace and a sense of calling that they were able to transmit and project to others, such as Jesus. Charisma is not a purely historical category, however, but it is only the rare charismatic leader, whose personal grace, sense of calling, duty, and devotion to some ultimate ideal is unshakeable, who would be able to impose such meaning on life. Hence, charisma was one of the key bases of authority that Weber identified. Charisma is typically something above the grubby theater of mundane political wheeling and dealing. In terms of modern life we might think of celebrities such as Princess Diana who were able to parlay their celebrity into a passionate social advocacy on behalf of the victims of landmines and in so doing, effect substantial changes in government policies. Here was a charismatic celebrity for the modern age—especially in her death. Clearly, in modern times, celebrity and charisma are closely related but by no means the same. Often they become confused: the cult of the personality that modern leaders like Mao, Stalin, and Hitler cultivated were deliberate attempts to use tight control of the state to fashion and stage-manage a celebrity that transcended the political sphere and which sought to invest them with charisma.

Tradition functions as another ideal type, best seen in claims to authority based on lineage or dynasty, in principles of 'blood and breeding' or exclusion from opportunities to claim authority because of the lack of such blood and breeding. The succession principles of monarchy, or inheritance rights attached to property, are prime examples of rule by tradition. Most pervasive, however, as a form of authority, was what Weber took as the quintessential modality of modern rule: rational-legal bureaucracy, where rule is conducted according to abstract principles. The dominance of what Weber referred to as the rational-legal does not overwhelm all other bases for rule and authority. However, its pervasiveness is often summed up today in a popular fear of 'economic rationalism,' referring to those situations where only a legally mandated economic calculus is applied to policy formulation (such as competition policy), without regard for other ways of accounting for the impact of such policies on the human spirit, communities, and the fabric of society.

Each principle of rule posits a particular type of authority and it is a simple matter to construct a matrix of these relations in order to generate some idea of what power might be. Weber defined power in terms of social action that is goal-oriented, even against the resistance of others. An evident example of power thus arises when there are cases of conflict between principles of rule (see Table 1). Here, power is seen in terms of the opposed principles or ultimate values.

2. Power as a Generic Concept

While Weber's definition of power is quite generic, the concept need not be premised on the conflict of rival, ultimate values, as per the examples of Table 1. Steven Lukes (1974), for instance, in one of the most influential books on the subject, regarded Weber's definition as essentially similar to the generic definition of power contained in the work of Robert Dahl (1957). Here power is defined as the situation that prevails when A gets a B to do something that they would not

otherwise do. This sounds a methodologically sparse and modern behavioral definition but it is one that reaches back into the earliest development of the modern era—to Thomas Hobbes's conception of power in a causal, mechanical world.

The Community Power Debate that formed around Dahl's (1961) work concerned the distribution of power in postwar USA. Dahl (1961) insisted that such questions could only be answered empirically, rather than through a broad ranging essaying of concepts that were then current, such as their being a 'military-industrial complex.' Additionally, he thought them better answered at the local, community level rather than at the national level. To determine the distribution of power required an operational definition for empirical investigation. For Dahl, power was indissoluble from its exercise: Someone had to be seen doing something to another for us to say that power had been exerted. Unlike Weber, he was not concerned with ultimate values but he did think it important that the exercise of power should be determined by competition over issues that people thought important. Basically, he sought to determine what the key issues were in a community and then establish the outcome of the politics surrounding these decisions, to see who gained or yielded power.

3. *Power, Real Interests and Hegemony*

By the time that Lukes (1974) published his book on *Power: A Radical View*, Dahl's approach had already been criticized by Bachrach and Baratz (1970) for leaving out of consideration 'what does not happen.' Nonissues, nondecision making, the construction of agendas, and the mobilization of bias through everyday routines, had all been introduced to show how power is accomplished through covert means, through inaction, rather than through overt exercise of causal power.

Lukes took the critique one step further. He argued that the critics had only sought to qualify Dahl's behaviorism rather than transcend it through radical critique. What was required, he argued, was a conception of how power worked not only against people's expressed and implicit preferences but also their 'real interests.'

The notion of real interests came from Marxist discourse—originally designed to address situations of 'false consciousness'—where a class of people were unable to recognize that the system within which they were embedded was the source of their entrapment. Lukes followed the Italian Marxist theorist, *Antonio Gramsci*, in introducing a 'cultural spin' with the notion of hegemony. Where people were deluded about what they wanted and what they might authentically be, because of their inability to think in concepts other than those of domination and the

dominated, then they suffered from hegemony. The basic idea was that they were subject to a 'ruling culture' disseminated through the media, the state, education, and other institutional arenas.

Theories of hegemony became variations upon the theme of working-class acquiescence secured through normative incorporation by dominant ideology (Abercrombie et al. 1980). The narrative was one of concealment: Marxist theory, in theory, should have provided revelation—but it did not. Only the theorist could see the truth of the mechanisms that, despite and even through resistance, secured and reproduced consent. The subjects of analysis had the nature of their real interests concealed from them. Western Marxism, the source of these insights, was impotent at best and at worst irrelevant, in lifting the veils of occlusion (Anderson 1976). In the West socialism became increasingly utopian (Bauman 1977); in the East increasingly dystopian (Fehér et al. 1983), as the collapse of 'dictatorship over needs' produced its implosion.

4. *The Discursive Turn*

The ideological arguments, with their fundamental denial of consumer sovereignty, receded in importance in the 1980s. One response to this situation within the Marxist camp came from Laclau and Mouffe 1985, who instead of thinking of hegemony as a state of mind regarded it as a set of practices. These practices were primarily discursive. Hegemony thus came to be seen not in terms of its substantive content but its form. Hegemony existed when the indefinite possibilities of signifying elements and their relations were strategically foreclosed (Laclau and Mouffe 1985). Again, the emphasis is on things that do not happen. Any stable set of articulations is thus hegemonic: Only permanent Cultural Revolution would not be—although the implications are not spelled out.

Real interests are what radical politics and theories claim to uncover and reveal. One way of doing this, suggested Haugaard (1998), might be to explain to actors the way in which their structured practices feed into the reinforcement of a regime of domination contrary to interests already known to them. What may be at issue are structural and institutional relations of domination rather than real but unknown interests. Once actors reflect upon their practical consciousness then they enjoy relative autonomy. Power thus consists not so much in the suppression of real interests but the denial of the relative autonomy required for reflection (Haugaard 1998, p. 141).

Haugaard (1998) notes that conceptions of interests form discursively, even to the extent that one can persuade some people that they have real but unrealized interests blocked by their present conditions of existence. Such discursive realities help shape them to

resist the sources of domination identified. Politics mobilize power and resistance through discursive expression of existing, past, or future conditions of existence. Thus, resistance is not endemic but depends on the practical consciousness of people and the categorical devices through which they lead their everyday lives in relation with the discursive consciousness proffered by various media of representation—parties, press, radio, television, the Internet, public intellectuals. Only in rare moments is such diversity represented or resisted in the name of some great unifying narrative or theme. Most probably these will be those relatively rare moments where a privileged identity is posited as a historical subject: the subject people, the subject party, the subject race, the subject nation, the subject church, the subject gender, the subject sexuality, the subject ecology, and so on.

A plurality of life-worlds, with interpretative resources, and translation devices for switching between them, minimizes the chances of hegemony. People from sophisticated cultures routinely have a multiplicity of interpretative horizons available to them, as a result of exposure to different life-worlds through various media, educational experiences, travel or migration, or interpersonal relations. Of course, to the extent that these experiences are compartmentalized, one way or another, then the ability to achieve critical distance from any posited identity is diminished.

The theorization of autonomy in terms of moves between interpretative horizons means that the ability of actors to transcend their social environments and structured context does not presuppose an ability to 'escape' the processes of socialization by adopting a neutral, culturally unbiased, position. Autonomy does not involve a flight into any transcendental realism of absolute Truth, nor does it involve the transcending of false consciousness. Critical faculty does not imply the objectivity of truth, nor does it entail an undersocialized concept of agency whereby autonomy is gained by transcending the basis of social culture. It requires only relative reflexivity on the part of subjects. Such a conception of the relation of power and knowledge, with its emphasis on conditions of relative autonomy, turns out to be a way of re-approaching the classical questions about power and community that were at the center of Dahl's work. The difference is that, this time, the approach has been through Foucault rather than Dahl.

5. Foucault and Power

Foucault provides a detailed history of some power practices and techniques that have characterized modernity in a way that is nonreductive and complements Weber (Dandekker 1991), although he understates the importance of Max Weber's (1978) argument concerning discipline. Nonetheless, there are common

themes (see Szakolczai 1998). Different emphases on power occur through the course of Foucault's work. First, there is his work on genealogy as the natural history of specific practices. Second, the work on the 'care of the self,' where histories are used to critique the present, by uncovering the constituted nature of what have been taken as the here-and-now; third, the discussion of government or 'governmentality.'

In Foucault's (1979) most influential genealogical model of 'surveillance and discipline,' power is constituted in an intimate relationship with liberty. Because the exercise of power presumes a free subject, in a well-known formula, there will always be resistance. Where liberty is extremely limited, because power is stable and hierarchical, then Foucault terms it domination. (The parallelism of this account with that of Weber's 1978 discussion of *Herrschaft*, is evident.) Two models of power are compared and contrasted: the repressive and the positive. (Foucault's 1979 concern with the facilitative and productive aspects of power relates to the kind of arguments made most clearly by Parsons 1963.) In general Foucault does not offer a critique of power: It would be impossible to do so because, to coin a phrase, power has no other: It is inescapable. Liberation from power is only a dream (or a nightmare). However, specific forms of power may and do change. The 'punitive' power of corporal excess may be contrasted with the 'positive' power of bureaucracy. However, there is no space outside of power, no knowledge that is pure truth. (In this way are utopia and dystopia avoided.) In reading Foucault it is necessary to uncouple a sociological from a philosophical question concerning truth. The former deals with what passes for truth rather than asking what is truth—with how truth-effects are produced, in the Foucauldian vernacular. It is the latter that concerns Foucault, rather than the former.

In Foucault's model of power relationships through government—the structuring of action through conduct—which he defines as the capacity of people to regulate their own behavior through various rationalities, his treatment of liberalism suggests that it is conditions of relative autonomy that define the nature of political community.

6. Community Politics after Foucault

Foucault suggests that democracy is neither an ideal with which to measure actual practice nor a method for producing utopias. It is the messy and politically expeditious ways of making sense of politics its participants routinely use. It is this fundamental approach that Flyvberg (1998) takes from Foucault. He provides a deeply documented, rich narrative of one complex set of issues and processes associated with the many ways in which political communities were constructed and deconstructed in the laborious

pursuit of local interests and issues surrounding a proposed traffic-plan for the city of Aalborg in Denmark.

Aalborg was not unusual in respect of its power relations: They were sufficiently asymmetric as to verge on domination. Here, the authority of technical rationality was a weapon of the politically weak—those without other resources at their disposal. Hence, not surprisingly, technically rational plans were constantly undercut and attempts at democracy secured on that technically rational basis proved weak. Neither a normative emphasis on democracy nor rationality proved sufficient to avoid domination by superior power. Such power defined reality and, not surprisingly, the rationality it accepted depended on power rather than legality. One of the main strategies of the dominant group was to use techniques of rationalization as power: Indeed, the greater the power the lesser the rationality. Thus, we might conclude that power is more ordinarily stable than conflictual and that its relations are constantly being produced and reproduced in process. Where power is stable then power relations are more characteristically rational.

7. Conclusion

In conclusion, we may note that in the genealogy of power that has been constructed here, power is not defined as only one thing, or in one set of terms. The terms shift with debate but there is a family resemblance among them. The emphases have changed with the values underlying scholarship. Lukes' importation of the Gramscian notion of hegemony seemed to take us a long way from the political communities that Dahl studied with such methodological scruples. Laclau and Mouffe's introduction of Foucault to the debate, somewhat surprisingly, brings debate back full circle to issues of political community because of the centrality, their conception of power placed on discourse and identity. The theme of the relative autonomy of political processes dominated the late work of Foucault and was taken up empirically by Flyvberg. Thus, nearly 50 years after the Community Power Debate first kicked off contemporary sociological interest in the conceptual relation of power and society, the concepts have come back to their empirical point of departure—but much richer and more sophisticated than when they first embarked on that journey.

See also: Authority, Social Theories of; Charisma: Social Aspects of; Democracy; Foucault, Michel (1926–84); Political Sociology; Power: Political; Weber, Max (1864–1920)

Bibliography

Abercrombie N, Hill S, Turner B S 1980 *The Dominant Ideology Thesis*. Allen and Unwin, London

- Anderson 1976 *Considerations on Western Marxism*. Verso, London
- Bachrach P, Baratz M S 1970 *Power and Poverty: Theory and Practice*. Oxford University Press, Oxford, UK
- Bauman Z 1977 *Socialism: The Active Utopia*. Allen & Unwin, London
- Clegg S R 1990 *Modern Organizations: Organization Studies in the Postmodern World*. Sage, London
- Dahl R A 1957 The concept of power. *Behavioural Science* 2: 201–5
- Dahl R A 1961 *Who Governs? Democracy and Power in an American City*. Yale University Press, New Haven, CT
- Dandekker C 1991 *Surveillance, Power and Modernity*. Polity Press, Cambridge, UK
- Fehér F, Heller A, Márkus G 1983 *Dictatorship Over Needs*. Blackwell, Oxford, UK
- Flyvberg B 1998 *Rationality & Power: Democracy in Practice*. Chicago University Press, Chicago
- Foucault M 1979 *Discipline and Punish*. Penguin, Harmondsworth, UK
- Haugaard M 1998 *The Constitution of Power*. Manchester University Press, Manchester, UK
- Laclau E, Mouffe C 1985 *Hegemony and Socialist Strategy*. Verso, London
- Lukes S 1974 *Power: A Radical View*. Macmillan, London
- Parsons T 1963 On the concept of political power. *Proceedings of the American Philosophical Society* 107: 232–62
- Szakolczai A 1998 *Max Weber and Michel Foucault: Parallel Life-Works*. Routledge, London
- Weber M 1978 *Economy and Society*. University of California Press, Berkeley, CA

S. R. Clegg

Power: Political

Leading political theorists, such as Machiavelli and Hobbes, have devoted much attention to 'political power.' As Robert A. Dahl noted (1957), 'power' has a certain universality, whether it is called 'power, influence, control, *pouvoir*, *puissance*, *Macht*, *Herrschaft*, *Gewalt*, *imperium*, *potestas*, *auctoritas*, *potentia*,' etc. This article focuses on the concept of 'power' defined as a type of social causation.

1. Power as Causation

From 1953 to 1975, several leading political scientists developed the view of power as social causation, converging with an earlier viewpoint of Max Weber. After rejecting alternative definitions of 'power,' Herbert Simon (1953) cited a definition by Lasswell and Kaplan (1950): 'The exercise of influence (influence process) consists in affecting policies of others than the self.'

In the formulation best known to political scientists, Dahl stated (1957): 'My intuitive idea of power, then, is something like this: A has power over B to the extent

that he can get B to do something that B would not otherwise do.' James G. March (1955) stated: '... the set of all influence relations is here defined to be that subset of all causal relations such that the behavior of an individual appears as the terminal point in the causal linkage. Alternatively, we can say that two individuals are in an influence relation if their behaviors are linked causally.'

This leads to distinctions among different types of causal relations as they might relate to power. First, 'political power' refers to causal relations among humans, not to inanimate forces changing human behavior, as a hurricane might, for instance. Second, a distinction causing much confusion, one needs to distinguish situations in which the changed behavior reflects the intentions of the powerful actor. One might change the behavior of another in a way not intended, as when an incompetent lobbyist changes a legislator's vote to one different from that favored by the lobbyist. Some political scientists, such as March, are more comfortable with 'power' as causing changes in the direction intended by the powerful actor, and referring to the more general category of causing any changes as 'influence.' However, the seminal articles by Simon and Dahl do not restrict 'power' to intended changes in others' behavior.

Other examples of such causal relations would include power-as-suggestion: when A proposes a behavior change, B adopts the change without resistance. Another category of causal relations occurs when B resists, and A exercises rewards and sanctions, perhaps even force, to get B to change his or her behavior. This type of power may be called coercion (McFarland 1969).

In regard to power as coercive causation, John Harsanyi (1962) emphasized the costs of exercising power and also the costs of compliance with the exercise of power. For instance, a lobbyist might change a legislator's vote with little cost if it just means arranging a meeting with one of the legislator's staff. However, the cost is high if the lobbyist must mobilize hundreds of persons to petition the recalcitrant legislator. From the standpoint of the legislator, a vote may be 'cheap,' offending almost no one, or it may endanger victory in the next election. Still another causal situation occurs when A's preferences cause changes in B's preferences, causing B to act differently. This is widely known as 'anticipated reactions,' in which B, perhaps to please A and to avoid A's exercise of sanctions, changes behavior in line with an anticipation of what A wants him or her to do, even though A did not explicitly state such preferences (Nagel 1975).

An important, implicit aspect of defining power as causation is its link to empiricism. A statement of power then implies observations of instances of A changing B's behavior. It rules out defining 'power' in terms of the resources (wealth or force, etc.) used to get B to comply. One first has to observe the

causal relationship; then one analyzes how power was wielded.

Such resources were called by Dahl (1957) a 'power base,' and both he and Simon (1953) stated that the power base should not be confused with 'power' itself. In addition, Dahl stated that power-as-causation, to be an empirically useful concept, had to be used with four concomitant dimensions of power: base, amount, scope, and domain. *Base*, then, refers to the resources or means that A uses to cause changes in others' behavior. *Amount* refers to the idea that some instances of power refer to greater changes in behavior than others. A lobbyist exercises more power when he or she changes the legislator's vote from 'no' to 'yes,' for instance, than if the change is only from 'abstain' to 'yes.' The *domain* of an actor's power consists of those persons subject to the actor's power. The *scope* of power consists of the matters subject to the actor's power.

Dahl's major contribution to the study of political power is his insistence that the domain and scope always be clearly stated in discussions of power. Or to put it another way, an actor's power in one situation should not be immediately generalized to postulating equivalent power over other persons in other situations. The pharmaceutical lobbyist may have power over just one or over a score of legislators regarding the regulation of generic drugs. The same lobbyist might not have equivalent power over votes about Medicare payments for prescriptions; this lobbyist may have no power at all over any legislators in regard to civil rights or to gun control.

Scope and domain of power were key issues in the still widely remembered community power debate of the 1960s. Following the example set by Dahl in *Who Governs?* (1961), political scientists studied political power in localities by conducting case studies of public policy-making and charting which persons changed the course of events. Such studies normally found that the domain and scope of power was fragmented; different persons had power in different areas of policy. On the one hand, other social scientists—tending to be sociologists—followed a different conception of power, conceiving power to be possession of wealth or high social status. Such observers usually described inequalities of wealth and status and stated this as a finding of 'power' in a community. Political scientists, on the other hand, for the most part did not consider wealth and status to be 'power' itself, but instead to be resources that could be used to win political power, if such wealthy or high-status persons intended to do so. But such intentions are an empirical question, in this view, and must be observed in case studies.

This famous (or notorious) controversy extended for at least a decade and was expressed in scores of books and hundreds of journal articles before the frustrated debate participants, unable to persuade one another, for the most part went on to study questions other than community power (Polsby 1980).

Proponents of power-as-resource generally overlook a strong criticism of the power-as-causation model. Social causation is ordinarily very complex; case studies are not easy to do; power as causation thus is ordinarily complex, requiring a great deal of effort to study.

Such writers as Simon (1957), March (1955), Harsanyi (1962), and Riker (1964) left us with elaborate statistical models for studying power, models which require costly and difficult data collection. Jack Nagel (1975) wrote a concluding work in this line of analysis, in which power-as-causation was stated as path analysis, a form of statistics.

2. Power and Policy-making

March (1966) and Riker (1964) were both noted for asking the question in respect to political analysis: 'what is the power of power?' Political scientists answered this question, not by flocking to do more studies of the holders of wealth and status, but by stating models of power structure applicable to political systems, both national and local. Dahl (1961) had pressed forward the definition of power as causation in the course of his empirical refutation of the power elite model (Mills 1956), as applied by Floyd Hunter (1953) to the city of Atlanta. Such writers were among those who described power by describing characteristics and activities of an elite group, whose actual exercise of power usually went undemonstrated. Scholars such as Dahl became known as stating a *pluralist* model of power, finding power to be fragmented among various scopes and domains, as opposed to the concentration of power within a single elite.

Scholars such as Edelman (1964), McConnell (1966), and Lowi (1979) agreed with Dahl that political power in America is fragmented, but argued that stable, unrepresentative elites often controlled policy within specific scopes of power—such as setting grazing fees on federal lands, developing nuclear power plants, setting airline and interstate trucking schedules and fees, and so forth. Such analyses, which flourished briefly, were also based on case studies of policy-making, and constitute a school which might be termed *multiple-elitist*, finding not a single power elite but hundreds of elites, each specific to a particular scope of action and not subject to control by elected officials. But multiple-elitism in turn faced a revisionist interpretation, *neopluralism*, as found in studies by James Q. Wilson (1980) and Jack Walker (1991), for instance, who found that while power within specific scopes of policy was normally unequal, it was nevertheless shared by so many organized interest groups, politicians, and governmental agencies that even within narrowly defined policy scopes, stable and controlling elites are seldom found.

The power-as-causation model was developed from 1953 to 1975, and the examples cited in this literature were almost always of power among individual persons. But given the great difficulties in gathering such data, political scientists seldom conducted such studies, and instead the analysis of political power flourished with the development of models of public policy-making just cited. Such scholars saw no need to worry about possible criticisms of ignoring the power bases of wealth and status, and continued to conduct case studies of policy-making, following Dahl's example, even as many argued with Dahl's pluralist model. Such case studies were not conducted according to the canons of path analysis or other forms of statistics; instead researchers used historical materials, interviews, and current documents to ascertain which groups, leaders, and government agencies caused changes in the course of public policy. Relatively simple empirical methods, assuming power-as-causation, have been enough to generate interesting models of power structure, but perhaps more complex methods will be needed to resolve differences among various schools of political power analysis.

The study of political power has developed satisfactorily since Simon's seminal article in 1953, but a great theoretical difficulty remains—the problem of *hegemony*. Hegemonic writers stress the great importance of the power to determine which issues are on the political agenda of a political system (Lukes 1974). Community power theorists Bachrach and Baratz (1962) argued that there are 'two faces of power,' one of which is power over the formation of political issues, and the other is the overt activity studied by Dahl. In arguing for the understanding of hegemony, the Marxist Antonio Gramsci argued (1971) that capitalism maintained its political power through socialization of persons into supportive values and norms, as much as through the machinations of 'bourgeois government.' Social values shape the preferences of the actors in their search for political power, but how can the scholar make empirical statements about one set of values displacing perhaps hundreds of conceivable alternative value systems? This seems to presuppose a comprehensive theory of social relations. Hegemony can be distinguished from the power of *anticipated reactions*, power exercised when B anticipates A's intentions, and acts according to them, perhaps fearing the costs of A's overt coercion. Republican members of Congress may have considered blocking appropriations for a government department, but anticipating President Clinton's effective reactions, they think better of the idea and do little but discuss the matter. But since the matter received some discussion, the theorist of political power may make an empirical observation. Similarly, in international relations, we speak of the deterrent power of weapons, even if these are never used. From the behavior of those deterred, scholars can find evidence that some probability of warlike behavior by one

nation was deterred (a change of behavior) through the possession of weapons by the other nation.

See also: Agendas: Political; Authority, Social Theories of; Causation (Theories and Models): Conceptions in the Social Sciences; Class: Social; Community Power Structure; Migration: Sociological Aspects; Power in Society

Bibliography

- Bachrach P, Baratz M 1962 Two faces of power. *American Political Science Review* 56: 947–52
- Dahl R A 1957 The concept of power. *Behavioral Science* 2: 201–15
- Dahl R A 1961 *Who Governs? Democracy and Power in an American City*. Yale University Press, New Haven, CT
- Edelman M 1964 *The Symbolic Uses of Politics*. University of Illinois Press, Urbana, IL
- Gramsci A 1971 Prison notebooks. In: Hoare Q, Nowell-Smith G (eds.) *Selections from the Prison Notebooks of Antonio Gramsci*. International Publishers, New York
- Harsanyi J C 1962 Measurement of social power, opportunity costs, and the theory of two-person bargaining games. *Behavioral Science* 7: 67–80
- Hunter F 1953 *Community Power Structure*. University of North Carolina Press, Chapel Hill, NC
- Lasswell H D, Kaplan A 1950 *Power and Society*. Yale University Press, New Haven, CT
- Lowi T 1979 *The End of Liberalism*, 2nd edn. Norton, New York
- Lukes S 1974 *Power: A Radical View*. Macmillan, London
- March J G 1955 An introduction to the theory and measurement of influence. *American Political Science Review* 49: 431–51
- March J G 1966 The power of power. In: Easton D (ed.) *Varieties of Political Theory*. Prentice-Hall, Englewood Cliffs, NJ
- McConnell G 1966 *Private Power and American Democracy*, 1st edn. Knopf, New York
- McFarland A S 1969 *Power and Leadership in Pluralist Systems*. Stanford University Press, Stanford, CA
- Mills C W 1956 *The Power Elite*. Oxford University Press, New York
- Nagel J H 1975 *The Descriptive Analysis of Power*. Yale University Press, New Haven, CT
- Polsby N W 1980 *Community Power and Political Theory*, 2nd edn. Yale University Press, New Haven, CT
- Riker W H 1964 Some ambiguities in the notion of power. *American Political Science Review* 58: 341–9
- Simon H A 1953 Notes on the observation and measurement of power. *Journal of Politics* 15: 500–16
- Simon H A 1957 *Models of Man*. Wiley, New York
- Walker Jr J L 1991 *Mobilizing Interest Groups in America*. University of Michigan Press, Ann Arbor, MI
- Wilson J Q (ed.) 1980 *The Politics of Regulation*. Basic Books, New York

A. S. McFarland

Practical Reasoning: Philosophical Aspects

Practical reasoning is reasoning that aims to resolve the question of what one is to do. Such reasoning is practical in at least two senses. First, it is practical in its subject matter, insofar as it is concerned with human action. But it is also practical in its issue, insofar as reflection about action can itself directly move people to act. Reasoning of this kind raises two sets of philosophical problems. One concerns the question of how reflection can succeed in being practical in its issue. What assumptions must be made both about agents and about the processes of reasoning they engage in to make sense of the fact that such reflection can give rise to action? Can justice be done to this fact while preserving the idea that practical reasoning is genuinely a form of reasoning? A second set of issues concerns the content of the standards that are brought to bear in practical reasoning. Which norms for the assessment of action are binding on us as agents? Do these norms provide resources for critical reflection about our ends, or are they exclusively instrumental? Under what conditions do moral norms yield valid standards for reasoning about action? The first set of issues is addressed in Sects. 1–3, while Sects. 4 and 5 cover the second set of issues.

1. Practical and Theoretical Reasoning

Practical reasoning takes place from a distinctive standpoint. When agents deliberate about action, they think about themselves and their situation in characteristic ways. What are some of the salient features of the practical point of view?

One way to interpret this point of view is to contrast it with the standpoint of theoretical reflection. Theoretical reflection is sometimes understood as reasoning about questions of explanation and prediction. Looking backward to events that have already taken place, it asks why they have occurred; looking forward, it attempts to determine what is going to happen in the future. In these ways, theoretical reflection is concerned with matters of fact and their explanation. Furthermore it treats these issues in impersonal terms that are accessible (in principle) to anyone. The natural and social sciences can be viewed as giving paradigmatic expression to theoretical reasoning, so construed.

Practical reasoning, by contrast, takes a distinctively normative question as its starting point. It typically asks, of a set of alternatives for action, none of which has yet been performed, what one ought to do, or what it would be best to do. It is thus concerned not with matters of fact and their explanation, but with matters of value, of what is to be done. In practical reasoning

agents attempt to assess and weigh their reasons for action, the considerations that speak for and against alternative courses of action that are open to them. Moreover, they do this from a distinctively first-personal point of view, one that is defined in terms of a practical predicament in which they find ourselves (individually or collectively—people sometimes reason jointly about what they should do together; see *Joint Action*).

There is, however, a different way of understanding theoretical reasoning, stressing the parallels rather than the differences between the two forms of reflection. According to this interpretation, theoretical reflection is concerned with a normative rather than a factual question, namely with the question of what one ought to believe. It attempts to answer this normative question by assessing and weighing reasons for belief, the considerations that speak for and against the particular conclusions one might draw about the way the world is. Furthermore, it does this from a standpoint of first-personal reflection: the stance of theoretical reasoning in this sense is the committed stance of the believer, not the stance of detached contemplation of one's beliefs themselves (Moran 1997). Seen in this way, the contrast between practical and theoretical reflection is essentially a contrast between two different systems of norms: those for the regulation of action on the one hand, and those for the regulation of belief on the other.

Theoretical reflection, interpreted along these lines, addresses the considerations that recommend accepting particular claims as to what is or is not the case. That is, it is reasoning with an eye to the truth of propositions, and the reasons for belief in which it deals are considerations that speak in favor of such propositions' being true, or worthy of acceptance. Practical reasoning, by contrast, is concerned not with the truth of propositions but with the value of actions. The reasons in which it deals are considerations that speak in favor of particular actions being good, or worthy of performance in some way. This difference in subject matter corresponds to a further difference between the two kinds of reasoning, in respect of their consequences. Theoretical reflection about what one ought to believe produces changes in one's overall set of beliefs, whereas practical reason gives rise to action; it is, as noted above, practical not only in its subject matter, but also in its issue.

Two observations are in order about this side of practical reasoning. First, the contrast just drawn might suggest that there is a categorical difference in the effects of theoretical and practical reasoning, insofar as the former produces changes in our mental states, whereas the latter gives rise to bodily movements. But this would be misleading. Practical reasoning gives rise not to bodily movements *per se*, but to intentional actions, which are intelligible as such only to the extent they reflect our mental states. It would thus be more accurate to characterize the consequences of both

theoretical and practical reasoning as *attitudes*, noting that the former leads to modifications of our beliefs, whereas the latter leads to modifications of our intentions (Harman 1986, Bratman 1987). Second, in neither case do the modifications in question occur infallibly. There is room for irrationality both in the theoretical and the practical sphere, which in its strongest form involves a failure to acquire the attitudes that one acknowledges to be called for by the reasons one has reflected on. Thus a person might end up watching another hour of television, while judging that it would be better on the whole to go back to work. Practical irrationality of this latter kind is known as *akrasia*, incontinence, or weakness of will, and its possibility makes clear that practical reasoning is not automatically practical in its issue. A better way to represent the consequences of practical reasoning would be to say that it generates appropriate intentions to the extent an agent is rational (Korsgaard 1996; see *Irrationality: Philosophical Aspects*).

2. Naturalism and Normativity

The connection of practical reasoning with intentional action raises large questions about its credentials as a genuine form of reasoning. As noted above, intentional action is not mere bodily movement, but reflects a mental state of the agent's, viz. intention. To be in this kind of mental state is to have a goal or aim of some kind, which one seeks to realize through one's action. Intention seems in this respect to be strikingly unlike belief. The propositions that give the content of beliefs have a representative function; they aim to fit the way the world is, so that if one discovers that the world is not how one previously took it to be, one's belief will automatically be modified in the process. With intentions, however, matters are otherwise (Smith 1987). The intention to go shopping on Tuesday, for instance, is not a state that would or should be abandoned upon discovering that one has not (yet) gone shopping on Tuesday; rather the person with such an intention will ordinarily try to bring the world into alignment with the intention, by going shopping when Tuesday comes around. Intentions are in this way more like blueprints than like sketches of an already completed structure.

Reflection on this contrast between belief and intention has led some philosophers to ask whether practical reasoning might not be something of a misnomer. The concern, in a nutshell, is how processes of thought that count as genuine reasoning could by themselves generate states with the peculiar, nonrepresentative function of intentions. Reasoning seems to consist of cognitive operations, whereas intentions are distinctively noncognitive states insofar as they do not aim to reflect the way things happen to be in the world.

Expressivist accounts give voice to this skeptical attitude about practical reasoning. These accounts offer interpretations of the normative language that distinctively figures in practical reflection about action. As was seen in Sect. 1, such reflection is an attempt to assess an agent's reasons for acting in one way or another; conclusions about such reasons are characteristically expressed in claims about what it would be *good* to do, or about the actions that one *ought* to perform. According to the expressivist, however, such claims do not represent genuine cognitive achievements, judgments that are literally capable of being true or false. Rather they give expression to desires, sentiments, and pro-attitudes, the kinds of noncognitive state that motivate people to action. Only if normative assertions are understood in this way—so the expressivist contends—can sense be made of the capacity of practical reflection to generate states with the peculiar structure and function of intentions.

Expressivism in this form can be seen as a naturalistic interpretation of practical reasoning, one that is appropriate to the enlightened commitments of the modern scientific world view. It is naturalistic metaphysically, insofar as no assumptions need be made about the objective existence in the world of such questionable entities as values, norms, or reasons for action. If normative claims do not represent genuine cognitive achievements, then their legitimacy does not depend on our postulating a realm of normative facts to which those claims might correspond. It is also naturalistic psychologically, insofar as it promises explanations of intentional human behavior that are basically continuous with explanations of the behavior of nonhuman animals. In both the human and the nonhuman case, behavior is seen as the causal product of noncognitive attitudes, operating in conjunction with the creature's factual representation of how things are in its environment. The difference is just that humans have much more sophisticated linguistic methods for giving voice to their motivating noncognitive attitudes. Indeed, many contemporary expressivists would contend that these expressive resources are sufficiently powerful to explain the features of practical reflection that initially give it the appearance of a genuine form of reasoning (Blackburn 1998, Gibbard 1990).

Other philosophers remain unimpressed with these naturalistic explanations. One ground for dissatisfaction with them is the following. The expressivist strategy relies on an initial contrast between practical reflection and the genuine forms of cognitive activity characteristic of theoretical reasoning. There has to be some basic sense in which such discourse does not come up to the standards of authentic cognitive discourse in the literal sense, otherwise the contention that normative discourse is expressive rather than cognitive will lack any significant content. But the contrast between theoretical and practical reflection required for this purpose seems elusive. As was seen in

Sect. 1, theoretical reasoning is no less a normative enterprise than practical reasoning. It concerns itself with reasons for belief: the evidence and other considerations that speak for and against particular conclusions about the way things are in the world. In this respect, theoretical and practical reasoning would seem equally problematic from the naturalistic perspective—assuming, that is, that it leaves no place for such normative considerations as reasons. But if naturalism calls in question the credentials of theoretical reason, it thereby undermines the contrast between genuine reasoning and noncognitive normative discourse on which expressivists themselves rely (see *Emotion and Expression; Intentionality and Rationality: An Analytic Perspective*).

3. Reasons and Motivation

The capacity of practical reasoning to generate intentional action divides even those philosophers who agree in rejecting the expressivist strategy discussed above. Such philosophers are prepared to grant that there are reasons for action, and to accept the cognitive credentials of normative claims about such reasons. But they differ in their accounts of the truth conditions of these normative claims. Two approaches may be distinguished. The first of these, frequently referred to as internalism, holds that reasons for action must be grounded in an agent's prior motivations (Williams 1981). According to this position, a given agent *s* can have reason to do *x* only if *x*-ing would speak to or further some element in *s*'s 'subjective motivational set.' There must be some rational route that connects *s*'s *x*-ing with the subjective motivations to which *s* is actually already subject; otherwise the claim that *s* has reason to *x* must be rejected as false or incoherent. Behind this internalist position lies the idea that practical reasoning is practical in its issue. Internalists contend that sense can be made of the capacity of such reasoning to generate new intentions only if it is conditioned by motivational resources that are already to hand. Reasoning about action is a matter of working out the implications of the commitments contained in one's subjective motivational set; in this sense, motivation is prior to practical reason, and constrains it. Externalists reject this assumption, contending that one can have reasons for action that are independent of one's prior motivations. Most of them agree that practical reasoning must be capable of generating new intentional actions, and hence of giving rise to motivation. They agree, in other words, that if agent *s* has reason to do *x*, it must be possible for *s* to acquire the motivation to *x* through reflection on the relevant reasons. But they deny that such reasoning must in any way be constrained by *s*'s subjective motivations prior to the episode of reasoning. Prac-

tical reasoning is not exclusively a process of working out the implications of one's existing commitments; it is also an attempt to get clear about what it would objectively be good to do. Normative considerations of this kind are independent of one's prior motivations, and can open up new motivational possibilities (Parfit 1997).

Underlying this dispute are diverging approaches to the explanation of intentional action. Internalists are impressed by the differences between intentions and the cognitive states that figure in paradigmatic examples of theoretical reasoning. Pointing to these differences, they ask how practical reasoning can succeed in producing new intentions if it is not grounded in something of the same basic psychological type: a motivation or desire that is already part of the agent's subjective motivational equipment. Many externalists find this contrast between intentions and cognitive states overdrawn. They observe that we need to postulate basic dispositions of responsiveness to reasons to account for the capacity of theoretical reflection to generate new beliefs, and question why these same dispositions cannot explain the fact that practical reasoning is practical in its issue. Cognitive or not, intentions belong to the broad class of attitudes that are sensitive to judgments, and this may account for the capacity of practical reflection to give rise to new intentions (Scanlon 1998). A third possibility is that intentions are products of the will, a capacity for self-determination distinct from the dispositions that render theoretical rationality possible. Depending on how it is developed, the theory of the will may offer a different way of accounting for the practical consequences of practical reflection, without assuming that reasons for action are grounded in an agent's subjective motivations (Wallace 1999).

4. Instrumental and Maximizing Rationality

Among the substantive norms of practical reasoning, those of instrumental rationality appear least controversial. Instrumental rationality, in its most basic form, instructs agents to take those means that are necessary to achieve their ends. In the modern era, this form of rationality has widely been viewed as the single unproblematic norm of practical reasoning. The instrumental principle makes no assumptions about the prospects for rationally scrutinizing peoples' ends. Rational criticism of this kind would seem to presuppose that there are objective values, standards for assessment of ends that are independent from psychological facts about what people happen to be motivated to pursue. In keeping with the naturalistic attitude sketched in Sect. 2, however, it may be doubted whether such standards can be reconciled with the metaphysical commitments of contemporary scientific practice (see *Value Pluralism*). A world that is shorn of objective values or norms leaves no room for

rational criticism of people's ends, but only for Weberian *Zweckrationalität*: the rational determination of means to the realization of ends that are taken as given, as a matter of psychological fact.

It is important to bear in mind, however, that instrumental rationality is itself the expression of an objective normative commitment. The instrumental principle tells us that we have *reason* to take the means that are necessary to achieve our ends; if the principle is valid, then people are open to rational criticism to the extent they fail to exhibit this kind of instrumental consistency, regardless of whether they want to comply with the principle or not. If naturalism really leaves no place for objective norms or values, it may be wondered how an exception can possibly be made for the instrumental norm. The consistent naturalist should perhaps reject even *Zweckrationalität* in favor of a skeptical attitude towards practical reasoning in all its forms (Hampton 1998). Further questions can be raised about the plausibility of the instrumental norm as a self-standing principle of practical reason. The norm says that one should take the means that are necessary to achieve one's ends. But how can the fact that a given means is necessary, relative to some end, give a person reason to choose the means, if the end is not itself something it would be good to achieve in some way? The instrumental principle seems to function as a binding standard for practical reasoning only if it is taken for granted that there are additional, independent standards for the assessment of our ends (Korsgaard 1997).

One way to make room for the rational criticism of individual ends, without departing from the spirit of *Zweckrationalität*, is to expand one's view to encompass the totality of an agent's ends. Thus, even if there are no values that are ultimately independent of an agent's given ends, it would appear possible to criticize particular intrinsic desires by reference to others in the agent's subjective motivational set. An agent's desire for leisure, for instance, might be subordinated insofar as its satisfaction would frustrate the realization of other goals that are subjectively more important to the agent. Practical reasoning, it might be suggested, is a holistic enterprise, concerned not merely with the specification of means to the realization of individual ends, but with coordinated achievement of the totality of an agent's ends.

Many philosophers view this holistic perspective as the most fruitful way of conceptualizing the tasks of practical reasoning. It defines an important and difficult problem for practical reasoning to solve, without departing from the metaphysically modest assumption that there is no court of appeal for the rational criticism of an agent's ends that is independent of those ends themselves. The most sophisticated and influential expression of this holistic perspective is the maximizing conception of rationality. According to the maximizing conception, the fundamental aim of practical reasoning is to determine which course of

action would optimally advance the agent's complete set of ends. Thus it has been suggested that the rational action is the one whose subjective expected utility—reflecting both the utility of possible outcomes, from the agent's point of view, and the agent's beliefs about the probability of those outcomes—is the highest.

The maximizing conception of rationality finds its most complete expression in decision theory and in the theory of rational choice (as developed, for instance, in economics). These disciplines articulate with mathematical precision the assumption that one is acting rationally to the extent one does what is likely to bring about the best state of affairs, given both one's preferences and one's beliefs about the probable consequences of the alternatives that are open to one. Proponents of these theories often claim for them the additional advantage of empirical adequacy, contending that they are flexible enough to accommodate the full range of behaviors that human agents engage in, both within the marketplace and outside of it. Especially if one operates with the notion of 'revealed preferences'—preferences, that is, that are inferred solely on the basis of actual behavior—then virtually anything an agent might choose to do could be interpreted as an attempt to maximize expected utility. Decision theory, on this account of it, becomes an all-encompassing framework for the interpretation of human behavior, according to which agents are always striving to produce outcomes that would be optimal, relative to their preferences and beliefs (see *Rational Choice Explanation: Philosophical Aspects*).

If decision theory is interpreted in this way, however, then its relevance to the understanding of practical reasoning begins to seem correspondingly tenuous. Maximizing rationality initially looks to be a norm of practical reflection. In this guise, its appeal lies in the idea that there can be rational requirements on action, stemming from the totality of an agent's preferences and beliefs, even if it is not assumed that there are independent, substantive standards for the critical assessment of individual ends. But this normative understanding of maximizing rationality is tenable only if it is at least conceivable that individual agents might occasionally fail to comply with its requirements—an 'ought' that it is not so much as possible to flout is not really an 'ought' at all. Proponents of the maximizing approach need to be attentive to the latent tension between the empirical and the normative aspirations of their theory, for maximizing rationality cannot claim both perfect empirical adequacy and normative authority for practical deliberation.

Assuming this tension can be resolved, further questions arise about the plausibility of the normative requirement to maximize expected utility. Doubts have been raised, for instance, with regard to the assumption that it is necessarily irrational to fail to select that action that would be optimal, relative to one's preferences and beliefs. Perfectly rational agents

often seem to be content with states of affairs that are 'good enough,' from the perspective of their aims and desires, even when they are aware of alternatives that promise a higher return; they 'satisfice' rather than seeking to maximize the value of the outcomes that they bring about (Slote 1989). They also treat their past intentions and plans as fixed constraints on their decision-making, rather than attempting to maximize subjective utility anew in every situation they confront (Bratman 1987). Some contend that maximizing rationality is flexible enough that it can accommodate alleged counterexamples of these kinds (Pettit 1984). If not, however, there may be grounds for doubting that it represents a fundamental and uncontroversial norm for reasoning about action.

5. Specification, Expressive Rationality, and Moral Reasons

If maximizing rationality is not the unproblematic requirement of practical reasoning that it initially seemed to be, what are the alternatives to it? Consider the assumption that critical assessment of an agent's individual ends is off-limits. This apparently truistic assumption has been questioned by some philosophers, who point out that a great many of one's aims in life are rather inchoate; people want, for instance, to be successful in their careers, and loyal to their friends, without being entirely clear about what these ends require of them. To the extent one's ends are indefinite in this way, they will not provide satisfactory starting points for instrumental, maximizing, or even satisficing reflection. One will need to specify such ends more precisely before one can begin to think about which means they require one to pursue, or to generate from them a rank-ordering of possible outcomes. Here is a possible task for practical reasoning that does not fit neatly into the categories of *Zweckrationalität*, however broadly construed (Wiggins 1987, Richardson 1994).

Practical reflection about ends is not an easy or well-defined activity. There are no straightforward criteria for success in this kind of reflection, and it is not even clear when it has been brought to a satisfactory conclusion. These considerations are no doubt partially responsible for the assumption—especially widespread in the social sciences—that there is no reasoning about final ends. On the other hand, how is one supposed to clarify one's largest and most important ends, if not by reasoning about them in some way? Rather than exclude such reflection because it does not conform to a prior and possibly inappropriate paradigm of proper reasoning, perhaps the conception of practical reasoning should be expanded to make room for clarificatory reflection about the ends of action. To do so would be to acknowledge that practical reasoning can be both objective and subjective at the same

time; by working out the meaning and implications of such commitments as loyalty or success, for instance, agents also help to get clear the values that define who they are (Taylor 1985; see *Hermeneutics, Including Critical Theory*).

A different point at which to question received models of practical reasoning is the theory of value to which they are committed. Maximizing and satisficing models alike seem committed to a consequentialist account of the relation of action to value: the thesis that value inheres ultimately in states of affairs, and that actions are to be judged rational to the extent they produce valuable states of affairs. This thesis seems at odds, however, with intuitive ways of thinking about value. It appears a distorted understanding of friendship, for instance, to say that what friends value fundamentally are states of affairs (involving, say, joint activities with the friend); what people value as friends are rather the concrete persons with whom they are befriended. This suggests a different way of conceptualizing the rationality of action. What makes actions rational, it has been claimed, is not the production of valuable states of affairs, but rather their success in expressing the attitudes that it is rational to adopt toward the true bearers of intrinsic value: people, animals, and things (Anderson 1993). A supposed advantage of this approach is its ability to explain the rationality of behaviors that seem intuitively sensible, but that are hard to fit into the consequentialist scheme (such as commitments deriving from one's past involvement in an activity or project, which can look like an irrational weighing of 'sunk costs' to the consequentialist; compare Nozick 1993).

Morality provides an especially fertile source of examples and problems for the theory of practical reasoning. One of the defining questions of moral philosophy is the question of the rational authority of moral norms: to what extent, and under what conditions, are people rationally required to comply with the demands of conventional morality? Reflection on this question has produced some of the most far-reaching and illuminating philosophical work on the nature of practical reasoning. Two divergent tendencies within this body of work should be singled out. Some approaches to moral reasoning proceed by relating it to patterns of reflection familiar from other, nonmoral domains, particularly the maximizing patterns canvassed in the preceding section. Thus it has been argued that, though morality imposes constraints on the direct pursuit of individual utility, these constraints can be justified in terms of the values expressed by ordinary economic rationality; a strategy of constrained maximization is recommended on grounds of enlightened self-interest, and this in turn accounts for the authority of moral norms to govern the practical reflection of individuals (Gauthier 1986). Other philosophers have sought to make sense of morality as a source of rational norms by assimilating

it more directly to the maximizing conception. For instance, proponents of consequentialism as an ethical theory—the theory that agents are morally required to maximize the good, impartially conceived—consider it an advantage of this account that it represents moral requirements in terms of the maximizing conception of rationality familiar from other contexts (Harsanyi 1982; see *Consequentialism Including Utilitarianism; Economics and Ethics*).

A different approach stresses the discontinuities between moral and nonmoral patterns of reasoning. According to this approach, morality is a source of constraints (such as prohibitions on murder and deception) that cannot be represented accurately within the framework of maximizing rationality (for example, Scanlon 1998). If this is correct, then those who wish to make sense of moral requirements, as norms that appropriately govern the reflection of individual agents, will need to expand their conception of the forms and possibilities of practical reasoning (see *Ethics and Values*).

See also: Ethics and Values; Free Will and Action; Intentionality and Rationality: A Continental-European Perspective; Intentionality and Rationality: An Analytic Perspective.

Bibliography

- Anderson E 1993 *Value in Ethics and Economics*. Harvard University Press, Cambridge, MA
- Blackburn S 1998 *Ruling Passions*. Clarendon Press, Oxford, UK
- Bratman M E 1987 *Intention, Plans, and Practical Reason*. Harvard University Press, Cambridge, MA
- Gauthier DP 1986 *Morals by Agreement*. Clarendon Press, Oxford, UK
- Gibbard A 1990 *Wise Choices, Apt Feelings*. Harvard University Press, Cambridge, MA
- Hampton J E 1998 *The Authority of Reason*. Cambridge University Press, Cambridge, UK
- Harman G 1986 *Change in View*. MIT Press, Cambridge, MA
- Harsanyi J C 1982 Morality and the theory of rational behavior. In: Sen A, Williams B (eds.) *Utilitarianism and Beyond*. Cambridge University Press, Cambridge, UK
- Korsgaard C M 1996 Skepticism about practical reason. In: Korsgaard C M (ed.) *Creating the Kingdom of Ends*. Cambridge University Press, Cambridge, UK
- Korsgaard C M 1997 The normativity of instrumental reason. In: Cullity G, Gaut B (eds.) *Ethics and Practical Reason*. Clarendon Press, Oxford, UK
- Moran R 1997 Self-knowledge: discovery, resolution, and undoing. *European Journal of Philosophy* 5: 141–61
- Nozick R 1993 *The Nature of Rationality*. Princeton University Press, Princeton, NJ
- Parfit D 1997 Reasons and motivation. *Proceedings of the Aristotelian Society* 77: 99–130
- Pettit P 1984 Satisficing consequentialism. *Proceedings of the Aristotelian Society* 58: 165–76
- Richardson H S 1994 *Practical Reasoning about Final Ends*. Cambridge University Press, Cambridge, UK

- Scanlon T M 1998 *What We Owe to Each Other*. Belknap Press of Harvard University Press, Cambridge, MA
- Slote M 1989 *Beyond Optimizing*. Harvard University Press, Cambridge, MA
- Smith M 1987 The human theory of motivation. *Mind* 96: 36–61
- Taylor C 1985 What is human agency? In: Taylor C (ed.) *Human Agency and Language*. Cambridge University Press, Cambridge, UK
- Wallace R J 1999 Three conceptions of rational agency. *Ethical Theory and Moral Practice* 2
- Wiggins D 1987 Deliberation and practical reason. In: Wiggins D (ed.) *Needs, Values, Truth*. Blackwells, Oxford, UK
- Williams B 1981 Internal and External Reasons. In: Williams B (ed.) *Moral Luck*. Cambridge University Press, Cambridge, UK

R. J. Wallace

Practice: Anthropological Aspects

The word practice carries a clear meaning in everyday English. A practice is a convention or institution, and in anthropological discourse the word often appears in its ordinary sense. Tibetans follow the practice of sky burial or South Indians practice cross-cousin marriage. But the word also enters anthropological discourse in more theoretized ways.

1. Sources

The idea that human action takes place against a background—constituted or tacit knowledge, *inter-subjective meanings*, a form of life, or practice—has sources in Husserl, Wittgenstein, and Polanyi. But the notion has mostly influenced students of human behavior, from Merleau-Ponty to Michel Foucault and Pierre Bourdieu, by way of Heidegger's *Being and Time*. For Heidegger, the epistemological question—how do we know our understanding of the world matches the world itself—cannot be asked as Descartes did, by making the question of truth a matter of beginning with what a conscious subject can know of the world with absolute certainty and working from there. The Cartesian subject on Heidegger's account can know that he or she exists only because of the presence of a set of background practices which make knowing (or acting, judging, laughing, and so on) possible. The intelligibility of the world in other words depends on the actor's prior 'being-in-the world.'

Heidegger argues that understanding human action requires recognizing that it is practical twice over—it depends on this background of practices and it is purposive. In Heidegger's language *intentionality* is always practical rather than cognitive, but practices are purposive in an extended sense of the word. The actor may not have any purpose in mind because

being-in-the world is a skilled activity that goes on whether the actor is aware or not (1962, p. 405). Intrinsic to human activity is the actor's comportment with things in the world, and everyday being resides in his or her doing, intending, and coping. Human beings are socialized into these practices relative to their historical moment and culture, yet the practices themselves cannot be made explicit. If such is the case, the idea of practice challenges hopes for a scientific account of behavior.

Mainstream political science, for instance, assumes that the *actor's point of view* is a thing in itself. It consists of beliefs, attitudes, and evaluations. These states of mind are an objective social reality and capable of furnishing the brute data on which the political scientist constructs a science of politics. The problem is that the actor's point of view cannot be reduced to that which can be attributed to him as a belief, attitude, or evaluation if the content of consciousness depends on intersubjective belief or practices that stand behind these mental states and make them possible (Taylor 1971). The intersubjective beliefs Japanese villagers hold about reaching public consensus, Taylor says, must be understood as residing in the relationship between actors and the society they inhabit. Critics of *positivism* argue that no amount of questioning could elicit these intersubjective beliefs from the Japanese themselves because they are ineffable and depend on still other practices. Yet without accounting for the bottomless matrix of practices in which those beliefs, attitudes, and evaluations are embedded, no comparative science of politics is possible.

2. Habitus, Practice, and Time

Another example of Heidegger's influence is Bourdieu's use of the word practice in the sense of *praxis*, *action*, or *agency*. While Taylor reacts to the assumption that social action can be explained by focusing on an actor's consciousness, Bourdieu denies that it can be understood as consciousness reduced to rules and norms. In his view, action is relational, recursive, and reflexive. Behavior has a cultural referent, but is considerably more complicated than the execution of a blueprint. Bourdieu's attempt to accommodate strategy and innovation as part of an account of social behavior appeared as nearly as his *Algeria 1960* which he organized around the distinction between dispositions and economic structures (English trans. 1979). He theorizes that relationship in a full-blown way in his *Outline of a Theory of Practice* (1977). Dispositions, Bourdieu writes, have two parts—the *habitus*—the ensemble of dispositions towards action and perception that operates from within social agents—and the practice which the habitus generates. The habitus is 'the durably installed generative principle of regulated improvisations' (1977, p. 78). Those structured, objectively unified practices

are subjective and changing, but they follow from the habitus in the same way a jazz band improvises on a familiar melody.

Where a traditional anthropological account of kinship focused on genealogical rules and the constraints those rules exert on some activity such as marriage choice, Bourdieu saw genealogical forces in relation to other variables (1977). To regard Kabyle marriage as the realization of a parallel cousin rule for arranging a match is to remove one source of determination from the relationship in which it operates. Any calculation that an anthropologist might make of how often Algerians marry their parallel cousins depends on the extent of the social unit in relation to which the calculation is made. The size of that social unit is itself a strategic variable (1990, p. 15). Other variables pertain to the couple (age and age difference, previous matrimonial history, sibling order, relationship to the family authority-holder), but a great many more are family matters (the economic and social status of the families, their social affiliations, ability to defend themselves, as well as sheer numbers). For Kabyles the habitus includes parallel cousin marriage as one disposition among many, and the strategies that emerge from all of those dispositions—weighing each in relation to others—constitute practice.

Families strategize by considering all of these considerations, and as they learn more about the other side, their strategy evolves. Because human behavior is recursive as well as relational, its analysis must take *time* into consideration. Bourdieu insists that de-temporalizing science fails to understand what constitutes social behavior. Time gives behavior 'its form, as the order of succession, and therefore its direction and meaning' (1990, p. 98). Where Lévi-Strauss thought the gift exchange was organized according to deep principles unknown to the actor, Bourdieu argued that it was driven by 100 small calculations, idiomatic to actors who re-evaluate and change their strategies as the situation unfolds in time. Those calculations may be intuitive, but theory misconstrues its object when it takes the actor's agency as simply applicative.

There is a third part of practice where Bourdieu turns Heidegger in a sociological direction. Practices are *reflexive*. They carry with them their own interpretation, which means that they organize the relationship between individuals and society in a way that is circular. Heidegger puts the matter in distinctive language—existence is an understanding of being shared by individuals and society. Because practices are shared, they are the basis of mutual intelligibility, but this same condition means that there is never anything like a subjective inner sphere (1977, pp. 80–1). Individuals, whatever their self-conception, are social creatures. Individuality cannot be absolute, but a variant on the local class habitus and other dispositions.

In this sense Bourdieu blurs the distinction between the individual and society. 'The habitus—embodied

history, internalized as a second nature and so forgotten as history—is the active presence of the whole past of which it is a product. As such, it is what gives practices their relative autonomy with respect to external determinations of the immediate presence. This autonomy is that of the past, enacted and acting, which ..., makes the individual agent a world within a world' (1990, p. 56). *Symbolic interactionists* and *ethnomethodologists* assume that social action derives from individuals responding to immediate circumstances, and Bourdieu too wants to account for circumstances. But he does so by insisting that the actor's course is guided by a dialectic. He is always a social being subject to the orientations and limits of the habitus' 'operations of invention' as they develop amidst those objective structures (1977, p. 95).

That practices are reflective of larger social understandings (of what it is to be a person, a teacher, a woman) has implications for the way Bourdieu understands agency: 'it is because subjects do not, strictly speaking, know what they are doing that what they do has more meaning than they know' (1977, p. 77). It also has implications for the way Bourdieu understands agency as organized in time. He says that gift exchange is one of the social games that cannot be played unless the players ignore 'the objective truth of the game ... and unless they are predisposed to contribute, with their efforts, their marks of care and attention, and their time, to the production of collective misrecognition. Everything takes place as if the agents' strategies and especially those that play on the tempo of action ... were organized with a view to disguising to themselves and from others the truth of their practice, which the anthropologist brutally reveals simply by substituting the interchangeable moments of a reversible sequence for practices performed in time and in their own time' (1990, pp. 105–6).

3. Practice and Domination

The notion that the actor's point of view entails his or her misrecognizing the social game derives from a tradition of concern for consciousness that has affinities to both Marx and Heidegger. What Bourdieu takes from Marx (and what separates him from Heidegger) is the idea of structure, and specifically the economic structures in which actors act. The world is prior to the actor's agency in the world, but it bears on individuals relative to their position in the social organization of production. It is the objective character of social conditions that produces a class habitus, the harmony of dispositions, ethos, taste, and what he calls body *hexis*—while individuals are thoroughly social, they are related to society in ways appropriate to their peculiar identities.

Bourdieu has studied how relations of *domination* shake an actor's subjective aspirations in an extraordinary variety of contexts—from the Kabyles of

Algeria, peasants in the region of Bearn in south-western France, the academic establishment, systems of classification in educational settings, museums, class relations, sports, to high fashion and the production of other cultural goods in modern industrialized societies. In its most cunning form domination operates in systems of classification. Whether he turns his attention to anthropological concerns with kinship, ritual, calendars, and domestic space or more sociological matters such as fashion, taste, and educational systems of classification, Bourdieu dismisses Kantian and Lévi-Straussian assumptions. All such classifications are not folk epistemologies but ideological mystifications of political realities.

Because Bourdieu assumes the nondistinction of the material and the symbolic, he rejects 'the spurious alternatives of social physics and social phenomenology' (1990, p. 140). In the same way that charisma is not a particular form of power, contra Weber, but the condition of all power, meaning grounds all human action. In studying *taste* as a discriminating practice in modern France, Bourdieu draws the study of culture into the realm of culture understood anthropologically (1984). The endless conversion of economic capital into *symbolic capital* uses taste to enunciate class differences but requires the complicity of the entire society. That conversion is the condition which makes possible the continuation of domination, yet it 'cannot succeed without the complicity of the whole group ... As Mauss puts it, the whole society pays itself in the coin of its dream' (1977, p. 195).

4. Practice and its Applications

Seen against the background of anthropological theory, practice theory was a response to the profession's recent discontent with the notion of culture, especially its neglect of power and strategy. Seen against a Heideggerian background, it was another critique of the error with which Western philosophy began, objectivism (1962, pp. 41–9). Anthropologists found much to admire in Bourdieu's work—his even-handed concern for both the actor's point of view and structure, efforts to transcend the conventional antinomies, and recognition of temporality as figuring in social life. Exploiting practice theory as a way to tie together consciousness and structure, action and its temporal horizon, as well as Weber and Marx, theorists have applied Bourdieu's thinking to case studies that range from gift-giving, saving, and consuming to gender, sexual difference, class relations, and colonial encounters.

If one counted anthropologists invoking the importance of practice, temporality, or simply Bourdieu's name, his influence would be striking. Counting only scholars who fully pursue his research program, the number would shrink to people who

publish in the journal he founded, *Actes de la recherche en sciences sociales*, or work with him at the Centre de Sociologie Européenne in Paris. The most vibrant tradition occupying the space between these extremes are anthropologists who have seen practice theory as a vehicle for coping with history and temporality. Marshall Sahlins' work on Hawaii and Fiji carries on Bourdieu's interest in social reproduction, here fixed on situations where European actors intruded on South Pacific systems of domination. In accounting for the circumstances that surrounded Captain Cook's arrival in Hawaii, Sahlins characterizes events in a way that melds agency and structure, showing how mutual misrecognition created changes that look in retrospect entirely predictable (1985). In dealing with Hawaiians's initially worshipful attitude to Cook and his eventual murder, Sahlins pushes Bourdieu's commitment to agency and structure in a historical direction, arguing that those incidents evolved in a way better understood not as misrecognition and disaster, but the reproduction of culture—Hawaiians thought Cook was their god Lono—and its transformation. Sherry Ortner melds practice theory, structure, and history in characterizing the relationship of 'big' people and 'small' people and monks and nuns in Sherpa society and relations among that society, the Nepal state, and the British Raj (1989). Her confidence in the Bourdieuan project also organized her important article evaluating the state of anthropological theory (1984).

Arjun Appadurai's attempt to theorize globalization has even more inspirations, but Bourdieu's work is critical. Approaching consumer capitalism as a system as much symbolic as material, Appadurai extends the analysis in several directions. He understands consumption as a transnational phenomenon, which finds its locus at the level of body hexis, the individual, class, and nation-state. Stressing the way the media have unleashed the imagination, Appadurai argues that in the transnational world the habitus has been replaced by self-conscious choice, justification, and representation, often intended for far-flung diasporic communities and capitalist frontiers (1996, p. 44). But the role agency, strategy, and innovation play in Appadurai's account of local societies negotiating modernity would strike Bourdieu as the operations of the habitus under new circumstances.

See also: Interactionism; Symbolic; Phenomenology in Sociology; Phenomenology: Philosophical Aspects

Bibliography

- Appadurai A 1996 *Modernity at Large*. University of Minnesota Press, Minneapolis, MN
- Bourdieu P 1977 *Outline of a Theory of Practice* [trans. Nice R]. Cambridge University Press, Cambridge, UK
- Bourdieu P 1979 [1963] *Algeria 1960* [trans. Nice R]. Cambridge University Press, Cambridge, UK

- Bourdieu P 1984 *Distinction* [trans. Nice R]. Harvard University Press, Cambridge, UK
- Bourdieu P 1990 *The Logic of Practice* [trans. Nice R]. Stanford University Press, Stanford, CA
- Heidegger M 1962 *Being and Time*. Harper and Row, New York
- Heidegger M 1981 *Basic Problems of Phenomenology* [trans. Hofstadter A]. Indiana University Press, Bloomington, IN
- Ortner S B 1984 Theory in anthropology since the sixties. *Comparative Studies in Society and History* 26(1): 126–66
- Ortner S B 1989 *High Religion: A Cultural and Political History of Sherpa Buddhism*. Princeton University Press, Princeton, NJ
- Sahlins M 1976 *Culture and Practical Reason*. University of Chicago Press, Chicago
- Sahlins M 1985 *Islands of History*. University of Chicago Press, Chicago
- Taylor C 1971 Interpretation and the sciences of man. *Review of Metaphysics* 25(1): 3–51

S. Kemper

Pragmatics: Linguistic

Pragmatics is the study of the relationship between the structure of a semiotic system (notably language) and its usage in context, and, along with semantics (see *Semantics*), forms part of the general theory of meaning. Within the theory meaning, pragmatics is concerned especially with implicit meaning, with inference and the unsaid, and with the way in which language structure trades on this background of the presumed and the inferred. Pragmatics has come to play an important part in general linguistic theory, partly because it has substantial intrinsic subject matter, partly because it promises explanations for other linguistic phenomena, and partly as a response to over-idealization in contemporary grammatical theory (see *Generative Grammar*). It has also become an area of interdisciplinary concern, with fundamental contributions from philosophy of language, linguistics, psychology, and the sociology of language.

1. History

The use of the term ‘pragmatics’ in modern semiotics derives from the philosophical work of C. S. Peirce and R. Carnap, reflected in Morris’s (1938) three divisions of semiotics, the study of sign systems: *syntax*, which investigates the relation of signs to signs; *semantics*, which investigates the relation of signs to the things referred to; and *pragmatics*, which studies the relation of signs to their users. Since then, the usage of the term has bifurcated into a broad use, which subsumes sociolinguistics (see *Sociolinguistics*) and discourse analysis (see *Conversation Analysis: Sociological, and Discourse, Anthropology of*), and a narrower use (associated especially with philosophy of

language and approaches to the study of meaning which derive from it) in which pragmatics deals with those aspects of meaning that are systematically context-dependent.

2. Scope

Contemporary linguistic pragmatics is focused on a number of special relations between linguistic meaning and context. On the narrower scope of pragmatics, concerned with context-dependent meaning, the following topics have come to be central: deixis, presupposition, speech acts, implicature, and conversational inference, which are discussed briefly in turn.

2.1 Deixis

Deixis, from the Greek word for ‘pointing or demonstrating’ (the equivalent philosophical term ‘indexicality’ is from the corresponding Latin) is the phenomenon whereby some linguistic expressions are systematically dependent on the context for their interpretation. Take, for example, the utterance ‘Put this book over there’—which book is being referred to, and in which place it is to be put, are determined by features of the context outside the utterance itself, here typically gestures: *this* and *there* act as place holders for contextually-specified parameters. A good test for whether an expression is deictic or not is whether an utterance in which it plays a part would be equally true (or false) regardless of speaker, addressee, time, or place of utterance: thus ‘The square root of nine is three’ has no obvious deictics, whereas ‘We are coming soon’ has only deictic expressions.

At first sight, deixis seems to be a simple phenomenon, a leftover from the direct ‘here and now’ relevance of animal communication systems. But the intersection of this context-dependence with the property of abstract symbolic representation in language (Hockett’s displacement feature of language design) leads to deep complexities, and the phenomena turn out to be very puzzling, both philosophically and psychologically, for deixis introduces context dependency into almost every utterance (in English, for example, nearly all sentences are tensed, and tense is deictic, as in ‘It is Tuesday’ where *is* locates the reference time as today). One of the central philosophical puzzles is that deixis makes possible self-reference in utterances, thus introducing, for instance, many paradoxes of the ‘Cretan liar’ type: ‘This sentence is false’ is true if and only if it is false. One of the central psychological puzzles is that deixis makes clear that there is a mismatch between what a sentence means and the thought corresponding to its utterance. For example, what exactly is the thought corresponding to the utterance ‘Today is Tuesday’? It is obviously not ‘Tuesday is Tuesday,’ nor (if the day of speaking is

31st December 2035) ‘31st December 2035 is a Tuesday’ since the speaker may have no idea of the date. Closer might be ‘The diurnal span in which I am now speaking falls on the calendrical second day of the week’—but now we have paraphrased one deictic term (*today*) in terms of others (*I, now*). And the puzzle goes on.

Languages incorporate deictic context-dependency in many different places in their grammars and lexica, but it is conventional to consider four main notional parameters.

First, there is *person deixis*, reflected in the traditional grammatical categories of first, second, and third person, where first and second person refer to members of the speech event (current speaker and current addressee, respectively) in contrast to the third person. These categories are reflected in pronoun systems, verbal agreement and elsewhere. Many languages make unfamiliar demands here: for example, in Southeast Asian languages such as Korean or Japanese have different word-forms for referring to entities such as kinsfolk, houses, or food, according to the relative status of speaker and addressee. Or in some Australian languages (such as Dalabon) pronouns make reference to whether the speaker and the referent, or the referents in the case of plural third person pronouns, are one vs. two generations apart genealogically. This encoding of the social relation between speaker, addressee, and third-party referents is often recognized as a separate deictic dimension, called *social deixis*.

A second important parameter is time. *Temporal deixis* is manifested directly in English in words such as *now*, and, as mentioned above, in tense. True tenses locate time relative to the moment of speaking, so that a past tense is used for events (or situations) preceding it, present for events including it, and future for events succeeding it. Not all languages have tenses (e.g., Malay or Chinese), and not all the grammatical categories that grammarians call tenses are truly deictic. Some languages (e.g., Bantu or Papuan languages) have up to six deictic tenses, however (specifying, for instance, that something happened the day before yesterday). There are many other manifestations of temporal deixis, from calendrical specifications like *yesterday* to more covert items like *ago* (as in *nine years ago*) or *Good night* which presumes that it is night and that the speaker is leaving.

A third parameter is *spatial deixis*, reflected in demonstratives such as *this* and *that*, or adverbs such as *here* and *there*, where spatial locations are indicated by reference to the place of speaking. Demonstrative systems vary greatly across languages, with some making distance from speaker primary, others making a contrast between ‘this near speaker’ vs. ‘that near addressee,’ and others invoking attentional issues, so that one has a contrast between, for instance, ‘this we are talking about’ vs. ‘that to which I want to draw your attention.’ Huge arrays of demonstratives are

found in some languages (e.g., Inuit) where these sorts of parameters intersect with, for example, the shape of things referred to. Languages often have pairs of verbs, such as *come* vs. *go*, *bring* vs. *take*, which are often interpreted according to whether the motion is toward or away from the place of speaking. Many other expressions in language are covertly place-deictic. For example, ‘the cat is behind the tree’ implies that the tree is between the speaker and the cat, and ‘Let’s go to the local pub’ implies that the pub is near to the place of speaking.

Lastly, *discourse deixis* involves the possibility of referring from one utterance backward or forward to others, as in ‘That was a good speech’ vs. ‘It made a sound like this: goooo.’ Discourse deixis grades into anaphora, the use of expressions such as pronouns to refer to entities introduced earlier, usually by fuller expressions. It is also closely related to *definiteness*, as in *the man* vs. *a man*, since all these phenomena rely on mental models of entities introduced into the discourse.

Considerable effort has gone into understanding how deixis works, and how it can be brought within a formal theory of meaning. But for the most part theoretical models underestimate the complexity and pervasiveness of deictic phenomena, and the richness of the contextual systems that support them. For example, good studies of actual deictic usage are rare, and the study of gesture is in its infancy (McNeill 2000).

2.2 Presupposition

Presupposition is a second major topic in pragmatics, and concerns the way in which propositions already presumed in a discourse context are usually not stated or questioned, but encoded in a more ‘background’ way. For example, ‘Has he stopped bothering you?’ presupposes the proposition that you and I know that he has been bothering you, and asks whether this has stopped. The classical test for presupposition is survival under negation: ‘He hasn’t stopped bothering me’ and ‘He has stopped bothering me’ both presuppose that he was bothering me. Languages have complex systems for foregrounding and backgrounding information in this way. Thus mental attitude verbs, such as *know* or *regret*, or change of state verbs such as *start* and *stop*, presuppose their complements, definite descriptions (such as ‘the King of Buganda’) presuppose the existence of the entities referred to, iteratives like *again* (as in ‘John did it again’) presuppose earlier occurrences, and so on. The phenomena and corresponding explanations are complex, and are dealt with in a separate article (see *Linguistic Presupposition*), but the relevance to pragmatics is that presupposition implies clearly that natural languages are built to trade on, and signal, the dependency of utterances on propositions already taken for granted.

The pragmatic aspects of the phenomena are often underplayed in semantic accounts of the phenomena. For example, consider the sentence ‘Sue cried before she finished her thesis’. This would normally presume that she finished her thesis, this being a presupposition from the *before*-clause. But the minimally different sentence ‘Sue died before she finished her thesis’ seems to make no such presumption, because of course we happen to know that the dead do not complete theses. This *defeasibility*, or cancelation of an inference in the context of contrary assumptions, is a hallmark of pragmatic inference.

2.3 Speech Acts

A third major topic in pragmatics concerns the general force or point of utterances, and is usually discussed under the rubric of speech acts. We can take a simple underlying proposition such as ‘The wine is (put) on the table,’ and we can embody it in utterances with different *illocutionary force* (in the terminology of the philosopher J. L. Austin). We can assert or vouch for the truth of the proposition, as in the assertion ‘The wine is on the table.’ We can ask whether it is the case, as in ‘Is the wine on the table?’ We can demand that the proposition be made true, as in ‘Put the wine on the table,’ and we can forbid anyone making it true, as in ‘Do not put the wine on the table.’ Languages typically indicate a handful of such illocutionary forces in the grammar of sentence types, as here in the English unmarked statement order, the marked inversion of yes–no questions, or the covert subjects of imperatives. Beyond statements, questions, and imperatives, most languages have in addition minor sentence types, e.g., exclamatives, as in ‘What a wonderful dress,’ or hortatives like ‘Let’s start,’ or imprecatives like ‘Damn you!’ and other more exotic types (e.g., special forms for blessings, curses, warnings). Thus the grammars of languages seem to presume a set of fundamental uses for language, and there has been some philosophical and theoretical interest in these categories. However, actual usage is much more complicated than this discussion would suggest. For example, one can use the statements ‘I suppose the wine is on the table’ to question whether it is, or ‘I would like the wine on the table’ to request that it be put there, or ‘The wine should not be placed on the table’ to forbid it. Some of these uses can be explained as secondary or derivative uses of one sentence type to perform a speech act prototypically associated with another. For example, *I would like ...* could be said to indirectly hint, rather than directly request, and can thus be described as an indirect speech act. What this discussion makes clear is that it is important to keep apart the grammar of sentence types from the illocutionary forces or speech acts they can be used to perform.

There has been much work on grammatical sentence types (and what the grammarians call *mood*), speech

acts, and indirect speech acts, and their relationship to one another. This is the subject of a separate article, but here one major point is that these phenomena indicate directly that natural languages are designed not just as abstract systems, but as tools for human communication, a point of view underlined in functional approaches to language structure (see *Grammar: Functional Approaches*). Speech acts can also be seen to be fundamentally context dependent. First, speech acts are dependent on, and contribute to, the context in which speech is taking place. One way of seeing this is to construe speech acts as operations on the context (conceived of as sets of propositions taken for granted): a statement adds a proposition to the context, a question requests whether such a proposition is to be added, a denial removes one, and so on. Note too how a permission presumes that a prohibition would otherwise be in force. Second, the interpretation of a sentence as performing a specific speech act is obviously context dependent. Consider the sentence ‘What are you doing tonight?’ in, say, the following discourse context where it has a straightforward informational query status:

A: ‘I can’t wait for this evening.’

B: ‘What are you doing tonight?’

Now compare its role in a different context:

A: ‘What are you doing tonight?’

B: ‘Oh I don’t know, not much.’

A: ‘Would you like to come to the movies?’

Here the utterance checks out a condition on an invitation, and in so doing can be seen as precluding such an invitation (a ‘pre-invitation’), which is then (given an appropriate response) expected to follow. Sequences of utterances thus set up specific expectancies, crucial in interpretation—a matter dealt with more fully below.

2.4 Conversational Implicature

So far we have considered ways in which the structures of languages build in presumptions of use, as in pronouns, demonstratives, verbs of presupposition, or sentence types as indicators of speech acts. In these cases, the semantics of the expressions trigger the pragmatic inferences, which are themselves subject to pragmatic resolution. But another very important aspect of pragmatics is concerned with inferences that are invited—or *implicated*—rather than required or triggered by the semantics: often, synonymous expressions will have different implicatures. The philosopher H. P. Grice noted that there are a number of different ways in which such invited meanings can be invoked, all of which point to a set of background assumptions about how language should be used, which he formulated as a set of maxims of conversation. He sketched four such maxims, with attendant sub-maxims: the maxim of Quality (‘Say what you believe to be true’); the maxim of Relevance (‘Make what you

say relevant and timely'); the maxim of Quantity ('Don't say more or less than is required'); and the maxim of Manner ('Be brief and clear'). Then he showed that if a speaker violates them overtly, the speaker is generally credited with some implicit, indirect meaning—for example, if it is raining, and the speaker says 'Thank God I didn't bother to bring my umbrella,' thereby clearly violating the maxim of Quality, an ironic interpretation is generated. (Any such interpretations which are suggested by reference to the maxims, Grice called *conversational implicatures* to distinguish them from logical implications). If on the other hand the speaker just seems to be following the maxims, then additional inferences are generated. For example, if you say 'There are five gallons in the tank,' I'll assume that there are no more than five gallons in the tank, because otherwise by the maxim of Quantity ('say as much as is required') you should have said so (note that if there are more than five gallons, then of course what you have said is certainly true, but is also misleading). Grice went on to note a number of interesting applications of his theory of implicature. For example philosophers had long disagreed about just how close the ordinary language words 'if,' 'or,' and 'and' are to their logical counterparts. Grice argued that the divergence between ordinary language and logic was due largely to the implicatures of language usage. For example, if I say 'Glenn was an astronaut or he was a senator,' I suggest that he was not both, but this does not mean that natural languages *or* is not equivalent to logical (inclusive) disjunction. That is because if I knew he was both, I should (following the maxim of Quantity) have said so; since I didn't, I imply that I believe he was one or the other, but not both.

Although modern versions of implicature theory use somewhat different 'maxims,' they follow Grice's idea that the presumption of background principles of language use can generate many detailed inferences that are suggested but not entailed by what has been said. There are two main branches of contemporary theory. One takes the line (following the trend in the cognitive sciences) that these background principles are innate cognitive mechanisms of information processing. The other branch follows Grice more closely, and argues that these principles follow from rational design characteristics of communication. The former line is more generally concerned with the nature of inference in communication, while the second 'neo-Gricean' line of argumentation has been developed more narrowly to give explanations of linguistic facts, and is thus of more direct relevance to linguistic theory, and so is the focus of the following remarks.

One central attraction of the Gricean approach is that it promises to simplify the kinds of meanings we attribute to expressions. For example, as mentioned above, English *or* seems normally to have an exclusive use (as in 'The book is either in the bedroom or the sitting room'), but obviously this is not always so

(consider 'I lent you my pen or my pencil, or both'). Grice's tactic is to avoid positing an ambiguity (between inclusive and exclusive *or*), by assuming that *or* has a wide, general meaning (e.g., the inclusive meaning), which is then specialized in context where appropriate by an extra pragmatic inference—namely the Quantity implicature 'not both.' By generalizing Grice's observations we can note, first, that any pair of expressions (such as *and*, *or*), where the use of one expression would entail the other (as in *p and q* implying *p or q*) but not vice versa, form a scale of informativeness—the use of *and* is more informative than *or*, because *p or q* allows not only for the circumstance that *p*, or that *q*, but also for the circumstance that *p and q*. In short, *p and q* rules out more alternative states of affairs. Then we can also note that, for any such pair of expressions, asserting the weaker or less informative member will implicate the inapplicability of the stronger expression (for if you had meant that, you should have said it). Hence asserting *p or q* implicates that one is not in a position to assert *p and q*. Now consider many other such ordered pairs, where each is of the form <strong, weak>:

<*and*, *or*>, <*all*, *some*>, <*certain*, *possible*>, <*four*, *three*>, <*must*, *may*>, <*hot*, *warm*>

For each of these, asserting the weaker, less informative expression will implicate that the stronger expression does not apply, as in 'Some of the students are punctual,' which, other things being equal, will implicate that not all of the students are punctual. Modern treatments of these so-called *scalar implicatures* are quite developed (Atlas 2000, Gazdar 1979, Horn 1989, Levinson 2000).

Many further systematic observations can be made. Consider the contrast between 'John could solve the problem' vs. 'John had the ability to solve the problem'—the first suggests that he did solve it, while the second suggests he didn't. The opposition is one between a direct lexical coding with the verb *can* and a more periphrastic expression *have the ability to*. More generally, the opposition is between any simple, direct encoding, which invites a rich interpretation to the stereotypical scenario, and a less direct, unusual, more prolix or *marked* expression, which suggests a complementary interpretation. Consider, for example, 'He went downtown and bought a coat,' which suggests a single complex action (a coat-buying expedition), vs. 'He went downtown and in addition he bought a coat,' which suggests two independent actions. These opposing tendencies of interpretation can be related to *iconicity* in language (see *Linguistics: Iconicity*), but what gives them reliability is mutual expectations between speaker and addressee—that is, that they follow from general principles or maxims.

Contemporary neoGricean theory recognizes two or three maxims: for example, a maxim of Quantity (or

Q-principle, giving us the scalar implicatures mentioned above); a maxim that maximizes information from direct, unmarked expressions (an Informativeness or I-principle); and a maxim of Manner or markedness (or M-principle) that curtails those informative interpretations (Levinson 2000). Three such principles suffice to capture many detailed inferences which interact closely with linguistic structure, and they can be thought of as stable heuristics which serve to amplify message content with default inferences.

In this way, Grice's ideas can be given both precision and generality. The description of many crucial aspects of meaning (for example, those associated with the natural language connectives, quantifiers, and modals) can be simplified and improved by taking into account these inferences which are closely associated with expressions by general principle, but not actually encoded in them. Such inferences have a default or presumptive character, and may be confused with the semantics and syntax of expressions.

These observations have considerable importance for linguistic theory. If they are pushed home, they imply a rather different view about the relation of language competence to nonlinguistic abilities than is usually assumed. For example, it turns out that pragmatics cannot interface with semantics in the way originally imagined: as a module or system of principles that, given the proposition coded in an utterance, would calculate additional noncoded inferences. The problem is that pragmatic inferences clearly play a central role in deriving the proposition in the first place. For example, in 'It is better to drive home and drink than to drink and drive home' the proposition expressed depends on the antecedent construal of *and* as temporarily asymmetrical (i.e., as 'and then'), which is a matter of implicature. Thus what is coded semantically and what is inferred pragmatically are often necessarily integrated in the extraction of propositional content. The implication is that there is a single level of representation to which quite different kinds of principles contribute—semantics and logical inference on the one hand, and pragmatics and presumptive reasoning on the other.

Another possible radical implication is that many correlations between form and meaning that have been attributed to syntax are more properly ascribed to pragmatics. For example, it is normally supposed that in a sentence like 'The chairman voted for him,' the chairman and the man voted for must be distinct by a rule of grammar. But in 'No-one else voted for the incumbent chairman—only the chairman voted for him' it is possible to understand the chairman as voting for himself. Arguably, this is because normally one would express such coreference by the reflexive *himself*, and the use of the nonreflexive pronoun thereby implicates a non-coreferential reading, unless there are other reasons to use it. Here there are such other reasons—'Only the chairman voted for himself'

would imply that everybody else voted for people other than themselves, and that may not be what is intended. Many such reductions of alleged syntactic patterns to general patterns in pragmatics may be possible (see Huang 1994, Levinson 2000).

2.5 Conversational Structure

Conversational inferences seem to be of two kinds: first the implicature type, which relies on general background presumptions or heuristics; and second, inferences that are tied to more specific details of conversational interchange. Conversation—or the normal, informal exchange of speech, where there is rapid alternation of speakers—can be contrasted with more formal, specialized types of speech exchange (as in chaired meetings, interviews, or religious services). Needless to say, conversation is the primary form of human verbal interaction, the context in which all primary language-learning is accomplished and many details of linguistic structure are tied to it intimately. For example, a primary property of conversation is rapid turn-taking, and with this turn-taking the deictic centering also alternates: what was 'I' is now referred to as 'you,' what was 'here' becomes 'there,' and so forth, an alternation that children initially find difficult.

The characteristics of turn-taking (and indeed conversation in general) have been studied primarily by sociologists (under the rubric of conversation analysis). In the model advanced by Sacks and Schegloff, conversational turn-taking is organized as a sharing system whereby the current speaker has rights only to the end of a sentential construction or turn-taking unit, at which point transition to another speaker is governed by an ordered set of rules—if the speaker has selected a next speaker, that party should speak next, if not, other speakers may start speaking, failing which the original speaker may continue. These rules explain many details of conversational patterning, such as just where overlaps occur (e.g., as competing starts). They also explain the differential significance of silence—if no-one decides to speak after a unit which selects no following speaker, silence is likely to be interpreted simply as a lapse in conversation, but after a next speaker is selected it is heard as a significant withholding of speech (Levinson 1983, p. 300, after Atkinson and Drew):

A: *Is there something bothering you or not?*

B: (1.0 second silence)

A: *Yes or no?*

B: (1.5 seconds silence)

A: *Eh?*

B: *No.*

Differential meanings attributed to silence eloquently make the point that the semiotic significance of utterances is very much conditioned by the locus in which they occur. Consider, for example, that after a

request, an acceptance or refusal is due, but these options are not equal: an acceptance is usually delivered immediately in simple form, whereas a refusal is typically delayed, indirect, and accompanied by excuses. Recollect too, that as noted above, many speech acts are preceded by an utterance that checks out whether the conditions for that action are met—thus one can hint at an upcoming request for an appointment by asking whether the addressee would be available at a certain time. Then we can readily understand the following, where a telephone caller C already interprets a two-second pause as a negative answer (from Levinson 1983, p. 320):

C: *So I was wondering would you be in your office on Monday*

(micro-second pause)

by any chance?

(2.0 second silence)

C: *Probably not*

Here the interpretation relies (a) on the asymmetry between what the conversation analysts call *preferred* and *dispreferred* responses to specific speech acts, and (b) on the understanding that the question is not idle, but functions as a prelude to a request.

The study of conversation thus shows that the attribution of meaning to utterances depends crucially on sequential location. Arguably, this is exactly what Grice had in mind by his maxim of Relevance, where the implicatures of an utterance would also derive from the expected goals of the speaker at a particular point in discourse. Goal-driven accounts of conversational understandings have been explored most thoroughly in artificial intelligence circles, where the construction of a machine that could enter into functional dialogues has been a long-term objective.

3. Conclusions

In addition to the topics reviewed here, there are pragmatic perspectives on many other phenomena, for example register, style, jargon, and other specializations of language, the contrastive use of languages by multilingual speakers (see *Code Switching: Linguistic*), and social constraints on the use of language (see, for example, *Politeness and Language*). From research on all these topics, an overall picture is emerging of the systematic relations between semiotic systems and their contexts of use. First, any sign system or language makes systematic provision for contextual features, as in deixis or presupposition. Second, what is actually coded in signs or linguistic expressions is only a small part of the meaning attributed in context, the remainder being generated by presumptions of use. Third, the coded meaning or semantics is usually specified much less fully than conventional wisdom implies. Consider a road sign depicting falling rocks—the depicted content is specific yet hardly conveys the full message: we must add the illocutionary force ('Beware!'), and the deictic inference ('in this vicinity'

or 'the next kilometer'). Such a sign contrasts with other warning signs—for instance, a simple exclamation mark—which indicates general danger but implies that the danger is other than rocks (since otherwise the rock sign would have been used). All such signs contrast with no sign, which implies no special foreseeable danger. In similar but much more complicated ways, pragmatic principles remain the essential basis for linguistic communication. Because lexical limitations and the bounds of learning and memory constrain natural language semantics deeply. In addition, the relatively slow articulation of speech compared with understanding processes puts a clear premium on pragmatic inference. A pragmatic perspective offers fundamental insights into both the structure and use of language.

See also: Communication and Social Psychology; Deixis; Lexical Semantics; Semantic Similarity, Cognitive Psychology of; Semiotics; Verbal Communication, Psychology of

Bibliography

- Atlas J in press *Logic, Meaning and Conversation*. Oxford University Press, New York
- Brown G 1995 *Speakers, Listeners and Communication*. Cambridge University Press, Cambridge, UK
- Clark H H 1996 *Using Language*. Cambridge University Press, Cambridge, UK
- Cole P (ed.) 1981 *Radical Pragmatics*. Academic Press, New York
- Davis S (ed.) 1991 *Pragmatics: A Reader*. Oxford University Press, Oxford, UK
- Fillmore C J 1997 *Lectures on Deixis*. CSLI Publications, Stanford, CA
- Frawley W 1992 *Linguistic Semantics*. Lawrence Erlbaum Associates, Hillsdale, NJ
- Gazdar G 1979 *Pragmatics: Implicature, Presupposition and Logical Form*. Academic Press, New York
- Green G M 1996 *Pragmatics and Natural Language Understanding*, 2nd edn. Lawrence Erlbaum Associates, Mahwah, NJ
- Grice H P 1989 *Studies in the Way of Words*. Harvard University Press, Cambridge, MA
- Horn L R 1989 *A Natural History of Negation*. University of Chicago Press, Chicago
- Huang Y 1994 *The Syntax and Pragmatics of Anaphora: A Study with Special Reference to Chinese*. Cambridge University Press, Cambridge, UK
- Hutchby I, Wooffitt R 1998 *Conversation Analysis: Principles, Practices and Applications*. Polity Press, Cambridge, UK
- Jarvella R J, Klein W (eds.) 1982 *Speech, Place and Action: Studies in Deixis and Related Topics*. Wiley, New York
- Kasher A (ed.) 1997 *Pragmatics: Critical Concepts*. Routledge, London
- Levinson S C 1983 *Pragmatics*. Cambridge University Press, Cambridge, UK
- Levinson S C 2000 *Presumptive Meanings*. MIT Press, Cambridge, MA
- McNeill D (ed.) 2000 *Language and Gesture: Window into Thought and Action*. Cambridge University Press, Cambridge, UK

- Mey J L (ed.) 1998 *Concise Encyclopedia of Pragmatics*. Elsevier, Oxford, UK
- Mey J L 1993 *Pragmatics: An Introduction*. Blackwell, Oxford, UK
- Morris C W 1938 Foundations of the theory of signs. In: Neurath O, Carnap R, Morris C (eds.) *International Encyclopedia of Unified Science*. University of Chicago Press, Chicago (reprinted in Morris C W 1971 *Writings on the Theory of Signs*. Mouton, The Hague)
- Nerlich B, Clarke D D 1996 *Language Action and Context: The Early History of Pragmatics in Europe and America, 1780–1930*. Benjamins, Amsterdam
- Turner K (ed.) 1999 *The Semantics/Pragmatics Interface from Different Points of View*. Elsevier, Oxford, UK
- Verschueren J 1999 *Understanding Pragmatics*. Arnold, London
- Yule G 1996 *Pragmatics*. Oxford University Press, Oxford, UK

S. C. Levinson

Pragmatism: Philosophical Aspects

Pragmatism is a US philosophical movement which began in the late nineteenth century and exerted significant influence until around the time of John Dewey's death in 1952. Pragmatism then fell from favor but began to enjoy a revival at the end of the twentieth century. The 'classical' pragmatists are C. S. Peirce, William James, and John Dewey. Their philosophies differed a good deal but all defended broadly empiricist views of knowledge which reject many traditional empiricist psychological views, and emphasize the role of thought in guiding action. Richard Rorty is the most influential contemporary figure who aligns himself with pragmatism.

1. 'Classical' Pragmatism

The most important pragmatist philosophers are C. S. Peirce, William James, and John Dewey. These three are often referred to as the 'classical' pragmatists. All three were born and worked in the USA. All were primarily philosophers but had strong interests in other disciplines. The views of the 'classical pragmatists' do have important common themes, but are also dissimilar in many ways.

Pragmatism is an empiricist view of thought and knowledge which rejects much of the psychological picture usually associated with empiricism. Pragmatists think traditional empiricism has had too passive a view of the mind, and pragmatists place great emphasis on the role played by thought in guiding action and solving practical problems. Pragmatism also rejects attempts to understand human knowledge by tracing its connections to some special set of

'foundational' beliefs. For pragmatism, both specific beliefs and methods of inquiry in general should be judged primarily by their consequences, by their usefulness in achieving human goals.

The nature of truth has been a central topic for pragmatists, and a source of much trouble for them. Pragmatists in general reject 'correspondence' theories of truth. Correspondence theories claim that a true belief or statement is one which represents the world as it really is. After rejecting correspondence, pragmatists have had a difficult time devising an alternative view of truth.

Pragmatism has an often complicated relationship with science, but it is fair to say that both Peirce and Dewey regarded science as a distinctive and superior way of acquiring knowledge. For both, the scientist is in some respects the model or ideal inquirer. Although pragmatists stress the role of knowledge in solving practical problems, it is definitely an error to see pragmatism as claiming that the direction of research in science, or choices between rival theories, should be guided solely by practical or commercial demands.

Pragmatists tend to argue that there is an important place in the universe for human choice and initiative; they oppose philosophical systems which regard the world as in some sense 'finished,' 'complete,' or impervious to the effects of human choice. This is a rather vague thing to oppose, but it is common for pragmatists to be constantly on the lookout for any hint of a denial of the significance of human choice. This concern pervades much pragmatist writing. Pragmatists also tend to hold generally 'humanistic' perspectives on matters involving morals and values, opposing both nihilistic views that reject all moral thinking as illusory, and views that locate moral and other evaluative facts outside the everyday world of human well-being.

2. Charles Sanders Peirce (1839–1914)

Peirce's career began well but ended very badly, without his ever having held a long-term academic job. Peirce never set out his ideas in an organized and systematic way. It is difficult and probably fruitless to try to form these ideas into a single coherent system. Peirce's most famous works are about the relations between doubt, inquiry, belief, and action. He also wrote a good deal about logic, the theory of signs (which he called 'semiology'), some speculative cosmological ideas, probability, measurement, and other topics in the philosophy of science.

Peirce argued that inquiry always begins with 'real and living doubt' which he distinguished from the feigned doubt of Descartes. Doubt prompts inquiry, which aims at belief. Peirce claimed that the essence of belief is the establishment of a habit of action. Alexander Bain had made this suggestion in the mid-

nineteenth century, within the framework of a more orthodox associationist psychology. Peirce and later pragmatists made this link between belief and action central to their theories of mind and knowledge.

Peirce believed that science is the most effective method for relieving doubt and acquiring useful habits of action. His defense of science stressed the social properties of scientific inquiry, especially the tendency scientific testing has to produce a 'convergence' of belief among people who start from different positions.

In his most influential paper, 'How to make our ideas clear,' Peirce argued for a principle which later became famous as 'the pragmatic maxim.' This principle is often seen as the single idea most central to pragmatism. Peirce said that to work out what the meaning of a 'conception' is, we should work out 'what effects, that might conceivably have practical bearings, we conceive the object of our conception to have. Then, our conception of these effects is the whole of our conception of the object.' This famous principle is quite obscure, and was later paraphrased by James, Peirce, and others in a number of conflicting ways. One reasonable view is to see the famous 'maxim' as expressing a fairly familiar empiricist view, that our concept of something is just our concept of its possible effects on our experience. Then the novelty in 'How to make our ideas clear' lies in Peirce's attempt to derive this empiricist principle from an unusual starting point—a theory of the role of belief in action.

In 'How to make our ideas clear' Peirce also gave a famous definition of truth. Peirce said that what we mean by the truth is the opinion which is 'ultimately fated to be agreed to by all who investigate,' or in a more modern formulation, the opinion that will be held at the limit of scientific inquiry. This is usually taken to be a novel *definition* of truth, in conflict with more familiar views such as the correspondence theory. In fact it is difficult to work out how Peirce intended his theory of truth to be taken. In some places it looks more like an expression of faith in the power of science to *find* the truth than an unorthodox philosophical theory of what it is for an idea to be true.

3. William James (1842–1910)

James made pragmatism famous, especially in his 1907 book *Pragmatism* (repr. 1979), which is the most widely read of all pragmatist writings.

James had a brilliant and influential career based at Harvard, in both philosophy and psychology. He wrote many important works before he began calling himself a 'pragmatist' and aligning his views with Peirce, in 1898. In these earlier works James argued that we have the right actively to choose sides on momentous, unresolved philosophical issues that will affect how we live our lives. Choices made on these issues will inevitably be strongly affected by our individual temperaments; we should not expect all individuals to find the same philosophies appealing.

Problems James applied this doctrine to include the existence of God, the freedom of the will, and the reality of moral facts. In this earlier period James also published a landmark work in psychology, his two-volume *Principles of Psychology* (1890, repr. 1950).

After 1898 James began to embed his characteristic themes within an overtly 'pragmatist' framework. Unlike Peirce, James took pragmatism to be continuous with the mainstream empiricist philosophies of English thinkers such as David Hume and J. S. Mill. James publicized pragmatism with great energy, and transformed it in ways that induced some distress in Peirce.

James presented pragmatism as a way to avoid the errors of the two key rival philosophies of his day—overly scientific materialism and empiricism on the one hand, and sentimental, over-optimistic religious idealism on the other. He saw pragmatism as a way of distilling the genuine human significance from obscure philosophical theories and debates. James also argued that there is an inevitable and appropriate role for values in choices about theoretical issues, defended the role of human choice as a factor in determining the ultimate nature and significance of the universe, and defended the legitimacy of individual religious experience.

James made pragmatism more 'practical' than Peirce did, linking it to problems of everyday life. James was a less scientific thinker than Peirce, and also more individualistic than Peirce or Dewey. James talked of assessing ideas in terms of their 'cash value,' an unfortunate phrase which led many to interpret pragmatism as vulgar and anti-intellectual. James also invited a stream of criticism through his tendency to discuss the nature of truth in a vague and simplistic way. He said at various times that true ideas are just those that can be verified, or those that are useful. Some of the criticisms of James rested on misinterpretations, but there is no question that he wrote in an imprecise way and was not the most consistent of philosophers. Despite this, James's writings have endured. His discussions of the role of individual temperament in philosophical choices never seem dated, for example. Another reason for James's continuing appeal is the way his attractive personality, imagination, and boundless energy seem to radiate from the pages of his work.

4. John Dewey (1859–1952)

Dewey's work is the high point of the pragmatist tradition to date. In the course of a remarkably long and productive career, Dewey published in virtually all areas of philosophy and in psychology, education, and politics. He wrote both popular, accessible works and dense, uncompromising ones. Dewey spent his most important years at the University of Chicago, where he worked in philosophy, education, and psychology, and after that at Columbia, where he

focused more on philosophy and was also active in politics.

Dewey's work went through several phases. Early in his career he accepted an idealist philosophy of the type influenced by Hegel, and also had strong Christian interests. Around the turn of the century he turned toward a more scientific approach to philosophy, influenced by Darwin and by James's *Principles of Psychology*. Around this time Dewey also dropped his commitment to Christianity, although his attitude to religious ideas and to the psychological and social role of religion remained complex.

The scientifically oriented philosophy Dewey developed was initially reminiscent of James's pragmatism. Later Dewey began to recast some pragmatist themes within a more 'naturalistic' framework—a framework which sets out from a biological description of living organisms and their relations to their environments. Dewey argued that intelligence is a means for humans to transform their environments, in order to deal with the problems posed by uncertainty and change in natural events.

In *Experience and Nature*, perhaps his greatest work, Dewey both defended this naturalistic view of mind and knowledge, and criticized the philosophical tradition for its postulation of false divides or 'gulfs' between mind and matter, thought and object, and so on. Dewey thought that the philosophical tradition is plagued by 'dualisms' which lead to pseudoproblems. The source of these dualisms is a 'split in being' established by the ancient Greeks, a split between the 'perfect, permanent, self-possessed' and the 'defective, changing, relational.' Dewey sought to replace these dualisms with a view based on various kinds of 'continuity,' between mind and nature, between organism and environment, and between cognition and simpler biological capacities. For Dewey, these natural continuities also provide the material needed to resolve other traditional oppositions such as that between the theoretical and practical, and between fact and value. Much of Dewey's later work is an unusual mixture of scrupulously careful system building based on assertions of 'continuity,' and sweeping historical surveys, tracing the histories of key philosophical errors back and forward through thousands of years.

Dewey also defended a social theory of mind, claiming that because thought is symbolic and symbolism is social, thought only exists within a language-using community. He argued that the primary role of science is to help human societies deal with problems and control their environments. But, like Peirce, Dewey thought that it was unscientific for scientists to direct their work according to *specific* practical problems. Rather, science is the study of a special class of properties of natural processes—relations and connections between things, which Dewey called 'instrumental' properties. Science is most successful in expanding our capacities for problem solving when it is directed on the study of the instrumental properties

of things in an open-ended way unconstrained by immediate practical applications.

Dewey thought that the proper role of knowledge is to enable humans to transform their situations in beneficial ways. This led Dewey to reject what he called the 'spectator theory of knowledge,' the view that the ideal knower is someone who registers what is going on but does not intervene. This also led Dewey into some of the same arguments James had with defenders of correspondence theories of truth. Dewey thought that theories of truth as correspondence were aimed at inventing magical relations between thought and the world in order to overcome what he saw as a nonexistent problem—the problem of how the mind and the external world could have any contact with each other. He also thought correspondence theories belong with a spectator theory of the role of knowledge.

Dewey wrote often of the role played by knowledge in 'transforming' or 'reconstructing' the world. This 'transformation' should not be understood as implying an idealist metaphysical view. Dewey thought that knowledge is a factor in changing things in the world because knowledge has a role in guiding *action*, which transforms things by means of ordinary physical relationships.

Dewey was a significant figure in US political and social thought. Here he defended a version of liberalism, and he wrote extensively about the proper structure of a democratic society. At various times he was attacked by both the right and the left. He hoped for a more 'democratic' economic order, but always kept Marxism and communism at arm's length. Dewey's best known political activity was his chairing of an international inquiry into Stalin's trial of Trotsky. The inquiry entitled its report 'Not guilty.'

At Chicago Dewey worked closely with George Herbert Mead, who became one of the founders of social psychology in the USA. Mead is sometimes grouped as a major pragmatist thinker (see, for example, Scheffler 1986, Joas 1993). Mead argued that human individuality is a product of social embedding; that cognition in general and specific mental developments such as a sense of self emerge as products of various kinds of symbolic behavior within a community. Dewey also wrote about art and took a lifelong interest in education. When he was at Chicago he established an experimental school. Dewey favored problem solving as an approach to all types of learning. His ideas remained influential in education for many years while his name had almost dropped out of mainstream English-speaking philosophy.

5. After Dewey

Dewey's ideas, and pragmatism in general, subsided from philosophical discussion for several decades after Dewey's death in 1952. A primary reason for this was

the growing focus on formal logic and the philosophy of language in US philosophy after World War II, prompted especially by the immigration of European philosophers and logicians who defended a rigorous 'analytic' style of philosophy. After Peirce, the pragmatists had not taken an interest in formal logic. Dewey also had idiosyncratic and rather opaque ideas on the philosophy of language, which rapidly came to seem dated.

More recently, interest in pragmatism has revived. In part, this has been due to the influence of Richard Rorty, who has both revived some of Dewey's ideas and become a significant pragmatist thinker in his own right. But aside from Rorty's explicit advocacy, quite a number of the key figures in mid- to late twentieth-century English-speaking philosophy have some connection with pragmatism. W. V. O. Quine is sometimes seen as a pragmatist, because of his opposition to foundationalist projects in epistemology, and his view of rational thought as a matter of making adjustments to a 'web of belief.' For Quine, these adjustments to the web are made in whatever ways seem to work best in dealing with experience; no beliefs are in principle immune to revision. In the closing sentences of his immensely influential paper 'Two dogmas of empiricism' (1953), Quine refers to these adjustments made to beliefs in the face of evidence as 'pragmatic.' But Quine's theory focuses on the role of beliefs in prediction, and does not discuss the link between thought and action in the detail associated with classical pragmatism.

Hilary Putnam is another influential defender of some broadly pragmatist ideas. Putnam's early work defended a strongly 'realist' set of views about the relation between thought and the external world. But Putnam changed his mind, and defends a position he calls 'internal realism,' which rejects the correspondence theory of truth and which Putnam sees as continuous with the pragmatist tradition. Putnam has defended views about truth which resemble (one reading of) Peirce's view, for example. Donald Davidson, another giant figure in late twentieth-century philosophy, is occasionally classified as a pragmatist, largely because of his opposition to some standard doctrines about the representational properties of language and thought, and his opposition to some versions of the correspondence theory of truth. But Davidson's connection to classical pragmatism is tenuous.

Rorty's work has carried the influence of pragmatism outside of philosophy into neighboring humanistic disciplines. Rorty views himself as carrying on the tradition of Dewey, and to a lesser extent James. Rorty is an 'antiessentialist' about traditional philosophical concepts such as truth, knowledge and justice. Antiessentialism in this sense holds that the only theory there can be of truth (etc.) is a theory of how the term 'true' functions in ordinary discourse; it is an error to give a philosophical analysis of truth that

goes beyond this. The degree of continuity between Rorty's ideas and Dewey's is a matter of controversy. As outlined above, Dewey's mature work includes a mixture of sweeping historical stories aimed at dissolving philosophical problems, and careful system building of his own. Rorty endorses the historical dissolving but not the system building. However, it can be argued that Dewey's critical points and his attempts to 'get over' standard debates tend to depend on his positive philosophical theories. And James used pragmatism to breathe new life into metaphysical problems, not to deflate them. As Rorty's views on traditional philosophical questions are so focused on dissolving and deflating, it is in some ways more appropriate to associate Rorty with the later work of Ludwig Wittgenstein rather than with pragmatism.

In recent discussion, the term 'pragmatist' has come to be used in a very loose way, to signify endorsement of any one of a wide range of ideas that are taken to have a connection with Peirce, James, or Dewey. These ideas include holism about justification, opposition to correspondence theories of truth, opposition to representationalist theories of thought and language, opposition to traditional views about logic and rationality, and almost any kind of emphasis on the practical, action-guiding side of reasoning and belief. Contemporary writers who identify themselves with 'pragmatism' often have very little in common. These loose uses of the term have led to confusion. Perhaps any contemporary 'pragmatism' should hold, at a minimum, that the role of thought in guiding action is of central importance to the philosophy of mind and epistemology, and as a consequence of this there is a need to rework standard views about knowledge and inquiry. But any summary one might give runs into the problem that even 'classical' pragmatism was a diverse and shifting collection of ideas.

See also: Dewey, John (1859–1952); Empiricism, History of; Irrationality; Philosophical Aspects; James, William (1842–1910); Mead, George Herbert (1863–1931); Social Science, the Idea of; Wittgenstein, Ludwig (1889–1951)

Bibliography

- Buchler J (ed.) 1955 *Philosophical Writings of*. Peirce. Dover, New York
- Dewey J 1969–72 *The Early Works* [ed. Boyston J A]. Southern Illinois University Press, Carbondale, IL
- Dewey J 1976–83 *The Middle Works* [ed. Boyston J A]. Southern Illinois University Press, Carbondale, IL, especially Vols. 13 and 14
- Dewey J 1981–90 *The Later Works* [ed. Boyston J A]. Southern Illinois University Press, Carbondale, IL, especially Vols. 1, 4, and 10
- James W 1950 *Principles of Psychology*. Dover, New York
- James W 1956 *The Will to Believe and Other Essays in Popular Philosophy*. Dover, New York

- James W 1979 *Pragmatism and the Meaning of Truth*. Harvard University Press, Cambridge, MA
- James W 1983 *The Varieties of Religious Experience*. Penguin, Harmondsworth, UK
- Joas H 1993 *Pragmatism and Social Theory*. University of Chicago Press, Chicago
- Mead G H 1934 *Mind, Self and Society* [ed. Morris C W]. University of Chicago Press, Chicago
- Peirce C S 1931–58 *Collected Papers* [Vols. 1–4 ed. Hartshorne C, Weiss P, Vols. 7–8 ed. Burks A]. Harvard University Press, Cambridge, MA, especially Vol. 5. A new edition of Peirce's works is in preparation
- Putnam H 1990 *Realism with a Human Face*. Harvard University Press, Cambridge, MA
- Quine W V O 1953 Two dogmas of empiricism. In: (ed.) *From a Logical Point of View*. Harvard University Press, Cambridge, MA, pp. 20–46
- Rorty R 1982 *Consequences of Pragmatism*. University of Minnesota Press, Minneapolis, MN
- Rorty R 1991 *Objectivity, Relativism and Truth: Philosophical Papers Volume I*. Cambridge University Press, Cambridge, UK
- Scheffler I 1986 *Four Pragmatists*. Routledge, New York
- Schillp P A, Hahn L E (eds.) 1989 *John Dewey*, 3rd edn. Open Court, La Salle, IL
- Thayer H S 1981 *Meaning and Action: A Critical History of Pragmatism*. 2nd edn. Hackett, Indianapolis, IN
- Westbrook J 1991 *John Dewey and American Democracy*. Cornell University Press, Ithaca, NY

P. Godfrey-Smith

Pragmatist Social Thought, History of

Pragmatism is the name usually given to the classical philosophical movement in the USA. It emerged in the last decades of the nineteenth century and achieved a certain hegemony in the country's intellectual life during, and immediately after, the so-called Progressive Era (1896–1914). From the 1930s, and even more after 1945, it was largely displaced by other currents of thought in philosophy, the social sciences, and public political discourse. American pragmatism also attracted considerable attention in Europe, especially around 1910, though exposition and critical discussion of pragmatism displayed widespread misunderstandings and the tendency to depreciate it by reducing it to an expression of alleged American national characteristics.

1. Pragmatism Defined

The everyday sense of the term 'pragmatic' certainly contributed to these misunderstandings, connoting as it does a kind of 'muddling through' which is oriented to immediate requirements, disregards theoretical or moral principles, and treats the given features of the

situation simply as part of a calculation. The term pragmatism has the same Greek roots as praxis, practical, etc. The founder of pragmatism, Charles Sanders Peirce (1839–1914), coined the term as a result of his reflections on Kant's use of the terms 'pragmatic' and 'practical.' Peirce's lectures and writings around 1878 are now regarded as the original documents of pragmatism. These first became known only to narrow circles of intellectuals in Cambridge, Massachusetts, and did not achieve wider notice until some 20 years later, when William James (1842–1910) referred to them in some lectures, but above all through his lectures on 'Pragmatism' published in 1907. Peirce came to distance himself from James's pragmatism and to call his own philosophy 'pragmaticism.'

Along with Peirce and James, the inner core of pragmatism is usually taken to include John Dewey (1859–1952) and George Herbert Mead (1863–1931). The differences between all these thinkers were so evident, not least to the 'pragmatists' themselves, that their subsumption into a single school or movement has constantly been questioned. The most widely accepted view, however, has been that despite the differences there is enough of a core of shared views to justify speaking of 'pragmatism' as a distinct philosophical orientation.

What are the basic themes of pragmatism? Peirce's approach starts from his critique of Descartes' methodological principle of radical doubt and the resultant program of making the thinking ego's certainty of itself into the firm foundation of a new philosophy. The pragmatist questions the meaningfulness of Cartesian doubt, not in order to defend unquestionable authorities against the emancipatory claim of the thinking ego, but arguing for a more substantial doubt, that is, the grounding of cognition in real problem situations. In pragmatism, the guiding idea of the doubting ego is replaced by the idea of a cooperative search for truth in order to overcome real problems of action. Real doubt occurs in action, conceived as a cyclical succession of phases. Thus every perception of the world and every action in it are anchored in an unreflective belief in self-evident conditions and successful habits. But these habitual ways of acting constantly come up against the resistance of the world, which is seen to be the source of the destruction of unreflective expectations.

The resultant phase of real doubt leads to a reconstruction of the interrupted context. Perception must grasp new or different aspects of reality; action must attach itself to other elements of the world or reorganize its own structure. This reconstruction is a creative accomplishment by the actor. If he or she succeeds, via a change in perception, in acting differently and thus getting going again, then something new has come into the world: a new way of acting which can be institutionalized and itself become an unreflectively followed routine. Thus pragmatists see all human action in the opposition between un-

reflective habits of action and creative accomplishments. This also means that creativity is seen here as characterizing achievements in situations which demand a solution, rather than the unconstrained creation of something new without a constitutive background in unreflective habits.

From this basic model of pragmatism, in which action and cognition are combined in a particular way, we can derive the other central claims of pragmatism. In the metaphysics of pragmatism, reality is not deterministic: rather, it enables and demands creative action. In the epistemology of pragmatism, knowledge is not the reproduction of reality but an instrument for dealing with it successfully. The semantics of pragmatism locates the meaning of concepts in the practical consequences for action resulting from their use or their difference from other concepts. Thus in pragmatism's theory of truth, the truth of sentences can only be determined by way of a process of agreement on the success of action based on them, and not on, say, their correspondence with an uninterpreted reality. The misunderstanding of pragmatism as principally a movement aiming at the destruction of the ideal of true knowledge resulted primarily from the isolation of individual sentences, such as some of William James's statements about truth, from the whole complex of pragmatist thought.

2. Main Variants of Historical Pragmatism

The leading representatives of pragmatism contributed to different fields of inquiry. Peirce was primarily interested in developing a general theory of scientific knowledge and a broadly conceived theory of signs of semiotics. His multifaceted work, which does not deal with questions of social theory, includes important thoughts on the intersubjective use of signs and on the creative production of hypotheses ('abduction'). This theory of signs, particularly in its emphasis on the 'discourse' of scientists in an experimental community, has become very influential.

William James at first worked mainly in psychology, in which he saw the prospect of a way out of the dilemma between a religiously based belief in the free will of the moral agent and the scientific image of the world as a universe governed by causal processes. The solution, for James, lay in the functionality for the survival of the human organism in its environment which he saw in the human capacity to pay deliberate attention to perceptual impressions and to choose between alternative courses of action. A 'functionalist' psychology could proceed by understanding all mental achievements in terms of their function for the organism's active mastery of its environment.

James elaborated his antideterminist stance in an influential article on *The Will to Believe*, in which he defended the legitimacy of religious experience against scientific skepticism. This essay also became the point of departure for his wide-ranging study of the varieties

of religious experience, in which he founded the empirical study of religion not on religious institutions or theological doctrines, but on the phenomenology of religious experience. In his last years, he began to use the phenomenology of experience as a guideline for the elaboration of a pluralist, nondeterminist metaphysics.

John Dewey, whose thought at first developed rather independently of the first pragmatists, abandoned his earlier neo-Hegelian approach during the 1890s and linked up with the logical and psychological aspects of pragmatism as developed by Peirce and James. He constructed a philosophy which extended the core ideas of pragmatism into all the traditional domains of philosophy (metaphysics, logic, ethics, aesthetics), but also and in particular the fields of educational theory, social and political philosophy. He gave more explicit emphasis than the others to a radical version of democracy as the normative core of pragmatism. Taking part in a great number of political controversies, he became one of the most influential American intellectuals ever and *the* representative thinker of the United States.

George Herbert Mead, a friend of Dewey, did most to develop James's strategy of the translation of pragmatist themes into the program of a biologically based empirical social science. His decisive contribution to social theory consists of his theory of the specific features of human communication and hence the attempt to thematize the constitution of personality structures in the dynamics of interpersonal relationships. Mead attacked the assumption of a presocial substantial self and replaced it by a theory of the genesis of the self in which even the interaction of a person with himself or herself is conceptualized as the result of social structures. Mead also pursued this line of thought in the direction of problems of cognitive development, such as the constitution of permanent objects in experience and the constitution of structures of time. His ethical theory and antideterminist philosophy of history remain rather neglected.

Thus it can be said that pragmatism has never been a clearly delimited school of thought with a fixed research program; it shows, however, the characteristics of an intellectual movement or, at least, constellation with the common elements of anti-foundationalism, the fallible nature of truth, the social nature of the self, the importance of inquiry in all fields of human life, and (metaphysical) pluralism (see Bernstein 1992). During the first decades, pragmatism exerted a considerable influence on a variety of disciplines in the United States (functionalist psychology; institutionalist economics; educational reform; law; theology; political theory; the Chicago school of sociology). After the 1930s this influence clearly waned, reaching a nadir in the 1950s and 1960s. (An indicator of this development is probably the fact that the 1968 edition of this encyclopedia did not contain an entry on pragmatism.) One of the few currents

which kept the heritage of pragmatist social thought alive was the school of symbolic interactionism in sociology, which produced a rich amount of empirical work mostly on interpersonal relationships, processes of emergence and definition of deviant and sub-cultural behavior, social movements, and the division of labor in professional organizations. There are other main figures of American sociology at that time in whose work pragmatist orientations can easily be detected (C. Wright Mills, Philip Selznick).

3. Contemporary Pragmatist Thinking

Since the late 1970s a renaissance of pragmatist philosophy has been underway. Apart from thinkers like Richard Bernstein and John Smith, who had continuously pursued a pragmatist agenda, two thinkers have probably contributed most to this renaissance. One is Richard Rorty who had for a long time tried to articulate pragmatist motifs within the tradition of analytical philosophy before breaking (more or less) with it and proposing a variant of postmodern thinking he calls pragmatism. Though Rorty is a perceptive social critic, he has no social theory and actively repudiates the need for it. His philosophy has sparked a vast critical literature and reappropriation of the historical pragmatists; his main rival in these debates is Hilary Putnam in whose work one can find a thorough and significant restatement of pragmatist philosophy.

The other main originator of this renaissance is Jürgen Habermas. Together with Karl-Otto Apel, these two German thinkers gave a new twist to the winded history of the European reception of pragmatism. From the beginnings of the twentieth century on, very diverse schools of European thinking had used elements from American pragmatism for their own purposes (e.g., French Catholic Modernism; Italian proto-Fascism; German Philosophical Anthropology; and others). Other European thinkers came close to pragmatism in some aspects of their work (e.g., Max Scheler, Arnold Gehlen, Martin Heidegger, Ludwig Wittgenstein) without spelling out their relationship with it. Peirce's ideas about the 'discourse' of scientists has been of decisive importance for the 'discourse ethics' developed by Apel and Habermas. Mead's understanding of symbolically mediated interaction has been crucial for the 'Theory of Communicative Action' by Habermas; he declares Mead the main inspiration for 'the paradigm shift from purposive to communicative action' which he proposes in sociological theory.

Habermas' faithfulness to the pragmatist heritage has been disputed almost as much as Rorty's. There are obviously other possibilities for pragmatist social thought than his combination of a theory of communication with systems theory or than his restriction of the tasks of moral theory to the program of discourse ethics. One can claim (Joas 1996, 2000) that the

most fundamental contribution of pragmatist social thought lies not in its emphasis on communication, but in its specific understanding of the creativity of human action. This understanding distinguishes pragmatist social thought from approaches focusing on either rational action or normatively oriented action, or from a systems theory or structuralism which disconnect the dynamics of human action from the development of social wholes. On the basis of pragmatism understood as a theory of the 'situated creativity' of human action, it becomes possible to reconceptualize rationality and normativity and to develop a theory of macrosocial processes and social change prepared by the political theory and philosophy of history in the works of the classical pragmatists.

See also: Aristotelian Social Thought; Communication: Philosophical Aspects; Critical Theory: Frankfurt School; Dewey, John (1859–1952); James, William (1842–1910); Marxist Social Thought, History of; Mead, George Herbert (1863–1931); Postmodernism: Philosophical Aspects; Sociology, History of; Utilitarian Social Thought, History of

Bibliography

- Bernstein R J 1971 *Praxis and Action*. Duckworth, Philadelphia, PA
- Bernstein R J 1992 The resurgence of pragmatism. *Social Research* 59: 813–40
- Blumer H 1969 *Symbolic Interactionism: Perspective and Method*. Prentice-Hall, Englewood Cliffs, NJ
- Dewey J 1967–87 *Early/Middle/Late Works*. Southern Illinois University Press, Carbondale, IL
- Dickstein M (ed.) 1998 *The Revival of Pragmatism. New Essays on Social Thought, Law, and Culture*. Duke University Press, Durham, NC
- Habermas J 1984/1987 *The Theory of Communicative Action*. Beacon Press, Boston, MA, 2 Vols. [English translation by McCarthy T]
- Habermas J 1992 *Postmetaphysical Thinking*. MIT Press, Cambridge, MA
- James W 1987/1992 *Writings 1878–1899/1902–1910*. Library of America, New York
- Joas H 1993 *Pragmatism and Social Theory*. University of Chicago Press, Chicago, IL
- Joas H 1996 *The Creativity of Action*. University of Chicago Press, Chicago, IL [English translation by Gaines J, Keert P]
- Joas H 2000 *The Genesis of Values*. Polity Press, Cambridge, UK [English translation by Moore G]
- Mead G H 1934 *Mind, Self, and Society from the Standpoint of a Social Behaviorist*. University of Chicago Press, Chicago, IL
- Mead G H 1964 *Selected Writings*. Bobbs-Merrill, Indianapolis
- Peirce C S 1931–60 *Collected Papers*. Harvard University Press, Cambridge, MA
- Rorty R 1979 *Philosophy and the Mirror of Nature*. Princeton University Press, Princeton, NJ
- Selznick Ph 1992 *The Moral Commonwealth: Social Theory and the Promise of Community*. University of California Press, Berkeley, CA

Thayer H S 1981 *Meaning and Action: A Critical History of Pragmatism*. Hackett, Indianapolis

H. Joas

Precautionary Principle

The precautionary principle is an intuitively appealing concept that has risen to prominence over the past decade. The most well-known definition of the principle is contained in the Rio Declaration of the 1992 United Nations Conference on Environment and Development (UNCED). Principle 15 of the Declaration states that 'Where there are threats of serious or irreversible damage, lack of full scientific certainty shall not be used as a reason for postponing cost-effective measures to prevent environmental degradation' (Cameron 1994, p. 269). The principle has been invoked so often in international environmental resolutions that it is considered by many to be a fundamental normative principle (Cameron 1994, Costanza and Cornwall 1992). Unfortunately, there is no consensual definition of the principle and few practical guidelines for its implementation in national and international environmental policy.

The precautionary principle's rapid rise to prominence can be attributed to several factors. First, the notion of precaution is intuitively appealing, perhaps because we have been told from an early age that 'an ounce of prevention is worth a pound of cure.' Second, with the growing realization that human activities are creating global problems, such as global warming, the depletion of the ozone layer and the reduction in biological diversity, such admonishments have become all the more urgent. As the scales of the problems we face have grown, so too have the stakes. The higher the stakes, the less we can afford to take chances with the future. Third, as the size and complexity of the problems have grown, so too have the uncertainties. Thus, traditional science is unable to provide unequivocal answers. Given the stakes involved and absent clear solutions, it seems only prudent to err on the side of safety and take action now rather than later. If we wait for clear scientific proof of harm, it may be too late to do anything.

The term precautionary principle (*Vorsorgeprinzip*) was first used explicitly with regard to environmental policy in the former Federal German Republic in the 1970s (O'Riordan 1994, p. 306, O'Riordan and Jordan 1995, p. 193, Boehmer-Christiansen 1994). Subsequently, the precautionary principle was incorporated into many national and international environmental policies and protocols. The earliest use of the precautionary principle in an international agreement was in the Ministerial Declaration of the Second International Conference on the Protection of the North Sea in London in 1987. This declaration stated

that 'in order to protect the North Sea from possibly damaging effects of the most dangerous substances, a precautionary approach is necessary which may require action to control inputs of some substances before a causal link has been established by absolutely clear scientific evidence' (Cameron 1994, pp. 267–8). In 1989, the Nordic Council's Conference on Pollution of the Seas called for a precautionary approach 'eliminating and preventing pollution emission when there is reason to believe that damage or harmful effects are likely to be caused, even where there is inadequate or inconclusive scientific evidence to prove a link between emissions and effects' (Cameron 1994, p. 268). In 1992, the precautionary principle was boosted to prominence by its incorporation as Principle 15 in the United Nations Conference on Environment and Development (UNCED). Similar language was used in the 1992 United Nations Framework Convention on Climate Change, and the 1992 UNCED Biodiversity Convention (Cameron 1994, p. 269). Many other similar definitions have been proposed. For example, a group of academics and environmental activists called the Wingspread Group have concluded that there is compelling evidence of harm from toxic substances, global warming, ozone depletion and nuclear materials, and that existing environmental regulations are inadequate. The group proposed that 'When an activity raises threats of harm to human health or the environment, precautionary measures should be taken even if some cause and effect relationships are not fully established scientifically' (Risk Policy Report 1998).

The precautionary principle comprises three core ideas. First, it is necessary to take protective actions in advance of clear scientific proof, because scientific proof is unlikely to be forthcoming in the near future, and delay in action may ultimately prove more harmful and more costly to correct after the fact. Advocates of the precautionary principle argue that the uncertainties inherent in problems such as global warming and ozone depletion necessitate taking action prior to clear scientific proof. Because the uncertainties are inherent, traditional scientific methods can never eliminate them, so waiting for clear scientific evidence is foolhardy. Wynne (1992) identifies four kinds of uncertainty—risk, uncertainty, ignorance and indeterminacy. He proposes that a precautionary approach that recognizes these uncertainties and the limitations of traditional scientific methods is necessary. Funtowicz and Ravetz (1992) argue that the nature of environmental problems changes as the decision stakes and levels of uncertainty increase. Applied scientists and professional consultants (such as engineers) are able to provide adequate solutions for environmental problems with moderate to low decision stakes and levels of uncertainty. As we move into the realm of problems such as global warming, the stakes and the uncertainties are so high that a new kind of science is necessary. Post-normal science relies on the use of 'extended facts' and an extended peer community for

review. Thus, advocates of the precautionary principle call for greater use of civic science based on negotiation, quasi-formal predictions and evaluations, and more attention to the process rather than the content of the analysis (O'Riordan and Jordan 1995, Hunt 1994).

Second, traditional methods of economic analysis, such as the use of discount rates and cost benefit analysis, are inadequate because of the nature and likelihood of the harm, the extent of uncertainties, and the delay before the harm becomes evident (Perrings 1995). Traditionally, economists assume that the growth in capital resulting from current investments will more than compensate future generations for possible harm. Unfortunately, this means that little value is placed on impacts 25 years or more in the future, and it remains unclear what the impacts will be on natural capital. Advocates of the precautionary principle argue that the likelihood, nature and magnitude of the future costs of environmental problems, such as global warming or biodiversity loss, are unknown and cannot be calculated using traditional economic methods (O'Riordan and Jordan 1995).

Third, there is a need to shift the burden of proof to the 'developer' to show that the technology or activity is 'safe.' Under many current legal and policy frameworks, the presumption is that a technology or activity is 'innocent until proven guilty.' The precautionary principle would change this to presume that a technology or activity is 'guilty until proven innocent' (O'Riordan and Jordan 1995, O'Riordan 1994).

O'Riordan and Jordan (1995) identify several other elements that characterize precautionary approaches. These include the need to: (a) maintain proportionality between the likely harm and the cost of preventive measures; (b) take preventive measures to safeguard 'ecological space' to prevent pushing natural systems beyond thresholds that result in irreversible change; (c) recognize that natural systems have intrinsic value and there is a moral imperative to protect them; (d) shift emphasis to meso-term planning (i.e., 25–100 years); and (e) pay for the ecological damage caused by those who were less than cautious in the past.

The precautionary principle has been criticized on several fronts. O'Riordan and Jordan (1995, p. 192) fear that the principle runs the risk of following the 'dangerously successful pathway pioneered by sustainability.' The principle has broad intuitive appeal but lacks rigorous consensual definitions, and while it has been adopted widely as a laudable goal for environmental policy there are no clear guidelines for its implementation. Like sustainability, the precautionary principle is neither a well-defined principle nor a stable concept. It has become the repository for a jumble of adventurous beliefs that challenge the status quo of political power, ideology and civil rights. Neither concept has much coherence other than it is captured by the spirit that is challenging the authority of science, the hegemony of cost-benefit analysis, the

powerlessness of victims of environmental abuse, and the unimplemented ethics of intrinsic natural rights and intergenerational equity (O'Riordan and Jordan 1995, p. 191).

Bodansky (1994) argues that the precautionary principle underlies many of the environmental laws enacted in the 1970s in the US, but that its impact has often been contrary to the intent of the framers of these laws. He argues that the precautionary approach of the original Clean Air and Clean Water Acts led to delays in the promulgation of regulations. He also suggests that shifting the burden of proof to the proponent of a technology or activity may do little to enhance safety, since harmful impacts are often discovered only belatedly, as in the case of DDT and CFCs. Furthermore, shifting the burden of proof may actually exacerbate harm by discouraging the introduction of beneficial technologies. Bodansky questions whether it is feasible to establish guidelines that determine what levels of harm and scientific certainty are necessary to trigger the precautionary principle, and whether we should always err on the side of safety even if the social costs are very high. Bodansky (1994, p. 225) concludes 'Simple rules promise simple answers. But, in the end, there may be no practical alternative to the imperfect, difficult, often dissatisfying task of muddling along.'

Perrings (1995) and Costanza and Cornwall (1992) are more optimistic about the practical application of the principle. Perrings (1995, p. 159) believes that in the absence of better (i.e., more certain) information the precautionary principle 'leads to a sequential approach in which the decision maker reserves judgment about the uncertain outcomes of each activity, assuming the worst-case outcome until evidence is provided to the contrary'. Thus, the decision-maker assigns worst-case values to the uncertain outcomes of current activities and attempts to minimize the maximum environmental costs. The pragmatic problem becomes choosing a plausible worst-case for decision-makers.

Costanza and Cornwall (1992) propose combining the precautionary principle with the polluter pays principle to develop an environmental deposit-refund or assurance bonding system. Under this system, an 'assurance bond,' equal to the best estimate of the largest potential future damages, would be levied on the proponent of the technology or activity. Portions of the bond would be returned as and when the proponent could show that the suspected damages had not occurred or they would be used to repair or compensate for any damages resulting. Costanza and Cornwall believe this system would shift the burden of proof and the cost of uncertainty from the public to the proponent or developer. The bond would also serve as an incentive for businesses to conduct thorough, publicly accessible research on the environmental impacts of their activities and to look for ways to minimize pollution. Unfortunately, many questions

about the practical value of such a system remain unanswered: who determines the worst-case scenario and the size of the bond, especially since there is so much uncertainty about the future environmental impacts of so many activities? Who decides if and when the bond should be returned, and how? Is it practicable to think of dealing with myriad environmental problems in this fashion? How would this be implemented at an international level?

In conclusion, the precautionary principle has enormous intuitive appeal and has become an integral part of many international environmental agreements. The future success of the principle, however, rests on the development of clear practical guidelines for its implementation at both the national and international levels.

See also: Environment and Development; Environmental and Resource Management; Environmental Challenges in Organizations; Environmental Economics; Environmental Policy; Environmental Policy: Protection and Regulation; Environmental Risk and Hazards; International and Transboundary Accords, Environmental; Planning Issues and Sustainable Development; Sustainable Development

Bibliography

- Bodansky D 1994 The precautionary principle in US environmental law. In: O'Riordan T, Cameron J (eds.) *Interpreting the Precautionary Principle*. Earthscan, London, pp. 203–28
- Boehmer-Christiansen S 1994 The precautionary principle in Germany—enabling government. In: O'Riordan T, Cameron J (eds.) *Interpreting the Precautionary Principle*. Earthscan, London, pp. 31–60
- Cameron J 1994 The status of the precautionary principle in international law. In: O'Riordan T, Cameron J (eds.) *Interpreting the Precautionary Principle*. Earthscan, London, pp. 262–89
- Costanza R, Cornwall L 1992 The 4P approach to dealing with scientific uncertainty. *Environment* 34: 12–42
- Funtowicz S, Ravetz J 1992 Three types of risk assessment and the emergence of post-normal science. In: Sheldon K, Golding D (eds.) *Social Theories of Risk*. Praeger, Westport, CT, pp. 251–74
- Hunt J 1994 Social construction of precaution. In: O'Riordan T, Cameron J (eds.) *Interpreting the Precautionary Principle*. Earthscan, London, pp. 117–25
- O'Riordan T 1994 The precautionary principle in environmental management. In: Ayres R U, Simonis U E (eds.) *Industrial Metabolism: Restructuring for Sustainable Development*. United Nations University Press, Tokyo, pp. 299–318
- O'Riordan T, Jordan A 1995 The precautionary principle in contemporary environmental politics. *Environmental Values* 14: 191–212
- Perrings C 1995 Reserved rationality and the precautionary principle: technological change, time, and uncertainty in environmental decision making. In: Krishnan R, Harris J M, Goodwin N R (eds.) *A Survey of Ecological Economics*. Island Press, Washington, DC, pp. 158–62
- Risk Policy Report 1998 Wingspread group offers new precautionary principle statement. *Risk Policy Report* 5: 22–3

Wynne B 1992 Uncertainty and environmental learning: reconciling science and policy in the prevention paradigm. *Global Environmental Change* 2: 111–27

D. Golding

Prefectorial Government

Governments must govern their territory and ensure that decisions reached in the national capital are implemented throughout the country. Different political traditions emphasize policy coherence and uniformity across their territory to differing degrees, and therefore invest varying levels of political and administrative resources into controlling subnational governments. For countries more concerned with uniformity (such as France and countries that have adopted the French administrative tradition) prefects are crucial instruments for central governments. France and several other countries utilize these administrative officials as one component of their attempts to control policy implementation through subnational units of government, and also exercise control over the field staffs of other ministries in central government.

Prefectorial government refers to the utilization of central government officials—the prefect or *prefet*—to monitor and control policy within the area of the country for which each is responsible. For example, each department in France has a prefect, along with several *sous préfets* and other administrators responsible for controlling policy within the department. These officials are employees of the Ministry of the Interior and must review actions taken by both field services of most other ministries, as well as many decisions of department and commune officials. The relationships between the prefect and field staffs may be rather complex. For example, the prefect has little control over the pedagogical activities of the Ministry of Education, but does control construction of new schools.

The office of prefect has a long history, in France being derived from royal intendants in the *ancien* regime, and having persisted in one form or another since the Napoleonic era. As the office was developed during the Napoleonic era, it was seen as the means of creating a common nationhood in the face of substantial internal diversity within the country. The office briefly was called Commissioner of the Republic during the early stages of decentralizing reforms during the 1980s but the traditional title soon was restored to reflect the power of the State. Also, the creation of the regional planning system and regional governments with limited powers during the 1960s produced another level of government at which there is a need for prefectorial control in France, with some analogous development in other systems. Also, the

continued decentralization and deconcentration of policymaking in France (and other countries) means there is an additional necessity of integrating and coordinating policies at the implementation stage (Dupuy and Thoenig 1985, Thorval 2000). The system has been copied, with predictable variations, at various times in numerous other countries such as Italy, Greece, Spain, and Japan, as well as continuing to be utilized in several former French colonies. Few of the systems copied from the original have provided these officials with the potent regulatory functions characteristic of French administration.

Formally prefects exercise substantial powers within their department. They can reverse decisions made by other officials within their department and impose their own decisions in place of those decisions they consider inappropriate, or at least refer the issues involved back to Paris for final determination. Given the strong tutelary and control functions exercised by prefects, they rarely if ever are natives of the territories within which they work. In addition, the prefect has tended to have a virtual monopoly over communication between the department and the central government in Paris, so that the field staffs of most ministries have been channeled through the prefecture. This monopoly of communication constitutes a major source of power for prefects, given that the field staffs responsible to other ministries might have few independent avenues for complaining about the decisions of the prefect. In addition to the powers over national ministries, prefects frequently are directly involved with local government and have been in essence the chief executive for local governments. With reforms in the administrative system these local governments have been given increased powers but the prefect still has numerous tutelary powers.

The French prefect, much as the President of the United States, can be conceptualized as fulfilling a number of functions (Bernard 1992):

- (a) The guarantor of the heritage of the nation and of the continuation society;
- (b) The guardian of public order, with ultimate responsibility to preserve security and safety;
- (c) The representative of the State within the geographical area, and the guarantor of equality of citizens;
- (d) The promoter of regional economic development and (increasingly) the link of the region with the European Union; and
- (e) The promoter of social consensus and cooperation within the region.

The political climate favoring decentralization and enhanced local autonomy has minimized the central position of prefects in most political systems, but the office remains an active component of administration in political systems that remain concerned about creating uniformity, including France. Indeed, the role of the prefect as a coordinator of local services may be even more crucial with the substantial deconcentration

of policymaking. The range of powers available to prefects is becoming more constrained, and local communities are more likely to assert their own policy and political concerns. Even with the increase in local democracy, prefects remain powerful actors in local government, preserving their political power in part through alliances with local political leaders. With the creation of regional governments concerned primarily with economic development there is now a new level of prefectorial government in France, although the degree of control available to these officials is substantially more limited. Further, the role that the prefect plays in the distribution of funds coming from the European Union has also enhanced the central role of the prefect in subnational policy-making.

All administrative systems that utilize this variety of administrative mechanism for control over local areas are not called prefectorial. For example, the governors appointed by central government in Scandinavian countries have lost much of their authority over local authorities, although there are some vestigial powers of that sort. Even local mayors in some administrative systems (Finland and the Netherlands) may be in effect central government officials responsible for ensuring policy delivery in the face of potentially recalcitrant local political forces.

Authoritarian governments also have employed a variety of administrative mechanisms to control local governments and policy implementation, often using party officials rather than administrators to exercise the same control functions exercised by prefects.

Likewise, all administrative systems that are called prefectorial do not function in the same ways. Fried (1963), for example, distinguished between integrated and unintegrated prefectorial systems, with the former being exemplified by the French model and the latter by the Italian. The Italian prefectorial system is substantially less restrictive than the French version of exercising administrative control through prefects. For example, Italian field staffs from line ministries have had their own lines of contact with the government in Rome without being dependent upon the prefect. Also, Italian prefects have not exerted direct control over local governments in as overt a manner as they might in France, nor have they supplied accounting and other managerial services to the local governments. The prefect in Italian (and most other) administrative systems is more of a monitor than an effective regulator of subnational governments.

There are in many ways two prefects existing simultaneously in the office, and there is a great deal of ambiguity in the role as it has evolved (de Montricher 2000). On the one hand, much of what has already been described reflects the legal version of the prefectorial role. This official is the formally powerful representative of the authority of the State, resplendent in their gold-braided uniform, playing a central part in local ceremonies to demonstrate the involvement of central government in those activities. On the

other hand, the prefect is a bureaucratic politician who must build coalitions with local officials in order to be successful in governing. Although prefectural systems originally were designed to exercise central control over subnational policies and administration, in practice local governments also have a great deal of political influence over the prefect, and may convince that official to promote their own local interests, rather than simply being forced to accept what they are being instructed from the center.

The prefect generally does not have to be persuaded very much to help the local community in pressing its demands on central government. The prefect is a career administrative official, generally coming from the very summit of the career structure. These officials naturally want to build a successful career, and prefects are also subject to removal from office more readily than are other civil servants. One measure of success for the prefect would be the ability to administer central policies without complaints from the localities, complaints that might be expressed through local political notables. Further, if prefects can construct a workable coalition with political elites in the localities over which they have responsibility they may be able to build an even more effective administration and thereby improve their career chances even more. Various examinations of the politics of these administrative relationships undertaken in France demonstrate that influence and control circulate in both directions and that most final policy decisions are the products of negotiation and bargaining rather than imposition of authority through the prefect (Crozier and Thoenig 1976, Worms 1966). Further, prefects engage in a number of consultation and consensus-building operations within their departments to both enhance the quality of the decisions made and improve the involvement of local citizens in their own government. In France the power of local politics is enhanced by the *cumul des mandats* through which the same politician will hold office at both the central and local levels.

Prefectoral control may appear almost quaint in the contemporary era that stresses local democracy and accepts greater diversity among policy outcomes. Further, prefectural administration may appear anti-democratic, given that this instrument of governing substitutes administrative controls for the judgment of locally elected public officials (Behar and Estebe 1996). The overwhelming concern with uniformity and control also may limit the capacity of local and regional governments to innovate, while there is growing evidence of the efficacy of policy innovation at lower levels of government. In addition, uniformity in policy and administration may not conform well to managerialist ideas devolving authority to lower echelons of organizations.

Prefectoral policymaking systems are concerned strongly with centralized control and can be contrasted with federal systems that tend to involve states or

provinces for implementation without the direct monitoring and control, and that are willing to accept wider disparities in policy. The prefectural systems may also be contrasted with other types of unitary government that rely on the functional field staffs of central ministries to provide sufficient uniformity throughout their territory. Those functional administrative systems do not consider it necessary to impose an extra level of oversight, at least not to impose *ex ante* controls. Thus, the choice of this form of governing reflects some fundamental choices about the way in which laws are to be implemented and the extent to which uniformity is a requirement for justice.

See also: Administration in Organizations; Financial Accounting and Auditing; Governments; Public Administration, Politics of; Public Bureaucracies

Bibliography

- Behar D, Estebe P 1996 *Villes, sciences sociales, professions*. L'Harmattan, Paris
- Bernard P 1992 *Le préfet de la République: Le chène et l'olivier*. Economica, Paris
- Crozier M, Thoenig J-C 1976 La regulation des systems organises complexes: Le cas du systeme de decision politico-administratif en France. *Revue française de sociologie* 16
- Dupuy F, Thoenig J-C 1985 *L'administration en miettes*. Fayard, Paris
- Fried R 1963 *The Italian Prefects*. Yale University Press, New Haven, CT
- de Montricher N 2000 The prefect and the reform of the state. *Public Administration*
- Thorval J 2000 L'avenir de la representation de l'Etat. *Pouvoirs Locaux* 44(1): 84-9
- Worms J-P 1966 Le prefet et ses notables. *Sociologie du travail* 3

B. G. Peters

Preference Models with Latent Variables

1. Introduction

Much of human behavior may be viewed as successive choices among more or less well-defined options or alternatives. In some situations, choices are made over a continuum of possibilities; for example how much of a particular resource to allocate for a task or how much to work. In other situations, choices are made among a finite (discrete) number of options. Familiar instances are a citizen deciding whether to vote and if so for whom, a shopper contemplating various brands of a product category, or a couple weighing marriage. Frequently, these decisions are made in a sequential manner. Examples include choice of employment, domicile, transportation mode and investment. With an emphasis on discrete types of choice problems,

research on preference and choice has revolved about the development of mathematical models that predict how both observed and unobserved attributes of decision makers and of choice options determine decisions. By treating the unobserved attributes as latent (random) variables, these preference models focus mainly on choice outcomes and to a lesser extent on the underlying decision processes. As a result, the main purpose of preference models with latent variables is to summarize the data at hand and to facilitate the forecasting of choices made by decision makers facing possibly new or different variants of the choice options.

The prevalence of behavioral models with their objective of choice prediction originated with the work of Thurstone (1927). By invoking the psychophysical concept of a sensory continuum, Thurstone explained the observation that choices by the same person may not be deterministic under seemingly identical conditions. He argued that options should be represented along this sensory continuum by random variables that describe the options' effects on a person's sensory apparatus. The choice process is then reduced to selecting the option with the greatest realization. One may question the use of randomness as a device to represent factors that determine the formation of preferences but are unknown to the observer. However, because, in general, it is not feasible to identify or measure all relevant choice determinants (such as all attributes of the choice options of the person choosing, or of environmental factors) the question of whether the choice process is inherently random or is determined by a multitude of different factors may be of little relevance in at least one aspect: either position arrives at the same conclusion that choices are best described in terms of their probabilities of occurrence.

Thurstone's early conceptualizations gave rise to an extensive body of methodological and empirical research, mainly in economics and psychology. Surveys about developments in choice theories (e.g. see *Decision and Choice: Random Utility Models of Choice and Response Time*; Luce's Choice Axiom; *Ranking Models, Mathematics of*, Marley 1992, Suppes et al. 1992) and choice modeling (Amemiya 1985, Fligner and Verducci 1993, Manski and McFadden 1981, McFadden 1984) indicate that the merging of the literatures in these two disciplines produced a flexible and powerful framework for the structural investigation of choice data. In particular, this framework has proved to be useful for analyses of individual preference differences at a particular point in time and over time, multidimensional representations of choice alternatives, and the incorporation of variables that describe choice options and decision makers. Each of these areas of research is outlined in the following sections. Because emphasis is on conceptual issues, more technical topics such as sampling, estimation and inference will not be discussed (Albert and Chib 1993).

2. Theoretical Framework

We define a population $\mathcal{P}$ of decision makers ($n \in \mathcal{P}$) and a finite set $\mathcal{C}$ of choice alternatives ($j \in \mathcal{C}$). For each decision maker n and choice alternative j , a vector $\mathbf{x}_{nj}$ of observed variables is available that partially describe the pair (n, j) . The underlying momentary utility of option j to person n is denoted by u_{nj} and a function of the observed and unobserved attribute values. Distributional assumptions are made to capture the variation of the unobserved attribute values and taste differences among the decision makers. In combination with a decision rule for selecting an option, the utilities determine the options' choice probabilities. As a result, the choice probabilities can be represented by an a priori specified member of a family of functions which is partially determined by the distributional assumptions about the unobserved attribute values and utility variations and indexed by a parameter vector θ .

Choice data may be collected in a variety of formats. For example, the *paired comparison* technique is the method of choice in laboratory settings. In large scale surveys or panel interviewer studies, incomplete and/or partial *rankings* or *preference ratings* are frequently collected because of their time-efficiency. In conjoint measurement studies (e.g., see *Measurement Theory: Conjoint*), the constant-sum technique is often used. In recurrent choice studies, we observe how often each choice option is selected within a certain time period. Each data type requires a different treatment to both investigate how decision makers differ in their preferences for choice options and to accommodate the distinct psychological processes that are specific to each choice task.

Consider, for example, a ranking experiment with the choice set $\mathcal{C} = \{i, j, k\}$ but without any attribute information. The first three columns of Table 1 represent the six possible ranking outcomes. Each number corresponds to one of the ranking positions of an option. For instance, the ranking outcome $\{2, 1, 3\}$ indicates that options i, j and k are ranked second, first and third, respectively. Here the rankings are obtained based on the ordering (from largest to smallest) of the underlying options' utilities. The remaining columns of Table 1 depict these orderings of the utility values

Table 1
Outcomes of a ranking experiment

Ranks			Ordering of utilities
i	j	k	
1	2	3	$u_i \geq u_j \geq u_k$
1	3	2	$u_i \geq u_k \geq u_j$
2	1	3	$u_j \geq u_i \geq u_k$
3	1	2	$u_j \geq u_k \geq u_i$
2	3	1	$u_k \geq u_i \geq u_j$
3	2	1	$u_k \geq u_j \geq u_i$

(without the subscript n) for each of the six ranking outcomes. For instance, $u_j \geq u_i \geq u_k$ indicates that the utility of j is larger than the utility of i which, in turn, is larger than the utility of k . The probability of observing this ranking outcome can be written as $\Pr(j \geq i \geq k | \theta) = \Pr((u_j - u_i) \geq 0 \cap (u_i - u_k) \geq 0)$. If the individual utilities follow a multivariate normal distribution with mean vector $\mu = (\mu_i, \mu_j, \mu_k)'$ and covariance matrix

$$\Sigma = \begin{bmatrix} \sigma_i^2 & \sigma_{ij} & \sigma_{ik} \\ \sigma_{ij} & \sigma_j^2 & \sigma_{jk} \\ \sigma_{ik} & \sigma_{jk} & \sigma_k^2 \end{bmatrix}, \quad (1)$$

then probability of observing this ranking can be computed by evaluating the cumulative distribution function of the bivariate normal density with upper integration limits $(\mu_j - \mu_i)/(\sigma_j^2 - 2\sigma_{ij} + \sigma_i^2)$ and $(\mu_i - \mu_k)/(\sigma_i^2 - 2\sigma_{ik} + \sigma_k^2)$ and correlation coefficient $(\sigma_{ij} - \sigma_{jk} - \sigma_i^2 + \sigma_{ik})/[(\sigma_j^2 - 2\sigma_{ij} + \sigma_i^2)^{1/2}(\sigma_i^2 - 2\sigma_{ik} + \sigma_k^2)^{1/2}]^{-1}$. It is important to note that because the ranking probabilities are based on the differences between the options' utilities, it is not possible to uniquely identify the underlying mean and covariance structure.

When relating attribute information to individual utilities, it is useful to distinguish between systematic and random sources of influence. Random sources of influence include (a) attributes of the choice options that are unobserved but influence preferences, (b) measurement errors, and (c) unobserved variables that account for taste variations among the decision makers (Manski & McFadden 1981). Systematic components of influence include the effects of the measured person and option characteristics x_{nj} . The implication of this view is that u_{nj} can be written as $u_{nj} = \mu_{nj} + \varepsilon_{nj}$, where μ_{nj} and ε_{nj} represent the effects of observed and unobserved determinants, respectively, of an option's utility. Without prior information, we can only determine the distribution of the unobserved values across the population of decision makers.

2.1 Modeling μ

The specific functional relationship between μ_{nj} and known option and person characteristics may depend on the particular application. For example, in situations where the measured attributes of the choice options are of a 'more-is-better' type, a linear function may prove sufficient to describe the attributes' effects. This is the case for the *vector model* (Carroll and DeSoete 1991) which assumes that a decision maker combines the attribute values in a linear fashion. In contrast, when the systematic utility component is a *single-peaked* function of the measured variables, an *ideal-point* representation is more appropriate (Coombs 1964).

The attribute effects may vary from person to person in a systematic fashion, for example, by being related

to such person-specific variables as age, gender, and educational achievement. However, in many cases reference to demographic variables is not sufficient to explain heterogeneity in preferences among individuals. Because psychological or economic theories provide little guidance about the role person-specific variables play in the formation of preferences, a random regression approach is required. Thus, it may be postulated that the attribute effects follow some a priori specified distribution. In particular, latent structure models are useful in capturing individual differences in the attribute effects by introducing the fairly unrestrictive assumption that individual variations can be described by a small number of 'prototypical' effect vectors (Böckenholt and Böckenholt 1991). Individual differences are accounted for by assigning each person to a preference class such that members of a preference class exhibit the same or similar attribute effects. For example, when preference data are collected by the method of first choices, a latent-class choice model is given by the following equation

$$\Pr_n(i|\theta) = \sum_{s=1}^S \pi_n(s) \Pr_n(i|\theta_s) \quad (2)$$

where

- $\Pr_n(i|\theta)$ = probability of person n choosing alternative i , and θ is a parameter vector;
- $\pi_n(s)$ = probability of person n belonging to class s ;
- $\Pr_n(i|\theta_s)$ = probability of person n choosing alternative i given n belonging to class s .

The latent classes may represent (a) different *decision rules* adopted by the individuals; (b) variable choice sets considered by the individuals; or (c) subgroups of the population with varying tastes or, sensitivity to the options and their attributes.

2.2 Modeling ε

Attributes that influence the utility of an alternative but are not measured are represented by ε_j . The distribution of ε_j is frequently specified to be multivariate normal with arbitrary covariance matrix. This specification is useful in analyzing *similarity* relationships between choice options that are introduced by the unmeasured attributes. Because the estimation of the elements of the covariance matrix is computationally complex, more restrictive representations are also of interest. For example, Stern (1990) proposed a family of choice models based on independent gamma random variables which includes Luce's (1959) choice model as a special case.

Careful checks are necessary to determine whether the independence assumption is valid or whether unobserved attributes may contain information useful in identifying systematic sources of individual differences and option characteristics. If respondents per-

ceive a common set of latent attributes but differ in the weights they assign to the attribute values, the covariances of ε_j can be decomposed according to a linear *factor-analytic model* (Takane 1987). Other possibilities include the application of ideal-point models which are based on the assumption that individuals choose the option that is closest to their ‘ideal option’ where closeness is a function of the distance between the choice option and the person-specific ideal option (Coombs 1964). Distance may be defined in various ways, for example, by an Euclidean distance in a continuous latent attribute space or by a latent tree structure in which both choice and ‘ideal’ options are represented by nodes. However, in all of these cases, the interpretation of taste differences is much simplified provided the assumption is correct that decision makers use the same set of attributes in assessing the choice options.

2.3 Modeling Preferences Over Time

When collecting time-dependent choice data, it is important to take into account that individuals not only differ in their preferences at a particular point in time but also in the way they change their preferences over time. For instance, current preferences may depend on such diverse factors as previous choices, *interchoice times*, or the length of time a particular preference state has been occupied. Decision makers may also differ in their familiarity with features of the choice options with the result that some decision makers are stable in their preferences over time while others may change their preferences according to some learning process. A complicating factor in this context is that choice environments outside a laboratory setting are rarely stable. Typically, attributes of choice options (e.g., price) and decision environments (e.g., changing tax code) vary over time and influence choice behavior.

The development of time-dependent choice models requires significant modifications of the statistical framework for cross-sectional choice data. However, because choice processes are inherently dynamic and are studied best by taking their temporal nature into account, approaches that focus on these aspects are more likely to lead to important insights about decision making. Conceptually, it is useful to distinguish between *heterogeneity* and *state dependence* in the analysis of recurrent choice data. According to the ‘heterogeneity approach,’ individual choices are assumed to be conditionally independent given the person-specific parameters. For example, *latent structure choice models* represent time-dependent individual differences by switches among latent preference states. If preferences are stable, membership in a latent preference state is time-homogeneous. In contrast, systematic choice variability over time may indicate changes in the perception of (some of) the choice options. Stability and shifts in preference states (for,

say, three equidistant time points) can be investigated by extending Eq. (2) to

$$\Pr_n(i_{t-1}, j_t, k_{t+1} | \theta) = \sum_{a=1}^S \sum_{b=1}^S \sum_{c=1}^S \pi_n(abc) \Pr_n(i_{t-1} | \theta_a) \Pr_n(j_t | \theta_b) \Pr_n(k_{t+1} | \theta_c), \quad (3)$$

where

- $\Pr_n(i_{t-1}, j_t, k_{t+1} | \theta)$ = probability of person n choosing alternative i at time point $(t-1)$, j at time point t , and k at time point $(t+1)$; and

- $\pi_n(abc)$ = probability of person n belonging to classes a , b , and c at time points $(t-1)$, t , and $t+1$, respectively.

Special cases of (3) are obtained by imposing constraints on $\pi_n(abc)$ to describe the switching process among the latent preference states (Böckenholt and Langeheine 1996).

Although time-varying taste heterogeneity is perhaps the single most important source of dependency in recurrent choice data, the assumption of conditionally independent decisions may require the estimation of a large number of preference states in order to be satisfied. A more parsimonious approach for modeling state dependency is obtained by considering the outcomes of previous choices as possible explanatory variables for current choices. For example, if only the choice outcome at $(t-1)$ influences the current choice, the state-specific choice probabilities are specified as $\Pr_n(j_t | i_{t-1}, \theta_s)$. This approach has been shown to be fruitful in capturing such psychological phenomena as habit persistence, learning, or variety seeking (Manski and McFadden 1981). Further generalizations are also possible by allowing for several sets of preference states, where each set follows a different switching process over time.

3. Conclusion

The formulation of choice models in terms of latent utilities has a long tradition in biometric, econometric, psychometric and sociometric disciplines. The numerous applications and further developments in each of these areas have led to many improvements in the statistical apparatus necessary for estimating and validating choice models. However, the unifying principle remains valid that observed individual choice behavior is an imperfect reflection of the underlying preferences of a person. This viewpoint leads to a probabilistic representation of the relation between the manifest and latent levels of choice analysis taking into account both the process by which an individual arrives at a choice and the factors that influence this process. Some of these factors are unobservable (such as the processed information about the choice alternatives), while other factors are known and measurable. By separating these two sources of influence due to measured and unobserved attributes at the individual level, a powerful modeling approach is obtained that

facilitates the identification and isolation of the net effects of choice option characteristics, individual characteristics, and interactions of option and individual characteristics as separate choice determinants.

Modifications of this choice modeling approach are necessary when the choice set is large or when a simultaneous presentation of the choice options is not feasible. In these cases, one may argue that decision makers tend to rely on *satisfying* rules and the idea of an *aspiration level* may be more appropriate than the principle of utility maximization (Simon 1957). In contrast to the large number of possible satisfying rules, however, only a few mathematical representations have been proposed. Perhaps the most popular model is due to Coombs' (1964) analysis of the 'pick any/m' task and Thurstone's (1959) work on the method of successive categories. In the 'pick any/m' task it is assumed that choice options are selected when their utilities exceed some specified minimum value. Thus, the underlying choice process is scalar valued and based on a binary comparison between the option's utility and a threshold value. More specifically, according to Coombs' (1964) unidimensional unfolding approach, the utility of item j for person n is given by the distance between two parameters representing j and n on a unidimensional continuum. This idea has been incorporated in several psychometric models by making the simplifying assumption that the perception of the item-specific parameters does not vary from person to person (Andrich 1997). In general, results of the *maximization* and 'aspiration level' hypotheses are not directly comparable because in the latter case we model the probability of belonging to some category (i.e., 'above the threshold') and in the former case we model order relations among the choice options.

In recent years, stochastic choice models have come under criticism because, frequently, they cannot capture context-sensitive behavior (Tversky and Kahneman 1991). For example, the models' underlying assumption that the assessment of an option's utility does not depend on comparisons drawn between it and other available alternatives has received little support in laboratory studies. Moreover, the deliberation process from the onset of the choice task to the final selection has been shown to play an important role that should be taken into account by the next generation of choice models (Busemeyer and Townsend 1993). Clearly, the latent utility approach has proved to be most useful in providing parsimonious descriptions of how judges differ in their preferences at a particular point in time or over time. However, by emphasizing interindividual differences, this approach can render only limited insights about cognitive processes influencing a choice.

See also: Bounded and Costly Rationality; Decision Research: Behavioral; Factor Analysis and Latent Structure: Overview; Unfolding and Vector Models

Bibliography

- Albert J, Chib S 1993 Bayesian analysis of binary and polychotomous data. *Journal of the American Statistical Association* **88**: 669–79
- Amemiya T 1985 *Advanced Econometrics*. Harvard University Press, Cambridge, MA
- Andrich D 1997 A hyperbolic cosine IRT model for unfolding direct responses of persons to items. In: van der Linden W J, Hambleton R K (eds.) *Handbook of Modern Item Response Theory*. Springer, New York, pp. 319–414
- Böckenholt U, Böckenholt I 1991 Constrained latent class analysis: Simultaneous classification and scaling of discrete choice data. *Psychometrika* **56**: 699–716
- Böckenholt U, Langeheine R 1996 Latent change in recurrent choice data. *Psychometrika* **61**: 285–302
- Busemeyer J R, Townsend J T 1992 Decision field theory: A dynamic cognition approach to decision making. *Psychological Review* **100**: 432–59
- Carroll J D, DeSoete G 1991 Toward a new paradigm for the study of multiattribute choice behavior. *American Psychologist* **46**: 342–52
- Coombs C 1964 *A Theory of Data*. Wiley, New York
- Fligner M A, Verducci J S (eds.) 1993 *Probability Models and Statistical Analyses for Ranking Data*. Springer, New York
- Luce R D 1959 *Individual Choice Behavior*. Wiley, New York
- Manski C F, McFadden D (eds.) 1981 *Structural Analysis of Discrete Data with Econometric Applications*. MIT Press, Cambridge, MA
- Marley A A J 1992 A selective review of recent characterizations of stochastic choice models using distribution and functional equation techniques. *Mathematical Social Sciences* **23**: 5–30
- McFadden D 1984 Qualitative choice models. In: Griliches Z, Intriligator M D (eds.) *Handbook of Econometrics*. MIT Press: Cambridge, Vol. II, pp. 1395–457
- Simon H A 1957 *Models of Man: Social and Rational, Mathematical essays on rational human behavior in a social setting*. Wiley, New York
- Stern H 1990 A continuum of paired comparison models. *Biometrika* **77**: 265–73
- Suppes P, Krantz D, Luce D, Tversky A 1992 *Foundation of Measurement. Geometric, Threshold, and Probabilistic Representations*. Academic Press, San Diego, Vol. II
- Takane Y 1987 Analysis of covariance structures and probabilistic binary choice data. *Communication and Cognition* **20**: 45–62
- Thurstone L L 1927 A law of comparative judgment. *Psychological Review* **34**: 273–86
- Thurstone L L 1959 *The Measurement of Values*. University of Chicago Press, Chicago
- Tversky A, Kahneman D 1991 Loss aversion in riskless choice: A reference-dependent model. *Quarterly Journal of Economics* **107**: 1039–61

U. Böckenholt

Prefrontal Cortex

The prefrontal cortex is the cortex of association of the frontal lobes. It plays a major role in the temporal organization of behavior, reasoning and language.

The underlying mechanisms are not fully understood, but are known to support at least two cognitive functions of temporal integration that are essential to that role: active short-term memory and preparatory set.

1. Anatomy

In evolution and in individual growth, the prefrontal cortex develops more slowly, yet further, than most other brain structures, in both size and weight (Fig. 1). Phylogenetically and ontogenetically, it is one of the last regions of the cortex (neocortex) to develop (Brodmann 1909). The human prefrontal cortex does

not reach full structural maturation of cells and connections until young adulthood. In the adult human, the prefrontal cortex constitutes almost one-third of the totality of the neocortex. For reasons that are not well understood, this region of the cortex, which is one of the latest to develop, is also one of the most vulnerable to involution and pathological aging (Raz 2000).

The primate prefrontal cortex (Petrides and Pandya 1994) is composed of three major parts or regions (Fig. 2): the dorsolateral prefrontal cortex, that is, the cortex of the dorsal and lateral convexity of the anterior portion of the frontal lobe (Brodmann cytoarchitectonic area 46, and lateral parts of areas 8, 9,

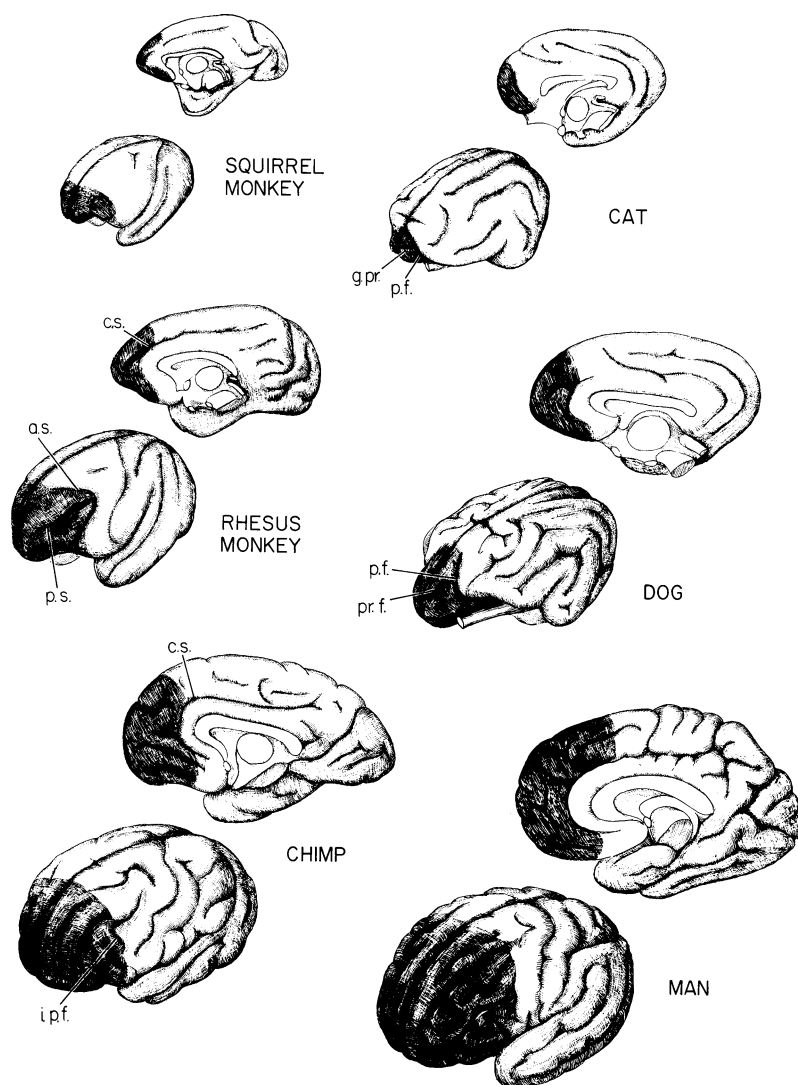


Figure 1
The prefrontal cortex (shaded) in six animal species

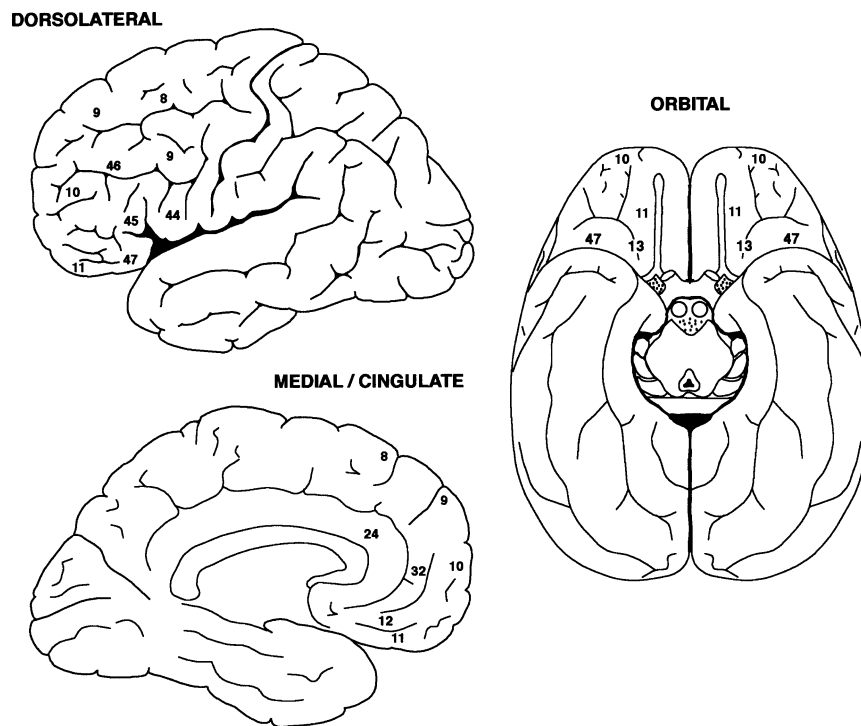


Figure 2

The three major regions of the prefrontal cortex in the human, with cytoarchitectonic areas numerated according to Brodmann's map

10, and 11); the medial and cingulate prefrontal cortex, which is nearly flat and faces the medial surface of the contralateral frontal pole (areas 12, 24 and 32, and medial parts of 8, 9, 10, and 11); and the inferior orbital prefrontal cortex, slightly concave and located directly above the orbit (areas 13, 47, and inferior parts of 10, 11, and 13).

Of the three regions, the dorsolateral develops last and most. This region is bordered posteriorly by the premotor cortex (area 6); the medial and orbital prefrontal cortices (often considered together and designated orbitomedial prefrontal cortex) are anterior and adjacent to the corpus callosum and to limbic structures (piriform cortex and amygdala, cingulate cortex, septum, and hypothalamus). In terms of cellular architecture, the posterior orbitomedial cortex is transitional, with features of both the limbic cortex and neocortex.

2. Functional Connectivity

The functions of the prefrontal cortex can be best understood as components of a general role of frontal cortex in the representation and execution of the actions of the organism. In its totality, the frontal cortex can be considered motor cortex in the broadest

sense. It represents and coordinates actions in all domains of adaptation of the organism to its internal and external environment. In order to coordinate such a wide gamut of activity, the prefrontal cortex receives inputs from, and sends outputs to, many other parts of the brain.

Each of the three principal prefrontal regions has its own set of fiber connections mostly reciprocal and topologically organized—not only with the other two prefrontal regions, but with other cerebral structures, cortical and subcortical. Especially noteworthy for their number and functional implications are the connections of the three prefrontal regions with the thalamus, limbic structures, the basal ganglia, and other parts of the cortex. The orbital and medial prefrontal regions are well connected with the medial and anterior nuclei of the thalamus, the piriform cortex, the hippocampus, the amygdala, and the hypothalamus.

This connectivity with the thalamus and limbic structures falls into three major categories: (a) gustatory and olfactory inputs for the processing of information on taste and olfaction; (b) inputs from the internal milieu of the organism, notably from the hypothalamus, related to motivation, visceral functions, and affective states; and (c) efferent outputs to thalamic and limbic structures for the control of

instinctual, emotional, and social behavior, as well as certain visceral functions.

The dorsolateral prefrontal cortex, on the other hand, is extensively interconnected with other areas of the frontal lobe, homolaterally and contralaterally (through corpus callosum), with the hippocampus, and with several areas of the neocortex of the temporal and parietal lobes, in the posterior aspects of the cerebral hemisphere (Pandya and Yeterian 1985). All this cortical connectivity is essential for cognition. It serves the integrative and associative functions that support the cardinal role of the prefrontal cortex in executive memory and in the temporal organization of behavior. In addition, the dorsolateral prefrontal cortex sends important efferent fibers to the basal ganglia, including caudate nucleus, and the cerebellum. This connectivity serves the motor output functions of the prefrontal cortex through structures that are involved in skeletal and ocular motilities.

3. Physiology

Drive, emotional behavior and visceral actions, as well as their affective connotations, are the physiological purview of medial and orbital prefrontal areas. The medial prefrontal cortex plays an important role in basic motivation, though this role is still poorly understood.

In the human, large lesions of this medial cortex lead to apathy, inattention, and lack of spontaneity in speech, behavior, and ideation (Cummings 1985). The orbital prefrontal cortex, on the other hand, seems essential for the inhibitory control of internal drives. Animals or humans with lesions of this cortex often manifest disinhibition of eating, sex, and aggression, in addition to related abnormalities of social and emotional behavior. The human subject with orbitofrontal damage commonly shows uninhibited instinctual behavior, impulsivity, reckless risk taking, coarse and tasteless humor, and poor moral judgement (Stuss and Benson 1986, Damasio et al. 1994). Oftentimes he or she runs into conflict with the law. Further, because internal impulse control is necessary for steady attention and normal goal directed behavior, the orbitomedial prefrontal cortex serves the general role of the prefrontal cortex in the organization of behavior and cognition.

The dorsolateral prefrontal cortex is involved in both the representation and the enactment of temporally organized behavior and of cognitive operations, which are required in language and logical reasoning (Luria 1966, Fuster 1997). These representations fall under the general category of motor or executive memory, which includes the representations of complex schemes of temporally organized action, such as plans, programs, or scripts of behavior and cognition. The qualifications of novelty and complexity should be emphasized because, after its acquisi-

tion and consolidation, routine behavior is no longer represented or enacted by the prefrontal cortex, but by neural structures at lower stages of the motor hierarchy (e.g., motor and premotor cortices, basal ganglia).

A subject of much debate is the functional specialization of areas within the dorsolateral prefrontal cortex. Given that this cortex can be anatomically subdivided into a number of cytoarchitectonically distinct areas (Petrides and Pandya 1994), each with a separate set of connections (Pandya and Yeterian 1985), the functional specialization of its areas can be, in principle, reasonably assumed. Accordingly, some single-cell studies with monkeys support a degree of seemingly modular commitment of various dorsolateral areas to the representation of different categories of sensory information, such as spatial location or visual configuration (Goldman-Rakic 1995). Other single-unit studies, however, point to the essentially associative character of all dorsolateral prefrontal areas.

In the pursuit of behavioral goals, dorsolateral prefrontal cells associate sensory information across modalities (e.g., sight and sound) and across time (Fuster et al. 2000). Furthermore, some cells associate sensory stimuli with subsequent motor actions (Rainer et al. 1998). These associative properties of dorsolateral prefrontal cells point to their essential temporal integrative functions. Thus, the apparent sensory or motor specificity of certain dorsolateral areas may be related to the specificity of their predominant inputs from sensory systems or outputs to motor systems. In any case, that specificity of areas—thus far fragmentary, and far from established—seems secondary to their basic integrative functions at the cognitive level. Those functions (see later) are crucial for the temporal organization of behavior.

It is likely, though not proven, that the connectivity of the prefrontal cortex with the hippocampus is essential for the formation of executive memory, including the schemas of behavioral and cognitive action. Once they have been formed, the executive memory networks of the cortex of the frontal lobe appear hierarchically organized. Represented in primary motor cortex, the lowest stage of that organization, are the simplest motor memories, elements of motor action defined by specific movement and muscle group (i.e., innate or phyletic motor memory). Immediately above, in the premotor cortex, are representations defined by trajectory and goal of movement. At the summit, in the dorsolateral prefrontal cortex, are representations of complex schemas and sequences of goal-directed action. In a manner still unknown, this cortical region represents the broad schemas of action and, in addition, is critically involved in the enactment of those schemas. A consistent component of the dorsolateral frontal-lobe syndrome in the human is the inability to formulate and to enact plans of action.

4. Temporal Structuring of Action

The main function of the dorsolateral prefrontal cortex is the temporal organization of behavior, language and reasoning. This cortical region appears to secure the orderly and timely execution of all new goal-directed sequences in those three domains of activity. As a precondition for its enactment, however, a new or recently learned sequence of acts is presumably represented in prefrontal networks in the form of a schema or plan of action (executive memory). The execution of that sequence is a highly complex process of temporal integration that probably engages many cortical and subcortical brain regions. The dorsolateral prefrontal cortex seems to orchestrate the functions of many other parts of the brain. It does so with the purpose and end-result of structuring behavior and cognition in time, of forming temporal gestalts (or melodies of action) toward the attainment of biological and cognitive goals.

In as much as intelligence is manifested and tested by performance of logical goal-directed directed tasks and sequences of behavior and language, it is reasonable to assume that the dorsolateral prefrontal cortex is an important component of the neural substrate of intelligence. Indeed, it has been argued on the basis of empirical data from the human, that this cortex is the foundation of general intelligence, as measured by Spearman (Duncan et al. 1996). This psychological metric reflects the proficiency of a subject in performance of a large variety of tests, all of which require the temporal integration of behavioral, verbal or cognitive contents.

Towards the end of the twentieth century, evidence has been accumulating for two temporally integrative functions of dorsolateral prefrontal cortex providing the neural and cognitive support to the process of temporal organization. Those two functions are temporally symmetrical and mutually complementary. The first is a temporally retrospective function of short-term memory; the second, a temporally prospective function of preparatory set. Both together help the organism to bridge time between sensory and motor events, and thus to mediate cross-temporal contingencies of behavior. Both seem to share the same or overlapping networks of dorsolateral prefrontal cortex. Both depend on the coordination of neural activity between this cortex and the association cortices of the posterior regions of the cerebral hemispheres.

4.1 Short-term Memory

Short-term memory is commonly conceptualized as the temporary retention of information before it becomes permanent or long-term memory. Recently, however, cognitive psychology has developed a different concept of short-term memory: the temporary

retention of information, old or new, for the execution of an action in the short term, as is required for the performance of complex behavior or the solution of problems. This kind of short-term memory has been called working memory (Baddeley 1986). There is considerable evidence that both types of short-term memory share the same cortical neurons and networks, and that working memory consists in the sustained activation of a wide cortical network representing information to be used prospectively in the near term. That information may be newly acquired or may be an item of long-term memory reactivated for the short term as the occasion demands. As that memory contributes to form a temporal structure of behavioral or cognitive action, the neurons of dorsolateral prefrontal cortex are part of the activated memory network. There is conclusive evidence that the dorsolateral prefrontal region is critically involved in all forms of active short-term memory toward a goal, or working memory (Fuster 1997).

The evidence comes chiefly from two methodologies, selective cortical lesion (or inactivation) and single-cell recording, in the nonhuman primate performing short-term memory or delay tasks (e.g., delayed response, delayed matching). In these tasks, the animal is presented on successive trials with an item of sensory information—which varies from trial to trial—and is required to retain it in memory for electing and executing the appropriate motor response a few seconds or minutes later. Each trial amounts to a specific selection from a set repertoire of actions, a selection that depends on the sensory cue that has preceded it by seconds or minutes. The relative novelty of the relevant sensory-motor association for that particular trial, and the need to retain a sensory item temporarily in memory, place both the trial and the memory under prefrontal control. Monkeys with lesions of dorsolateral prefrontal cortex perform poorly in delay tasks. They seem unable to keep the working memory of the stimulus for each trial. In a normal animal performing a delay task, neurons of that cortex fire persistently—and often in stimulus-specific manner—during the delay (memory period) of every trial.

These observations denote the importance of cellular dynamics in dorsolateral prefrontal cortex for working memory and the temporal organization of behavior. Recent neuroimaging methods have provided further evidence, at a more global level, of the involvement of the dorsolateral prefrontal cortex in working memory (Smith and Jonides 1999). Unlike microelectrode studies, however, and because of its constraints of temporal and spatial resolution, neuroimaging does not yet reveal the dynamics of neuronal populations of this cortex in working memory. What defines prefrontal memory in the active state, or working memory, is the teleological quality of that memory, which is mobilized for the construction of future action. As noted in Sect. 3, one source of

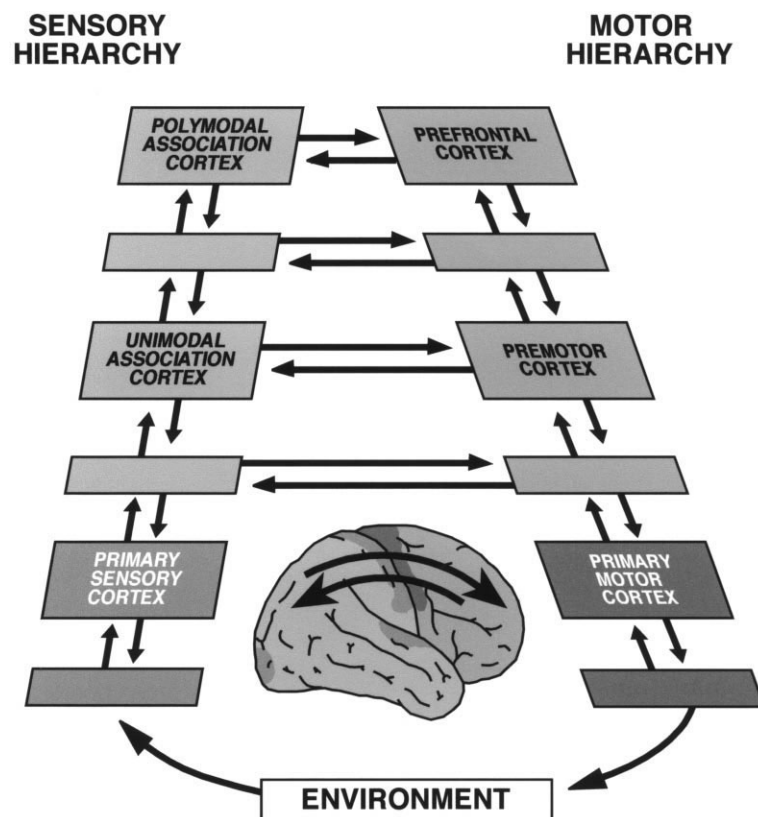


Figure 3

Cortical anatomy of the perception-action cycle depicted around a lateral view of the human brain. Unlabelled regions represent sub-areas of labelled regions or intermediate areas between them. All the arrows represent connections that have been demonstrated in the monkey

apparent areal specialization in dorsolateral prefrontal cortex is the nature of sensory inputs leading to associated action (visual, spatial, etc.). The other is the nature of the associated actions themselves (skeletal movements, ocular movements, etc.). Sensory- or motor-specific cell groupings and areas within that cortex are probably components of teleological associations, and thus the paths to prospective action. In any case, the integrative functions of dorsolateral cortex, namely memory and set, are supramodal and transcend the sensory or motor properties of any given prefrontal area.

4.2 Set

Also from the teleological character of the temporal structuring of action derives the prefrontal function of preparatory set. Setting receptors and effectors for prospective actions is the temporal mirror image of active short-term memory. This second integrative function, preparatory set, is the other side of the coin

of the superordinate prefrontal function of temporal organization. It prepares the organism for anticipated percepts and actions (memory of the future) that will complete the bridging of a cross-temporal contingency. Although the mechanisms of set are not yet known, they probably include the priming of sensory and motor structures via efferent prefrontal connections. In the monkey and the human, the activation of dorsolateral prefrontal cortex in preparatory set is manifested by a progressive increase of cell discharge and slow surface-negative potentials before a stimulus-contingent act. The cellular evidence indicates that memory and set occur largely at the same time. Both seem to be processed simultaneously in the period between two mutually contingent events.

While some prefrontal cells engage in working memory of the cue, others in the same region do it in the preparation of the upcoming behavioral response. The activity of those cells shows that, while the representation of sensory information tends to wane over a delay, the representation of the prepared response tends to increase (Quintana and Fuster 1999).

The choice of one motor act among alternatives is, on the motor side, the equivalent of focusing sensory attention. Thus, preparatory set, the second cognitive function of the dorsolateral prefrontal cortex, can be rightly considered attention, the focusing of attention on a motor act to ensure its prompt and efficient execution.

In summary, the symmetry of the two temporal integrative functions of prefrontal cortex, memory and set, is paralleled by a symmetry of attentive processes, one sensory and the other motor. Active short-term memory is attention focused on the representation of a recent sensory event, whereas preparatory set is attention focused on a consequent and subsequent act. Although attention—especially perceptual attention in the human—is usually associated with consciousness, neither of the two temporal integrative functions of the prefrontal cortex—active memory or set—need be conscious.

5. Perception–Action Cycle

The organization of new and complex behavior commonly requires a continuous succession of temporal integrations of percepts and actions. In this succession, each integration depends to some degree on previous ones. Each act produces changes in the environment that will determine and modify subsequent perceptions, and these will determine and modify subsequent acts. The continuous operation of this cybernetic cycle of interactions of the organism with its environment is a basic principle of biology, the so-called perception–action cycle.

As complex new behavior evolves toward a goal, percepts and acts that are mutually contingent are often separated by time and by other intervening acts and percepts. Further, the pursuit of that goal requires the intervening attainment of lesser or subordinate goals. Thus, the temporal organization of the behavior requires a continuous mediation of contingencies across time. To mediate those cross-temporal contingencies is a function of the biological substrate of the perception–action cycle. The dorsolateral prefrontal cortex is at the apex of that substrate, which basically consists of two parallel hierarchies of neural structures, one sensory and the other motor, extending from the spinal cord to the cerebral cortex. Corresponding levels of each hierarchy are reciprocally interconnected, while each structure in either hierarchy is reciprocally interconnected with the one above and the one below. Figure 3 shows the cortical stages of those hierarchies. The contingencies of automatic and routine behavior (e.g., walking, reflex acts) are mediated at lower, subcortical, stages of the neural organization of the perception action cycle.

Complex new behavior, however, engages the cortical stages of the cycle neural hierarchies. Because this kind of behavior usually contains multiple and complex contingencies across time, the highest neocortical

stages are called into action for its implementation, in particular the dorsolateral prefrontal cortex. This cortex, in functional coordination with the sensory association cortices of the posterior hemisphere, mediates cross-temporal contingencies at the top of the cycle. The prefrontal cortex supports that mediation across time through its two temporal integrative functions of active memory and preparatory set. Both require the functional interaction of prefrontal with posterior cortices, as suggested by combined use of behavioral, cortical-cooling and cell-recording methods (Fuster et al. 1985).

The understanding of the mechanisms of that interaction will lead to more precise knowledge than we now have of the functions of the prefrontal cortex in the temporal organization of behavior. For a number of empirical and theoretical reasons, the role of this cortex in the organization of language and reasoning is probably based on similar functions and mechanisms.

See also: Cognitive Control (Executive Functions): Role of Prefrontal Cortex; Prefrontal Cortex Development and Development of Cognitive Function; Short-term Memory, Cognitive Psychology of; Short-term Memory: Psychological and Neural Aspects; Working Memory, Neural Basis of; Working Memory, Psychology of

Bibliography

- Baddeley A 1986 *Working Memory*. Clarendon Press, Oxford, UK
- Brodmann K 1909 *Vergleichende Lokalisationslehre der Grosshirnrinde in ihren Prinzipien dargestellt auf Grund des Zellenbaues*. Barth, Leipzig, Germany
- Cummings JL 1985 *Clinical Neuropsychiatry*. Grune and Stratton, Orlando, FL
- Damasio H, Grabowski T, Frank R, Galaburda A M, Damasio A R 1994 The return of Phineas Gage: Clues about the brain from the skull of a famous patient. *Science* **264**: 1102–5
- Duncan J, Emslie H, Williams P 1996 Intelligence and the frontal lobe: the organization of goal-directed behavior. *Cognitive Psychology* **30**: 257–303
- Fuster J M 1997 *The Prefrontal Cortex: Anatomy Physiology, and Neuropsychology of the Frontal Lobe*, 3rd edn. Lippincott-Raven, Philadelphia, PA
- Fuster J M, Bauer R H, Jervey J P 1985 Functional interactions between inferotemporal and prefrontal cortex in a cognitive task. *Brain Research* **330**: 299–307
- Fuster J M, Bodner M, Kroger J 2000 Cross-modal and cross-temporal association in neurons of frontal cortex. *Nature* **405**: 347–51
- Goldman-Rakic P S 1995 Architecture of the prefrontal cortex and the central executive. *Proceedings of the National Academy of Science* **769**: 71–83
- Luria A R 1966 *Higher Cortical Functions in Man*. Basic Books, New York
- Pandya D N, Yeterian E H 1985 Architecture and connections of cortical association areas. In: Peters A, Jones E G (eds.) *Cerebral Cortex*. Plenum Press, New York, Vol. 4, pp. 3–61
- Petrides M, Pandya D N 1994 Comparative architectonic

- analysis of the human and the macaque frontal cortex. In: Boller F, Grafman J (eds.) *Handbook of Neuropsychology*. Elsevier, Amsterdam, pp. 17–58
- Quintana J, Fuster JM 1999 From perception to action: Temporal integrative functions of prefrontal and parietal neurons. *Cerebral Cortex* 9: 213–21
- Rainer G, Assad WF, Miller E 1998 Selective representation of relevant information by neurons in the primate prefrontal cortex. *Nature* 393: 577–9
- Raz N 2000 Aging of the brain and its impact on cognitive performance: Integration of structural and functional findings. In: Craik FIM, Salthouse TA (eds.) *Handbook of Aging and Cognition*. Erlbaum, Mahwah, NJ, pp. 1–90
- Smith EE, Jonides J 1999 Storage and executive processes in the frontal lobes. *Science* 283: 1657–61
- Stuss DT, Benson DF 1986 *The Frontal Lobes*. Raven Press, New York

J. M. Fuster

Prefrontal Cortex Development and Development of Cognitive Function

Prefrontal cortex (PFC) is by far the largest cortical area in the human brain (see *Prefrontal Cortex*). It appears to be required when a task is novel or complicated and thus requires concentration. An example would be when you need to guide your actions by information you are holding in mind, and must pay attention so that you act according to that information and not your natural inclination. Its period of maturation is particularly protracted. PFC shows significant developmental changes as early as the first year of life, but does not reach full maturity in humans until young adulthood. In this article, developmental periods are broken up into 0–1, 1–3, 3–7, and 7–22 years. For each period, changes in cognitive abilities thought to depend on PFC are briefly described, as is evidence of maturational changes in PFC.

1. Behavioral Evidence of Improvements in Cognitive Functions that Depend on PFC During the First Year of Life

Piaget's 'A-not-B' task has been widely used to study infant cognitive development (Wellman et al. 1987). Under the name 'delayed response,' the almost identical task has been widely used to study the functions of a subregion of PFC in rhesus monkeys (Goldman-Rakic 1987). The subregion is dorsolateral PFC (DL-PFC). A participant watches as a reward is hidden to the left or right in one of two identical hiding places. A few seconds later the participant is encouraged to find the hidden treat. The participant must hold in mind over those few seconds where the treat was hidden

and, over trials, must update his or her mental record to reflect where the treat was hidden last. The participant is rewarded for reaching correctly by being allowed to retrieve the treat. In this way, the behavior of reaching to that location is reinforced and hence the tendency to make that response is strengthened. When the reward is then hidden at the other location, the participant must inhibit the natural tendency to repeat the rewarded response and instead respond according to the representation held in mind of where the reward was hidden last. Hence, this task requires an aspect of working memory (holding information in mind) plus inhibition of a prepotent action tendency (the tendency to repeat a positively reinforced response).

By roughly 7½ to 8 months of age, infants reach correctly to the first hiding location with delays as long as two or three seconds (Diamond 1985). When the reward is then hidden at the other hiding place, however, infants err by going back to the first location (called the 'A-not-B error'). Infants show marked improvements in their performance of the A-not-B/delayed response task at 7½ to 12 months of age. For example, each month they can withstand delays approximately two seconds longer, so that by 12 months of age delays of roughly 10 seconds or longer are needed to see the A-not-B error (Diamond 1985).

In a transparent barrier detour task called 'object retrieval' (Diamond 1991) a toy is placed within easy reach in a small, clear box, open on one side. There is a very strong pull to reach straight for the toy through the side one is looking, which must be inhibited when an infant is looking through a closed side of the box. At 6 to 8 months of age, infants reach only at the side through which they are looking. They must look through the opening, and continue to do so, to reach in and retrieve the toy. As they get older, the memory of having looked through the opening is enough; infants can look through the opening, sit up, and reach in while looking through a closed side. By 11 or 12 months of age, infants do not need to look along the line of reach at all. Infants progress through a well-demarcated series of five stages in performance of this task from 6 to 12 months of age (Diamond 1991).

Although the A-not-B/delayed response and object retrieval tasks share few surface similarities, human infants improve on these tasks during the same age period (6–12 months) and so do infant rhesus monkeys (1½–4 months; Diamond 1991). Despite wide individual differences in the rate at which infants improve on these tasks, the age at which a given infant reaches 'Phase 1B' on the object retrieval task is remarkably close to the age at which that same infant can first uncover a hidden object in the A-not-B/delayed response paradigm (Diamond 1991).

There is no behavioral task more firmly linked to DL-PFC than the A-not-B/delayed response task (Goldman-Rakic 1987, Fuster 1989). This is one of the strongest brain-behavior relations in all of cognitive

neuroscience. DL-PFC lesions in the monkey also disrupt performance on the object retrieval task (Diamond 1991). MPTP (1-methyl-4-phenyl-1,2,3,6-tetrahydropyridine) injections, which reduce the level of dopamine in PFC, also produce deficits on the task (Taylor et al. 1990, Schneider and Roeltgen 1993). (MPTP also affects the level of dopamine in the striatum, but lesions of the striatum do not impair performance on the object retrieval task (Crofts et al. 1999).)

Human infants of 7 to 9 months, infant monkeys of $1\frac{1}{2}$ to $2\frac{1}{2}$ months, adult monkeys in whom DL-PFC has been removed, infant monkeys of 5 months in whom DL-PFC was removed at 4 months, and adult monkeys who have received MPTP injections to disrupt the prefrontal dopamine system all fail the A-not-B/delayed response and object retrieval tasks under the same conditions and in the same ways (Diamond 1991). In human infants, changes in the pattern of electrical activity detected by electroencephalogram (EEG) over frontal cortex and in the coherence of electrical activity detected by EEG over frontal cortex and parietal cortex (A-not-B: Bell and Fox 1992, object retrieval: N A Fox, personal communication 1992) correlate closely with developmental improvements on the A-not-B/delayed response and object retrieval tasks. This does not prove that developmental improvements on these tasks rely in part on maturational changes in DL-PFC, but it is consistent with that hypothesis.

2. Anatomical and Biochemical Evidence of Maturational Changes in PFC During the First Year of Life

Nerve cells consist of axons, dendrites, and a cell body. Dendrites, which branch off from the cell body, represent the largest part of the receptive surface of a neuron. In humans, the period of marked growth in the length and extent of the dendritic branches of pyramidal neurons in layer III of DL-PFC is $7\frac{1}{2}$ to 12 months (Koenderink et al. 1994), exactly coinciding with the period when human infants are improving on the A-not-B/delayed response and object retrieval tasks. Pyramidal neurons of this PFC region have a relatively short dendritic extent in $7\frac{1}{2}$ -month-old infants. By 12 months of age, they have reached their full mature extension. The surface of the cell bodies of these neurons also increases between $7\frac{1}{2}$ and 12 months (Koenderink et al. 1994). The level of glucose metabolism in DL-PFC increases during this period as well, and approximates adult levels by 1 year of age (Chugani et al. 1987).

Dopamine is an important neurotransmitter in PFC. During the period that infant rhesus monkeys are improving on the A-not-B/delayed response and object retrieval tasks (1.5–4 months), the level of dopamine is increasing in their brain (Brown and

Goldman 1977), the density of dopamine receptors in their PFC is increasing (Lidow and Rakic 1992), and the distribution within their DL-PFC of axons containing the rate-limiting enzyme (tyrosine hydroxylase) for the production of dopamine is changing markedly (Rosenberg and Lewis 1994).

Acetylcholinesterase (AChE) is an enzyme essential for metabolizing another neurotransmitter, acetylcholine. In humans, the pattern of AChE staining in DL-PFC changes dramatically during the first year of life (Kostovic et al. 1988).

3. Behavioral Evidence of Improvements in Cognitive Functions that Depend on PFC From Age 1–3 Years

This is the period for which we know the least about changes in PFC anatomy or in cognitive functions dependent on PFC. Kochanska et al. (2000) report that the ability to inhibit a prepotent response in order to perform a modulated or different response improves markedly from 22 to 33 months of age and that the consistency across their various measures of inhibition also increases between 22 and 33 months.

Using a lazy susan to bring a toy within reach requires relating the lazy susan and its movement to the toy and its movement. It also requires inhibition of trying to reach on a direct line of sight (as the younger children try to do) and inhibition of the tendency to push the lazy susan in the direction you want the toy to go (you must push left to get the toy to approach on the right). Children improve in their ability to do this between 12 and 24 months (Koslowski and Bruner 1972). Case (1992) similarly reports marked improvements on a simple balance beam problem between $1\frac{1}{2}$ and $2\frac{1}{2}$ years.

4. Evidence of Maturational Changes in PFC From Age 1–3 Years

Almost nothing is known about changes in prefrontal cortex during this period. AChE reactivity of layer III pyramidal neurons begins to develop at this time (Kostovic et al. 1988), but that is surely not the only change in prefrontal cortex between 1 and 3 years of age.

5. Behavioral Evidence of Improvements in Cognitive Functions that Depend on PFC From Age 3–7 Years

Clear improvements in tasks that require holding information in mind plus inhibition are seen between 3 and 7 years of age. Three-year-olds make an error reminiscent of infants' A-not-B error, but with a more difficult task. On this task, 3-year-olds sort cards

correctly by the first criterion (whether it is color or shape: Zelazo et al. 1995), just as infants and prefrontally lesioned monkeys are correct at the first hiding place, and just as adults with PFC damage sort cards correctly according to the first criterion (Wisconsin Card Sort test: Milner 1964, Stuss et al. 2000). Three-year-olds err when they must switch to a new sorting criterion, e.g., when cards previously sorted by color must now be sorted by shape, just as infants of 7½ to 9 months and prefrontally lesioned monkeys err when required to switch and search for the reward at the other location, and just as adults with PFC damage err when required to switch to a new sorting criterion. Although 3-year-old children fail to sort by the new sorting criterion, they can correctly state the new criterion (Zelazo et al. 1996), as is sometimes seen with adult patients who have sustained damage to PFC (Luria and Homskaya 1964, Milner 1964). Infants, too, sometimes indicate that they know the correct answer on the A-not-B task, by looking at the correct well, although they reach back incorrectly to the well that was previously correct (Hofstadter and Reznick 1996).

Another example of apparently knowing the correct answer, but not being able to act in accord with it is provided by work with 'go/no-go' tasks (Tikhomirov 1978, Livesey and Morgan 1991). Here, the participant is to respond to one stimulus but do nothing when shown another. Children between 3 and 4 years old repeat the instructions correctly, but they cannot get themselves to act accordingly; they respond even to the 'no-go' stimulus. By 5 years, they perform well on the card sort and go/no-go tasks.

The problem on the card sort task appears to be in (a) integrating two dimensions of a single stimulus (e.g., if children are used to focusing on whether stimuli are red or blue, they have trouble refocusing and concentrating instead on whether the same stimuli are trucks or stars) and (b) inhibiting the tendency to repeat their previously correct response. Similarly, children of 3 years have difficulty with 'appearance-reality' tasks (e.g., Flavell 1986) where, for example, they are presented with a sponge that looks like a rock. Three-year-olds typically report that it looks like a rock and really is a rock, whereas children of 4 or 5 years correctly answer that it looks like a rock but really is a sponge. The problem for the younger children is in relating two conflicting identities of the same object (e.g., Rice et al. 1997) and in inhibiting the response that matches their perception. Manipulations that reduce perceptual salience, by removing the object during questioning, enable children of 3 or 4 years to perform much better (e.g., Heberle et al. 1999).

'Theory of mind' and 'false belief' tasks are other tasks that require holding two things in mind about the same situation (the true state of affairs and the false belief of another person) and inhibiting the impulse to give the veridical answer. Here, as well, manipulations that reduce the perceptual salience

of the true state of affairs aid children of 3 to 4 years (e.g., Zaitchik 1991). Carlson et al. (1998) reasoned that pointing veridically is likely to be a well-practiced and reinforced response in young children, and that children of 3 to 4 years have trouble inhibiting that tendency when they should point to the false location on false belief tasks. Carlson et al. (1998) found that when they gave children a novel response by which to indicate the false location, children of 3 to 4 years performed much better.

Many of the advances of Piaget's 'preoperational' child of 5 to 7 years over a child of 3 to 4 years, who is in the stage of 'concrete operations,' similarly reflect the development of the ability to hold more than one thing in mind and to inhibit the strongest response tendency of the moment. For example, children of 3 or 4 years fail tests of liquid conservation (they do not attend to both height and width, attending only to the most perceptually salient of the two dimensions) and they fail tests of perspective-taking where they must mentally manipulate a scene to indicate what it would look like from another perspective and must inhibit the tendency to give the most salient response (i.e., their current perspective). By 5 or 6 years, they can do these things. Since part of the difficulty posed by Piaget's liquid conservation task is the salience of the visual perception that the tall, thin container contains more liquid, placing an opaque screen between the child and the containers before the child answers enables younger children to perform better (Bruner 1964).

In the 'delay of gratification' paradigm, when faced with the choice of a smaller, immediate reward or a later, larger reward, children of 3 to 4 years are unable to inhibit going for the immediate reward although they would prefer the larger one. By 5 or 6 years of age, children are better able to wait for the preferred reward (Mischel and Mischel 1983).

Between 4 and 5 years of age children show improvement on several tests in the CANTAB battery that are designed to assess frontal lobe function (Luciana and Nelson 1998). For example, their improved ability to retain temporal order information in memory is demonstrated by better performance at 5 years on (a) a test of spatial memory span that is similar to the children's game, Simon, or the Corsi block task (the child must touch boxes displayed on a computer screen in the order in which he or she just saw them change color); and on (b) a spatial self-ordered pointing task where the child must keep track of which squares he or she has already touched.

The 'day-night' task (Gerstadt et al. 1994) requires that children hold two rules in mind and inhibit the tendency to say what the stimuli really represent; instead they must say the opposite ('Say "night" when shown a white card with a picture of the sun, and say "day" when shown a black card with a picture of the moon and stars'). Children of 3½ to 4½ years find the task very difficult; by 6 or 7 years it is trivially easy. If

abstract designs are used as the stimuli, even children of $3\frac{1}{2}$ have no difficulty correctly saying 'day' to one and 'night' to the other (Gerstadt et al. 1994). Hence, the need to learn and remember two rules is not in itself sufficient to account for the poor performance of young children on the task.

Luria's 'tapping' test (Luria 1966) also requires (a) remembering two rules and (b) inhibiting a prepotent response, making the opposite response instead. Here, one needs to remember the rules, 'Tap once when the experimenter taps twice, and tap twice when the experimenter taps once,' and one needs to inhibit the tendency to mimic what the experimenter does. Adults with large frontal lobe lesions fail this task (Luria 1966). Children improve on the task over the same age range at which they improve on the day-night task (Diamond and Taylor 1996). Moreover, performance on the two tasks is correlated so that children whose performance on the day-night task is delayed or accelerated show a corresponding delay or acceleration in their performance on the tapping task (Diamond et al. 1997, Diamond 2001).

The 'counting span' and 'spatial span' tasks of Case (1972) require finding target stimuli, counting them or remembering their locations, holding that information in mind while finding target stimuli in new arrays, and keeping track of the answers computed on each trial so that they can be repeated back in order at the end. Thus, these tasks require temporal order memory and resisting interference from prior and interpolated activity. A meta-analysis of 12 cross-sectional studies revealed strong linear improvements on both tasks between $4\frac{1}{2}$ and $6\frac{1}{2}$ years (Case 1992). Performance on these tasks is highly correlated with performance on Piaget's 'balance beam' task, 'Raven's matrices,' and 'concept acquisition.' On the balance beam and Raven's matrices tests, 4-year-olds tend to answer accordingly to what is perceptually salient. By 7 or 8 years of age, they are able to take two dimensions, or two perspectives, into account and relate one to the other. They can mentally manipulate, combine, recombine, order, and reorder information.

6. Anatomical Evidence of Maturational Changes in PFC From Age 3–7 Years

The density of neurons in human DL-PFC is highest at birth and declines thereafter. At 2 years of age it is 55 percent above the adult mean, but by 7 it is only 10 percent above adult levels (Huttenlocher 1990). Thus there is a dramatic change in neuronal density in DL-PFC between the ages of 2 and 7. The synaptic density of layer III pyramidal cells in DL-PFC increases after birth and reaches its maximum at about 1 year of age; by 7 years of age the decrease in synaptic density is significant, though not yet down to adult level (Huttenlocher 1979). Another change during this period is a marked expansion in the dendritic trees of

layer III pyramidal cells in human DL-PFC between 2 and 5 years of age (Mrzljak et al. 1990). In addition, the density of neuropeptide Y-immunoreactive neurons in human DL-PFC increases between the ages of 2 and 4 years and 6 to 7 years (DeLalle et al. 1997).

7. Behavioral Evidence of Improvements in Cognitive Functions that Depend on PFC From 7 Years to Adulthood

Some abilities dependent on DL-PFC appear to reach their adult levels by about 11 years of age. The Wisconsin Card Sort test is one of the classic tests for studying PFC function in adults (Milner 1964, Stuss et al. 2000). Here, the participant must deduce the rule for sorting cards, which can be sorted by color, shape, or number. Children do not begin to reach adult levels of performance on this task until they are about 11 years old (Chelune and Baer 1986). Continuous, marked improvement is seen on the counting and spatial span tasks through age 8, and then improvement is less steep until performance asymptotes around 10 or 11 years of age (Case 1972, 1992). Pascual-Leone's tasks (the compound stimulus visual information and the digit placement tests) show similar developmental progressions (Pascual-Leone 1970, Case 1972). Hale et al. (1997) report that resistance to interference improves from 8 to 10 years, when it approximates adult levels.

Other abilities dependent on PFC continue to show improvement into adulthood. Zald and Iacono (1998) report that spatial working memory improves from 14 to 20 years of age. They found that 20-year-olds were significantly more accurate than 14-year-olds at indicating the location of an object in space using memory even after brief delays of only half a second. (There was no age difference in the ability to accurately indicate an object's location using visual feedback.) The improvement with age was in the ability to accurately get the information into working memory, not in the ability to hold it in mind for a longer period. Indeed, they found little difference over age in the rate of degradation of the memory.

Another task that requires precise spatial memory is the pattern span task (Wilson et al. 1987). On any given trial some of the squares in a grid pattern are filled and some unfilled. After a quick look, the participant is presented with the grid with one change; the participant is to find that change. Performance on this task improves dramatically between 5 and 11 years, when it reaches adult levels (Wilson et al. 1987).

Several bodies of work indicate that the ability to inhibit a prepotent response tendency continues to improve until early adulthood. In the 'directed forgetting' paradigm, participants are directed to forget some of the words they are shown and to remember others. Even children as old as roughly 11 years show more intrusions of the to-be-forgotten words than do

adults (e.g., Harnishfeger and Pope 1996, Lehman et al.). The 'anti-saccade' task requires the participant to suppress the tendency to reflexively look to a visual stimulus in the periphery, and instead look away in the opposite direction. It depends especially on the frontal eye fields, and the supplementary eye fields and DL-PFC (O'Driscoll et al. 1995, Luna et al.). Performance on the task improves continuously from 8 until about 20 to 25 years of age (Munoz et al. 1998, Luna et al. (in press)). Luna et al. (in press) report that while activation in the frontal eye fields, supplementary eye fields, and DL-PFC increased during anti-saccade performance in participants of all ages, increased activation of the thalamus, striatum, and cerebellum was seen only in adults, suggesting that the circuit connecting PFC with subcortical regions might mature late.

8. Anatomical Evidence that PFC is not Fully Mature Until Adolescence or Early Adulthood

One of the functions of glial cells is to provide an insulating sheath, called myelin, around the axons of neurons, which increases the speed of transmission of communication between neurons and which acts as an electrical insulator, decreasing lateral transmission and interference, thus improving the signal:noise ratio. Myelination of PFC is protracted and does not reach adult levels until adolescence (Huttenlocher 1970, Giedd et al. 1999). For example, using MRI and following the same children longitudinally, Giedd et al. (1999) were able to show that the amount of white matter (i.e., myelinated axons) increased linearly in frontal cortex from 4 to 13 years of age.

Portions of the neuron that are unmyelinated, such as the cell body, have a gray appearance. In their longitudinal study, Giedd et al. (1999) found that gray matter in frontal cortex increased until adolescence, reaching its maximum extent at 12 years of age for males and 11 years for females. However, in cross-sectional volumetric studies, Jernigan et al. (1991) and Sowell et al. (1999a) report *reductions* in gray matter volume between childhood and adolescence, with the most dramatic changes occurring in dorsal frontal and parietal cortex. Sowell et al. (2001) related these gray matter changes to cognitive performance and found that, between 7 and 16 years of age, gray matter in frontal cortex (which included in their analyses not only PFC, but motor cortex, and all cortex in between) decreased in size, and the ability to accurately remember which words had and had not been presented earlier (i.e., the ability to remember which words had been seen in the present context and to discriminate them from other familiar words) improved. Impressively, gray matter thinning in frontal cortex was significantly correlated with this source memory, independent of age. Indeed, while the relation between frontal cortex gray matter thinning and this ability

remained significant controlling for age, the relation between age and source memory was no longer significant controlling for frontal gray matter changes.

Synaptogenesis occurs concurrently with myelination. Huttenlocher (1979) reports that the synaptic density of layer III pyramidal cells in DL-PFC increases until about the age of 1 year, and then decreases, finally reaching adult levels at about 16 years of age. Huttenlocher and Dabholkar (1997) report that the formation of synaptic contacts in DL-PFC reaches its maximum after 15 months of age, and synapse elimination occurs late in childhood, extending to mid-adolescence for DL-PFC.

Developmental changes in PFC continue into adulthood. Sowell et al. (1999b) report a reduction in the density of gray matter in frontal cortex between adolescence (12–16 years) and adulthood (23–30 years). Kostovic et al. (1988) report that acetylcholinesterase (AChE) reactivity of layer III pyramidal cells in DL-PFC, which begins to develop after the first postnatal year, reaches its peak intensity in young adults.

See also: Brain Development, Ontogenetic Neurobiology of; Cognitive Development in Childhood and Adolescence; Cognitive Development in Infancy: Neural Mechanisms; Piaget's Theory of Child Development; Prefrontal Cortex

Bibliography

- Bell M A, Fox N A 1992 The relations between frontal brain electrical activity and cognitive development during infancy. *Child Development* **63**: 1142–63
- Brown R M, Goldman P S 1977 Catecholamines in neocortex of rhesus monkeys: Regional distribution and ontogenetic development. *Brain Research* **127**: 576–80
- Bruner J S 1964 The course of cognitive growth. *American Psychologist* **19**: 1–15
- Carlson S M, Moses L J, Hix H R 1998 The role of inhibitory processes in young children's difficulties with deception and false belief. *Child Development* **69**: 672–91
- Case R 1972 Validation of a neo-Piagetian capacity construct. *Journal of Experimental Child Psychology* **14**: 287–302
- Case R 1992 *The Mind's Staircase: Exploring the Conceptual Underpinnings of Children's Thought and Knowledge*. L Erlbaum Associates, Hillsdale, NJ
- Chelune G J, Baer R A 1986 Developmental norms for the Wisconsin Card Sorting Test. *Journal of Clinical and Experimental Neuropsychology* **8**: 219–28
- Chugani H T, Phelps M E, Mazziotta J C 1987 Positron emission tomography study of human brain functional development. *Annals of Neurology* **22**: 487–97
- Crofts H S, Herrero M T, Del Vecchio A, Wallis J D, Collins P, Everitt B J, Robbins T W, Roberts A C 1999 Excitotoxic lesions of the caudate nucleus in the marmoset: Comparison with prefrontal lesions on discrimination learning, object retrieval and spatial delayed response. *Society for Neuroscience Abstracts* **25**: 891
- Delalle I, Evers P, Kostovic I, Uylings H B M 1997 Laminar distribution of neuropeptide Y-immunoreactive neurons in

- human prefrontal cortex during development. *The Journal of Comparative Neurology* **379**: 515–22
- Diamond A 1985 Development of the ability to use recall to guide action, as indicated by infants' performance on ABBAR. *Child Development* **56**: 868–83
- Diamond A 1991 Neuropsychological insights into the meaning of object concept development. In: Carey S, Gelman R (eds.) *The Epigenesis of Mind: Essays on Biology and Cognition*. L Erlbaum Associates, Hillsdale, NJ, pp. 67–110
- Diamond A 2001 A model system for studying the role of dopamine in prefrontal cortex during early development in humans. In: Nelson C A, Luciana M (eds.) *Handbook of Developmental Cognitive Neuroscience*. MIT Press, Cambridge, MA, pp. 433–72
- Diamond A, Taylor C 1996 Development of an aspect of executive control: Development of the abilities to remember what I said and to 'Do as I say, not as I do'. *Developmental Psychobiology* **29**: 315–34
- Diamond A, Prevor M B, Callender G, Druin D P 1997 Prefrontal cortex cognitive deficits in children treated early and continuously for PKU. *Monographs of the Society for Research in Child Development* **62**(4): Monograph 252
- Flavell J H 1986 The development of children's knowledge about the appearance–reality distinction. *American Psychologist* **41**: 418–25
- Fuster J M 1989 *The Prefrontal Cortex*, 2nd edn. Raven Press, New York
- Gathercole S E 1998 The development of memory. *Journal of Child Psychology and Psychiatry* **39**: 3–27
- Gerstadt C L, Hong Y J, Diamond A 1994 The relationship between cognition and action: Performance of 3.5–7 year old children on a Stroop-like day-night test. *Cognition* **53**: 129–53
- Giedd J N, Blumenthal J, Jeffries N O, Castellanos F X, Liu H, Zijdenbos A, Paus T, Evans A C, Rapoport J L 1999 Brain-development during childhood and adolescence: A longitudinal MRI study. *Nature Neuroscience* **2**: 861–3
- Goldman-Rakic P S 1987 Circuitry of primate prefrontal cortex and regulation of behavior by representational memory. In: Plum F (ed.) *Handbook of Physiology*. American Physiological Society, Bethesda, MD, Vol. 5, pp. 373–417
- Guittin D, Buchtel H A, Douglas R M 1985 Frontal lobe lesions in man cause difficulties in suppressing reflexive glances and in generating goal-directed saccades. *Experimental Brain Research* **58**: 455–72
- Hale S, Bronik M D, Fry A F 1997 Verbal and Spatial Working Memory in School-Age Children: Developmental Differences in Susceptibility to Interference. *Developmental Psychology* **33**: 364–71
- Harnishfeger K K, Pope R S 1996 Intending to forget: The development of cognitive inhibition in directed forgetting. *Journal of Experimental Child Psychology* **62**: 292–315
- Herberle J F, Clune M, Kelly K 1999 *Development of young children's understanding of the appearance–reality distinction*. Paper presented at the Biennial Meeting of the Society for Research in Child Development, Albuquerque, NM
- Hofstadter M, Reznick J S 1996 Response modality affects human infant delayed-response performance. *Child Development* **67**: 646–58
- Huttenlocher P R 1970 Myelination and the development of function in immature pyramidal tract. *Neurology* **29**: 405–15
- Huttenlocher P R 1979 Synaptic density in human frontal cortex—developmental changes and effects of aging. *Brain Research* **163**: 195–205
- Huttenlocher P R 1990 Morphometric study of human cerebral cortex development. *Neuropsychologia* **28**: 517–27
- Huttenlocher P R, Dabholkar A S 1997 Regional differences in synaptogenesis in human cerebral cortex. *Journal of Comparative Neurology* **387**: 167–78
- Jernigan T L, Trauner D A, Hesselink J R, Tallal P A 1991 Maturation of human cerebrum observed in vivo during adolescence. *Brain* **114**: 2037–49
- Kochanska G, Murray K T, Harlan E T 2000 Effortful control in early childhood: Continuity and change, antecedents, and implications for social development. *Developmental Psychology* **36**: 220–32
- Koenderink M J Th, Uylings H B M, Mrzljak L 1994 Postnatal maturation of the layer III pyramidal neurons in the human prefrontal cortex: A quantitative Golgi analysis. *Brain Research* **653**: 173–82
- Koslovski B, Bruner J S 1972 Learning to use a lever. *Child Development* **43**: 790–99
- Kostovic I, Skavic J, Strinovic D 1988 Acetylcholinesterase in the human frontal associative cortex during the period of cognitive development: Early laminar shifts and late innervation of pyramidal neurons. *Neuroscience Letters* **90**: 107–12
- Lehman E, Srokowski S A, Hall L C, Renkey M E (submitted) Directed forgetting of related words: Evidence for the inefficient inhibition hypothesis
- Lidow M S, Rakic P 1992 Scheduling of monoaminergic neurotransmitter receptor expression in the primate neocortex during postnatal development. *Cerebral Cortex* **2**: 401–16
- Livesey D J, Morgan G A 1991 The development of response inhibition in 4- and 5-year-old children. *Australian Journal of Psychology* **43**: 133–7
- Luciana M, Nelson C A 1998 The functional emergence of prefrontally-guided working memory systems in four- to eight-year-old children. *Neuropsychologia* **36**(3): 273–93
- Luna B, Merriam E P, Minshew N J, Keshavan M S, Genovese C R, Thulborn K R, Sweeney J A (in press) Development of brain function for voluntary response suppression from childhood through adolescence. *Neuroimage*
- Luria A R 1966 *The Higher Cortical Functions in Man*. Basic Books, New York
- Luria A R, Homskaya E D 1964 Disturbance in the regulative role of speech with frontal lobe lesions. In: Warren J M, Akert K (eds.) *The Frontal Granular Cortex and Behavior*. McGraw Hill, New York, pp. 353–71
- Milner B 1964 Some effects of frontal lobectomy in man. In: Warren J M, Akert K (eds.) *The Frontal Granular Cortex and Behavior*. McGraw Hill, New York, pp. 313–34
- Mischel H N, Mischel W 1983 The development of children's knowledge of self-control strategies. *Child Development* **54**: 603–19
- Mrzljak L, Uylings H B M, van Eden C G, Judas M 1990 Neuronal development in human prefrontal cortex in prenatal and postnatal states. In: Uylings H B M, van Eden C G, de Bruin J P C, Corner M A, Feenstra M G P (eds.) *The Prefrontal Cortex: Its Structure, Function, and Pathology. Progress in Brain Research*. Elsevier, Amsterdam, Vol. 85, pp. 185–222
- Munoz D P, Broughton J R, Goldring J E, Armstrong I 1998 Age-related performance of human subjects on saccadic eye movement tasks. *Experimental Brain Research* **121**: 391–400
- O'Driscoll G A, Alpert N M, Matthysse S W, Levy D L, Rauch S L, Holzman P S 1995 Functional neuroanatomy of anti-saccade eye movements investigated with positron emission tomography. *Proceedings of the National Academy of Sciences (USA)* **92**: 925–9
- Pascual-Leone J 1970 A mathematical model for transition—in Piaget's developmental stages. *Acta Psychologica* **32**: 301–45

- Rice C, Koinis D, Sullivan K, Tager-Flusberg H, Winner E 1997 When 3-year-olds pass the appearance-reality test. *Developmental Psychology* **33**: 54–61
- Rosenberg D R, Lewis D A 1994 Changes in the dopaminergic innervation of monkey prefrontal cortex during late postnatal development: A tyrosine hydroxylase immunohistochemical study. *Biological Psychiatry* **15**: 272–7
- Schneider J S, Roeltgen D P 1993 Delayed matching-to-sample, object retrieval, and discrimination reversal deficits in chronic low dose MPTP-treated monkeys. *Brain Research* **615**: 351–4
- Sowell E R, Delis D, Stiles J, Jernigan T L 2001 Improved memory functioning and frontal lobe maturation between childhood and adolescence: a structural MRI study. *Journal of the International Neuropsychological Society* **7**: 312–22
- Sowell E R, Thompson P M, Holmes C J, Batth R, Jernigan T L, Toga A W 1999a Localizing age-related changes in brain structure between childhood and adolescence using statistical parametric mapping. *Neuroimage* **9**: 587–97
- Sowell E R, Thompson P M, Holmes C J, Jernigan T L, Toga A W 1999b In vivo evidence for post-adolescent brain maturation in frontal and striatal regions. *Nature Neuroscience* **2**: 859–61
- Stuss D T, Levine B, Alexander M P, Hong J, Palumbo C, Hamer L, Murphy K J, Izukawa D 2000 Wisconsin Card Sorting Test performance in patients with focal frontal and posterior brain damage: Effects of lesion location and test structure on separable cognitive processes. *Neuropsychologia* **38**: 388–402
- Taylor J R, Elsworth J D, Roth R H, Sladek J R Jr., Redmond D E Jr 1990 Cognitive and motor deficits in the acquisition of an object retrieval detour task in MPTP-treated monkeys. *Brain* **113**: 617–37
- Tikhomirov O K 1978 The formation of voluntary movements in children of preschool age. In: Cole M (ed.) *The Selected Writings of A R Luria*. M E Sharpe, White Plains, NY, pp. 229–69
- Wellman H M, Cross D, Bartsch K 1987 Infant search and object permanence: A meta-analysis of the A-not-B error. *Monographs of the Society for Research in Child Development* **51**: 1–67
- Wilson J T L, Scott J H, Power K G 1987 Developmental differences in the span of visual memory for pattern. *British Journal of Developmental Psychology* **5**: 249–55
- Zaitchik D 1991 Is only seeing really believing?: Sources of the true belief in the false belief task. *Cognitive Development* **6**: 91–103
- Zald D H, Iacono W G 1998 The development of spatial working memory abilities. *Developmental Neuropsychology* **14**: 563–78
- Zelazo P D, Reznick J S, Piñon D E 1995 Response control and the execution of verbal rules. *Developmental Psychology* **31**: 508–517
- Zelazo P D, Frye D, Rapus T 1996 Age-related dissociation between knowing rules and using them. *Cognitive Development* **11**: 37–63

A. Diamond

Prehistoric Art

As far as we know, people have been making objects, forms and images that we consider to be ‘art’ for at least 35,000 years and perhaps even longer. Some

claim that evidence of red ochre, as a kind of coloring material, indicates human engagement with artistic activities (such as body painting) for more than 200,000 years. Others have pointed to individual objects, dated to even earlier than that, which they are convinced were modified by ancient humans with an intent to represent something, suggesting that ‘art’ has long been a part of the cultural repertoire of humans, even of different biological species than we anatomically modern humans. And when most lay people learn that they are talking to an archaeologist, they inevitably ask, ‘What is the most fantastic thing you have ever found?’, hoping we will be able to tell them about a gold decorated necklace, a carved figurine, or magnificent painted jar: they want to know about ‘prehistoric art.’

1. Definitions

The very term, ‘prehistoric art,’ is difficult to define. Technically speaking, it refers to the ‘art’ made by prehistoric peoples; that is, by people in those societies without, and usually before the development of, written texts. There have in recent times been human societies without writing systems (nonliterate) and with what might be called ‘the arts,’ but these are more likely to be what we call ‘ethnographic arts’ because they are still ‘living’ arts and can be described ethnographically. There are some problems with the way in which the term ‘prehistory’ is used, including the implication that, without writing systems, prehistoric groups have no history, and experienced little change. But we know this is not the case at all, and that there are myriad ways other than in writing by which ‘history’ is reckoned, transmitted, and experienced. Also, in relation to the idea of ‘prehistoric art,’ there are some archaeologically known societies, such as the Maya or Egyptians, who had writing systems, but whose ‘arts’ are sometimes referred to as prehistoric or, more often, as ancient art. In fact, it is perhaps better to think of the arts that we are interested in as the archaeological arts, rather than prehistoric ones. With such a term, we realize we are dealing with the art(s) from past human societies that we study by using the wide range of archaeological methods, theories, and techniques.

Then there is the term ‘art,’ which is itself notoriously difficult to define. Furthermore, it is a term that has had differing meanings even within the last two centuries of Western scholarship and culture. Since the nineteenth century, ‘we’ in the Western tradition have associated the term, ‘art,’ with aesthetics and aesthetically-pleasing attributes; as something that is usually produced by specialists who are often considered to be gifted individuals with admirable degrees of creativity, and who are often separated from real-life, everyday activities. Our present-day notions of art include the idea that art is to be held in awe while, at the same time, the production

and products of art are embedded in the capitalist system, with economic as well as cultural value; art objects function in their own domain, that of high culture and aesthetics.

However, all this is unlikely to be applicable to the art-making and art images and forms that we are concerned with in prehistoric and archaeological contexts. It is a well-known truism in anthropology that in most cultures which have been studied ethnographically, there is rarely a word, or even a concept, equivalent to our word and concept, 'art.' It is not really possible to try to separate out 'art' from other aspects of cultural life, such as art as something distinct from religion, social relations, and/or politics. And yet there are images, forms, and cultural practices in other cultural contexts, both ethnographic and archaeological, that we acknowledge unproblematically as 'art,' mostly because, if they were produced by our own culture, we would classify them readily as art: cave and rock paintings, sculpted figurines or statues, decorated bronze goblets, and the like.

Is there a better term: visual culture? Material culture? Images? Representations? Material symbols? The term or label we use, such as 'art,' is not itself the real problem; what is more of an issue is that if we use such a term in our own sense, without recognizing its historical specificity, then what we mean by it is often left unexamined, and the analysis or interpretation of the phenomenon is led in certain directions because of the label. If what we have is 'art,' then there must be artists and aesthetic meanings. However, as Morphy (1999) has so eloquently noted: we need to develop a concept of art that 'neither distances the objects of other societies from Western cultures nor appropriates the object to European categories.'

Books that treat the topic of 'prehistoric art' as a single entity are rare; most book-length treatments are concerned with the art and imagery of specific regions (e.g., Keyser 1992; see also *Sub-Saharan Africa, Archaeology of*), time periods (e.g., Leroi-Gourhan 1967) or in specific media (Dowson 1992). In 1966, Powell authored an overview of prehistoric art, which was divided into four chapters: the art of the hunters, that of the cultivators, that of the metalworkers, and what he called the 'art of a barbarian nation,' referring to the Celts of Europe (Powell 1966). His coverage was limited to Europe, and he treated each 'kind' of art by apparent categories of 'lifestyle,' but it is just as much a book about the 'progress' of art from the Paleolithic period (the hunters), through the Neolithic (the cultivators), the Bronze and Iron Ages (the metalworkers), to the 'archaic state' (Celts). As such, he presents a now-problematic evolutionist view, a view of 'progress' through different chronocultural periods. Indeed, for much of the nineteenth and twentieth century, the ethnographic (and archaeological) arts have been embedded in a progressivist notion (Errington 1998); we used to talk unquestioningly of 'primitive art' (Price 1989).

More recently, Bahn (1998) has authored an illustrated history of prehistoric art, but it is almost exclusively about 'rock art,' with a short chapter on 'body art' and one on 'objets d'art,' or portable art (items that could circulate). He returns again and again to a baseline of the Ice Age imagery of Europe, although this book is an important source for appreciating that rock art—paintings, engravings, and petroglyphs on rock surfaces, rock shelters and in caves—is truly global in its distribution. Bahn's map of sites is a map of the world, while Powell's is one of Europe. This is indicative of a real shift of scope in our understanding and treatments of 'early art' and/or rock art from all times. As we will see below, the search for the earliest art has been one of the major preoccupations of many who study 'prehistoric art.'

2. How to Study Prehistoric Art

First, it is always important to recognize that the corpus of materials that we consider as 'prehistoric art' is certainly only a sample of what was made in the human past. There is always the problem of preservation, and that what we have in the present is only a sample of past artistic production. It is no coincidence that rock art, ceramic sculpture and decorated pots, metal objects, and images made out of ivory, bone, antler, shell, and beads of various sorts are some of the primary forms of prehistoric art. The wood, hide, basketry, cloth, netting and twining, feathers, and body art are rarely, if ever, preserved. If the arts of ethnographic groups are any indication, we have an impoverished archaeological sample of prehistoric visual and material culture.

But given the preservation of at least some portion of prehistoric artistic production, how do we then go about studying it? First, we need some theoretical or conceptual frameworks within which even to think about what the images and forms might be about, especially if we cannot directly apply our own notions of 'art,' artists, and aesthetics. Much of this has to come from related fields, such as the anthropology of art, the ethnographic arts, and art history. We need to draw on understandings of what visual and material representation is all about (e.g., Gombrich 1961, Bryson 1983), of symbolism (e.g., Geertz 1983), and of humans as materialists and symbolists (Conkey 1993); of how technology and materials are integral to the artforms (e.g., Lechtman 1984), of style (Conkey and Hastorf 1990); and of the social contexts and production of art (e.g., Wolff 1981), and of the 'social life of things' (e.g., Appadurai 1986). Studies of 'art' in ethnographic contexts (e.g., Morphy 1991), general overviews of the anthropology of art (e.g., Layton 1981), and ethnoarchaeological studies (see *Ethnoarchaeology*) provide important resources.

Second, we need to recognize the importance of the archaeological context of the so-called art forms, and

we need to take context into consideration when we try to interpret or understand the meanings, uses, or even the making of the art. Do the preconceptions that female figurines—from about 7,000 years ago from sites in such places as southeast Europe—are objects of worship and a female-centered world change when we recognize that some, if not many, of the figurines have been recovered from trash deposits, often broken? Are the Ice Age paintings of southwestern Europe found deep in caves, away from daylight, comparable to those found on daylight shelters? Is there a difference in the uses and meanings between objects found in everyday places and living sites, and those found in architectural structures that are so different in size, form, and location that they must be other-than-'houses'? The individual images and objects may well take their specific meanings from the locations where they were made and/or found, or from the other artifacts or structures with which they are associated.

We can use the most up-to-date techniques and methods, especially those of materials science, to make inferences about the images and objects: we can identify the sources of the raw materials, such as clay, stones, or pigments, which in turn can inform on trade, exchange, circulation of goods, and the social relations that effect such practices. We can identify different ways in which pigments were processed by Ice Age cave 'artists' in Europe (e.g., Clottes 1993), and therefore suggest the social relationships that must have obtained in order to have images in different sites made by the same pigment 'recipes.' We can show how certain fibers were used to apply pigment in Australian rock paintings (e.g., Cole and Watchman 1992). We can infer the stages of manufacture of metal earspools (e.g., Lechtman 1984) and how these stages resonate with other cultural principles and practices. We can use infrared photography to 'see' color and images that have long since faded away. We can use high powered microscopic analysis to detect the precise strokes of a stone tool on the surface of a bone object; from such analyses have come the suggestion that lunar calendars may have been made at least 15,000 years ago (Marshack 1991). These are just a few examples of how our increasingly sophisticated methods of materials science that are common in the analysis of other archaeological materials can be applied to the study of prehistoric art images and objects.

Both of these approaches—knowing the archaeological contexts of the art, and the application of scientific techniques—have featured prominently in the increasing ability to date many objects and images. It has long been the case that objects that are found in stratigraphic context—that is, in the geological/archaeological layers of an archaeological site—can, if excavated and recorded carefully, be attributed to at least a relative date (see *Chronology, Stratigraphy, and Dating Methods in Archaeology*). They can also be given more secure dates if other artifacts or materials

in the same stratigraphic layer are suitable for some of the many techniques of archaeological or chronometric dating (see *Archaeometry and Chronology, Stratigraphy, and Dating Methods in Archaeology*). However, many art forms, such as rock art or objects that are unfortunately out-of-context, cannot be dated in this associative manner. And many of these images themselves cannot be dated at all, although recent advances have changed this somewhat. Earlier generations of scholars, for example, dated all the cave art of Europe or the varying rock art styles of Australia by means of stylistic criteria, often based on assumptions as to what 'ought' to be earlier image-making and what 'ought' to be later. At the same time, many researchers were reluctant even to consider the art as a subject worthy of archaeological study, primarily because it could not be dated (see Whitley and Loendorf 1994).

That much art cannot be dated has, in combination with the fact that the interpretation of art is never straightforward (if only because it is the product of symbolic cultural practices that are not easily comprehended through the usual archaeological methods), led many archaeologists to dismiss art and images as viable archaeological data. This has been the case especially when it comes to objects and images from many thousands of years ago, where there are no real historic links between the archaeological 'cultures' and those known ethnohistorically or ethnographically, where there is no direct historical method that can be applied (see *Theory in Archaeology*). Some have suggested that the recent explosion of interest in the global study of rock art is due, in part, to new dating methods that have allowed researchers to begin to date imagery with more security. For this genre of prehistoric art—rock art—the most important new technique has been the development of accelerator mass spectrometry (AMS) carbon-14 dating, which can be done using a very minute sample of material, thus not destroying a painting or image, although it can only be used if there are organic materials in the pigments, such as charcoal (see *Archaeometry*).

3. Some Major Questions

At least two major questions come to mind when we sit down to address the subject of prehistoric art: first there is the question of the origins or the 'birth' of art; second, there is the challenging issue of how we can interpret the art and approach what it might mean. Until recently, it was thought that the cave paintings and other arts of the Upper Paleolithic period (c. 30,000–10,000 years ago) in Europe comprised the 'first' art, even though images from ancient cultures had been found all over the globe (see Bahn 1998 on the 'discovery' of prehistoric art). This was a compelling idea because there were many different caves

with images and many other objects and art forms; because it seemed to be associated with the earliest evidence for anatomically modern humans—our own species, *Homo sapiens*; and because it was in Europe, which allowed for the idea of a continuous tradition of art—a crucial essence of what it means to be human—from deep prehistory up into the twenty-first century.

However, work during the last few decades of the twentieth century has undermined seriously the idea that there is any single ‘birthplace’ of art: both southern Africa and Australia have yielded rock art images that can be attributed to the same general time period as the European Paleolithic imagery. And there are those who point to more isolated finds (see Bahn 1998, pp. 86–7), although many are debated hotly as being ‘art’ or intentional representations at even earlier dates. And we now recognize that anatomically modern humans have a deeper history, perhaps as much as 200,000 years longer, than previously thought (see *Sub-Saharan Africa, Archaeology of*). However, the exact nature of the relationship between being anatomically modern and the appearance of representational and/or symbolic imagery is hardly resolved, but the fact that this kind of imagery appears in multiple locations, in varying forms and amounts, calls for a more contextual approach: what is it about life and social relations at these times that set a context within which ‘art’—at least in more durable forms that have been preserved—became a part of cultural practices? While the search for ‘the origins’ of many phenomena in the human past is always a seductive question, as far as prehistoric art is concerned, it is now clear that there was no single ‘center’ for art-making, but that the preserved evidence around the globe attests the engagement of ancient humans with materials, images, and technologies (of working stone, of coloring materials, etc.) for cultural purposes that still elude us.

And thus, the final, but most difficult question is the interpretive one: how can we infer some (if any) of the meanings of prehistoric art? This question is both the most difficult and most complex question, and what everyone wants to know. Since all of archaeology is interpretive, while drawing on as much empirical evidence as possible, and since human behavior is so multivocal, with so many potential meanings and ambiguities, it is unlikely to expect a single answer, a single interpretation. Even in known or ethnographic contexts, the so-called meaning of art-making or of the art itself is ever-shifting, contextual, and multiple, dependent upon viewers, users, makers, and situations. Do we assume that prehistoric peoples were somehow ‘simpler’ and therefore a (simple) single account can suffice?

Despite the complexity of the challenge, there are some things that we can say about interpreting prehistoric art. In some instances, we can infer information clearly about clothing from human depictions, about activities as depicted in ‘scenes,’ preferred

symbols or forms, and what kinds of object were traded or used in certain contexts. In other instances, we can advance general theories, such as the role of shamans in making images that derive from their visions (e.g., Lewis-Williams and Dowson 1989, Turpin 1994), such as the ways in which ritual landscapes were created with artforms and megaliths, such as the idea that some arts appear to be ‘about’ interpersonal social relations, whereas other arts appear to be ‘about’ maintaining the corporate entity (e.g., Layton 1985). While we are unlikely ever to fix, with certainty, what a particular picture or object ‘meant,’ recent past decades have witnessed enormous gains, thanks to the application of scientific techniques, the innovative and expanded theories for interpretation, and a more widespread recognition that the archaeological arts are just as crucial to our understanding of past societies as are other kinds of archaeological data.

Bibliography

- Appaduri A 1986 *The Social Life of Things*. Cambridge University Press, Cambridge, UK
- Bahn P 1998 *The Cambridge Illustrated History of Prehistoric Art*. Cambridge University Press, Cambridge, UK
- Bryson N 1983 *Vision and Painting: The Logic of the Gaze*. Yale University Press, New Haven, CT
- Clottes J 1993 Paint analyses from several Magdalenian caves in the Ariège region of France. *Journal of Archaeological Science* 20: 223–35
- Cole N, Watchman A 1992 Painting with plants: Investigating fibers in Aboriginal rock paintings at Laura, North Queensland. *Rock Art Research* 9
- Conkey M 1993 Humans as materialists and symbolists. In: Rasmussen D T (ed.) *The Origins and Evolution of Humans and Humanness*. Jones and Bartlett, Boston, MA, pp. 95–118
- Conkey M, C Hastorf (eds.) 1990 *The Uses of Style in Archaeology*. Cambridge University Press, Cambridge, UK
- Dowson T 1992 *Rock Engravings of Southern Africa*. Witwatersrand University Press, Johannesburg, South Africa
- Errington S 1998 *The Death of Authentic Primitive Art and Other Tales of Progress*. University of California Press, Berkeley, CA
- Geertz C 1983 Art as a cultural system. In: Geertz C (ed.) *Local Knowledge: Further Essays in Interpretive Anthropology*. Basic Books, New York
- Gombrich E 1961 *Art and Illusion: A Study in the Psychology of Pictorial Representation*, 2nd edn. Pantheon, New York
- Keyser J D 1992 *Indian Rock Art of the Columbia Plateau*. University of Washington Press, Seattle, WA
- Layton R 1981 *The Anthropology of Art*. Columbia University Press, New York
- Layton R 1985 The cultural context of hunter-gatherer rock art. *Man (N.S.)* 20: 434–53
- Lechtman H 1984 Andean value systems and the development of prehistoric metallurgy. *Technology and Culture* 25(1): 1–36
- Leroi-Gourhan A 1967 *Treasures of Prehistoric Art*. Harry Abrams, New York
- Lewis-Williams J D, Dowson T 1989 *Images of Power: Understanding Bushman Rock Art*. Southern Book Publishers, Johannesburg, South Africa

- Marshack A 1991 *The Roots of Civilization*. McGraw-Hill, New York
- Morphy H 1991 *Ancestral Connections: Art and an Aboriginal System of Knowledge*. University of Chicago Press, Chicago
- Morphy H 1999 Traditional and modern visual art of hunting and gathering peoples. In: Lee R B, Daly R (eds.) *The Cambridge Encyclopedia of Hunters and Gatherers*. Cambridge University Press, Cambridge, UK, pp. 441–8
- Powell T G E 1966 *Prehistoric Art*. Praeger, New York
- Price S 1989 *Primitive Art in Civilized Places*. University of Chicago Press, Chicago
- Turpin S (ed.) 1994 *Shamanism and Rock Art in North America*. Rock Art Foundation, San Antonio, TX
- Whitley D, Loendorf L 1994 *New Light on Old Art: Recent Advances in Hunter-gatherer Rock Art Research*. Monograph No. 36, UCLA Institute of Archaeology, Los Angeles, CA
- Wolff J 1981 *The Social Production of Art*. St. Martin's Press, New York

M. W. Conkey

Prejudice in Society

Prejudice is defined as cognitive beliefs, affects, and discriminatory behaviors towards members of a group on account of their membership to this group. When race, other groups, sex, or sexual preferences, for instance, are the targets of prejudice, it will be labeled racism, ethnocentrism, sexism, or homophobia. Many authors add that prejudice relies on undue generalizations and is irrational. Such addition implies that one knows what the objective truth is. It also overlooks the sophisticated ways that prejudice can take.

1. Classical Explanations

Personality was one of the first explanations to be proposed. After World War II, researchers explained fascism, ethnocentrism, and antisemitism by socialization (Adorno et al. 1950). Because children were obliged to obey parental authority blindly and to repress their hostile ideas, they later had to find scapegoats as a source of catharsis and represented the world into simplified categories: the good and the bad ones. This 'right-winged' authoritarian personality later was expanded to include all dogmatisms (Rokeach 1960). Whether authoritarian or dogmatic, personality cannot explain rapid changes in prejudice. Moreover, it is unable to account for regional differences in prejudice and it is mute about why certain groups, but not others, are the targets of prejudice.

Other psychologists influenced by behaviorism and psychoanalysis proposed a general theory of aggression that focused on situations (Dollard et al. 1939). According to them, aggression was a necessary consequence of frustration, and frustration was the neces-

sary and sufficient antecedent of aggression. Applied to prejudice, this theory predicted that socio-economic factors (e.g., the price of cotton) would be determining in eliciting aggressive discriminatory behaviors (e.g., lynching of Blacks in the US). Besides being unfalsifiable, this theory is not better than personality ones in order to predict which group will be selected as a scapegoat.

Lack of contact was a third explanation (Allport 1954). It played a crucial role when the US Supreme Court ruled segregation in schools illegal. Socio-psychological experts testified that prejudice was due to lack of contacts between the groups. These psychologists were not naive to the point of thinking that contact *per se* was sufficient to eliminate prejudice. They listed a series of conditions under which contact would be effective. Unfortunately, in everyday life, these conditions are sometimes difficult to meet, such as contact restricted to atypical members of the stigmatized group and collaborative institutional environment. That will be discussed under different headings later on, but lack of contact is still a predominant explanation of prejudice. Sherif (1966) was one of the opponents to the contact hypothesis. For him, life is a competition for scarce resources and mere contact will render competition even harder. In order to be successful, Sherif argued, reduction of tension depends on supra-ordinate collaboration, that is, a collaboration that needs the efforts of the two otherwise competing groups.

In reaction to Sherif's conceptualization, proponents of Social Identity Theory stressed the fact that resources need not be real to induce conflict (Tajfel 1981). Symbolic resources may suffice. For these theoreticians, mere categorization into groups will lead to ethnocentrism, also called in-group favoritism bias, if the latter brings out a positive social identity to the discriminators.

2. Modern Forms of Prejudice

Prejudice or at least its expression has decreased over the years. When the measure of stereotypes, that is, measure of general beliefs about groups' behaviors and personality, was discovered in the early 1930s, it was a good thing for white Americans to express negative stereotypes about African-Americans. Whites gave more negative stereotypes about African-Americans when responding publicly than when answering anonymously. Now, the targets of open prejudice or stereotypes become rarer although they still exist (e.g., integrists). Such a situation does not mean that prejudice has disappeared. According to the authors, it is either the form or the conditions of expression that have changed. Here, we review briefly the different conceptualizations (see Dovidio and Gaertner 1986).

The most pessimistic view is probably the one of symbolic racism. For its proponents, prejudice has

remained the same but its expression has become subtler because of the pressure from the society. The most optimistic view is represented by regressive racism, which suggests that prejudice has decreased but that it comes again to the surface in stressful situations. Between these two extremes, aversive racism defends the idea that people persuade themselves that they are not prejudiced and try to rationalize what could be considered prejudiced in their behavior. Under the label of ambivalent amplification racism, other theoreticians hypothesize that people have ambivalent feelings about others. According to the circumstances, these ambivalent feelings are amplified either towards the absence of prejudice or towards its expression.

In fact, prejudice itself may take ambivalent forms. To the same extent that sociability and competence are the two dimensions organizing person perception, they may also guide the perception of groups. Groups may be perceived as likeable but incompetent (e.g., housewives) or they may be considered competent but unlikable (e.g., feminists). Competence would be triggered by the status of the group in the society (e.g., feminists are perceived as intellectuals). Likability, on the other hand, would result from the outcome dependence between the dominant and the dominated groups in society (e.g., the males 'need' housewives). Because of these differences in the perception of groups, people will try to explain them. The most radical explanation is to attribute them a different (genetic, cultural, linguistic, religious) essence. Because group members are all and foremost concerned with their in-group, they will attribute to it the human essence. It follows that they will attribute an infrahuman essence to 'others.' This infrahuman essence is characterized by a lack of typically human attributes. According to the circumstances, these deficient attributes can be the intelligence, the language, some typically human emotions such as love or sorrow, etc. (Leyens et al. 2000).

Finally, to the same extent that lack of contact was thought to reduce prejudice, many authors have argued that lack of cognitive resources may facilitate the expression of prejudice. All people know the negative stereotypes held by different groups and many do not share them. However, when they are overloaded by a demanding environment and in need of cognitive capacities, they may let prejudice leak. Such perspective makes a clear distinction between beliefs or stereotypes, on the one hand, and prejudice on the other (Devine 1989). It is obvious that, in normal circumstances, the attribution of an infrahuman essence is only possible under lack of cognitive control. In war times, on the contrary, this attribution often constitutes a conscious strategy.

Besides their differences, these different conceptualizations share one common point. Increasingly, prejudice is considered less a reaction against some other groups of individuals but more a measure of

protection of the in-group. In general, people are less concerned by the other groups than by the well being of their own group.

3. *Implicit and Explicit Measures of Prejudice*

Prejudice typically was measured by questionnaires. To give but one example, the 'social distance scale' proposed different items of the form: 'I would accept a member of X (the group concerned) in the country, in the street, at work, as a friend, as a member of my family, etc.' A response at one level (e.g., at work) implied acceptance of the lower levels (e.g., country and street). As the expression of prejudice changed, its measures had to be adapted. From being blatant, questionnaire items have become more subtle but it is still easy to guess their purpose and, therefore, to control one's answers. These questionnaire measures are called explicit.

Some behaviors are obviously a more valuable source of information than questionnaires. Prosocial behaviors in the street or hiring rates in high-status positions on the job market may be taken as valid indicators of nonprejudice.

Laboratory implicit measures have also proliferated. Some of these measures have the advantage that the participants to the studies are not at all aware of what is measured. For instance, participants have to decide whether the string of letters they see on the screen of a computer is a word or a nonword. Among the words there are some positive or negative words. Before the presentation of some of the words, a social category is primed subliminally. The typical finding of such procedure is that the reaction times will be shorter for positive words when the social category is liked and for negative words when the social category is disliked.

The problem still to be solved concerns the correlations between these different measures. The questionnaire measures usually correlate highly between them. They rarely correlate, however, with behaviors and implicit measures. Also, different implicit measures do not necessarily correlate between themselves, which means that they probably measure different aspects of prejudice.

4. *Automatic and Controlled Prejudice*

The pervasive expression of prejudice and the development of implicit measures have launched an important and provocative debate about automatic vs. controlled behavior. One is reminded here of the unconscious in psychoanalysis.

Not long ago, most social psychologists believed that people made errors in their social judgments and, therefore, showed prejudice because they were unmotivated or incompetent to attend correct infor-

mation. Intelligence (or information) and effort were thought to be able to counterbalance these 'errors.' At present, some researchers do not hesitate to claim that people's first reactions are automatic and that neither abilities nor motivation can help. Automaticity does not imply inevitability of prejudice. It only means that people with prejudice will show it in their first reactions no matter how hard they try not to.

Automatic behavior will also depend on what is activated. A Black person is not only Black; she may also be tall, attractive, a physician, a woman, etc. The activation of the concept 'woman' may induce seduction. The activation of 'physician' may lead to help-seeking behavior. Among many persons, however, there is a limit to this flexibility. A lot of individuals will see a physician when this Black physician brings good news, and they will see a Black person when the physician is the bearer of bad news.

This link between a category and its attributes has implications. Active suppression of a category (e.g., skinhead) may, paradoxically, lead to a rebound effect. Stated otherwise, people who are instructed not to think of someone as a skinhead may, later on, perceive him as more prototypical of a skinhead. Indeed, by an ironic process, the monitoring of 'not-thinking-skinhead' activates attributes pertaining to the category. When the suppression is abandoned, all the activated attributes pop up; it is the rebound effect.

Up to now, we have reasoned as if activation of a category and of its attributes always led to an assimilation process (i.e., prejudice). It is true that people who are primed with the concept of professors do better at a Trivial Pursuit game than those who have been primed by the concept 'hooligan.' Priming specific exemplars of a category, like Einstein, may lead to a contrast effect. Indeed, everyone realizes that they are not Einstein and far from resembling him. This example leads us naturally to the attempts at reducing prejudice.

5. Reduction of Prejudice

One can distinguish between two approaches in the studies that looked at the reduction of prejudice. One approach focuses on information about individuals and the other concentrates on categorical representation of the world.

5.1 The Focus on Individuals

This approach relies mostly on the intellectual (and somewhat emotional) abilities to take into account evidence contradictory with one's prejudice. Three models have been proposed as possible ways to reduce prejudice:

5.1.1 The conversion model. According to this model, which could also be called the Sidney Poitier or Mandela model, the encounter with a very atypical member of a disliked category could provoke a conversion that would generalize to the whole category. Although anecdotal evidence is not lacking in favor of this model, no experimental demonstration has yet succeeded.

5.1.2 The book-keeping model. For the proponents of this model, people would integrate the different pieces of information that they receive from various members of the stigmatized group and that contradict their beliefs. Reduction of prejudice would be proportionate to the contradicting information. Empirical evidence in support of this model exists but it is not overwhelming.

5.1.3 The subtyping model. When confronted with contradicting information, people would try to subcategorize the exceptions in order to maintain their representation of the whole category intact. 'These women are unemotional; but they are not real women, they are businesswomen.' Needless to say that this strategy is often used.

5.2 The Focus on Categories

It is not surprising that this approach has been more elaborated than the previous one by psychologists because it goes back to the definition of prejudice, that is, reaction against individuals because of the category they belong to.

5.2.1 Decategorization. The early proponents of the contact hypothesis were members of the human relations movement and believed that solutions to conflicts would come from considering anybody as an individual rather than part of a given category. In spite of the disaster brought up by the compulsory desegregation of schools in the US, decategorization can be effective in eliminating the in-group favoritism bias.

5.2.2 Recategorization. It is often unrealistic to transform group members into mere individuals. Supporters of recategorization inspired themselves of Sherif's (1966) superordinate goal. This goal could only be achieved when all parties put their forces together and create some kind of new entity. Re-

categorization also eliminates the in-group favoritism bias, but it does it differently and more advantageously than decategorization. Whereas decategorization brings the attraction to the in-group down to the level of the out-group, recategorization carries the attraction of the out-group up to the level of the in-group. Recategorization transforms the 'we' and 'they' into 'us.'

5.2.3 Cross-categorization. It is said that Belgium never had bloody conflicts despite its history of linguistic problems because language was not the only divider in the country; religion and political orientations also played a role. Language cross-cuts religion and ideology. The problem with cross-categorization is that some people will find similarities with the other side, but that others will be completely different.

5.2.4 Harmonious differentiation. European researchers especially are aware that it may be neither feasible nor desirable to get away with categories. Some of them defend a blend of decategorization while at the same time proposing means to respect existing categories. This stream of research is still in need of convincing results.

6. The Targets' Reactions to Prejudice

Members of stigmatized group may suffer from prejudice and react to it in different ways (Crocker et al. 1998). When treated badly, for example, some may prefer to attribute the situation to discrimination rather than to their misconduct. Such a solution preserves their self-image. Others prefer to condemn their own behavior because this behavior can change whereas they will remain forever in their stigmatized group. The conditions that stir up the first or the second reaction are not yet well known.

Another solution is to find innovative ways to affirm a valuable identity. A typical example is the case of the African-Americans who invested in sport when the slogan 'Black is beautiful' lost its momentum. Simultaneously to this investment in a new domain, members of stigmatized groups may disengage themselves from goals that are denied to them. School or work may become unimportant for some of them. Indeed, very often, immigrants' children do poorly in school and their peer group is likely to disavow those who do well.

Besides the lack of social support, the atypical members of stigmatized groups suffer an additional burden in competitive situations like exams or job interviews. Not only do they feel the anxiety linked to the situation, but they have the additional anxiety to demonstrate that they are atypical of their group. It is,

for instance, the case of women and mathematics, or of children from poor families and school achievement in general. In a way, it is a fulfilling prophecy.

See also: Adorno, Theodor W (1903–69); Allport, Gordon W (1897–1967); Anti-Semitism; Attributional Processes: Psychological; Authoritarian Personality: History of the Concept; Collective Beliefs: Sociological Explanation; Discrimination; Discrimination: Racial; Ethnic Identity, Psychology of; Ethnicity, Sociology of; Gender, Class, Race, and Ethnicity, Social Construction of; Identity: Social; National Socialism and Fascism; Racism, History of; Racism, Sociology of; Stereotypes, Social Psychology of

Bibliography

- Adorno T W, Frenkel-Brunswick E, Levinson D J, Sanford R N 1950 *The Authoritarian Personality*. Harper and Row, New York
- Allport G W 1954 *The Nature of Prejudice*. Doubleday Anchor, New York
- Crocker J, Major B, Steele C 1998 Social stigma. In: Gilbert D T, Fiske S T, Lindzey G (eds.) *The Handbook of Social Psychology*, 4th edn. McGraw-Hill, Boston, Vol. 2, pp. 504–53
- Devine P G 1989 Stereotypes and prejudice: Their automatic and controlled components. *Journal of Personality and Social Psychology* **56**: 5–18
- Dollard J, Doob L, Miller N E, Mowrer O, Sears R 1939 *Frustration and Aggression*. Yale University Press, New Haven, CT
- Dovidio J F, Gaertner S L 1986 *Prejudice, Discrimination, and Racism*. Academic Press, New York
- Eberhardt J L, Fiske S T 1998 *Confronting Racism. The Problem and the Response*. Sage, Thousand Oaks, CA
- Leyens J-P, Paladino M P, Rodriguez R T, Vaes J, Demoulin S, Rodriguez A P, Gaunt R 2000 The emotional side of prejudice: The attribution of secondary emotions to ingroups and outgroups. *Personality and Social Psychology Review* **4**: 186–97
- Leyens J-P, Yzerbyt V, Schadrone G 1994 *Stereotypes and Social Cognition*. Sage, London
- Rokeach M 1960 *The Open and the Closed Mind*. Basic Books, New York
- Sherif M 1966 *In Common Predicament*. Houghton Mifflin, Boston
- Tajfel H 1981 *Human Groups and Social Categories: Studies in Social Psychology*. Cambridge University Press, Cambridge, UK

J.-P. Leyens

Pre-motor Cortex

The frontal lobe of primates is formed by two large regions, a rostral region involved in cognitive functions (prefrontal lobe) and a caudal one devoted to motor functions. Histologically, the caudal region is

characterized by the absence of granular cells. It is referred to also as the agranular frontal cortex (AFC). The AFC consists of two main sectors: area 4 (the primary motor cortex) and area 6. Recent data showed that area 6 is neither anatomically nor functionally unitary, but is formed by a mosaic of independent 'premotor' areas (see He et al. 1993, Rizzolatti et al. 1998). The aim of the present article is to review the general organization of the premotor areas in monkeys and to discuss in detail the operations of two of them. A comparison between the monkey and human premotor cortices will be made in the last section.

1. The General Organization of the Premotor Area: Classes and Connectivity

The existence of a cortex with motor functions lying rostral to the primary motor cortex (the 'premotor cortex') has been postulated since the beginning of the twentieth century. However, following the stimulation studies of Woolsey and Penfield, the prevalent view was that AFC functionally consists of two large motor areas: the 'primary motor cortex' located on the lateral brain convexity (MI) and the 'supplementary motor area' (SMA) located on its mesial surface (see Porter and Lemon 1993). Anatomical and functional evidence collected from 1980–2000 showed that this view is incorrect. The AFC is formed by seven independent areas (F1–F7). F1 corresponds to the Brodmann area 4 (the 'primary motor cortex'); the other areas are located within the Brodmann area 6 (the 'premotor areas'). This new subdivision of the AFC is shown in Fig. 1.

The premotor areas can be subdivided into two main classes: areas that receive their predominant cortical input from the parietal lobe (parieto-dependent premotor areas) and areas that receive their predominant cortical input from the prefrontal and the cingulate cortex (fronto-dependent premotor areas). Areas F2, F3, F4, and F5 belong to the first class, areas F6 and F7 to the second (Rizzolatti et al. 1998).

The organization of corticospinal projections is in accord with this subdivision (He et al. 1993). The parieto-dependent areas have direct access to the spinal cord, whereas the fronto-dependent areas send their output to the brain stem. Furthermore, the parieto-dependent areas have direct connections with F1, whereas the fronto-dependent areas lack this link.

Within the parieto-dependent group of premotor areas, each area is the target of a different set of parietal areas. Typically, each premotor area receives strong afferents from one parietal area. Conversely, each parietal area projects predominantly (but not exclusively) to a single premotor area. Thus, the parietal and premotor areas form a series of ana-

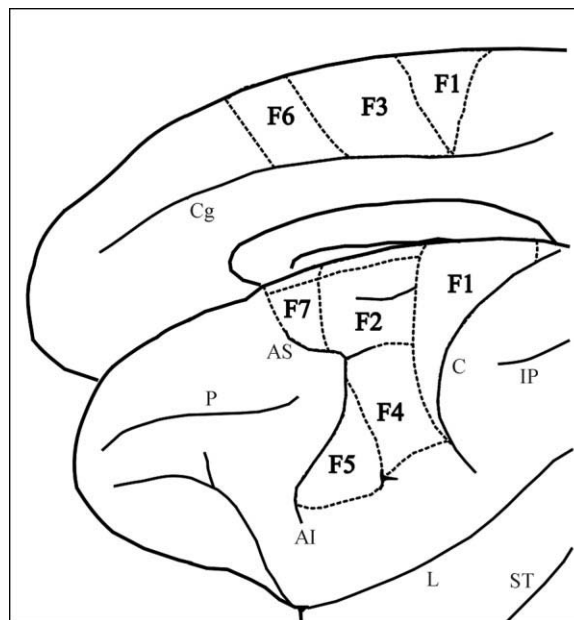


Figure 1

Frontal lobe of the monkey brain. The mesial and the lateral surfaces are shown in the upper and lower part of the figure, respectively. Dotted lines indicate the borders between the various areas forming the agranular frontal cortex (F1–F7). Area F1 roughly corresponds to area 4 of Brodmann. Some authors refer to areas F2 and F7 as caudal and rostral dorsal premotor cortex, respectively. Areas F4 and F5 are also collectively referred to as ventral premotor cortex.

Abbreviations: AI = inferior arcuate sulcus;

AS = superior arcuate sulcus; C = central sulcus;

Cg = cingulate sulcus; IP = intraparietal sulcus;

L = lateral sulcus; P = principal sulcus; ST = superior temporal sulcus

tomical circuits largely independent one from another. The functional correlate of this anatomical organization is that each parieto-premotor circuit is dedicated to a specific sensorimotor transformation (see Rizzolatti et al. 1998). The result of this operation is the generation of a 'potential motor action,' that is of a representation of an action that may or may not be executed.

The fronto-dependent premotor areas (F6 and F7) have only weak connections with the parietal lobe. F6 is mostly connected with the prefrontal area 46 and the rostral cingulate cortex (area 24c). F7 is richly connected with the dorsal sector of the prefrontal cortex. The functions of prefrontal and cingulate areas are much less known than those of the parietal cortex. There is consensus, however, that these areas are not involved in the analysis of sensory stimuli, but rather

play a role in cognitive functions such as working memory, temporal planning of actions, and motivation. It is clear, therefore, that the functions of the fronto-dependent premotor areas ought to be different from those of the parieto-dependent ones. The hypothesis presented here is that the fronto-dependent premotor areas relay information on the individual's motivations, long-term plans, and memory of past action from the prefronto/cingulate cortex to the parieto-dependent premotor areas. On the basis of this information, the potential motor actions, formed in the posterior premotor areas, are either executed or not.

2. The Parieto-dependent Premotor Areas (Fig. 1: F2, F3, F4, and F5)

As a model of information processing of the parieto-dependent areas, the functional properties of one of them (area F5) will be reviewed in some detail (see Rizzolatti et al. 1999). The properties of the other parieto-dependent areas will only be dealt with in short.

2.1 Area F5

Area F5 occupies the rostroventral part of the AFC (Fig.1). Its electrical stimulation evokes hand and mouth movements. F5 neurons typically discharge in association with goal-directed motor actions, such as object grasping, holding, tearing, and manipulating. The neuron firing correlates with specific actions (e.g., grasping) or action segments (e.g., hand opening, hand closure during grasping) rather than with the individual movements forming the action. Most of the F5 neurons code specific types of hand prehensions such as precision grip or whole-hand prehension.

A considerable part of F5 neurons respond to visual stimuli. Visually responsive F5 neurons fall into two classes: 'canonical neurons' and 'mirror neurons.' Canonical neurons discharge to the presentation of 3-D objects. Typically, their response is also present when there is no subsequent action on the presented object. Most canonical neurons are selective for one or a small set of visual objects similar in size and shape. Typically, the visual selectivity is congruent with the motor selectivity. A neuron that is selective for precision grip responds also to the presentation of small, but not of large objects. Mirror neurons do not discharge in response to object presentation but to action upon objects (see Sect. 2.2).

The parietal area that predominantly sends its input to area F5 is the anterior intraparietal area (AIP). Neurons of this area have been subdivided into three main classes: 'motor dominant,' 'visual dominant,'

and 'visual and motor' neurons. 'Motor dominant' neurons are functionally similar to F5 motor neurons. Like F5 neurons, their discharge is associated with actions rather than individual movements. Often they are selective for object size, shape, and/or orientation. 'Visual dominant' neurons are active only when the task is performed in light, but not in dark. They respond also to object presentation in the absence of movements. Many of them are selective for size, shape, and/or orientation of 3-D objects. 'Visual and motor' neurons discharge in both light and dark, but less in the last condition. Like F5 canonical neurons, 'visual and motor' neurons show congruence between their visual and motor properties (Sakata et al. 1995).

From this summary of the AIP neuron characteristics, it is clear that the AIP and F5 have many functional properties in common. There are, however, also some important differences between the two areas. First, in F5 there are no 'visual dominant' neurons. Second, purely motor neurons are much more frequent in F5 than in the AIP. Third, most AIP neurons discharge during the whole action leading to the grasping of objects, being active often also during object holding. By contrast, most F5 neurons are active only during a specific phase of the grasping/holding action, not during the whole action.

On the basis of these findings and inactivation data, it has been proposed that areas AIP and F5 form the key elements of a circuit that transforms visual information on intrinsic (physical) properties of objects into grasping movements (Jeannerod et al. 1995). The following model has been proposed to explain how this transformation occurs.

AIP neurons code visual information on the object's intrinsic features. This information is then sent to specific sets of F5 neurons that transform the received information into potential hand movements appropriate to the size, shape, and orientation of the presented object. In turn, F5 neurons transmit this information to F1 (and the spinal cord) for movement execution. The movement occurs, however, only if there is a concomitant 'go' signal coming from the fronto-dependent premotor areas.

Objects can be grasped in different ways according to the use one wants to make of them. Because different AIP neurons are selective for different object features, F5 receives a multiple descriptions of each seen object from the AIP. The decision on how to grasp that object is determined by the motivations of the acting agent. This decision comes from the fronto/cingulate cortex and reaches the AIP via F5. Thus, a corollary discharge from F5 activates the AIP. In this view, the AIP 'visual and motor' neurons are memory neurons that maintain the representation of the selected object feature active during the entire execution of movement. Finally, the 'motor dominant' neurons would represent an intermediate stage in the transmission of F5 corollary discharge to the AIP 'visual-and-motor' neurons.

2.2 Mirror Neurons

F5 mirror neurons are neurons that discharge when the monkey makes a specific hand action and when it observes another individual making a similar action. The simple presentation of 3-D objects does not activate them. Similarly, the observation of actions that are not directed toward an object, or the observation of actions made using tools, fail to activate them (for a review, see Rizzolatti et al. 1999).

The visual and motor properties of mirror neurons are usually congruent. A neuron that discharges when the monkey grasps an object discharges also when the monkey observes another individual making the same action. In some neurons the way in which the object is precisely grasped does not matter (broad congruence). Other neurons fire only if the observed action is identical to that coded by that neuron (e.g., executed precision grip and observed precision grip; strict congruence).

There is evidence that mirror neurons are mostly located in the caudal part of F5. This sector is anatomically connected with parietal area PF, where, unlike in area AIP, there are neurons with mirror properties. It appears, therefore, that mirror neurons are part of a parieto-premotor circuit separate from that of canonical F5 neurons.

The discharge evoked by the observation of actions is most likely, as in the case of canonical neurons, not a pictorial description of the stimuli, but the motor representation of the action coded by the neuron. However, while canonical neurons use this representation to guide successive motor actions, this is not true for mirror neurons. This, obviously, raises the question of what this motor representation is for. Monkeys, unlike humans, are unable to imitate hand actions. Thus, the imitation hypothesis can be ruled out (at least for hand F5 neurons). The most accepted interpretation of mirror neuron activity is that it is used for action understanding. The assumptions underlying this hypothesis are: (a) individuals understand actions made by other individuals because they are able to react to them internally; and (b) the individuals know the outcome of their actions.

In conclusion, mirror neurons data indicate that the parieto-premotor circuits intervene in higher order cognitive functions and, very interestingly, that the same mechanisms that are at the basis of sensorimotor transformation underlie cognitive functions.

2.3 Areas F2, F3, and F4

Area F2 forms the caudal part of superior area 6 (Fig. 1). F2 receives a rich somatosensory input from area 5. Its ventro-rostral sector also receives visual information from the parietal areas MIP (medial intraparietal area) and V6A (see Rizzolatti et al. 1998). The dorsal part of F2 is related to hind limb movements, whereas the ventral part is related to forelimb

movements. Single neuron studies of the arm field showed that some F2 neurons discharge during the preparation of reaching movements, others during movement execution. F2 neurons code the direction of reaching movement rather than a specific position in space (Weinrich et al. 1984). It is likely that F2 uses information on limb position in order to organize (and/or control) appropriate limb trajectories.

Area F3 (or SMA proper) is located on the mesial cortical surface. Its main input originates from somatosensory parietal areas (mostly mesial area 5). F3 contains a complete representation of body movements, with a prevalence of axial and proximal, arm and leg, movements. Electrical stimulation of F3 frequently evokes complex movements, involving both proximal and distal joints. Single joint movements are rarely evoked (Luppino et al. 1991). It has been proposed that F3 uses the knowledge of actual limb position to prepare the posture necessary for the incoming movements, thus playing a fundamental role in predictive postural adjustments. SMA also appears to contribute to the initial stages of the learning of motor sequences improving their performance (Hikosaka et al. 1999) and, according to others, in the execution of certain motor sequences (Tanji 1994).

Area F4 forms the caudal part of inferior area 6. F4 receives its main input from the ventral intraparietal area (VIP), a parietal area that contains visual and bimodal, visual and tactile, neurons. Electrical stimulation of F4 evokes mostly head and proximal arm movements. One of the most interesting characteristics of F4 is the presence of neurons with tactile and visual receptive fields. The tactile receptive fields are typically large and located on the face, arms, and chest. The visual receptive fields are limited in depth and are located in the space immediately around the animal. They are registered with the tactile fields. The visual receptive fields remain anchored to the tactile fields regardless of the eye and the body part position. These properties show that the F4 visual receptive fields are coded in egocentric coordinates. It has been suggested that F4 plays a crucial role in reaching movements, transforming spatial locations coded in egocentric coordinates into movement directed toward those locations (Graziano et al. 1994, Fogassi et al. 1996).

3. The Fronto-dependent Premotor Areas (F6, F7)

3.1 Area F7

Very little is known on the functional properties of area F7, except for its dorsal sector where eye movements are represented (supplementary eye field, SEF). The SEF is connected with the frontal eye field and other centers controlling eye movements. In contrast to the SEF, the F7 ventral part is linked to the premotor areas controlling body movements. In both

the SEF and ventral F7 many neurons respond to visual stimuli. The precise role of F7 in motor control is unknown.

Interestingly, joint ablation of F7 and (rostral) F2 determines a profound deficit in learning arbitrary stimulus-movements associations. This finding stresses once more that the premotor areas are involved in cognitive, as well as motor functions (Passingham 1993).

3.2 Area F6

Area F6 (or pre-SMA) corresponds to the rostral part of the SMA of Woolsey and Penfield. Motor responses can be evoked from it only with rather high current intensities. Typically, they consist of slow, complex arm movements (Luppino et al. 1991). Visual responses are common in F6, while somatosensory responses are rare. F6 neurons, recorded during active arm movements, show frequently a long leading activity preceding movements (Tanji 1994).

Studies carried out in a naturalistic condition where monkeys interacted with objects presented close to and far from them, showed that F6 neurons exert a global control on arm actions, rather than firing in association to specific, distal (like F5), or proximal (like F4) actions. Some discharge at visual presentation of graspable objects independent of their location, but increase their firing as the stimulus moves toward the monkey. Others, on the contrary, are inhibited at the presentation of the graspable objects, but start to discharge as soon as an object is brought close to the monkey. Summing up: there are several facilitation/inhibition combinations that characterize different F6 neurons. What appears to be constant is the modulation of the neuron discharge according to whether an object can or cannot be grasped (see Rizzolatti et al. 1998).

Recent data showed that F6 plays an important role in the early stages of motor learning, precisely when the cognitive aspect of the procedure to be learned is acquired. This conclusion is based on the behavior of single neurons and on inactivation experiments showing that the number of errors for learning new sequences increases after muscimol injections of this area. In contrast, the performance of learned sequences is not affected by pre-SMA inactivation (Hikosaka et al. 1999).

The data reviewed above suggest that F6 plays a double role in motor control. A first role consists in controlling the posterior premotor areas. Normally, even when the posterior areas are activated, movement does not start. Only when external contingencies and motivational factors allow it, F6 renders the movement onset possible. A second possible role of F6 is more cognitive. It consists in the control of visuomotor associations underlying the declarative phase of motor learning.

4. The Premotor Cortex in Humans

Data from brain imaging studies indicate that the organization pattern of the AFC in humans is similar to that of non-human primates. As in monkeys, there is a multiplicity of premotor areas in humans. Some of them become active during the execution of simple movements (Fink et al. 1997); others require more complex tasks.

The distinction between premotor areas that elaborate sensory information for action (parieto-dependent areas) and areas that control action execution (fronto-dependent areas) appears to hold true also for humans. For example, recent data (Binkofski et al. 1999) showed that in humans there is a specific circuit for hand-object interactions. As in the monkeys, this circuit includes a parietal area—the probable homologue of monkey area AIP—and a ventral premotor area, area 44, the probable homologue of monkey F5. Conversely, data on the pre-SMA (F6) indicate that this premotor area is involved in humans, as it is in the monkey, in complex actions, rather than in the execution of simple movements (Picard and Strick 1996, Hikosaka et al. 1999).

See also: Motor Control; Motor Cortex; Neural Representations of Intended Movement in Motor Cortex

Bibliography

- Binkofski F, Buccino G, Posse S, Seitz R J, Rizzolatti G, Freund H J 1999 A fronto-parietal circuit for object manipulation in man: Evidence from an fMRI-study. *European Journal of Neuroscience* **11**: 3276–86
- Fink G R, Frackowiak R S J, Pietrzyk U, Passingham R E 1997 Multiple nonprimary motor areas in the human cortex. *Journal of Neurophysiology* **77**: 2164–74
- Fogassi L, Gallese V, Fadiga L, Luppino G, Matelli M, Rizzolatti G 1996 Coding of peripersonal space in inferior premotor cortex (area F4). *Journal of Neurophysiology* **76**: 141–57
- Graziano M S A, Yap G S, Gross C G 1994 Coding of visual space by premotor neurons. *Science* **266**: 1054–7
- He S Q, Dum R P, Strick P L 1993 Topographic organization of corticospinal projections from the frontal lobe: Motor areas on the lateral surface of the hemisphere. *Journal of Neuroscience* **13**: 952–80
- Hikosaka O, Sakai K, Nakahara H, Lu X, Miyachi S, Nakamura K, Rand M K 1999 Neural mechanisms for learning of sequential procedures. In: Gazzaniga M S (ed.) *The New Cognitive Neurosciences*. MIT Press, Cambridge, MA, pp. 553–72
- Jeannerod M, Arbib M A, Rizzolatti G, Sakata H 1995 Grasping objects: the cortical mechanisms of visuomotor transformation. *Trends in Neurosciences* **18**: 314–20
- Luppino G, Matelli M, Camarda R, Gallese V, Rizzolatti G 1991 Multiple representations of body movements in mesial area 6 and the adjacent cingulate cortex: an intracortical microstimulation study. *Journal of Comparative Neurology* **311**: 463–82

- Passingham R E 1993 *The Frontal Lobe and Voluntary Action*. Oxford University Press, Oxford, UK
- Picard N, Strick P L 1996 Motor areas of the medial wall: A review of their location and functional activation. *Cerebral Cortex* 6: 342–53
- Porter R, Lemon R 1993 *Corticospinal Function and Voluntary Movement*. Clarendon Press, Oxford, UK
- Rizzolatti G, Fogassi L, Gallese V 1999 Cortical mechanisms subserving object grasping and action recognition: A new view on the cortical motor functions. In: Gazzaniga M S (ed.) *The New Cognitive Neurosciences*. MIT Press, Cambridge, MA, pp. 539–52
- Rizzolatti G, Luppino G, Matelli M 1998 The organization of the cortical motor system: New concepts. *Electroencephalography-and-clinical-neurophysiology* 106: 283–96
- Sakata H, Taira M, Murata A, Mine S 1995 Neural mechanisms of visual guidance of hand action in the parietal cortex of the monkey. *Cerebral Cortex* 5: 429–38
- Tanji J 1994 The supplementary motor area in the cerebral cortex. *Neuroscience Research* 19: 251–68
- Weinrich M, Wise S P, Mauritz K M 1984 A neurophysiological study of the premotor cortex in the rhesus monkey. *Brain* 107: 385–414

G. Rizzolatti and G. Luppino

Prenatal and Infant Development: Overview

Infancy extends from birth through 2 years of age. It has long been regarded as a special period of life in which numerous important transformations of structure take place, and the essential processes of information exchange with the world are laid down. Over this period, the infant is transformed from a condition of utter dependence on the caregiver for survival to a condition of relative independence or life-preserving agency—a transformation accompanied by marked physical, behavioral, and cognitive changes. While no characterization of this transition is entirely satisfactory, several elements set the infant apart from children of other ages. Despite disclaimers that what happens very early in life can be reversed or improved upon, most developmental psychologists celebrate the very considerable and enduring influence of early experience. This article paints a broad picture of infant development, its prenatal antecedents, and its relation to the behavior of older children and adults.

1. Prenatal Development

1.1 Physical and Cognitive Development

The ‘prenatal period’ extends from conception to birth. During the ‘germinal stage’ (fertilization–2 weeks) and ‘embryonic stage’ (2–8 weeks), the fertiliz-

ed egg rapidly divides and attaches to the uterine wall, the major organs and systems of the body differentiate, and interconnections form between the developing components of the central and peripheral nervous systems. During the fetal stage (8 weeks–birth), the organism grows rapidly and the physical features that identify it as human become increasingly apparent.

During the prenatal period, the mechanisms that will facilitate the organism’s adaptation postnatally become functionally mature. Infants born prematurely, for example, who readily learn responses that increase the amount of quiet sleep, reveal this anticipatory maturation. The characteristics of the mother’s voice are also learned during the prenatal period. After birth, hungry infants orient and root in the direction of their mother’s voice, thereby increasing their efficiency of breast-feeding irrespective of whether the mother is holding them on the left or right. Newborns will also learn to activate a tape-recording of their mother’s voice—but not of another woman’s voice—by sucking non-nutritively on a modified nipple. Moreover, ‘prenatal learning’ is highly specific: newborns will suck harder to listen to their mother reading a Dr. Seuss passage that she had read aloud daily during the last 6 weeks of gestation than to hear her reading a novel Dr. Seuss passage (DeCasper and Spence 1986). This specificity continues postnatally throughout the first half year. Similar learning prenatally familiarizes organisms with the linguistic environment into which they eventually will be born.

1.2 The Jacksonian Principle

The development and dissolution of prenatal motor function is described by the ‘Jacksonian principle,’ which states that the last reflexes to develop are the first to disappear when the organism undergoes demise, whereas the first to develop are the last to disappear (Jackson 1884, cited in Taylor 1958). The contralateral (avoiding) reflex, which moves a stimulated site away from a tactile stimulus, is supplanted weeks later by an ipsilateral reflex (Hooker 1952). This sequence is reversed by fetal anoxia. The reversion occurs because reflexes that develop first are well practiced and require less oxygen to be initiated than reflexes that are more recent. Similarly, the order in which sensory functions develop (tactile, vestibular, thermal, chemical, auditory, and visual) is subsequently reversed during injury, disease, or aging.

2. Mileposts of Infant Development

2.1 Zero–6 Months

Birth marks the organism’s transition from a parasitic, waterborne existence in a thermally constant environment with constant access to nutrition to an in-

dependent, land-dwelling existence in a desiccating and cold environment with discontinuous access to food. The occasion of birth is unremarkable, important primarily for triggering the onset of the 'regulatory mechanisms' (respiration, blood flow, digestion) that are critical for this transition. These regulatory arrangements and their interactions are the major constraints on the newborn's behavior and the environmental conditions it can tolerate (Adolph 1968). The evolutionary requirement of all newborn mammals is to grow, which necessitates maximizing energy intake and minimizing energy expenditure. Not surprisingly, therefore, newborns primarily eat and sleep. Early postnatal learning also reflects this necessity: Newborns readily learn where and when food is accessible. Also, because full-term newborns do not physiologically thermoregulate before 4–9 weeks of age, they will not learn energetically costly responses that compete with behavioral thermoregulation.

Over the first 3 months of life, the birthweight of the infant doubles. During this period, brain development is very rapid and influenced appreciably by the kinds of experience to which the infant has been subjected. While most information about dendrite proliferation, the directionality of neuronal connections, cell death, and the effects of deprivation on neural and behavioral development comes from mammals other than humans, the universality of the effects described by neuroanatomists and neuropsychologists is impressive.

In the period following birth, much of the infant's behavior is reflexive, that is, specific stimuli call forth specific reactions (e.g., touches to the mouth elicit rooting and mouthing, abrasions of the skin elicit retraction of the injured body part). This early behavior is principally under the control of the subcortical portions of the brain, and the elicited behaviors are typically swift and obligatory.

Between 2 and 4 months of age, cortical maturation—especially dendritic proliferation—is increasingly evident, and neonatal reflexes disappear or weaken. Also during this period, activity undergoes a major transition from elicited to voluntary. The developmental sequence in which reflexes emerged during the prenatal period is recapitulated during this transitional period, and the sequence in which the postnatal reflexes appear is similarly reversed when the organism is challenged (Humphrey 1969). The postnatal repetition of fetal activity sequences is again due to the lower oxygen requirement of more primitive, hence more practiced, reflexes. Practicing neonatal reflexes reduces their oxygen requirement and speeds the transition.

The transitional process has been implicated in the sudden infant death syndrome (SIDS). McGraw (1943) characterized the period when reflexes are neither reliably elicited nor under voluntary control as a period of 'disorganized behavior.' The transitional period of the respiratory occlusion reflex corresponds

to the time when SIDS is most likely to occur (95 percent of all crib deaths occur between 2–5 months). By 6 months, the occlusion reflex has come under cortically mediated voluntary control, and the infant is once again able to defend successfully against respiratory occlusion (Lipsitt 1979).

Also, over the first 3 months of age the behavioral repertoire of the infant sharply expands. Infants can acquire both classically conditioned and operantly conditioned responses at birth and can remember what they have learned for a surprisingly long time (Lipsitt 1969). An eyeblink reflex that is classically conditioned at 10 days of age, for example, is remembered for at least 10 days. Newborns can habituate to a repeatedly presented stimulus and discriminate it from others that they have not previously encountered. Imitation, a particularly efficient form of learning, can be demonstrated within hours of birth. Not only can newborns imitate an adult's facial gestures immediately after they were modeled, but by 6 weeks of age, infants can imitate facial gestures after a 24-hour delay. Piaget (1962) described early facial imitation as stimulus-bound and reflexive and hypothesized that it disappeared along with many other neonatal reflexes. However, it was later shown not to disappear (e.g., Meltzoff and Moore 1994). Infants' early imitation of facial expressions foreshadows their later dyadic interactions with adults and siblings.

By 3 months of age, infants can acquire arbitrary categories, learn the serial order of an arbitrary list, and pick up incidental information about events in their visual surround—including information about correlated features ('what goes with what'). Although some aspects of visual processing (e.g., accommodation and convergence, and feature search) are adult-like by 3–4 months, others (e.g., visual acuity and conjunction search) do not achieve adult levels for 1 year. Infants can detect depth by 2–3 months but do not fear depth until they begin to crawl. They can also recognize faces of live adults after 24 hours by 6 weeks of age and can recognize them from photos by 5 months.

Observers often remark that infants blossom into 'social creatures' between 2 and 4 months of age, cooing and smiling widely at individuals who interact with them, reaching to other persons, and intensely watching the behavior of others. By their fourth month, infants follow conversations between adults with their eyes, differentiate social from nonsocial objects (which they previously did not), and imitate vocalizations. Not surprisingly, social stimuli are highly effective in establishing and maintaining behavior at this time. The features that define a stimulus as social, however, change with age. Although newborns differentiate between facial expressions and prefer to look at faces than at other visual stimuli, this preference is not based on a social dimension but on contrast and movement—critical features of the human face which influence infants' scanning patterns over the

first 8 weeks. By 4 weeks of age, however, infants display greater pupillary dilation to social than non-social stimuli; by 6 weeks, they display differential cardiac responses to mothers and strangers; and by 12 weeks, they display differential pupillary responses to them and smile more frequently at the mother than at a stranger.

2.2 Six to 12 Months

During the second half-year, a normal infant becomes a walker, a speaker, and a social being who responds emotionally to parents and differentially to familiar persons and strangers. The simultaneous appearance of strong attachment to caregivers and wariness of unfamiliar individuals at 9–10 months of age is one of the most consistent benchmarks of this period (Ainsworth 1979). At the same time, infants begin to look at the target of a pointing finger rather than at the finger *per se* and begin to exhibit social referencing—a strategy by which infants assess the emotional valence of a new situation or person by checking how the caregiver has responded and responding in kind.

For a half-century, another developmental benchmark was attainment at 10 months of object permanence—the concept that an unseen object or person continues to exist. In the classic object search task, infants younger than 10 months consistently failed to retrieve an object they had just watched an adult hide. New paradigms using looking instead of reaching measures, however, have now revealed that even 3-month-olds can locate where an object was hidden.

The types of learning of which infants are capable and their understanding of the physical world rapidly expands over the last half-year. Infants can individuate objects, solve problems requiring causal reasoning, form learning sets ('learning how to learn'), comprehend some spoken speech, repeat a couple of syllables and perhaps produce a word, plan and execute a detour, and track objects through occlusion—a finding that has fueled heated debate over whether infants can count. Before infants begin to crawl (7–9 months), they have already learned what happens in what places. Once infants can locomote, they begin to form a cognitive map or representation of the spatial relationships between these places, which facilitates navigation from one place to another.

2.3 Twelve to 24 Months

Over this period, infants' linguistic, motoric, and cognitive skills continue to improve and broaden. Between 12 and 18 months of age, for example, infants can use landmarks to locate objects instead of using their own body as a reference point, and between 18 and 24 months, they can use geometric cues in the

environment to reorient. Although 18 months was set as the minimum age for deferred imitation of actions on an object (Piaget 1962), the use of tasks particularly suited to younger infants has lowered this age to 6 months. Still, infants' ability to imitate televised demonstrations lags well behind their ability to imitate live demonstrations. The same live demonstration that 6-month-olds can imitate after 24 hours cannot be imitated from television before 18 months, although 15-month-olds can imitate some televised actions after 24 hours if the viewing conditions are more naturalistic. Generally speaking, the pervasiveness of reports that infants' cognitive and motoric performance is superior in natural settings calls into question most if not all minimum age limits that have been imposed by laboratory tests.

Within the last five years, research on the ontogeny of memory has revealed that retention increases linearly over the infancy period, and reminders protract retention even longer. At all ages, a single reactivation or reinstatement reminder, for example, doubles the life of a memory—a remarkable result given that memories of older infants are initially so long. The finding that a few reinstatements can maintain a memory acquired at 6 months through 2 years of age reveals that early experiences can be enduring and raises new questions about infantile amnesia (adults' inability to recollect experiences from early childhood). The two memory systems (declarative, non-declarative) found in older children and normal adults are also functional by 3 months (Rovee-Collier et al. 2001).

3. Risk Factors and Individual Differences

In recent years, advances in the successful delivery and survival of very small and premature babies, often born in jeopardy, have fueled new debate over the importance of early experience. Because the lower threshold for survival is very new, it is inevitable that time must pass before the numerous longitudinal studies, now under way, will yield critical information about the relation between specific prenatal deficits and neonatal hazards and particular developmental outcomes. Such investigations are unlikely to rule out the possibility that some affected infants will be indistinguishable from their age-matched peers in developmental competencies. It is a question of proportions and resiliency.

Although infants born 38 weeks after conception are, on average, less mature than those born at 40 weeks are, the level of physiological maturity can vary among infants born at 40 weeks by as much as 3–4 weeks. Sex differences are also present at birth, with girls being 1–2 weeks ahead of boys in neurological maturation. In normal infants, some of this variability disappears by the end of the first year, but many

individual differences persist, the full range and significance of which are not yet fully appreciated.

Early attempts to predict later intellectual performance from infant behavior focused on sensory and motor abilities and was unsuccessful. Modest-to-good correlations have now been found between measures of information processing beginning at 5–6 months (e.g., habituation rate, novelty preference, 1-week retention ratios) and scores on various intelligence tests at 3–5 years. The Fagan Test of Infant Intelligence (Fagan and Detterman 1992), which correlates later intelligence with infant novelty preference scores, is sensitive to early exposure to environmental toxins, substances of abuse, and nutritional deficits and supplements, as well as to maternal HIV infection, intraventricular hemorrhage, bronchopulmonary dysplasia, genetic anomalies, and various neurological abnormalities.

4. Past Trends and Future Trajectory

The origins of research on infant development can be traced to baby diaries published by Darwin and Preyer in the late nineteenth century. Over the first half of the twentieth century, detailed diaries of their own children's behavior stimulated Morgan, Stern, Guillaume, Valentine, and Piaget to propose theoretical accounts of early cognitive development (Wallace et al. 1994). At the same time, Pavlov's seminal studies spawned longitudinal studies of classical conditioning with infants in Russia and other Eastern-block countries. These early studies provided the first systematic data on infant learning and documented the sequential onset of sensory function from the age at which stimuli from different modalities became effective conditioned stimuli. During this period, Western researchers produced only scattered reports of infant conditioning and retention, emotional development, activity, and color vision.

In 1958, federal funding of a long-term prospective study of infant development ushered in the era of modern infancy research. Using sophisticated apparatus and innovative methods, researchers at 12 sites across the US began to systematically document the infant's basic sensory and learning abilities. These initial inquiries sparked subsequent generations of research on all aspects of perceptual, cognitive, and behavioral development across the entire span of infancy. By the end of the twentieth century, a large amount of developmental data was amassed on infant attention, perception, cognition, motor skills, and the sequelae of many early risk conditions.

The twenty-first century will witness intensive research that elaborates this knowledge and challenges other long-held beliefs about infant development. Also in this century, advances in cognitive and behavioral neuroscience should facilitate attempts to relate de-

velopmental changes in infant cognition to specific changes in the developing brain. This effort will require the development of new, noninvasive techniques and the adaptation of current ones (e.g., functional magnetic resonance imaging, positron emission topography) for use with very young infants. Finally, methodological advances should allow new insights into prenatal development and its contribution to postnatal development.

See also: Cognitive Development in Infancy; Neural Mechanisms; Prefrontal Cortex; Prefrontal Cortex Development and Development of Cognitive Function

Bibliography

- Adolph E F 1968 *Origins of Physiological Regulations*. Academic Press, New York
- Ainsworth M D S 1979 Attachment as related to mother-infant interaction. In: Rosenblatt J S, Hinde R A, Beer C, Busnel M (eds.) *Advances in the Study of Behavior*. Academic Press, New York, Vol. 9
- DeCasper A J, Spence M J 1986 Prenatal maternal speech influences newborns' perception of speech sounds. *Infant Behavior and Development* **9**: 133–50
- Fagan J F III, Detterman D K 1992 The Fagan Test of Infant Intelligence: A technical summary. *Journal of Applied Developmental Psychology* **13**: 173–93
- Hooker D 1952 *The Prenatal Origin of Behavior*. University of Kansas Press, Lawrence, KS
- Humphrey T 1969 Postnatal repetition of human prenatal activity sequences with some suggestions of their neuro-anatomical basis. In: Robinson R J (ed.) *Brain and Early Behaviour: development in the fetus and infant*. Academic Press, London
- Lipsitt L P 1969 Learning capacities of the human infant. In: Robinson R J (ed.) *Brain and Early Behaviour: development in the fetus and infant*. Academic Press, London
- Lipsitt L P 1979 Critical conditions in infancy: A psychological perspective. *American Psychologist* **34**: 973–80
- McGraw M B 1943 *The Neuromuscular Maturation of the Human Infant*. Columbia University Press, New York
- Meltzoff A N, Moore M K 1994 Imitation, memory, and the representation of persons. *Infant Behavior and Development* **17**: 83–99
- Piaget J 1962 *Play, Dreams, and Imitation in Children*. Routledge and Kegan Paul, London
- Rovee-Collier C, Barr R (in press) Infant cognition. In: Wixted J (ed.) *Stevens' Handbook of Experimental Psychology*. Wiley New York, Vol. 4
- Rovee-Collier C K, Hayne H, Colombo M 2001 *The Development of Implicit and Explicit Memory*. Benjamins Publishing, Amsterdam
- Smotherman W P, Robinson S R 1988 *Behavior of the Fetus*. Telford Press, Caldwell, NJ
- Taylor J E (ed.) 1958 *Selected Writings of John Hughlings Jackson*. Hodder and Stoughton, London, Vol. 2
- Wallace D B, Franklin M B, Keegan R T 1994 The observing eye: A century of baby diaries. *Human Development* **37**: 1–29

L. P. Lipsitt and C. Rovee-Collier

Copyright © 2001 Elsevier Science Ltd. 11997
All rights reserved.

Presidency: United States

1. Meaning

The term 'presidency' is applied to several distinct perspectives on the office of the President of the United States. Most commonly, and in its narrowest sense, it is employed as a label to refer to the incumbency of one particular President, as in the presidency of Abraham Lincoln or the presidency of Franklin Roosevelt. More broadly, the term applies to the historical development and legacy of a succession of presidencies since the office was established in 1789. In a more precise constitutional sense, the presidency is that branch of government established by Article II of the Constitution of the United States encompassing a set of constitutional and statutory powers and duties of the President together with additional 'implied' or 'inherent' powers and responsibilities not specifically enumerated in the Constitution or statute law. Finally, the term 'presidency' is used to denote the institutional structure of the office of President embracing the Executive Office of the President—a collection of staff units, originating from proposals contained in the Brownlow Report of 1937, that provide the President with the capacity to manage the work of the executive branch of government and to liaise with the principal constituents of the President, namely the Congress, the media, and interest groups.

These different usages of the term present no significant analytical difficulties. In the literature on the US presidency, the context is usually sufficient to clarify the sense in which the term is being employed and in none of its various forms is it conceptually or theoretically problematic. This is partly because the academic study of the presidency has not been encumbered by rigid theoretical or conceptual frameworks—a matter of concern to some presidency scholars who wish it were so, and a relief to others who see value in the broad, disparate, and eclectic approach that has characterized the study of the presidency to date.

2. Intellectual Context

The academic study of the US presidency has essentially been the preserve of historians and political scientists. Historians have generally confined their attention to individual presidents and presidencies whereas political scientists have traditionally concentrated on the constitutional and political power of the presidency and its institutional structure and development over time. Few historians, with some notable exceptions (see, e.g., Schlesinger 1973, McDonald 1994) have focused on the broader institutional development of the American presidency and only a handful of political scientists (see, e.g., Milkis 1993, Skowronek 1993) have borrowed from the discipline of history to further what political science has had to

say about the presidency. Given the vast amount of scholarship on US presidents and the presidency produced by historians and political scientists, there has been little significant cross-fertilization between the two disciplines and it would not be unfair to say that the study of the US presidency, as distinguished from the study of US presidents, has been dominated by political scientists.

Until the publication in 1960 of Richard Neustadt's seminal work, *Presidential Power*, scholarship on the US presidency had primarily focused on the constitutional and statutory powers of the President, the various roles and functions of the President, and its relationship to the other branches of government as exemplified in the classic texts of Edward S. Corwin (1957), Clinton Rossiter (1956), and Wilfred Binkley (1962). There has also been a strong normative tradition in the literature the presidency, but this has produced various and divergent diagnoses of the weaknesses of the presidency and no uniform prescription for change, although the British system of cabinet government or variants of it has tended to be among the more popular reform proposals.

Neustadt's *Presidential Power* brought the study of the presidency into contact with the behavioral revolution in postwar political science. Neustadt defined power in terms of the president's persuasive or bargaining ability derived from his 'professional reputation' and 'public prestige.' Asserting that 'powers are no guarantee of power,' he proceeded to show that the leadership ability of presidents was not inherent in the office but was highly personal and dependent on the President's own capabilities as a seeker and wielder of effective influence upon others involved in governing the country. Neustadt's immediate impact was to shift the concerns of presidency scholars away from the formal, legal, and institutional boundaries of presidential power toward a more behavioral and theoretical effort to locate the reality of power and leadership in the White House.

While Neustadt's work was a major break with the tradition of presidency scholarship it did not represent a departure from the basic values of the leading presidency scholars who were firm in their belief that the presidency was and ought to be the center of action in American politics and who advocated a strong and dominant presidency as the only institution capable of responding to the social and political problems facing the country. They represented a generation reared during the presidency of Franklin D. Roosevelt and who made FDR, or what was often referred to as 'the New Deal presidency,' the normative model for their scholarship.

3. The Changing Focus of Presidency Scholarship

Notwithstanding Neustadt's impact, the political science of the American presidency lagged behind the behavioral revolution for much of the 1960s and 1970s

and not until the 1980s did the community of presidency scholars begin to exercise the kind of introspection and self-criticism on which behavioralism in America had thrived. That introspection produced a clear but belated appeal for the study of the presidency to connect with the basic tenets of the behavioral revolution—a more rigorous scientific methodology, the need for quantitative measures, and a greater concern for the conceptual and theoretical underpinnings of presidency scholarship (see, e.g., Edwards and Wayne 1983). That appeal did result in a redirection of the focus of presidency scholars and, since the mid-1980s, the political science of the US presidency has taken a quantum leap theoretically, methodologically, and conceptually.

The fundamental values of presidency scholars also changed around the same time. The combined effect of the Vietnam War and Watergate led to a significant reassessment of the New Deal model of presidential power and even those who continued to advocate strong presidential leadership became more sensitive to notions of accountability and restraint. By the end of the century, presidency scholarship had launched an assault on what Mark Peterson called the ‘presidency-centered perspective’ on American government (see Peterson 1990, also Jones 1994), a timely shift in a new era of divided government.

4. Current Theory and Research

The epistemological transformation of presidency scholarship from the mid-1980s could not properly be called a revolution because it did not succeed in displacing older, established approaches entirely. The ‘public law’ tradition, e.g., has flourished during the last two decades partly because of the practical relevance of issues such as the war power, the legislative veto, impoundments, presidential nominations, and impeachment. Scholars following in the footsteps of Corwin (see, e.g., Fisher 1995) have made significant contributions to the study of the presidency that have not been dependent on theoretical sophistication or quantitative methodology. Similarly, the ‘institutional’ school has also been reinvigorated by political developments in recent times. Watergate and the Iran-contra affair, for example, acted as a catalyst for substantial research on the role of the presidential staff and the structure and organization of the Executive Office of the President (see Burke 1992, Hart 1995, Walcott and Hult 1995) and, although much of this work is theoretically and conceptually grounded, it also builds on the descriptive and empirical approach that underlies the ‘institutional’ tradition in presidency scholarship.

The breadth of current research on the US presidency would be difficult to categorize. In addition to those areas mentioned above, it encompasses the study

of presidential psychology and character, leadership style, decision-making processes, popularity, professional reputation, bargaining and persuasion, media management, rhetoric, the impact of public opinion on presidents, the impact of presidents on public policy, and much more.

Current research also embraces a variety of approaches and techniques which, when combined with the range of subject-matter, makes presidency scholarship a broad, diverse, and pluralistic area of study open to the criticism that it lacks coherence in focus and methodology. But it would also be true to say that much current research on the presidency builds upon and extends the seminal research of the past and, while there may not be a single, coherent approach to the political science of the American presidency, there is continuity of research concerns co-existing with a willingness to utilize new methodological and theoretical developments.

Rational choice theorists, however, have staked a claim for the comprehensiveness of their approach and suggested that it may offer presidency scholarship the analytical and theoretical coherence it presently lacks. Applications of rational choice theory to the study of the presidency are still few in number but recent work by rational choice theorists, particularly in the field of presidential–congressional relations, has made a significant impact (see especially Krehbiel 1998, Cameron 2000) and laid the foundations for an extension of that approach to other hitherto less theoretically receptive aspects of the presidency

5. Methodological and Theoretical Problems

Although the cutting-edge of presidency research is in the hands of the rational choice theorists and their close companions, the new institutionalists, their approach has been contested by other presidency scholars who are less convinced of its general utility across the spectrum of presidency scholarship and who question the desirability of imposing a conceptual and methodological uniformity on future research. The debate is exacerbated by the tendency of the new institutionalists to distinguish between the ‘institutional’ and the ‘personal’ presidency, rejecting the latter in the interests of parsimony and theoretical clarity. While it is true that behavioral studies of the ‘personal’ aspects of the presidency, as exemplified by James David Barber’s pioneering work on presidential personality (Barber 1972), revealed how difficult it is to theorize on the basis of a large number of complex variables, that has yet to be accepted as a sufficient argument for ignoring the complexity of the US presidency.

A more longstanding methodological problem for presidency scholars is the lack of access to their subject of study. Presidents of the USA generally do not

give interviews to political scientists, nor do they respond to questionnaires, nor, for the most part, can they be observed at work, neither do their memoirs or papers usually provide the kind of data that presidency scholars require. This is a principal reason why the study of the presidency has lagged behind other branches of the US government in terms of scientific rigor and why the gaps in the state of knowledge about the presidency are still quite substantial.

6. Future Directions

During the last 20 years, research on the presidency has been driven in large measure by the behavior of presidents themselves and the imperatives of the political science profession. In all probability, these two forces are likely to determine the direction of presidency research in the foreseeable future.

Presidency scholars have often been caught unprepared by the behavior of recent presidents and the events surrounding their presidencies. The Vietnam War was a catalyst for substantial research on the war power. The growth and expansion of the presidential staff system was barely studied before Watergate. The weakness of Jimmy Carter as President provided the impetus for Polsby's (1983) landmark study of the reformed presidential nomination process and its impact on the nature of the presidency. Ronald Reagan's public popularity prompted numerous studies of 'going public' strategies of presidential leadership (see especially Kernell 1986). As long as occupants of the White House continue to surprise scholars of the presidency they are likely to influence the range and nature of future research.

Similarly, the professional emphasis on theoretical sophistication and scientific rigor is likely to determine the priorities of future research on the presidency as it has done in the past. The danger is that such imperatives may overwhelm and marginalize those aspects of presidency scholarship where theory and quantitative methods are less applicable and less suitable to the object of study.

See also: Behavioralism; Political; Executive Branch, Government; Leadership: Political; Leadership, Psychology of; Presidential Government; Presidential Personality; Presidential Popularity

Bibliography

- Barber J D 1972 *The Presidential Character: Predicting Performance in the White House*. Prentice-Hall, Englewood Cliffs, NJ
Binkley W E 1962 *President and Congress*. Vintage Books, New York
Burke J P 1992 *The Institutional Presidency*. Johns Hopkins University Press, Baltimore, MD

- Cameron C M 2000 *Veto Bargaining: Presidents and the Politics of Negative Power*. Cambridge University Press, New York
Corwin E S 1957 *The President: Office and Powers 1787–1957*, 4th rev. edn. New York University Press, New York
Edwards G C, Wayne S J (eds.) 1983 *Studying the Presidency*. University of Tennessee Press, Knoxville, TN
Fisher L 1995 *Presidential War Power*. University Press of Kansas, Lawrence, KS
Hart J 1995 *The Presidential Branch: From Washington to Clinton*. Chatham House, Chatham, NJ
Jones C O 1994 *The Presidency in a Separated System*. Brookings Institution, Washington, DC
Kernell S 1986 *Going Public: New Strategies of Presidential Leadership*. CQ Press, Washington, DC
Krehbiel K 1998 *Pivotal Politics: A Theory of US Lawmaking*. Chicago University Press, Chicago
McDonald F 1994 *The American Presidency: An Intellectual History*. University Press of Kansas, Lawrence, KS
Milks S M 1993 *The President and the Parties: The Transformation of the American Party System Since the New Deal*. Oxford University Press, New York
Neustadt R E 1960 *Presidential Power: The Politics of Leadership*. John Wiley, New York
Peterson M A 1990 *Legislating Together: The White House and Capitol Hill from Eisenhower to Reagan*. Harvard University Press, Cambridge, MA
Polsby N W 1983 *Consequences of Party Reform*. Oxford University Press, New York
Rossiter C L 1960 *The American Presidency*, 2nd edn. Harcourt, Brace & World, New York
Schlesinger A M 1973 *The Imperial Presidency*. Houghton Mifflin, Boston
Skowronek S 1993 *The Politics Presidents Make: Leadership from John Adams to George Bush*. Belknap Press, Cambridge, MA
Walcott C E, Hult K M 1995 *Governing the White House: From Hoover Through LBJ*. University Press of Kansas, Lawrence, KS

J. Hart

Presidential Government

Democratic governance takes two basic forms: parliamentarism and presidentialism; some people would add semipresidentialism, leaving aside the very unique Swiss institutional structure. Each of them presents a variety of forms that lead to question the dichotomy. There are a few defining characteristics from which many consequences can be derived. Each is identified with a paradigmatic case: the British Westminster parliamentarism, US presidentialism and the semipresidentialism of France. However, as Sartori (1994) notes, parliamentarism takes a variety of forms like the German Federal Republic—the Kanzler democracy—and the assembly parliamentarism of the Third and Fourth French Republics, and contemporary Italy.

Presidential systems are based on an executive with considerable powers in the constitution, and generally with full control of the composition of the cabinet and the administration, elected by the people (directly or by an electoral college elected for that purpose) for a fixed period of time, and are not dependent on a formal vote of confidence by the democratically elected representatives in a parliament. The president is not only the holder of executive power, but is also the symbolic head of state who cannot be dismissed between elections except in rare cases of impeachment.

Two features stand out in presidential systems:

(a) Both the president, who controls the executive and is elected by the people (or an electoral college elected by the people for that sole purpose), and an elected legislature (unicameral or bicameral) enjoy democratic legitimacy. It is a system of 'dual democratic legitimacy.'

(b) Both the president and congress are elected for a fixed term, the president's tenure in office is independent of the legislature, and the survival of the legislature is independent of the president. This leads to what we characterize as the 'rigidity' of the presidential system.

Most of the characteristics and problems of presidential systems flow from these two essential features. Some other nondefining features of presidentialism are often associated with it, such as term limits or no re-election, automatic succession by a vice-president, more or less extensive rule making, and emergency powers. One characteristic, so normal that it can be included in the definition, is that the presidency is a unipersonal office. There have only been two cases of directly elected pluripersonal 'presidencies': the two-person Cypriot administration (1960–1963) and the Uruguayan Colegiado (which governed twice — 1918–1933 and 1952–1967).

The basic characteristic of the president is the full claim to democratic legitimacy. The claim very often has strong plebiscitary components, although sometimes it is based on fewer popular votes than are received by many prime ministers in parliamentary systems heading minority Cabinets. The presidency gives the incumbent, who combines the qualities of head of state representing the nation and the powers of the executive, a very different aura and self-image, and creates very different popular expectations than those redounding to a prime minister with whatever popularity he might enjoy after his party receiving the same number of votes.

The most striking fact is that in a presidential system, the legislators, particularly when they represent well-organized, disciplined parties that constitute real, ideological, and political choices for the voters, also enjoy a democratic legitimacy, and it is possible that the majority of a legislature might represent a different political choice from that of the voters supporting a president. Who, on the basis of democratic principles, is better legitimated to speak

the name of the people: the president, or the congressional majority that opposes his policies? Since both derive their powers from the vote of the people, a conflict is always latent and sometimes erupts dramatically; there is no democratic principle to resolve it, and the mechanisms that might exist in the constitution are generally complex, highly technical, legalistic, and therefore of doubtful democratic legitimacy for the electorate. It is no accident that in some of those situations the military intervenes as 'poder moderador.'

It could be argued that such conflicts are normal in the USA and have not led to serious crisis. It would exceed the limits of this essay to explain the uniqueness of American political institutions and practices that have limited the impact of such conflicts, including the unique characteristics of the American political parties that lead many American political scientists to ask for a more responsible, disciplined ideological party system (Riggs 1998).

In a unipersonal presidency there is only one winner who gains full control of the executive. It has been argued that in a Westminster parliamentary system a party winning an absolute majority is also (or even more) a 'winner take all' while in a presidential system it is less likely that a president will have an absolute majority in congress for the entire period of his mandate. However, it is a president, one person that wins control of the executive rather than a party, and that person gains that power irrespective of the size of his plurality or in a runoff in a narrowed competition in which necessarily one candidate will have a majority. The loser in a presidential election ends with no public office in contrast to the leader of a defeated party in a parliamentary election likely to hold a seat in Parliament. Part of the unipersonal election is, in many presidential systems, the succession by a vice-president chosen by the nominee using criteria that would not have been those of the electorate or a party leadership.

The plurality rule remains the most common method for electing presidents. However, the fear of a president being elected by a small plurality has led to introduce a majority runoff election among the two leading candidates (when no candidate has received a majority). The plurality rule (Shugart and Carey 1992) favors a two-party competition or at least the formation of broad pre-election coalitions and discourages candidates from minor parties. The runoff election, on the other hand, encourages a larger number of candidates in the first run, either in the hope to make it to the runoff (even with a modest number of votes) or place themselves in the condition to bargain for their support by one or the other of the two front runners.

Each procedure can have negative consequences. The plurality rule in a fragmented party system can lead to the election with the support of a small minority. The runoff in addition to encouraging party

fragmentation in the first run, can create false coalitions (which will not hold after the election), or mortgages the candidate with allies which may exact a price or distort his image. It also generates in the winner the feeling that he represents the majority of the people, though in the first run he had weak support and his opponent might have been ahead.

In a multiethnic society the fear that a candidate from a populous state could be elected against those from other ethnic-religious communities, has led to a complex system of concurrent pluralities, to assure the election of a president with appeal across ethnic groups (as in Nigeria).

There are a few instances in which the popular vote, in the absence of a decisive majority, does not determine the winner. In Chile, before 1973, the Congress chose the president in the event that no candidate received a majority of popular votes among the top two contenders. Traditionally the one with the largest plurality was elected. In the USA, when in the nondeliberative electoral college no candidate obtains the vote of the majority of electors at the first count, the decision goes to the House of Representatives.

The fixed term of office of both presidents and congress introduces a particular rigidity. A failed president cannot be (normally) forced to relinquish power, but is obliged to continue governing, and he cannot (constitutionally) dismiss a congress committed to oppose his policies. A failed parliamentary prime minister can lose a vote of confidence, be replaced by another leader of his party or coalition, and if he feels that the opposition is wrong, call for an election hoping to discipline his party or coalition and/or revalidate his authority. The rigidity and the length of some presidential mandates leads to the frequent rule of no-reelection, or of no continuous reelections.

The dual legitimacy resulting from separate election of president and the legislature and their mutual independence can lead to 'divided government.' The congress and the presidency may be controlled by different parties (or factions of the same party) creating an impasse. How to govern?

Ackerman (2000) has developed three scenarios. The first emphasizes accommodation, based on bargaining, they may be called the Madisonian hope. Fiorina (1996) argues that American voters desire a divided government because they are uncertain about endorsing an ideological direction offered by either party. Mayhew (1991) argues that divided government has not impeded major legislation. The second involves constitutional breakdown, in which one or another power assaults the system and installs itself as a single lawmaker. A situation of this type is in Latin America, where traditionally the military as 'poder moderador' intervened temporarily, or more recently assuming complete power. Referring to my analysis (Linz 1994), Ackerman calls this breakdown the 'Linzian nightmare.' A third less dramatic alternative

is backbiting, mutual recrimination, and partisan deadlock, which he calls 'crisis of governability.' To the use of constitutional or extraconstitutional decree powers, emergency legislation, patterns described in detail by Carey and Shugart (1998) and their collaborators in a chapter significantly entitled 'Calling Out the Tanks or Filling Out the Forms,' I would add a fourth alternative that explains the absence of recorded conflicts: the application of the 'rule of anticipated reactions.' The avoidance of conflict by not initiating policies is likely to lead to conflict even when desirable, necessary, and proposed by the president at the time of his election.

The disjunction between the image of the voters and the president himself as an initiator of major policies, reforms, and change, and the gridlock, mutual blaming, and/or of a frustrated president or congress confronted with *decretazos*, an avalanche of decrees as *faits accomplis*, reducing the legislature to impotence, does not contribute to political efficacy and the legitimacy of either or both institutions. The breakdown of democracy and the turn to unconstitutional rule or authoritarianism is only one possible outcome. 'Formal' democratic continuity, particularly in recent crises in Latin America (or the ex-USSR countries) is no proof that the system works.

The situation that Ackerman calls of full authority, when the presidency, the House and Senate are controlled by the same party and there is a favorable majority in the Supreme Court is rare and may last a short time. The result is a burst of legislation, law-making initiatives that will be strong in symbolic statement and difficult to change afterwards, with less concern about implementation. In a parliamentary system, where laws can more easily be changed with electoral defeat, governments are more interested in statutes framed in mid-sized concepts keyed to operational realities, to make them work.

Although the separation of executive and the legislative power, more accurately separate institutions sharing powers, implies that the bureaucracy (and the military) are the domain of the presidency, the fact is that they have two masters. Congress is perhaps increasingly adept to intervene in details of public administration and defense thanks to the prerogative of 'oversight,' of financing departments and programs, the long-standing freedom to investigate. A result is the dependence on both the presidency and the congress and indirectly the separation from both, and subordination to neither. A response to this situation has been 'politicized professionalism': the growing number of political appointees, the large number of in-and-outers, and the short tenure of those appointed who prefer to move to the private sector (and consequently the lack of accumulated experience and continuity in policies). The president turns to 'loyalists' rather than experienced civil servants who would be concerned about their relations with powerful congressmen or women.

In the ongoing debate over the success or failure of presidentialism it is important to distinguish:

(a) The basic theoretical issues derived from the defining characteristics of presidentialism (Linz 1994, Shugart and Carey 1992).

(b) The problems for presidentialism not derived from the basic characteristics; like no reelection, tensions with vice-presidents, mid-term congressional elections, etc.

(c) The problems with presidentialism and the actual breakdowns of presidential compared to parliamentary systems. A comparison plagued by a number of methodological problems; time period covered, regional comparisons, level of economic development (OECD vs. developing countries), and increasingly formal stability with 'defective' democracies (some actually authoritarian regimes) (Stepan and Skach 1994, Shugart and Carey 1992, Thibaut 1996).

(d) The absence of a possibility of a transition to parliamentarism and therefore the limited relevance of a theoretical comparison and the need to consider 'improvements,' modifications of presidentialism that correct some features contributing to some of its failures (Nohlen and Fernandez 1991).

(e) The need to keep the problems of semi-presidentialism, offered as a more viable alternative to presidentialism, separate. In this case, the lack of consensus on which countries are semipresidential and the small number of cases (at least before the new postcommunist regimes), makes an analysis like (c) difficult.

The basic argument for the separation of powers is that it provides mutual checks, enhancing freedom, preventing the concentration of power enjoyed by a British prime minister. Mutual checks, the greater difficulty in enacting policies, the opportunity for all interests to be heard and taken into account, the slowing down of the process, and the opportunity for vetoes, generate weak executives and block many of their initiatives, and this is presumably what the voters want when they split their vote for congress and the president. Perhaps this is an advantage in a society that works well independently of public policies, in a wealthy market economy with no major need for redistributive policies, or with independent nonpolitical agencies, but is unlikely to fit societies facing serious problems of underdevelopment and social injustice, requiring more effective government action.

Among the positive aspects of presidentialism its advocates list:

The separation of powers of congress and the executive allows the 'representatives' to act as representatives of interests, particularly of those with a geographical basis, since they are not bound to support the policies of the executive and no party discipline is needed to support a government. In the absence of proportional representation, they enjoy considerable independence from the party leadership and are more likely to represent interests rather than broad policy or

ideological alternatives. All politics become local. As Sartori (1994, p. 84) writes: three factors 'ideological unprincipledness, weak and undisciplined parties and locality centered politics, contribute to the working of US presidentialism.'

The separate election, particularly mid-term elections, and the limited influence of the president in many cases in the nomination or the making of lists of candidates (especially in federal systems) and in their electoral success, makes the representatives independent. Even their own partisans will owe them almost nothing and often doubt their usefulness in the elections. The selection by primaries has probably reinforced that tendency. This makes members of congress, compared to parliamentarians, better representatives of interests of their constituents but less reliable in supporting a party (and presidential) policy program. In fact, that 'representative' function is seen (Shugart and Carey 1992) as one of the positive aspects of presidentialism.

One presumed advantage is the identifiability of presidents, compared to the incertitude of who will head the government in some parliamentary systems based on coalition-making among multiple parties after the election. However, in many parliamentary systems the voter knows well who will head the government if the party he votes for wins. Although presidential candidates identify themselves with certain policy goals, they are less bound to a program than a party and the prime minister it supports, or minor parties that make their coalition support dependent on certain demands (for example, national minority parties).

Shugart and Carey (1992, p. 44-45) consider the personal identifiability the basis for greater accountability. This might be true, except when reelection is barred and for the blaming game between the president and congress (particularly in the case of divided government).

Presidential elections allow the citizen to choose one individual to put his trust in—on account of real or perceived qualities and commitments without the mediation of complex institutions: a party, a legislature, or a plurality of actors like parties forming a coalition. That personalization simplifies the choice. It also permits outsiders, not identified with a party, who may not have held any office before, to appeal personally and directly to the voter: Collor, Fujimori, Tyminski, Perrot. That appeal requires access to the media, which in turn may involve access to funding—in some cases personal wealth—and considerable misunderstanding of what is necessary to govern. It may also appeal to 'being above parties,' perceived as divisive, somewhat like a monarch. The election of military leaders: Mannerheim, Hindenburg, De Gaulle, Eanes, even Eisenhower, reflects this desire. TV is particularly congruent with these patterns.

Latin American presidents greatly vary in terms of their constitutional powers, both reactive (veto power,

for example) and proactive (decree legislation) and in terms of their partisan support in congress. The possible combinations (Mainwaring and Shugart 1997, pp. 40–54) are too complex to summarize here.

The difficulties with presidentialism in Latin America have generated two opposite responses: attempts to curtail the powers of the president out of fear of the abuse of power and authoritarian predispositions and efforts to strengthen the presidency as an answer to the weakness of the presidency, particularly in relation to congress.

Many presidential systems have a vice-president, with limited functions except when he has to step in to succeed the president for reasons of death, resignation, impeachment, or incapacity. He generally is not elected independently but on a ticket with the presidential candidate, chosen by him, often for electoral or personal reasons without concern about his or her qualifications. The frequent destructive conflicts between presidents and vice-presidents have led to recent constitutions to not include that office.

1. Semipresidentialism

Semipresidentialism as a distinctive form of organizing democratic political life, gained importance as a constitutional model and political reality with the consolidation of the French Fifth Republic defined by Sartori (1994, p. 132):

(a) The head of state (President) is elected by popular vote—either directly or indirectly—for a fixed term of office.

(b) The head of state shares the executive power with a Prime Minister, thus entering a dual authority structure whose three defining criteria are (c) to (e).

(c) The President is independent from parliament, but is not entitled to govern alone or directly and therefore his will must be conveyed and processed via his government.

(d) Conversely, the Prime Minister and his Cabinet are president-independent in that they are parliament-dependent: they are subject to either parliamentary confidence or no-confidence (or both), and in either case need the support of a parliamentary majority.

(e) The dual authority structure of semipresidentialism allows for different balances and also for shifting prevalences of power within the executive, under the strict condition that the ‘autonomy potential’ of each component unit of the executive does subsist.

Shugart and Carey (1992) prefer to call this system ‘premier presidentialism’ and distinguish it from ‘presidential parliamentary’ (which others include under semipresidentialism) on the basis of the constitutional position of the president.

Since its inception in the constitution of the Weimar Republic (1918) and the Finnish Constitution, it has been the object of considerable dispute. Different

terms were used to characterize it: bipolar executive, parliamentary presidential, and Duverger established the term semipresidentialism. The ambiguities surrounding it lead some authors to treat it as a variant of presidentialism, while others see it as a form of parliamentarism. This author, as Sartori, considers it a form *sui generis* distinct from both presidentialism and parliamentarism.

Bartolini (1984) has shown that semipresidentialism was introduced in countries that achieved independence and sought a symbol for the new nation (Ireland, Iceland, and earlier in Finland). In Portugal, it was an attempt to overcome the uncertainties of a transition to democracy.

The French semipresidentialism of the Fifth Republic after the introduction in 1962 of direct presidential election is a two-headed system. It makes possible the alternation—according to some authors—between presidential and parliamentary phases depending on the party or parties of the president having a majority in the chamber. Sartori, rightly, prefers to speak of ‘oscillation,’ since, in the case of split majorities—described by the French as ‘cohabitation’—the president retains some specific powers, particularly in foreign affairs and is not reduced to the status of a president head of state in a parliamentary republic. Nor can a president cum government, that has to govern with and through another body, be assimilated to a ‘pure’ president that governs alone, that is the government. The oscillation assures considerable flexibility and until now ‘cohabitation’ has worked relatively well.

Elgin (1999, p. 14) lists countries that he considers semipresidential: 6 in Western Europe: France, Finland, Portugal, Austria, Iceland, and Ireland; 5 in Central and Eastern Europe: Poland, Bulgaria, Romania, Croatia, Macedonia, and Slovenia; 11 in the former USSR: Russia, Ukraine, Lithuania, Moldavia, and Belarus (of them in the Asian belt: Armenia, Azerbaijan, Kazakhstan, Kyrgyzstan, and Uzbekistan; 5 in Asia and the Middle East: Lebanon, Maldives, Mongolia, South Korea, and Sri Lanka; 3 in America: Dominican Republic, Guyana, and Haiti; and 11 in Africa. A total of 42 countries, among which only 18 in 1999 were rated from 1 to 3 on the Freedom House index of political rights; 6 in Western Europe, 4 in Central and Eastern Europe, 1 (Lithuania) in the ex-USSR; 2 in Asia; 1 in the Americas; and 4 in Africa.

The conclusion about regimes that from a formal constitutional perspective and definition can be classified as semipresidential is that the material constitution (the actual working over time) varies enormously. In some, the president, although popularly elected and with considerable legal powers, acts more like a president in parliamentary regimes. One difficulty is that a large number of countries that have adopted semipresidential institutions are new democracies, and are politically unsettled (and doubtful democracies), and it is premature to characterize those

politics. Moreover, much depends on the personality and political goals of the founding presidents and their successors.

Though all conceptual classification of forms of government cover a wide range of variations, this is even truer about semipresidential given the ambiguities in the constitutions about the powers of presidents and prime ministers, the role of leadership and personality, and the fact that the extended list of such regimes includes many countries dubiously democratic.

On the basis of the long-time actual working of formally semipresidential regimes, Linz (1994) has excluded Austria, Ireland, and Iceland from the list and he and Sartori (1997) would also exclude Portugal after 1982.

The dual legitimacy resulting from the separate direct election of president and parliament persists in semipresidential regimes, but the principle that a prime minister requires the confidence of a majority in the parliament reduces—perhaps even eliminates—the conflicts derived from ‘divided government.’ The French constitutional practice in the situations of ‘cohabitation’ supports that interpretation, though one might still ask to what extent it reflects a long tradition of parliamentarism (in addition to other favorable conditions).

We might add to the complexity (and perhaps confusion) a distinction between formally semipresidential regimes in which the president exercises significant powers (and is expected to do so) and those where he has not done so (and is not expected to do so). We obviate the argument that the formal institutional structure is not irrelevant, and that in fact it might at some point become politically relevant. The formally but not materially semipresidential regimes on that account are not parliamentary though they may work as parliamentary. More complicated is the issue—not considered in the earlier studies of this regime type—of the president enjoying such formal and actual powers that the prime minister and his cabinet are fully dependent on the president and the parliament, though formally empowered to question that dominance, is unable to do so. Should such regimes be considered presidential rather than semipresidential? The question can be raised about Russia and Sri Lanka. Again, the complex constitutional rules cannot be ignored but so can the actual working of the system under Yeltsin not be fitted into a ‘semipresidential mode.’ With such a narrowing down we still would have a number of working stable pure semipresidential regimes.

Presidents generally are identified with a political party, a leader of a party, or after winning the nomination of a party become its leader (yet when defeated lose that position). However, in Russia, Yeltsin, though elected with the support of the parties, opted for not identifying with a party and the constitution formally forbids the president from being

a party member. The same is true for other post-communist presidential or semipresidential constitutions.

The idea of president above party, not identified with a party and not campaigning for a party, is likely to make it more difficult for a president to have party support, and does not contribute to the building of strong parties.

Semipresidentialism can work well when the presidential majority and the majority in parliament are of the same party or coalition and when, without being the same, there is a majority in the chamber, supporting a prime minister of a different party (the French ‘cohabitation’). The system becomes more problematic as Kaltefleiter (1970) has analyzed in detail when the parliament is unwilling or unable to provide support to the choice of prime minister or of the president but is also unable to support someone of its own choice. This can be the case of a highly fractioned parliament or of a polarized pluralism in which the extremes make it impossible for the center to govern. Parties might be ready to tolerate weak presidential cabinets (abdicating their responsibility) and the president governing by decree. The president, on the other hand, having the power of dissolution, may be tempted to call for elections to solve the impasse and construct a majority. Repeated elections in a climate of crisis might deepen the crisis and contribute to strengthen the antisystem disloyal or semiloyal oppositions. These patterns developed in Germany under the presidency of Hindenburg and his presidential cabinets, contributed to the breakdown of the Weimar Republic.

See also: Electoral Systems; Executive Branch, Government; Leadership: Political; Legitimacy: Political; Parliamentary Government; Parliaments, History of; Party Systems; Presidency: United States; Presidential Personality; Presidential Popularity

Bibliography

- Ackerman B 2000 The new separation of powers. *Harvard Law Review* 113(3): 634–729
- Carey J M, Shugart M S (eds.) 1998 *Executive Decree Authority*. Cambridge University press, Cambridge, UK
- Duverger M 1978 *Echec on Rev.* Albin Michel, Paris
- Duverger M 1980 A new political system model: Semi-presidential government. *European Journal of Political Research* 8(2): 165–87
- Elgie R (ed.) 1999 *Semi-presidentialism in Europe*. Oxford University Press, Oxford, UK
- Fiorina M 1996 *Divided Government*. Allyn and Bacon, Boston, MA
- Kaltefleiter W 1970 *Die Funktionen des Staatsoberhauptes in der parlamentarischen Demokratie*. Westdeutscher Verlag, Köln
- Linz J, Stepan A (eds.) 1978 *The Breakdown of Democratic Regimes*. Johns Hopkins University Press, Baltimore
- Mainwaring S 1990 Presidentialism in Latin America. *Latin American Research Review* 25: 157–79

- Mainwaring S 1993 Presidentialism, multipartism, and democracy: The difficult combination. *Comparative Political Studies* 26(2): 198–228
- Mainwaring S, Shugart M S (eds.) 1997 *Presidentialism and Democracy in Latin America*. Cambridge University Press, Cambridge, UK
- Mayhew D R 1991 *Divided We Govern: Party Control, Law-making, and Investigations, 1946–1990*. Yale University Press, New Haven, CT
- Sartori G 1994 *Comparative Constitutional Engineering: An Inquiry into Structures, Incentives, and Outcomes*. New York University Press, New York
- Serraforo M D 1999 El poder y su sombra. los vice presidentes. Editorial de Belgrono, Buenos Aires
- Shugart M S, Carey J M 1992 *Presidents and Assemblies: Constitutional Design and Electoral Dynamics*. Cambridge University Press, New York
- Stepan A C, Skach C 1994 Presidentialism and parliamentarism in comparative perspective. In: Linz, Valenzuela, p. 36
- Thibaut B 1996 *Präsidentalismus und Demokratie in Lateinamerika*. Leske and Budrich, Opladen, Germany

J. J. Linz

Presidential Personality

The structure of the American constitutional system and the realities of American politics combine to make it a virtual certainty that the personal qualities of the chief executive will have significant consequences for the nation and the world. The Constitution accords the president coequal status with the legislative and judicial branches of government. Barring impeachment by the House of Representatives and conviction by the Senate, he (and at some possible point she) is entitled to remain in office until the end of his term. Because the occupant of the Oval Office has power to veto legislation and is his nation's chief diplomatic officer, the commander and chief of its armed forces, and the administrator of its laws, he is a force to contend with even when his standing with the public and Congress is at low ebb.

One could imagine a nation in which the highest political leader had great formal powers but was of little personal consequence, because that individual was merely the agent of a powerful political party or some other extra-governmental force. That is emphatically not the case in the USA, where the political parties are highly decentralized and there are many centers of power. American presidents have placed their stamp on public policy since the founding of the Republic, but for much of American history the president tended to defer to Congress for the initiation of policy, the federal government played only a modest part in the lives of citizens, and the nation was inactive in global affairs. Then came the Great Depression,

World War II, and the incisive response to these challenges of Franklin D. Roosevelt. During the three terms and three months of the Roosevelt presidency, the USA became a world power and a nascent welfare state, and the presidency itself underwent a metamorphosis. The president replaced Congress as the principal source of legislative proposals; presidents began to make an increasing amount of public policy by executive action; and the Executive Office of the President was created, providing the chief executive with the organizational support needed to carry out his obligations.

The qualities that distinguish one chief executive from another have their greatest potential impact in the realm of national security, where the president is custodian of a vast nuclear arsenal. No president other than Harry S. Truman has employed nuclear weapons, but there have been many presidentially-initiated conventional military commitments, the largest of which have been in Korea in the 1950s, Vietnam in the 1960s and 1970s, and the Persian Gulf in 1991. The president also is a major domestic force, not only because of his formal powers, but also because his great political visibility places him in a unique position to seek public support for his policies.

The two most influential efforts to identify the personal qualities that shape presidential performance are Neustadt's *Presidential Power* (1960), and Barber's *The Presidential Character* (1972). Neustadt's book, which has been described as the American equivalent of Machiavelli's *The Prince*, is an exercise in political psychology with the emphasis on the political. Taking it as his premise that the American political order places severe constraints on the chief executive, Neustadt argues that a president can best overcome the constraints on him by using his formal powers as levers for bargaining and persuasion and making it evident to the other members of the policy-making community that he has the support of the public and is a skilled, determined political operator. From this it follows that an effective president should be well grounded in the ways of American politics, sensitive to power relations, and self-assured about the merit of his policies and his ability to direct his nation.

Barber's book is a study of political psychology with the stress on the psychological. Barber identifies five broad determinants of a president's White House performance. Two relate to the political context—the contemporary *power situation* (e.g., the partisan balance in Congress) and *climate of expectations* (e.g., the state of public opinion). The other three bear on the president's inner characteristics—his *political style* (habitual way of carrying out his responsibilities), *world view* (political beliefs and assumptions), and *character* (the deeper layers of his psyche). In principle, this is an excellent framework within which to analyze presidential personality, in that it reminds the analyst that a president's actions will be a joint function of his personal predispositions and his political context

(Greenstein 1969). In practice, however, Barber's focus is virtually exclusively on character, which he classifies in terms of whether a president is active or passive in the conduct of his responsibilities and positive or negative in the feelings he invests in politics.

In the resulting four-part classification, the two passive types parallel David Riesman's distinction between other- and inner-directed personalities. (Reisman 1950). These are the *passive-positive* presidents, who take their clues from others, and the *passive-negative* presidents, whose motivations for public service are grounded dictates of their consciences. Neither type, Barber argues, is well suited for the activist demands of the modern presidency. Barber's two active types echo another important distinction. The *active-positive* presidents are marked by what Lasswell (1951) refers to as the democratic character structure, which is to say that they are free of emotional perturbations that impede their leadership. The *active-negative* presidents are of a piece with the class of politicians Lasswell identifies as emotionally troubled power seekers—those for whom politics is a means of venting inner insecurities. As Barber puts it, the active-negative presidents channel great energy into their political actions, 'but it is an energy distorted from within' (Barber 1972, p. 9). Examples of such distortions include Woodrow Wilson's stubborn refusal to make the modest compromises that would have permitted the ratification of the Versailles Treaty and Lyndon Johnson's persistent escalation of the war in Vietnam, despite mounting protest at home and lack of success in the field.

Barber performs a valuable service by directing attention to the deeper strata of presidential personality, but his assumptions about human motivation are overly simple. Indeed, the very dichotomies that determine Barber's typology fail to account for important complexities. Rather than being either positive or negative in their feelings about politics, presidents may be subject to mood swings, as was the case of Theodore Roosevelt and Lyndon Johnson. Rather than being either active or passive, they may be active in some spheres and passive in others. Thus, Ronald Reagan was a tireless public communicator but had a hands-off approach to the internal workings of his presidency.

Barber's taxonomy is an example of what A. L. George and J. L. George (1998) characterize as a 'file drawer' approach to personality analysis. As an alternative, they propose elucidating presidents and other political figures in all of their complexity, drawing on existing clinical typologies as sources of hypotheses about a subject's inner dynamics. Theirs is the strategy taken in the most persuasive presidential psychobiographies, including their own sharply etched portrait of Woodrow Wilson (A. L. George and J. L. George 1956), Glad's (1980) subtle account of Jimmy Carter, and Renshon's richly psychoanalytic interpretation of Clinton (Renshon 1996).

A related track can be found in Greenstein's *The Presidential Difference* (2000), in which each of the eleven presidents from FDR to Clinton is characterized on his own terms but also compared with the others on the basis of six criteria: *public communication ability, organizational capacity, political skill, policy vision, cognitive style, and emotional intelligence*. A number of instructive patterns emerge. Only three of those presidents prove to have been consistently outstanding public communicators, Roosevelt, John F. Kennedy, and Reagan. In each case the president's effectiveness was much enhanced by his rhetorical prowess. Organizational capacity proves to have been even less common than eloquence, but it is at least as important. The most dramatic example of what can result in its absence is the chaotically organized Kennedy administration's failed effort to land anti-Castro insurgents at Cuba's Bay of Pigs.

The presidency of Jimmy Carter illustrates the critical importance of political skill for presidential effectiveness. Carter displayed impressive political ability in his rise to the White House, but upon becoming president he resisted the standard reciprocities of Washington politics, with negative results for the success of his program. The presidency of Lyndon Baines Johnson offers a complementary illustration. Between November 1963 and the end of 1965, Johnson displayed prodigious political skills, presiding over a tidal wave of legislative enactments. But in 1965 Johnson also embarked on an open-ended military intervention in Vietnam, using his skill to minimize public debate and controversy. What Johnson lacked was a realistic policy vision, at least in international affairs. He failed to establish the precise goals of the intervention, its probable duration, and the likely troop requirements. By 1968 there were a half million American troops in Vietnam at which point Johnson halted further troop increases, entered into negotiations with the communists, and withdrew himself from contention for a second elected term.

The cognitive and emotional qualities of presidents also are central of their leadership. One way presidents differ in their cognitive styles is in their capacity for abstract thinking. Jimmy Carter, for example, had an impressive ability to master specifics, but was unable to embed them in a larger rationale. Richard Nixon, by way of contrast, had an outstanding capacity to envisage broad patterns and possibilities. Before taking office, Nixon established three major international policy goals: extricating the nation from the fighting in Vietnam, arriving at an accommodation with the Soviet Union, and establishing relations with the People's Republic of China. By the end of his first term he had accomplished them all. But during the same period Nixon initiated the program of domestic political espionage and sabotage which led to the destruction of his presidency.

Nixon was one of several modern presidents who had outstanding cognitive strengths, but lacked what

has come to be called emotional intelligence (Goleman 1995), a term that captures the ability to manage one's emotions constructively, rather than being under their domination. Others include Johnson, Carter, and Clinton. Nixon and Johnson presided over Vietnam and Watergate—two of the unhappy episodes in American history. The presidencies of Carter and Clinton were not marked by comparably dire events, but each of them suffered from the character flaws of the chief executive.

The American political system is said to be one of institutions rather than individuals. But it was not the institution of the presidency that authorized the use of the atom bomb against Japan or led the USA into Vietnam. It was Harry S. Truman and Lyndon B. Johnson. American history is replete with episodes in which presidents took actions that might not have been taken by another White House incumbent, some of them of momentous significance. It follows that the qualities that distinguish one president from another demand close and continuing attention.

See also: Clinical versus Actuarial Prediction; Executive Branch, Government; Leadership: Political; Leadership, Psychology of; Presidency: United States; Presidential Government; Presidential Popularity

Bibliography

- Barber J D 1972 *The Presidential Character: Predicting Performance in the White House*. Prentice-Hall, Englewood Cliffs, NJ
- George A L, George J L 1956 *Woodrow Wilson and Colonel House: A Personality Study*. John Day Co., New York
- George A L, George J L 1998 *Presidential Personality and Performance*. Westview, Boulder, CO
- Glad B 1980 *Jimmy Carter, In search of the Great White House*. W. W. Norton, New York
- Goleman D 1995 *Emotional Intelligence*. Bantam Books, New York
- Greenstein F I 1969 *Personality and Politics; Problems of Evidence, Inference, and Conceptualization*. Markham Pub. Co., Chicago
- Greenstein F I 2000 *The Presidential Difference: Leadership Style from FDR to Clinton*. Martin Kessler Books, New York
- Lasswell H D 1930 *Psychopathology and Politics*. The University of Chicago Press, Chicago
- Lasswell H D 1951 *Democratic Character*, in *The Political Writings of Harold D. Lasswell*. Greenwood Press, Westport, CT
- Neustadt R E 1960 *Presidential Power, The Politics of Leadership*. Wiley, New York
- Riesman D 1950 *The Lonely Crowd; A Study of the Changing American Character*. Yale University Press, New Haven, CO
- Renshon S 1996 *High Hoper: The Clinton Presidency and the Politics of Ambition*. New York University Press, New York

F. I. Greenstein

Presidential Popularity

A president's 'popularity' consists of the public's affection for him and endorsement of his policies. Popularity becomes significant to the extent that it can be used by a president to achieve his policy goals. For much of the twentieth century, commentators discussed the 'mandate' that a president has received in an electoral victory for him and his party in Congress. A more recent development has been the attempt to analyze and measure a president's popularity, defined as a resource enabling him to pursue goals that were not a product of that mandate. It is this latter point on which most recent scholarship has been focused following the publication of Richard Neustadt's *Presidential Power*, arguably the most influential book on studies of the presidency. Neustadt stated that a president's standing with the public, along with his professional reputation, was the base on which his influence and success would lie.

The need that White House officials have shown to organize and mobilize a president's popular standing stemmed from significant changes in the political environment during the second half of the twentieth century. Previously, the most important talent a president needed to gain support in the political system was to be able to bargain and negotiate with the independent political powers of Congress, the political parties, and the states. Without this collective mobilization of resources he did not have the ability to frame government policy. Important exceptions to this constitutional weakness of the office occurred during declared international wars such as World War I, when President Woodrow Wilson was able to direct policy without the normal deliberations and specific legislative powers granted by Congress, and during the extreme internal threat of the Civil War when President Lincoln asserted similar plenary authority.

The changes that led observers of the presidency to recognize the importance of his popularity as a major political resource occurred because of the erosion of hierarchy and authority among those in the political system whose support he needed to achieve his goals. After 1960, the fragmentation of institutional power led presidents and their advisers to try to gain and maintain the allegiance of a majority among the politically involved portions of the public. The diminished ability of a president to make arrangements within the small sphere of those who continuously engaged as primary actors in the political system required that he seek public support for almost all-important domestic policies before they could be instituted.

Beginning with the Administration of Richard Nixon, the White House established and continued to develop institutions at the service of the president so that he could maintain continuous support for his conduct of the office, could gather public support for

his policy direction, and combat his adversaries in the political system and the media whose activities weakened the president's status with the public. As Nuestadt elucidated, presidential power was now dependent not only on his professional prestige among national political influentials but on his popularity and prestige among the public at large.

Changes in technology have also played a central role in the importance of presidential popularity as a major foundation on which a president depends. Until the third decade of the twentieth century, however, a president had no way of reaching a national audience whose attention he could command directly and simultaneously. This changed dramatically with the arrival of radio, film with sound in the ubiquitous newsreels in the third and fourth decades of the twentieth century, and the establishment of television as the primary universal channel of communication and news in the fifth and sixth decades, when a president could sell his policies directly to an audience made up of large segments of the public, an audience that could increase his influence with formal apparatus of power in the system.

An early prediction of the importance of presidential popularity may be found in the pre-presidential work of Woodrow Wilson. In series of lectures delivered in 1908 while still in the academic world, Wilson stated that the president 'is the one person about whom a definite national opinion is formed, and, therefore, the one person who can form opinion by his own direct influence and act upon the whole country at once' (Wilson 1961, p. 127.) Other major commentators on the presidency through the 1950s put little or no emphasis on the issue of popularity. Among the best-known authors of texts on the presidency during this era, Clinton Rossiter provided only passing mention of the consequences for a president of his standing with the public (Rossiter 1954). Rossiter did suggest that communication with the public is an important tool of presidential leadership. Edward Corwin, author of the most widely used text on the presidency, made no mention at all of popularity (Corwin 1989).

By the 1960s and 1970s, political scientists, journalists, and other observers recognized that there was an important relationship between what a president had been able to accomplish and his personal support among the public. Especially after the end of the Cold War in the late 1980s, a president could no longer use his prestige as defender against a recognized adversary to gain support for his efforts.

Studying these phenomena, however, proved both difficult and elusive. What exactly was a president's popularity? Did it stem from his personal skills in presenting himself, or from the benefits the public believed could be attributed to him? Did events drive his popularity up or down, or did his popularity enable him to undertake difficult and risky objectives, which, if successful, would make him loved and respected by

the public? Furthermore, although you could show statistical correlations between the president's standing with the public as measured by polls, surveys, and particular events of his administration, causal links would be much harder to establish. Attempts to analyze the hows and whys of the level of popular support a president received during his term in office did not necessarily establish the consequences of this support.

The issue of presidential popularity now consists of a body of work divided here into two categories:

(a) The first encompasses research into how to define and measure presidential popularity, and how to explain changes in that popularity over time. Some variables that are commonly offered as potential explanations for changes in presidential standing with the public are presidential actions, the president's abilities as a communicator, news events, and the presentation of the news by the media as well as the general level of stability and prosperity in the nation.

(b) The second category includes work that addresses the question of how presidential popularity serves as a resource or constraint on a president in his dealings with other political actors. In particular, researchers in this area focus on a president's institutional and personal efforts to use his popularity to garner support from other institutional and political supporters in the system.

1. *Measuring Presidential Popularity*

Efforts by scholars to measure and obtain data on the public's perception and approval of a president have relied on polls, particular those published by the Gallup organization, which has provided results since the 1930s. From 1975 through 2000, Richard Brody provided the most prolific and important analysis summarizing public attitudes toward the president (Brody 1991; this book provides the best bibliographic source on works by Brody and others through 1991). In particular, Brody looked for correlations in the popular standing of the president, and public support for the president's ability to handle economic, social, and international situations. Following the course of a president's popularity during several administrations, he found that 'news events are dynamic ingredients in the processes of opinion change.' Further, 'evidence of policy success and failure contained in daily news reports in the mass media' are the basis on which the American people form and revise their collective evaluation of the president. (Brody 1991, p. 9)

Students of presidential popularity have paid a great deal of attention to the so-called 'rally phenomenon,' first put forward by John Mueller in 1970 (Mueller 1985). He suggested that poll data indicated that presidents received a significant boost in their popular rating from a public affected by patriotism when news events suggested that an international

crisis could involve the safety of the nation or its military forces. The widespread dissemination of this theory during the Cold War led to the political charge that a president might undertake military action deliberately to distract the public from other failings. Subsequent research, however, has indicated that a president receives such a boost in support only if there is no counterelite opposing his action. Since the end of the Cold War, many international events have been greeted with debate rather than support by the opposition and the rally phenomenon seems much more ephemeral.

A continuing problem of this research has been the difficulty of determining the proximate causes of fluctuations in a president's popularity. For example, although one can point out that a particular level of popularity was measured at the same time as certain events occurred, it is not clear whether there was support for a president's policy because he was popular or because the public endorsed the policy. Further, as suggested by Paul Gromke and Brian Newman, poll standing may reflect an 'aggregation of personal preferences' rather than the 'preferences of aggregates' (see Gronke and Newman 2000; this paper contains an excellent bibliography of recent works in the field.)

2. The Fabric of Presidential Popularity

The second line of study involves descriptive research emphasizing the action and/or statements of a president and his entourage as they attempt to bolster his standing with the public. Using interviews, observations, and reviewing documents, these researchers present a portrait of the importance of popularity as a factor determining a president's status in the political system. In particular, they have focused on the methods the White House employs to enhance a president's success in gaining approval for his administration. They also examine the activities of those in politics or working for news organizations who provide an alternative and frequently critical version of what is going on.

In *Going Public*, Samuel Kernell describes the process by which presidents have needed to gain public support in a political universe where persuading Congress or the bureaucracy involves approaching them with a powerful backing (Kernell 1997). In a study of the relationship between news organizations and the White House, Michael Grossman and Martha Kumar describe the institutionalization of White House operations designed to maximize the presi-

dent's standing through the news media and the efforts by news organizations to present an independent and alternative viewpoint (Grossman and Kumar 1981). In *Congress and the Presidency*, Nelson Polsby (1976) describes the difficult problem a president has in converting his public support to a body that has its own connection with the public: 'It is hard for a President to change all this while explaining the intricacies of a particular issue over television. He can, however, communicate his concern about the issue ... and, to a certain extent, set the terms in which it will be debated' (p. 187).

Researchers of this type also have a problem determining whether a president's standing is due to their efforts or other factors in the political environment. In spite of this lack of generally applicable tools of measurement, however, all researchers agree that what they have found has provided important information about how a president's standing affects his success.

See also: Campaigning: Political; Executive Branch, Government; Media, Uses of; Party Identification; Political Parties; Presidency: United States; Presidential Government; Presidential Personality; Public Opinion: Political Aspects

Bibliography

- Brody R A 1991 *Assessing the President: The Media, Elite Opinion and Public Support*. Stanford University Press, Stanford, CA
- Corwin E S 1989 *The President: Office and Powers 1787-1984*. New York University Press, New York
- Edwards G C 1989 *At the Margins: Presidential Leadership of Congress*. Yale University Press, New Haven, CT
- Gronke P, Newman B 2000 FDR to Clinton. A 'state of the discipline' review of presidential approval. Paper presented at the American Political Science Association Annual Meeting, September
- Grossman M B, Kumar M J 1981 *Portraying the President: The White House and the Media*. Johns Hopkins University Press, Baltimore, MD
- Kernell S 1997 *Going Public: New Strategies of Presidential Leadership*, 3rd edn. Congressional Quarterly Press, Washington, DC
- Mueller J 1985 *Wars, Presidents, and Public Opinion*. Lanham, MD
- Polsby N 1976 *Congress and the Presidency*. Englewood Cliffs, NJ
- Rossiter C 1954 *The American Presidency*. Harcourt Brace, New York
- Wilson W 1961 *Constitutional Government in the United States*. Columbia University Press, New York

M. B. Grossman

Copyright © 2001 Elsevier Science Ltd.
All rights reserved.

Pri

Primary Elections

One of the major functions that political parties perform in democratic political systems is the recruiting and selection of candidates for public office. In most democracies, this task is performed by a small number of official party members, meeting together as a caucus, committee, or convention. In the USA, by contrast, candidate selection is done almost entirely by party primaries: preliminary elections, held under governmental auspices, in which virtually every interested voter is allowed to participate, regardless of their past or current support for the party and its purposes.

1. History and Development

From the 1830s through the early 1900s, American political parties nominated their candidates for public office through a multitiered system not unlike those still used in most other democracies. The system began with gatherings of the party membership in small local units such as towns, wards, or precincts, which would select delegates to county or district conventions. These county and district conventions would nominate some candidates themselves and might also select delegates to participate in a state convention.

In the first decade and a half of the twentieth century, however, this system was almost entirely abandoned, as most states enacted legislation requiring the major political parties to nominate their candidates through primary elections. The rapid and widespread adoption of primaries was the confluence of two distinct trends within the American political system. The first was a movement to subject the parties to a greater measure of legal regulation. Though parties clearly played a major role in electoral politics, they were, in the beginning, regarded as entirely private entities, such that fraud, intimidation, and corruption in party conventions was seen as a purely internal matter. As Key (1958) put it, 'it was no more illegal to commit fraud in the party caucus ... than it would be to do so in the election of officers of a drinking club.' Shortly after the Civil War, however, a number of states began passing laws aimed at outlawing bribery and fraud and preserving some measure of order and procedural integrity in party affairs. By the end of the nineteenth century, almost every state had enacted some legislation of this kind (for a detailed discussion of this development, see Merriam 1908).

Even as this trend was sweeping the field, however, a second demand was being made: that the parties needed to be 'democratized.' To quote Key again, 'through the history of American nomination practices runs a persistent attempt to permit popular participation in nominations and thereby to limit or to destroy the power of party oligarchies.' In the early 1900s, this viewpoint had a particular resonance with the Progressive movement, which saw popular government as being threatened by an alliance of party machines and powerful economic interests. The push for party 'democracy' was given further impetus by the increasing sectionalism of the American party system. By the late 1890s, many parts of the country had become so completely dominated by one party or the other that receiving the major party's nomination was tantamount to being elected.

Though a number of southern Democratic parties had already used primary elections on their own initiative, and other states had made them available on a voluntary basis, the first comprehensive, mandatory, statewide primary law was adopted by Wisconsin in 1903. Wisconsin's example was quickly emulated in a number of other western and midwestern states, and then spread to the rest of the country. By 1917, 'the direct primary had been adopted for most nominations in all save a handful of states' (Key 1958). A small number of states resisted the trend, but the tides of historical circumstance were clearly against them. The last holdout was Connecticut, which finally enacted a primary law in 1955. As of the early twenty-first century, primary elections are clearly the dominant way of making nominations throughout the USA.

It is perhaps worth adding that most state election laws require nomination by primary election only for so-called major parties (usually defined as parties that received some minimum percentage of the vote in the last election). Minor and third parties, such as the Libertarians or the Socialists, still generally use a caucus-convention process to nominate their candidates for public office.

Students of comparative politics have generally regarded primary elections as a uniquely American institution, strikingly different from the nomination systems employed in every other democracy (see, for example, Ranney 1981). While this still holds true as a generalization, there are some signs that other countries are beginning to experiment with primaries. In 1999, for example, the PRI in Mexico nominated its presidential candidate through an open, national primary. In these other countries, however, the de-

cision to use a primary is generally made by the party itself; only in the USA is the primary imposed upon the major parties by law.

2. Presidential Primaries

The trend toward primaries also partially transformed the presidential nomination process. Since the 1830s, both major parties in the USA had nominated their presidential candidates at national conventions, to which each state sent its own delegation. These state delegations, in turn, were selected at state party conventions or by state party committees. Beginning in 1912, however, a sizable number of states started using primary elections to choose their national convention delegates. Some Progressives (including President Woodrow Wilson) wanted to go further, scrap the national conventions entirely (at least as a device for making nominations), and choose the parties' presidential candidates through a national primary.

Yet, for a variety of reasons, primaries never became as thoroughly institutionalized in the presidential nomination process as they did in making state and local nominations. Proposals for a national primary never came close to being enacted; and while 25 states had passed presidential primary legislation by 1916, over the next 15 years eight of them repealed these statutes. The result, when all the dust had settled, was a distinctive system for conferring presidential nominations that modern scholars generally call the 'mixed system.' Under the mixed system, about one-third of the states selected their national convention delegations via presidential primary, while the remaining two-thirds used state conventions or state party committees. National conventions thus had a partly popular, partly organizational basis, and remained significant decision-making bodies, though the presidential nomination itself was increasingly decided in a single ballot (the best account of presidential primaries during this period is Davis 1967).

The mixed system came to an end in the early 1970s. In 1968, at a remarkably divisive and chaotic national convention, the Democrats voted to create a commission to re-examine their presidential selection process. In just four years, this Commission on Party Structure and Delegate Selection (often called the McGovern-Fraser Commission, after the two men who served as its chairmen) completely rewrote the party's rules governing delegate selection—so much so that most scholars see the years since 1968 as constituting a new era in presidential nominations, often referred to as the plebiscitary system. While the plebiscitary system differs from the old, mixed system in a variety of ways, one of the most important differences has been a huge increase in the number of presidential primaries. Where the 1968 election fea-

tured 17 Democratic presidential primaries and 16 Republican primaries, by 2000 there were 39 Democratic primaries and 43 Republican primaries. Under the mixed system, it was quite possible for someone to win a presidential nomination without contesting a single primary; under the new system, the nomination essentially goes to whichever candidate wins the most primaries (on these points, see Hagen and Mayer 2000). Though both parties still hold national conventions, they are increasingly empty affairs, put on for the benefit of television, in which all the important questions have already been decided.

3. Who Can Participate?

Since primaries are a product of state rather than national legislation, primaries vary across states in a host of major and minor ways. One of the most important differences, for both its practical and theoretical consequences, concerns who is allowed to vote in the primaries. At one extreme are the so-called closed primaries, in which every registered voter is asked if they are supporters of or affiliated with any particular party, and then only those who have expressed an affiliation with a party in advance are allowed to vote in that party's primary. (Even in this 'extreme' case, it should be noted, the level of commitment required from the voter is exceedingly light. Voters are only asked to affirm a kind of general, psychological tie to the party; they are not required, in any state, to show that they have performed any previous service for the party or to make a binding promise of future support.)

A somewhat more permissive regime, that might be called semiclosed primaries, prohibits registered Democrats from voting in Republican primaries and vice versa, but allows unaffiliated or independent voters to vote in either party's primary. A related provision, found in many state statutes, allows voters to change their party affiliation when they show up at the polling place on election day. More permissive still are what might be called declarative systems, where there is no party registration and any voter may vote in either party's primary simply by asking for that party's ballot.

Declarative systems at least ask the voters to state publicly that they wish to participate in the Democratic or Republican party primary. Open primaries, by contrast, relieve the voters of even this encumbrance. In open primaries, each voter is, in effect, given a ballot for both parties and then allowed to make his or her decision about which primary to participate in in the privacy of the voting booth. Since open primaries at least require a person to vote in only one party's primary on any given election day, a number of western states have recently sought to remove even this last imposition on the long-suffering voter. So-

called blanket primaries allow voters to participate in both party's primaries on the same day, casting a vote in, say, the Democratic primary for the US Senate, the Republican primary for the US House, the Democratic primary for state representative, and so on. (In June 2000, the US Supreme Court declared the California blanket primary law unconstitutional.)

States frequently change their primary legislation in large and small ways, and many of the systems that emerge are not easily placed in one of the categories listed above. But according to a recent analysis by Jewell and Morehouse (2001), 11 states had closed primary laws, 15 used semiclosed systems, 10 had declarative systems, 10 used open primaries, and 3 (including California) had opted for the blanket primary. As has already been suggested, none of these systems demand very much from the voter in the way of past or future support for the party, and none except the closed primaries even do much to prevent members of one party from helping select the other party's candidates.

Does it matter what type of participation standard a state chooses to adopt? Much of the concern about having open or declarative primaries originally focused on the possibility of party 'raiding': cases where large numbers of persons affiliated with one party would vote in the other party's primary, with the deliberate intention of nominating the weakest candidate. But there is little evidence that this sort of thing occurs: it is difficult enough to get most American voters to the polls at all, much less to convince them to carry out a complicated strategic maneuver. A more likely consequence of open primaries is that they make it more difficult for the parties to present a cohesive ideology and set of candidates. Independents and members of the opposition party tend to be less committed to a party's core symbols and principles, and thus more willing to support a candidate who is out of step with large segments of his or her own party.

4. *Who Does Participate?*

In numerical terms, voter turnout in primary elections is substantially greater than the turnout at local caucuses (where these are still held), but smaller than the turnout at general elections. The best venue for comparing turnout rates is presidential elections, for besides having reliable data on the number of votes cast in presidential primaries and in the November general election, there are, as noted earlier, still about 10–15 states that continue to select their national convention delegates through a caucus-convention procedure. In general, contested presidential primaries attract about 20–35 percent of a state's voting age Democrats or Republicans, compared to a 50–55 percent turnout in the general election and a participation rate of just 1–3 percent in the presidential caucuses. In addition, there is far greater variation in

primary election turnout rates, depending on how many seriously contested races are on the ballot.

One major exception to these generalizations occurred in the South when that region was still a one-party enclave. Since winning the Democratic nomination virtually guaranteed a candidate's triumph in the general election, southern Democratic primaries regularly drew a far larger turnout than the general elections that followed.

It was once widely believed that the people who voted in a party primary were not very representative of the party's typical voter or identifier. More recent research, however, suggests that this fear has been exaggerated. While there are some differences between primary voters and general election voters (the former are, for example, somewhat older and better educated), the differences are relatively small and do not appear to have much effect on either the type of candidate who gets nominated or the ideological views that get expressed.

5. *Consequences*

The widespread use of primary elections as a nominating device has had several major consequences for American politics. First, it greatly weakened party organizations by taking away from them the single most important function they once performed. Prior to the Progressive Era, it was very difficult for a candidate to get nominated by claiming to be an 'outsider' or insurgent, running in avowed opposition to the party establishment. Since the adoption of the primary, however, such candidacies have become quite common. This is not to say, of course, that candidates endorsed or supported by the regular party organization invariably lose primaries; but primaries have greatly increased the odds that a non- or anti-organization candidate will be successful.

Primaries also greatly lengthen the public phase of American election campaigns, since they generally guarantee that, in one party or the other, there will be two distinct campaigns: one for the nomination and one for the general election. Primary campaigns are also more expensive than those required in a caucus-convention system, since they must be targeted at a much larger and less attentive audience.

Though the point is disputed, there is also a long-standing concern that primaries make it more difficult for the parties to present a united front during the general election. According to this argument, often referred to as the 'divisive primary' hypothesis, primaries require the parties to air all their grievances and divisions in public, rather than settling them behind closed doors. In some cases, indeed, candidates may go out their way to create or emphasize divisions, as a way of distinguishing themselves from the other contenders in a primary. Primaries, it is further argued, may be good at assessing the first choices of party

voters, but are not very good at discovering compromise candidates that are acceptable to a variety of factions, or at promoting 'balanced tickets' that offer some measure of reward or recognition to a diverse set of interests. Other scholars, however, claim that primaries do not so much create divisions as reveal the ones that already exist (for a review of these issues, see Mayer 1996). At the very least, one can say that sharp intraparty conflict was certainly not absent from the days when American parties selected their candidates through caucuses and conventions, and that the primary has not prevented most Republican and Democratic candidates from receiving a substantial majority of the votes cast by their fellow partisans.

See also: Electoral Systems; Key, Valdimer Orlando (1908–63); Participation: Political; Party Systems; Political Money and Party Finance; Political Parties; Political Parties, History of; Voting: Turnout

Bibliography

- Davis J W 1967 *Presidential Primaries: Road to the White House*. Thomas Y. Crowell, New York
- Hagen M G, Mayer W G 2000 The modern politics of presidential selection. In: Mayer W G (ed.) *In Pursuit of the White House 2000: How We Choose Our Presidential Nominees*. Chatham House, New York
- Jewell M E, Morehouse S M 2001 Political parties and elections in American States, 4th edn. *Congressional Quarterly*, Washington, DC
- Key V O 1958 *Politics, Parties, and Pressure Groups*, 4th edn. Thomas Y. Crowell, New York
- Mayer W G 1996 *The Divided Democrats*. Westview Press, Boulder, CO
- Merriam C E 1908 *Primary Elections*. University of Chicago Press, Chicago
- Ranney A 1981 Candidate selection. In: Butler D, Penniman H R, Ranney A (eds.) *Democracy at the Polls*. American Enterprise Institute, Washington, DC

W. G. Mayer

Primary School Education

1. Introduction

The primary grades are the first few grades of elementary school, in which students aged approximately 5–8 are socialized into the student role and taught basic tool skills. Evolving views on what constitutes appropriate curriculum and instruction during these grades have been influenced by developments in theory and research on the needs, interests, capabilities, and other characteristics of children in this age range. Emphasis has shifted from maturation

of generic stages of cognitive development to learning of domain-specific knowledge networks, so that readiness to learn is now seen as dependent more on relevant prior knowledge and experience than on age or developmental level.

2. From Generic Developmental Stages to Domain-specific Learning

The term 'primary school education' refers typically to the first two or three grades of elementary school, in which curriculum and instruction emphasize basic literacy and numeracy skills. These grades also emphasize socialization into the student role: teaching children to pay attention to lessons and to work persistently on assignments; to follow prescribed conventions for participating in whole-class activities (e.g., raise hand and wait to be called on); to collaborate productively with peers in pairs and small groups; and to take good care of personal possessions and classroom equipment. In many countries, formal schooling in the primary grades is preceded by a transitional year (called kindergarten in the USA). It is a year in which students are oriented to school routines, engaged in activities that feature artwork and other multisensory representations that call for listening, speaking, counting, building vocabulary, making visual discriminations, exercising fine motor skills, or in other ways developing readiness for the more formal academic learning emphasized in subsequent grades.

At all levels of schooling, curriculum development involves negotiating a reasonable balance among four general approaches: (a) equipping students with knowledge that is lasting, important, and fundamental to the human experience (emphasizing but not limited to the academic disciplines), (b) following the natural course of child development by matching content to students' interests and learning needs at each grade level, (c) meeting society's needs by preparing students to assume adult work roles and citizenship responsibilities, and (d) combating social injustice and promoting social change by addressing social policy issues (Kliebard 1994). The principle of meeting children's developmental needs and interests has special relevance for the primary grades because substantial cognitive development occurs during these years. Children aged five–eight (especially five–six) are often considered to be not merely less knowledgeable, but qualitatively different in several respects from older children and adults. Philosophers traditionally have maintained that children reach the *age of reason* at about age seven, when developments in their understanding of cause-and-effect relationships, reciprocity and the Golden Rule, and other key concepts and principles enable them to distinguish right from wrong and engage in principled moral reasoning. Researchers studying children's cognitive development speak of a

developmental shift that occurs between ages five and seven as children become less like animals and infants and more like older children and adults in their performance on a broad range of perceptual and cognitive tasks (Sameroff and Haith 1996).

Jean Piaget (1983) described children aged two–seven as egocentric in purview and unstable in their cognitive structures and thought patterns. Within his comprehensive stage theory of cognitive development, such children were labeled as preoperational because they had not yet acquired reversibility of thinking, conservation, classification, seriation, and other *concrete operations* that stabilize older children's cognitive structures, and enable them to engage in systematic reasoning (see *Cognitive Development: Child Education*). Piaget's work implied that progression from the preoperational to the concrete operational stage was determined primarily by maturation and thus not open to significant acceleration through educational interventions. Consequently, primary schooling models influenced by his ideas tended to emphasize play, exploration, discovery learning, and sensorimotor activities over more formally academic and verbally mediated instruction. Ideas drawn from Arnold Gesell, Maria Montessori, and others who emphasized developmental stages or the presumed special needs and characteristics of young children similarly led to early childhood and primary-grade education models that emphasize short and varied activities, hands-on learning, concrete and integrated topical units rather than systematic treatment of disciplinary content, and attempts to be responsive to children's learning initiatives rather than to move them through preplanned curricula in lock-step fashion. Contemporary notions of developmentally appropriate practice continue this tradition (NAEYC 1991), as do the British Infant School model and the Italian Reggio Emilia model.

In recent years, the predominant theoretical emphasis in developmental psychology has shifted from the stage theory of Piaget to the sociocultural learning theory of Lev Vygotsky (1978), and research on children's cognition has shifted emphasis from general developmental stages to domain- and situation-specific learning. Findings indicate that many aspects of cultural learning, including much of what is taught in elementary school, can be acquired through instruction, in which teachers provide modeling and explanations for students and coach and 'scaffold' their learning efforts. Instead of emphasizing limitations in cognitive capacity, pending the unfolding of genetically programmed maturation of internal structures, contemporary thinking emphasizes connecting with and building on students' prior knowledge in the domain. Instruction focuses on the *zone of proximal development*: the range of knowledge and skills that students are not yet ready to learn on their own but can learn with help from teachers (Tharp and Gallimore 1988).

A related notion is that instead of being organized into hierarchical levels, knowledge is organized within networks structured around key ideas. This implies that, instead of having to begin with the lowest cognitive level (knowledge) and proceed systematically through successively higher levels (comprehension, application, analysis, synthesis, and evaluation), instruction can begin anywhere in the network and develop understandings from there. Initial possibilities will be limited by the extent of learners' prior knowledge in the domain, but in principle, learners can accomplish any domain-related learning goals if they experience whatever instruction and learning opportunities may be required (without having to wait for supposed maturation of cognitive structures to occur).

Studies of children's domain-specific knowledge have verified that children's knowledge and learning capacities in many areas are much more advanced than previously suspected, although children find certain concepts difficult to grasp and certain misconceptions are commonly developed (e.g., that the sun rotates around the earth instead of vice-versa). It appears that Piaget's notions about children's learning capacities were too pessimistic and that his stage concepts are more applicable to developments in children's mathematical and scientific thinking than to developments in their thinking about issues addressed in the humanities and social sciences. Children are able to understand historical and literary content that is encoded within narrative structures built around the goals and motives of central characters, as well as cultural comparisons or other social science content that focuses on food, clothing, shelter, transportation, or other cultural universals with which they have had personal experience (Downey and Levstik 1991).

Research supports the following generalizations about difficulty level of curriculum content: (a) difficulty increases with the number of items that must be addressed simultaneously (it is harder to compare or interpret three things than two things); (b) abstract content is more difficult than more concrete content; (c) fine distinctions are more difficult than gross ones; (d) learning from print alone is more difficult than learning from multiple media; (e) it is easier to develop skills in thinking about matters that students see as closely related to their own lives than about other matters; and (f) providing structure, cues, and props makes thinking easier (Fair 1977).

In summary, much of what Piaget claimed about stage-related limitations in children's cognitive capacities is incorrect or limited to certain knowledge domains. However, Piaget was correct in suggesting that young children are qualitatively different from older children and adults in certain respects, so that it is not possible (or at least, not cost effective) to teach them some of the more abstract and complex aspects of each school subject. Primary-grade children appear to be ready for certain forms of systematic instruction that go well beyond what usually are described as

‘developmentally appropriate practices,’ but not yet ready for discipline-based courses that emphasize abstract content organized conceptually.

3. Curriculum and Instruction in the Primary Grades

Children show the capacity for operational thinking at five or six, but usually do not become functionally operational (use operational thinking most of the time) until a year or two later. Thus, in effect, the preoperational period extends until second or third grade for most students. Preoperational students are ready to learn about familiar experiences and observable phenomena, especially through exploration and manipulation of concrete objects. They also learn basic literacy and mathematics skills and those aspects of content area knowledge that can be linked to their experience base. Instruction in basic skills should connect with knowledge about language, print, counting, and numbers that children develop through their preschool experiences. These tool skills should be taught in an integrated way that stresses explanations for meaningful understanding and opportunities for authentic applications (Adams 1990, Hiebert and Carpenter 1992, Hooper and DeFraim 1980). Execution of basic skills should gradually become automatized, but these skills should be learned as strategies, so that procedural knowledge (about what to do and how to do it) is developed in connection with related propositional knowledge (about the nature and purposes of the skill) and conditional knowledge (about when and why to use it).

Primary-grade students require demonstration and explanation of skills, guided practice with feedback, and frequent opportunities to develop mastery through application. Good demonstrations and visual aids provide concrete models for children to watch or imitate, so that they do not have to struggle to follow purely verbal instruction. Complex tasks are broken into subtasks that children can master in sequences that gradually connect until the ultimate task is learned. When they work on their own, they may need cues, reminders, or other learning *scaffolds* that will help them overcome their limited attention spans and difficulty in following spoken or written directions. Models of completed tasks can be placed in learning centers and worksheets can include features such as spacing and lines to keep things separate; boxes or arrows to indicate where responses should be placed or where to go next; and division of assignments into modular units that can be presented, reviewed, and discussed separately.

Teaching in subject areas other than basic skills should either emphasize familiar content or make the strange familiar by using concrete props or other media. Actual objects are best, but substitutes such as videos or photos are also useful. Because young

children have limited experience and major gaps in their vocabularies, they may need such props to enable them to understand what the teacher is talking about. They are most likely to have difficulty when content is organized around the conceptual structures used in the academic disciplines rather than around their own emerging knowledge structures. Children tend to learn specifics first and build up abstract generalizations only gradually as they accumulate experience; they have trouble following the logical patterns (flowing from general rules to specific examples) that adults find helpful. They are predisposed to learn in various domains, and although they are inexperienced, they reason facily with the knowledge they have (Bransford et al. 1999). However, it is important to focus instruction on familiar and observable events, or at least events that are easily assimilated into the students’ existing schemes. For example, children have an intuitive sense of the past and an interest in people that can provide a basis for instruction in history and social studies. However, such instruction has to reach the operative level rooted in students’ concrete experiences, e.g., by having them investigate their own family histories, explore the geography of the neighborhood, or examine photos and artifacts from another culture. If the instruction stays at a purely figurative level (associating tipis with Native Americans, sombreros with Mexicans, etc.), it will have little meaning or potential for inclusion in networks of knowledge structured around powerful ideas (and will engender stereotypes, rather than layered understandings).

There has been noteworthy cross-national variation in the age at which children begin school and in the scope of the primary-grade curriculum. As more has been learned about children’s knowledge and learning capacities, the trend has been towards earlier school initiation and more demanding curricula. Kindergarten classes in the USA, for example, now feature less sensorimotor play and more verbally mediated instruction than in the past. Even so, primary-grade teachers tend to be nurturant individuals who are oriented more toward teaching children than teaching subject matter, and who are concerned about supporting their students’ personal development and adjustment to school, along with meeting their academic learning needs. Subject matter specialization usually does not begin until at least the fourth grade, so primary-grade teachers tend to be generalists who stay with their home-room class all day and teach all subjects.

Some schools even use looping arrangements that keep intact classes together for two consecutive years with the same teacher, as part of a larger effort to provide emotionally safe and supportive classroom environments for young learners. The children often work in learning centers or sit together on a rug instead of always sitting at desks or in chairs. Teacher aides, adult volunteers, or older students often assist the teacher by working with individuals or small

groups. Assessment is mostly informal, without emphasis on tests or grades. Curricula often present an idealistic and optimistic view of the world, sanitized to eliminate disturbing or depressing content, in an effort to 'preserve childhood innocence' by delaying confrontation with threatening realities. In general, most primary education approaches focus on orienting young children to school and equipping them with basic knowledge and skills, but within a relatively informal and highly supportive learning community. This ethos has persisted through developments from penpals to KIDSNET; from mathematics flash cards to number lines and manipulables; from penmanship drills to authentic writing and publishing; from fixed desks to flexibly arrangeable children's chairs and tables, etc., apparently because it is well-suited to the educational needs and capacities of young students.

See also: Cognitive Development: Child Education; Cognitive Development in Childhood and Adolescence; Education (Primary and Secondary Schools) and Gender; Piaget's Theory of Child Development; Schooling: Impact on Cognitive and Motivational Development; Vygotskij's Theory of Human Development and New Approaches to Education

Bibliography

- Adams M 1990 *Beginning to Read: Thinking and Learning about Print*. MIT Press, Cambridge, MA
- Bransford J, Brown A, Cocking R (eds.) 1999 *How People Learn: Brain, Mind, Experience, and School*. National Academy Press, Washington, DC
- Downey M, Levstik L 1991 Teaching and learning history. In: Shaver J (ed.) *Handbook of Research on Social Studies Teaching and Learning*. Macmillan, New York, pp. 400–10
- Fair J 1977 Skills in thinking. In: Kurfman D (ed.) *Forty-seventh Yearbook: Developing Decision-making Skills*. National Council for the Social Studies, Arlington, VA, pp. 29–68
- Hiebert J, Carpenter T 1992 Learning and teaching with understanding. In: Grouws D A (ed.) *Handbook of Research on Mathematics Teaching and Learning*. Macmillan, New York, pp. 65–97
- Hooper F, DeFrain J 1980 On delineating distinctly Piagetian contributions to education. *Genetic Psychology Monographs* **101**: 151–81
- Kliebard H 1994 *The Struggle for the American Curriculum 1893–1958*, 2nd edn. Routledge, New York
- National Association for the Education of Young Children (NAEYC) 1991 Guidelines for appropriate curriculum content and assessment in programs serving children ages three through eight. *Young Children* **46**(3): 21–38
- Piaget J 1983 Piaget's theory. In: Mussen P (ed.) *Handbook of Child Psychology*, 4th edn. Wiley, New York, Vol. 1, pp. 703–32
- Sameroff A, Haith M 1996 *The Five to Seven Year Shift: The Age of Reason and Responsibility*. University of Chicago Press, Chicago
- Tharp R, Gallimore R 1988 *Rousing Minds to Life: Teaching, Learning, and Schooling in Social Context*. Cambridge University Press, Cambridge, UK

Vygotsky L 1978 In: Cole M, John-Steiner V, Scribner S, Souberman E (eds.) *Mind in Society: The Development of Higher Psychological Processes*. Harvard University Press, Cambridge

J. Brophy

Primate Behavior: Significance for Understanding Humans

Interest in the behavioral similarities between human and nonhuman primates can be traced to Huxley's (1863) assertion of a phylogenetic relationship between humans and African apes. From the outset, primate studies have been incorporated into diverse fields of inquiry, including biology, ecology, psychology, and anthropology. Physical anthropologists were intensely interested in the degree to which primate behavior could be viewed through an interpretive lens for insights into the behavior and mental capacities of ancestral humans. Contemporary primate studies continue to inform a broad array of fields in both the human and nonhuman behavioral sciences; this integrative approach contributes to our understanding not only of human evolution, but of the behavior of extant humans as well. In this article we briefly outline the history of primate behavioral studies within the field of physical anthropology, discussing past uses of primatology for building hypotheses of human evolution. We then discuss changes in the theoretical approach to the study of primate behavior and their effect on the understanding of social systems. We conclude by presenting current issues in primatology as they relate to the study of human social evolution and behavior.

1. Early Primate Models of Ancestral Hominin Behavior

Early primate models for the behavior of ancestral humans were based on either: (a) assumptions of similarity in ecological adaptation between extant savannah-living primates and ancestral humans, or (b) shared behavioral traits within the clade of African hominoids (including humans). Raymond Dart, the discoverer of the *Australopithecus* specimen at Taung, popularized the hypothesis of a saltatory shift to human bipedalism, tool use, and hunting as a result of moving from forests to savannahs (Dart 1957). Dart drew explicit parallels to baboons as contemporary savannah-dwellers that might share behavioral similarities to early humans. Brief field studies of wild baboons revealed distinct differences from modern humans. Nonetheless, early descriptions of baboon sociality, which emphasized male dominance and

aggression as important influences on social structure, were used to typify the human condition (Kinzey 1987). The historical prejudices and methodological weaknesses that gave rise to these models have received considerable critical scrutiny (Cartmill 1993). Despite their methodological weaknesses, these early studies planted the seeds for future primate field studies by asking the following questions: (a) In what way do ecological parameters shape social systems? (b) What were the relevant ecological parameters that shaped early human evolution?, and (c) What can extant primates tell us about human behavior and social evolution?

While ecological affinities led to the baboon model, phylogenetic affinities between human and nonhuman apes prompted field studies of chimpanzees, gorillas, orangutans, and bonobos (Hamburg and McCown 1979). Taking a comparative approach, Louis Leakey sponsored well-publicized field studies of three great ape species. Anthropologists focused most attention on the common chimpanzee, the closest living relative to humans (Sibley and Alquist 1987). Behaviors observed in both humans and chimps were assumed to be phylogenetically conservative and therefore likely to have been expressed in our shared common ancestor. Tool use, hunting/meat eating, and the savannah/woodland habitat of chimps at Gombe reinforced the confidence of some anthropologists in the importance of this suite of characters for the evolution of humans (Kinzey 1987).

Contrary to expectation, these studies revealed that, neither baboon nor chimpanzee models have strong explanatory power in reconstructing the behavior of ancestral humans. First, extant species evolved under unique selective pressures and are not satisfactory typological specimens of ancestral species. Second, there is significant intraspecific variation in the behavior of chimpanzees and baboons, with no adequate criteria for selecting which behaviors are applicable to models of ancestral humans. Third, if behaviors common to chimpanzees and humans are assumed to be phylogenetically conservative, there are significant problems with interpreting morphological evidence from the fossil record that is not shared by either modern chimpanzees or humans (see Potts, in Kinzey 1987).

These criticisms apply to the assumptions and methods employed by these modeling approaches and not to the usefulness of primate behavioral studies *per se* in elucidating past and current human behavior. Past mistakes have served to refine the approaches by which primate studies are conducted and applied. At the beginning of the twenty-first century, few primate studies make explicit use of typological models. Contemporary studies of primate behavior emphasize a comparative approach in which the functions of behaviors are investigated under different ecological and social conditions. The next section presents important conceptual changes in the study of

primate behavior and their influence on the application of primate models to the understanding of human behavior.

2. Conceptual Shifts in Primate Behavior Studies

During the same time that primate behavior was being used to construct models of ancestral hominines, it was also an active area of interest in the zoological sciences, particularly in the subfield of behavioral ecology. Behavioral ecology investigates behavioral adaptations to ecological and social conditions. Although behavioral ecologists directly observe phenotypic traits, they use these observations to test falsifiable predictions generated within an evolutionary framework which states that individuals act in ways that increase their inclusive fitness. For example, studies testing optimal foraging theory use economic cost-benefit models to generate predictions of behavioral responses to ecological conditions that affect an individual's energy budget. Other studies, particularly those that focus on reproductive behavior such as mating and offspring care, examine an individual's decisions when situated in a particular social milieu (Krebs and Davies 1991). This approach to the study of primate behavior is distinctly different from that which was used previously in two important ways: (a) its goal is to test and further develop general evolutionary principles for behavior rather than create models of specific extinct species, and (b) comparative outgroups can be taken from all animals rather than those restricted to the primate order.

Within the field of anthropology, human behavioral ecologists applied evolutionary concepts to test economic models of the ecology of contemporary human hunter-gatherers. Human behavioral ecologists focused on contingencies that had been found to be important in non-human primates: predation, reproductive opportunities, and the distribution and predictability of resources. However, the application of primate studies to the interpretation of human behavior encountered significant difficulties following the publication of E. O. Wilson's *Sociobiology* (1975). Criticism of early sociobiological theory is well documented and expressed two main concerns. First, the effusion of speculations on the application of sociobiological principles to human behavior was untestable and therefore unfalsifiable (Caro and Borgerhoff Mulder 1987). The second body of criticism, which objected to sociobiological depictions of humans as inflexible executors of the interests of selfish genes, stemmed from a poorly explained differentiation between genotypic determinism and the plasticity of behavioral phenotypic expression. Suites of behaviors resulting from complex socioenvironmental conditions (e.g., 'rape') were incorrectly treated as simple units of evolutionary analysis (Sunday and Tobach 1985). Disagreement over the application of

sociobiological theory to the study of human behavior resulted in a division between cultural anthropologists, who emphasized human uniqueness, and those biological anthropologists who studied primate behavior with an eye toward understanding that of humans. While it is a truism that genetic make-up influences behavior, and genetic relatedness influences to whom particular behaviors may be expressed, genetic determinism has no place in behavioral ecology.

The contemporary use of primate behavior within anthropology has shifted from the use of primates as typological models of human origins to field studies and experiments that test predictions developed within a behavioral ecology framework. Significant methodological changes have accompanied this theoretical shift. First, standard methods for the collection of quantifiable data facilitate interpretable statistical tests of predictions and comparison across primate studies (Altmann 1974). Second, more primate species are studied, thereby providing a larger and more diverse sample for comparison. Third, long-term field studies provide information on seasonal and inter-populational variability in behavior, the ontogeny of and change in relationships among individuals, and the causes and consequences of demographic processes in long-lived species. Fourth, a more diverse representation of scientists expands the field of inquiry to include aspects of social life previously assumed ancillary to social organization, particularly female reproductive strategies and the energetics of offspring care (Altmann 1980, Hrdy 1999). Analyses of the resulting data sets demonstrate rich intra- and inter-specific variability in behavior and relationships that parallel the intraspecific variability observed in humans.

3. Current Primate Studies in Anthropology

3.1 Social Behavior

Two general principles are commonly invoked to explain group size and composition in birds and mammals. First, female reproductive success is limited primarily by the availability of ecological resources, particularly food. With whom females form groups, and the competitive regime between females within those groups, is a function of the distribution of ecological resources in space and time (Wrangham 1980, van Schaik 1989). Male distribution, on the other hand, is patterned upon that of females (Andelmann 1986). The diversity of grouping patterns observed among primates across a range of environments supports this general framework.

Whether of the scramble or contest type, competition for resources underlies this ecological model (Krebs and Davies 1991). Competition, however, can be expressed in multiple ways. Aggression between group members certainly occurs, and individuals

obtain inclusive fitness benefits by aiding kin in contests, but not all competition results in direct conflict. Unrelated individuals also engage in cooperative and affiliative behaviors, and reconciliation often occurs after conflicts (de Waal 1983). Several benefits accrue from affiliative interactions with non-kin. First, individuals who engage in post-conflict reconciliation experience more rapid reductions in stress response than those that do not (Aureli 1992). Second, initiating affiliative behaviors garners favor with powerful group members and develops relationships with potential future allies (de Waal 1983). Third, males who invest in nonsexual affiliative behavior toward females obtain future matings with their 'friend' (Smuts 1985). 'Peaceful' interactions are essential to the social cohesion of groups and provide an alternative means of exerting influence and eliciting support. Long life spans and the opportunity for repeated encounters between known individuals provide the conditions under which individuals can predicate present behavior on anticipated future social interactions (Byrne and Whiten 1988).

The study of behavioral ecology in humans is likewise concerned with the tactics employed by modern humans to survive and reproduce within a particular socioecological context. Using principles developed in the field of animal behavior, human behavioral ecologists collect data from modern hunter-gatherers to test predictions derived from ecological optimization models (Winterhalder and Smith 2000). Human behavioral ecologists are able to measure directly the economic results of behavioral traits thought to be important to human evolution such as the intersexual division of labor, long juvenile dependence, and females' lengthy post-reproductive life span. The results of such studies have led to important challenges to the primacy of hunting in human history, suggesting instead that the economics of female provisioning and care of offspring provide a more parsimonious explanation for the evolution of human social systems (Hawkes et al. 1997, O'Connell et al. 1999).

Because of their general applicability, the use of behavioral ecology models raises the question of whether primate studies in particular are any more informative than studies of other taxa. Several lines of evidence combine to produce a positive answer. While studies of all taxa, particularly social mammals, serve to test theories that may be relevant to humans regarding the causes and consequences of social systems, the complex nature of social relationships observed in nonhuman primates make them particularly useful as comparative species for understanding the function of human social relationships. In addition, the use of nonhuman primates as comparison species controls for traits more phylogenetically conservative than behavior that may influence behavioral options (e.g., life history parameters, functional morphology, and physiology). Confining the study of

behavioral ecology to humans within the field of anthropology limits the perspective from which to understand human behavior. Cross-cultural studies can demonstrate intraspecific variation, but they do nothing to reveal which traits or behaviors are uniquely human. Interspecific comparative studies among primates reveal the extent and degree to which certain behavioral patterns are confined to humans. Primate studies therefore both focus and expand the field of vision in the study of human behavior.

3.2 Functional Morphology

Anatomists have a rich tradition of inferring the behavior of extinct species based on fossil skeletons, and tooth morphology figures prominently in the reconstruction of hominin diets (Jolly 1970). Interpretation of fossils is only possible, however, with studies of analogous structures and their function in extant species. Recently, anatomists have teamed up with primate behavioral ecologists to combine direct observations of primate feeding behavior with dental morphology. This partnership helps refine the techniques used to determine ancestral diet and feeding patterns. For example, field studies of the feeding ecology of four sympatric Sumatran primates allowed Ungar (1996) to test whether incisor size corresponds with diet. His finding that incisor size and shape best reflect the mode of ingestion rather than diet *per se* resolves some conflicts in interpretation of the fossil record. Dental microwear analyses can reveal more detailed information about diet than can analysis of gross dental morphology. Recent field studies examine primate feeding behavior, measure the physical properties of foods, and cast the teeth of wild primates in order to better understand the relationship between food type, mode of ingestion, and microwear (Teaford 1994). Similar studies are being conducted on humans (Ungar and Spencer 1999). With further refinement of this new technique, analysis of dental microwear can play an important role in the reconstruction not only of the diet of ancestral hominines, but the use of teeth for non-feeding functions as well (Teaford 1994).

Studies of positional behavior investigate the relationship between postcranial musculoskeletal anatomy and locomotion. The means by which wild great apes travel in their natural environment reveal the function of bone and muscle in the limbs (Fleagle 1999), and comparative anatomy between great apes and fossil hominines provides a basis for inferring locomotor behavior in hominines. This approach has shown that bipedal hominines descended from an arboreal primate that moved primarily by climbing rather than brachiating (Fleagle 1999). Likewise, *Australopithecus afarensis*, the hominin most likely responsible for the famous Laetoli footprints (Leakey 1987), probably spent a good proportion of the day moving arboreally as well as bipedally (Stern and Susman 1983). Field observations also provide clues

to possible sex differences in behavior. In apes, sexual dimorphism corresponds with differential use of arboreal and terrestrial substrates for moving, with males spending more time on the ground than females (Doran 1996). If the sexually dimorphic *A. afarensis* moved on the ground and in the trees, sex differences in substrate use may have likewise occurred.

3.3 Intelligence/Cognition

The comparative study of intelligence and cognition among nonhuman primates in the laboratory spans over half a century (Russon et al. 1996). The discovery of tool manufacture and use in wild chimpanzees (Goodall 1968) put to rest concerns that the development of the cognitive capacities necessary to use tools by laboratory primates was an artifact of intensive interaction with humans. Wild orangutans also manufacture and use tools with considerable complexity (Fox et al. 1999). In contrast, the proficiency with which capuchin monkeys use tools in both field and laboratory settings is significantly lower than that of great apes (Visalberghi and Trinca 1989). Great apes have crossed the line into a domain of intelligence once postulated to be uniquely human; furthermore, the intelligence required for ape-level tool use and manufacture evolved before the orangutan–African ape split (van Schaik et al. 1996).

Tool use behavior, however, is only one demonstration of intelligence. The decision-making processes required for large-bodied apes to travel arboreally is another (Povinelli and Cant 1995). Complexity in social interactions such as deception evinces cognition as well (Byrne and Whiten 1988). Primate studies provide empirical proof that many purported aspects of human uniqueness, such as intelligence and culture, are unique to humans only insofar as they are defined as such (Cartmill 1990). The corpus of work in this field demonstrates that, broadly defined, intelligence has evolved by degrees rather than as a result of a single selective pressure on early hominids. Diverse levels and types of cognitive capacities in an array of primates elicit two questions about humans: (a) is the higher intelligence demonstrated in humans an added *degree* of intelligence or is it a difference of *kind*, or both? and (b) what selected for human-level intelligence? Language is an obvious difference between humans and nonhuman primates, and it will undoubtedly figure prominently in answering these questions.

4. Future Directions in Research

Although most primates live in groups, a group is comprised of individuals with interests that often differ from those of their group-mates. In addition, those interests change as individuals grow to maturity, reproduce, and then, in some species, shift to a post-

reproductive phase of life. Building upon earlier research, future studies are needed to investigate the complexity of factors that influence individual decisions within the context of group life, in particular: (a) the strategies that individuals employ to satisfy their reproductive interests, even when those interests conflict with other group members, and (b) reproductive decisions that incorporate not just matings or fertilizations, but offspring survival as well.

Male aggression or threat of aggression toward females is observed among most species of primates, including humans (Smuts 1992, Smuts and Smuts 1993). Some of this conflict is with current or potential mating partners. However, significant variation in rates of sexual aggression occurs between and within species. For example, woolly spider monkey females copulate without aggression from their mates or other males, while male sexual aggression in rhesus macaques occurs frequently. In baboons, particular males develop 'friendships' with females while others do not (Smuts and Smuts 1993). There is even variation in individual males: some human males with pair-bonded mates force copulations with other females, and some pair-bonded males alternately exhibit affiliative and aggressive behavior toward their mate (Ellis 1989). Despite the seemingly widespread occurrence among primate species, however, extreme forms of direct sexual aggression in the form of forced copulations are uncommon in primates other than humans and orangutans. Some hypotheses to explain this striking contrast have been proposed and deserve further exploration.

What is the function of male sexual aggression? Under what conditions does it occur? Feminist scholars and biological and social scientists have long recognized that conflict between female and male reproductive interests influence individual behavior and shape the social systems in which individuals live (Hrdy 1977, MacKinnon 1989). Theoretical treatments of intersexual mating conflict in human and nonhuman primates provide a basis for predicting the conditions under which male sexual aggression, and conversely female countertactics to thwart males, occur (Smuts 1992, Smuts and Smuts 1993). Recent studies provide some preliminary basis for comparison (Smuts and Smuts 1993), but to strengthen the comparative method more studies conducted across a greater spectrum of primates are necessary. Of particular importance is quantifying the costs of sexual coercion, such as reduced foraging efficiency, in addition to infant or female mortality. New non-invasive hormonal assay techniques for field-collected samples are starting to be used (Whitten et al. 1998) and will allow researchers to measure endocrine response by females who are subject to male sexual aggression. Using these techniques, research can reveal the degree to which male aggression produces physiological effects that may negatively impact female fecundity.

Little is understood about why females resist mating by some males but not others. Equally little is known about female tactics to circumvent male attempts to restrict female reproductive choice. More studies are needed to determine the degree to which females select mates based on variation in quality among males (direct mate choice) or whether other factors, including male aggression, sufficiently constrain females to result in facultative, or indirect, female mate choice (Gowaty 1992, Wiley and Poston 1996). Whether for genetic quality, the quality of services, or constraints placed by males, with whom a female mates has an important reproductive outcome. Females are therefore predicted to exercise tactics, including sperm selection, that optimize their ability to retain control over reproductive choice and the timing of reproduction. Techniques for genetic analysis using non-invasive sampling methods have recently been developed (Borries et al. 1999) and will in the future facilitate evaluation of the success of both male and female reproductive strategies.

Variability among male reproductive decisions is likewise important. In contrast to the previously pervasive view within evolutionary studies and behavioral ecology, a recent shift recognizes that the acquisition of multiple mates is not the sole factor motivating male reproductive decisions. Whether they are excluded by other males, fail to attract mates, or are constrained by the need to allocate time to nonmating activities, male primates are limited in the number of females with whom they can actually mate and produce successful young. Males therefore optimize reproductive effort within operating constraints. While limitations on male mating opportunities have traditionally been viewed in terms of male-male competition and male mating aggression toward females, this is not always the case. Male choosiness for mates is likely to be one important tactic in species that have long periods of infant and juvenile dependency (Altmann 1997). Differences in female fertility, for example, may guide males away from young females as mates. Furthermore, in primates conception and birth are merely the first steps to reproductive success. Offspring must survive into adulthood in order to reproduce, since reproductive success is most accurately measured in the second generation (Clutton-Brock 1988). Offspring survival is critically dependent on maternal competence, including mothers' ability to obtain access to resources necessary to rear young. Males are therefore predicted to prefer as mates those females who are higher ranking, older, and experienced (Altmann 1997).

Females with insufficient resources to rear young may abandon current offspring in order to invest in future reproductive opportunities. Alternatively, in order to meet the demands of rearing offspring, mothers in some species share the task of offspring care. Fathers sometimes significantly assist females, and the need for paternal care constrains male prom-

isculity in some primates as it does in many species of birds (Clutton-Brock 1991). Likewise, related or unrelated individuals sometimes assist mothers. Related individuals gain inclusive fitness benefits resulting from the survival of grandchildren, sisters, and cousins, while both related and unrelated individuals may gain needed experience necessary for the survival of their future offspring (Hrды 1999). Long juvenile dependency and the need for assistance in rearing offspring play an important role in the lives of extant hunter-gatherers, and it figures prominently in models of human social evolution (O'Connell et al. 1999). The function of extended juvenile dependency, how mothers meet the demands of extensive infant care, and the consequences for reproductive success are among the most fascinating questions of human and nonhuman primate life-history. Yet because primate life spans are so long, these questions are also among the hardest to study. As data from uninterrupted long-term field studies become available, questions concerning the fitness advantages to variation in parental and alloparental care can be addressed.

See also: Comparative Neuroscience; Comparative Psychology; Primate Families, Structure of; Primate Socioecology; Primates, Evolution of; Primates, Social Behavior of; Recognition Memory (in Primates), Neural Basis of

Bibliography

- Altmann J 1974 Observational study of behavior: Sampling methods. *Behaviour* **49**: 227–67
- Altmann J 1980 *Baboons Mothers and Infants*. Harvard University Press, Cambridge, MA
- Altmann J 1997 Mate choice and intrasexual reproductive competition: Contributions to reproduction that go beyond acquiring more mates. In: Gowaty P A (ed.) *Feminism and Evolutionary Biology*. Chapman and Hall, New York
- Andelmann S J 1986 Ecological social determinants of cercopithecine mating patterns. In: Rubenstein D I, Wrangham R W (eds.) *Ecological Aspects of Social Evolution: Birds and Mammals*. Princeton University Press, Princeton, NJ
- Aureli F 1992 Post-conflict behaviour among wild long-tailed macaques (*Macaca fascicularis*). *Behavioral Ecology and Sociobiology* **31**: 329–37
- Borries C, Laundhardt K, Epplen C, Epplen J T, Winkler P 1999 DNA analyses support the hypothesis that infanticide is adaptive in langur monkeys. *Proceedings of the Royal Society London* **266**: 901–4
- Byrne R W, Whiten A (eds.) 1988 *Machiavellian Intelligence: Social Expertise and the Evolution of Intellect in Monkeys, Apes and Humans*. Clarendon Press, Oxford, UK
- Caro T M, Borgerhoff Mulder M 1987 The problem of adaptation in the study of human behavior. *Ethology and Sociobiology* **8**: 61–72
- Cartmill M 1990 Human uniqueness and theoretical context in paleoanthropology. *International Journal of Primatology* **11**: 173–92
- Cartmill M 1993 *A View to a Death in the Morning: Hunting and Nature Through History*. Harvard University Press, Cambridge, MA
- Clutton-Brock T H (ed.) 1988 *Reproductive Success: Studies of Individual Variation in Contrasting Breeding Systems*. University of Chicago Press, Chicago, IL
- Clutton-Brock T H 1991 *The Evolution of Parental Care*. Princeton University Press, Princeton, NJ
- Dart R A 1957 The osteodontokeratic culture of *Australopithecus prometheus*. *Transvaal Museum (Pretoria) Memoir* **10**: 1–105
- DeVore I, Washburn S L 1963 Baboon ecology and human evolution. In: Howell F C, Bourlière F (eds.) *African Ecology and Human Evolution*. Aldine, Chicago, IL
- Doran D M 1996 Comparative positional behavior of the African apes. In: McGrew W C, Marchant L F, Nishida T (eds.) *Great Ape Societies*. Cambridge University Press, Cambridge, UK
- Ellis L 1989 *Theories of Rape: Inquiries into the Causes of Sexual Aggression*. Hemisphere, New York, NY
- Fleagle J G 1999 *Primate Adaptation and Evolution*, 2nd edn. Academic Press, San Diego, CA
- Fox E A, Sitompul A F, van Schaik C P 1999 Intelligent tool use in wild Sumatran orangutans. In: Parker S T, Miles H L, Mitchell R W (eds.) *The Mentalities of Gorillas and Orangutans*. Cambridge University Press, Cambridge, UK
- Goodall J v L 1968 The behavior of free-living chimpanzees in the Gombe Stream Reserve. *Animal Behaviour Monographs* **1**: 161–311
- Gowaty P A 1992 Evolutionary biology and feminism. *Human Nature* **3**: 217–49
- Hamburg D A, McCown E R (eds.) 1979 *The Great Apes*. Benjamin/Cummings, Menlo Park, CA
- Hawkes K, O'Connell J F, Blurton Jones N G 1997 Hadza women's time allocation, offspring provisioning, and the evolution of long post-menopausal life spans. *Current Anthropology* **38**: 551–77
- Hrды S B 1977 *The Langurs of Abu: Female and Male Strategies of Reproduction*. Harvard University Press, Cambridge, MA
- Hrды S B 1999 *Mother Nature*. Pantheon, New York
- Huxley T H 1863 *Evidence as to Man's Place in Nature*. Williams and Norgate, London
- Jolly C J 1970 The seed eaters: A new model of hominid differentiation based on a baboon analogy. *Man* **5**: 5–26
- Kinzey W (ed.) 1987 *The Evolution of Human Behavior: Primate Models*. State University of New York Press, Albany, NY
- Krebs, J R, Davies N B 1991 *Behavioural Ecology*. Blackwell Scientific, Oxford, UK
- Leakey M D 1987 Introduction (to the hominid footprints). In: Leakey M D, Harris J M (eds.) *Laetoli: A Pliocene Site in Northern Tanzania*. Clarendon, Oxford, UK
- MacKinnon C A 1989 *Toward a Feminist Theory of the State*. Harvard University Press, Cambridge, MA
- O'Connell J F, Hawkes K, Blurton Jones N G 1999 Grandmothering and the evolution of *Homo erectus*. *Journal of Human Evolution* **36**: 461–85
- Povinelli D J, Cant J G H 1995 Arboreal clambering and the evolution of self-conception. *Quarterly Review of Biology* **70**: 393–421
- Russon A E, Bard K A, Parker S T (eds.) 1996 *Reaching into Thought: The Minds of Great Apes*. Cambridge University Press, Cambridge, UK
- Schaik C P van 1989 The ecology of social relationships amongst female primates. In: Standen V, Foley R A (eds.) *Comparative*

- Socioecology: The Behavioural Ecology of Humans and other Mammals*. Blackwell Scientific, Oxford, UK
- Schaik C P van, Fox E A, Sitompul A F 1996 Manufacture and use of tools in wild Sumatran orangutans. *Naturwissenschaften* **83**: 186–8
- Sibley C G, Alquist J E 1987 DNA hybridization evidence of hominoid phylogeny: results from an expanded data set. *Journal of Human Evolution* **26**: 99–121
- Smuts B B 1992 Male aggression against women: an evolutionary perspective. *Human Nature* **3**: 1–44
- Smuts B B, Smuts R W 1993 Male aggression and sexual coercion of females in nonhuman primates and other mammals: evidence and theoretical implications. *Advances in the Study of Behavior* **22**: 1–63
- Stern J T, Susman R L 1983 The locomotor anatomy of *Australopithecus afarensis*. *American Journal of Physical Anthropology* **53**: 69–80
- Sunday S R, Tobach E (eds.) 1985 *Violence Against Women: A Critique of the Sociobiology of Rape*. Gordian Press, New York
- Teaford M F 1994 Dental microwear and dental function. *Evolutionary Anthropology* **3**: 17–30
- Ungar P S 1996 Relationship of incisor size to diet and anterior tooth use in sympatric Sumatran anthropoids. *American Journal of Primatology* **38**: 145–56
- Ungar P S, Spencer M A 1999 Incisor microwear, diet, and tooth use in three Amerindian populations. *American Journal of Physical Anthropology* **109**: 387–96
- Visalberghi E, Trinca L 1989 Tool use in capuchin monkeys: distinguishing between performing and understanding. *Primates* **30**: 511–21
- Waal F de 1983 *Chimpanzee Politics: Power and Sex Among Apes*. Harper and Row, New York
- Whitten P L, Brockman D K, Stavisky R C 1998 Recent advances in noninvasive techniques to monitor hormone-behavior interactions. *Yearbook of Physical Anthropology* **41**: 1–24
- Wiley R H, Poston J 1996 Indirect mate choice, competition for mates, and coevolution of the sexes. *Evolution* **50**: 1371–81
- Wilson E O 1975 *Sociobiology*. Harvard University Press, Cambridge, MA
- Winterhalder B, Smith E A 2000 Analyzing adaptive strategies: Human behavioral ecology at twenty-five. *Evolutionary Anthropology* **9**: 51–72
- Wrangham R W 1980 An ecological model of female-bonded primate groups. *Behaviour* **15**: 262–300

E. A. Fox and J. Altmann

Primate Families, Structure of

We often talk about animal ‘families’ when referring to the breeding units of animals. The word ‘family’ however, is derived from the Latin word for ‘household’ in which a man controls the resources of his wife and offspring. Such entities seem absent in non-human primates. In contrast, most female primates and their

dependent offspring forage largely independently of the males. Furthermore, mating and breeding occur in a large variety of ways. Biologists use the term ‘mating system’ to describe the patterns of behavior related to mate choice and reproduction. Such mating systems are intimately linked to the social system in which the animals live, but these systems are not identical. Mating partners include, but need not be restricted to, social partners. A key characteristic of female primates is their monthly ovulation. Breeding seasons do exist in many primates, but overall periods of female fertility are more frequent than in other mammals. Accordingly, sexual behavior is frequent and has acquired additional functions in many primates.

1. Diversity of Primates and Their Mating Systems

1.1 Types of Mating Systems

Primates include two hundred species ranging in size from mouse-sized prosimians and pigmy marmosets weighing less than 100 g to orangutans and gorillas with males weighing over 200 kg. Most primates are tree-dwelling but some, such as gorillas and baboons, are terrestrial. While some of the prosimians (a group of relatively ‘primitive’ primates such as galagos, pottos, lorises, and lemurs) are nocturnal and mostly solitary, the vast majority of primates are diurnal and extremely social. In fact sociality and sexuality are key primate adaptations. Primate mating systems represent a wide spectrum ranging from polyandry, where a single female mates with several males (South American tamarins), to monogamous pairs (owl monkeys, gibbons), to multi-male, multi-female groups (chimpanzees, spider monkeys, red colobus monkeys, and baboons) to harem type groups where a single male has exclusive reproductive access to multiple females (gorillas and langurs). Single-male groups can experience temporal visits by groups of bachelor males (Guenons) and solitary males and females can temporarily gather (orangutans). Size differences between males and females (called sexual dimorphism) seem to correlate with the degree to which males of a given species are polygamous. The males and females of the (serially) monogamous gibbons are the same size but males of the polygamous gorillas and mandrills can weigh more than twice as much as females. Primate mating systems are far more flexible and dynamic than previously thought. The same species may show different mating systems in different locations or at different times in the same location (Dunbar 1988). Changes in resource distribution or social changes can cause animals to alter the ways in which they find and choose mating partners.

1.2 Reproductive Costs for Male and Female Primates

Sexual reproduction refers to the process by which male and female germ cells fuse and produce individuals with unique combinations of genes. The germ cells carry half of the reshuffled genes that the parents inherited from the grandparents. Female germ cells (ova) are large and costly to produce whereas male germ cells (sperm) are tiny and much cheaper to produce. In animals with internal fertilization and gestation, this discrepancy in reproductive effort is exaggerated. Female mammals bear an even more disproportionate cost of reproduction because of lactation (the physiological cost of which is higher than gestation itself). Male investment often ends with successful insemination, after which males can look for further mates. The average number of offspring born to male and female species is roughly the same for both sexes, but the variance is much higher in males, where some highly successful males sire large numbers of offspring and many others none at all.

Biologists have repeatedly pointed at the imbalance between male and female reproductive costs to explain why males have a natural tendency to be promiscuous while females tend to be coy and choosier (Hrdy & Williams 1983). This caricature of male emphasis on quantity and female emphasis on quality is in part responsible for the fact that students of animal behavior have long overlooked the importance of female choice and multiple paternity (Gowaty 1992). Females, much rather than just being passive resources, for which males compete, are pursuing their own reproductive interests very actively. Their strategies are often less conspicuous and may involve long-term patterns, which explains why they have so often been overlooked.

Male primates can undoubtedly increase their lifetime reproductive success by maximizing the number of matings with fertile females. In primates, where mating often has more functions than simple fertilization, there may be strong advantages to multiple matings for females as well. Primates have long periods of development and infants are very dependent on parental assistance. Parental care is crucial to primate reproduction and is mostly the responsibility of mothers, though, in many cases males are also involved. The tamarins of the Amazon forest even have cooperative paternal care for twins born to the female. The young of many primates benefit from the protection and care from males with whom their mothers associate. Infant baboons and gibbons are often carried by males and chimpanzee males are known to adopt orphaned male youngsters. Wherever male help with the rearing of offspring is needed and there are several males present in the group, uncertainty of paternity can cause males to assist offspring that they did not sire, resulting in a strong motivation for females to mate with many different males. Too much

uncertainty about paternity is potentially dangerous, as males can be a threat to offspring that they perceive as not their own.

1.3 Sex as a Social Tool

In bonobos, apes known for their frequent sexual interactions, sex is directly used to reduce tension in the group and is often exchanged for food. Chimpanzee females often exchange sex for meat that tends to be controlled by males, which do most of the hunting and use prey as a social currency. A female chimpanzee will copulate more than six thousand times with a very large number of males (10 to 30) while she can only produce a maximum of six surviving offspring. Similarly, the high frequency of same-sex sexual contact in bonobos plays important, entirely social roles. In fact, bonobos appear to be the one species of primates, where frequent female–female sex facilitates the existence of strong female alliances which are strong enough to keep male aggression in check.

1.4 Sexual Advertising in Multi-male Groups

Females in many species with multi-male groups exhibit conspicuous pink swellings of their genital area during several days around the time of their ovulation. These swellings are known as estrous swellings and they represent a formidable ‘tool’ for female primates. A female with a swelling becomes the focus of attention of all males in the group. Her social status increases and males will compete for access to her. Several females can have estrous swellings at the same time, thus making it virtually impossible for any given male to monopolize access to all of them. Estrous swellings appear to play an important role for species where females disperse and join new social groups after they become sexually mature. Swellings are a kind of ‘social passport’ which greatly facilitates a female’s acceptance by unfamiliar males. There is evidence for certain amounts of disconnect between the advertising of fertility and actual time of ovulation. Such a disconnect could work to the advantage of the female because she could escape the controlling efforts of certain males and exert a freer choice of paternity for her offspring. The topic of ‘concealed ovulation’ takes on considerable importance as many scientists have suggested that humans have attained a stage of perfectly concealed ovulation (concealed even from a large proportion of women themselves). To what degree the inability of human males to detect the moment of peak fertility in women has played a role in leading to the high frequency of (serial) monogamy in our species is a much-debated issue.

Some primates use vocalization as a way to attract mates. Notable cases are the orangutan, gibbons, and

siamangs. In the orangutan, an interesting case of multiple male mating strategy includes sexually mature but small males, that do not sing to attract females but rather rely on forced matings between several males and a very reluctant female. Female criteria for choosing their mates remain poorly understood. Females can certainly learn about future social partners through their yearlong interactions with males, but for detecting 'good genes' they must rely on indirect signs such as body symmetry and relative absence of disease and parasites.

2. Determinants of Sociality in Primates

2.1 The Need for Social Grouping

The distribution of resources and the presence of predators play key roles for the distribution of individuals across space. When resources are clumped, groups of animals can live around a particular cluster of resources and defend a certain territory. Such groups can form cohesive social units. When resources are few and very dispersed, no single cluster of resources can be defended even by large groups of animals. Such environments tend to harbor animals living in much smaller groups or as dispersed individuals. Even highly social populations of chimpanzees are known to scatter into much smaller sub-groups during periods of low food abundance. The solitary orangutans on the other hand can gather in large numbers during periods of plenty (mast fruiting).

There are many good reasons for living socially. More eyes can see better and more ears can hear better. Living socially facilitates the search for food and the detection of predators. Another advantage is the phenomenon of the 'selfish herd' where the presence of other individuals spreads the risk of predation, as any of the others is as likely to be preyed upon as oneself. Also, cooperation in anti-predator defense is only possible in groups. Obviously, social life facilitates finding mates, even if it will increase the competition for them. Animals in groups can cooperate in defending resources against competitors. Such cooperative resource defense, when done by males, may have the added benefit of attracting females to the group, which may represent unwelcome competition to the females already resident in the group. Finally, the presence of individuals possessing different capabilities (such as hunting, raiding beehives, remembering the time and place certain foods can be harvested) in a group will benefit all group members. Unique individual experience and history shape large-brained, long-lived organisms such as primates. The mating strategies within a social group can vary dramatically between individuals. Some males, for example, may not show any ambition with respect to the dominance hierarchy. They spend their

whole life as subordinate males. Their affiliative behavior and friendship with certain females in the group may in the end translate into numerous offspring. Certain females may be much more successful at forming social bonds with dominant males. Sheer force can allow a male to ascend in the dominance hierarchy, but the majority of animals in his group may not follow him. The cognitive capacities of primates appear to have coevolved with their tendency for sociality. Long-lasting relationships with many group members require good memory and social intelligence.

2.2 Costs of Living Socially

There are multiple costs to sociality. Competition between males can create social instability and conflict. Such conflict is often violent and can lead to harm to bystanders or even infanticide by males hoping to remove the offspring of competitors and gain rapid access to fertile females (the loss of an infant causes lactating females to resume ovulatory cycles). Competition occurs not only between males for access to females, but also between females for access to males, access that may lead to support (shared resources) and male protection during social conflicts. Competition between female may be the factor that precludes the formation of strong female coalitions against males in many species. Infanticide perpetuated by females has been observed in some wild primate groups and may represent the darkest side of female-female competition for resources. The study of female dominance hierarchies in natural primate populations is still in its infancy. Such hierarchies tend to be more subtle, nevertheless, female social rank can play an important role for reproductive success and change the way females invest in sons or daughters. Finally, higher risk of infection by pathogens and parasites is an important cost to sociality.

2.3 Most Primate Groups are Female Bound

In most primates, females and their female relatives remain in their native groups while most males disperse to other groups when they reach sexual maturity. Some notable exceptions to female residence are found in colobus and spider monkeys and in our closest living relatives, chimpanzees and bonobos, where females disperse upon reaching sexual maturity and males remain in their natal communities. The driving force behind dispersal is thought to be inbreeding avoidance and/or competition with animals of the same sex (Moore & Rauf 1984). Sexual mimicry in white-cheeked gibbons, where adolescent females look like males, and in red colobus, where young males have pseudo-swells, indicates that same-sex competition is important in those species. The mechanisms

underlying inbreeding avoidance are poorly understood. Whether primates are able to recognize their paternal kin by features inherited from the father or whether they simply avoid mating with individuals from the same age cohort is still unclear (Alberts 1999). In humans, close proximity at an early age seems to be the crucial mechanism that triggers avoidance as sexual partners.

2.4 Do Males Follow Female Spacing?

The way females deploy themselves across time and space tends to determine how males can space themselves around females. Only if the ecosystem is capable of feeding a large number of females with enough tolerance for each other can males attempt to control such groups, either as the solitary reproductive male in a harem or as a group of males cooperating in female and territory defense and sharing the reproductive access to those females. All multi-female groups create the potential for female–female competition over access to resources and access to males and the support they may provide. Competition between potential mates may create the opportunity to enhance the visibility of positive traits or weaknesses. Because competition involves stress and stress is known to impact an animal's immune system, competition can exaggerate the visible differences in male quality. If female choice incites increased male competition, then female choice may indirectly lead to an increased size difference between males and females. Because strong and assertive males can be more successful at guarding resources that they share with females, females may favor the evolution of such males. Females need male protection from unwanted attention of other males, including, in the extreme case, forced copulation, as well as against the threat of male infanticide; both are part of male mating strategy in many primate species (Smuts 1993). Strong and protective males by the same token will be more successful at restricting the independence of females.

3. The Importance of Genes

3.1 The Benefit of Multiple Paternity to Females

Mating with multiple males can be advantageous for a female in various ways. As mentioned above, confusing paternity of her offspring can effectively protect them from male infanticide. She may gain support from her mating partners, both in terms of social support during conflicts with other group members and in terms of material support in the form of shared resources. More recently, the possibility of a genetic benefit has been proposed and some evidence is starting to accumulate (Zeh & Zeh 1996, 1997).

Because the primate fetus develops inside the mother, the female and the fetus need to achieve a type of 'immunological détente.' That is, the fetus must not be rejected by the mother's powerful immune system but, at the same time, the demands of the fetus must be kept in check, not allowing it to exact too high a toll on maternal resources. Multiple paternity will produce more diverse offspring. Such diversity may have much to do with defense against disease. The vertebrate immune system is an immensely sophisticated and very costly system, in order to allow long-lived organisms such as primates to survive in a world in which very short-lived, rapidly evolving parasites and pathogens (ranging from worms and protozoa to bacteria and viruses) are a constant threat.

The immune system relies on genetic variation present in all populations, as well as generated anew in each individual in order to produce a flexible repertoire of molecular probes (antibodies and T-cell receptors) which detect specific intruding pathogens. It seems critical that each fetus inherits genes coding for two rather different immunological 'tool kits' in order for the offspring to be well equipped for its countless future encounters with pathogens. Conceiving offspring with different fathers would allow females to produce sets of offspring with more variation in their immune repertoires, strongly reducing the risk of losing all or most offspring to the same disease. The same obviously applies to males as well. Furthermore, paternal and maternal genes have to work in harmony in order to produce healthy offspring. Some combinations may simply fail to work and yet others may work but put the offspring at a disadvantage from the beginning. The detection of embryos with strongly limited immune repertoires (i.e., two almost identical sets of multi-histocompatibility complex or 'MHC' alleles) inherited from both parents, can cause spontaneous abortions in many mammal species including our own. Because of the importance of compatibility between genomes, the most promising genes for each female are found in different males. Finally, multiple paternity may allow females to reduce inbreeding where females cannot recognize close kin.

3.2 Molecules and Behavior

Ova and sperm engage in intense molecular dialogue, a process which is important for creating reproductive barriers between species but which is also suspected to involve an additional level of (post-mating) female choice. Female primates are known to produce a large quantity of antibodies against sperm in their genital tracts, a process which strongly favors diversity in paternity as the antibodies would be formed against the sperm of the 'usual' mating partner leaving sperm from unknown males at an advantage. The role of chemoreception (pheromones) in regulating some primate behavior remains very poorly understood.

The fact that primates may be able to smell important immune molecules (MHC) points towards potentially important roles of such molecules and their detection in mate choice and sexual behavior (Wedekind & Furi 1997).

3.3 Post-mating Female Choice and Sperm Competition

Certain sperm may be more likely to be taken up by the egg as they approach it. Such cryptic female choice would represent a choice of individual sperm, possibly based on the genes carried by that sperm. Certain genes of crucial importance in early embryonic development are known to be expressed on the surface of sperm. Males can compete for fertilization of females via a process called 'sperm competition.' When females mate with many males within a short period of time, the sperm of different males will be available for fertilizing her egg. This exerts pressure on males to produce larger quantities of sperm in order to succeed in preference to other males. In chimpanzees, where multiple matings with many males give rise to strong sperm competition, males have conspicuously large testicles and produce large amounts of sperm. They even produce 'copulatory plugs,' a gelatinous mass blocking the entry to the uterus. In gorillas, on the other hand, where silver back males face very little competition in their harems, males have very small testicles. Female-female competition may take place via the depletion of sperm supply for other females.

3.4 Genomic Conflicts

Because different males can sire subsequent offspring, a conflict of interest can arise between the paternal and the maternal genes. A female will spread her total reproductive efforts across all her subsequent offspring. Males are likely to have sired offspring with other females and thus are spreading their reproductive effort across several females. Paternal genes in the embryo are likely to cause each embryo to be over-demanding from the point of view of the mother. Such paternal genes will increase the male's reproductive success at a detriment to each female, who will have invested too much in that particular offspring. It is this conflict which seems to be at the root of a phenomenon called 'imprinting.' Imprinting occurs when, among the two variants of each gene present in the embryo (the paternal and maternal alleles), only one of the two is expressed. Often, only the allele inherited from the mother is expressed, while the paternal allele appears to be silenced. Imprinting may have several functions but one of them may be female control of over-demanding genes from the father (Mochizuki et al. 1996, Spencer et al. 1998).

Primate mating systems represent the outcome of a delicate balance between various constantly changing forces. These include:

- (a) ecological factors such as the distribution of resources and pressure by predators;
- (b) female foraging and mating strategies, including the need to find multiple fathers of high genetic quality and good genetic compatibility and one or more supportive social partners;
- (c) male mating strategies that may entail the access to the largest number of fertile females often in severe competition with other males; and,
- (d) the need for parental care of infants and juveniles.

Compromises between male and female strategies seem to occur at every level, from gene expression to behavior. The existence of life-long friendships between adult primates of both sexes introduces an additional dimension, that of emotional attachment and trust between individuals.

4. Future Directions

Long-term field studies of primates in their natural habitat are being complemented by powerful molecular methods of analysis. Genetic analyses are providing detailed information on pedigrees and kinship, while hormone studies can provide knowledge of female reproductive state. Studies of hormones, such as oxytocin and vasopressin and their receptors, which seem to have dramatic effects on male social attachment behavior, are likely to produce new insights (Young et al. 1999). The study of genomic conflicts and imprinting is in its infancy. Similarly, further insight into the determinants of genetic compatibility and the maximizing of immune repertoires will produce much badly needed knowledge. The large number of longitudinal studies of natural primate communities under way is bound to reveal additional ways in which males and female primates strike compromises and successfully reproduce in ever-changing environments. Understanding mating systems is not only essential if one wants to understand evolution, because many primates are facing extinction, such an understanding is also crucial for conservation efforts in the wild and in captivity.

See also: Group Processes, Social Psychology of; Primate Behavior: Significance for Understanding Humans; Primates, Social Behavior of

Bibliography

- Alberts S C 1999 Paternal kin discrimination in wild baboons. *Proceedings of the Royal Society of London, Biology* **266**: 1501-6
- Dunbar R I M 1988 *Primate Social Systems*. Cornell University Press, Ithaca, New York

- Gowaty P A 1992 Evolutionary biology and feminism. *Human Nature* 3: 217–49
- Hrdy S B, Williams G C 1983 Behavioral biology and the double standard. In: Wasser S K (ed.) *Social Behavior of Female Vertebrates*. Academic Press, New York, pp. 3–17
- Mochizuki A, Takeda Y, Iwasa Y 1996 The evolution of genomic imprinting. *Genetics* 144: 1283–95
- Moore J J, Rauf A, Ali R 1984 Are dispersal and inbreeding avoidance related? *Animal Behaviour* 32: 94–112
- Smuts B B 1993 Male aggression and sexual coercion of females in non-human primates and other mammals: Evidence and theoretical implications. In: Slater P J B, Rosenblatt J S, Milinski M, Snowden C T (eds.) *Advances in the Study of Behavior*. Academic Press, New York, pp. 1–63
- Spencer H G, Feldman M W, Clark A G 1998 Genetic conflicts, multiple paternity and the evolution of genomic imprinting. *Genetics* 148: 893–904
- Wedekind C, Furi S 1997 Body odour preferences in men and women: Do they aim for specific MHC combinations or simply heterozygosity? *Proceedings of the Royal Society of London, Biology* 264: 1471–9
- Young L J, Nilsen R, Waymire K G, MacGregor G R, Insel T R 1999 Increased affiliative response to vasopressin in mice expressing the V1a receptor from a monogamous vole. *Nature* 400: 766–8
- Zeh J A, Zeh D W 1996 The evolution of polyandry I: Intra-genomic conflict and genetic incompatibility. *Proceedings of the Royal Society of London, Biology* 263: 1711–7
- Zeh J A, Zeh D W 1997 The evolution of polyandry II: Post-copulatory defences against genetic incompatibility. *Proceedings of the Royal Society of London, Biology* 264: 69–75

P. Gagneux

Primate Socioecology

Sociality is the primary behavioral adaptation of the anthropoid primates, the monkeys and apes. All of the anthropoid primates, except orangutans, spend most of their lives in stable groups of familiar individuals. However, the social organization among the anthropoid primates encompasses great diversity, ranging from the pair-bonded groups of gibbons and owl monkeys, to the polyandrous groups of tamarins; one-male groups of howlers and blue monkeys; multimale, multifemale groups of cebus monkeys and baboons; and fission-fusion communities of chimpanzees and spider monkeys. This great diversity in social organization shapes the behavioral options available to males and females and profoundly influences the evolution of social behavior and reproductive strategies in both sexes.

1. The Evolution of Social Organization

The great diversity in primate social organization has prompted researchers to ask a number of related

questions: Why do primates live in groups? How should groups be structured? How big should groups be? These questions were first raised in the 1960s and 1970s by primatologists who conducted comparative analyses searching for broad ecological correlates of social organization (reviewed by Janson 2000). These comparative efforts yielded some consistent associations between diet, home range size, group size, body weight, and group composition. But the analyses were unsatisfying because they could not account for much of the observed variation in social organization among primates. Moreover, while these studies revealed a number of robust associations among various ecological variables, they did not produce much insight about the causal processes underlying these correlations. Finally, these comparative studies did not consider the possibility that some of the similarities between species were the result of common phylogenetic history, not independent adaptations to ecological conditions, a problem that has become a central methodological concern in comparative studies in behavioral ecology.

In the 1980s researchers returned to these questions, equipped with the theoretical and methodological tools of contemporary behavioral ecology (Krebs and Davies 1993). Behavioral ecologists try to explain how evolutionary processes shape behavior in particular ecological contexts. Sociality is presumed to reflect a dynamic balance between the advantages and disadvantages of living in close proximity to conspecifics. Sociality is beneficial because animals who live in social groups are better able to acquire and control resources or less vulnerable to predators. On the other hand, sociality is disadvantageous because group-living animals may face more competition over access to food, become more vulnerable to disease, and become more susceptible to cuckoldry, infanticide, and cannibalism. The size and demographic composition of the groups that we see in nature reflects a compromise between the costs and benefits of sociality for individuals who occupy particular ecological niches.

1.1 Why do Primates Live in Groups?

For primates, the primary advantages of sociality are thought to be enhanced access to resources, reduced vulnerability to predation, or some combination of both these factors. Many primates feed on foods, such as fruit, that occur in patches large enough to feed several individuals at the same time. Taking note of this fact, Wrangham (1980) suggested that primates aggregate because groups are more successful in defending access to resources than lone individuals, an argument that has also been applied in other taxa. Thus, the potential for between-group competition over resources favors sociality.

van Schaik (1983, 1989) questioned Wrangham's explanation of the evolution of sociality, arguing that the benefits accrued in between-group competition would be offset by the costs incurred from increased within-group competition. Instead, van Schaik argued, grouping evolved as a defense against predators. Primates are preyed upon by a variety of predators, including leopards, raptors, snakes, and other primates. Grouping provides safety from predators for three reasons. First, levels of vigilance are higher because there are more individuals scanning for predators. Second, groups are more effective in actively mobbing or deterring predators. Third, any single individual is less likely to be caught in a group if predators strike at random. Thus, primates aggregate to obtain protection from predators, and within-group competition for food or safe positions is an unavoidable side effect of sociality.

Initial efforts to evaluate the relative importance of between-group competition and predation among primates provided some support for both models (Janson 2000). Thus, the between-group competition model was supported by evidence that hostile intergroup interactions are common among primates, and larger groups are typically able to supplant smaller groups in intergroup encounters. In some populations, large groups are better able to defend higher quality home ranges than smaller groups. However, within-group competition seems to reduce foraging efficiency consistently in gregarious species (Janson 1992).

The predation model was supported by evidence that group sizes are reduced when predators are absent and that members of large groups spend less time scanning for predators and detect predators sooner than members of small groups do (Janson 1992). Nonetheless, vigilance levels do not decline monotonically as group size increases (Janson 2000), suggesting that groups are larger than needed for predator defense.

1.2 *The Nature of Competition and Female Social Relationships*

As researchers continued to debate the role of resource competition and predation in the evolution of sociality, they also began to think more carefully about the effects of competition on primate females. The distribution of food resources determines the distribution of females, and this in turn determines the distribution of males. Females come first in this scenario because female fitness depends mainly on their nutritional status (Wrangham 1980), while male fitness depends primarily on their ability to obtain access to fertile females.

Wrangham (1980) hypothesized that variation in the extent of intergroup competition would produce two distinct kinds of social arrangements. When resources were clumped and abundant enough to feed

more than one individual, females would benefit from banding together to collectively defend access to them. Coalitionary alliances among kin are expected to be more beneficial than alliances among nonkin because kin share a common genetic interest (Hamilton 1964); thus females were expected to form nepotistic alliances. Female philopatry preserves matrilineal networks, and strong affiliate bonds strengthen alliances. Thus, female-bonded kin groups are expected to be associated with between group competition over resources. On the other hand, if resources are not clumped or occur in patches that are too small to feed several individuals, then the collective defense of resources will not be profitable. Nepotistic alliances will not enhance females' access to resources. In this situation, selective pressures favoring female philopatry and strong female bonds will be relaxed. When there is little between-group competition, non-female-bonded groups are expected to occur. Wrangham linked dietary specializations to social organization, arguing that fruit occurs in discrete patches that are often large enough to make collective defense profitable. He predicted that female-bonded groups would be mainly frugivorous.

Although it was rather easy to classify primate species as female-bonded or non-female-bonded, a number of empirical and theoretical objections to Wrangham's model were raised. However, all subsequent efforts to link ecology and social behavior have relied on Wrangham's insight that resource competition among females drives the evolution of social organization among primates.

van Schaik (1989) considered how the distribution of resources influences the nature of competition among females. Scramble competition occurs when individuals cannot effectively monopolize access to resources, and consequently do not compete over them directly. Contest competition occurs when resources are limited and can be monopolized profitably, generating direct confrontations over access to them. van Schaik suggested that within group contest competition (WGC), within group scramble competition (WGS), and between group contest competition (BGC) would have different effects on social organization (Table 1). (Between group scramble competition is assumed to be an automatic consequence of population density, but is not expected to influence social relationships directly).

When females compete over access to resources with group members, access to resources will be a function of individual dominance rank. Van Schaik predicted that WGC would be associated with strict linear dominance hierarchies. Since dominance rank influences female access to resources and ultimately affects their fitness, females will provide coalitionary support to their kin in order to enhance their dominance rank. This will lead to matrilineal dominance hierarchies and create strong selective pressures favoring female philopatry. Nepotistic support will also enhance the

Table 1

Effects of competitive regime on social organization and behavior

	Competitive regime			
	WGC only	WGC + BGC	WGS + BGC	WGS only
Characteristics of dominance hierarchy	Strictly linear	Moderately linear	Egalitarian	Egalitarian
	Nepotistic	Nepotistic	Individualistic	Individualistic
	Very stable	Fairly stable	Unstable	Unstable
Coalitionary support	Nepotistic	Nepotistic in WGC, non-nepotistic in BGC	Non-nepotistic	Absent
Nature of female bonds	Highly differentiated	Highly differentiated	Poorly differentiated	None
Dispersal from natal group	Males only	Males only	Males only	Males/Females
Social organization category ^a	Resident-nepotistic	Resident-nepotistic-tolerant	Resident-egalitarian	Dispersal-egalitarian

a Following Sterck et al. 1997.

stability of dominance hierarchies by dampening the effects of temporary variations in individual fighting ability. Affiliative interactions with coalition partners will reinforce alliances, so females are expected to form close and well-differentiated social relationships.

When WGS is the primary form of resources competition, female reproductive success will depend on the size of the female's social group, not her own rank. This means that there will be little incentive to aid relatives in disputes with other group members, and dominance hierarchies will not be nepotistic or strictly linear. If WGS is combined with strong BGC, females will rely on coalitionary support in intergroup encounters from all group members. Because females benefit more from supporting kin than nonkin, female philopatry will be favored. Since females' fitness depends on collective group action, females are expected to cultivate social ties with all other group members. Thus, females are expected to form relatively poorly differentiated social relationships.

If WGS is associated with weak BGC, then selective pressures favoring female philopatry will be relaxed. Female residence in social groups should reflect the balance between the benefits of living in groups of particular sizes and the costs of emigration from groups of suboptimal size.

1.3 Testing Socioecological Models in the Field

In the 1990s, many researchers set to work testing the predictions derived from socioecological models. This body of work confirmed many of the predictions derived from socioecological models. For example, Sterck et al. (1997) showed that coalitionary support,

linear dominance hierarchies, and female philopatry are linked in a sample of 27 primate taxa. Others have compared closely related species, that differ in key features of their ecology. For example:

(a) Mitchell et al. (1991) compared two closely related species of squirrel monkeys in Costa Rica and Peru that had similar diets, formed groups of similar sizes, and experienced similarly high levels of predation. The two species differed mainly in the extent of within-group competition over access to food resources. As predicted by socioecological models, the species that experienced high within-group competition also formed stable, linear dominance hierarchies, kin-based coalitions, and exhibited strict female philopatry. In the species that experienced low levels of within-group competition, a dominance hierarchy was not detected among females, female coalitions were not observed, and females sometimes left their natal groups.

(b) Barton et al. (1996) compared the behavior and ecology of baboons at two sites. In Laikipia, Kenya, where within-group competition for food and predation pressures were high, groups tended to be large, females formed strong bonds with one another, and females rarely transferred from their natal groups. In the Drakensburg Mountains of South Africa, food was more evenly distributed and predation pressure was relaxed. There, groups were typically smaller, bonds among females were weaker, female dispersal was more common, and bonds between females and males were strengthened.

(c) Isbell and her colleagues have examined the relative importance of phylogeny and ecology in shaping the behavior of female monkeys in two closely related species, patas (*Erythrocebus patas*) and vervets

(*Cercopithecus aethiops*). If phylogeny is the primary factor influencing social organization, then patas and vervets should be very similar. However, if ecological factors are crucial, then patas and vervets should differ because their diets differ. Patas rely more heavily than vervets on arthropods and other food items that require little handling time and are consumed very quickly. On the other hand, vervets rely more heavily than patas on foods, such as gums, seeds, and fruits, that have longer processing times and take longer to eat. Patas have less linear dominance hierarchies than vervets do, and dominance rank has a less pronounced effect on nonagonistic behavior tightly linked to ecology (such as ranging, diet, and activity budgets) among patas than among vervets (Isbell et al. 1999).

(d) Koenig et al. (1998) compared two populations of Hanuman langurs (*Presbytis entellus*) living at different sites in India. The langurs of Ramnagar relied heavily on just three food plants. These plants were low in abundance and clumped in their distribution, conditions that are expected to lead to contest competition. The langurs of Kanha relied on a larger array of plant foods, and one of their most highly preferred foods is superabundant, conditions that are expected to lead to scramble competition. The langurs of Ramnagar behave much as expected when WGC is strong: they have linear dominance hierarchies and females are philopatric. The langurs of Kanha have poorly developed dominance hierarchies and females frequently emigrate from their natal groups.

1.4 Distribution of Males

In socioecological models, males generally play a secondary role. Ecological factors shape the distribution of females, and males go where females are (Kappeler 2000). Social factors are generally thought to be more important than ecological factors in determining the distribution of males because intrasexual competition over access to females contributes more to variation in male reproductive success than intrasexual competition over access to food. When females live in groups, they become a defensible resource for males. The more females there are, the more difficult it is for a single male to monopolize them. This difficulty is compounded if females are widely dispersed or if females become receptive simultaneously. Comparative analyses indicate that the number of males in primate groups is generally correlated with the number of females present and with the extent of synchrony in female receptivity (Mitani et al. 1996, Nunn 1999).

The availability of females explains some, but not all of the variation in the number of males in social groups. This raises the possibility that the distribution of males is also influenced by other factors, including ecological pressures. Comparative analyses suggest

that the distribution of males in social groups is directly linked to predation pressure. Hill and Lee (1998) conducted a systematic survey of the Old World monkeys in an effort to assess the effects of predation risk on the distribution of males. They found that predation risk consistently influenced group size, as groups that faced the highest risks of predator attacks lived in the largest groups. In single-male groups, high predation risk was associated with an increase in the number of females, but in multimale groups predation risk influenced the number of males in social groups, but not the number of females. Thus, the distribution of males may be a joint product of the distribution of females and the intensity of predation pressures.

2. Beyond Socioecology

Socioecological models of primate social organization gained considerable currency in the 1990s. However, in the last few years, socioecological models of social organization have been subjected to a new set of empirical and theoretical challenges (Janson 2000). Not all the observed variation in social organization and behavior fit predictions derived from socioecological models. For example the comparison of squirrel monkeys by Mitchell et al. was complicated by subsequent observations of squirrel monkeys in Suriname (Boinski 1999). There, monkeys face intense WGC, but females have weak bonds with one another and do not form coalitions with other females. Comparative evidence indicates that males generally go where females are, but it has been surprisingly difficult to explain why individual males move from one group to another. Male baboons do not always enhance their access to females when they emigrate (Altmann 2000).

The role of environment in shaping social organization was challenged by new comparative studies which took phylogeny explicitly into account (di Fiore and Rendall 1994). Selected pairwise comparisons of closely related primate species provide support for socioecological models, but these comparisons may not be representative. Most of the characteristics associated with WGC, including linear dominance hierarchies, nepotistic alliances, and female philopatry are found in almost all of the extant Cercopithecae species, even though they now occupy an extremely diverse range of habitats. Di Fiore and Rendall's analysis indicates that there is considerable inertia in social evolution. The social systems categorized by socioecologists may represent different peaks in the adaptive landscape (Pope 2000). If so, this seriously constrains the possible paths for change in social evolution.

Socioecological models assume that ecological factors play a primary role in shaping social organization, but a strong case can be made that demographic and social factors may also shape social systems. A

growing body of evidence suggests that the risk of infanticide may affect primate social systems (van Schaik 2000). There is compelling evidence that infanticide is an evolved reproductive strategy in many animal taxa. In many primate species, males selectively kill infants whose deaths enhance their own reproductive prospects (van Schaik and Janson 2000). The presence of multiple males in social groups throughout the year may be linked to infanticidal risks faced by females (van Schaik 2000).

In summary, present evidence suggests that a broad range of ecological, historical, demographic, and social factors influence the evolution of primate groups. It was probably naïve to expect simple models to explain the full range of variation in primate social organization and behavior. It is, however, essential to begin with simple models before constructing more elaborate ones. The task now is to elaborate and extend socioecological models to accommodate new data and ideas about why primates live in groups.

See also: Communication and Social Psychology; Evolutionary Social Psychology; Group Processes, Social Psychology of; Primate Behavior: Significance for Understanding Humans; Primate Families, Structure of; Primates, Evolution of; Social Psychology; Social Psychology, Theories of

Bibliography

- Altmann J 2000 Models of outcome and process: Predicting the number of males in primate groups. In: Kappeler P M (ed.) *Primate Males*. Cambridge University Press, Cambridge, UK
- Barton R A, Byrne R W, Whiten A 1996 Ecology, feeding competition, and social structure in baboons. *Behavioral Ecology and Sociobiology* **38**: 321–9
- Boinski S 1999 The social organizations of squirrel monkeys: Implications for ecological models of social evolution. *Evolutionary Anthropology: Issues, News, and Reviews* **9**: 101–12
- di Fiore A, Rendall D 1994 Evolution of social organization: A reappraisal for primates by using phylogenetic methods. *Proceedings of the National Academy of Science USA* **91**: 9941–95
- Hamilton W D 1964 The genetical evolution of social behaviour I, II. *Journal of Theoretical Biology* **7**: 1–52
- Hill R A, Lee P C 1998 Predation risk as an influence on group size in cercopithecoïd primates: implications for social structure. *J. Zool. Lond.* **245** (part 4): 447–56
- Isbell L A, Pruettz J D, Lewis M, Young T P 1999 Rank differences in ecological behavior: A comparative study of patas monkeys (*Erythrocebus patas*) and vervets (*Cercopithecus aethiops*). *International Journal of Primatology* **20**: 257–72
- Janson C H 1992 Evolutionary ecology of primate social structure. In: Smith E A, Winterhalder B (eds.) *Evolutionary Ecology and Human Behavior*. Aldine de Gruyter, New York, NY
- Janson C H 2000 Primate socioecology: The end of a golden era. *Evolutionary Anthropology: Issues, News, and Reviews* **9**: 73–86

- Kappeler P M 2000 Primate males: history and theory. In: Kappeler P M (ed.) *Primate Males*. Cambridge University Press, Cambridge, UK
- Koenig A, Beise J, Chalise M K, Ganzhorn J U 1998 When females should contest for food—testing hypotheses about resource density, distribution, size, and quality with Hanuman langurs (*Presbytis entellus*). *Behavioral Ecology and Sociobiology* **42**: 225–237
- Krebs J R, Davies N B 1993 *An Introduction to Behavioural Ecology*, 3rd edn. Blackwell, Oxford, UK
- Mitchell C L, Boinski S, van Schaik C P 1991 Competitive regimes and female bonding in two species of squirrel monkeys (*Saimiri oerstedii* and *S. sciureus*). *Behavioral Ecology and Sociobiology* **28**: 55–60
- Mitani J C, Gros-Louis J, Manson J H 1996 Number of males in primate groups: Comparative tests of competing hypotheses. *American Journal of Primatology* **38**: 315–32
- Nunn C L 1999 The number of males in primate social groups: A comparative test of the socioecological model. *Behavioral Ecology and Sociobiology* **46**: 1–13
- Pope T R 2000 The evolution of male philopatry in neotropical monkeys. In: Kappeler P M (ed.) *Primate Males*. Cambridge University Press, Cambridge, UK
- Sterck E H M, Watts D P, van Schaik C P 1997 The evolution of female social relationships in nonhuman primates. *Behavioral Ecology and Sociobiology* **41**: 291–309
- van Schaik C P 1983 Why are diurnal primates living in groups? *Behaviour* **87**: 120–44
- van Schaik C P 1989 The ecology of social relationships amongst female primates. In: Standon V, Foley R A (eds.) *Comparative Socioecology: the Behavioural Ecology of Humans and Other Mammals*. Blackwell, Oxford, UK
- van Schaik C P 2000 Social counterstrategies against infanticide by males in primates and other mammals. In: Kappeler P M (ed.) *Primate Males*. Cambridge University Press, Cambridge, UK
- van Schaik C P, Janson C H (eds.) 2000 *Infanticide by Males and its Implications*. Cambridge University Press, Cambridge, UK
- Wrangham R W 1980 An ecological model of female-bonded primate groups. *Behaviour* **75**: 262–300

J. B. Silk

Primates, Evolution of

The Primates, one of 20 extant orders of placental mammals, are of special interest because they include the human species (*Homo sapiens*). It should be noted at once that tree shrews, included in the Primates for most of the last hundred years, are now generally thought to have only a distant connection (if at all) and are hence allocated to their own order Scandentia. A recent review (Rowe 1996) portrayed 238 living primate species (excluding tree shrews) and the number is now some 250 and rising, partly reflecting the discovery of entirely new species but also following subdivision of existing species because of new information, notably molecular evidence. This is a dramatic increase over the 150 species listed in a leading

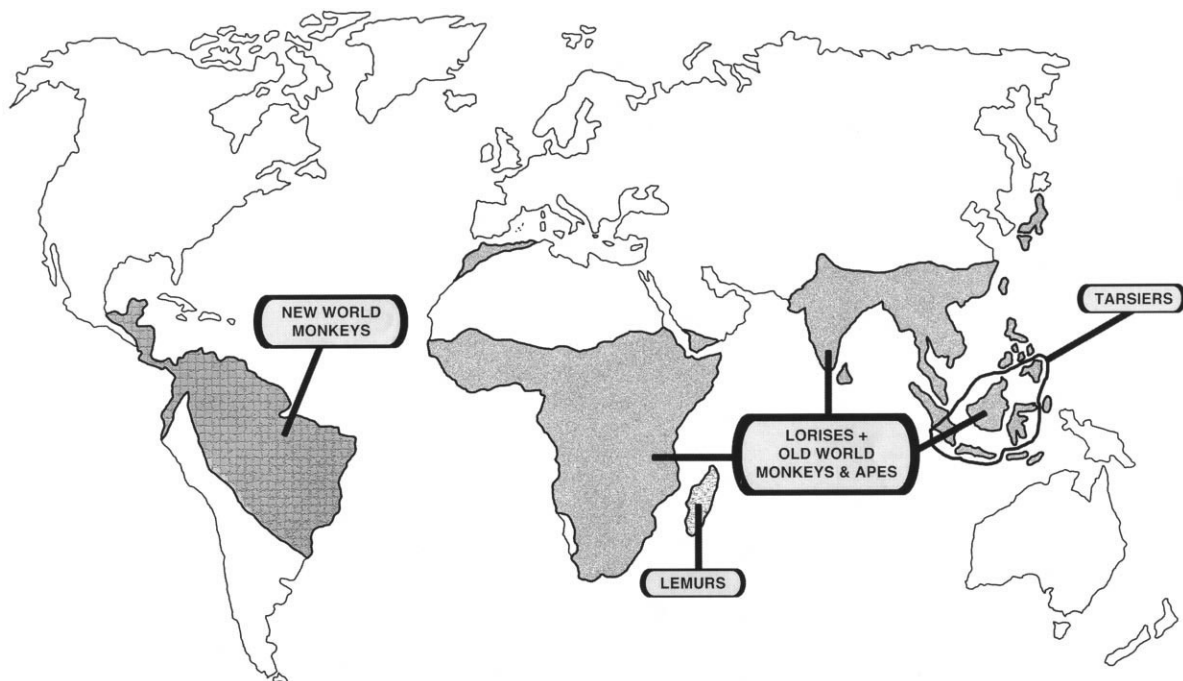


Figure 1

Geographical distribution of living nonhuman primates (excluding tree shrews). Shaded areas indicate the approximate ranges of the five natural groups, with the exception of the loris group (which occupies most but not all of the range occupied by Old World monkeys and apes)

handbook in the 1960s (Napier and Napier 1967). The order Primates also includes almost 450 fossil species (also progressively increasing), documenting the evolution of this group.

The evolutionary history of primates is a striking example of adaptive radiation, in which descendants from an original common ancestor with a particular set of characteristics diverged and adapted to occupy different geographical areas and ecological niches. Following Darwin's *Origin of Species*, this evolutionary radiation was reconstructed initially almost entirely from morphological (anatomical) features, but the comparative framework has been expanded gradually to include physiological, behavioral, ecological, chromosomal, and molecular characters.

1. Living Primates

The logical starting point for reconstruction of primate evolution is the array of living species, partly because recognition of the group arose from the study of living representatives, and partly because many features (notably chromosomal and molecular characters) are not fossilized. Known fossil relatives of living primates can be included subsequently in the overall framework, providing the benefit of a geological timescale.

Later inclusion of fossil species in a preliminary phylogenetic tree based on living forms also permits use of palaeontological evidence to test its validity.

Living nonhuman primates can be divided into five 'natural groups', each with a restricted geographical range (Fig. 1). As geographical isolation plays a fundamental part in evolutionary divergence, these groups are inherently likely to reflect basic subdivisions within the primate tree. Indeed, they are recognizable immediately in all widely accepted phylogenetic reconstructions and classifications. The first such group contains the Madagascar lemurs (infraorder Lemuriformes). These are the only primates present on that island, and document a substantial local radiation (Tattersall 1982). There are at least 35 living species and the modern array also includes at least 17 subfossil species that died out over the last few millennia, following the first human inhabitation of Madagascar. The second group contains the lorises and bushbabies (infraorder Lorisiformes). They are distributed widely in Africa, South Asia and South-east Asia, but the total number of species (about 20) is smaller than for lemurs, representing only a limited radiation. This applies even more drastically to the third group, containing the five closely related tarsier species of the South east Asian archipelago (infraorder Tarsiiformes). The fourth group contains the

monkeys of South and Central America (infraorder Platyrrhini), which are the only New World primates. This is a relatively large group of some 80 species, divided into claw-bearing marmosets and tamarins (family Callitrichidae), and true monkeys (family Cebidae), again documenting a major local radiation (Kinzey 1997). The fifth group, containing the Old World monkeys and apes of Africa, South Asia and South east Asia (infraorder Catarrhini) is the largest, with approximately 105 species. In fact, humans also belong to this group, but have expanded far beyond this original geological range. Within this fifth group, which is most directly relevant to human evolution, there is a clear division between Old World monkeys (superfamily Cercopithecoidea) and apes and humans (superfamily Hominoidea), despite extensive overlap in geographical distribution.

The first three natural groups of primates (lemurs, lorises and bushbabies, tarsiers) have remained relatively primitive and are collectively labeled prosimians (literally 'before the monkeys'). They are comparatively small (modal body weight approximately 500 g) and most species are nocturnal. Diurnal species (certain relatively large-bodied members of the families Lemnidae and Indridae) occur only among the lemurs. In fact, some lemurids (species of *Eulemur* and *Hapalemur*) show a highly unusual cathemeral pattern (Tattersall 1988), involving a mixture of nocturnal and diurnal activity. This is just one example of the great diversity of special adaptations found among the lemurs of Madagascar.

By contrast, members of the fourth and fifth natural groups of primates (monkeys, apes, and humans) share a distinctive set of advanced features in their dentitions, jaws, brain, and reproductive system, and can be labeled collectively simians. They are comparatively large (modal body weight approximately 5 kg), and virtually all species are diurnal. The only nocturnal simians are the New World owl monkeys (*Aotus*). The tenfold difference in modal body size between simians and prosimians, undoubtedly connected with the general distinction between nocturnal and diurnal habits, is just one indication of the pervasive importance of body size in primate biology. Availability of a compilation of body weights for almost all primate species, very largely derived from wild animals, therefore deserves special mention (Smith and Jungers 1997). Overall, living primates cover a 300-fold range of body sizes, extending from 40 g for the smallest of the mouse lemur species (*Microcebus*) to 120 kg for an adult male gorilla.

2. Definition of Living Primates

It is often stated that living primates represent a graded series, ranging from relatively primitive to very advanced, and that it is therefore difficult to demarcate them from other orders of mammals (notably Insecti-

vora). However, this view was influenced heavily by the inclusion of tree shrews in the order Primates. Reexamination excluding tree shrews revealed that there is in fact a substantial set of characteristics shared by all living primates, which separates them fairly clearly from other placental mammals, leading to the following definition (Martin 1986):

Primates are typically arboreal inhabitants of tropical and subtropical forest ecosystems. Their extremities are essentially adapted for prehension, rather than grappling of arboreal supports. A widely divergent hallux (big toe) provides the basis for a powerful grasping action of the foot in all genera except *Homo*, while the hand usually exhibits at least some prehensile capacity. The digits typically bear flat nails rather than bilaterally compressed claws; the hallux always bears a nail. The ventral surfaces of the extremities bear tactile pads with cutaneous ridges (dermatoglyphs) that reduce slippage on arboreal supports and provide for enhanced tactile sensitivity in association with dermal Meissner's corpuscles. Locomotion is hindlimb-dominated, with the center of gravity of the body located closer to the hindlimbs, such that the typical walking gait follows a diagonal sequence (forefoot precedes hindfoot on each side). The foot is typically adapted for tarsi-fulcrimation, with at least some degree of relative elongation of the distal segment of the calcaneus, commonly resulting in reverse alternation of the tarsus (calcaneo-navicular articulation).

The visual sense is greatly emphasized. The eyes are relatively large and the orbits possess (at least) a postorbital bar. Forward rotation of the eyes ensures a large degree of binocular overlap. Ipsilateral and contralateral retinofugal fibres are approximately balanced in numbers on each side of the brain and organised in such a way that the contralateral half of the visual field is represented. Enlargement and medial approximation of the orbits is typically associated with ethmoid exposure in the medial orbital wall (though there are several exceptions among the lemurs). The ventral floor of the well-developed auditory bulla is formed predominantly by the petrosal. The olfactory system is unspecialized in most nocturnal forms and reduced in diurnal forms. Partly because of the increased emphasis on vision, the brain is typically moderately enlarged, relative to body size, in comparison to other living mammals. The brain of living primates always possesses a true Sylvian sulcus (confluent with the rhinal sulcus) and a triradiate calcarine sulcus. Primates are unique among living mammals in that the brain constitutes a significantly larger proportion of fetal body weight at all stages of gestation.

Male primates are characterised by permanent precocial descent of the testes into a postpenial scrotum; female primates are characterised by the absence of a urogenital sinus (i.e. urinary and reproductive tracts entirely separate). In all primates, involvement of the yolk-sac in placentation is suppressed, at least during the latter half of gestation. Primates have long gestation periods, relative to maternal body size, and produce small litters of precocial neonates. Fetal growth and postnatal growth are characteristically slow in relation to maternal size. Sexual maturity is attained late and life-spans are correspondingly long relative to body size. Primates are, in short, adapted for slow reproductive turnover.

The dental formula exhibits a maximum of 2.1.3.3/2.1.3.3. The size of the premaxilla is very limited, in association with the reduced number of incisors, which are arranged more

transversely than longitudinally. The cheek teeth are typically relatively unspecialized, though cusps are generally low and rounded and the lower molars possess raised, enlarged talonids.

In fact, an additional distinctive feature should be added to this definition (Sussman 1999): all living primate species live in fairly elaborate social networks. Although nocturnal primate species have commonly been labeled 'solitary,' numerous field studies have shown that they have well established patterns of social interactions based on shared ranges, communal nesting, and regular encounters during the night. There are no immediately recognizable social groups of the kind found in all diurnal primates except the orang-utan, but there are nonetheless long-lasting associations between individuals. The pronounced social tendency now recognized for all primates is presumably linked to the high investment in individual offspring (including intensive parental care), and relatively long lifespans that characterize primates.

The distinctive set of characteristics shared by living primates reflects a suite of adaptations developed in their common ancestor. Taken in combination with various primitive features inferred for the ancestral condition, such as relatively small body size and nocturnal habits (Martin 1990), they yield a plausible reconstruction of the earliest primates. These were probably small-bodied, nocturnally active arboreal creatures that were particularly well adapted for exploitation of the fine branch niche, where they fed on a mixed diet of insects and fruit (the latter accounting for certain modifications of their molar teeth). Leaping and running in the fine terminal branches of trees and among lianas would account for their typical nail-bearing, grasping extremities, notably the prehensile foot, and for emphasis of the visual sense, with large, forward facing eyes and specialization of the brain for highly effective visual orientation. Development of prehensile extremities also involved a shift toward use of the hands rather than the snout in grasping food items, thus explaining shortening of the snout and the associated reduction in number of incisors. Adaptations for low reproductive turnover and high investment in individual offspring can be interpreted as responses to life under the relatively stable and predictable conditions of tropical forests combined with comparatively low predation levels among fine branches.

3. Phylogenetic Relationships among Living Primates

In their broad outlines, reconstructions of the phylogenetic tree of primates based on morphological evidence have been confirmed by analyses based on chromosomal or molecular data, such that a general consensus is now emerging (Martin 1990, Fleagle

1999). Although certain relationships in the tree remain controversial, notably in cases where very early divergences are involved, such confirmation derived from largely independent analyses of new data provides valuable support for earlier phylogenetic reconstructions based exclusively on morphological evidence.

The substantial body of evidence now available indicates that the five natural groups of living primates (lemurs; lorises and bushbabies; tarsiers; New World monkeys; Old World monkeys, apes, and humans) are, in fact, phylogenetically distinct subunits within the tree (Fig. 2). In other words, each of the five natural groups is monophyletic (derived from a specific common ancestor within the tree). Although certain cranial features led some authors to suggest that the mouse and dwarf lemurs (family Cheirogaleidae) are more closely related to lorises and bushbabies than to other lemurs, chromosomal and molecular evidence indicates overwhelmingly that the lemurs are monophyletic (Yoder 1997). Furthermore, it has consistently emerged that the Old World monkeys, and the apes and humans, respectively, belong to two separate monophyletic groups at a lower level. The overall pattern of relationships between the major groups of primates has also become relatively clear. On the one hand, it has emerged consistently that the Lemuriformes and the Lorisiformes are sister groups (that is, their individual common ancestors were derived from an earlier common ancestor). Together, the lemurs, lorises, and bushbabies constitute a larger-scale monophyletic group known as strepsirrhines, because of their common retention of a rhinarium (a naked, moist area of skin surrounding the nostrils in the primitive mammalian condition). On the other hand, it has emerged equally consistently that the simian primates (monkeys, apes, and humans) together constitute a second larger-scale monophyletic group. There is also a considerable body of evidence indicating that tarsiers and simians are sister groups, together constituting a high-level monophyletic group known as haplorhines, sharing complete suppression of the rhinarium (unique among mammals). Although some investigators have questioned an ancestral link between tarsiers and the simians, an overall review of evidence (Martin 1990) indicated that haplorhines do, indeed, constitute a monophyletic group (Fig. 2).

4. The Primate Fossil Record

The known record of reliably identified fossil primates dates back to the beginning of the Eocene epoch, about 55mya. An assemblage of earlier fossil forms mainly from the Palaeocene epoch (65–55mya), collectively labeled 'archaic primates' (Plesiadapiformes), has traditionally been allocated to the order Primates.

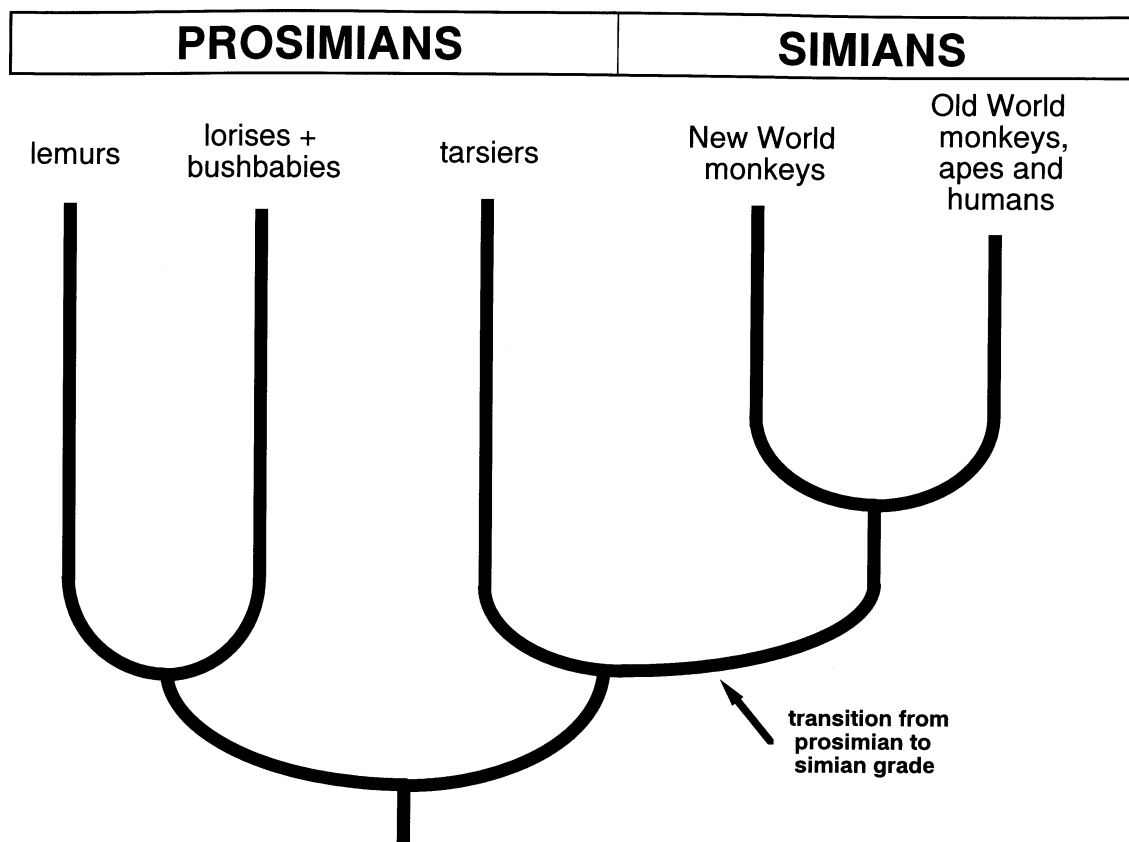


Figure 2

Outline phylogenetic tree of living nonhuman primates (excluding tree shrews) showing the likely pattern of relationships among the five monophyletic natural groups. Traditional classifications distinguish between the generally more primitive prosimians and the simians, which are characterized by a set of more advanced features

However, these early fossil forms (some of which survived into the Eocene) show few, if any, of the features in the definition given above and recent evidence indicates that their connection with primates is questionable. Even if Plesiadapiformes did have some early connection with primates, they surely diverged at such an early stage that they are of little relevance to the later adaptive radiation leading to modern primates and their direct fossil relatives ('primates of modern aspect').

The currently documented primate fossil record, excluding Plesiadapiformes, is divided fairly sharply into two segments separated by a gap of at least 6 million years in the middle of the Oligocene epoch (35–25mya). Approximately equal numbers of fossil primate species (about 225) are known from each half of the record. Initially, primates of the Eocene epoch (55–35mya) were documented only by early prosimians from Europe and North America. On both continents, they can be divided fairly clearly into two families, generally larger lemur-like Adapidae (e.g.,

Adapis in Europe, *Notharctus* in America), and generally smaller tarsier-like Omomyidae (e.g., *Necrolemur* in Europe, *Shoshonius* in America), and many authors have linked them to modern strepsirrhines and haplorrhines, respectively. Almost the entire skeleton is known for *Notharctus*; it is very lemur-like and certainly demonstrates that many defining features of modern primates were present at that stage.

This initial picture from America and Europe has been modified radically by subsequent fossil finds from North Africa, Asia and South east Asia. In North Africa, fossils from the latest Eocene and early Oligocene, notably from the Egyptian Fayum site, have documented the presence of a variety of early simians (e.g., *Aegyptopithecus* and *Apidium*) along with some potential relatives of modern strepsirrhines and tarsiers. More recently, Chinese middle Eocene fossil deposits have yielded evidence of early simians (*Eosimias*) and a putative direct ancestor of modern tarsiers (*Tarsius eoceanus*), along with adapids and omomyids. Furthermore, late Eocene sites in Myan-

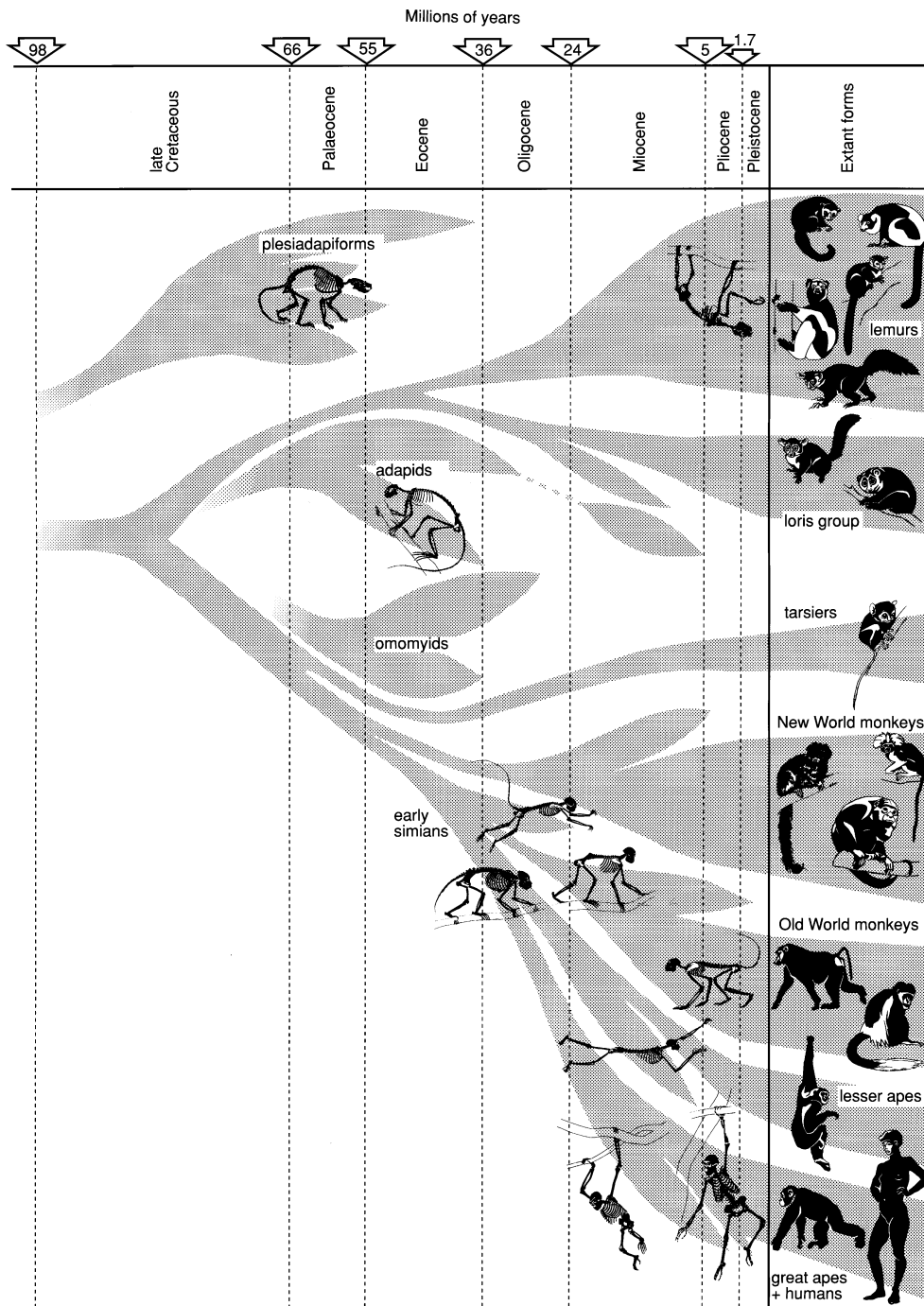


Figure 3

Overall phylogenetic tree of primates (excluding tree shrews) showing the likely pattern of relationships among the five monophyletic natural groups of living primates. This tree also includes major groups of fossil representatives (excluding plesiadapiforms). Note that the initial date of divergence for 'primates of modern aspect' is set at about 80 mya. Original illustration by Lukrezia Bieler-Beerli. (Reproduced by permission of *Nature* (www.nature.com) (Martin R D 363: 223–34) © (1993) Macmillan Magazines Ltd.)

mar and Thailand have yielded several fossil simians (e.g., *Amphipithecus* and *Siamopithecus*). Thus, both prosimians and simians were present in the Eocene.

The situation is entirely different in the second half of the primate fossil record, extending from the late Oligocene to the present day. Direct fossil relatives of New World monkeys are known from South America, beginning in the latest Oligocene (*Branisella*) and extending into the Miocene (e.g., *Homunculus*). In the Old World (notably Africa), early Miocene deposits have yielded fossil relatives of bushbabies and lorises (e.g., *Progalago*), of tarsiers (albeit just one molar tooth from Thailand), and of monkeys and apes (e.g., *Victoriapithecus* and *Proconsul*). Thus, four of the five modern natural groups of primates can be traced back to about 25mya, and one (tarsiers) can be tentatively traced back to 40mya. Nevertheless, there are also notable gaps in the known fossil record. The most striking gap is in Madagascar, where no single fossil from the major radiation of lemurs is known prior to the recently extinct subfossils. Similarly, in the New World, no single fossil form can be linked reliably to the marmosets and tamarins, a substantial proportion of the modern species.

Gaps in the fossil record influence directly the inferred date for the origin of primates. The standard practice has been to assume that primates diverged during the Palaeocene epoch, perhaps 60–65mya, just a few million years prior to the first known fossil primates. However, direct inference from the known fossil record will lead to serious underestimation if there are substantial gaps in the fossil record, as is true of primates (Martin 1993). Several recent studies that calibrated molecular phylogenies for mammals using various dates for fossils of well documented groups outside the primates have all concluded that primates diverged from other mammals about 90mya. Given some allowance for the elapse of time between this initial divergence and initiation of the adaptive radiation leading to modern primates, their common ancestor can be dated at about 80mya (Fig. 3).

See also: Human Cognition, Evolution of; Primate Socioecology; Primates, Social Behavior of

Bibliography

- Fleagle J G 1999 *Primate Adaptation and Evolution*, 2nd edn. Academic Press, New York
- Kinzey W G (ed.) 1997 *New World Primates: Ecology, Evolution, and Behavior*. Aldine de Gruyter, New York
- Martin R D 1986 Primates: A definition. In: Wood B A, Martin L B, Andrews P (eds.) *Major Topics in Primate and Human Evolution*. Cambridge University Press, Cambridge, UK, pp. 1–31
- Martin R D 1990 *Primate Origins and Evolution: A Phylogenetic Reconstruction*. Princeton University Press, Princeton, NJ
- Martin R D 1993 Primate origins: Plugging the gaps. *Nature, London* **363**: 223–34

- Napier J R, Napier P H 1967 *A Handbook of Living Primates*. Academic Press, London
- Rowe N 1996 *The Pictorial Guide to the Living Primates*. Pogonias Press, East Hampton, New York
- Smith R J, Jungers W L 1997 Body mass in comparative primatology. *Journal of Human Evolution* **32**: 523–59
- Sussman R W 1999 *Primate Ecology and Social Structure. Vol. 1: Lorises, Lemurs and Tarsiers*. Pearson Custom Publishing, Needham Heights, MA
- Szalay F S, Delson E 1979 *Evolutionary History of the Primates*. Academic Press, New York
- Tattersall I 1982 *The Primates of Madagascar*. Columbia University Press, New York
- Tattersall I 1988 Cathemeral activity in primates: A definition. *Folia Primatologica* **49**: 200–2
- Yoder A D 1997 Back to the future: A synthesis of strepsirrhine systematics. *Evolutionary Anthropology* **6**: 11–22

R. D. Martin

Primates, Social Behavior of

The distribution of individuals within a species is related to such factors as the distribution and abundance of resources, predation pressure, and genetically determined behavioral propensities. In some animal species, individuals spend most of their time alone, whereas in others most or all of their time is spent in groups. With one exception, all diurnal primates live in social groups. In many animal taxa temporary aggregations form under certain conditions (e.g., schools of fish, flocks of birds, herds of ungulates) but in most of these cases, aggregations are made up of different individuals each time the animals congregate. One of the most common features of primate groups is the relative constancy of group membership. There are many theories attempting to explain why diurnal primates live in groups and the reasons for different types of groups (see Janson 2000).

1. Definition of Terms

There is a great deal of debate concerning what factors may have led to group living among diurnal primates and the relationships that may exist between primate social structure and ecology. However, before discussing some of these issues, it is necessary to define some terms.

1.1 The Term 'Social Group'

If one searches the primate literature for a definition of the term 'social group,' it is quite difficult to find. It seems that primatologists know a social group when

they see one but are reluctant to define the term generally. In the literature, there are few attempts to formally define the primate 'social group' (e.g., Altmann 1962, Kummer 1967, Struhsaker 1969, Sussman 1999) and few critical evaluations of these definitions. Altmann (1962) defines a group as consisting of conspecific, intercommunicating individuals bounded by frontiers of far less frequent communication. Kummer (1967) considers that a group as a 'spatial aggregation of primates that travel and rest together and at the same time avoid the proximity of other such aggregations is, at closer inspection, also a functional and reproductive unit' (p. 359). These are both broad definitions that emphasize social proximity, social interaction, and synchrony of activity. However, all of these criteria are not equally useful in defining the social group.

Reproduction was once thought to be a major *raison d'être* of primate groups but a number of factors indicate that it is not a major common denominator of the primate group. For example, in many species breeding is seasonal but the group is stable throughout the year. In fact, in some species, behavior during the breeding season actually disrupts the social group. In the Madagascar sifaka, *Propithecus*, groups break up during the short mating season and reformulate after breeding takes place. Most species of primates have mechanisms that ensure continuous gene flow between groups. Groups exist within a breeding population (or deme) and genetic exchange is common between groups, at higher or lower frequencies, depending on the species.

Struhsaker (1969) offered a more detailed definition of the social group. He included in his definition the following factors: (a) membership in the same social network; (b) the majority (at least 80 percent) of nonaggressive social interactions take place within this social network; (c) membership is temporally stable; (d) group members occupy the same home range; (e) group members have distinct social roles; (f) progressions between members are synchronized; (g) group members exhibit different behavior toward nongroup conspecifics; and (h) social groups are relatively closed to new members.

Using some but not all of the above criteria, the following definition of a 'social group' in primates is proposed: a number of individuals who interact socially more frequently among themselves than with other individuals; group members exhibit different behavior towards nongroup members and occupy the same home range. Group membership is relatively stable over time and group members have distinct social roles but, although these two factors are characteristics of the group, the particulars of group stability and of social roles are extremely variable from species to species and are not necessary components of this definition.

It is also true that most social groups are relatively closed to new membership but this is also highly

variable and would be difficult to include in a general definition. In fact, the only social groups that were thought to be closed to new membership are monogamous pairs, and this does not appear to be as true as was previously believed (Fuentes 1998, Strier 2000); individual exchange between groups is a common occurrence in most diurnal primate species. Synchronized progressions, mentioned in the definitions of both Kummer and Struhsaker, are not common to all primate groups. A number of groups divide into subgroups for various periods of time and these subgroups do not necessarily synchronize movements or activity while separated. Because of the dynamics of individual social relationships, Dunbar (1988, p. 12) preferred to view primate groups or social systems in terms of their constituent relationships. 'Where the sets of relationships of several individuals coincide and are reciprocated, a formal group emerges as a stable cohesive unit. ... These clusterings are sufficiently distinctive in their cohesion in time and space that we can define them as levels of groupings.'

There are many different types of social group found among the diurnal primates. At a very general level these include: monogamous pairs (i.e., one adult male, one adult female, and their immature offspring); multimale groups (groups containing more than one adult male); one-male groups (groups containing only one adult male); and fission-fusion groups (groups that form dispersing and coalescing subgroups). Of course, each of these superficial categories does not include groups of identical social structure or social organization. For example, multimale groups of lemurs are quite different from those of baboons in structure, function, evolutionary history, and ecological advantage. This is even more obvious when we compare species with 'fission-fusion' group structures.

1.2 The Terms 'Social Structure' and 'Social Organization'

The terms social structure and social organization have now been introduced. These terms are often used interchangeably; however, it is useful to distinguish between them. Following definitions similar to those proposed by Rowell (1972), the social structure of a species can best be divided into two components: group structure and population structure. Group structure is the size and composition (as to age and sex) of a primate group. Population structure is the demographic character of the population, population density, and the density and dispersal of groups within the breeding population or deme. One measures social structure using censusing techniques. Social organization is the pattern of social interactions that occurs between individuals within and between groups: it is a description of social behavior.

Thus, social structure refers to the components of the social system, whereas social organization refers to

the patterns of interaction between the components. Both patas monkeys and Hamadryas baboons live in one-male groups that can have essentially similar age and sexual composition. However, in Hamadryas one-male groups, the male is the center of attention of the group and directs group travel. In the patas monkey, the male takes a peripheral position in the group and the females initiate and direct group travel. Thus the patas monkey and the Hamadryas baboon have groups with similar structure but the social organization of the two species is very different. Furthermore, differences in degrees of tolerance between adult males result in very different distribution patterns of one-male groups. Hamadryas one-male groups often sleep and move together, whereas patas males are intolerant of each other and one-male groups are usually widely dispersed. Although the group structure is similar, the structure of the populations of these two species is radically different.

1.3 Spacing Between and Within Groups

When dealing with animals living in groups, two levels of social spacing must be considered: between groups of the same species and between individuals living within the same group. Each species has developed complex communication patterns for both intergroup and intragroup spacing.

These patterns are part of the social organization of a species and they help to maintain the species-specific social structure. In group-living animals, such as diurnal primates, the terms home range, core area, area of exclusive use, and territory relate to intergroup spacing, whereas dominance hierarchies, roles, kinship lines, and peer groups are usually pertinent to spacing patterns within a group.

In the above definition of a 'social group,' group members were said to occupy the same home range. Home range can be defined as the area over which a group normally travels in pursuit of its routine activities. Other terms such as day range and monthly range are used to denote the area over which the group travels in a given time period. If a group uses a particular part of its home range more frequently, or with greater regularity, than other parts, this zone is referred to as a core area. The core area often contains many of the major resources of the group. In some species of primates, the boundaries of different groups overlap only slightly, whereas in others home ranges overlap a great deal and each group may have a very small area that is not shared by other conspecific groups (referred to as an area of exclusive use). In a number of primate species, home ranges of groups may overlap completely.

Mechanisms used for intergroup spacing also vary between species. In some species, spacing between groups involves the maintenance of exclusive use of a particular area, but in others it seems to involve the

maintenance of intragroup integrity and a patterned use, by two or more groups, of the same space. However, a unitary definition of territoriality and a simple dichotomy between territorial and nonterritorial species might do more to confuse the issue than to aid in an understanding of ecological parameters of spatial relationships between individuals and groups of conspecifics (see Sussman 1999). For example, active, ritualized, agonistic defense behaviors are used by many species of primates to maintain the integrity of home range boundaries or areas of exclusive use. However, in other species, spatial integrity is maintained between groups by exchange of vocalizations or with no apparent outward behavioral spacing mechanisms. In fact, very few species of primates could be considered territorial in the strict sense of the term.

A large number of species that do not maintain exclusive areas employ ritualized postures and gestures, and/or vocalizations, to maintain group integrity and to regulate spatial relationships within a shared home range. Some of these species were thought to be highly territorial until recent studies have shown this not to be the case. Thus, there appears to be two separate phenomena: (a) the behavioral mechanisms that control spacing and (b) the degree of isolation or overlap between groups. The use of the term territoriality often confounds these two factors and a more clear definition might be the active defense of individual or group home range boundaries by actual or ritualized agonistic encounters, thereby maintaining essentially exclusive use of the home range. This, then, describes a very specific set of behaviors that control spacing and a particular pattern of spacing. However, even using this strict definition may conceal important variation in conditions of spacing between two species labeled as territorial.

2. Natural Selection and Primate Behavior

An important question relating to group-living among diurnal primates is why, in fact, do they live in groups? What advantage does this habit confer on individuals? The theory of natural selection predicts that evolution will act to increase the frequency of traits that improve an individual's ability to survive and to reproduce more successfully than other members of its species. Traits that confer survival and reproductive advantages upon an individual are referred to as adaptations. There were, however, a number of characteristics that Darwin found difficult to explain when he developed the theory of natural selection. These included morphological and behavioral differences between the sexes that occur in many species, such as larger canines or body size, or greater aggressiveness in males than in females. In many cases it was not obvious how these secondary sexual characteristics gave any direct advantage to individuals in their attempt to obtain resources.

To explain these traits, Darwin formulated the theory of sexual selection. He proposed that many morphological and behavioral differences between the sexes evolved because they contributed directly to an individual's reproductive success, either by improving his or her competitive ability against members of the same sex, or by increasing attractiveness to the opposite sex. Some of the traits Darwin believed to be sexually selected were subsequently shown to proffer direct adaptive advantage to individuals. However, sexual selection is not an important mechanism in evolution.

A basic assumption of sexual selection theory is that there is a direct relationship between particular traits and genetics. If dominance in males is directly related to increased fitness, then males must be able to pass on this trait to their offspring. However, dominance is complex and involves a number of morphological and behavioral characteristics as well as a social and ecological context. This is true of many other traits proposed to be sexually selected. It is often extremely difficult, if not impossible, to explain the genetics of these complex traits.

Another phenomenon scientists found difficult to explain was cooperative behavior. Why would animals cooperate if this did not increase their evolutionary fitness? If an animal avoids being eaten by a predator by joining forces with others, cooperation easily fits within the rubric of individual selection. However, there are many cooperative behaviors, especially among group-living animals, that appear to be of no benefit or are even dangerous to the individual. To explain these behaviors, the concepts of kin selection and reciprocal altruism were formulated.

Darwin was unaware of Mendelian genetics and thus genetic inheritance did not figure in his explanation of natural selection. Hamilton (1964) examined the potential genetic consequences of cooperative behavior and developed the theory of kin selection to explain altruistic behavior between related individuals. An individual's traits, and the genes responsible for them, are more likely to be shared by close relatives than by unrelated individuals. Thus, by sacrificing oneself for a number of one's kin, an individual may be ensuring the survival of genetically inherited traits. Since more closely related individuals are more likely to share copies of the same genes, altruism is expected to be selectively directed toward kin, and more close kin will accrue more costly altruism (Silk 1987).

However, many altruistic acts are performed by unrelated individuals. Trivers (1971) formulated the theory of reciprocal altruism to explain cooperation and assistance in these cases. He believed that altruistic acts would only occur between unrelated individuals if they were likely to reciprocate in the future and had the opportunity to do so, if the benefits to the recipient were greater than the costs to the altruist, and if there was little chance of cheating. Thus, both parties improve their individual fitness in the long term. In

practice both types of altruistic interactions are often difficult to identify, and predictions derived from these theories are difficult to evaluate empirically (see Silk 1987 for an excellent review).

These concepts formed the basis of the subdiscipline of Sociobiology (Wilson 1975), which was developed to explain the influences of biology on social behavior. Sociobiology was very controversial when first formulated because of the questionable genetic nature of many of the behaviors it attempted to explain. Were behaviors such as aggressiveness or dominance inherited or the result of a complex interaction between genetics and environment? However, many of the theories currently in vogue in socioecology are also based on sexual selection, kin selection, and reciprocal altruism. These include theories attempting to interpret variation in social structure, social organization, and social behavior in group-living primates, and the relationships between ecology and social systems. For example, one of the major theories attempting to relate individual spacing to that of resource distribution is based on sexual selection. In most vertebrates, females must invest a great deal of time and energy in successfully reproducing offspring, whereas males invest a small amount. Because of this, the distribution of females in the environment is presumed to be more closely related to resources than is that of males. The distribution of males is related to the dynamics of male competition for females (Wrangham 1987). Thus, females compete for energy resources and males compete for females, and the social system of a primate species can be understood by studying these relationships. In fact, this theory was proposed when it was believed that, in most primates, females remained in the same home range throughout their lives and males dispersed. Now researchers have shown that, in the majority of primate taxa, either females or both sexes disperse (Strier 2000).

To date it still is not possible to relate patterns of resource distribution to patterns of social spacing among primates. Very little is known about the factors that underlie the enormous amount of variability in primate social systems. These theories so far have not led to a better understanding of the relationships that might exist between ecology and social structure.

3. Ecology and Social Behavior Among Primates

Many of the theories currently attempting to explain the relationship between ecology and social structure among primates are one-dimensional (e.g., distribution of resources, predation pressure, infanticide avoidance). However, there are probably a multitude of factors interacting to cause the great variety of social groups among these animals.

We would expect that there will be some relationship between the spacing of resources in the environment and the spacing of those individuals using the re-

sources. However, to explain the adaptive significance of specific differences between the types of social structure and organization found among primates, diachronic as well as synchronic factors must be considered.

A complete understanding of the relationship between ecology and social behavior among primates will most likely have to take into account at least the following factors: (a) the resources and periodicity of these resources within the total geographical range of the species, as well as those in any particular area; (b) the selectivity of the species—those resources chosen by the species vs. those available; (c) the evolutionary history of the species leading to its current phylogenetic (genetic) propensities (these should be related to the present and past geographic distribution of the species and the environments and resources which exist(ed) within these ranges); (d) the history of the particular population and group, since the conditions in a given area may select for population-specific or group-specific adaptations. This latter factor should take into account stochastic and demographic phenomena which can have a major influence on a population at any particular time (Altmann and Altmann 1979, Chan 1992, Rowell 1993).

See also: Communication and Social Psychology; Evolutionary Social Psychology; Group Processes, Social Psychology of; Primate Behavior: Significance for Understanding Humans; Primate Families, Structure of; Primate Socioecology; Primates, Evolution of; Social Psychology; Social Psychology, Theories of

Bibliography

- Altmann S A 1962 A field study of the sociobiology of Rhesus monkeys, *Macaca mulatta*. *Annals of the New York Academy of Sciences* **102**: 338–435
- Altmann S A, Altmann J 1979 Demographic constraints on behavior and social organization. In: Bernstein I S, Smith E O (eds.) *Primate Ecology and Human Origins*. Garland, New York
- Chan L K 1992 Problems with socioecological explanations of primate social diversity. In: Burton F D (ed.) *Social Processes and Mental Abilities in Non-human Primates*. Edwin Mellen Press, Lewiston
- Dunbar R I M 1988 *Primate Social Systems*. Comstock, Ithaca, NY
- Fuentes A 1998 Re-evaluating primate monogamy. *American Anthropology* **100**: 890–907
- Hamilton W D 1964 The genetical evolution of social behavior. *American Nature* **97**: 354–6
- Janson C H 2000 Primate socio-ecology: The end of a golden age. *Evolutionary Anthropology* **9**: 73–86
- Kummer H 1967 Dimensions of a comparative biology of primate groups. *American Journal of Physical Anthropology* **27**: 357–66
- Rowell T E 1972 *Social Behaviour of Monkeys*. Penguin Books, Harmondsworth, UK

- Rowell T E 1993 Reification of primate social systems. *Evolutionary Anthropology* **2**: 135–7
- Silk J B 1987 Social behavior in evolutionary perspective. In: Smuts B B, Cheney D L, Seyfarth R M, Wrangham R W, Strushaker T T (eds.) *Primate Societies*. University of Chicago Press, Chicago
- Strier K B 2000 *Primate Behavioral Ecology*. Allyn and Bacon, Boston
- Strushaker T T 1969 Correlates of ecology and social organization among African cercopithecines. *Folia Primatologica* **11**: 80–118
- Sussman R W 1999 *Primate Ecology and Social Structure, Volume 1: Lorises, Lemurs and Tarsiers*. Pearson Custom Publishing, Needham Heights, MA
- Trivers R L 1971 The evolution of reciprocal altruism. *Quarterly Review of Biology*. **46**: 35–57
- Wilson E O 1975 *Sociobiology: The New Synthesis*. Harvard University Press, Cambridge, MA
- Wrangham R W 1987 Evolution of social structure. In: Smuts B B, Cheney D L, Seyfarth R M, Wrangham R W, Strushaker T T (eds.) *Primate Societies*. University of Chicago Press, Chicago

R. W. Sussman

Priming, Cognitive Psychology of

Priming refers to the phenomenon whereby the identification of a word or an object can be enhanced either by the presentation of a related stimulus or by a prior processing episode involving that word or object. In the case of priming with a related stimulus, enhancement is assessed relative to identification performance following an unrelated stimulus or some neutral stimulus. Priming effects are used to support inferences about the time course of component processes that contribute to stimulus identification and about the structure of knowledge representation. Priming induced by an earlier processing episode involving the target is referred to as *repetition priming*. Enhancement due to repetition priming is assessed relative to performance on targets that were not included in any earlier processing episode. Repetition priming effects provide evidence regarding the influence of memory on skilled task performance.

1. Priming with Related Stimuli

The widespread use of related primes to examine word and object identification has its origins both in the classic era of verbal learning and in the more recent study of visual attention.

Experiments on verbal learning tasks such as memorizing word lists showed that prior learning of a target word's associate (e.g., *joy* is an associate of *happy*) improved subsequent learning of that target

(Cofer 1967). These early studies foreshadowed some of the well-established aspects of priming obtained in more recent research (e.g., priming dissipates rapidly over time).

In the visual attention domain, it has been proposed that presentation of a prime stimulus that is of the same type as the target could benefit the processing of the target through two mechanisms (Posner and Snyder 1975). First, the prime automatically can activate the pathway used in processing the target. Second, the observer can be induced consciously to attend to activation produced by the prime, creating an expectancy and amplifying the benefit of the prime but also creating a cost if the prime is unrelated to the target and the expectancy is not met.

1.1 Semantic Priming

Building on the seminal work of Meyer and Schvaneveldt (1971), who provided one of the first demonstrations that a pair of semantically related words (e.g., *nurse–doctor*) was identified faster than a pair of unrelated words, early investigations of semantic priming examined automatic and conscious contributions to priming. These studies were motivated in part by the idea that semantically related concepts could automatically activate one another (Collins and Loftus 1975). This automatic spreading activation was the hypothesized basis for the facilitative effect of presenting a prime prior to identifying a semantically related target word. Experiments demonstrating this effect typically employed a word identification task such as lexical decision, in which word and nonword targets are presented and the observer is required to classify each target as a word or nonword. Accuracy in this task is high, so the aspect of performance that is of greatest interest is the time taken to make a response. The benefit of a related prime is revealed by a shorter response latency relative to when a neutral prime such as *XXX* is used.

By varying the time interval between the onset of the prime stimulus and the subsequent onset of the target (stimulus onset asynchrony, or SOA) it was shown that automatic activation of a target word by a semantically related prime can occur within 250ms of the onset of the prime (Neely 1977). At an SOA of 400ms or more, however, not only was there a facilitative priming effect, there also was an interference effect, such that relative to the neutral prime condition, a longer response latency was found when an unrelated prime was used. This pattern of benefit and cost suggested that in the earliest stages of processing a prime word, semantic information related to that prime was automatically activated. Shortly thereafter, attention was directed to the activated information allowing the benefit of activation to be maintained but also leading to a cost if the target was unrelated to the expected domain of semantic knowledge.

1.2 Models of Semantic Priming

The idea that semantic priming may be amplified by conscious attention, perhaps owing to expectations about what related target word might follow a particular prime, is generally accepted. In contrast, however, the proposal that activation spreads automatically from the prime concept to semantically related concepts in semantic memory has been challenged. One type of alternative theory, compound-cue theory, proposes that rather than activation spreading from the prime, the prime stimulus combines with the target to form a compound cue that is matched with the contents of long-term memory (Ratcliff and McKoon 1988). Related prime–target pairs are likely to be represented in memory owing to a prior history of co-occurrence, so this compound cue has a high degree of familiarity and permits a relatively rapid lexical decision. Unrelated prime–target pairs have lower familiarity and more time is needed to make a lexical decision. This theory correctly predicts that presenting an unrelated stimulus between the prime and target eliminates the priming effect.

Whittlesea and Jacoby (1990) extended the compound-cue idea by proposing that processing a target can induce automatic retrieval of a related prime event, producing an integrated or compound prime–target cue. They demonstrated that forming a prime–target compound is situation dependent. When a target is degraded (e.g., by using alternating case: dOcToR), the prime is more likely to form a compound with the target because independent identification of the target is more difficult. The result is a larger priming effect with degraded targets.

A second alternative account of semantic priming derives from neural network models of word identification. In many of these models, knowledge about words is represented by a pattern of activation across a collection of processing units. It is the instantiation of this pattern in the network that supports a response to a word target on tasks such as lexical decision. Related words have similar patterns of activation. Presentation of a related prime moves the network into a pattern similar to that associated with the target, giving the network a head start at instantiating the target's pattern of activation, thereby creating a priming effect. As in the compound-cue theory, presenting an unrelated word between the prime and target is expected to reduce or eliminate the priming effect, although neural network theory attributes this reduction to disruption of the pattern of activation created by the prime (Masson 1995).

Although semantic priming can be disrupted by an intervening unrelated word, prime events appear to modify long-term memory in a manner that supports semantic priming even when a period of minutes intervenes between prime and target. The survival of semantic priming across such time intervals requires that the task performed on the target demand sub-

stantial semantic processing, such as deciding whether the target represents a living or a nonliving object (Becker et al. 1997). A neural network model developed by Becker et al. accounts for this effect by assuming that the prime event modifies connection weights in the network to favor related words, which is equivalent to modifying long-term memory.

Recent work on semantic priming, then, is largely consistent with the idea that short and long timescale priming effects arise from different mechanisms: transient patterns of activation or compound cueing of memory in the former case and incremental learning in the latter.

1.3 Phonological and Other Forms of Priming

The role of *phonological recoding* in the reading of words has been studied by presenting phonologically related primes with target words. Perfetti and Bell (1991) used nonword primes that were phonologically related to target words (*baik-bake*) and control primes that were graphemically but not phonologically related to the targets (*bawk-bake*). In this paradigm, the target word was presented briefly (about 15–30ms) and was *followed* by the prime, which served as a backward mask and was in view for a similarly brief time. Finally, a pattern mask consisting of a row of Xs was presented. Subjects were more accurate in their report of the target word when phonologically related primes were used, even though the primes were usually not consciously perceived. Phonological information derived from the prime can thus contribute to the process of identifying a target. Results such as this one, which have been replicated using Hebrew and Chinese materials, suggest that phonological recoding of words is the default processing mode in reading.

Other types of primes bearing various relations to target words, such as orthographic (*cord-card*) or morphological similarity (*swim-swam*), have also been used to determine whether these knowledge sources play a role in reading. Related primes of these types usually improve performance on target identification tasks relative to primes that bear no relation to the target, implicating these sources of knowledge in reading. In addition, priming paradigms have been used to study clinical populations in an effort to understand the knowledge retrieval processes associated with particular disorders. For example, semantic priming of alcohol-related words is enhanced among subjects who suffer from alcohol dependence. Medications have been shown to reduce this priming effect, suggesting that such interventions are capable of reducing the retrieval of alcohol-related memories.

2. Repetition Priming

Stimulus identification can be enhanced not only by priming with a related item but also by an earlier

presentation of the target itself. Repetition priming of words or objects has been demonstrated in experiments in which a preliminary presentation of an item is followed minutes, hours, or even days later by a repetition of that same item. Improved processing typically is evidenced either by a reduction in the time taken to identify the item or increased accuracy in its identification. In a particularly striking demonstration of this phenomenon, Kolers (1976) showed that even when over 1 year intervened between the first and second reading of text passages presented in inverted typography, subjects were able to read repeated passages faster than ones they had not read earlier. An important aspect of repetition priming that separates it from many other studies of memory is the fact that subjects are often unaware that their performance is being influenced by prior experience with the target stimulus (see *Implicit Memory, Cognitive Psychology of*).

2.1 Long-term Priming

The fact that repetition priming can persist over relatively long periods of time provides an important theoretical constraint. The proposal that repetition priming results from sustained activation of the mental representation of a target item has been criticized on the grounds that it would be implausible for activation to be sustained over long periods of time. Moreover, repetition priming effects involving novel stimuli such as nonwords and novel shapes indicate that activation of pre-existing mental representations is not a necessary condition for priming.

A more plausible account proposes that a representation of the first encounter with an item is encoded into long-term memory and is recruited, usually without conscious intention, when that item is repeated on a later occasion. To the extent that the processing operations performed on the item are similar on the two occasions, recruitment of memory for the original processing episode will benefit current task performance—a principle known as *transfer-appropriate processing*.

Because many of the tasks used to study repetition priming involve the identification of a stimulus, perceptual operations that support identification have been of particular interest in this work. Consistent with the transfer-appropriate processing principle, it has been shown that altering the perceptual characteristics of an item between its first and second presentations (e.g., spoken to printed presentation, or male to female voice) reduces or eliminates repetition priming.

The perceptual specificity of repetition priming has also been taken as support for the proposal that this form of priming is mediated by a neurologically distinct memory system that represents perceptual forms. This memory system is assumed to be in-

dependent of other memory systems, such as the one responsible for supporting conscious recollection of prior episodes. The independent system assumption receives support from the finding that amnesic subjects who are severely impaired on tasks requiring conscious recollection nevertheless show normal repetition priming on various tasks.

Although the processing and memory systems accounts of repetition priming constitute potentially opposing views of memory organization and function, these two approaches are seen by many researchers as complementary (Roediger 1990). In particular, consideration of the memory systems account has encouraged the integration of informative findings from cognitive neuroscience studies, particularly those involving cases of amnesia, into behavioral theories of priming.

2.2 Masked Priming

A second form of repetition priming appears to be short-lived. In *masked repetition priming*, a prime is presented briefly (e.g., 60 ms) and is both preceded and followed by a mask. The second mask is in fact the target item that the subject is required to identify. On repetition priming trials, the target is identical with the prime, except for a change in case (*house*–*HOUSE*), whereas on unrelated trials the prime is unrelated to the target. Presentation of a repetition prime enables subjects to identify the target item more quickly or accurately. This kind of repetition priming has been dissociated from long-term repetition priming in two ways. First, it appears to dissipate within less than 1 s if a delay is introduced between the prime and target. Second, masked repetition priming is equally strong for common, high-frequency words and for uncommon, low-frequency words, whereas in long-term priming low-frequency words show a larger priming effect than high-frequency words.

Dissociations between masked and long-term repetition priming have supported the view that whereas the latter type of priming may be due to a form of episodic memory, the former is due to temporary activation of lexical knowledge (Forster and Davis 1984). The apparent lack of masked priming for nonwords in the lexical decision task also is consistent with this proposal. More recent evidence, however, has challenged two of these dissociations. First, there are cases in which masked repetition priming is stronger for low-frequency words, just as in long-term priming. Second, masked repetition priming of nonwords can occur when the experimental design allows subjects to overcome the tendency to classify fluently processed repetition-primed nonwords as words.

Similarities between masked and long-term repetition priming encourage the view that both phenomena may reflect the operation of a form of episodic

memory. Presentation of a prime is assumed to create an episodic representation that can later be recruited as a resource to assist in the identification of that item on a subsequent presentation. This interpretation of repetition priming fits with the retrieval interpretation of semantic priming described above and promises to provide a fruitful framework for interpreting priming effects in general.

2.3 Negative Priming

There is an interesting circumstance under which repetition of an item can, paradoxically, interfere with task performance. *Negative priming* arises when a pair of items is presented on a *prime* trial, and another pair is presented on the subsequent, *probe* trial. On each trial, the subject is cued to respond to one member of the pair and to ignore the other (e.g., a drawing of a dog in red and a drawing of a violin in green; name the green item). If the ignored item on the prime trial then appears as the cued item on the probe trial, a longer response time is observed than when unrelated items appear on both trials (Tipper 1985). One interpretation of this result is that the act of ignoring an item on the prime trial inhibits the mental representation of that item, rendering it more difficult to identify it on the subsequent probe trial.

The inhibition interpretation of negative priming has prompted some researchers to apply the technique as a measure of inhibitory processing in studies of special populations such as older adults and clinical groups (e.g., individuals with obsessive-compulsive disorder). These applications appear to be premature, however, given recent evidence for an interpretation of negative priming that does not involve inhibition.

On this alternative view, negative priming does not result from selecting against an item on the prime trial. Rather, the appearance of the target on the probe trial automatically retrieves memory for its appearance on the prime trial as in semantic and repetition priming. Because the target must be selected over a distracter item, however, the subject must discriminate between the current perceptual representation of the target and the retrieved memory representation of the target's appearance on the prime trial, among other targets. It is this discrimination process that slows responding and produces negative priming (Milliken et al. 1998). Two important findings support this proposal. First, negative priming is found even when a single stimulus that requires no response is presented on the prime trial, eliminating the need for any inhibition of or selection against that stimulus. Second, when no selection is required on the probe trial, negative priming is not obtained. Thus, in keeping with the memory retrieval account, it is selection on the probe trial, not the prime trial, that induces negative priming.

3. Conclusion

Although the initial theoretical frameworks for understanding priming with related stimuli and repetition priming advocated the idea that priming reflects the relationship between stable mental structures such as a mental lexicon or semantic memory and the activation of those representations, an alternative class of theories is emerging in both domains. These alternative theories emphasize the role of specific episodes in modifying memory representations and in constructing resources that are recruited when a target event is to be identified. The emergence of this new class of theories holds significant promise for the development of a unified account of how memory and perception interact to enable skilled identification of words and objects.

See also: Attention, Neural Basis of; Cognitive Psychology: History; Cognitive Psychology: Overview; Implicit Learning and Memory: Psychological and Neural Aspects; Implicit Memory, Cognitive Psychology of; Semantic Knowledge: Neural Basis of

Bibliography

- Becker S, Moscovitch M, Behrmann M, Joordens S 1997 Long-term semantic priming: a computational account and empirical evidence. *Journal of Experimental Psychology: Learning, Memory, and Cognition* **23**: 1059–82
- Cofer C N 1967 Conditions for the use of verbal associations. *Psychological Bulletin* **68**: 1–12
- Collins A M, Loftus E F 1975 A spreading-activation theory of semantic processing. *Psychological Review* **82**: 407–28
- Forster K I, Davis C 1984 Repetition priming and frequency attenuation in lexical access. *Journal of Experimental Psychology: Learning, Memory, and Cognition* **10**: 680–98
- Kolers P A 1976 Reading a year later. *Journal of Experimental Psychology: Human Learning and Memory* **2**: 554–65
- Masson M E J 1995 A distributed memory model of semantic priming. *Journal of Experimental Psychology: Learning, Memory, and Cognition* **21**: 3–23
- Meyer D E, Schvaneveldt R W 1971 Facilitation in recognizing pairs of words: Evidence of a dependence between retrieval operations. *Journal of Experimental Psychology* **90**: 227–34
- Milliken B, Joordens S, Merikle P M, Seiffert A E 1998 Selective attention: A re-evaluation of the implications of negative priming. *Psychological Review* **105**: 203–29
- Neely J H 1977 Semantic priming and retrieval from lexical memory: Roles of inhibitionless spreading activation and limited-capacity attention. *Journal of Experimental Psychology: General* **106**: 226–54
- Perfetti C A, Bell L 1991 Phonemic activation during the first 40 ms of word identification: Evidence from backward masking and priming. *Journal of Memory and Language* **30**: 473–85
- Posner M I, Snyder C R R 1975 Attention and cognitive control. In: Solso R L (ed.) *Information Processing and Cognition: The Loyola Symposium*. Erlbaum, Hillsdale, NJ
- Ratcliff R, McKoon G 1988 A retrieval theory of priming in memory. *Psychological Review* **95**: 385–408
- Roediger H L 1990 Implicit memory: Retention without remembering. *American Psychologist* **45**: 1043–56
- Tipper S P 1985 The negative priming effect: Inhibitory priming by ignored objects. *Quarterly Journal of Experimental Psychology* **37**: 571–90
- Whittlesea B W A, Jacoby L L 1990 Interaction of prime repetition with visual degradation: Is priming a retrieval phenomenon? *Journal of Memory and Language* **29**: 546–65

M. E. J. Masson

Primitive Society

In popular usage, 'primitive society' distinguishes ourselves from other people who have made little progress towards what we understand as civilization. There are very few such people left today: they live in scattered communities in deserts or rainforests, and they interest us mainly because we think of them as the living relics of our own distant ancestors. The comparison is usually unflattering. Until the beginning of the nineteenth century 'primitive' simply meant 'first' or 'earliest,' but as the word was applied to the original inhabitants of territories colonized by European and American states, it acquired the connotations of inferior, backward, rude.

In the evolutionary social science which developed towards the end of the nineteenth century, 'primitive society' came to be regarded as the defining subject matter of the emerging discipline of anthropology. The new natural science of biology provided the basic inspiration, and physical anthropology remains one of the four subfields of the discipline. However, although physical (or 'biological') anthropologists are much concerned with comparisons among primates they no longer distinguish human bodies as more or less primitive. We vary from one population to another in average body size, skin color, hair type, etc., and people within each community (neighbors or family members, young and old) look unlike, but unless we wish to be regarded as 'racist' we studiously avoid implying that any of these differences connotes inferiority or even some sort of adaptive advantage. Physically we are neither less 'primitive' nor more 'modern' than our ancestors 50–100,000 years ago.

It is, therefore, somewhat ironic that Darwinian hypotheses of the 'ascent of man' by natural selection have provided such a tenacious metaphor for distinguishing among human societies and cultures. For most social scientists for most of the twentieth century, being 'primitive' was a relative judgment about how people think and behave, not about their genes, brains, or bodies. Nineteenth-century explanations of social evolution construed Western civilization not just as the most successful human adaptation, but also as an inevitable and morally necessary path of progress.

While metabolizing or copulating might have changed little over the millennia, there were certainly more and less civilized ways of dining or marrying. And it was significant that only civilized modern man had the capacity to understand and explain the difference.

The first generation of modern anthropologists sought a general, empirically valid definition of the 'primitive,' and used the word without inhibition. Tylor's article defining anthropology in the famous 11th edition of *Encyclopedia Britannica* (1911) may represent the apogee of the technical use of the term. Thereafter, 'primitive' continued to appear in the title of ethnographic monographs, more as a generic label for subject matter than as an analytical concept. By the middle of the twentieth century, even the label had gone out of favor in the social sciences. The dissolution of the European empires and the emergence of new, transnational political-economic interests brought a shift in intellectual attitudes. Ideas of 'underdevelopment' as a natural, primitive condition which would be relieved by technically-enhanced progress towards modernity, soon yielded to ideas of 'underdevelopment' as a political-economic predicament inflicted on citizens of the 'Third World' by the expansion of industrial capitalism from its metropolitan centers. Peoples in the new states who were once classified as 'primitive' were now identified as the least privileged members of a widely distributed 'Fourth World,' which included the impoverished citizens of 'developed' countries.

The only entry under 'primitive' in the 1968 counterpart of this *Encyclopedia* referred to naive styles in art. 'Primitive society' has resurfaced in the present work as a historical curiosity rather than as current terminology. The phrase says more about changing interests, attitudes, and theories in anthropology than about the peoples to whom it once referred. In trying to come to terms with their designated subjects during a century of radical social change, anthropologists have buried 'primitive' society in layers of definition. These include specification of a distinctive sort of social structure, mode of livelihood, spatial distribution, and mentality, all defined (tacitly or explicitly) by a contrast with, and presumed progress towards, civilization as we understand it. Very occasionally an anthropologist will resort to 'primitive' as a portmanteau term: Hallpike (1976) uses it 'simply as a convenient abbreviation for "non-literate, small-scale, face-to-face-societies with subsistence economies."' Although the pejorative implications are resisted sternly, there is still an evident need to refer contrastively to people who have not yet been absorbed within the ambit of 'literate, large-scale, impersonal societies with industrial economies.' Rejecting one word around which a discipline was constructed is not enough to dismiss the assumptions about evolutionary progress which linger on in our understanding of the differences between 'Us' and the sorts of people who are 'Our' anthropological subjects.

1. Livelihood

The simplest and least obviously offensive definition of the primitive has been the material circumstances of their lives. Anthropology as a discipline is still organized around a sequence of social types based on technical developments in the means of production. The prototypical human population of 'hunter-gatherers' lived in nature, taking what they needed day-by-day, whereas subsistence farmers were obliged to work on nature with tools and skills, taking possession of land and other resources, and thereby setting in motion the larger political and economic organization of the state. Technical complexity increased the division of labor, which produced societies in which larger numbers of people depended on more specialized functional relationships with each other. This was especially true of the second great technical revolution which began with the industrialization of cloth production in north-west Europe towards the end of the eighteenth century. With capitalism as its social vector, industrialism swept out to transform the world, creating among other things a new and palpable contrast between the modern and the primitive. It also produced the first generation of social scientists whose role in the ever-expanding division of labor was to explain what was going on.

2. Social Structure

In the pattern of evolutionary contrasts which emerged in the writing of Morgan, Tylor, Rivers, and the other early anthropologists, the social world of the primitive was the small, technically, and socially 'simple' community. In this evolutionary formulation, 'primitive society' was almost a contradiction in terms, "society" being seen as the multiplication of roles and relationships in expanding populations, with heterogeneous individuals bound by new forms of rational contract into complex entities like the modern nation state. The primitive, by contrast, was a 'society of clones' (Kuper 1988), held together 'mechanically' (Durkheim's term) by the relations of reproduction. Social anthropologists pieced together the logic of systems of descent and marriage which apparently gave these 'homogeneous' communities their cohesion. For example, the simple rule 'reckon that you are the child of your father rather than your mother' was enough to tease a whole population out into mutually-exclusive (patri-) lineages, conferring distinct political identities on people who might never actually meet each other. Debates raged about whether paternity or maternity was the more primitive mechanism of social solidarity. The 'elementary structures of kinship' (clans, demes, moieties, etc.) conferred statuses and roles by birth, but these 'ascribed' relationships were less useful as people were drawn into much larger fields of interaction in technically more advanced societies where

survival and progress depended on geographical movement and the exercise of choice and imagination.

Governed by the biologically-derived moral mechanisms of kinship, the primitive tribe could manage without officials and rulers. From an imperial perspective, these emerging definitions of the 'tribe,' boosted by romantic writings of the period, were both intellectually attractive and administratively useful. The tribe suggested a prototypical democracy, a predominantly male political structure, and a paradigm of nationalism—the great seductive and destructive passion of the twentieth century. After two world wars which revealed the barbarity of civilization and cast doubts on who was more primitive than whom, the natural logic of the 'tribe' (as expounded, e.g., by Evans-Pritchard for *The Nuer* (1940) of the southern Sudan) was appealing. In the 1968 *Encyclopedia*, 'Primitive society' was broken down into seven cross-references: Anthropology, Culture, Evolution, Hunting and Gathering, North American Indians, and—the principal entry—Tribal Society. However, the 'Tribal' label could not stick. The year 1968 also brought the revival of the intellectual left, and the association of 'tribalism' with the ideologies and practices of imperialism quickly made it taboo. Anthropologists were now being accused of complicity in the task not just of identifying but inventing the tribes which had been the units of European and North American colonial government.

At this time, the other substitution for 'primitive' was traditional, the preferred counterpart for modern in the evolutionary scheme. But very soon this too was found to be pejorative, mainly because the innumerable books juxtaposing the two words in their titles seemed to presume an ethnocentric convergence on a Euro-American paradigm of modernity. 'Traditional societies' were scattered and diverse, stuck in the inertia of custom until history drew them inexorably into the global monotype of 'modern society.' In the 1980s such teleologies of modernization were rejected by scholars who defined themselves philosophically as postmodern. They wrote critically of our historic and ideological self-centeredness and in praise of a diverse category of 'The Other', which stood mostly against the hegemony of (Western) scholarship itself. An emblematic study was Clastres's (1977) account of the Tupi-Guarani Indians of the Amazon, whose profoundly primitive social structure supposedly set up antibodies for the development of inequality, leadership, capitalism, the state, and the other vices of modern civilization. Such writing paved the way for a revival of sympathetic interest in the 'primitive.' Postmodern characterizations of 'Otherness' invoked a much older assumption about primitive society: its spatial isolation. Earlier anthropologists had pursued the prototypical primitive to the ends of the earth, especially Australia and Oceania, and ethnographic accounts of these tiny encapsulated communities in turn fleshed out the definition of 'primitive.' Post-

modern anthropologists pointed out that the (relatively) concerted forces of worldwide capitalism had tended to marginalize people physically remote from the metropolitan centers, making 'peripheralness' a common denominator of what had hitherto been called traditional, tribal, or primitive. The desire to see 'Others' entirely in their own terms as just 'different' not only meant abandoning pernicious comparisons with 'Us' but also the assumption that the 'Others' had anything specific in common among themselves—some generic sort of 'traditional' culture or 'primitive' social structure.

Most problematically, it was not even safe to assume that everyone had essentially the same sort of body, because biology or physiology themselves were only historically specific sets of ideas for coming to terms with the human condition. This deliberate affront to science has exacerbated a split in anthropology, set in motion by the invention of 'sociobiology' in the 1970s. With 'evolutionary psychology' as the link, a new genre of anthropology is pursuing genetically endowed physical mechanisms which shape culture, for example, through the processes of language acquisition or mate selection. From this point of view, 'primitive society' as represented by today's vanishing hunter populations are living testimony to the 'Environment of Evolutionary Adaptedness' in which modern man was physically constructed and mentally endowed. Virtual banking or an appreciation of Beethoven is fundamentally as 'primitive' as gnawing bones. These exciting speculations have captivated public attention, putting the denser, relativistic theories of idealist anthropologists at a disadvantage.

However, it is ironic that the radically Cartesian postmodern approaches to society have taken the psychic unity of human beings entirely for granted: it is not our minds which differ, but the way they are culturally made up. This returns us to some very old and persistent mental assumptions underlying the meaning of 'primitive' in its various transmutations.

3. *Mentality*

In drawing analytical metaphors from biology or botany the social sciences and humanities have not always kept pace with theoretical advances in the natural sciences themselves. The nineteenth century evolutionist idea of recapitulation was absorbed by the emerging social sciences, and lingers as a latent assumption about human behavior in forms which biological scientists would now barely recognize. The original formula, as specified by Ernst von Haeckel, was that the life history of the individual organism repeats, from the embryo onwards, the stages of its species evolution ('ontogeny recapitulates phylogeny'). Transposed to the social domain, this fostered

the notion that children are primitive—and, by a facile converse, that primitives are children. Thus, Lubbock (1875) was fascinated by ‘The similarity of savages and children,’ and struggled to prove that the childish trait of phonetic reduplication (words like *bonbon*) was an infantile characteristic of primitive languages, absent from civilized English, French, German, and Greek. For social philosophers like Henry Sumner Maine, primitive society was ‘the infancy of man,’ and as late as 1953 Robert Redfield commended the human race for its tendency ‘to grow up.’

The metaphoric transfer of this early biological concept to society and culture continues to create much confusion. Although it is implausible that a modern child could be reliving the traumatic experiences of ancestral adults, the idea that children relive primitive urges remains at the heart of Freudian psychology. Indeed, in his study of *Totem and Taboo* Freud (1938 [1919]) pursued the argument to ‘resemblances between the psychic lives of savages and neurotics.’ The logic of recapitulation leads inexorably to the conclusion that the metamentality of ‘culture’ is, for primitive people, childish and crazy, and that their world is a kindergarten which can only benefit from some civilized parenting—a good excuse for colonization and imperialism. This elaborate equation of the mad, the childish, and the primitive is extraordinarily tenacious, despite many attempts to sort out the confusion. Margaret Mead, for example, sought to demonstrate that primitive Melanesian children were actually ‘more rational’ than their ‘animistic’ parents. This subverts Enlightenment assumptions that rationality increases as a function of both of individual and species development. But for such a devout idealist and anti-evolutionist as Mead, it also raises disturbing question about how children everywhere come ready-equipped with this sort of scientific rationality, and think straight until ‘enculturation’ fills their primitive minds with superstitious rubbish.

Anthropologists have always traded on public fascination with the magic and superstition we associate with primitive (and childish) thought. Given its persistence in modern society, religion itself has not been an acceptable marker of the primitive. The discriminating variable here is, once again, scale: the universalistic ‘World Religions’ (Christianity, Judaism, Islam) deal with the mass of individual souls, while the narrow-minded particularistic religions of primitive society connect people spiritually to community and locality (‘animism’, ‘fetishism’) often through the biosocial idiom of descent (‘ancestor worship,’ ‘totemism’). These peculiarities of primitive thought have interested French anthropologists particularly, immersed as they are in the idealist tradition. Inspired by Franz Boas, Kroeber, Mead, Benedict and other American anthropologists tried very explicitly to reject the evolutionary premises, and pursued a relativistic interpretation of culture (rather than ‘society’) which

sought to avoid the prejudices of a superior western civilization. However, the central conundrum of relativism seems inescapable: how can we distinguish one culture from another if we have no agreed definition (some sort of ‘metaculture’) of what culture is in the first place? Simply to insist that all human beings have the mental capacity to ‘do’ culture evades the issue of how that capacity is physically embodied, how it constrains behavior, and (the evolutionary question) how and why we need it.

The most fundamental belief, which very few modern anthropologists would dare challenge, is of psychic unity, a conception of the human mind as inhabiting a standard brain which may determine how but not what we think. In the idealist tradition, what we think is determined outside of ourselves in an integral, self-reproducing system of ideas—culture, with a ‘mind’ of its own. The trouble is how to reconcile the enormous differences in culture with the primitive sameness of our minds. A symptom of lingering evolutionism has been the association of this primitive mentality with a closeness to nature.

This is captured in the old term *savage*, a synonym for primitive which appeared in anthropological titles well into the twentieth century, though increasingly in ironic quotation marks (Lévi-Strauss (1966) *The Savage Mind*; Goody (1971) *The Domestication of the Savage Mind*; LaFontaine (1981) *The Domestication of the Savage Male*, etc.). Although in French ‘sauvage’ still has a more benign, Rousseauian ring, the English connotations of fierceness and intemperateness latterly made it at least as stigmatic as ‘primitive.’

Seeking a general and objective axis of intellectual difference, many anthropologists have settled on literacy as a key technical distinction between primitive and modern mentality. Thus, in the early 1930s, Margaret Mead uses the word ‘primitive’ to refer to those peoples who are completely without the benefits of written tradition, and in 1953 Robert Redfield declared ‘I shall use “primitive” and “preliterate” interchangeably’ (he also used a range of cognate terms including ‘folk’ and ‘precivilized’). By the 1970s pre- or nonliterate had become the favored term for the societies traditionally discussed by anthropologists, and it remains a key universal classification for those working in the idealist tradition.

4. Nostalgia

A final reason for the persistence of notions of the primitive within and beyond anthropology is plain old-fashioned nostalgia—the old image of the noble savage as a counterpoint to the vices of civilization. Anthropologists are notorious for their populist commitment to the people among whom they have their main (usually first) ethnographic experience. Indeed, it has often been said that anthropologists are people

seeking a special sort of redemption in the notion that the moral virtue lacking in so-called civilized society may be rediscovered among simple people who are more in tune with nature. Defense of primitive peoples, usually in conjunction with the protection of their forest or desert environments, has become a respectable political commitment for many anthropologists. Despite Marx's rage against populist utopianism, his own material laws of history imagined a return to the perfect altruism of primitive society. This image of our natural selves before we were corrupted by the rise of capitalism was inspired by the ethnographic vision of 'Ancient Society' in North America constructed in the mid-nineteenth century by one of the founding fathers of anthropology, Lewis Henry Morgan. Marxists have been rebuked for urging the violent recovery of this old society without offering any plausible account of its structure. That it should somehow combine the benefits of the scale, technical expertise, occupational specialization and intellectual achievements of modern industrial society directly contradicts most of what has been learned about primitive society.

Today, the idea of the primitive lives on in a schizophrenia of admiration and disgust. On the one hand, popular writers like Robert Kaplan have promised the return of the modern world to barbarism, while anthropologists like Paul Richards explain that what appears as primitive lapses in Sierra Leone or Cambodia is a wholly modern circumstance—the global failures of industrial capitalism. In whatever guise, the concept of the primitive seems irrepressible in our imagination of ourselves, lurking in the shadows as we struggle to detach our vile, ancient bodies from the dream of a transcendent modern mind.

See also: Evolution, History of; Evolutionary Approaches in Archaeology; Evolutionism, Including Social Darwinism; Hunter–Gatherer Societies, Archaeology of; Hunting and Gathering Societies in Anthropology; Tradition, Anthropology of; Tribe

Bibliography

- Clastres P 1977 *Society Against the State*. Blackwell, Oxford, UK
- Durkheim E 1933 *The Division of Labor in Society*. Free Press, Glencoe, IL
- Evans-Pritchard E E 1940 *The Nuer: A Description of the Modes of Livelihood and Political Institutions of a Nilotic People*. Clarendon Press, Oxford, UK
- Freud S 1938 [1919] *Totem and Taboo: Resemblances Between the Psychic Lives of Savages and Neurotics*. Penguin, Harmondsworth, UK
- Goody J 1990 *The Oriental, the Ancient and the Primitive: Systems of Marriage and the Family in the Pre-industrial Societies of Eurasia*. Cambridge University Press, Cambridge, UK
- Hallpike C R 1976 Is there a primitive mentality? *Man* 11: 253–70

- Kaplan R D 1994 The coming anarchy: How scarcity, crime, overpopulation, and disease are rapidly destroying the social fabric of our planet. *Atlantic Monthly* February: 44–76
- Kuper A 1988 *The Invention of Primitive Society: Transformations of an Illusion*. Routledge, London
- Lewis I M 1968 *Tribal Society*. *International Encyclopedia of the Social Sciences*. Free Press, Glencoe, IL, pp. 146–51
- Lowie R 1920 *Primitive Society*. Boni and Liveright, New York
- Lubbock J 1875 *The Origin of Civilisation and the Primitive Condition of Man: Mental and Social Conditions of Savages*, 3rd edn. Longmans, Green, London
- Maine H J S 1893 [1874] *Lectures on the Early History of Institutions*, 6th edn. Murray, London
- Mead M 1932 An investigation of the thought of primitive children, with special reference to animism. *Journal of the Royal Anthropological Institute* 62: 173–90
- Redfield R 1953 *The Primitive World and its Transformations*. Cornell University Press, Ithaca, NY
- Stocking G W Jr 1982 *Race, Culture, and Evolution: Essays in the History of Anthropology*, Phoenix edn. University of Chicago Press, Chicago
- Terray E 1972 *Marxism and 'Primitive' Societies*. Monthly Review Press, New York
- Tylor E B 1913 [1871] *Primitive Culture*, 5th edn. Murray, London

A. F. Robertson

Printing as a Medium

Printing is a technique for producing all kinds of paper-based records (books, periodicals, musical scores, ephemera, etc.) by means of ink, a press, and reusable characters. It permits the mechanical production of many identical copies, thereby eliminating the corruption inevitable in hand-copying. Printing has had profound consequences in all spheres of social life. In particular, it has often been ascribed a major role in the transition to modernity that took place in the Western world after the Renaissance.

1. The History and Impact of Printing

1.1 The Invention of Printing

Techniques for the mechanical reproduction of texts are not a uniquely Western invention. Such techniques existed in Korea and China hundreds of years before they did in the West. Yet it is in the West that the press had its greatest impact. It is customary to divide the history of printing there into two periods: the hand-press era, lasting from the mid-fifteenth century to the early nineteenth; and the industrial era, beginning shortly after 1800 with the invention of the steam press and lasting into the present. Social scientists' arguments about the effects of printing as a medium tend to

concentrate on the former, in an effort to help explain modernization in terms of a putative 'printing revolution.' But, as will be suggested below, they would often be more aptly applied to the latter.

There is no plausible route by which Asian technologies may have influenced Johann Gutenberg (c. 1398–1468), a goldsmith from the Rhineland town of Mainz who built the first European press, so he may legitimately be called its inventor. At first, his intent was anything but revolutionary. When he constructed his first press, Gutenberg did so as a calculated intervention in the world of commercial manuscript production that already existed. His earliest printed books were designed to resemble manuscripts, and he intended to sell them as such. He also swore his workmen to confidentiality about his new technique, apparently, in the manner of apprentices in any late-medieval craft. But the strategy did not last. In 1462 Mainz was besieged and sacked, and Gutenberg's printing house broke up. His workers went to other towns and set up their own operations. Before long, printers had appeared in a score of German, Dutch, Italian, and French cities, often fostering legends that printing had been invented locally. Venice, Avignon, Strasbourg, and Haarlem created the most plausible of such histories, as a result of which determining the true origins of the press remained a notoriously intractable problem well into the nineteenth century. William Caxton set up his printing house in London at the end of this first wave of expansion. By that time, in 1480 or so, a new European craft was firmly established.

1.2 The Practice of Early Printing

That craft centered on the *press*—a simple, largely wooden machine, capable of being made by semi-skilled workmen and, if small enough, amenable to concealment and transport from place to place. This press would normally be operated in a place called a printing house. Here, a compositor would set type into a frame called a *form*, which would then be placed in the press itself, which two pressmen would operate to make up to 1,000 impressions per day. The resulting sheets would then be dried, and collated to make books. It sounds, and in essence it was, a simple process. But in practice it was hard to sustain. Most printing houses were severely under-capitalized, and depended for their day-to-day survival not on the printing of books, but on that of ephemera and pamphlets—anything, in fact, that furnished regular and frequent influxes of cash. Almost all of these products were destroyed soon after manufacture. The population of books that has lasted to the present day to fill our great research libraries therefore represents an extreme distortion of the reality of early modern print culture.

This being so, printing raised an immediate question: whether culture could be allowed to be defined by

commercial necessity. Early modern Europe attempted to answer that question by constraining printing on three social levels: the household, the craft, and the polity (where the polity might be a religious one, such as the Catholic church).

In the first place, a printing house, like most places of work in early modern Europe, was normally supposed to be a residence as well. The master printer generally lived in the same building—as, quite frequently, did his journeymen and apprentices. There was a reason why this was held to matter. It was conventional to believe that the moral structure of the patriarchal household largely guaranteed the legitimacy of practices pursued within that household, and printing was to be subjected to the same moral oversight.

In the second place, printing was also organized into guilds, one of which existed in every major city. To be a master printer was to be a respected member of such a group. Each maintained its own conventions, expected to be honored by all members of the book trade. Violations were identified and remedied by book-traders themselves, generally behind closed doors, so that for outsiders the image of an intrinsically harmonious order of print was sustained. Again, there was nothing strange about this; it was how all early modern crafts operated.

And in the third place, such guilds operated in fragile harmony with regulations imposed by church and state. Chief among these were 'privileges' (patents) and licenses. Privileges were grants of literary property, to use an anachronistic phrase, given out by the monarch or equivalent authority. They conflicted in principle with titles allocated by the guilds internally, which was one reason why they were eventually replaced by copyright conventions. A license, on the other hand, was a permission to print (in Latin, *imprimatur*, 'it may be printed'), granted by an authorized individual to the printer or bookseller. In principle, all works had to be vetted by such an official before publication, but in practice this was rather less severe than the censorship sometimes portrayed by modern historians. Most books were never licensed, and only a tiny minority were pursued for lack of an *imprimatur*. The Counter-Reformation Catholic Church extended this principle of licensing into a systematic *Index of Prohibited Books*, which listed titles the ecclesiastical licensors had decreed too unsound for general readers. Again, while this undoubtedly affected the availability of certain books and authors (Machiavelli and Copernicus, for example) in Catholic lands, protestant publishers quickly realized that appearing on the *Index* virtually guaranteed best-seller status in their own countries, and they rushed to print large impressions.

In these ways, then, the social order of early modern Europe sought to accommodate the potentially revolutionary new craft. It was never an entirely stable arrangement, and by the late seventeenth century it

was breaking down. In England, first, licensing lapsed in 1695. With it went much of the authority of the local guild, called the Stationers' Company. Nobody wanted outright deregulation, however, and the major result was a hard-won new compromise between state, authorial, and trade interests. It was articulated in the so-called 'Copyright Act' (1710—the term copyright never actually appears in it). This Act became the basis for the entire subsequent history of copyright and intellectual property law. By the late eighteenth century, agitation for reform of printing and literary property on similar lines had occurred across Europe, and in the years around 1800 broadly similar copyright provisions were established in all major Western states; they extended beyond Europe and the colonies in the nineteenth century. The modern conventions of printed authorship thus owe their origin to the existence and instability of early modern attempts to define and regulate printing: to licensing, patents, and the craft practices of printers themselves (Chartier 1992).

1.3 Industrial Printing

The technology of printing remained virtually unchanged from soon after Gutenberg's invention until around 1800. At that time, however, change came rapidly. Not only did the legal and conventional practices governing the industry alter once and for all; so did the technology of printing itself. First came Stanhope's metal press. This was soon followed by the steam press, which at a stroke increased production rates by an order of magnitude. Stereotyping finally made the dream of truly identical copies a reality. And at the same time, the technology of papermaking changed dramatically with the success of Henry and Sealy Fourdrinier's machines in Hertfordshire. Although it never had the iconic status of the steam press, the papermaking machine had as great a role in the changes of this period, and Brunel himself hailed it as 'one of the most splendid inventions of our age.' With the advent of all these new devices occurring within a generation, between 1800 and 1840 printing underwent its own industrial revolution.

The first commercial steam printing was done at the office of *The Times* in London. It vastly increased the print run of the newspaper, quite literally overnight (the management had the machines installed in secrecy to prevent unrest among the soon-to-be-redundant workers, and unveiled them only when they produced their first day's issue). Nor did the improvements cease there: by 1827 *The Times* had new presses that were producing 5,000 copies an hour. Steam printing was understandably slower to take hold in smaller publishing markets, since it required a large capital outlay—Cambridge University Press, for example, was unusually avant-garde among specialist houses in adopting the technology, yet still did not do so for more than a

generation. And in general the book publishing business was less immediately affected by industrialization than the newspaper and periodical press. But the advent of industrial technology nonetheless changed printing forever, and vastly increased its social impact.

In fact, many of the attributes that historians and social scientists attribute to the hand press may more properly be assigned to the steam press, in conjunction with mechanical papermaking and stereotyping. Typographical fixity in particular was never as clear a component of hand printing as we tend, anachronistically, to assume—in an edition of perhaps 300–1,000 copies, any number of variations would occur across different copies, as correction would be done on the fly. This kind of variability now ceased. Print runs 10 or 100 times as great as had previously been viable also represented a quantitative leap as great as that associated with Gutenberg. They made possible, indeed in economic terms virtually necessitated, the first national daily press. The development of mass literacy, too, was really much more a phenomenon of the nineteenth and especially twentieth centuries than of the sixteenth and seventeenth; mass education of the kind parodied by Dickens in *Hard Times* relied on steam printing. At the same time, since industrial printing required a great deal more concentration of capital than had hand printing, it encouraged agglomeration and uniformity within the publishing industry itself. The press barons of the twentieth century, with their incalculable influence on social and political life, owed their baronies to the printing techniques introduced in the nineteenth century. In effect, printing had become a branch of factory culture.

2. A Printing Revolution?

Gutenberg's invention had profound social and cultural consequences, and modern historians and social scientists have tried hard to discover what they were. First and foremost, the press produced texts in unprecedented quantities. Lucien Febvre and Henri-Jean Martin (1958), in the first modern work to focus on the history and impact of printing, calculated that the number of printed books soon exceeded the number ever produced in manuscript up to that time. Sheer quantity of this order had its own consequences. Books were suddenly available in unprecedented numbers, in more places, and at lower costs. As they spread through social ranks, people—especially town-dwellers—made these new objects their own. Literacy increased markedly, and that alone constitutes one of the major turning-points in the development of modern social order. But mere numbers cannot explain the more specific cultural consequences of the press. Major *qualitative* changes were also brought about by print. These have been described most exhaustively by Elizabeth Eisenstein (1979). They were of two major

kinds: transformations in the making and appearance of the page, and innovations in the uses of books by writers and readers.

After Gutenberg's brief experiment with imitation, the printed page soon began to look very different from its manuscript predecessor. Typefaces and layouts became standardized. Cross-referencing, indexes, errata, and notes came into use, according to conventions that themselves grew more uniform. Woodcuts and engravings allowed the use of repeatable images, permitting diagrams, charts, and scientific illustrations to be communicated with a security never before possible (Ivins 1953, McKenzie 1986). Hitherto, it had been very difficult to express sophisticated claims in pictorial form, so rapidly did images degrade in the hands of copyists. The simple ability to juxtapose and compare reliable representations of competing claims revolutionized the possibilities for learned work.

This textual 'fixity,' as Eisenstein calls it, affected a number of cultural realms, among which religion and science may be singled out. It is not a coincidence that Luther's reformation lasted, while previous attempts at reform (such as those of Hus, or the Lollards) had withered or been wiped out. The permanence and cornucopic plenty of print made Luther's program impossible to suppress. Indeed, it is in the furor over Lutheranism that some of the most characteristic forms of mass communication were invented, such as the pamphlet and the polemical cartoon (Scribner 1981). And, most fundamentally, only such a machine could truly hope to permit a 'priesthood of all believers' by making the Bible available as an object to innumerable readers.

In the sciences too the advent of print permitted great upheavals. The contemplative natural philosophy of the schoolmen gave way before a newly active interrogation of natural processes, modeled, allegedly, on the practices of craftsmen. Print not only circulated this new natural knowledge but also helped to create it, since printers' homes became 'polyglot households' where scientists of all nationalities could meet and correspond. Print thus helped forge the norm of open communication that is so central to science. And men like Henry Oldenburg secured the status of the new science in novel forms of printed communication. Oldenburg's *Philosophical Transactions*, launched in 1665, survived to rank today as the oldest of all scientific journals.

But printed texts alone did not change the world; printed texts plus new ways of using them did. Not least, the sheer mass of printed matter required new classification techniques—in effect, a science of bibliography—to master a worldwide 'library without walls.' But more profound was the change in reading practices that print facilitated. From its early days, the *Philosophical Transactions* was the central plank of a new, transnational readership of naturalists, gentlemen, mathematicians, and scholars—the ancestor of

today's scientific community. The common readership of such periodicals was rapidly recognized as uniting a novel kind of collective, called the *republic of letters* or, rather later, the *public sphere*. This 'public' was defined by its engagement with print. And with Baconian ambition, it took all knowledge to be its province (Habermas 1962). It marked a decisive break from traditional and absolutist societies. In Habermas' very influential account, it therefore laid the foundations for true socio-political transformation, most notably in the form of the French revolution. In that sense, the 'printing revolution' ushered in by Gutenberg warranted its name.

3. Theories about Print

It was in the context of the Jacobins that printing was first recognized as an explicitly revolutionary force. True, Francis Bacon's salute to the press (along with gunpowder and the mariner's compass) was much cited; yet it was little more than an aphorism. The first fully articulated and contextual interpretation of printing's impact was that of the Marquis de Condorcet (1795), who was seeking to explain the overthrow of the French monarchy. Condorcet sketched out a history of 'the human spirit' in a series of stages, with the invention of the press constituting a major turning-point and effectively making the deposition of the Bourbons into 'the revolution that the discovery of printing must bring about.' In composing this sequence, he became the first writer to chart a trajectory of print-based modernization that saw its effects manifested first in the sciences, and only later in general social life. That trajectory has retained its persuasive power in modern times, but its meaning has been construed in very different ways.

Heirs to Condorcet, modern historians, and social scientists have continued to concentrate their interpretative efforts on the typographical *ancien régime* that came to an end with the dual inventions of copyright and the steam press. Compared with this, the period of industrial printing has received comparatively slight attention. There have been many empirical histories of printing, publishing, and reading in the nineteenth and twentieth centuries, but nothing as influential in social-scientific terms as Habermas or Eisenstein. Indicative is the reception of Habermas's work on the public sphere: most Anglophone readers neglect its second half, which tells a depressing story of the replacement of Enlightenment public culture with industrial mass culture. His story is far more elegiac than celebratory, but one would not realize this to read most of the Anglo-American commentators on *The Structural Transformation of the Public Sphere* (1962).

Where they do not ignore the later history of printing, most scholars simply conflate it with the earlier history, and speak of 'print' as though it were a

single, uniform entity. Probably the most famous and notorious of construals along these lines has been that of Marshall McLuhan (1911–1980), the Canadian literary critic and scholar. McLuhan's *The Gutenberg Galaxy* (1962) and *Understanding Media* (1964) set the pattern for a series of works published in the 1960s that made extravagant claims for the cultural and even psychological impact of print. His works used what he called 'probes'—unsystematic, radical aphorisms delivered scattergun fashion to jolt readers out of the formulaic thinking that, he believed, typography inevitably instilled. The probes worked, in at least one sense. McLuhan himself became a cult figure: Tom Wolfe remarked that he sounded like 'the most important thinker since Newton, Darwin, Freud, Einstein, and Pavlov.'

What McLuhan achieved was to make 'media' the subject of analysis in their own right. He argued that these media, of which print was the archetype, brought the world together, dissolving boundaries. The most important of these boundaries were socio-psychic in character. Indeed, McLuhan's very definition of a medium was that it acted as an 'extension of man,' transcending the material constitution of the human frame. In effect, he believed that the state of media determined not only how humans lived, but what humans were. This he presented as a matter of evolution. In particular, McLuhan portrayed what he called 'typographic man,' namely the kind of being who lived according to the cultural logic of print. Typographic man thought in terms of linear logic and objectivity, because his reference points were fixed texts, in what later became the Eisensteinian sense of fixity. Typographic man was both individualist (he could be sure who he was) and nationalist (he could see the bounds of his community, and tell the difference consistently between it and others). There had been no such being before about 1450. At that point, typographic man had begun to displace 'tribal man,' who had been restricted to local contacts and communication by hand. And in his turn typographic man was now becoming extinct, to be replaced by something else that McLuhan himself left nameless, but who could easily be called 'electronic man.'

McLuhan remarked that it was in the nature of electronic media to integrate with the nervous system itself, establishing humans as nodes on a worldwide network. In effect, the skin itself would dissolve as a social constraint, as neurology became inseparable from sociology. Written in the early 1960s, before the first computer network, it is easy to see why McLuhan has recently been rediscovered by the digerati and labeled a 'prophet of the Web.'

Yet McLuhan's arguments were ultimately too faddish in their tone, and too determinist in their content, to remain convincing beyond their bare outlines. Criticisms soon started to mount up (an influential example is Jonathan Miller's short 1971 introduction, which leaves little doubt where its author

stands). There are few social scientists or historians today who would openly acknowledge McLuhan as a leading influence on their representations of print and its consequences. But that influence is nonetheless real. For the most part, it is refracted through the historical analysis of Eisenstein. Eisenstein's careful arguments about the impact of the hand press gave empirical weight to McLuhan's aphorisms. Through her, his proposals have gained academic respectability. An example is Benedict Anderson's widely-praised account of nationalism (1983), which rests on an explicitly Eisensteinian argument linking print to time reckoning, and thence to consciousness of national identity. Similarly once-removed arguments of this kind can be found in many of the analyses of 'print culture' to have appeared in the 1980s and 1990s.

Yet there is one important respect in which current trends in sociological and historical research are starting to depart consciously from this approach. This is the move towards an empirical history of the practice of reading. The immediate origins of this trend lie in France, where the postwar enterprise of the history of the book was born. The spur came from the responses of both history and sociology to the dominant *Annales* school. In both fields, a realization took hold in the later 1970s that quantitative social-scientific accounting failed to capture something fundamental about the cultural impact of print. At the same time, Pierre Bourdieu (1979) was emphasizing the importance of active cultural appropriation by readers of the press rather than passive cultural reception by those readers (Hoggart (1957) had earlier made similar arguments for an English readership). By the early 1980s it looked as though attention to the varied ways of using books, building on Bourdieu's approach, might stand a chance of seeing what the quantitative *Annalistes* were clearly missing. Since then the history of reading has become a burgeoning field in its own right.

The leading proponents of the history of reading have been cultural historians like Roger Chartier in France and Robert Darnton in the USA, who have pioneered the reinterpretation of the French Revolution in these terms. Chartier in particular objects to the determinism about print that is implicit in the work of Habermas, and would direct attention more to changes in reading practice in the eighteenth century (Chartier 1990). In particular, he argues that it is hard to understand how the scatological and pornographic literature that admittedly flooded eighteenth-century France was intrinsically more devastating to the priesthood, say, than the rich seam of libels that appeared in Luther's day. Instead, Chartier would point to a newly skeptical and extensive practice of reading that was devoted to these materials, making them far more damaging in use. This practice contributed substantially to the desacralization of kingship. It hence made regicide possible, and with it the origins of modernity. In this light, the printing revolution

needs to be redefined as part of a broader reading revolution (Cavallo and Chartier 1999).

As a result, research into printing as a medium is today leaving behind quasi-determinist accounts of print exerting some kind of cultural 'logic' on societies, and towards empirical and historical research into the different ways in which societies have made use of the technology and its products. Important examples include the national histories of the book now being published in Britain, the USA, and other countries (Chartier and Martin 1982–1986; Amory and Hall 1999; Hellinga and Trapp 1999). D F McKenzie's 1984 study of a decisive cultural encounter mediated by print in his native New Zealand shows how this modern approach can influence anthropological understandings as well as historical. In McKenzie's terms, excessively schematic talk of 'print logic' is being replaced by a properly contextual 'sociology of texts' (McKenzie 1986). It is interesting to speculate whether, with print logic thus banished, the use of 'print culture' as a catch-all explanatory device may eventually fall into abeyance too. And that may happen, not coincidentally, just as the advent of electronic media signals the end of half a millennium in which print was the predominant communicative medium.

See also: Communication: Electronic Networks and Publications; Literacy and Illiteracy, History of; Mass Media, Political Economy of; Media and History: Cultural Concerns; Media Effects; Media, Uses of; News: General; Radio as Medium; Television: General

Bibliography

- Amory H, Hall D D (eds.) 1999 *A History of the Book in America: Vol. I, The Colonial Book in the Atlantic World*. Cambridge University Press, Cambridge, UK
- Anderson B 1983 *Imagined Communities: Reflections on the Origin and Spread of Nationalism*. Verso, London
- Bourdieu P 1979 *La Distinction: Critique Sociale du Jugement*. Éditions de Minuit, Paris [Also published as 1986 *Distinction: A Social Critique of the Judgement of Taste*. Routledge, London]
- Cavallo G, Chartier R (eds.) 1999 *A History of Reading in the West*. Polity Press, Cambridge, UK
- Chartier R 1990 *Les Origines Culturelles de la Révolution Française*. Seuil, Paris [also published as 1991 *The Cultural Origins of the French Revolution*. Duke University Press, Durham, NC]
- Chartier R 1992 *L'Ordre des Livres: Lecteurs, Auteurs, Bibliothèques en Europe entre XIVe et XVIIIe Siècle*. Alinea, Paris [also published as 1994 *The Order of Books*. Polity, Cambridge, UK]
- Chartier R, Martin H-J (eds.) 1982–1986 *Histoire de l'Édition Française*. 4 vols. Promodis, Paris
- Condorcet M J A N, de Caritat, Marquis de 1795 *Esquisse d'un Tableau Historique des Progrès de l'Esprit Humain*. Agasse, Paris [also published as 1955 *Sketch for a Historical Picture of the Progress of the Human Mind*. Noonday Press, New York]

- Darnton R 1982 *The Literary Underground of the Old Regime*. Harvard University Press, Cambridge, MA
- Eisenstein E L 1979 *The Printing Press as an Agent of Change: Communications and Cultural Transformations in Early-modern Europe*. Cambridge University Press, Cambridge, UK
- Febvre L P V, Martin H-J 1958 *L'Apparition du Livre*. Michel, Paris [also published as 1984 *The coming of the Book: the Impact of Printing 1450–1800*. Verso, London]
- Habermas J 1962 *Strukturwandel der Öffentlichkeit: Untersuchungen zu einer Kategorie der bürgerlichen Gesellschaft*. Luchterhand, Berlin [also published as 1989 *The Structural Transformation of the Public Sphere: An Inquiry into a Category of Bourgeois Society*. MIT Press, Cambridge, MA]
- Hellinga L, Trapp J B (eds.) 1999 *The Cambridge History of the Book in Britain: 1400–1557*. Cambridge University Press, Cambridge, UK; New York
- Hoggart R 1957 *The Uses of Literacy: Aspects of Working Class Life with Special Reference to Publications and Entertainments*. Chatto and Windus, London
- Ivins W M 1953 *Prints and Visual Communication*. Harvard University Press, Cambridge, MA
- McKenzie D F 1984 The sociology of a text: Orality, literacy and print in early New Zealand. *The Library* 6(6): 333–65
- McKenzie D F 1986 *Bibliography and the Sociology of Texts*. British Library, London
- McLuhan M 1962 *The Gutenberg Galaxy: The Making of Typographic Man*. Routledge and Kegan Paul, London
- McLuhan M 1964 *Understanding Media: the Extensions of Man*. Routledge and Kegan Paul, London
- Miller J 1971 *Marshall McLuhan*. Viking, New York
- Scribner R W 1981 *For the Sake of Simple Folk: Popular Propaganda for the German Reformation*. Cambridge University Press, Cambridge, UK

A. Johns

Prison Life, Sociology of

The sociology of prison life concentrates on thick descriptions of the actions and accounting practices that make up the distinctive social organization of a prison. Chief amongst its preoccupations have been how administrators and staff strive to maintain an authority which can never be wholly legitimate in the eyes of their captives; how, in turn, those captives contend individually and collectively with conditions of deprivation and adversity, including the adversities inflicted by their fellow inmates (Morris and Morris 1966, pp. 168–9); how a negotiated social order can emerge out of the exchanges between the two; and how, as a result, prisons manage to persist at all.

The proper study of prison had to await the emergence of universities and other institutions that could sustain social research, deploy some form of fieldwork as part of its methodology, and appreciate

that the prison is a strange institution which 'challenges the anthropological imagination' (Sparks et al. 1996, p. 33). It was not until the 1940s and beyond that those three strands converged in what are commonly held to be the first significant pieces of writing about the sociology of prison life, Clemmer's (1940) *The Prison Community* and, perhaps more important, Sykes's (1958) *The Society of Captives* and Goffman's (1961) *Asylums*. Because they set a pattern and foil for much of what was to come, it is useful to recite their principal arguments.

The Prison Community reported research conducted in the 1930s in a penitentiary for adult felons by a sociologist attached to the mental health staff. Clemmer used the methods of anthropology to study what he called the 'facts of social life,' the groupings, patterns of leadership, controls, leisure pursuits, sex, and work of prison, and to illustrate the complex relations, and the informal relations in particular, of a predatory, suspicious and atomized world of inmates banded publicly against authority but privately riddled by informers and mistrust.

The Society of Captives depicted inmates suffering 'the pains of imprisonment' and striving for the goals of better food, more rewarding work, fuller information, access to sex and pornography, security, status, and autonomy, formally denied them by an administration that purported to treat all prisoners more or less alike and certainly completely distinct from staff. It was in this sense that prison was held to engender anomic pressures that led, in their turn, to the production of illegitimate stratagems, informal social structures and a code of conduct that could satisfy a number of the inmates' aspirations. Inmates were enjoined not to interfere in one another's business, not to 'lose their head,' not to exploit one another, not to weaken, and not to trust the guards. Like many another code, it was largely exhortatory, a form of 'figurative action' rather than a reliable predictor of behavior and its accompanying social relations generated neither 'a perfect solidarity ... [nor] a warring aggregate' (Sykes 1958, p. 83). In practice, some prisoners were forced or encouraged to service the sexual demands of others; some asserted their social standing by violence, defiance or conspicuous example; some supported illegal markets in the provision of goods; and overlapping hierarchies emerged.

The ensuing social order was ostensibly clandestine and based on a rejection of the rejecters, and the staff were supposed to suppress it, but the staff confronted their own difficulties. They themselves wielded great power in principle but little in effect, they were constrained by rules and the practical exigencies and limits of control, they were outnumbered, troubled by the precariousness of their authority, and discouraged from drawing their weapons lest they be disarmed by inmates. Their reply was to replace a formal equality of treatment by a *de facto* recognition and rewarding

of the stronger inmates and inmate groupings, and they did so in exchange for a devolved social order in which inmates effectively regulated one another. The consequence was what Sykes called 'functional corruption,' but it could as readily be seen in functionalist language as a drift towards system integration, equilibrium, or homeostasis, an exemplification of the doctrine that the latent functions of action can work in seemingly paradoxical fashion to promote stability in social systems.

Goffman (1961) worked as a participant-observer in a mental hospital in Washington, DC, to construct a model of asylums, prisons, and like organizations as 'people-processing' 'total institutions' that were substantially cut off from the larger world by symbolic and physical boundaries; a continuous round of activity lived publicly within the walls; a relative lack of internal differentiation of space, role, and temporal schedule; an unbridgeable divide between staff and inmates; and, through what he called a 'mortification of the self,' the discarding of old identities and the assumption of new, degraded, and formally homogeneous identities peculiar only to the institution.

Five main processes have shaped the character and development of much of the work that followed. First has been the adoption of a somewhat critical stance. It was unavoidable that criminologists, penologists, and sociologists came motivated to the study of prison life, sometimes wishing to reform prisons, sometimes to abolish them, and sometimes—more rarely—to analyze them with the dispassion that they might extend to any other social institution. But it has been a strong distaste for prisons that has colored much of what they have written, and a number have been avowedly partisan and, indeed, polemical in their work. That stance has, in its turn, led to methodological difficulties in winning access to, and trust from, a number of key groups in prison, and prison officers especially.

Work has, second, selectively incorporated interpretations of transformations occurring in the political, administrative, and demographic conditions of penal institutions. In the United States, for instance, the sociology of prison life was born in descriptions of autocratic, *laissez-faire* regimes and the resistance that was waged by conservative inmates who feared they might lose power in the face of administrative reform. It moved on through a rehabilitation of American penology's flirtation with rehabilitation in the early 1960s and the dilemmas that arose when the particularistic ethos of treatment confronted the universalism of disciplinary control. It looked at the radical politicization of prisoners, and particularly of black prisoners, in the late 1960s and early 1970s. It analyzed the efforts at legalistic reform, the organizational instabilities, fragmentation of inmate communities, and new waves of riots of the 1970s and early 1980s. And it has culminated most recently in an analysis of the new, racialized politics of mass incarceration and incapacitation that, in Simon's phrase,

is the consequence of 'governing through crime' (Simon 1993).

Internally, American prisons were to become 'Balkanized' into antagonistic racial camps based in some measure on a prior membership of street gangs. Externally, Simon has claimed, the logic of the penal system was to be extended out from the prisons through parole, policing, and probation to subject large blocks of the population, and particularly the young, black male population of the cities of the United States, to a tight control based on actuarial and risk-driven calculations of propensities to offend. Prison life has, in effect, spread beyond the walls of the American penitentiary to the ghetto.

Third, like the sociology of almost any other substantive area, the study of prison life has tended to edit and import some of the more general and evolving theoretical preoccupations of sociology at large as they have appeared to illuminate, and be illuminated by, those structural changes. It has already been observed, for instance, that work was launched inside a context set by the structural-functionalism and anomie theory of the 1940s and 1950s, a context suited to the study of willful, autocratic repression. Current work focuses on issues of resistance and authority, gender and power, 'governmentality' and 'structuration,' and other matters now engaging a number of the wings of sociology.

Fourth, like any other method, and aided by those transformations, interpretations and theoretical pre-occupations, its focus has advanced restlessly from one phenomenon to another, continually introducing new 'facts' with which to re-interpret prison life. Above all, it has produced what is in effect a running series of footnotes or qualifications to the foundational models of the 1950s and 1960s.

It was easy enough at the beginning to overlook some of the finer and more nuanced details of structure. The social anthropologist in the American penologist would have been fascinated by the inmate community's vividly phrased argot and roles. What was less often reported—because it was less obtrusive and less interesting at first sight—was the presence of those who were relatively disengaged and played no part in the clandestine life of the prison, the 'retreatists' and others: who tried to serve their time uneventfully and discreetly and who eventually became interesting precisely because of their earlier neglect. What was also partially obscured at first was the lives of those apart from the inmate. Sociologists moved away from the inmate and towards chaplains, medical officers, and others: for example, the male prison officer beset by problems of violence, and the female prison officer and her gendering of control.

Studies report not only the features that were neglected in the earliest work, but also new features that have emerged since that work was done, but much remains unexplored, for instance, the structural repercussions of the social class of prisoners, aging and the

age composition of prison populations, the world of prison administration, and such groupings as gay prisoners and the prison untouchables (the 'nonces' or nonsenses of English prison argot) imprisoned for sexual and violent offences against children or the elderly.

Most important, the political control of discreditable information and the absence of a sense of anthropological curiosity and of a robust ethnographic tradition in much of the world have ensured that there have been almost no studies of prison life outside Britain, North America and a few small pockets of northern Europe. Sociologists began to emphasize how the early models had tended to extrapolate from one form of time-bound regime alone, the American, authoritarian, maximum security prison for men. It is a model that works well enough where power is arbitrary and formally unchecked, controllers are outnumbered and resistance is fragmented, and it applies broadly to German concentration camps and Soviet labor camps as well as to some of the prisons of the United States. But doubts have been raised about how far or strictly it should otherwise be generalized. It is clearly a limited model. Take Ketchum (1965), who described how in Ruhleben, an internment prison camp established in Berlin for the citizens of countries hostile to Germany in the First World War, there could be seen almost none of the informal responses to incarceration listed by Clemmer, Goffman or Sykes. To the contrary: the inmate population behaved cooperatively, peacefully, and without mutual exploitation, and its conduct invites questions about the impact of class, shame, and pre-existing patterns of organization on the social construction of prison life.

Again, take research on a number of therapeutic or rehabilitative regimes outside the United States. Sparks (1994) studied Barlinnie, a small unit for long-term prisoners in a Scottish prison, and Genders and Player (1995) examined Grendon Underwood, an English therapeutic prison to which serving prisoners in other establishments could elect to go. The three authors declared that neither Barlinnie nor Grendon was critically troubled by the legitimization crises or conflicts of other prisons, but both institutions did benefit from being regimes within regimes—inmates could always be transferred back to the larger and more coercive system from which they had come—and it would be wrong to assume that therapy alone is a proof against disorder or confrontation. So it was that Mathieson's (1965) *Defences of the Weak* described a medium-secure Norwegian prison, organized on therapeutic lines, whose disunited and cynical inmates challenged those who guarded them in a language of 'ensoriousness' that turned therapeutic ideology into a medium of criticism, defense, and attack. Theirs was a response very dissimilar to that of other models.

A second, allied form of qualification to the foundational model attacked the proposition that prisons are

total institutions which sever biographies and create new men and women within an indigenous culture (Sparks et al. 1996, p. 44). Instead of the fractured identities described by Goffman, study after study has revealed that inmates actually bring with them beliefs, associations, and styles of behavior from the outer world.

The importation into prison of social forms from the street gangs and antagonistic racial politics of America has already been noted, and there are other borrowings that have presumably been colored by local and national traits in the world outside the walls. Thus, Irwin and Cressey (1962) talked in America of *thieves* who carried a wider criminal culture into prisons, treating imprisonment as a career contingency; *convicts* whose behavior was a localized adaptation to the special travails of prison life; and *square Johns* who had an allegiance to neither convict nor thief but to a show of obedience to authority. Also, Cohen and Taylor (1972) talked about the affinities between the manner in which long-term inmates in Durham Prison in England managed problems of serving time and the offenses for which they had been convicted, armed robbers worrying about physical deterioration and concentrating on body-building exercises, and fraudsters engaging in copious attempts at litigation.

The discovery of the gendered character of imprisonment is as revealing. The very first studies dwelt on institutions for men. Whilst it was recognized from the outset that male inmates confronted largely male problems stemming from a loss of liberty, status, and heterosexual contact, gender itself was not treated as especially problematic or interesting. It was only when ideas about sexual identity were applied in explanations of the behavior of women inmates in the 1960s, women who were said to form themselves into pseudo-families in response to a loss of emotional sustenance, that the influence of gender itself became notable. Other sociologists have subsequently extended and modified that early work, but the thesis that women combine together differently remains. Later, ideas about masculinity were themselves turned back on the social organization of men's prisons to translate them into places ordered by a 'hegemonic masculinity.' Male prisons, it was argued, must be understood as institutions that are dominated by a very masculine exercise of violence, reliance on hierarchy, ritualized combativeness, and the suppression of displays of emotion and vulnerability.

Last, it must be stressed how closely, although never absolutely, the study of prison life has been bound up with the use of ethnographic methods (Terence and Pauline Morris said of their base in Pentonville Prison, London, that it was 'like the anthropologist's hut on the village street' (Morris and Morris 1966, p. 8)). Its history has, by extension, roughly paralleled the rise and fall of sociological ethnography itself. It was probably at its most vigorous between the 1950s and

early 1970s, when, according to Janowitz, it 'contain[ed] some of the most outstanding research monographs in sociology' (Foreword to Jacobs 1977, p. ix). Its decline may be linked in part to the seemingly inevitable half-life enjoyed by any body of theory and research, and to problems of access and publication that began to surface in the 1970s when prisons were turbulent and their managers became politically defensive and defeatist. But it may also be attributed to the ascent to influence of Michel Foucault (1979), whose political economy of discipline prompted scholars to turn *en bloc* to reflections about what they called 'penality' in a markedly more abstract and empirically disengaged manner.

Foucault took Bentham's plans for a concentric, panopticon prison to be an especially telling representation of the new technologies of control that were emerging in society at large. Prisoners in the brightly lit rim could never know whether they were being observed by guards in the shadowy inspection tower at the center, and the result was supposed to be that they would discipline themselves, colluding in their own subordination in a decentralized, self-sustaining, economical and dispersed system of power. The panopticon was never finally constructed in England, and Foucault's history was flawed, but so seductive and wide-reaching was his metaphor that it became an archetype for a large new family of ideas, and sociologists came to adopt Foucault rather than Clemmer, Sykes or Goffman as their mentor. Jonathan Simon (1999, p. 4) observed that, 'the great tradition of studies of prison social organization that begins with Clemmer and Sykes ... has largely ceased in the last twenty years ... Much of the research now is conducted from afar [and] ... the new distance introduces important interpretive problems ...'

See also: Bentham, Jeremy (1748–1832); Control: Social; Courts and Adjudication; Crime and Class; Crime and Ethnicity (Including Race); Crime, Sociology of; Criminal Justice, Sociology of; Delinquency, Sociology of; Foucault, Michel (1926–84); Goffman, Erving (1921–82); Imprisonment: Sociological Aspects; Law: Economics of its Public Enforcement; Police, Sociology of; Prisons and Imprisonment; Punishment, Comparative Politics of; Punishment: Social and Legal Aspects; Rehabilitation of the Offender

Bibliography

- Clemmer D 1940 *The Prison Community*. Holt, Rinehart and Winston, New York
Cohen S, Taylor L 1972 *Psychological Survival*. Penguin, Harmondsworth, UK
Foucault M 1979 *Discipline and Punish*. Penguin, London
Genders E, Player E 1995 *Grendon: A Study of a Therapeutic Prison*. Oxford University Press, Oxford, UK
Goffman E 1961 *Asylums*. Doubleday Anchor, Garden City, NY

- Irwin J, Cressey D 1962 Thieves, convicts and the inmate culture. *Social Problems* 10: 145–55
- Jacobs J 1977 *Stateville: The Penitentiary in Mass Society*. University of Chicago Press, Chicago
- Ketchum J 1965 *Ruhleben: A Prison Camp Society*. Oxford University Press, Oxford, UK
- Mathieson T 1965 *Defences of the Weak*. Tavistock, London
- Morris T, Morris P 1966 *Pentonville: A Sociological Study of an English Prison*. Routledge, London
- Simon J 1993 *Poor Discipline: Parole and the Social Control of the Underclass*. University of Chicago Press, Chicago
- Simon J 1999 The 'society of captives' in the era of hyperincarceration. Unpublished paper
- Sparks R 1994 The Barlinnie Special Unit as prison and escape. *Prison Service Journal* 95: 2–6
- Sparks R, Bottoms A, Hay W 1996 *Prisons and the Problem of Order*. Clarendon Press, Oxford
- Sykes G 1958 *The Society of Captives*. Princeton University Press, Princeton, NJ

P. Rock

Prisoners' Dilemma, One-shot and Iterated

The Prisoners' Dilemma is the formal structure of a common sort of collective action problems. It is also a vivid scenario embodying that structure. It offers a useful framework for the analysis of interaction where there are certain conflicts of interests.

1. Two Agents, One Meeting

Adam and Eve have been arrested. The District Attorney (DA) now tells them this. They can either talk or keep silent. If they both talk, they will each get a 10-year sentence. If one of them talks (confessing for both) and the other does not, the one who talked will go free; the other will get a 20-year sentence. If they both keep silent, each will get a one-year sentence on some trumped-up charge. The DA lets them discuss it. Then he puts them in separate cells and goes to take their pleas.

Say that neither Adam nor Eve cares about how the other makes out or thinks he will meet the other again. Each considers only the length of his or her own sentence. For each, that person's talking (*D*) dominates that person's being silent (*C*); that is, whatever Eve will do, Adam prefers the outcome of his talking to that of his *not* talking, and correspondingly for Eve. So if both Adam and Eve are rational, both of them will talk. But since each prefers the outcome of *C, C* (both of them silent) to that of *D, D*, both will be sorry for it.

This is the familiar prisoners' scenario, first presented by Albert Tucker in 1950. The story speaks of two people as agents, but the structure of the in-

teraction allows the agents to be teams or nations—some of the earliest applications were to the Cold War race (on these matters, see Poundstone 1992). The arms story is meant to demonstrate that, in certain conflicts of interests where there is some agreement too, rational agents won't cooperate though it would be better for each if they both did than if they did not.

2. Two Agents, Many Meetings

Where the agents are rational and they think they won't meet again, there is no way out for them. Neither party will cooperate (*C*). Both will defect (*D*), and both will be sorry. But say that they think they *will* meet again, that their present interaction is only the first of many just like it. (In place of the possible jail-sentences, let the payoffs be sums of money reflecting the same ranking of the outcomes.) Here it may seem that the prospect each faces of having to live with the other's resentment ought to deter defection. Still, it often is argued that, where the number of rounds is finite and known to both of the agents, if they both are rational (and think each other to be rational ...), both will defect from the start to the finish.

The argument first appeared in Luce and Raiffa (1957). Suppose that the number of rounds is known by both Adam and Eve to be *n*, and that they both are rational. At the *n*th round, each will know that there will be no further meetings, no need to guard against penalties, so they both will defect. Let Adam think that Eve is rational—let him think this throughout. He will then think in the (*n*–1)st round that Eve will defect in the *n*th round whatever *he* does in the (*n*–1)st, that his cooperating in the (*n*–1)st, wouldn't be rewarded by Eve in the *n*th. He will therefore defect in the (*n*–1)st round, and Eve, thinking likewise about Adam, will too. The same in round *n*–2: each expecting the other to defect in the round that follows, each will here defect, though we must now also assume that Adam thinks that Eve thinks *him* rational—let him think this too throughout—and that Eve thinks he thinks *her* rational. So we move stepwise back to round 1 (both agents defecting all the way), though with a heavier load of assumptions at each preceding stage.

This backward induction rests on assumptions about these people's beliefs about each other. It calls for supposing that, from the start to the finish, Adam believes Eve to be rational, believes she believes *him* to be rational, believes she believes he believes *her* to be rational, etc., and so too with the names transposed—call this Iterated Mutual Belief, or IMB. Some writers have noted that IMB isn't credible (Basu 1977, Bicchieri 1989, Pettit and Sugden 1989). Where in fact it fails, rational people might cooperate in at least some of the rounds; they need *not* defect throughout. Both parties may thus want it to fail, and one or the other may act in some way he thinks would get it to fail.

What sort of action would do that? If either cooperated in any round, the other would have to make sense of that, and this alone would suffice. Let Adam cooperate at $n-1$. Then Eve will either think (a) that Adam isn't rational (that he thinks this best for him however Eve responds to it) or (b) that he thinks *she* isn't rational (that he expects her to respond by cooperating at n). If he cooperates at $n-2$, she will think (a) or (b) or (c), that he thinks *she* thinks *he* isn't rational (that he expects her to respond by cooperating at $n-1$ in the hope of getting *him* to cooperate at n), etc. Whether Eve thinks (a), (b), or (c) (or the next-level d at $n-3$), IMB will have failed.

But would it ever be rational for Adam or Eve to cooperate? More generally, how should two rational people act in a sequence of Prisoners' Dilemmas (the number of rounds being known)? The answer depends on what each thinks the other party would do. Let B be the strategy of starting with C and then tit-for-tatting, doing in every subsequent round what the other did just before. Also let B_x be starting with C , then tit-for-tatting in all but the last x rounds, and then defecting in those. It would be rational for Adam to adopt B_x if he thought Eve will defect in round 1 but would then adopt B_{x-1} , and *that* would be rational for Eve if she took Adam's cooperating in round 1 to mean that he was pursuing B_{x-2} . There may be many such pairs of courses. In each, both Adam and Eve cooperate in some (or many or almost all) of the rounds, and what they are doing is rational. For each person's course is the best response to what that person thinks is the other's. But they still will defect in some (perhaps in many) rounds.

3. Two Agents, Equilibria

Facing a sequence of interactions, people seldom know how many there will be. So let us turn to cases in which their number is unknown to the agents. Here the backward-induction argument cannot get a foothold. We cannot begin with the round both parties know will be the last—none is here known to be last. But neither can we argue for the possibility of cooperation as we just did, for that too assumed that the agents reason backwards from some known final round. In game-theoretic terms, a Prisoners' Dilemma is a *game*. Where two people consider what strategy of action they should follow in some sequence of such games, they are facing what is called a Prisoners' Dilemma *supergame*. We are here speaking of *open-ended* supergames of this sort.

The fullest treatment is Taylor's (1987). Let each agent have a number of open-ended strategy options. If he is rational, he will choose among them in light of how he ranks the outcomes of their different pairings with the other agent's options, every such outcome being the sequence of the results of each agent's acting in each round as his strategy requires. Write each

outcome of, say, C, C as $[C, C]$. The outcome of Adam's taking B where Eve does too (of both starting with C and then tit-for-tatting) is thus $[C, C], [C, C], [C, C] \dots$. This is also the outcome of their both cooperating every time—of their both taking C^∞ . Let B' be starting with D and then tit-for-tatting; where Adam takes B and Eve takes B' , the outcome is $[C, D], [D, C], [C, D] \dots$. Where Adam takes D^∞ (defecting each time) and Eve takes B , the outcome is $[D, C], [D, D], [D, D] \dots$.

How Adam ranks these outcomes depends on how he values future benefits and costs. The more the future now matters to him—the smaller his *discount rate*—the more will it figure in his rankings. (If $1-d$ is Adam's discount rate, $0 \leq d \leq 1$, he values some benefit r rounds in the future at d^r times its worth to him now.) Where the future matters enough, he will prefer $[C, C], [C, C], [C, C] \dots$ to $[D, C], [D, D], [D, D] \dots$, the prospect of the long run of $[D, D]$'s offsetting the appeal of the initial $[D, C]$. Thus if the future matters enough, he will prefer his taking B where Eve does too to his taking D^∞ where Eve is taking B .

If he is rational, what strategy will he follow? A supergame is a game, and the central concept of game theory is that of (Nash) equilibria. An equilibrium in a two-person game is a pair of options, each a best response to the other. The idea is that, if Adam knows the situation (including the operative discount rates), he will choose a strategy in an equilibrium pair. What if there are several equilibria? Call an equilibrium *admissible* if the outcome of no other equilibrium is either preferred to it by both agents or preferred by one of them while the other is indifferent. Taylor suggests that a rational person will choose an option in an admissible equilibrium pair.

He notes that several equilibria are possible in a supergame of the Prisoners' Dilemma: D^∞, D^∞ is always an equilibrium, and some other pairings of strategies are equilibria where the agents' discount rates meet certain conditions. His basic conclusion is that where the discount rates of both agents are sufficiently small, B, B and D^∞, D^∞ are equilibria and all other strategy pairs he thinks would be considered are not. Since only B, B is admissible (both agents prefer its outcome to that of D^∞, D^∞), each of the agents will choose B . It follows that, where their discount rates are sufficiently small, both agents will cooperate in every round.

This last may overstate the case. Whether it is rational to cooperate in every round depends not only on the discount rates but also on the strategies considered. Where only $B, B', C^\infty, D^\infty$, and certain others are in view, Taylor's conclusion holds. But still others might come up. Let the agents consider E^m (being an *exploiter*): taking B only after defecting m times while the other cooperated. Let them also consider P^m (being *patient*): cooperating m times and then taking B . The E^m, P^m pair is an admissible equilibrium (where the future matters enough to the

patient)—if Adam followed E^m and Eve followed P^m , they would both be rational. But Adam would be exploiting Eve in each of the first m rounds.

4. Many-person Interactions

Say that there are many people each of whom must either cooperate (pay their dues, pull their weight, etc.) or defect (*not* pay, *not* pull). Each would be better off defecting, whatever the others did. But if all defected, or more than a certain number, all would be worse off than they would be if they all had cooperated. Here we have a *many-person Prisoners' Dilemma*.

A frequently cited scenario is Hardin's (1968) commons story. Each farmer in the village sends his cows to graze on the just barely adequate commons. Each knows that his sending out a new cow would degrade that commons. Since each knows too that his share of the cost would be less than the benefit to him of having an extra cow, each will send out another cow and the commons will collapse. This is the logic of the underprotection (and underprovision) of *public goods*, of the *free rider* problem, the commons (the oceans, the forests, clean air) being the public good, the overgrazing (overfishing, polluting) being free riding.

Refinements are possible here. In a many-person Prisoners' Dilemma, each agent would be better off defecting than cooperating, whatever the others did. Each would also be better off if all cooperated than if all defected. But say that there is some lone defector who exploits all the others. Where these others remain better off than they would be if all defected, the defector is a mere nuisance—he is a free rider proper. Where he leaves some others worse off than they would be if all defected, the defector is injurious—he is a *foul dealer*. (For this distinction, see Pettit 1986.)

Let all the parties expect to be facing the same sort of cases many times over and let each consider what to do in the whole sequence of them. They are then facing a many-person Prisoners' Dilemma supergame. Much has been written about this (see Taylor 1987 for references). It appears that here, as above, cooperating in each round may be rational. This too is argued by Taylor, who shows that everyone's cooperating in round 1 and then also in all the rest if each of the others cooperated just before is an admissible equilibrium if each agent's discount rate is sufficiently small. But so is a lone defector's taking E^m (moving to B only after defecting m times while all the others cooperated) and the others taking P^m —the conditional cooperation after round m being dependent on everyone's cooperating in the round just before. So here too rationality allows for extensive exploitation.

The question of the rationality of cooperation in such many-person contexts has some resonance in political theory. The claim that it *isn't* rational (and so can't be expected) provides the principal backing for the idea that people need a strong governing state. Hobbes and Hume based their case for authority, for a

Sovereign on this, and Hardin argues for strict enforcement of access rights to commons (for limiting procreation and entry into wealthy countries). Taylor uses his proof to the contrary to argue for the viability of voluntarism and for communal rule. So some large issues are involved.

5. Evolution and Stability

The above has been about questions of the rationality of various strategies. There is also a second sort of question. How well in the long run would various strategies fare in different contexts of others? Which (if any) would gain more adherents—which would *evolve*? And which (if any), if pursued by all, would resist invasion by others—which would then be *stable*? Much of the recent discussion derives from Axelrod (1984), a study relating Prisoners' Dilemma to current work in evolutionary biology.

Axelrod ran computer simulations of two round-robin tournaments of two-agent supergames, each supergame pairing a strategy with either itself or some other in the pool. Strategy B won both tournaments, and Axelrod notes that it would have won too in many other contexts (other strategy pools). Suppose that a tournament were repeated many times, the agents in each a new *generation*. Letting the change in the numbers of those pursuing the different strategies in each generation be in proportion to the relative fitness (the yield divided by the average yield) of those who pursued them in the one just before, Axelrod also ran a simulation of a many-fold repetition of tournaments. He found that B was the strategy pursued most often in the end. Conditional cooperation can thus *evolve*.

Would any strategy, if pursued by all, resist an invasion by others? Would any strategy be *stable*? Axelrod held that B would be stable (if the future mattered enough). This has been challenged by Boyd and Lorberbaum (1987), who hold that no strategy is ever stable. Which of these claims is correct?

Resistance to invasion might be *weak*, the invaders being kept from multiplying. Or it might be *strong*, their being pushed to extinction. Biologists speak only of *monotone* dynamics, in which greater fitness implies faster increase in numbers. Under such a dynamic, no strategy is *strongly* stable: none can strongly resist invasion. Say that the natives now pursue B . Let $B2$ be tit-for-two-tats, that is, starting with C and then cooperating unless the other defected in the preceding two rounds. $B2$ is a *neutral mutant* of B : it yields exactly what B yields in all interactions with either B or $B2$. Since B can't gain on $B2$, can't *strongly* resist an invasion by it, it isn't strongly stable. And neither is any other strategy, for every strategy has neutral mutants.

What about weak stability? Let us allow for *heterogeneous* invasions, not all the invaders pursuing the same strategy, and let us allow for (monotone) dynamics other than that of proportional fitness. Say

that *B* is the native strategy, that some of the invaders follow *B2* and the rest *B'*, and that the dynamic is that of imitating the preceding tournament's winner. *B* and *B2* do well against themselves and against each other, *B2* doing slightly worse against *B'*, against which *B* does badly. *B'* does badly against *B*, very badly against itself, and slightly better than *B* against *B2*. *B2* will win the first tournament, and having won, will be imitated and become the only strategy followed. (It in turn will be vulnerable to the alternation of cooperation and defection!) So *B* isn't weakly stable under the imitation dynamic, and neither is any other strategy, for there always is some other, inept (like *B'*) against itself, against which certain neutral-mutant invaders do better than the natives.

No strategy is strongly stable under any monotone dynamic, and (allowing for heterogeneous invasions) none is weakly stable under every such dynamic. Thus Boyd and Lorberbaum are partly right. Still, it can be shown that *B* is weakly stable under the proportional-fitness dynamic: it is (weakly) stable in that special context. So Axelrod is partly right too. (These results are in Bendor and Swistak 1998.)

The basic thesis here is that greater fitness implies faster increase in numbers. This is true of the genes that render their carriers fitter to reproduce. But are we here speaking of genes? Are there genes (or alleles) for *B'*ing and for *B2'*ing and the like? If not, the thesis must refer to the numbers of people *X'*ing or to the numbers of replicas of the *X'*ing-meme (for memes, see Dawkins 1976 and Dennett 1995). Increases in these are independent of reproductive fitness. They depend on imitation (by progeny or by others), and people's reproductive fitness needn't increase the number of their imitators.

What sort of fitness promotes imitation? There may be no single answer, for sets of memes co-adapt to their situations. 'Natural selection among genes chooses those that prosper in the presence of certain other genes ... [and] gene pools come to consist of genes that do well in each others' company ...' (Dawkins 1980, p. 354). The same holds for selection among memes. So the prospects of *X'*ing may hinge on the support of co-adapting memes, among them the agents' ideologies, their concepts of morality. In the context of some common ideologies of behavior toward others, *B'*ing may be imitated. In the context of more patient ideologies, *B2'*ing or *B3'*ing or *P^m'*ing may be imitated. Where it goes along with an *exploiter* ideology, even *E^m'*ing may be imitated. What fits a strategy for imitation depends (at least partly) on the co-evolving ideology. Whether or not this is good news depends on the ideology.

See also: Economics and Ethics; Evolution, Natural and Social: Philosophical Aspects; Game Theory: Noncooperative Games; Rational Choice Explanation: Philosophical Aspects; Surprise Examination

Bibliography

- Axelrod R M 1984 *The Evolution of Cooperation*. Basic Books, New York
- Basu K 1977 Information and strategy in iterated Prisoner's Dilemma. *Theory Decision* 8: 293–8
- Bendor J, Swistak P 1998 Evolutionary equilibria: Characterization theorems and their implications. *Theory and Decision* 45: 99–159
- Bicchieri C 1989 Self-refuting strategies of strategic interaction: A paradox of common knowledge. *Erkenntnis* 30: 69–85
- Boyd R, Lorberbaum J P 1987 No pure strategy is evolutionarily stable in the repeated Prisoner's Dilemma game. *Nature* 327: 58–9
- Dawkins R 1976 *The Selfish Gene*. Oxford University Press, Oxford, UK
- Dawkins R 1980 Good strategy or evolutionarily stable strategy? In: Barlow G W, Silverberg J (eds.) *Sociobiology: Beyond Nature/Nurture*. Boulder, Westview, CO
- Dennett D C 1995 *Darwin's Dangerous Idea*. Simon & Schuster, New York
- Hardin G 1968 The tragedy of the commons. *Science* 162: 1243–8
- Luce R D, Raiffa H 1957 *Games and Decisions*. Wiley, New York
- Pettit P 1986 Free riding and foul dealing. *Journal of Philosophy* 83: 361–89
- Pettit P, Sugden R 1989 The backward induction paradox. *Journal of Philosophy* 86: 169–82
- Poundstone W 1992 *Prisoner's Dilemma*, 1st edn. Doubleday, New York
- Taylor M 1987 *The Possibility of Cooperation*. Cambridge University Press, Cambridge, UK

F. Schick

Prisons and Imprisonment

Imprisonment is the principal form of punishment in most countries for people convicted of serious crimes. People charged with crimes are often held in pretrial confinement until the charges are resolved. The institutions in which convicted offenders and pretrial detainees are held are usually called prisons. In the United States, however, the word 'prisons' refers only to state and federal institutions for offenders sentenced to terms of a year or more. Pretrial detainees and offenders sentenced to shorter terms are held in 'jails' operated by local governments. This article examines the history of prisons, historical and comparative patterns of imprisonment, and the collateral and direct effects of imprisonment on prisoners, their families, and their communities, and on crime rates and patterns.

1. Origins

The modern prison dates from the period 1780–1840, though recognizable precursors existed. Imprisonment as punishment for convicted criminals dates from classical times, but such incarceration occurred only

intermittently and on a small scale. By the sixteenth and seventeenth centuries, many Western countries had created debtors' prisons and houses of correction for paupers, and gaols for holding people pending trial.

Development of the Penitentiary is usually attributed to American Quakers. Penitentiaries were to be places of solitude and introspection where, guided by religious instruction, convicted offenders would repent, realize their errors, and prepare for law-abiding futures. The earliest such regime was established in 1790 in the Walnut Street Jail in Philadelphia. The first purpose-built American penitentiaries were in New York, and in Pennsylvania (see *Punishment: Social and Legal Aspects*).

In its extreme form, the Quaker regime, known as the 'silent system' and associated with Pennsylvania prisons, required prisoners to live in solitary confinement with only limited contacts with staff and clergy. Later, work was permitted in the one-person cells. Because the silent system led to the mental breakdown of many prisoners, it was soon abandoned for the 'congregate system' associated with New York prisons, in which inmates lived alone but worked together in enforced silence.

Europeans including Charles Dickens and Alexis de Tocqueville visited America to examine the penitentiaries and to consider their feasibility. Millbank, the first national penitentiary in England, was built in 1816, although the first on the American model, Pentonville, was not built until 1842. The first French penitentiary, La Petite Rocquette, opened in 1836. Penitentiaries soon existed in every European country.

A second wave of innovation occurred in the late nineteenth century, exemplified by the 1877 opening of New York's Elmira Reformatory. This prison, for male offenders up to the age of 30, was premised on rehabilitative ideas. It strove to reform offenders by providing a wide range of educational and vocational programs together with discipline and religious instruction. These rehabilitative ideas led to the first parole release systems, since there was no point in holding prisoners after they had been reformed. They also precipitated creation of indeterminate sentencing systems in which judges, prison managers, and parole boards exercised broad, standardless discretion over the lives of offenders.

Historians point out that the bursts of enthusiasm were short, that even in their own eras the innovative institutions were atypical, and that most prisons at most times have been brutal, inhumane places. Marxist accounts tied the development of the prison to the emergence of the industrial economy and portrayed the prison as a mechanism for labor discipline (see *Marxism and Law*). Michel Foucault in *Discipline and Punish* (1978) argued that the prison, like public education and the army, was a reaction to the breakdown of the *ancien régime* and served as a

socializing institution, producing disciplined individuals suited to bureaucratic and conformist roles provided by the modern state. David Rothman, in *The Discovery of the Asylum* (1971), described prison developments in Jacksonian America as an attempt by philanthropic reformers to respond to social dislocation and rapid change by trying to reform the by-products of a disordered society: the deviant, disorderly, and deranged. In *Conscience and Convenience* (1980), Rothman considered why new Progressive Era institutions were created and why they endured. 'Conscience,' an idealistic wish to improve the lives of troubled people, explains their creation, and 'convenience,' the autonomy and flexibility given managers by indeterminate sentencing, explains their durability. Michael Ignatieff, in *A Just Measure of Pain* (1978), emphasized the humanitarian intentions of prison reformers but, like Foucault, saw the penitentiary as a product of a series of nineteenth-century social crises and the emergence of the modern state.

During most of the twentieth century, researchers, the media, and the public knew little of life inside prisons. The laws, traditions, and rehabilitative rationale of indeterminate sentencing gave prison and parole officials broad authority over the day-to-day lives of prisoners. In the United States, a 'hands-off doctrine' led courts to refuse even to consider prisoners' complaints, reasoning that prisoners were 'slaves of the state' and that courts should defer to the professional expertise of prison administrators.

Developments in the 1960s and 1970s undermined indeterminate sentencing and transformed American prisons. The prison's rehabilitative rationale lost credibility. Reviews of rehabilitative programs concluded that there was little credible evidence that correctional treatment programs reduced recidivism. The civil rights movement and social ferment of the 1960s led prisoners, especially minority prisoners, to become more assertive. This resulted in high-visibility prison riots, most famously in New York's Attica Prison in 1971, and led to abandonment of the hands-off doctrine. Conditions in many American prisons were then racist, squalid, and brutal. By the late 1970s, federal courts declared conditions in many jails and prisons to be unconstitutional. Prisons in more than 40 states came under federal court control in the 1980s.

Beginning in the 1960s, crime and punishment became central political issues in the United States. Many politicians ran for office on 'law and order' platforms, and supported harsher policies. Fifteen states abolished parole release. Many states enacted laws to make conditions in prisons less comfortable, including reducing or eliminating access to televisions and recreational facilities; others eliminated recreational and treatment programs. In the 1990s, nearly 30 states enacted 'three-strikes-and-you're-out' laws requiring mandatory life sentences for people convicted of a third serious crime. Most prosecutors requested

harsher penalties, most judges imposed them, and most parole boards released fewer prisoners.

Other changes affected prisons. By the late 1990s, private corporations operated hundreds of institutions and provided comprehensive services, such as medical care, to others. Efforts to build new facilities before 1980 were often stymied by not-in-my-backyard movements; communities thereafter competed for new prisons as local economic development initiatives. In the 1960s, there were no labor unions for prison guards; in California in the 1990s, the guards' union was a major campaign contributor and an active lobbying group.

Most developments affecting American prisons also occurred elsewhere. Compared to the USA, prison-guard unions developed earlier and were stronger in Australia, England, and France. Prison privatization occurred extensively in Australia, New Zealand, and England, though nowhere else in Europe. Riots precipitated by prisoners' complaints occurred in many countries. In England and in some Australian states, conservative legislators in the 1990s enacted harsher laws and mandated more Spartan prison conditions, but nowhere as harshly as in the United States. Prisoners' rights movements arose in many countries but, because US constitutional traditions give more legal protections to citizens than are recognized in most countries, no other courts intervened extensively in prisons. The exception is in Europe, where prisoner grievances can be redressed under the European Convention on Human Rights, and where the European Convention Against Torture and Inhuman Treatment influences improvements in prison conditions (see *Torture: Legal Perspectives*).

In 2000, prisons outside the Western countries were typically inhumane, unhygienic, overcrowded, and understaffed, with notable exceptions in a few Asian countries, especially Japan. The United Nations Declaration on Human Rights pertains to prisons, but no international body has authority to enforce such provisions. As a result, reform pressures must come from local groups, with occasional support from human rights groups such as Amnesty International and Human Rights Watch.

2. Imprisonment Patterns

Countries vary widely in how they use imprisonment. In the United States, from 1925 to 1972 the incarceration rate fluctuated between 90 and 130 per 100,000 residents (state and federal prisons only), leading some scholars to argue that self-balancing processes maintained relative stability over time in the incarceration rate. However, since 1972, American incarceration rates have increased every year, quintupling by the century's end. Trends in other countries vary widely, though generally the 1990s were a period of rising incarceration rates. In Holland, the incar-

ceration rate increased continually after 1976, although the 1998 rate was still only 80 per 100,000. In Finland, however, the incarceration rate dropped continually after 1975 and in 1998 was around 60 per 100,000, only 40 percent of the rate in the mid-1960s. In Japan, the incarceration rate fell rapidly in the 1950s and then fluctuated around 35 per 100,000. In Southern European countries, including France, Italy, and Spain, incarceration rates varied erratically with sharp rises and declines.

National incarceration rates vary greatly. The United States' rate in 1998, 680 per 100,000 (prison and jail inmates combined), was among the highest in the world. Only Russia had a comparable rate (685 per 100,000 in 1998). As a region, Eastern Europe had the world's highest rates in the late 1990s, typically varying between 200 and 500 prisoners per 100,000. England, Ireland, Australia, Canada, and New Zealand had rates of 100–130 per 100,000. Most Western and Southern European countries had rates between 50 and 100 per 100,000, and most Scandinavian countries incarcerated approximately 60 people per 100,000. With the notable exception of Singapore (465 per 100,000 in 1997), most Asian countries had incarceration rates of 20–50 per 100,000.

Very different penal strategies can lie behind incarceration rates. Some Scandinavian countries, notably Sweden and Finland, have low absolute rates (55–60 per 100,000 in the 1990s), but large percentages of convicted offenders receive short-term prison sentences. In Germany, a much smaller percentage of offenders are sentenced to prison, and sentences shorter than six months are disapproved. In most European and English-speaking countries, only 1 to 3 percent of prison sentences are for 2 years or longer. In the United States in 1991, 56 percent of state prisoners were serving sentences of 10 years or longer. In part, these differences result from different penal ideologies. Many American policy makers believe crimes are primarily the product of wrongful moral choices of criminals, that punishment policies can affect crime rates, and that prison sentences should be harsh to provide deterrent disincentives, or lengthy to incapacitate incorrigible offenders. Many European policy makers believe that decisions to commit crimes are powerfully influenced by social conditions and inadequate socialization, and that punishment policies are unlikely to influence crime rates.

Imprisonment is used in most countries primarily as a punishment for violent crimes, serious property crimes, and drug trafficking, and seldom for white collar, regulatory, or environmental crimes. In many countries, prison sentences for violent and drug offenders are getting longer, and populations of drug offenders are increasing. Most prisoners in every country are young, male, unmarried, and economically and socially disadvantaged, and have extensive criminal records. Large percentages are drug dependent and committed their crimes under the influence of

alcohol or other drugs. Women constitute only 3 to 8 percent of prisoners in most countries but their numbers and percentages are increasing faster than men's (see *Crime and Gender*). Disadvantaged minority groups are disproportionately present in most prison systems, though the identity of the minority groups varies widely. Relative to population, blacks in the USA and Afro-Caribbeans in England are eight times more likely than whites to be in prison. In Australia and Canada, the aboriginal/nonaboriginal differences are as high as 15 and 20 to 1. Similar disproportions affect North Africans in Belgium, Holland, France, and Italy, Turks and Eastern Europeans in Germany and Scandinavia, and Koreans in Japan (see *Race and the Law*).

3. Collateral Effects of Imprisonment

At least five kinds of collateral effects can be identified. First, what are the effects on prisoners' later lives? Economic and ethnographic literature in the United States and Europe show that imprisonment reduces ex-offenders' subsequent incomes and employment. In the United States, state and federal laws deny ex-offenders the right to vote or to hold office, the opportunity to engage in certain occupations, and the right to receive various public benefits and services.

Second, what are the effects on offenders' spouses or partners, and their children? The relevant literature is thin and fragmented, and the most pressing task is to pull together existing knowledge in order to formulate plausible, testable hypotheses. Hypotheses would address the effects of imprisonment on the financial and social stability of prisoners' families while they are in prison and afterwards, on the maintenance of prisoners' relationships with families, and on the short- and long-term well-being and social functioning of prisoners' children. Sometimes families and partners benefit from the removal of abusive, disordered, or dysfunctional parents and spouses. However, given the strong negative effects on children's well-being of being raised in disadvantaged, single-parent households, the major effects of imprisonment on spouses and children are often likely to be negative.

Third, what are the effects on prisoners' later physical and mental health? Controlling for age, sex, and social variables, suicide rates are higher in prison than elsewhere. Psychological literature on coping and adaptation concludes that even long-term imprisonment has few lasting mental health effects, although some researchers doubt this. It would be surprising if established adverse effects on income, employment, and family functioning were unrelated to former prisoners' mental and physical health. However, a long-term, multiple-measure, longitudinal study is needed to answer such questions, and none has ever been done.

Fourth, what are the effects on prisoners' later crime involvement? The negative effects on ex-prisoners' incomes, employment prospects, and family involvement are predictive of increased offending probabilities. At least since the time of John Howard, the eighteenth-century prison reformer, the proposition has been put forward that prisons are 'schools for crime,' that inexperienced prisoners are socialized into antisocial and oppositional attitudes and leave prison more likely to commit crimes than when they entered. Sociologists refer to this as 'prisonization.' No informed person doubts that this process sometimes happens, and at least partly offsets any crime-reductive effects of imprisonment achieved through deterrence, incapacitation, and rehabilitation. Many European scholars accept as proven that prison is criminogenic. They contend that for crime-prevention reasons, prison terms should be avoided whenever possible. Such arguments explain why German and Austrian policies impose prison terms only for serious crimes and discourage imposition of short sentences (under six months).

Fifth, what are the effects on the larger community? Plausible hypotheses can be offered about effects of American prison expansion on the diversion of public resources (from higher education particularly, and from other public services, to prisons), on economic development (from inner cities where most offenders and victims live, to rural communities where prisons are built), and on community cohesion (in disadvantaged minority communities, large fractions of the young men are or have been in prison and are thereby disabled from working, parenting, marrying, or otherwise contributing to the community). In some disadvantaged communities, imprisonment is so common that it may no longer carry meaningful stigma, and it may have lost whatever deterrent effects it once had.

4. Crime-control Effects of Imprisonment

The literature on crime control effects is inconclusive, small, and for the most part old. First, has increased imprisonment reduced crime rates? Most people probably assume that a quarter-century's quintupling of the American prison and jail population has reduced crime rates. There has, however, been relatively little research in recent years on deterrence and incapacitation effects (see *Deterrence: Legal Perspectives*). Most authoritative reviews of the subject, including the famous 1978 report of the National Academy of Sciences Panel on Research on Deterrent and Incapacitative Effects, conclude that such effects exist, but are modest (Blumstein et al. 1978). Similar conclusions were reached after an exhaustive survey of deterrence research carried out in 1998 by Andrew von Hirsch et al. for the English Home Office.

Second, can prison treatment programs reduce recidivism rates? The pessimism associated with the 'nothing works' findings of the 1970s and 1980s has passed, and there are grounds for cautious optimism about the positive effects, under some conditions, of some cognitive-skills, drug-treatment, vocational training, educational, and other programs in adult prisons. The conditions, however, are stringent: program eligibility must be carefully matched to prisoners' needs and risks, programs must be well implemented and adequately funded, and compatible aftercare programs in the community must sustain treatment.

Third, without distinguishing among deterrent, incapacitative, and rehabilitative effects, does increased imprisonment reduce crime rates in general, and have the increases of recent years done so? The prevalent answer appears to be yes, but the literature, much of it by economists, is technical and inaccessible to nonspecialists. Moreover, research mostly dates from the 1970s and suggests that effects are more modest than is widely understood.

Fourth, in cost-benefit terms is increased imprisonment worth it? This is the smallest and most primitive of the prison effect literatures, and provides few policy-relevant findings. The work to date is useful mostly for methodological reasons, by showing how not to pursue such inquiries. For example, the most widely publicized recent work greatly exaggerates the costs of crime by means of inflated imputed costs of victim 'pain and suffering,' and takes no account of suffering by offenders or pains of imprisonment borne by their partners, children, or communities. It may be, as Franklin Zimring and Gordon Hawkins (1991) suggested, that cost-benefit assessments of the effects of imprisonment require weighing of inherently incommensurable values and have reached a dead end.

5. Prisoners and Prison Staff

In 1970, a sociological literature on prison and prisoner subcultures was the best known body of prison scholarship. Since then, prisons have changed a great deal, as have inmate and staff subcultures and interactions between them, and the learning of earlier times may no longer be valid. American prisons are much larger, inmate populations and staffs are more disproportionately black and Hispanic, more line and management staff are women, gangs are larger and their influence more pervasive, more prison staff are unionized, a large and growing fraction of prisons is under private management, and judicial oversight and intrusion are greater. Because larger fractions of prisoners are serving very long sentences, we have witnessed increased need for medical care and geriatric services.

Behavioral, cultural, and social changes in the larger society impinge on life inside prisons. AIDS and HIV, for example, are more prevalent inside prisons than outside. The proportions of mentally ill and mentally defective prisoners increased as a result of the 1960s and 1970s movement to deinstitutionalize the mentally ill. The civil rights and women's movement have affected prisons, as has the political conservatism of recent years. Programs premised on restorative justice ideas are beginning to appear inside prison walls.

Studies of women prisoners and life inside women's prisons are conspicuously underdeveloped. There is a small quality literature, but a paucity of research on life inside prisons since 1975 has had an especially impoverishing effect on traditionally under-studied subjects.

See also: Deterrence; Deterrence: Legal Perspectives; Deterrence Theory: Crime; Imprisonment: Sociological Aspects; Legal Issues: Public Opinion; Prison Life, Sociology of; Punishment, Comparative Politics of; Rehabilitation of the Offender

Bibliography

- Adams K 1992 Adjusting to prison life. In: Tonry M (ed.) *Crime and Justice: A Review of Research*. University of Chicago Press, Chicago, Vol. 16, pp. 275–359
- Blumstein A, Wallman J (eds.) 2000 *The Crime Drop in America*. Cambridge University Press, Cambridge, UK
- Blumstein A, Cohen J, Nagin D (eds.) 1978 *Deterrence and Incapacitation: Estimating the Effects of Criminal Sanctions on Crime Rates*. National Academy Press, Washington, DC
- Foucault M 1978 *Discipline and Punish*. Pantheon, New York
- Ignatieff M 1978 *A Just Measure of Pain: The Penitentiary in the Industrial Revolution, 1750–1850*. Pantheon, New York
- King R, McDermott K 1995 *The State of Our Prisons*. Clarendon Press, Oxford, UK
- Morris N, Rothman D (eds.) 1995 *The Oxford History of the Prison: The Practice of Punishment in Western Society*. Oxford University Press, Oxford, UK
- Rothman D 1971 *The Discovery of the Asylum*. Little, Brown, Boston
- Rothman D 1980 *Conscience and Convenience: The Asylum and its Alternatives in Progressive America*. Little, Brown, Boston
- Stern V 1998 *A Sin Against the Future: Imprisonment in the World*. Northeastern University Press, Boston
- Tonry M (ed.) 1998 *The Handbook of Crime and Punishment*. Oxford University Press, New York
- Tonry M, Petersilia J (eds.) 1999 *Prisons*. University of Chicago Press, Chicago
- von Hirsch A, Bottoms A, Burney E, Wikström P-O 1998 *Criminal Deterrence and Sentence Severity: An Analysis of Recent Research*. Hart Press, Oxford, UK
- Walmsley R 1996 *Prison Systems in Central and Eastern Europe*. European Institute for Crime Prevention and Control, Helsinki, Finland

Zimring F, Hawkins G 1991 *The Scale of Imprisonment*. University of Chicago, Chicago

M. Tonry

Privacy: Legal Aspects

Serious social and legal concern with privacy is a relatively modern matter, though themes of privacy and its importance form part of all known human civilizations.

1. *The Meaning of Privacy*

The definition of privacy is a matter of some controversy. Its characterizations move from the all-encompassing 'right to be left alone,' covering a general interest in not being interfered with in any way that violates human dignity, advocated by Edward Bloustein (1964), to the more limited interest in control over personal information and accessibility, argued for by Westin Allen (1967) and Parent (1983). Somewhere in the middle we find characterizations of privacy as only those aspects of liberty that relate to intimate and highly personal aspects of one's life (e.g., Rubenfield 1989), such as those involved in sexual preferences and decisions about whether and when to procreate. The uses of 'privacy' in everyday language and in various legal systems are broad enough to cover all of these approaches, so the analysis of ordinary language cannot, by itself, decide the issue. The controversy is found in philosophical analyses of privacy as well as in the contexts of its legal protection. It has led some scholars to argue that the term be abolished in order to promote clarity of thought and analysis (Prosser 1960, Thomson 1975).

Conflating privacy with all claims for noninterference, or with the value of an 'inviolable personality,' is indeed much too broad, and threatens the distinctiveness and the practical utility of privacy. All concede that there are ways to violate dignity (e.g., subjecting one to torture) and to interfere with one's important liberties (e.g., freedom of speech or freedom of religion) without raising concerns of privacy. It is also possible to violate privacy (by watching or by eavesdropping) without limiting the person's liberty. The need to stress this distinction between privacy and liberty is one of the main reasons for defining privacy in terms of control over information, or, in the somewhat broader analysis suggested by Gavison (1980), of accessibility. According to this approach, the hard core of concerns and interests covered by the label 'privacy' is related to the interest of individuals in limiting access to themselves. Privacy can be lost through the invasion of a person's physical space by either presence, wiretapping, or clandestine super-

vision; the acquisition and dissemination of information about a person; or the 'lifting' of a person's anonymity, which in turn may encourage loss of privacy in either of these other ways. Privacy in this sense is thus distinct from liberty, but it is connected to liberty and to dignity in a variety of ways. These connections to liberty and to 'inviolable personality,' to be discussed below, are in part responsible for the importance of privacy as a central value in modern life.

The need to distinguish privacy from a general interest in liberty is challenged by those who identify issues such as abortion, contraception, and choice of sexual expression as matters of privacy. While scholars like Rubenfield (1989) and DeCew (1997) concede that these claims concern liberty, they argue that the special strength of these claims derives from the highly personal and intimate nature of these decisions, and that these features of the decisions are what makes them matters of privacy. This argument, which is supported by our linguistic and moral intuitions, is based on an important ambiguity in the term 'privacy' and the related distinction between 'private' and 'public.' There is a sense in which the 'private' is that which is intimate and personal. But 'private' is also that which is held, or used, or owned, or financed by individuals (as compared to that which is held, used, or financed by the public or at least by some group). In addition, 'private' also means that which one is entitled to keep secret or inaccessible, but also that which the individual is entitled to claim as belonging to her own realm of autonomous decision-making. Abortions and sexual practices are both matters of great intimacy and personal import and matters that most people desire to keep unknown and unpublicized. In addition, the principle of offense suggests that we may prohibit the public performance of activities that are fully legitimate when conducted 'in private' (e.g., sexual intercourse).

Against this background, any attempt to provide one definition of privacy and to reject others must be arbitrary to some extent. Those who use the term in descriptive or normative statements should be aware of these ambiguities and clarify the sense(s) in which they use the terms. In the remainder of this essay, only concerns of privacy in terms of accessibility and information will be discussed. Issues of personal autonomy, important as they are, are better discussed under different headings. The importance of stressing the distinctive definition of privacy has grown as modern technology has created a variety of new challenges to privacy. Modes of surveillance, eavesdropping, photographing, and processing of such information, as well as the development of computerized data banks with huge abilities of storage, retrieval, and processing and linking of information, and the growing possibilities of quick dissemination of such information, both verbal and pictorial, have made the threats of loss of privacy more serious and

acute. It is thus important to connect these urgent practical concerns with the reasons for privacy's importance.

2. The Importance of Privacy

While most individuals seek privacy, at least some of the time and for some purposes, it is not clear whether this wish for privacy should always be protected or granted. Some scholars point out that privacy, at least in the sense of keeping personal information unknown, is simply the scoundrel's best friend (notably, Posner 1978). While these scholars may concede the coherence of the concept of privacy, they argue that privacy is in fact undesirable, encouraging tendencies to individualism and downplaying the importance of community and accountability. Some feminist writers argue that treating privacy as an ideal is just a way to perpetuate the exclusion of women from public life, and to facilitate their continued exploitation and vulnerability within the realm of the private (notably MacKinnon 1989 but see also Gavison 1993).

Proponents of privacy (see, e.g., Bloustein 1964, Westin 1967, Gavison 1980, Fried 1968, Benn 1971) do not argue that all wishes to limit access to information about one or of one's realm should be respected. They may well concede that total privacy is both impossible and highly undesired and undesirable. They do insist, however, that interests in privacy are important, and that privacy is related, and contributes, to other important values such as freedom, autonomy, and intimacy. These values are an integral part of our ideals of personhood and society, and our normative landscape will be impoverished if we do not recognize both the uniqueness and the importance of privacy claims.

It is easier to see the many ways privacy functions in our lives if we try to imagine a world in which we have lost all the privacy we enjoy at present. Imagine the loss in terms of intimacy, friendship, trust, and love that we would have suffered if we could not choose in whom to confide, and with whom to share personal information about ourselves. Or if we had known that whatever we said to any individual would immediately be broadcast on the Internet. Imagine the loss in learning and acquisition of skills we would have suffered if all our drafts and early attempts at mastering a skill had been made public, open to the judgment and ridicule of all. Imagine the loss in terms of self-respect and autonomy that would ensue from a life in which we would not have the benefit of experimenting and trying things alone, or in a protected and supportive environment, which could teach us and give us the courage to face audiences less supportive. Imagine democracy without a secret ballot and the freedom to develop ideas and associations without constant police surveillance, or rehabilitation without the ability to start afresh. It is no accident that

people have sought privacy from the inception of human society, and that extensive loss of privacy (of the kind that is suffered in total institutions like prisons or mental hospitals) is regarded by many as degrading and humiliating.

Even if this argument is conceded, and people accept that some claims to privacy should be morally protected, this does not say that claims to privacy or limitation of accessibility are absolute. Like all other claims, they must be balanced against other important interests and rights. The importance of privacy means only that we should be aware that the costs we pay in terms of privacy are important, and that we should try, as far as possible, to decrease them. While total loss of privacy is dangerous and bad, some claims of privacy are merely self-serving and instrumental. Naturally, a candidate for a job does not want her prospective employer to know her weaknesses. A con-man does not want his victims to know of his past record. In all these cases we may well prefer the interest in knowledge and thus refuse to defend the claim of secrecy or privacy. As a matter of fact, in many situations of conflict, privacy gives way to other claims. Some scholars, have concluded from this fact that it is therefore misleading to talk about a 'right to privacy' in morality or in law (Prosser 1960, Thomson 1975, Davis 1959). They argue that privacy claims can never trump other claims on their own. The cases in which we feel that claims of privacy have the upper hand are ones in which there is another interest involved, such as the interest in reputation, or in physical seclusion, or in property. It is the presence of these other interests, they argue, not the claim of privacy, that explains (and justifies) the fact that we find the privacy argument persuasive.

A closer analysis of cases of conflict, in both mortality and law, does not support this generalization. Norms defending the rights of prominent gay people not to be 'outed,' and laws prohibiting the publication of rape victims' names, are two examples in which claims of privacy are seen as morally justifying, on their own, the limitation of the freedom of others to disseminate information. Moreover, even if it is the case that in many situations privacy claims win only when they are reinforced by other interests, this does not mean that the costs in terms of privacy are not important. It may mean only that claims of privacy, on their own, cannot defeat some conflicting claims. It is misleading to conclude that the costs in terms of privacy are nonexistent.

3. Legal Protection of Privacy and its Limits

Part of the debate around the concept of privacy and the justification for granting a right to privacy was triggered by claims for legal protection against various sorts of losses of privacy. One of the consequences of the conceptual ambiguity mentioned above is that the

extension of the term 'privacy' may be different in different legal systems. Thus, the initial decisions concerning the use of contraceptives and abortion in the US were made in terms of privacy (Griswold 1965 and Roe 1973). While the US Supreme Court has moved to discuss abortion as an issue of liberty rather than one of privacy (Planned Parenthood 1992), there are still scholars seeking to justify dealing with issues, such as the liberty to behave in certain intimate matters, in terms of privacy (Rubinfeld 1989, DeCew 1997, Allen 1988). Other scholars see these decisions as based on conceptual and normative confusion between issues of liberty and issues of privacy (Gross 1971, Henkin 1974, Gavison 1980).

Legal systems vary a great deal in their positive protection of privacy. Most Western countries now try to regulate computerized data banks in some ways. The legal frameworks seek to make sure that data is accurate, that individuals know what information about them is held in such data banks, and that acquisition and dissemination of information are controlled. The first two issues, while of extreme practical importance, are not—strictly speaking—privacy concerns. My privacy is not enhanced by the fact that I know what information about me is available to others. In fact, such knowledge may threaten my sense of privacy, because it may reveal to me that I am in fact more exposed than I had thought. Most legal systems include in their protection of property rights the liberty and the power to exclude others from entrance. Most legal systems now contain regulation of practices such as wiretapping and electronic eavesdropping. While the details of legal regulation differ, all systems seek to reduce abuses resulting from the existence of extensive data banks. The law has been slow to respond to the implication of technological innovations in the realm of dissemination of information. Cable TV and satellite technologies, as well as the Internet, create fascinating and difficult problems for imposing legal liability for the dissemination of information, including privacy-violating publications.

Legal systems differ also in the extent to which they are willing to grant legal protection to individuals' concerns that true information about them not be published against their will. In the European continent, legal systems protect this interest through a 'right to personality.' England still has no recognition of a general right to privacy, and no legal liability for such publication, while in the US a right to privacy is recognized, but it usually gives way to the paramount interest in free speech. Finally, the question of the extent to which public figures are entitled to the protection of their privacy remains controversial, with different answers given by different legal systems and cultures. The US seems to take the lead in denying its public figures their right to privacy, and in a willingness not to stop short of explicit and detailed public discussion of their most intimate affairs. However, this

area of privacy claims illustrates that, very frequently, social realities are not dictated only by the law. The actual privacy enjoyed by public figures in Western countries depends on cultural and journalistic norms much more than on the legal arrangements existing in them.

Violations of privacy, especially in the field of dissemination and publication of personal information, also illustrate the limits of law in effective protection of values. Once the information is published, or the embarrassing picture aired, there is nothing that can delete it from memory or consciousness. Money awards cannot undo the harm, and the publicity of the trial just adds to the magnitude of the original injury. We encounter here a paradox of law in some areas: Laws may, to some extent, serve as the basis of injunctions, or they may deter people from violations. If they fail in doing that, they are unlikely to be invoked to impose subsequent liability.

See also: Confidentiality and Statistical Disclosure Limitation; Information Society; Personal Privacy: Cultural Concerns; Privacy of Individuals in Social Research; Confidentiality; Public Sphere and the Media; Telecommunications and Information Policy

Bibliography

- Allen A 1988 *Uneasy Access: Privacy for Women in a Free Society*, Totowa, NJ
- Benn S 1971 Privacy, freedom and respect for persons. *Nomos* **13**: 1
- Benn S Gaus G (eds.) 1983 *The Private and Public in Social Policy*, NY
- Bloustein E 1964 Privacy as an aspect of human dignity. **39** *New York University Law Review* 962
- Davis F 1959 What do we mean by the right to privacy? *South Dakota Law Review*. **4**
- DeCew J 1997 *In Pursuit of Privacy: Law, Ethics and the Rise of Technology*. Cornell
- Ferdinand S (ed.) 1984 *Philosophical Dimensions of Privacy*, Cambridge University Press, Cambridge, UK
- Fried C 1968 Privacy, *Yale Law Journal* **77**: 475
- Gavison R 1980 Privacy and the limits of law. *Yale Law Journal* **89**: 421
- Gavison R 1993 Feminism and the private-public distinction. *Stanford Law Review* **45**: 1
- Griswold vs. Connecticut 1965 381 US 479
- Gross H 1971 Privacy and autonomy. *Nomos* **13**: 169
- Henkin L 1974 Privacy and autonomy. *Columbia Law Review* **74**: 1410
- MacKinnon C 1989 *Towards a Feminist Theory of the State*. Cambridge, MA
- Parent W 1983 Privacy, morality and the law. *Law and Philosophy* **2**: 269–88
- Pennock R, Chapman J (eds.) 1971 Privacy, *Nomos* **13**
- Planned Parenthood vs. Casey 1992 505 US 833
- Posner R 1978 The right to privacy. *Georgia Law Review*. **12**: 393
- Prosser W 1960 'Privacy', *California Law Review* **48**: 383
- Roe vs. Wade 1973 410 US 113

Rubinfeld J 1989 The right of privacy. *Harvard Law Review* **102**: 73

Thomson, J J 1975 The right of privacy. *Philosophy and Public Affairs* **4**: 295–314

Westin Allan 1967 *Privacy and Freedom*. NY

R. Gavison

Privacy of Individuals in Social Research: Confidentiality

1. Definitions

The word privacy here refers to the state of the individual, including individuals' right to control information about them at times. Confidentiality refers to the state of information, that is, information in an individual's record may not be disclosed without the individual's consent except under certain conditions. Security here refers to the state of physical or electronic records, especially protection against unauthorized access to the record.

2. An Ethical and Legal Context

People may provide information in social surveys, randomized field trials, and in other studies. Their doing so may be voluntary. Or it may be mandatory, that is, required by law.

Generally, the researcher is under an explicit ethical obligation to assure the privacy of the individual, the confidentiality of information generated on the individual, and the security of the record. These obligations may be reflected in professional codes of ethics. The obligations may be verified through government regulations and review processes that foster assurance of privacy, such as Institutional Review Boards in the United States. Or, the obligation may be verified in law. Statutes that govern the activity of census bureau interviewers and statisticians in countries, such as Sweden, the United States, Canada, and elsewhere, are plain on this account and include penalties for failure to meet the obligation.

3. Assessing Privacy in Surveys and Randomized Field Trials

The social scientist's statutes must be met in a variety of ways at each stage in the process of a study. At the study's design stage, for instance, one minimizes the information that is asked of a respondent and minimizes the intrusiveness of questions, that is, one restricts questions so as to meet the study's object. One

also minimizes the number of people who are engaged in a study. This is determined partly by technical concerns such as the level of statistical precision or statistical power that is needed.

Under many codes of ethics and certain broad legal regulations, individuals to whom questions are put must be told whether their participation in the research is voluntary. People usually must be told the purpose of the research, and the disposition and functions of the data that they provide. Usually, they are assured that the information they provide will remain confidential, that is, the assurance is that information will be used only for research purposes and that the information they provide will not be disclosed with their identity attached to anyone else.

Under certain circumstances, the information provided by individuals can be assured of confidentiality even in the face of government or private threat. In the United States, for example, researchers who are involved in studies on alcohol and substance abuse, and criminological research can receive a certificate of confidentiality. The rules are such that researchers are prohibited under penalty of law from disclosing identifiable microrecords, and are legally immune from any obligation to provide microrecords to law enforcement agencies.

More generally, statutes that govern statistical agencies in the United States and other countries require that the agencies refrain from disclosing records on individuals to other people, that is, the statutes say that records on identifiable individuals cannot be appropriated legally by anyone else, including the courts, legislatures, and law enforcement agencies.

The researcher is under an ordinary obligation to cooperate with courts and other government agencies in providing information. This obligation becomes problematic if the researcher made promises of confidentiality and then, when confronted with a government demand, discovers that the promise has no legal standing. Keeping track of such episodes and their frequency is important. Shield laws that assure that the social scientist cannot be compelled to disclose information about identifiable individuals who have agreed to participate in the research are then important.

4. Privacy and Researchers' Use of Statistical Data Bases

Producing high quality data in social research can be costly irrespective of the benefits of the information. To the extent that the same data set can be put to work frequently to benefit the public, the cost/effectiveness ratio decreases.

For instance, good scholars undertook multiple reanalysis of data from the statewide class size randomized trial in Tennessee. These analyses informed

policy debates about reducing the size of classrooms in 20 states of the United States.

Making local, regional, or national data available to statisticians, however, may put one ethical standard against another: public interest in the results of statistical analysis versus the possibility that records will be used for nonresearch purposes.

Learning how to make individual records available for statistical analysis in ways that do not compromise the privacy of the individuals who are the source of the record information is not easy. Since the 1970s, serious research on this topic has been done by federal statistical agencies in many different countries. Virtually all such agencies concern themselves with surveys rather than active randomized trials on the effects of social programs. Nonetheless, the lessons from the survey research community are applicable to people who attempt to mount evaluations of program effects, such as randomized field trials.

The approaches that have been developed to assure researchers' access to existing data sets without compromising privacy concerns can be classified into two categories. The first involves simple procedures that focus on assuring that records cannot be used for learning about an individual. The second involves legal arrangements, including statutes, regulations, and contractual agreements.

Consider, first, the approaches to making data available and protecting individual privacy. Typically, the data custodian who releases a file for the public's use deletes the direct and clear identifiers from data files, that is, the addresses and other unique information about individuals are deleted. Further, because individual records are often detailed and external collateral information is available, sophisticated statistical approaches may be used to eliminate the possibility of deductive disclosure. Such approaches include deletion of certain variables from the file, eliminating certain microrecords that allow disclosure on small samples of particular individuals, reducing the precision of entry by making the reporting categories crude, introducing controlled random perturbation, controlled random swapping, and other techniques. These technical approaches fall under the rubric of statistical disclosure limitation technology. Advances in this technology by statisticians in the Netherlands, the United Kingdom, the United States, Israel, and Luxemburg are given in Fienberg and Willenborg (1998).

Data enclaves are organizations created to (a) permit intensive sophisticated analyses by people other than the original data custodian, and (b) protect the confidentiality of microrecords with the same rules that govern the original custodian. At the University of California at Los Angeles and at Berkeley, for instance, enclaves have been established that permit university-based researchers to analyze data that would be accessible ordinarily only to federal civil servants. The rules that govern the data's use, in-

cluding security procedures and the threat of criminal and civil sanctions, apply to the enclave's researcher as it does to civil servants.

A variant on the enclave idea involves licensing agreements. A specific example of this arrangement, developed by the National Center for Education Statistics, is a license granted to particular individuals to analyze specific data sets that would not otherwise be accessible on privacy grounds. The justifications for such privilege must be made. The licensing agreement includes requirements for security of the microrecords, sanctions and rules of disclosure, and so on.

5. Privacy and the Use of State Data Bases

In the United States, each state of the 50 states has laws governing third party access to certain administrative record systems and administrative data bases. The state laws differ appreciably in their provisions for permitting a researcher (and certain others) to access individually identifiable microrecords. The administrative records that are laws and regulations also differ across systems within schools, hospitals, child welfare and adult welfare agencies, crime and justice agencies, runaway shelters, and others.

There exists no national compendium for identifying which administrative record systems are legally open to legitimate researchers and which systems are closed to them. A compendium for each country and for each political jurisdiction and agency within each country is important as a way of understanding how privacy of individuals and confidentiality of microrecords can be assured in a research context.

Despite the absence of a national, regional, and local compendium, there exist useful compendiums that have been constructed for special purposes. They identify the State and its laws on particular record systems and summarize access rules for certain individuals including perhaps researchers (see, for instance, the Office of Juvenile Justice and Delinquency Research 1995 on missing children and the relevant databases and the National Criminal Justice Information and Statistics Service (1978, 1988, 1992) on criminal history information. Among federal agencies in the United States, the Bureau of Justice Statistics (BJS) has been consistent in preparing and updating related information. It has produced eight reports since 1972; more recent reports supplant earlier ones. They all focus on state laws regarding criminal history record information.

In the BJS (1992) compendium, each state's laws are cited, the full text of laws being given in a national repository that is itself accessible to researchers. The compendium outlines information about each state law on 28 dimensions. One such dimension is whether the particular law contains 'provision for and regulation of' access to criminal justice information by outside researchers (BJS 1992, p. 12).

The compendium identifies 29 states that have laws or regulations that permit research access and the relevant statute or regulation. No analysis is provided. But researchers can obtain full text documents for the statutes and regulations from the criminal justice clearinghouse.

6. Privacy and Record Linkage

Linking records from different databases in the interest of research presents privacy problems. The rules governing one data custodian, such as a census bureau, may differ from the rules of another statistical agency such as an education statistics or a labor statistics bureau. Such rules differ across agencies for a less centralized government such as the United States. They may not differ across agencies for a more centralized government such as Canada, Sweden, or Japan. The matter is very complex and generalizations are hard to make.

Certainly, the case for linkage has to be made. Identifying the discernable benefits to society of a linkage and the absence of discernable costs to individual privacy is important. This can be done through example and systematic study. An early example of an effort to discover the benefits and costs of linkage for administrative purposes, not research purposes, is given in USGAO (1986a, 1986b). The bottom line was that certain administrative linkages are not worth the effort. Recent examples, more to the point for social scientists, are given in the work of the National Research Council (1999).

See also: Confidentiality and Statistical Disclosure Limitation; Ethical Dilemmas: Research and Treatment Priorities; Research Conduct: Ethical Codes; Research Ethics, Cross-cultural Dimensions of; Research Subjects, Informed and Implied Consent of

Bibliography

- Boruch R F 199 *Randomized Experiments for Planning and Evaluation*. Sage, Thousand Oaks, CA
- Boruch R F, Cecil J S 1979 *Assuring the Confidentiality of Social Research Data*. University of Pennsylvania Press, Philadelphia, PA
- Bureau of Justice Statistics 1992 *Criminal History Record Information: Compendium of State Privacy and Security Legislation, 1992, Current Status by Category and State*. US Department of Justice, Washington, DC
- Cecil J S, Wetherington G T (eds.) 1996 Court ordered disclosure of academic research: A clash of values of science and law. *Law and Contemporary Problems* 59(3): 1–7
- Donaldson M S, Lohr K (eds.) 1994 *Health Data in an Information Age: Use, Disclosure, and Privacy*. National Academy Press, Washington, DC

- Duncan G T, Jabine T, deWolf V A (eds.) 1993 *Private Lives and Public Policies: Confidentiality and Accessibility of Government Statistics*. National Academy Press, Washington, DC
- Fienberg T, Willenborg L C R J (eds.) 1999 Disclosure limitation methods for protecting the confidentiality of statistical data. *Journal of Official Statistics* 14(4)
- Fox J A, Tracey P E 1986 *Randomized Response*. Sage, Beverly Hills, CA
- National Research Council 1999 Committee on National Statistics Workshop on Confidentiality of and Access to Research Files, October 14–15, 1999 Work Shop Papers. National Research Council, Washington, DC
- Sigma Xi, The Scientific Research Society. *The Responsible Research: Paths and Pitfalls*. Sigma Xi, Research Triangle Park, NC
- Soecken K L 1987 Randomized response methodology in health research. *Evaluation Health Professions* 10(1): 58–66
- Takas M, Bass D (eds.) 1995 *Addressing Confidentiality of Records in Searches for Missing Children: Final Report*. Office of Juvenile Justice and Delinquency Prevention, Washington, DC
- US General Accounting Office 1986a *Computer Matching: Factors Influencing the Agency Decision-making Process* (PEMD – 87 – 3 BR). USGAO, Washington, DC
- US General Accounting Office 1986b *Computer Matching: Assessing Its Costs and Benefits* (PEMD – 87 – 2). USGAO, Washington, DC

R. F. Boruch

Privatization of the Economy

State ownership has characterized most market and, *a fortiori*, planning economies in the twentieth century, the United States being the most notable, albeit partial, exception. To an extent that has somewhat varied across countries and decades, financial, strategic, political, and technological reasons have been advanced to justify either the nationalization of existing private-sector firms or the *ex novo* creation of state-owned enterprises (SOEs). State ownership of utilities has been justified on the ground of technological conditions leading to natural monopolies, external economies, and diverging social and private discount rates. Moreover, many governments argued that only the state could overcome critical ‘bottle-necks’ in terms of physical, financial, and human capital investments required by industrialization. A further motive has been the desire to keep domestic control on monopoly rents produced exploiting non-renewable natural resources. On weaker economic grounds, SOEs have been established to build a coalition in support of new regimes, to assure national independence in strategic industries, or to avoid that unpopular minorities could eventually control the economy. SOEs were used for different distributional

aims—to make available essential goods and services, to create new jobs, to reduce geographical concentration of economic power—but they were also crucial elements for playing patronage politics through jobs and the servicing of constituencies. A further explanation for the diffusion of state ownership in semiperipheral countries, especially in Latin America, was advanced by the dependency literature. While accepting the primacy of private ownership, ‘bureaucratic-authoritarian regimes’ considered that the role of the state had to go beyond that of a subsidiary character to that of supplying the complementary inputs to the process of private capital accumulation (Evans 1978).

Since the rise of oil prices in the 1970s, the whole postwar model of economic development has been heavily criticized, and SOEs have not been exempted from this swing in the ideological pendulum from the structuralist to the neo-liberal end. The appropriateness of public management of productive units was put in doubt, and SOEs were found to have negative income distribution effects, to worsen budget and trade balances, and to divert resources from much more needed social goals, such as education and health expenditures. The efficiency of SOEs was found to be lower *ceteris paribus* than in the private sector. Inefficient SOEs do not face the risk of bankruptcy, and the market for corporate control, because of the state’s tight grip through majority stakes, cannot act as an adequate device for disciplining managers. In theory, this concentration of ownership in the hands of a single majority investor could have served to circumvent the collective action problem that impedes small, dispersed shareholders in listed companies to wield efficient monitoring on corporate managers. This relative advantage of public ownership, however, has often been offset by agency problems stemming from the multiplicity of ties linking the government, ministries with specific competencies, Parliament, political parties, and the management of SOEs, all with different goals, possibly at cross-purpose. Moreover, SOEs’ managers are often appointed for political reasons, rather than for their management skills. As regards the utilities more specifically, fast technological advances have substantially weakened the argument that public ownership of integrated monopolies is necessary to assure investment growth, service quality, and price declines in real terms.

Over the last two decades, the initially timid progress recorded in some pioneering countries (United Kingdom, Chile, New Zealand) were matched and sometimes even surpassed by new countries joining the privatization bandwagon, first in the developing world (Argentina, Mexico, Malaysia), then in Europe (France, Italy), and finally in transition economies. (Possibly the first large-scale sale of a public enterprise was the Volkswagen issue in the early 1960s. Nonetheless, the dramatic fall of the share price after markets

were hit by the Cuban missile crisis had a lasting negative impact on equity ownership culture in Germany.) Yet, while a central issue in the debate for almost two decades, privatization has been a rather restricted phenomenon from a quantitative point of view (Shirley 1998). Selling SOEs is part of the political game: it requires the creation of supportive coalitions and it impacts on the distribution of power resources. Strong vested interests, such as organized labor, the military, parts of the urban bourgeoisie, and most economic groups, mount in defense of nationalization. Moreover, privatization is a typical example of a collective action situation: actors more likely to gain from it are widely dispersed, and it is thus simpler for beneficiaries of the status quo to organize an opposition than it is for future beneficiaries to create a support group. Finally, in developing countries where capital markets are thin and scarcely organized, and where it is thus difficult to realize popular capitalism through public placing of SOEs, the potentialities of privatization in terms of coalition building are rather remote.

In practice, the intrinsic allure of free-market economics has proven less important a factor for the success of privatizations than more conjunctural conditions (Armijo 1998). The nature and the seriousness of crisis deeply influence both the perceived need of radical changes and the central themes of the political debate. The tighter the fiscal constraint, and the greater the contribution of SOEs to the deficit, the quicker the recourse to divestiture to solve problems. An important issue is therefore the choice of firms to be sold and the method of sale. Understanding the economic goals to be met through divestiture, in particular the choice between an immediate reduction of macroeconomic imbalances and the medium-term achievement of efficiency improvements by liberalization and more effective regulation, is crucial. The least productive enterprises are harder to dispose of, which might make it necessary to concede easy terms well beyond a low price, when not liquidating them altogether. At the same time, countries in dire economic straits need quick ‘leading cases’ to build their reputation *vis-à-vis* foreign investors. The choice has thus often been to start from the best managed and more profitable SOEs, such as telecommunications, at the risk of making learning mistakes in those very cases where state assets should be more coveted by international investors. Moreover, such industries present technological and tariff complexities, requiring sophisticated tools for asset valuation and regulation that developing countries seldom possess.

Timing and organizational structure also matter. New political incumbents have higher chances of implementing reforms, thanks not only to the ‘post-election honeymoon,’ but also because, according to the electoral cycle theory, there are less incentives to put off adopting risky and painful actions. Likewise, although the relationship between regimes and econ-

omic policies is indeterminate, being able to use emergency powers opens smoother roads on the reforms' path. 'Insulated change teams' or 'technopols' are important, especially where governments came to power with the backing of the lower class and need to prove themselves credible in the face of international investors' fears. However, not all technopols are alike, and the priority they assign to privatization (and the resources they demand politicians to earmark for this as opposed to other structural measures) also depends on their ideological values and on their appraisal of their country's riddle.

Third, given the short-term adjustment costs of economic reforms, it is necessary to introduce compensatory mechanisms. While scholarly attention has been focused on specific, targeted measures to face the reduction of fiscal subsidies to the poor, large business groups that dominate the industrial sector in most developing countries are hit by falling trade barriers, positive real interest rates, and more careful awarding of public credit. In order to win their support, some countervailing measures may be needed, such as cheap (relative to book value) sales of SOEs (Schamis 1998).

Studies of privatization in industrial countries have reported very positive results in firms' performance (e.g., Meggison and Netter 1999). As for developing economies, there are significant increases in profitability, operating efficiency, capital spending, output, and employment, which are usually greater in countries with higher per capita income (Boubakri and Cosset 1998). (In the case of transition economies, empirical tests of the relationship between enterprise performance and ownership generally refute the hypothesis that privatization *per se* is associated with improved performance (e.g., Estrin and Rosevear 1999). The belief that mass privatization, by providing powerful incentives for efficient restructuring, would release entrepreneurial endeavors has also proved naïf (e.g., Spicer et al. 1999).) The substitution of private to public ownership, however, has been accompanied by the emergence of new regulatory challenges for policy-makers. First, in the hitherto natural monopolies that constituted the core of public enterprises (telecoms, energy, water, transport), how to prevent the new private owners from simply pocketing monopoly rents. Especially in the electricity industry, weaknesses in the regulatory framework have sometimes reduced the benefits of privatization and deregulation (Nicoletti 2000). Public authorities must devise sectoral policies that introduce and maintain competition; establish and maintain a sound regulatory framework for the remaining monopolies, public and private; keep transparency in transactions and convince investors that their investments are secure; negotiate, monitor, and enforce contracts with private suppliers of management and financing; ensure that resources from privatization sales are put to productive uses; and manage the inevitable political and social tensions that arise as enterprise reforms are

implemented, especially the critical issues of foreign ownership and labor layoffs. There is, however, much less agreement on how to approach the next set of challenges (second-generation issues) for countries facing the consolidation of initial reforms. In general, these issues are related to postprivatization disputes and renegotiations between governments and the private sector and to the mechanisms necessary to promote competition in the reformed industries.

Second, in view of the goal of creating a people's capitalism, how to provide appropriate mechanisms of corporate governance that protect small shareholders, while also allowing management the required flexibility to pursue long-term corporate goals. Advances in the theory of corporate finance and industrial organization, as well as the ongoing debate over the 'best' capitalist model, have made the analysis of corporate governance—the mechanisms whereby economic systems (in their broadest possible sense) cope with the information and incentive problems inherent in financing investments and facilitate the intertemporal transfer of income claims—a burgeoning theme in comparative institutional economics. Key elements of this literature are the sources of financial resources for corporations, the concentration of ownership, the relevance of listing, the role and composition of boards of directors, the rules governing the market for corporate control and the obstacles that they may pose to corporate control activity, and the relative importance of the voice and exit mechanisms (boardroom pressures and takeovers, or internal and external control, respectively) in disciplining managers.

Links between privatization and corporate governance are of two main kinds: selling SOEs exposes them to the takeover and bankruptcy threats, thereby easing the corporate governance problems proper of public ownership, and it provides an opportunity to modify the distribution of ownership rights among different classes of investors by extending public listing among large firms, increasing the number of small shareholders, and reducing ownership concentration. But while creating 'people's capitalism' has always ranked high among governments' goals, tackling the issues that remain after ownership is transferred from public to private hands and when no (absolute) majority shareowner emerges has seldom been an overriding concern. Potential improvements in technical efficiency following control transfer may be jeopardized if corporate control is not contestable: in this case, and especially if conduct regulation proves insufficient to open up protected markets, managers can exploit rents accruing from market position without having to worry about the threat of takeovers. Privatization policies should take such elements into account, making it imperative to introduce reforms and redress perceived inefficiencies. Governments have great discretion in pricing the SOEs they sell, especially those being sold via public share offering, and they use

this discretion to pursue political and economic ends. Most experiences, however, suggest the limited power of privatization in changing the modes of governance which are prevalent in each country's large private companies. Further, those countries which have chosen the mass (voucher) privatization route have done so largely out of necessity and face ongoing efficiency problems as a result. In the UK, a country whose privatization policies are often referred to as a benchmark, 'control [of privatized companies] is not exerted in the forms of threats of take-over or bankruptcy; nor has it for the most part come from direct investor intervention' (Bishop et al. 1994, p. 11). After the steep rise experienced in the immediate aftermath of privatizations, the slow but constant decline in the number of small shareholders highlights the difficulties in sustaining people's capitalism in the longer run. In Italy, for example, privatization was accompanied by a legislative effort aimed at providing noncontrolling shareholders (i.e. both individual and collective investors) with more adequate safeguards and at introducing the necessary conditions to allow them to monitor managers. But successive governments were unsuccessful in broadening the number of large private business groups, whereas enhancing the mobility of control to investors outside the traditional core of Italy's capitalism was explicitly included among the authorities' strategic goals. On the other hand, experiences such as those of the UK and Chile underscore that mass sell-offs require the development of new institutional investors such as pension funds that may later play an active role in corporate governance. (In Chile, for example, the takeover of the country's dominant electricity utility, Enersis, one of the largest in emerging markets, was stalled for some months in 1998 as pension funds' questioned lucrative additional terms that the management had negotiated for themselves based on important agreements concerning the future strategic direction of Enersis that they had never told other shareholders about.)

To conclude, although privatization's promise has been frequently oversold, not least by international organizations, its ills have also been greatly exaggerated. When ownership transfer has been accompanied by market liberalization and proper implementation, in OECD and non-OECD countries alike, consumers and end-users have benefited in terms of choice, quality, and prices. If public opinion and policy-makers wish so, much remains to be sold and the future challenges for the regulatory state remain substantial to ensure maximization of welfare benefits.

See also: Business Law; Capitalism: Global; Corporate Finance: Financial Control; Corporate Governance; Corporate Law; International Law and Treaties; International Trade: Commercial Policy and Trade Negotiations; Law and Economics; Law and Economics: Empirical Dimensions; Liberalism; Markets

and the Law; Monetary Policy; Multinational Corporations; Regulation, Economic Theory of; Regulation: Working Conditions; Smith, Adam (1723–90); Socialism; Venture Capital

Bibliography

- Armijo L 1998 Balance sheet or ballot box? Incentives to privatize in emerging democracies. In: Oxhorn P, Starr P (eds.) *The Problematic Relationship between Economic and Political Liberalization*. Lynne Rienner, Boulder, CO
- Bishop M, Kay J, Mayer C 1994 Introduction: privatization in performance. In: Bishop M, Kay J, Mayer C (eds.) *Privatization and Economic Performance*. Oxford University Press, Oxford, UK
- Boubakri N, Cosset J-C 1998 The financial and operating performance of newly privatized firms: evidence from developing countries. *Journal of Finance* **53**: 1081–110
- Estrin S, Rosevear A 1999 Enterprise performance and ownership: The case of Ukraine. *European Economic Review* **43**: 1125–36
- Evans P 1978 *Dependent Development*. Princeton University Press, Princeton, NJ
- Meggison W, Netter J 1999 From state to market: a survey of empirical studies on privatization. *Fondazione Eni Enrico Mattei Working Paper*, No. 01-99
- Nicoletti G 2000 The implementation and the effects of regulatory reform: past experience and current issues. *Mimeo*. OECD Economics Department, Paris
- Schamis H M 1998 Distributional coalitions and the politics of economic reform in Latin America. *World Politics* **51**: 236–68
- Shirley M M 1998 Bureaucrats in business: the roles of privatization versus corporatization in state-owned enterprise reform. *World Development* **27**: 115–36
- Spicer A, McDermott G, Kogut B 1999 Entrepreneurship and privatization in Central Europe: the tenuous balance between destruction and creation. The Wharton School. *Reginald H. Jones Center Working Paper*, No. 99-04

A. E. Goldstein

Probabilistic Grammars and their Applications

1. Introduction

Natural language processing is the use of computers for processing natural language text or speech. Machine translation (the automatic translation of text or speech from one language to another) began with the very earliest computers (Kay et al. 1994). Natural language interfaces permit computers to interact with humans using natural language, for example, to query databases. Coupled with speech recognition and speech synthesis, these capabilities will become more important with the growing popularity of portable computers that lack keyboards and large display

screens. Other applications include spell and grammar checking and document summarization. Applications outside of natural language include compilers, which translate source code into lower-level machine code, and computer vision (Fu 1974, 1982).

Most natural language processing systems are based on formal grammars. The development and study of formal grammars is known as computational linguistics. A *grammar* is a description of a language; usually it identifies the sentences of the language and provides descriptions of them, for example, by defining the phrases of a sentence, their interrelationships, and perhaps also aspects of their meanings. *Parsing* is the process of recovering a sentence's description from its words, while *generation* is the process of translating a meaning or some other part of a sentence's description into a grammatical or well-formed sentence. Parsing and generation are major research topics in their own right. Evidently, human use of language involves some kind of parsing and generation process, as do many natural language processing applications. For example, a machine translation program may parse an input language sentence into a (partial) representation of its meaning, and then generate an output language sentence from that representation.

Modern computational linguistics began with Chomsky (1957), and was initially dominated by the study of his 'transformational' grammars. These grammars involved two levels of analyses, a 'deep structure' meant to capture more-or-less simply the meaning of a sentence, and a 'surface structure' which reflects the actual way in which the sentence was constructed. The deep structure might be a clause in the active voice, 'Sandy saw Sam,' whereas the surface structure might involve the more complex passive voice, 'Sam was seen by Sandy.'

Transformational grammars are computationally complex, and in the 1980s several linguists came to the conclusion that much simpler kinds of grammars could describe most syntactic phenomena, developing Generalized Phrase-Structure Grammars (Gazdar et al. 1985) and Unification-based Grammars (Kaplan and Bresnan 1982, Pollard and Sag 1987, Shieber 1986). These grammars generate surface structures directly; there is no separate deep structure and therefore no transformations. These kinds of grammars can provide very detailed syntactic and semantic analyses of sentences, but even today there are no comprehensive grammars of this kind that fully accommodate English or any other natural language.

Natural language processing using handcrafted grammars suffers from two major drawbacks. First, the syntactic coverage offered by any available grammar is incomplete, reflecting both our lack of understanding of even relatively frequently occurring syntactic constructions and the organizational difficulty of manually constructing any artifact as complex as a grammar of a natural language. Second, such grammars almost always permit a large number of

spurious ambiguities, that is, parses which are permitted by the rules of syntax but have unusual or unlikely semantic interpretations. For example, in the sentence 'I saw the boat with the telescope,' the prepositional phrase 'with the telescope' is most easily interpreted as the instrument used in seeing, while in 'I saw the policeman with the rifle,' the prepositional phrase usually receives a different interpretation in which the policeman has the rifle. Note that the corresponding alternative interpretation is marginally accessible for each of these sentences: in the first sentence one can imagine that the telescope is on the boat, and in the second, that the rifle (say, with a viewing scope) was used to view the policeman.

In effect, there is a dilemma of coverage. A grammar rich enough to accommodate natural language, including rare and sometimes even 'ungrammatical' constructions, fails to distinguish natural from unnatural interpretations. But a grammar sufficiently restricted so as to exclude what is unnatural fails to accommodate the scope of real language. These observations led, in the 1980s, to a growing interest in stochastic approaches to natural language, particularly to speech. Stochastic grammars became the basis of speech recognition systems by outperforming the best of the systems based on deterministic handcrafted grammars. Largely inspired by these successes, computational linguists began applying stochastic approaches to other natural language processing applications. Usually, the architecture of such a stochastic model is specified manually, while the model's parameters are estimated from a *training corpus*, that is, a large representative sample of sentences.

As explained in the body of this article, stochastic approaches replace the binary distinctions (grammatical vs. ungrammatical) of nonstochastic approaches with probability distributions. This provides a way of dealing with the two drawbacks of nonstochastic approaches. Ill-formed alternatives can be characterized as extremely low probability rather than ruled out as impossible, so even ungrammatical strings can be provided with an interpretation. Similarly, a stochastic model of possible interpretations of a sentence provides a method for distinguishing more plausible interpretations from less plausible ones.

The introduction sets out formally various classes of grammars and languages. Probabilistic grammars are introduced in Sect. 2, along with the basic issues of parametric representation, inference, and computation.

2. Grammars and Languages

The formal framework, whether used in a transformational grammar, a generalized phrase-structure grammar, or a more traditionally styled context-free grammar, is due to Chomsky (1957) and his co-

workers. This section presents a brief introduction to this framework. But for a thorough (and very readable) presentation we highly recommend the book by Hopcroft and Ullman (1979).

The concept and notation of a formal grammar is perhaps best introduced by an example:

Example 1: Define a grammar, G_1 , by $G_1 = (T_1, N_1, S, R_1)$, where $T_1 = \{\text{grows, rice, wheat}\}$ is a set of words (a 'lexicon'), $N_1 = \{S, \text{NP}, \text{VP}\}$ is a set of symbols representing grammatically meaningful strings of words, such as clauses or parts of speech (e.g., S for 'Sentence,' NP for 'Noun Phrase,' VP for 'Verb Phrase'), and $R_1 = \{S \rightarrow \text{NP VP}, \text{NP} \rightarrow \text{rice}, \text{NP} \rightarrow \text{wheat}, \text{VP} \rightarrow \text{grows}\}$ is a collection of rules for rewriting, or instantiating, the symbols in N_1 . Informally, the nonterminal S rewrites to sentences or clauses, NP rewrites to noun phrases and VP rewrites to verb phrases. The language, L_{G_1} , generated by G_1 is the set of strings of words that are reachable from S through the rewrite rules in R_1 . In this example, $L_{G_1} = \{\text{rice grows, wheat grows}\}$, derived by $S \Rightarrow \text{NP VP} \Rightarrow \text{rice VP} \Rightarrow \text{rice grows}$, and $S \Rightarrow \text{NP VP} \Rightarrow \text{wheat VP} \Rightarrow \text{wheat grows}$.

More generally, if T is a finite set of symbols, let T^* be the set of all strings (i.e., finite sequences) of symbols of T , including the empty string, and let T^+ be the set of all nonempty strings of symbols of T . A language is a subset of T^* . A *rewrite grammar* G is a quadruple $G = (T, N, S, R)$, where T and N are disjoint finite sets of symbols (called the *terminal* and *nonterminal* symbols respectively), $S \in N$ is a distinguished nonterminal called the *start* or *sentence symbol*, and R is a finite set of productions. A *production* is a pair (α, β) where $\alpha \in N^+$ and $\beta \in (N \cup T)^*$; productions are usually written $\alpha \rightarrow \beta$. Productions of the form $\alpha \rightarrow \varepsilon$, where ε is the empty string, are called *epsilon productions*. This entry restricts attention to grammars without epsilon productions, that is, $\beta \in (N \cup T)^+$, as this simplifies the mathematics considerably.

A rewrite grammar G defines a *rewriting relation* $\Rightarrow_G \subseteq (N \cup T)^* \times (N \cup T)^*$ over pairs of strings consisting of terminals and nonterminals as follows: $\beta A \gamma \Rightarrow \beta \alpha \gamma$ iff $A \rightarrow \alpha \in R$ and $\beta, \gamma \in (N \cup T)^*$ (the subscript G is dropped when clear from the context). The reflexive, transitive closure of $\Rightarrow$ is denoted $\Rightarrow^*$. Thus $\Rightarrow^*$ is the rewriting relation using arbitrary finite sequences of productions. (It is called 'reflexive' because the identity rewrite, $\alpha \Rightarrow \alpha$, is included.) The language generated by G , denoted L_G , is the set of all strings $w \in T^+$ such that $S \Rightarrow^* w$.

Rewrite grammars are traditionally classified by the shapes of their productions. $G = (T, N, S, R)$ is a *context-sensitive* grammar if for all productions $\alpha \rightarrow \beta \in R$, $|\alpha| \leq |\beta|$, that is, the right-hand side of each production is not shorter than its left-hand side. G is a *context-free* grammar if $|\alpha| = 1$, that is, the left-hand side of each production consists of a single nonterminal. G is a *left-linear* grammar if G is context-free

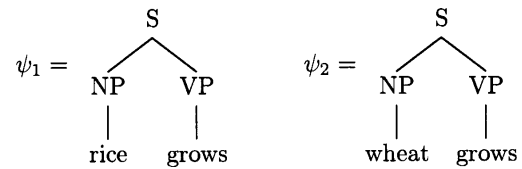
and β (the right-hand side of the production) is either of the form Aw or of the form w where $A \in N$ and $w \in T^*$; in a *right-linear* grammar β always is of the form wA or w . A right or left-linear grammar is called a *regular* grammar. G_1 , in Example 1, is context sensitive and context free.

It is straightforward to show that the classes of languages generated by these classes of grammars stand in strict equality or subset relationships. Specifically, the class of languages generated by right-linear grammars is the same as the class generated by left-linear grammars; this class is called the *regular* languages, and is a strict subset of the class of languages generated by context-free grammars, which is a strict subset of the class of languages generated by context-sensitive grammars, which in turn is a strict subset of the class of languages generated by rewrite grammars.

It turns out that context-sensitive grammars (where a production rewrites more than one nonterminal) have not had many applications in natural language processing, so from here on we will concentrate on context-free grammars, where all productions take the form $A \rightarrow \beta$, where $A \in N$ and $\beta \in (N \cup T)^+$.

An appealing property of grammars with productions in this form is that they induce tree structures on the strings that they generate. And, as Sect. 3 shows, this is the basis for bringing in probability distributions and the theory of inference. We say that the context-free grammar $G = (T, N, S, R)$ generates the labeled, ordered tree ψ , iff the root node of ψ is labeled S , and for each node n in ψ , either n has no children and its label is a member of T (i.e., it is labelled with a terminal) or else there is a production $A \rightarrow \beta \in R$ where the label of n is A and the left-to-right sequence of labels of n 's immediate children is β . It is straightforward to show that w is in L_G if G generates a tree ψ whose *yield* (i.e., the left-to-right sequence of terminal symbols labeling ψ 's leaf nodes) is w ; ψ is called a *parse tree* of w (with respect to G). In what follows, we define Ψ_G to be the set of parse trees generated by G , and $\mathcal{Y}(\cdot)$ to be the function that maps trees to their yields.

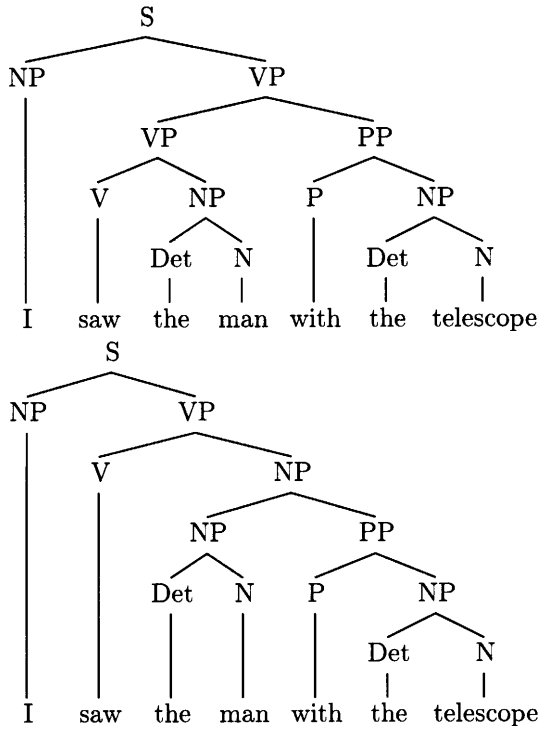
Example 1 (continued): The grammar G_1 defined above generates the following two trees, ψ_1 and ψ_2 .



In this example, $\mathcal{Y}(\psi_1) = \text{'rice grows'}$ and $\mathcal{Y}(\psi_2) = \text{'wheat grows.'}$

A string of terminals w is called *ambiguous* if w has two or more parse trees. Linguistically, each parse tree of an ambiguous string usually corresponds to a distinct interpretation.

Example 2: Consider $G_2 = (T_2, N_2, S, R_2)$, where $T_2 = \{\text{I, saw, the, man, with, telescope}\}$, $N_2 = \{S, NP, N, Det, VP, V, PP, P\}$ and $R_2 = \{S \rightarrow NP, VP, NP \rightarrow I, NP \rightarrow Det\ N, Det \rightarrow \text{the}, N \rightarrow NP, PP, N \rightarrow \text{man}, N \rightarrow \text{telescope}, VP \rightarrow V, NP, VP \rightarrow VP, PP, PP \rightarrow P, NP, V \rightarrow \text{saw}, P \rightarrow \text{with}\}$. Informally, N rewrites to nouns, Det to determiners, V to verbs, P to prepositions and PP to prepositional phrases. It is easy to check that the two trees ψ_3 and ψ_4 with the yields $\mathcal{Y}(\psi_3) = \mathcal{Y}(\psi_4) = \text{'I saw the man with the telescope'}$ are both generated by G_2 . Linguistically, these two parse trees represent two different syntactic analyses of the sentence. The first analysis corresponds to the interpretation where the seeing is by means of a telescope, while the second corresponds to the interpretation where the man has a telescope.



3. Probability and Statistics

Obviously broad coverage is desirable—natural language is rich and diverse, and not easily held to a small set of rules. But it is hard to achieve broad coverage without massive ambiguity (a sentence may have tens of thousands of parses), and this of course complicates applications like language interpretation, language translation, and speech recognition. This is the dilemma of coverage that we referred to earlier, and it sets up a compelling role for probabilistic and statistical methods.

We briefly review the main probabilistic grammars and their associated theories of inference. We begin in Sect. 3.1 with probabilistic regular grammars, also known as hidden Markov models (HMM), which are the foundation of modern speech recognition systems—see Jelinek (1997) for a survey.

In Sect. 3.2 we discuss probabilistic context-free grammars, which turn out to be essentially the same thing as branching processes. Finally, in Sect. 3.3, we take a more general approach to placing probabilities on grammars, which leads to Gibbs distributions, a role for Besag's pseudolikelihood method (Besag 1974, 1975), various computational issues, and, all in all, an active area of research in computational linguistics.

3.1 Regular Grammars

We will focus on right-linear grammars, but the treatment of left-linear grammars is more or less identical. It is convenient to work with a *normal form*: all rules are either of the form $A \rightarrow bB$ or $A \rightarrow b$, where $A, B \in N$ and $b \in T$. It is easy to show that every right-linear grammar has an equivalent normal form in the sense that the two grammars produce the same language. Essentially nothing is lost.

3.1.1 Probabilities. The grammar G can be made into a probabilistic grammar by assigning to each non terminal $A \in N$ a probability distribution p over productions of the form $A \rightarrow \alpha \in R$: for every $A \in N$

$$\sum_{\substack{\alpha \in (N \cup T)^+ \\ \text{s.t. } (A \rightarrow \alpha) \in R}} p(A \rightarrow \alpha) = 1 \quad (1)$$

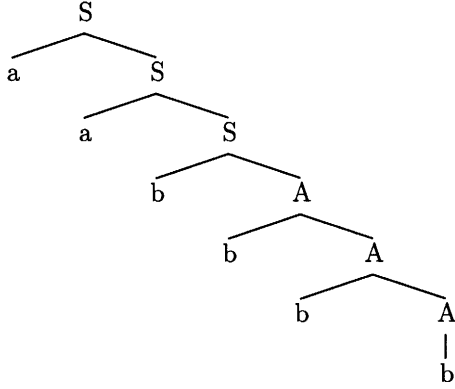
Recall that Ψ_G is the set of parse trees generated by G (see Sect. 2). If G is linear, then $\psi \in \Psi_G$ is characterized by a sequence of productions, starting from S . It is, then, straightforward to use p to define a probability P on Ψ_G : just take $P(\psi)$ (for $\psi \in \Psi_G$) to be the product of the associated production probabilities.

Example 3: Consider the right-linear grammar $G_3 = (T_3, N_3, S, R_3)$, with $T_3 = \{a, b\}$, $N_3 = \{S, A\}$ and the productions (R_3) and production probabilities (p):

$S \rightarrow aS$	$p = 0.80$
$S \rightarrow bS$	$p = 0.01$
$S \rightarrow bA$	$p = 0.19$
$A \rightarrow bA$	$p = 0.90$
$A \rightarrow b$	$p = 0.10$

The language is the set of strings ending with a sequence of at least two bs . The grammar is ambiguous: in general, a sequence of terminal states does not uniquely identify a sequence of productions. The sentence $aabbbb$ has three parses (determined by the placement of the production $S \rightarrow bA$), but the most

likely parse, by far, is $S \rightarrow aS$, $S \rightarrow aS$, $S \rightarrow bA$, $A \rightarrow bA$, $A \rightarrow bA$, $A \rightarrow b$ ($P = 0.8 \cdot 0.8 \cdot 0.19 \cdot 0.9 \cdot 0.1$), which has a *posterior* probability of nearly 0.99. The corresponding parse tree is shown below.



3.1.2 Inference. The problem is to estimate (see *Estimation: Point and Interval*) the transition probabilities, $p(\cdot)$, either from parsed data (examples from Ψ_c) or just from sentences (examples from L_c). Consider first the case of parsed data ('supervised learning'), and let $\psi_1, \psi_2, \dots, \psi_n \in \Psi$ be a sequence taken iid according to P . If $f(A \rightarrow \alpha; \psi)$ is the counting function, counting the number of times transition $A \rightarrow \alpha \in R$ occurs in ψ , then the likelihood function (see *Likelihood in Statistics*) is

$$L = L(p; \psi_1, \dots, \psi_n) = \prod_{i=1}^n \prod_{A \rightarrow \alpha \in R} p(A \rightarrow \alpha)^{f(A \rightarrow \alpha; \psi_i)} \quad (2)$$

The maximum likelihood estimate is, sensibly, the relative frequency estimator:

$$\hat{p}(A \rightarrow \alpha) = \frac{\sum_{i=1}^n f(A \rightarrow \alpha; \psi_i)}{\sum_{i=1}^n \sum_{A \rightarrow \beta \in R} f(A \rightarrow \beta; \psi_i)} \quad (3)$$

The problem of estimating p from sentences ('unsupervised learning') is more interesting, and more important for applications. Recall that $\mathcal{Y}(\psi)$ is the yield of ψ , that is, the sequence of terminals in ψ . Given a sentence $w \in T^+$, let Ψ_w be the set of parses which yield w : $\Psi_w = \{\psi \in \Psi: \mathcal{Y}(\psi) = w\}$. Imagine a sequence $\psi_1, \dots, \psi_n$, iid according to P , for which only the corresponding yields, $w_i = \mathcal{Y}(\psi_i)$ $1 \leq i \leq n$, are observed. The likelihood function is

$$\begin{aligned} L &= L(p; w_1, \dots, w_n) \\ &= \prod_{i=1}^n \sum_{\psi \in \Psi_{w_i}} \prod_{A \rightarrow \alpha \in R} p(A \rightarrow \alpha)^{f(A \rightarrow \alpha; \psi_i)} \end{aligned} \quad (4)$$

As is usual with hidden data, there is an EM-type iteration for climbing the likelihood surface—see Baum (1972) and Dempster et al. (1977):

$$t+1(A \rightarrow \alpha) = \frac{\sum_{i=1}^n E_{\hat{p}_t}[f(A \rightarrow \alpha; \psi) | \psi \in \Psi_{w_i}]}{\sum_{A \rightarrow \beta \in R} \sum_{i=1}^n E_{\hat{p}_t}[f(A \rightarrow \beta; \psi) | \psi \in \Psi_{w_i}]} \quad (5)$$

Needless to say, nothing can be done with this unless we can actually evaluate, in a computationally feasible way, expressions like $E_{\hat{p}_t}[f(A \rightarrow \alpha; \psi) | \psi \in \Psi_{w_i}]$. This is one of several closely related computational problems that are part of the mechanics of working with regular grammars.

3.1.3 Computation. A sentence $w \in T^+$ is *parsed* by finding a sequence of productions $A \rightarrow bB \in R$ which yield w . Depending on the grammar, this corresponds more or less to an interpretation of w . Often, there are many parses and we say that w is ambiguous. In such cases, if there is a probability p on R then there is a probability P on Ψ , and a reasonably compelling choice of parse is the most likely parse:

$$\arg \max_{\psi \in \Psi_w} P(\psi) \quad (6)$$

This is the *maximum a posteriori* (MAP) estimate of ψ —obviously it minimizes the probability of error under the distribution P . (Of course, in those cases in which Eqn. (6) is small, P does little to make w unambiguous.)

What is the probability of w ? How are its parses computed? How is the most likely parse computed? These computational issues turn out to be more-or-less the same as the issue of computing $E_{\hat{p}}[f(A \rightarrow \alpha; \psi) | \psi \in \Psi_w]$ that came up in our discussion of inference. The basic structure and cost of the computational algorithm is the same for each of the four problems—compute the probability of w , compute the set of parses, compute the best parse, compute $E_{\hat{p}}$. In particular, there is a simple dynamic programming solution to each of these problems, and in each case the complexity is of the order $n \cdot |R|$, where n is the length of w , and $|R|$ is the number of productions in G —see Jelinek (1997), Geman and Johnson (2000). The existence of a dynamic programming principle for regular grammars is a primary reason for their central role in state-of-the-art speech recognition systems.

3.2 Context-free Grammars

Despite the successes of regular grammars in speech recognition, the problems of language understanding and translation are generally better addressed with the more structured and more powerful context-free grammars. Following our development of probab-

ilistic regular grammars in the previous section, we will address here the interrelated issues of fitting context-free grammars with probability distributions, estimating the parameters of these distributions, and computing various functionals of these distributions.

The context-free grammars $G = (T, N, S, R)$ have rules of the form $A \rightarrow \alpha, \alpha \in (N \cup T)^+$, as discussed previously in Sect. 2. There is again a normal form, known as the Chomsky normal form, which is particularly convenient when developing probabilistic versions. Specifically, one can always find a context-free grammar G' , with all productions of the form $A \rightarrow BC$ or $A \rightarrow a, A, B, C \in N, \alpha \in T$, which produces the same language as G : $L_{G'} = L_G$. Henceforth, we will assume that context-free grammars are in the Chomsky normal form.

3.2.1 Probabilities. The goal is to put a probability distribution on the set of parse trees generated by a context-free grammar in Chomsky normal form. Ideally, the distribution will have a convenient parametric form, that allows for efficient inference and computation.

Recall from Sect. 2 that context-free grammars generate labeled, ordered trees. Given sets of nonterminals N and terminals T , let Ψ be the set of finite trees with:

- (a) root node labeled S ;
- (b) leaf nodes labeled with elements of T ;
- (c) interior nodes labeled with elements of N ;
- (d) every nonterminal (interior) node having either two children labeled with nonterminals or one child labeled with a terminal.

Every $\psi \in \Psi$ defines a sentence $w \in T^+$: read the labels off of the terminal nodes of ψ from left to right. Consistent with the notation of Sect. 3.1 we write $\mathcal{Y}(\psi) = w$. Conversely, every sentence $w \in T^+$ defines a subset of Ψ , which we denote by Ψ_w , consisting of all ψ with yield w ($\mathcal{Y}(\psi) = w$). A context-free grammar G defines a subset of Ψ , Ψ_G , whose collection of yields is the language, L_G , of G . We seek a probability distribution P on Ψ which concentrates on Ψ_G .

The time-honored approach to probabilistic context-free grammars is through the production probabilities $p: R \rightarrow [0, 1]$, with

$$\sum_{\substack{\alpha \in N^2 \cup T \\ \text{s.t.}, (A \rightarrow \alpha) \in R}} p(A \rightarrow \alpha) = 1 \quad (7)$$

Following the development in Sect. 3.1, we introduce a counting function $f(A \rightarrow \alpha; \psi)$, which counts the number of instances of the rule $A \rightarrow a$ in the tree ψ , i.e., the number of nonterminal nodes A whose daughter nodes define, left-to-right, the string α . Through f , p induces a probability P on Ψ :

$$P(\psi) = \prod_{(A \rightarrow \alpha) \in R} p(A \rightarrow \alpha)^{f(A \rightarrow \alpha; \psi)} \quad (8)$$

It is clear enough that P concentrates on Ψ_G , and we shall see shortly that this parameterization, in terms of products of probabilities p , is particularly workable and convenient. The pair, G and P , is known as a probabilistic context-free grammar, or PCFG for short. (Notice the connection to branching processes—Harris (1963): Starting at S , use R , and the associated probabilities $p(\cdot)$, to expand nodes into daughter nodes until all leaf nodes are labeled with terminals—i.e., with elements of T .)

3.2.2 Inference. As with probabilistic regular grammars, the production probabilities of a context-free grammar, which amount to a parameterization of the distribution P on Ψ_G , can be estimated from examples. In one scenario, we have access to a sequence $\psi_1, \dots, \psi_n$ from Ψ_G under P . This is ‘supervised learning,’ in the sense that sentences come equipped with parses. More practical is the problem of ‘unsupervised learning,’ wherein we observe only the yields, $\mathcal{Y}(\psi_1), \dots, \mathcal{Y}(\psi_n)$.

In either case, the treatment of maximum likelihood estimation is essentially identical to the treatment for regular grammars. In particular, the likelihood for fully observed data is again Eqn. (2), and the maximum likelihood estimator is again the relative frequency estimator Eqn. (3). And, in the unsupervised case, the likelihood is again Eqn. (4) and this leads to the same EM-type iteration given in Eqn. (5).

3.2.3 Computation. There are again four basic computations: find the probability of a sentence $w \in T^+$, find a $\psi \in \Psi$ (or find *all* $\psi \in \Psi$) satisfying $\mathcal{Y}(\psi) = w$ (‘parsing’); find

$$\arg \max_{\substack{\psi \in \Psi \text{ s.t.} \\ \mathcal{Y}(\psi) = w}} P(\psi)$$

(‘maximum *a posteriori*’ or ‘optimal’ parsing); compute expectations of the form $E_{\hat{p}}[f(A \rightarrow \alpha; \psi) | \psi \in \Psi_w]$ that arise in iterative estimation schemes like (5). The four computations turn out to be more-or-less the same, as was the case for regular grammars (Sect. 3.1.3) and there is again a common dynamic-programming-like solution—see Lari and Young (1990, 1991), Geman and Johnson (2000).

3.3 Gibbs Distributions

There are many ways to generalize. The coverage of a context-free grammar may be inadequate, and we may hope, therefore, to find a workable scheme for placing probabilities on context-sensitive grammars, or perhaps even more general grammars. Or, it may be preferable to maintain the structure of a context-free grammar, especially because of its dynamic programming principle, and instead generalize the class of probability distributions away from those induced (parameterized) by production probabilities. But

nothing comes for free. Most efforts to generalize run into nearly intractable computational problems when it comes time to parse or to estimate parameters.

Many computational linguists have experimented with using Gibbs distributions, popular in statistical physics, to go beyond production-based probabilities, while nevertheless preserving the basic context-free structure. Next we take a brief look at this particular formulation, in order to illustrate the various challenges that accompany efforts to generalize the more standard probabilistic grammars.

3.3.1 Probabilities. The sample space is Ψ_G , the set of trees generated by a context-free grammar G . Gibbs measures are built from sums of more-or-less simple functions, known as ‘potentials’ in statistical physics, defined on the sample space. In linguistics, it is more natural to call these features rather than potentials. Suppose, then, that we have identified M linguistically salient features $f_1, \dots, f_M$, where $f_k: \Psi_G \rightarrow R$, through which we will characterize the fitness or appropriateness of a structure $\psi \in \Psi_G$. More specifically, we will construct a class of probabilities on Ψ_G which depend on $\psi \in \Psi_G$ only through $f_1(\psi), \dots, f_M(\psi)$. Examples of features are the number of times a particular production occurs, the number of words in the yield, various measures of subject-verb agreement, and the number of embedded or independent clauses.

Gibbs distributions have the form

$$P_\theta(\psi) = \frac{1}{Z} \exp \left\{ \sum_{i=1}^M \theta_i f_i(\psi) \right\} \quad (9)$$

where $\theta_1, \dots, \theta_M$ are parameters, to be adjusted ‘by hand’ or inferred from data, $\theta = (\theta_1, \dots, \theta_M)$, and where $Z = Z(\theta)$ (known as the ‘partition function’) normalizes so that $P_\theta(\Psi) = 1$. Evidently, we need to assume or ensure that $\sum_{\psi \in \Psi_G} \exp \{ \sum_{i=1}^M \theta_i f_i(\psi) \} < \infty$. For instance, we had better require that $\theta_1 < 0$ if $M = 1$ and $f_1(\psi) = |\mathcal{Y}(\psi)|$ (the number of words in a sentence), unless of course $|\Psi_G| < \infty$.

3.3.2 Relation to probabilistic context-free grammars. Gibbs distributions are much more general than probabilistic context-free grammars. In order to recover PCFG’s, consider the special feature set $\{f(A \rightarrow \alpha; \psi)\}_{A \rightarrow \alpha \in R}$. The Gibbs distribution Eqn. (9) takes on the form

$$P_\theta(\psi) = \frac{1}{Z} \exp \left\{ \sum_{A \rightarrow \alpha \in R} \theta_{A \rightarrow \alpha} f(A \rightarrow \alpha; \psi) \right\} \quad (10)$$

Evidently, then, we get the probabilistic context-free grammars by taking $\theta_{A \rightarrow \alpha} = \log_e p(A \rightarrow \alpha)$, where p is a system of production probabilities consistent with Eqn. (7), in which case $Z = 1$. But is Eqn. (10) more general? Are there probabilities on Ψ_G of this form

that are not PCFGs? The answer turns out to be no, as was shown by Chi (1999) and Abney et al. (1999): given a probability distribution P on Ψ_G of the form of Eqn. (10), there always exists a system of production probabilities p under which P is a PCFG.

3.3.3 Inference. The feature set $\{f_i\}_{i=1, \dots, M}$ can be accommodate arbitrary linguistic attributes and constraints, and the Gibbs model Eqn. (9), therefore, has great promise as an accurate measure of linguistic fitness. But the model depends critically on the parameters $\{\theta_i\}_{i=1, \dots, M}$, and the associated estimation problem is, unfortunately, very hard. Indeed, the problem of unsupervised learning appears to be all but intractable.

Let us suppose, then, that we observe $\psi_1, \psi_2, \dots, \psi_n \in \Psi_G$ (‘supervised learning’), iid according to $\hat{P}_\theta$. In general, the likelihood function, $\prod_{i=1}^n P_\theta(\psi_i)$, is more or less impossible to maximize. But if the primary goal is to select good parses, then perhaps the likelihood function asks for too much, or even the wrong thing. It might be more relevant to maximize the likelihood of the observed parses, *given the yields* $\mathcal{Y}(\psi_1), \dots, \mathcal{Y}(\psi_n)$:

$$\prod_{i=1}^n P_\theta(\psi_i | \mathcal{Y}(\psi_i)) \quad (11)$$

Maximization of Eqn. (11) is an instance of Besag’s remarkably effective pseudolikelihood method (Besag 1974, 1975), which is commonly used for estimating parameters of Gibbs’s distributions. The computations involved are generally much easier than what is involved in maximizing the ordinary likelihood function. Take a look at the gradient of the logarithm of Eqn. (11): the θ_j component is proportional to

$$\frac{1}{n} \sum_{i=1}^n f_j(\psi_i) - \frac{1}{n} \sum_{i=1}^n E_\theta [f_j(\psi) | \mathcal{Y}(\psi) = \mathcal{Y}(\psi_i)] \quad (12)$$

and $E_\theta [f_j(\psi) | \mathcal{Y}(\psi)]$ can be computed directly from the set of parses of the sentence $\mathcal{Y}(\psi)$. (In practice there is often massive ambiguity, and the number of parses may be too large to feasibly consider. Such cases require some form of pruning or approximation.)

Thus gradient ascent of the pseudolikelihood function is (at least approximately) computationally feasible. This is particularly useful since the Hessian of the logarithm of the pseudolikelihood function is non-positive, and therefore there are no local maxima. What is more, under mild conditions pseudolikelihood estimators (i.e., maximizers of Eqn. (11)) are consistent; (Chi 1998).

Bibliography

Abney S, McAllester D, Pereira F 1999 Relating probabilistic grammars and automata. In: *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*, Morgan Kaufmann, San Francisco, pp. 542–609

- Baum L E 1972 An inequality and associated maximization techniques in statistical estimation of probabilistic functions of Markov processes. *Inequalities* 3: 1–8
- Besag J 1974 Spatial interaction and the statistical analysis of lattice systems (with discussion). *Journal of the Royal Statistical Society, Series D* 36: 192–236
- Besag J 1975 Statistical analysis of non-lattice data. *The Statistician* 24: 179–95
- Chi Z 1998 *Probability models for complex systems*. PhD thesis, Brown University
- Chi Z 1999 Statistical properties of probabilistic context-free grammars. *Computational Linguistics* 25(1): 131–60
- Chomsky N 1957 *Syntactic Structures*. Mouton, The Hague
- Dempster A, Laird N, Rubin D 1977 Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B* 39: 1–38
- Fu K S 1974 *Syntactic Methods in Pattern Recognition*. Academic Press, New York
- Fu K S 1982 *Syntactic Pattern Recognition and Applications*. Prentice-Hall, Englewood Cliffs, NJ
- Gazdar C, Klein E, Pullum G, Sag I 1985 *Generalized Phrase Structure Grammar*. Blackwell, Oxford, UK
- Geman S, Johnson M 2000 Probability and statistics in computational linguistics, a brief review. Internal publication, Division of Applied Mathematics, Brown University
- Harris T E 1963 *The Theory of Branching Processes*. Springer, Berlin
- Hopcroft J E, Ullman J D 1979 *Introduction to Automata Theory, Languages and Computation*. Addison-Wesley, Reading, MA
- Jelinek F 1997 *Statistical Methods for Speech Recognition*. MIT Press, Cambridge, MA
- Kaplan R M, Bresnan J 1982 Lexical functional grammar: A formal system for grammatical representation. In: Bresnan J (ed.) *The Mental Representation of Grammatical Relations*. MIT Press, Cambridge, MA, Chap. 4, pp. 173–281
- Kay M, Gawron J M, Norvig P 1994 *VerbMobil: A Translation System for Face-to-face Dialog*. CSLI Press, Stanford, CA
- Lari K, Young S J 1990 The estimation of stochastic context-free grammars using the inside-outside algorithm. *Computer Speech and Language* 4: 35–56
- Lari K, Young S J 1991 Applications of stochastic context-free grammars using the inside-outside algorithm. *Computer Speech and Language* 5: 237–57
- Pollard C, Sag A 1987 *Information-based Syntax and Semantics*. CSLI Lecture Notes Series No.13. University of California Press, Stanford, CA
- Shieber S M 1986 *An Introduction to Unification-based Approaches to Grammar*. CSLI Lecture Notes Series. University of California Press, Stanford, CA

S. Geman and M. Johnson

Probabilistic Thinking

1. Introduction

Some people have enjoyed playing games that involved at least a component of chance ever since civilizations—as we know them—emerged in human history.

Many of these people apparently had a fairly good intuitive idea of what we might now term the ‘probabilities’ or ‘odds’ of various outcomes associated with these games. Alternatively, at least some people (whom in hindsight we might now label ‘superstitious’) believed that the outcomes of these games could indicate something about the favorability of the world in general toward the gambler (a belief that may exist even today—at least on an implicit basis—for chronic gamblers), specifically the favorability or unfavorability of a particular god that was involved with a particular game.

In the last 600 years or so people have proposed systematic ways of evaluating probabilities—e.g., in terms of equally likely outcomes in games of chance or in terms of subjective belief. In the last 150 years or so scientific endeavor has started to make specific use of probabilistic thinking, and in the last 100 years or so probabilistic ideas have been applied to problems of everyday life. (see Gigerenzer et al. 1989). Currently, for example, jurors are often asked to determine manufacturer or company liability on the basis of differential rates of accidents or cancers; in such judgments, jurors cannot rely on deterministic reasoning that would lead to a specification of exactly which negative outcome is due to what; nor can they rely on ‘experience’ in these contexts in which they have had none. Thus, it is extraordinarily important to understand—if indeed generalizations about such thinking are possible—how it is that people think about and evaluate probabilities, and in particular the problems that they (we) have. The major thesis presented here is that both the thinking and the problems can be understood in terms of *systematic deviations* from the normative principles of probabilistic reasoning that have been devised over the past five or so centuries. These deviations are not inevitable; it is certainly possible for people to think about probabilities coherently (at least on occasion); nevertheless, *when* deviations occur, they tend to be of the systematic nature to be described here; they occur most often when people are engaged in what is termed ‘automatic’ or ‘associational’ thinking, as opposed to ‘reflective’ and ‘comparative’ thinking. (These terms will be made more precise later in this article.)

2. A Brief History (of Thought)

Beginning in early infancy, we develop intuitions about space, time, and number (at least to the extent of ‘one, two, three, many’). These intuitions are essentially correct. They remain throughout our lives basically unaltered except for small variations, and they form the foundation of far more sophisticated mathematical systems that have been developed over the past thousands of years. Moreover, their importance in our lives has been clear throughout those millennia.

In contrast, probabilistic thinking is often unintuitive, and it generally can be taught only after we are familiar with fractions, so that analogs to relative frequency are clear. Moreover, our intuitions about probabilities are often just plain wrong. For example, belief in the ‘gambler’s fallacy’ (negative recency) can even be explicitly rejected, but nevertheless implicitly influence decision making—such as when people demand a greater amount of money in preference to gambling on an outcome that is ‘due’ (e.g., a tail after five straight heads) than in preference to gambling on one that isn’t (e.g., an additional head); see Gold (1997).

The fairly recent, and somewhat ‘iffy,’ use of probabilistic judgment does not imply, however, that people did not gamble throughout the centuries. As mentioned previously, the outcomes of games of chance could provide not just monetary winners and losers but ‘signals’ from ‘the gods’ about favorability toward the gambler; this belief is well captured by the narrative of the novelist Robert Graves (1943) about Caligula’s gambling shortly before his assassination. This tyrant is both watching gladiator games, often putting to death people he has bribed to lose, and rolling a form of dice called ‘astragali’—which are four-sided dice made from pig or lamb bones.

Caligula is playing with his Uncle Claudius, who in order to avoid winning (a dangerous outcome for Claudius) hands the emperor beautiful new dice loaded to come up with a ‘Venus roll,’ which consists of four separate sides (generally with the numbers 1, 3, 4, and 6). Caligula’s subsequent success convinces him that Venus is favorable toward him that day, and he does not take the usual precautions upon leaving the games. Throughout his life, Claudius feels guilty that the loaded dice sent the wrong message to Caligula about how Venus felt about him that day—an inappropriate intervention between a god and a man. (According to Hacking 1975, the interpretation of such events as ‘signals’ from ‘on high’ is common up to the Renaissance, even after people in Europe had given up polytheism for monotheism; the reader might also note how often in the *Old Testament* people decide to ‘cast lots’ to decide what to do—apparently in the belief that God would determine the outcome.)

Of course, not everyone believed in such intervention. For example, Cicero is quoted as claiming that ‘Nothing is so unpredictable as a throw of the dice [modern translation],’ and that ‘both now and then indeed he will make it [the Venus roll] twice and even thrice in succession. Are we going to be so feeble-minded then as to aver that such a thing happened by the personal intervention of Venus rather than by pure luck?’ (see David 1962, p. 25).

Among those not so superstitious (such as Cicero, who was himself assassinated), the intuitive understanding gained through experience of which bets are good ones and which bets are bad ones led to fairly good implicit estimates of probability. Even now,

however, Wagenaar (1988) finds chronic gamblers to be more reliant on intuition and experience than on mathematical analysis.

In contrast, explicit probabilistic thinking goes beyond experience. We can, for example, indicate what are good or bad bets in gambles we have just now constructed, with which we have had no previous experience. Probabilistic thinking can also be extended to ‘personal belief’ about uncertain events, it can be used to evaluate risks, and it can serve as a guide for social policy.

Systematic, probabilistic thinking appears to have started with questions about gambling (in particular about gambles that have not yet been played multiple times). Despite the fact that people gambled a lot, however, the formal theory did not begin until the Italian Renaissance, specifically in correspondence and argumentation between Geronimo Cardano (1501–76) and his associates. The approach that has been ascribed to Cardano involved *comparative* rather than associational thinking. He proposed that the likelihood of success of a gamble could be *computed*, or at least approximated, by comparing the number of outcomes of the device favorable to the gambler divided by the total number of outcomes. In situations where the gambling device is set up to make each possible outcome equally likely, this ratio assessment of probability is used today. The ratio assessment leads to certain general principles, often termed ‘axioms,’ which in turn provide a more abstract notion of probability, and allow different interpretations of probability to be developed—just provided they lead to measures consistent with the axioms. For example, Ramsey (1931) demonstrated that when explicit probabilities are used to determine fair gambling odds, then the gambler will follow the basic principles of probability theory if and only if it is not possible to make a ‘Dutch Book’ (see de Finetti 1974–5, p. 87) such that the gambler always loses; that is, the gambler will be consistent if and only if it is impossible for a highly intelligent opponent to make a series of bets such that the opponent always wins.

After Cardano, the idea of probabilistic inference developed quite rapidly. Many French and Dutch intellectuals corresponded with each other about what constituted better or worse bets—not just in games that had been played chronically, but in games of their own invention. Interestingly, when they discussed probability, they often did so in terms of the expected value of the game, and questions were often posed as to which of the two games had a higher expected value. Then there was rapid development of probability theory—for example, Bernoulli’s Law of Large Numbers,’ (see *Statistical Methods, History of: Pre-1900* and *Statistical Methods, History of: Post-1900*) which specified the relationship between the probability of a single outcome and the probability distribution of multiple outcomes of the same type (for example, the relationship between the probability a fair coin comes

up heads and the probability that it comes up heads at least three times in 10 trials); the 'law' itself derived the limit of the proportion of heads as the number of trials becomes indefinitely large. To jump ahead to recent times, von Neumann and Morgenstern published their classic work *Theory of Games and Economic Behavior* (1944/1947), in which again the question was one of choosing between different gambles, but in terms of expected utility rather than expected value. This work was quite consistent with the earliest work, in which the question of probability was secondary to the question of choice between alternative courses of action.

3. Probability Assessments

All probabilistic analysis is based on the idea that (suitably trained and intelligent) people can at least recognize good probabilistic arguments presented by someone else, or discovered or thought of by themselves, but not necessarily generate good assessments. The very fact that there was correspondence about the gambles—and occasionally some disputes about them—indicated that people do not automatically assess probabilities in the same way, or accurately (e.g., corresponding to relative frequencies, or making good gambling choices).

The von Neumann and Morgenstern work, however, does involve some *psychological* assumptions that people can engage in 'good' probabilistic thinking. First, they must do so implicitly, so that the choice follows certain 'choice axioms' that allow the construction of an expected utility model—i.e., a model that represents choice as a maximization of implicit expected utility; that in turn requires that probabilities at the very least follow the standard axioms of probability theory. It also implies considering conditional probabilities in a rational manner, which is done only when implicit or explicit conditional probabilities are consistent with Bayes' theorem. Thus, the von Neumann and Morgenstern work required that people be 'Bayesian' in a consistency sense, although that term is sometimes used to imply that probabilities should at base be interpreted as degrees of belief. Another way in which probability assessment must be 'good' is that there should be some at least reasonable approximation between probabilities and long-term relative frequencies; in fact, under particular circumstances (of interchangeability and indefinitely repeated observations), the probabilities of someone whose belief is constrained by Bayes' theorem must approximate relative frequencies (Dawid 1982).

Are people always coherent? Consider an analogy with swimming. People do swim well, but occasionally we drown. What happens is that there is particular systematic bias in attempting to swim that makes it difficult. We want to hold our heads above water.

When, however, we raise our heads to do so, we tend to assume a vertical position in the water, which is one of the few ways of drowning (aside from freezing or exhaustion or being swept away in rough waters).

Just as people drown occasionally by trying to hold their heads above water, people systematically deviate from the rules of probabilistic thinking. Again, however, the emphasis is on 'systematic.' For example, there is now evidence that people's probabilistic judgments are 'subadditive'—in that when a general class is broken into components, the judgmentally estimated probabilities assigned to disjoint components that comprise the class sum to a larger number than the probability assigned to the class. That is particularly true in memory, where, for example, people may recall the frequency with which they were angry at a close friend or relative in the last month and the frequency with which they were angry at a total stranger, and the sum of the estimates is greater than an independent estimate of being angry period (even though it is possible to be angry at someone who is neither a close friend or relative nor a total stranger; see Mulford and Dawes 1999). The clever opponent will then bet against the occurrence of each component but on the occurrence of the basic event, thereby creating a Dutch Book.

4. The Heuristics and Biases Approach to Problems of Probabilistic Thinking

The basic conception is that we all use a series of 'heuristics' in our everyday estimation (explicit or implicit) of probabilities, and these heuristics lead to biases. These heuristics provide a type of 'bounded rationality' (Simon 1955, 1956, 1979)—in that they substitute for a very detailed or logical analysis, an analysis that may be beyond our 'computational capacity.' Just as associational thinking serves us well in many contexts, however, so do the heuristics. Occasionally, however, they lead to grossly erroneous—systematically erroneous—probabilistic thinking. The conception that heuristics have some type of 'generalized validity,' but that they often lead us astray, was well argued in a classic *Science* article by Tversky and Kahneman (1974); unfortunately, the principle has been lost on some critics, who can point out many situations in which these heuristics 'work.' Naturally, they do. If they didn't have some sort of generalized utility, they would not have survived over generations of people trying to make probabilistic judgments.

One of the most important of these heuristics is termed the 'representativeness' heuristic, which was first proposed by Tversky and Kahneman in an article concerning *Belief in the Law of Small Numbers* (1971). The idea supported in that article is that we expect very small samples to 'represent' the populations from

which they were drawn, where the ‘we’ includes even people with a fair amount of formal training in statistics. That expectation leads to judgment of probability on the basis of the degree to which an instance or a sample *matches* a population or general idea, rather than on a judgment of how it can be generated. For example, when tossing a coin, a result consisting of six heads appears to be much less likely than one of a head followed by two tails, followed by two heads and finally a tail. But of course each of the sequences is equally likely; each occurs with probability $1/64$. In fact, the sequence of a head followed by two tails followed by two heads and a tail is seen as more likely than a sequence of strict alternation of heads and tails—even though in general people believe that there is more than a chance amount of alternation when they judge what is truly ‘random’ (and in fact are most apt to pick as random the sequence that has a transition probability of $2/3$ rather than $1/2$). This incorrect expectation of alternation leads to such erroneous beliefs as that in the ‘hot hand’ in basketball when observing clusters of similar outcomes (e.g., successful shots). See, for example, Gilovich et al. (1985).

Another heuristic is termed an availability one. Briefly, we estimate frequencies on the ease with which instances ‘come to mind,’ or on our judgment of the ease with which we can recall or generate instances. It is, in fact, generally true that when there are more instances as opposed to fewer in the world, we find it easier to recall or generate more as opposed to fewer. The problem is, however, that there are determinants of ease of recall or generation that are independent of—or sometimes even in conflict with—actual relative frequency in our environment. For example, vivid instances are more readily recalled than pallid ones. (See again the result on ‘subadditivity’ in memory.) Moreover, instances that are easily imaginable (even if we haven’t actually experienced them) are more readily ‘brought to mind’ than are those difficult to imagine.

The availability heuristic creates a bias when we are unaware of the factors other than frequency that can result in ease of recall or generation, but use it as a way of estimating actual frequency. Sometimes, the bias is so strong that we even ignore instances that we *cannot* sample—a ‘structural availability’ problem. For example, clinical psychologists often claim that no child sexual abusers stop on their own without therapy. When asked to substantiate this claim, they will cite many of the child abusers with whom they have had contact, often mandated by a judge to enter therapy with them. The problem is, however, that they would not have had contact with any who had stopped on their own, and by definition they were seeing people only in a therapy setting, hence could not observe any of them stopping ‘without therapy.’ Nevertheless, the experience with such people is so salient and compelling that even while acknowledging the logical problem with the conclusion, many of these clinical

psychologists simply revert to the belief a day or a week after having publicly recognized the availability bias (see Dawes 1988, p. 102).

Following the ideas of Fischhoff and Beyth-Marom (1983) as elaborated by Dawes (1998), the standard cognitive biases and heuristics can be understood by alluding to simple forms of Bayes’ theorem. Consider, for example, the relationship between a symptom S and a disease D , and suppose a diagnostician observes this symptom S . If the probability of the disease is assessed on the basis of a pure matching or association between the symptom and the disease, independent of considerations of conditional probability, there is no normative structure to which the judgment corresponds. More often, however, the judgment will be made on the basis of the conditional probabilities—a normatively correct judgment if it is based on the conditional probability of the disease given the symptom, $P(D/S)$, but a representative judgment if it is the probability of the symptom given the disease, $P(S/D)$. Unfortunately, there is a lot of evidence that the judgment is often made on the basis of the latter relationship—when conditional probabilities are considered at all.

The relationship between these two probabilities is given by rewriting Bayes’ theorem as:

$$P(D/S) = \frac{P(S/D)P(D)}{P(S)} \quad (1)$$

It can also be written in terms of the ‘ratio rule’:

$$\frac{P(D/S)}{P(S/D)} = \frac{P(D)}{P(S)} \quad (2)$$

Because the ratio rule follows from the very definition of conditional probability, it is simply incoherent to equate the probability of the disease given the symptom with the probability of the symptom given the disease in the absence of a simultaneous belief that the probability of the disease and the probability of the symptom are identical. (Again, if someone were to make a series of bets based on ‘fair betting odds’ in such a belief, an opponent could make a Dutch Book against that person.) Classic representative thinking—e.g., ‘she gave a typical schizophrenic response, therefore she must be schizophrenic’—embraces such an identity. What is critical, of course, are the *base rates* of D and of S , yet extensive research has shown that people underutilize such base rates in their intuitive judgments—or at least fail to incorporate them to a sufficient extent into their prior beliefs, given that they are often seen as ‘not relevant.’ (Why, people often ask, should the *general* probabilities of these symptoms and diseases be relevant to a particular judgment about a particular person with a particular symptom? Such an argument is sometimes even followed by

bland assertions to the effect that ‘statistics do not apply to the individual’—a wholly erroneous conclusion, as can be attested to by the heavy cigarette smoker who develops lung cancer.)

People often do not ignore these base rates *entirely* in the judgments, particularly if the probabilistic nature of the judgment is made clear (see Gigerenzer et al. 1989). Nevertheless, the ‘underutilization’ of such base rates is ubiquitous. In the psychological literature it was first decried by Meehl and Rosen (1955), who noted that in case conferences they attended, clinical psychology and psychiatry judges tended to ignore base rates completely; later, underutilization has been established in a variety of experimental situations using a wide variety of subjects, especially by Kahneman and Tversky (1972, 1973). For example, when asked to judge whether a description is of a particular engineer or a particular lawyer, subjects make the same judgment whether the experimenter states that the description was drawn from a pool of 70 percent engineers and 30 percent lawyers, or vice versa. As Bar-Hillel (1990) points out, the 70/30 split does not appear to be ‘relevant,’ but it’s relevance would most likely be clear if the split were 99/1, and certainly if it were a 100/0. (Again, we can hypothesize that people can recognize probabilistic arguments—although it’s possible to be fooled on occasion—but often not make them spontaneously.) Thus, for example, someone who is said to be interested in sailing and carpentry and sharing model train sets with a child is judged to be much more likely to be an engineer than a lawyer, whether this person was drawn from a pool of 70 percent engineers or 70 percent lawyers. Making the sampling procedure absolutely ‘transparent’ does, however, reduce that tendency.

The use of Bayes’ theorem to understand the relationship between symptoms and disease leads naturally to its use to consider the general relationship between any hypotheses and any bit of evidence. Let e refer to some bit of evidence about whether a hypothesis h is or is not true; for example, a symptom may be regarded as a bit of evidence, and having a particular disease a hypothesis; or the hypothesis may concern who is going to win the election in the United States in the year 2020 and a particular bit of evidence may be the margin of victory or defeat of the potential candidate in a more local election. Bayes’ theorem expressed in terms of evidence and hypotheses is presented as

$$P(h/e) = \frac{P(e/h)P(h)}{P(e)} \quad (3)$$

This form of Bayes’ theorem, while true, often leads to complications when trying to evaluate the probability of the evidence. There is, however, a way to avoid having to estimate that probability. What we can do is to consider *odds* that the hypothesis is true. These odds

are the probability the hypothesis is true given the evidence divided by the probability that this hypothesis is false given the evidence—that is, by the ratio $P(h/e)/P(-h/e)$. Once we know the odds, it is trivial to compute the probability. The advantage of considering odds is that the denominator in Bayes’ theorem cancels out when we compute these odds. That is

$$\frac{P(h/e)}{P(-h/e)} = \frac{P(e/h)P(h)}{P(e/-h)P(-h)} \quad (4)$$

Now we are in a position to categorize the most common forms of representative thinking. Pseudodiagnosticity refers to making an inference about the validity of hypothesis h on the base of evidence e without considering hypothesis $-h$. Another way of stating pseudodiagnosticity is that it involves considering only the numerator of the first term on the right-hand side in Eqn. (4). People do that when they state that a bit of evidence is ‘consistent with’ or ‘typical of’ some hypothesis without concerning themselves about alternative hypotheses, or in particular the negation of the hypothesis. For example, in attempting to diagnose whether a child has been sexually abused, alleged experts in court often refer to symptoms ‘typical’ of such abuse—without concerning themselves with how typical these symptoms are of children who have *not* been abused, or the frequency with which children have been sexually abused.

Another type of representative thinking involves considering the probability of the evidence given the hypothesis or hypotheses without looking at the second terms in Eqn. (4), that is, the prior odds. The reason that these odds are important is that they yield the *extent* of the evidence and hypotheses considered. For example, being a poor speller may be evidence of dyslexia in the sense that it has a very high probability given dyslexia. However, in order to diagnose dyslexia on the basis of poor spelling, we have to know something about the base rates of poor spelling and the base rates of dyslexia in the population from which the person whom we wish to diagnose was chosen.

Dilution effects occur when evidence that does not distinguish between hypotheses in fact influences people to be less convinced of a most likely hypothesis; the problem is similar to that found in pseudodiagnosticity in that people do not realize that evidence that is not very likely given the hypothesis may be equally unlikely given the negation of that hypothesis, but may in fact believe a hypothesis less as a result of collecting evidence that is unlikely if it is true. Dilution is simply the converse of pseudodiagnosticity.

Finally, it is possible to describe *availability* biases as well by Eqn. (4). People believe that they are sampling evidence given the hypothesis, when in fact they are sampling this evidence given the hypothesis *combined with* the manner in which they are sampling. For example, when private practice clinical psycho-

logists claim that they are sampling characteristics of people who fall into a certain diagnostic category on the basis of their experience, what they are really sampling is people who fall in that category *and who come to them*. Similarly, when we sample our beliefs about 'what drug addicts are like,' most of us are sampling on the basis of how the media presents drug addicts—both in 'news' programs and in dramatizations (Dawes 1994b). Doctors often sample on the basis of their contact with addicts when these addicts are ill, perhaps gravely so, while police are often sampling on the basis of their experience with these same addicts during arrests and other types of confrontation. It is not surprising, then, that doctors are in favor of a 'medical' approach to drug addiction including such policies as sterile needle exchanges, while police are in favor of much more punitive policies. Both are sampling evidence *combined with* their exposure to it.

All these anomalies have been demonstrated experimentally. It is important to point out, however, that the investigation of these anomalies did not spring simply from understanding Bayes' theorem and then from creating very clever experimental situations in which subjects' inferences systematically violate it. The motive to study these anomalies of judgment arises from having observed them on a more informal basis outside the experimental setting—and trying to construct an experimental setting, which yields a greater measure of control, that will allow them to be investigated in a systematic manner. The reader is referred, for example, to the essay of Meehl (1977) or the book of Dawes (1994a) for descriptions of these biases in the clinical judgment of professional psychologists and psychiatrists who rely on their own 'experience' rather than (dry, impersonal) 'scientific' principles for diagnosis and treatment.

5. Analyzing an Important Example

This section ends with a discussion of pseudodiagnosticity, a bias that is prevalent in 'expert' proclamations encouraging 'child sex abuse hysteria' (Dawes 2001). The main problem here is that hypotheses are not compared: instead, single hypotheses are evaluated in terms of the degree to which evidence is 'consistent with' them; in addition, evidence is often *sought* in terms of its consistency with or inconsistency with 'favorite hypotheses'—rather than in terms of its ability to distinguish *between* hypotheses. This type of pseudodiagnosticity has made its way into the legal system, where experts are allowed to testify that a child's recanting of a report of sexual abuse is 'consistent with' the 'child sexual abuse accommodation syndrome' (Summit 1983). The finding is that many children who are *known* to have been sexually abused deny the abuse later (recant), therefore, the

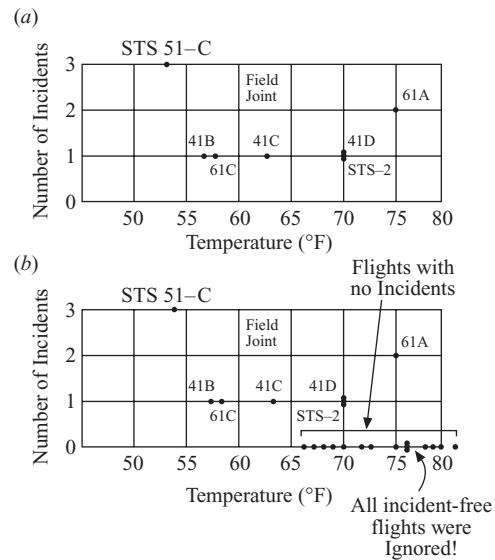


Figure 1

Partial and total analyses of the relationship between (a) data reviewed prior to launch and (b) the full story from past flights (source: Russo and Schoemaker 1989, pp. 197–8)

probability of the evidence (the child recants) given the hypothesis (actual abuse) is not as low as might be naively assumed; therefore, recanting is 'consistent with' having been abused (i.e., can be considered part of a 'syndrome' of accommodation to abuse).

The problem with this pseudodiagnostic reasoning is that it compares the probability of the evidence given the hypothesis to the probability of the *negation of the evidence* given the hypothesis, whereas a rational comparison is of the probability of the evidence given the hypothesis to the probability of the evidence given the *negation of the hypothesis*. When considering this latter comparison, we understand immediately that the recanting would be diagnostic of actual abuse only if the probability of recanting given actual abuse were *higher* than the probability of recanting given such abuse had not occurred—a highly implausible (not to mention paradoxical) conclusion.

Such pseudodiagnostic reasoning is involved in other disasters as well, for example, the explosion of the US Space Shuttle Challenger on January 22, 1986. When considering the possibility that cold temperatures might cause the O-rings to malfunction, 'the worried engineers were asked to graph the temperatures at launch time for the flights in which problems have occurred' (Russo and Schoemaker 1989, p. 197). Thus, these engineers attempted to evaluate the hypothesis that there would be O-ring problems by asking what the temperature was given the existence of these problems. As indicated in Fig. 1(a) (Fig. 6 in

Russo and Schoemaker 1989, p. 198), there appeared to be no relationship. When, however, the temperature of launches in which there were *no* problems were considered, a very clear relationship between problems and temperature is evident (see Fig. 1(b), from Russo and Schoemaker 1989, Fig. 7, p. 198.)

Moreover, people actively seek evidence compatible or incompatible with a hypothesis rather than evidence that distinguishes between hypotheses, as has been extensively studied by Doherty and his colleagues (see for example Doherty et al. 1979). Consider, for example, subjects who are asked to play the role of medical students (and actual medical students prior to specific training, see Wolf et al. 1985) who must determine whether patients have one of two conditions. The conditions are explained, and subjects are told that a patient has two symptoms. The judges are then presented with the conditional probability of one of these symptoms given one of the diseases (e.g., that it is very likely that the patient has a high fever given the patient has meningitis). They can then choose to find out the probability of the other symptom given the same disease, the probability of the first symptom given the second disease, or the probability of the second symptom given the second disease. While virtually no subjects choose a different symptom and a different disease, the majority choose to find the probability of the *second* symptom given the first disease (the 'one that is focal' as a result of being told the first symptom). But for all these subjects know, of course, the first symptom may be more or less typical of the *alternative disease*, as may the second symptom. Finding out that the probability of these two symptoms given only one disease in no way helps to distinguish between that disease and some other. *Of course, in real medical settings doctors may have prior knowledge of these other disease/symptom relationships*, but because the diseases are not even identified in the current setting, such knowledge would be of no help.

Occasionally, pseudodiagnosticity has been confused with what has been termed a 'confirmation bias'—that is, a desire to search for evidence that is very probable given a focal hypothesis. As Fischhoff and Beyth-Marom (1983), Trope and Bassok (1982), and Skov and Sherman (1986) point out, however, such a search does not necessarily imply pseudodiagnosticity (or any other particular irrationality) because it may be possible that the evidence is *not* found—that is, that instead of finding the anticipated evidence, the diagnostician finds its negation. In fact, Skov and Sherman (1986) demonstrate that in searching for evidence—at least in their experimental context—subjects often became quite rational *when they are informed of the conditional probabilities*. The irrationality occurs in the search for which probabilities to seek, and in 'jumping' to a conclusion on the basis of only one probability (i.e., on the basis of the first term in the numerator in the odds form of Bayes' theorem).

The cognitive explanation of pseudodiagnosticity and the other heuristic-based biases covered in the section is simple. People often think by association; in fact, whole computer programs have been built to simulate human judgment on the basis of 'matching' characteristics to schemas. The point of pseudodiagnosticity is that only one 'schema' (i.e., hypothesis) is considered, whereas *coherent diagnostic judgment is always comparative*. In fact, all the biases covered here are based on associational thinking, whereas clear normative probabilistic thinking always involves comparisons.

6. Summary and Implications

When we deal with probabilistic situations in everyday life, we can often 'muddle through,' but occasionally not appreciating the comparative nature of valid probabilistic thinking can lead to judgmental disasters. Similarly, experience with games of chance can lead to some pretty good intuitive estimations of the likelihood of various payoffs, but once again there can be systematic biases, such as belief in the gambler's fallacy. Explicit probabilistic thinking has developed in our cultures only in the last 500 years or so, and in the last 150 years or so it has become increasingly important in both scientific reasoning and in our day-to-day living—and even in legal contexts (where as jurors we may be asked to judge the liability of a particular company by comparing the probability of developing cancer among those exposed to it or its product as opposed to the 'base rate' probability, or may be asked to evaluate what an expert witness states about the relationship between a symptom and a diagnosis). Here, if we rely on our intuitive grasp of probabilities, we often find that we have systematic biases, much as novice swimmers may attempt to keep their heads above water and thereby drown. These biases—which have been briefly described and catalogued in this article—involve explicitly defined deviations from the conclusions of valid probabilistic reasoning, which follow from the simplest forms of Bayes' theorem. They are not random deviations, which could be modeled as analogous to an engineering context involving a 'signal' plus 'noise.'

Of course, we are also capable of precise probabilistic thinking—just as swimmers are capable of assuming a horizontal position in the water and consequently not drowning. The point is, however, that when we deviate from the rules ('axioms') of probability theory or from coherent judgments, these deviations follow the patterns described here—as has been documented in everyday observation and studied experimentally by those advocating the 'heuristics and biases' approach to such judgment. For a greater elaboration of this approach than is found here, see the chapters in Kahneman et al.'s book *Judgment Under Uncertainty: Heuristics and Biases* (1982). These

systematic departures themselves in turn may be connected to thinking in terms of associations, whereas good probabilistic judgment always requires comparative thinking.

See also: Information Theory; Information, Economics of; Probability and Chance: Philosophical Aspects; Probability: Interpretations; Risk, Sociological Study of

Bibliography

- Bar-Hillel M 1990 Back to base-rates. In: Hogarth R M (ed.) *Insights in Decision Making: A Tribute to Hillel J. Einhorn*. University of Chicago Press, Chicago
- Dawes R M 1988 *Rational Choice in an Uncertain World*. Harcourt Brace Jovanovich, San Diego, CA
- Dawes R M 1994a *House of Cards: Psychology and Psychotherapy Built on Myth*. Free Press, New York
- Dawes R M 1994b AIDS, sterile needles, and ethnocentrism. In: Heath L, Tinsdale R S, Edwards J, Posavac E, Bryant F B, Henderson-King E, Suarez-Balcazar Y, Myers J (eds.) *Social Psychological Applications to Social Issues. III: Applications of Heuristics and Biases to Social Issues*. Plenum, New York, pp. 31–44
- Dawes R M 1998 Behavioral decision making and judgment. In: Gilbert D, Fiske S, Lindzey G (eds.) *The Handbook of Social Psychology*, 11. McGraw-Hill, Boston, MA, pp. 497–548
- Dawes R M 2001 *Everyday Irrationality: How Pseudoscientists, Lunatics and the Rest of Us Systematically Fail to Think Rationally*. Westview Press, Boulder, CO
- David F N 1962 *Games, Gods, and Gambling*. Hafner Pub. Co., New York, Chap. 4
- Dawid A P 1982 The well-calibrated Bayesian. *Journal of the American Statistical Association* 77(379): 605–10
- de Finetti B 1974–5 *Theory of Probability: A Critical Introductory Treatment*. John Wiley and Sons, London, Vol. 1
- Doherty M E, Mynatt C R, Tweney R D, Schiavo M D 1979 Pseudodiagnosticity. *Acta Psychologica* 43: 111–21
- Fischhoff B, Beyth-Marom R 1983 Hypothesis evaluation from a Bayesian perspective. *Psychological Review* 90: 239–60
- Gigerenzer G, Swijtink Z, Porter T, Daston L, Beatty J, Kruger L 1989 *The Empire of Chance: How Probability Changed Science and Everyday Life*. Cambridge University Press, Cambridge, UK
- Gilovich T, Vallone R, Tversky A 1985 The hot hand in basketball: On the misperception of random sequences. *Journal of Personality and Social Psychology* 17: 295–314
- Gold E 1997 The Gambler's Fallacy, Doctoral dissertation submitted in partial fulfillment of the requirement for the degree of Doctor of Philosophy, Carnegie Mellon University
- Hacking I 1975 *The Emergence of Probability: A Philosophical Study of Early Ideas About Probability, Induction and Statistical Inference*. Cambridge University Press, New York
- Kahneman D, Tversky A 1972 Subjective probability: A judgment of representativeness. *Cognitive Psychology* 3: 430–54
- Kahneman D, Tversky A 1973 On the psychology of prediction. *Psychological Review* 80: 237–51

- Kahneman D, Slovic P, Tversky A (eds.) 1982 *Judgment Under Uncertainty: Heuristics and Biases*. Cambridge University Press, Cambridge, UK
- Meehl P E 1977 Why I do not attend case conferences. In: Meehl P E (ed.) *Psychodiagnosis: Selected Papers*. W. W. Norton, New York, pp. 225–302
- Meehl P E, Rosen A 1955 Antecedent probability in the efficiency of psychometric signs, patterns, or cutting scores. *Psychological Bulletin* 52: 194–216
- Mulford M, Dawes R M 1999 Subadditivity in memory for personal events. *Psychological Science* 10(1): 47–51
- Ramsey F 1931 Truth and probability. In: Ramsey F P (ed.) *The Foundations of Mathematics and other Logical Essays*. Harcourt, Brace and Co., New York, pp. 156–98
- Russo J E, Schoemaker P J H 1989 *Decision Traps: Ten Barriers to Brilliant Decision Making and how to Overcome Them*. Simon and Schuster, New York
- Simon H 1955 A behavior model of rational choice. *Quarterly Journal of Economics* 69: 99–118
- Simon H 1956 Rational choice and the structure of the environment. *Psychological Review* 63: 129–38
- Simon H 1979 *Models of Thought*. Yale University Press, New Haven, CT, Vol. 1
- Skov R B, Sherman S J 1986 Information-gathering processes: Diagnosticity, hypothesis-confirmatory strategies, and perceived hypothesis confirmation. *Journal of Experimental Social Psychology* 22: 93–121
- Summit R C 1983 The child sexual abuse accommodation syndrome. *Child Abuse and Neglect* 7: 177–93
- Trope Y, Bassok M 1982 Confirmatory and diagnosing strategies in social information gathering. *Journal of Personality and Social Psychology* 43(1): 22–34
- Tversky A, Kahneman D 1971 Belief in the law of small numbers. *Psychological Bulletin* 76: 105–110
- Tversky A, Kahneman D 1974 Judgment under uncertainty: Heuristics and biases. *Science* 185: 1124–31
- von Neumann J, Morgenstern O 1944/1947 *The Theory of Games and Economic Behavior*, 2nd edn. Princeton University Press, Princeton, NJ
- Wagenaar W 1988 *Paradoxes of Gambling Behavior*. Lawrence Erlbaum Associates, Hillsdale, NJ
- Wolf F M, Gruppen L D, Billi J E 1985 Differential diagnosis and the competing hypothesis heuristic—a practical approach to judgment under uncertainty and Bayesian probability. *Journal of the American Medical Association* 235: 2858–62

R. M. Dawes

Probability and Chance: Philosophical Aspects

The central philosophical issue involving probability concerns the interpretation of ordinary probability claims, such as ‘there is a 30% probability that it will rain today.’ *Epistemic* interpretations of probability understand this to be a claim about belief under uncertainty: it describes the extent to which a certain type of rational agent would (or ought to) believe that

it will rain, given the available evidence. *Objective* interpretations of probability read the claim as describing some quantifiable feature of the objective world, sometimes called ‘chance.’ The concept of chance is thought by some to be intimately connected with *indeterminism*. A variety of epistemic and objective interpretations of probability will be discussed below.

1. The Problem of the Interpretation of Probability

The problem of the interpretation of probability can be posed using a familiar example: what does a meteorologist *mean* when he or she says that there is a 30 percent probability of rain on a given day? To give an account of what a proposition means is to say what must be the case in order for the proposition to be true. If the meteorologist says that it *will* rain, this assertion is true if and only if it does rain. Whether it rains or not, however, neither outcome by itself renders the *probabilistic* claim true. So what must be the case in order for this claim to be true?

An adequate solution to this problem must provide some way of making sense of the numerical values that appear in probability claims (and must do so in such a way that those values are in accord with the basic principles of probability—see *Probability: Formal*). The claim that 30 percent of the balls in a given urn are blue has a well-understood meaning: the number of blue balls in the urn, divided by the total number of balls in the urn, is equal to 0.3. In the case of the weather report, however, no comparable formula is immediately available.

Proposed interpretations of probability tend to fall within one of two broad camps, *objective* (or *empirical*) and *epistemic* (or *subjective*) interpretations. Objective interpretations interpret probability claims as describing some feature of the world. Epistemic interpretations of probability, by contrast, interpret probability claims as claims about the extent to which the claim is or ought to be believed. Note that this standard terminology is somewhat misleading: some epistemic interpretations (such as the logical interpretation) take probabilities to be objective in the sense that they do not vary from person to person. The various interpretations of probability canvassed below are often presented as *rivals*, but it may be more appropriate to think of them as complementary. Most researchers who have grappled with interpretive issues have been of the opinion that the concept of probability has both an objective and an epistemic dimension. Accordingly, philosophers often talk of ‘objective probabilities’ and ‘epistemic probabilities.’ Objective probabilities are sometimes called ‘chances.’

Philosophers (as well as many philosophically minded statisticians) have become interested in the

problem of interpretation of probability in part because of the ubiquity of probability claims in science as well as ordinary life, and in part because of the utility of probability theory in analyzing a variety of concepts of interest to philosophers, such as causation (see *Causation: Physical, Mental, and Social*), confirmation of hypotheses by evidence (see *Atomism and Holism: Philosophical Aspects*), and rationality (see *Rational Choice Explanation: Philosophical Aspects*).

This article will focus on the philosophical, rather than the mathematical, dimensions of the problem of interpreting probability. Readers interested in this topic would also do well to read the article on *Probability: Interpretations*. Other surveys of this topic include Salmon (1967), Sklar (1993), and Howson (1995).

2. Determinism and Chance

In a famous thought experiment, the French physicist Pierre Simon de Laplace (1819) imagined a demon who could observe the exact position and momentum of every particle in the universe at a given time. Such a demon could plug this information into Newton’s equations of motion and predict the state of the universe at all future times. In this thought experiment, Laplace was articulating the concept of *determinism*. A system is *deterministic* if its state at a given time, together with the laws of evolution for that system (see *Causes and Laws: Philosophical Aspects*), uniquely determine the state of the system at later times. For Laplace, determinism meant that all *objective* probabilities would have to have values of zero or one. Laplace thus thought that probabilistic claims, such as the meteorologist’s prognostication described above, reflect human ignorance about which outcome is determined to occur.

At present, there is good reason to believe that certain physical systems described by *quantum mechanics* behave indeterministically. For example, if a vertically polarized photon strikes a polarizer that is oriented with an angle of θ , quantum theory entails that the photon has a probability of $\cos^2\theta$ of passing through. According to the orthodox interpretation of quantum theory (which is not controversial), this probability is *irreducible*: there is no further fact about the photon that determines what it will do on this occasion. If this is correct, then the probability value $\cos^2\theta$ does not merely reflect human ignorance about what the photon will do: it is an objective feature of the world. (For a thorough discussion of philosophical issues involving determinism, see Earman (1986).) One important philosophical issue is whether only such irreducible probabilities are genuine chances. If so, then it would be safe to infer that *none* of the probability claims made within the social sciences describe genuinely objective probabilities.

The word ‘chance’ is also sometimes used as a singular noun to refer to a chance that is not equal to zero or one. For example, the thesis of determinism is sometimes expressed by saying that there is no genuine chance in the world; ‘the USA soccer team has a chance of winning the next world cup’ says that the chance of their winning is greater than zero and less than one; to ‘take a chance’ is to embark on a venture that has a chance of success that is greater than zero but less than one; and so on.

3. Epistemic Probability

Laplace claimed that probability claims do not describe objective features of the world, but the state of human knowledge (and ignorance). Epistemic interpretations of probability all share this general feature.

3.1 The Classical Interpretation

If asked the probability of rolling an odd number on a six-sided die, most people would unhesitatingly respond with the answer one-half: there are six possible outcomes of the die roll, three of which are odd. This illustrates the *classical interpretation* of probability, often associated with Laplace (1819). According to this interpretation, the possible outcomes of a particular trial are divided into ‘equally possible’ cases, and the probability of outcome *F* on a particular trial is equal to the number of such cases in which *F* is true, divided by the total number of possible cases. Which cases are equally possible is determined by the *Principle of Indifference*, which states that two cases are equally possible in the absence of any reason to prefer one outcome over the other.

There are a number of difficulties with the principle of indifference. When the weather reporter claims that the probability of rain is 0.3, there seem to be at least two different types of outcome (rain and no rain) but it is hard to see what the 10 equally possible outcomes (three of which involve rain) could be. Even in much simpler cases, the principle of indifference is not precise enough to specify the equally possible cases. An urn contains both red balls and blue balls; some of the red balls are light red, and some are dark red. What is the probability that the next ball drawn is blue: are there two equally possible cases (red and blue) or three (light red, dark red, and blue)? This problem is further exacerbated when there are infinitely many possible outcomes.

3.2 Subjective Interpretations

According to the subjective interpretation of probability (whose proponents are variously called subjectivists’ personalists, or Bayesians), probability

measures the degrees of belief or *credences* of a certain type of idealized rational agent. Subjectivists regard probability theory as analogous to *formal logic*. Logic alone never tells an agent whether to believe a proposition, but it does require that an agent’s beliefs be consistent with one another. For example, the rules of logic forbid an agent to believe ‘it will rain today,’ ‘if it rains today, it will not snow today,’ and ‘it will snow today’ at the same time. According to subjectivists, probability theory is an extension of logic that allows an agent to have degrees of confidence in a proposition that range in value between zero and one. Probability theory alone will not tell the agent how much confidence to place in the proposition ‘it will rain today,’ but rather imposes a consistency-like constraint (usually called *coherence*) on the totality of the agent’s degrees of belief.

The numerical value of an agent’s degree of belief in some proposition is connected with the agent’s evaluation of bets on the truth of that proposition. For example, if an agent has a degree of belief of 0.3 for the proposition ‘it will rain today,’ that means that he or she would find a bet on rain at 3:7 odds—a bet that costs the agent three units, and buys a chance to win 10 units (for a profit of seven units) if it rains—to be fair. This connection between degrees of belief, utility, and evaluation forms the foundation of a theory of rational choice wherein an agent always acts so as to maximize expected utility (see *Rational Choice Explanation: Philosophical Aspects; Elicitation of Probabilities and Probability Distributions; Game Theory and its Relation to Bayesian Theory*).

This connection between degrees of belief and evaluation is often deployed in arguments that rational degrees of belief conform to the formal rules of probability (see *Probability: Formal*). When an agent’s degrees of belief conform to the probability calculus, then his or her overall system of evaluations will have a number of desirable features. For example, Ramsey (1926) showed that having one’s degrees of belief conform to the rules of probability is a necessary condition for having one’s overall system of *preferences* satisfy a set of plausible constraints. One such constraint is transitivity: if the agent prefers A to B and B to C, then the agent prefers A to C. A second class of results comprises the so-called *Dutch Book* theorems. (See, for example, de Finetti (1964) and Kemeny (1955).) These results show that if and only if an agent’s degrees of belief violate the rules of probability, the agent will be vulnerable to a ‘Dutch Book.’ A Dutch Book is a system of bets such that the agent finds each individual bet fair, but where the agent is guaranteed of suffering a net loss no matter what the outcome. For example, a bet on heads at 2:1 odds together with a bet of equal stakes on tails at 2:1 odds would constitute a Dutch Book, for the agent purchasing the two bets would be guaranteed of losing 2 units on one bet, and winning 1 unit on the other. (For more details, see *Probability: Interpretations*.)

The requirement that rational degrees of belief satisfy the rules of probability theory is a *synchronic* constraint on one's degrees of belief: it concerns an agent's degrees of belief at one particular time. Some subjectivists impose a further *diachronic* constraint on allowable methods for changing one's degrees of belief over time. Suppose that a rational agent's degrees of belief at some particular time are represented by the probability function P . If the agent then acquires some new piece of information I , then the agent's new degree of belief in any proposition A is equal to the *conditional probability* $P(A/I)$ (see *Probability: Formal* for the definition of conditional probability). This rule of updating degrees of belief by *conditionalizing* can also be motivated by a form of Dutch Book argument (proved by David Lewis and reported in Teller (1973)).

This rule for changing degrees of belief has been championed by so-called *Bayesians* as a theory of confirmation. According to this theory, a hypothesis H is confirmed or supported by a piece of evidence E just in case conditionalizing on E increases the probability of H . This probabilistic approach to confirmation theory has proven to be very fruitful. (See *Truth, Verification, Verisimilitude, and Evidence: Philosophical Aspects*; see also Howson and Urbach (1993).)

The various theorems invoked by subjectivists are *formal* results, and there is considerable debate over their interpretation. In particular, there is disagreement about how, or even whether, the Dutch Book theorems really show that coherence or updating by conditionalization are canons of rationality. Critics of subjectivism point out that even an incoherent agent would be capable of following the proofs of these theorems, and hence of foreseeing the dangers of Dutch Books; thus coherence and updating by conditionalization are not really necessary for avoiding financial loss (see Howson and Urbach (1993) for further discussion).

A second focus of criticism is the *subjective* nature of degrees of belief. Rational agents may differ in the degree to which they believe a proposition such as 'it will rain today'; indeed their degrees of belief may range all the way from zero to one. Moreover, the constraints of synchronic coherence plus updating by conditionalization do not rule out the possibility that one's degrees of belief are *dogmatic* in a certain respect. It is consistent with these constraints that an agent have degree of belief of 0.99 that each toss of a given coin will land heads, and after observing a thousand tails in a row, *still* believe that the *next* toss will almost certainly be heads. Subjectivists have been able to prove many interesting results about how an agent's degrees of belief *would* evolve in the face of new evidence *if* the agent's original degrees of belief (or *prior probabilities*) were nondogmatic in various ways (de Finetti's (1964) representation theorem is one important example), but there is nothing in sub-

jectivism that would *require* the agent to have such nondogmatic prior probabilities. For many, this latitude makes subjective degrees of belief ill-suited as a foundation for theories of rational choice or rational inference.

3.3 Logical Interpretations

Logical interpretations aim to avoid the problems of subjectivist interpretations by imposing further constraints on an agent's prior probabilities. Prior probabilities are determined by the logical structure of the space of outcomes, much as they are in the classical interpretation. Unlike the classical interpretation, however, probability assignments can reflect the bearing of evidence that discriminates among the possible outcomes. Thomas Bayes (1763) and Laplace (1819), who are commonly associated with the subjectivist and classical approaches (respectively), both endorsed a version of the logical approach to probability. Twentieth century proponents include Carnap (1950), Jeffreys (1961), and Jaynes (1973).

Logical interpretations of probability inherit many of the difficulties of the classical interpretation discussed above: logical considerations alone do not determine a unique prior probability distribution. Moreover, many of the most plausible candidates for logical probabilities—such as one in which a sequence of coin tosses is *independent and identically distributed* (see *Probability: Formal or Distributions, Statistical: Special and Discrete*)—turn out to be dogmatic in just the manner described in Sect. 3.2.

3.4 The Need for Objective Interpretations

Although few would doubt that our ordinary concept of probability has an epistemic dimension, it seems implausible that *all* probability claims can be interpreted epistemically. Consider the claim that a photon has a certain probability of passing through a polarizer; it would be very odd indeed if the truth of this claim (and hence of quantum mechanics) depended on the extent to which some agent(s) may believe that the photon will pass through or on the extent to which available evidence confirms this hypothesis. At best, epistemic interpretations of probability will have to be supplemented by an objective interpretation of at least some probability statements.

4. Objective Interpretations of Probability

In practice, probabilities are often inferred from frequency data. To estimate the probability that a given coin will land heads when tossed, for example, one could toss it many times and observe the relative frequency with which the coin lands heads. Such

relative frequencies are *objective* features of the world. Hence relative frequencies provide a natural resource for objective interpretations of probability claims.

4.1 Finite Relative Frequency Interpretation

Perhaps the simplest interpretation of a probability claim would be to understand it as a claim about an actual finite relative frequency. For example, if a coin were tossed exactly 10 times and then destroyed, resulting in 'heads' on six of those 10 tosses, then the probability that the coin lands heads when tossed would be 0.6. The central difficulty with this definition of probability is that it rules out unrepresentative samples by definition. For example, if a coin is fair, i.e., has a probability of 0.5 of landing heads when tossed, that should not be taken to rule out the possibility of it landing heads on six of the 10 occasions on which it is tossed. According to the finite frequency definition of probability, however, if the probability of heads is 0.5, then *by definition* the coin will land heads on exactly half of its tosses. Thus, while finite relative frequencies provide indispensable evidence for probabilistic claims, probabilities cannot be identified with finite relative frequencies.

4.2 Limiting Relative Frequency Interpretation

In order to avoid these problems with finite frequencies, some philosophers have suggested that probabilities be identified with frequencies in an *infinite* sample. If the first n tosses of a coin yields $f(n)$ heads, then the relative frequency of heads in the first n tosses is $f(n)/n$. The *limiting relative frequency* of heads in the infinite sequence of tosses is the limit as n goes to infinity of $f(n)/n$. This approach to interpreting probability was pioneered by Venn (1866); leading twentieth century proponents of this interpretation (hereafter called *frequentists*) include Richard von Mises (1957) and Hans Reichenbach (1949).

One important problem with limiting relative frequencies is that they seldom, if ever, exist: no coin is actually tossed an infinite number of times. Thus the limiting relative frequencies that determine the probability of an outcome are really *hypothetical limiting relative frequencies*: to say that the probability of heads on a toss of a certain coin is 0.5 is to say that *if* the coin were to be tossed an infinite number of times, the limiting relative frequency of heads *would be* 0.5. This is a hypothetical or *counterfactual* claim; by itself, it says nothing about what features of the *actual world* make the probabilistic claim true. The problem of interpretation of counterfactuals is itself a thorny philosophical problem (see *Counterfactual Reasoning, Quantitative: Philosophical Aspects; Counterfactual Reasoning, Qualitative: Philosophical Aspects*).

One important issue that divides frequentists is whether it is possible to ascribe a probability value to the outcome of a single trial. One tradition, following von Mises (1957), views probability as a property of a sequence of trials, or a 'collective.' Any single coin toss belongs to infinitely many different sequences: the sequence of all coin tosses, the sequence of all tosses of that particular coin, the sequence of all tosses of that coin by imparting a particular force from a particular height above the floor, and so on. The limiting relative frequency of 'heads' may be different in each of these sequences. Thus, there is no unique probability of 'heads' on a given toss, but only a probability for 'heads' in a collective of tosses. While frequentists in this tradition deny that there is one privileged sequence of coin tosses, not just any sequence of tosses is fit to bear probability values. For example, if an infinite sequence of coin tosses were to yield an alternating sequence of 'heads' and 'tails,' the relative frequency of 'heads' would be 0.5. The regularity of the pattern of outcomes would nonetheless render this sequence inappropriate for characterizing a probability value. Rather, probabilities are limiting relative frequencies in a *random* sequence of trials. There is no universally accepted mathematical characterization of randomness, although this is currently an area of active research (see Earman 1986, Chap. VIII). For a more detailed exposition of this version of the limiting relative frequency interpretation, see *Probability: Interpretations*.

One feature of this approach is that it allows chances to take on values other than zero or one even if determinism is true. It may be that on any particular toss of a coin, the linear and angular momentum imparted to the coin together with its height above the floor and so on determine the outcome of the coin toss. Nonetheless, there may be a random sequence of coin tosses where the limiting relative frequency is not equal to zero or one, for the tosses comprising the collective need not be identical with respect to such determining factors.

The relegation of probabilities to properties of collectives would seem to be a drawback of this approach to probability. When deciding whether to bet on a coin toss, or to postpone a picnic, one wants to know the probability of heads *on that particular toss*, or the probability of rain *on this particular day*. Some frequentists, following Reichenbach, believe that it is possible to assign probability values to the outcomes of individual trials. This requires that a choice be made from among all of the sequences to which the trial in question belongs. For example, a particular day belongs in an infinite number of sequences of days, and the limiting relative frequency of rain need not be the same in all of them, so which sequence determines the probability of rain on the day in question? Presumably not just the sequence of all days, for then the probability of rain would not vary from day to day. A natural thought is that what is

wanted is a sequence of days *just like this one*. No two days will be exactly alike in *all* respects, however; the best that can be hoped for is a sequence of days that are alike in all *relevant* respects. The key problem, then, is to say what makes some respects of similarity between days relevant and others irrelevant. One detailed proposal is developed in Salmon (1984, Chap. 3).

On this approach, determinism is compatible only with chances of zero or one. Suppose, for example, that on a given day, climactic conditions determine that it will rain. Then the appropriate sequence will presumably consist of days in which the same climactic conditions prevail, and hence the limiting relative frequency of rain within that sequence will be one.

4.3 Propensity Interpretation

The propensity interpretation of probability, first introduced by Karl Popper (1959), is motivated by some of the difficulties with the limiting relative frequency interpretation. According to this interpretation, a coin, in conjunction with the conditions in which it is tossed, has a certain *disposition* or *propensity* to land heads. The probability that the coin will land heads is a measure of the strength of this propensity. This propensity is a property of the particular experimental set-up, hence, it is appropriate to assign a probability to an individual trial. Moreover, this propensity provides grounds for the truth of claims about what the coin *would do* if tossed infinitely many times in identical (or relevantly similar) conditions.

While this is certainly a useful way of thinking about single case probabilities within a frequentist framework, critics wonder whether anything has really been added. A coin is about to be tossed: there is some property of this experimental set-up in virtue of which the coin *would* land heads with a certain limiting relative frequency if this experiment were repeated infinitely often; calling this property a 'propensity' of a certain strength does little to indicate just what this property is. Moreover, if propensities are *P*not simply identified with hypothetical limiting relative frequencies, it is hard to see what is supposed to ground the claim that propensities have the mathematical structure of probabilities.

4.4 Best System Interpretation

The philosopher David Lewis (1986) has proposed an interpretation according to which an objective probability claim is true when it is entailed by a probabilistic law (see *Causes and Laws: Philosophical Aspects*). On this view, objective probability is intimately connected with indeterminism. A probabilistic claim is a law if it belongs to the set of propositions that provides the best overall systematization of the world. In order to be the best systematization, a set of propositions must

obtain the optimal balance between informativeness, simplicity, and fit with data. For example, consider the claim that a photon will pass through a polarizer with a probability of $\cos^2\theta$. Although the frequency with which photons do pass through when the experiment is performed may diverge from this number, the probability value is firmly grounded in the principles of quantum mechanics, which overall provides an excellent fit with experimental data, and is a powerful and elegant theory.

Unlike the frequency interpretations, however, the truth conditions for probabilistic claims provided by the best system interpretation are not very precise. It is not clear what makes one set of propositions simpler than another, nor how simplicity is to be balanced against informativeness and goodness of fit.

4.5 Theoretical Role Interpretations

One further line of interpretation is influenced by the philosophical school of pragmatism (see *Pragmatism: Philosophical Aspects*). In this approach, objective probability claims are to be understood in terms of their role in our overall conceptual system. In particular, to understand a probability claim is to know what kinds of evidence would support or undermine that claim, and how to act accordingly if one accepts that claim as true. This approach has been developed by Mellor (1971), Skyrms (1980), and Lewis (1986).

Consider an agent whose degrees of belief conform to the rules of probability (see Sect. 3.2), and let *Ch* be a function that assigns numerical values to propositions. The agent takes *Ch* to be objective chance precisely if his or her degrees of belief are guided by beliefs about the values of *Ch*. For example, suppose that the agent believes *Ch* ('it will rain today') to be 0.3. If the agent understands *Ch* to be objective chance, then his or her degree of belief in the proposition 'it will rain today' will also be 0.3. This degree of belief, together with the agent's utility function, can then be used to decide whether or not to postpone a picnic. In this way, then, the agent's beliefs about objective chances influence his or her decisions. Moreover, since objective probability claims are now propositions about which the agent has degrees of belief, the apparatus of Bayesian confirmation theory shows how those degrees of belief will change in the face of new evidence.

This approach has a number of virtues, not least of which is that it explains why objective chances have the formal properties of probability theory: in order for beliefs about chances to guide degrees of belief in the appropriate way, objective chances must have the same formal structure as rational degrees of belief. There is, however, an important drawback to this approach. In analyzing the meaning of objective probability claims in terms of their theoretical role instead of their truth conditions, this account does not

answer what many have taken to be the central question of objective probability: what must be the case in the world in order for the (objective) probability of rain to be 0.3?

5. Conclusion

Interest in the philosophical aspects of probability dates back to the origins of the formal theory of probability in the seventeenth century. Epistemic and frequency interpretations of probability have been around for well over a century. In the second half of the twentieth century, interest in objective interpretations of single-case probabilities has dramatically increased the range of interpretive options available. Given the current lack of consensus, and the increasingly widespread use of probabilistic concepts in both philosophy and everyday life, it is safe to predict that interest in the problem of the interpretation of probability will continue into the twenty-first century.

See also: Bayesian Statistics; Bayesian Theory: History of Applications; Decision Theory: Bayesian; Probabilistic Thinking; Probability: Formal; Probability: Interpretations; Problem Solving and Reasoning, Psychology of; Subjective Probability Judgments; Utility and Subjective Probability: Contemporary Theories

Bibliography

- Bayes T 1763 An essay towards solving a problem in the doctrine of chances. *Philosophical Transactions of the Royal Society* **53**: 370–418
- Carnap R 1950 *Logical Foundations of Probability*. University of Chicago Press, Chicago
- de Finetti B 1964 Foresight: its logical laws, its subjective sources. In: Kyburg H E, Smokler H E (eds.) *Studies in Subjective Probability*. Wiley, New York
- Earman J 1986 *A Primer on Determinism*. Reidel, Dordrecht, Chaps. II, VIII
- Howson C 1995 Theories of probability. *British Journal for the Philosophy of Science* **46**: 1–32
- Howson C, Urbach P 1993 *Scientific Reasoning: The Bayesian Approach*, 2nd edn. Open Court, Chicago
- Jaynes E T 1973 The well-posed problem. *Foundations of Physics* **3**: 477–93
- Jeffreys H 1961 *Theory of Probability*, 3rd edn. Clarendon Press, Oxford, UK
- Kemeny J G 1955 Fair bets and inductive probabilities. *Journal of Symbolic Logic* **20**: 263–73
- Laplace P S de 1819 *Essai Philosophique sur les Probabilités*. Mme. Ve Courcier, Paris
- Lewis D 1986 *Philosophical papers*. Oxford University Press, Oxford, UK, vol. II, see especially the introduction and Chap. 19
- Mellor D H 1971 *The Matter of Chance*. Cambridge University Press, Cambridge, UK
- Popper K 1959 The Propensity interpretation of probability. *British Journal for the Philosophy of Science* **10**: 25–42
- Ramsey F 1926 Truth and probability. In: Mellor D H (ed.) *Philosophical Papers*. Cambridge University Press, Cambridge, UK
- Reichenbach H 1949 *The Theory of Probability*. University of California Press, Berkeley, CA
- Salmon W C 1967 *The Foundations of Scientific Inference*. University of Pittsburgh Press, Pittsburgh, PA, see especially Chap. V
- Salmon W C 1984 *Scientific Explanation and the Causal structure of the world*. Princeton University Press, Princeton, NJ, see especially Chap. 3
- Sklar L 1993 *Physics and Chance*. Cambridge University Press, Cambridge, UK, see especially Chap. 3
- Skyrms B 1980 *Causal Necessity*. Yale University Press, New Haven, CT
- Teller P 1973 Conditionalization and observation. *Synthese* **26**: 218–58
- Venn J 1866 *The Logic of Chance*. Macmillan, London
- von Mises R 1957 *Probability, Statistics, and Truth*, 2nd rev. English edn. Allen and Unwin, London

C. Hitchcock

Probability Density Estimation

Associated with each random variable, X , measured during an experiment is the cumulative distribution function (cdf), $F(x)$. The value of the cdf at $X = x$ is the probability that $X \leq x$. Features in a set of random data may be computed analytically if the cumulative distribution function is known. The probability that a random sample falls in an interval (a, b) may be computed as $F(b) - F(a)$. This formula assumes the values of X are continuous. Discrete random variables taking on only a few values are easy to handle; thus, this article focuses on continuous random variables with at least an interval scale.

The same information may be obtained from the probability density function (pdf), $f(x)$, which equals $\Pr(X = x)$ if X takes on discrete values, or the derivative of $F(x)$ if X is continuous. If either $F(x)$ or $f(x)$ is known, then any quantity of interest such as the variance may be computed analytically or numerically. Otherwise, such quantities must be estimated using a random sample $x_1, x_2, \dots, x_n$. The relative merits of estimating $F(x)$ vs. $f(x)$ are an important part of this article.

Density estimation takes two distinct forms—parametric and nonparametric—depending on prior knowledge of the parametric form of the density. If the parametric form $f(x|\vec{\theta})$ is known up to the p parameters $\vec{\theta} = (\theta_1, \dots, \theta_p)$, then the parameters may be estimated efficiently by maximum likelihood or Bayesian algorithms. When little is known of the parametric form,

nonparametric techniques are appropriate. The most commonly used estimators are the histogram and kernel methods, as discussed below.

1. Histograms and CDFs

Grouping and counting the relative frequency of data to form a histogram may be taught at the beginning of every statistics course, but the concept seems to have emerged only in 1662 when John Graunt studied the Bills of Mortality and constructed a crude life table of the age of death during those times of plague (Westergaard 1968). Today the histogram and its close relative, the stem-and-leaf plot, provide the most important tools for exploring and representing structure in a set of data. Studying the shape of a well-constructed histogram can verify an assumption of normality, or reveal departures from normality such as skewness, kurtosis, or multiple modes. The recent study of the theoretical properties of histograms has led to concrete methods for computing well-constructed figures as discussed below.

With only a random sample of data, the cdf may be estimated by the empirical cdf (ecdf), denoted by $F_n(x)$. The value of the ecdf at x is the fraction of the sample with values less than or equal to x , that is, $\{\#x_i \leq x\}$. Now the probability any one of the x_i is less than x is given by the true cdf at x . Hence, on average, the value of $F_n(x)$ is exactly $F(x)$, and $F_n(x)$ is an unbiased estimator of $F(x)$. The inequality $\{X_i \leq x\}$ is a binomial random variable, so that the variance of

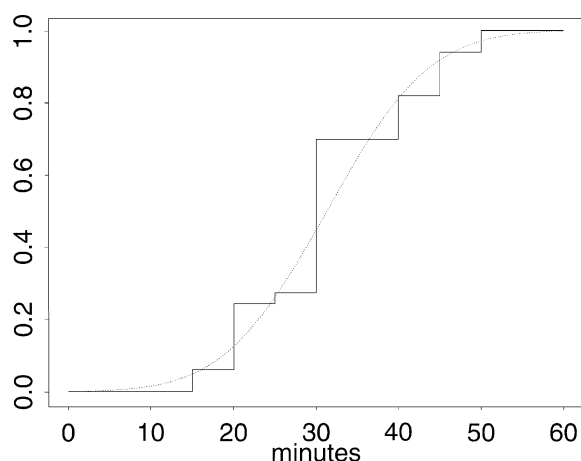


Figure 1

Empirical cumulative distribution function of physicians' estimated treatment times for a patient suffering migraine headaches. The dotted line is the cdf of the normal fit

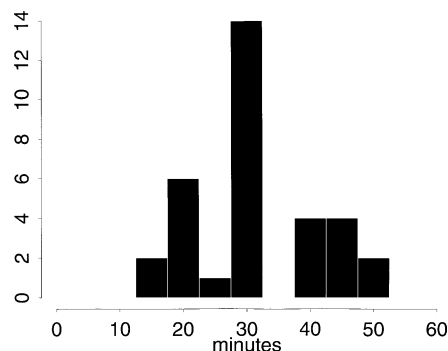


Figure 2

Frequency histogram of treatment times

$F_n(x)$ may be computed as $F(x)(1 - F(x))/n$. In fact, there is no other unbiased estimator with smaller variance.

The graph of the empirical cdf looks like a staircase function. Figure 1 displays the ecdf of the treatment times estimated by 33 physicians when presented with a hypothetical patient suffering from extreme migraine headaches (Hebl and Xu 2000). The dotted line is the cdf of the fitted normal density $N(31.3, 9.9^2)$, which appears to fit reasonably well. Figure 2 displays a histogram of the same data, which does not appear to be normal after all. Social scientists often are taught to favor the ecdf over the histogram, but in fact aside from computing probabilities, the ecdf is not well-suited for the easy understanding of data.

The probability density function, $f(x)$, carries the same information as the cdf. Since the derivative of the empirical cdf $F_n(x)$ is not well-defined, estimators that rely on grouping have been devised. A convenient interval, (a, b) , containing the data is chosen, and subdivided into m 'bins' of equal width $h = (b - a)/m$. The value h is called the bin width (and is often called the smoothing parameter). A 'well-constructed' histogram must specify appropriate values for a , b , h , and m . Given such parameters, the number of samples falling into each of the m bins is tallied. Label the m bin counts $v_1, v_2, \dots, v_m$ from left to right. If the interval (a, b) contains all the samples, then $\sum_{k=1}^m v_k = n$. A *frequency histogram* is simply a plot of the bin counts, as in Fig. 2. Although the values are all multiples of 5 min, the data may be treated as continuous. However, the bin width should only be a multiple of 5 and care should be exercised to choose a and h so that no data points fall where two bins adjoin. Here, $a = 2.5$, $b = 57.5$, and $h = 5$. The histogram suggests a departure from normality in the form of three modes. Re-examine the ecdf in Fig. 1 and try to identify the same structure.

A true probability density is a non-negative function that integrates (sums) to 1. Hence, the *density his-*

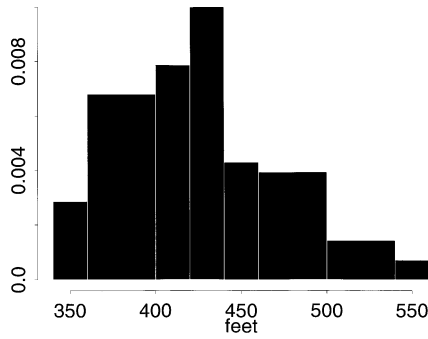


Figure 3
Density histogram of the distances of Mark McGwire's 70 home runs in the 1998 season

togram is a simple normalization of the frequency histogram in the k^{th} bin,

$$\hat{f}(x) = \frac{v_k}{nh} \quad (1)$$

The normalization is important when the data are heavily skewed and unequal bin widths are constructed. In such cases, the interval (a, b) is partitioned into m bins of width $h_1, h_2, \dots, h_m$. The bin counts $v_1, v_2, \dots, v_m$ should *not* be plotted as they are not relatively proportional to the true density histogram, which is now given by $\hat{f}(x) = v_k/(nh_k)$ in the k th bin. Figure 3 displays a density histogram of the distances of the 70 home runs hit by Mark McGwire during his record-setting 1998 American baseball season. A bin width of 20ft was chosen, and then several adjacent bins were collapsed to make the histogram look unimodal. The long right tail reflects the booming home runs for which McGwire is known.

2. Bin Width Construction

Rules for determining the number of bins were first considered by Sturges in 1926. Such rules are often designed for normal data, which occur often in practice. Sturges observed that the m th row of Pascal's triangle, which contains the combinatorial coefficients $\binom{m-1}{0}, \binom{m-1}{1}, \dots, \binom{m-1}{m-1}$, when plotted, appears as an ideal frequency histogram of normal data. (This is a consequence of the central limit theorem for large m .) The sum of these 'bin counts' is 2^{m-1} by the binomial expansion of $(1+1)^{m-1}$. Solving the equation $n = 2^{m-1}$ for the number of bins, we obtain Sturges' rule

$$m = 1 + \log_2 n \quad (2)$$

This rule is widely applied in computer software.

The modern study of histograms began in Scott (1979) and Freedman and Diaconis (1981). The optimization problem formulated was to find the bin width h to make the average value of the 'distance' $\int [\hat{f}(x) - f(x)]^2 dx$ as small as possible. The solution for the best bin width is

$$h^* = \left[\frac{6}{\int f'(x)^2 dx} \right]^{1/3} n^{-1/3} \quad (3)$$

While the integral of the derivative of the unknown density is required to use this formula, two useful applications are available. First, the *normal reference rule* gives

$$h^* = 3.5\sigma n^{-1/3} \quad (4)$$

since the value of the integral for a normal density is $(4\sqrt{\pi}\sigma^3)^{-1}$. The second application was given by Terrell and Scott (1985) who searched for the 'easiest' smooth density (which excludes the uniform density) and found that the number of bins in a density should always exceed $\sqrt[3]{2n}$ or that the bin width should always be less than

$$h^* \leq 3.73\sigma n^{-1/3} \quad (5)$$

These choices are called *oversmoothed rules*, as the optimal choices will never be more extreme for other less smooth densities. Notice that the normal reference rule is very close to the upper bound given by the oversmoothed rule. For the McGwire data, the oversmoothed rule gives $h^* \leq 42'$.

More sophisticated algorithms for estimating a good bin width for a particular data set are described in Scott (1992) and Wand (1997). For example, a cross-validation algorithm (Rudemo 1982) that attempts to minimize the error distance directly leads to the criterion

$$CV(h) = \frac{2}{(n-1)h} - \frac{n+1}{(n-1)n^2h} \sum_{k=1}^m v_k^2 \quad (6)$$

The bin counts are computed for a selection of bin widths (and bin origins), the criterion computed, and the minimizer chosen, subject to the over-smoothed bound. The criterion tends to be quite noisy and Wand (1997) describes alternative plug-in formulae that can be more stable.

All of these rules give bin widths which shrink at the rate $n^{-1/3}$ and give many more bins than Sturges' rule. In particular, most computer programs give histograms which are *oversmoothed*. The default values

should be overridden, especially for large datasets. Similarly, the error of the best histogram decreases to zero at the rate $n^{-2/3}$, which can be improved to $n^{-4/5}$ as described below. However, the histogram remains a powerful and intuitive choice for density estimation and presentation.

3. Advanced Algorithms

Density estimators with smaller errors are continuous functions themselves. The easiest such estimator is the frequency polygon, which connects the midpoints of a histogram. A separate error theory exists. For example, the oversmoothed bin width for a histogram from which the frequency polygon is drawn is given by

$$h^* \leq 2.33\sigma n^{-1/5}$$

Note that this bin width is wider than the rules for the histogram alone.

Many other estimators exist, based on splines, wavelets, Fourier series, or local likelihood, but all are closely related to the kernel estimator of Rosenblatt (1956) and Parzen (1962), which is described here. The kernel can be any probability density function, but the normal density is a common and recommended choice. Here the smoothing parameter h is the standard deviation of the normal kernel. If the normal density $N(\mu, \sigma^2)$ at a point x is denoted by $\phi(x|\mu, \sigma)$, then the kernel estimator is

$$\hat{f}_K(x) = \frac{1}{n} \sum_{i=1}^n \phi(x|x_i, h)$$

This form shows that the kernel estimator is a mixture of normal densities which are located at the data points. The oversmoothed smoothing parameter value is

$$h^* \leq 1.14\sigma n^{-1/5}$$

An example is given below.

4. Visualization

The primary uses of density estimation are to display the structure in a set of data and to compare two sets of data. The histogram is adequate for the first task, but overlaying two histograms for comparison is difficult to interpret. The continuity and improved error properties of the kernel estimators excel for this purpose.

In the great home run race of 1998, Sammy Sosa finished second with 66 home runs. Figure 4 displays the empirical cdfs of the distances for the two home run kings (Keating and Scott 1999). This graph clearly shows that McGwire hits longer home runs than Sosa

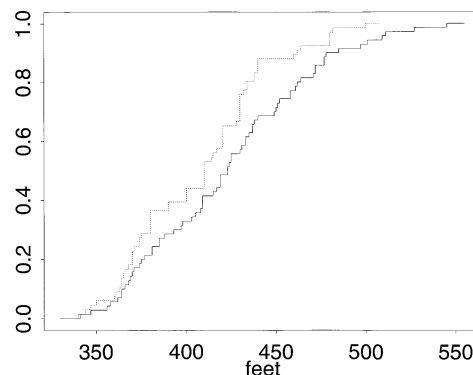


Figure 4

Empirical cumulative distribution functions of the distances of Mark McGwire's 70 (solid line) and Sammy Sosa's 66 (dotted line) home runs in 1998

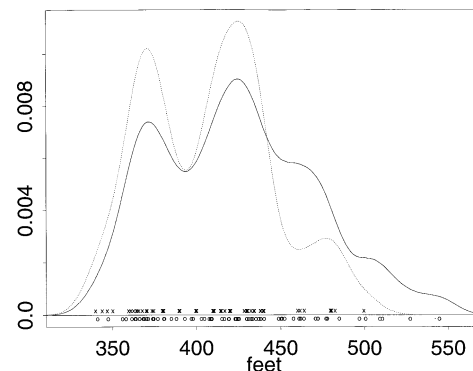


Figure 5

Kernel density estimates of the distances of Mark McGwire's 70 (solid line) and Sammy Sosa's 66 (dotted line) home runs in the 1998 season. McGwire's data are plotted below the horizontal line and Sosa's above

at every distance. In Fig. 5, kernel estimates of the McGwire and Sosa data with $h = 10$ are overlaid. The possible structure in these data is reinforced by the similarity of the two densities. Each has two primary modes at 370 ft and 425 ft, with smaller feature around 475 ft. In addition, McGwire's density clearly shows the right tail of his longest homers. In contrast, the modes in the original histogram of Fig. 3 did not appear convincing, and were eliminated by locally adjusting the bin widths. The continuity of the kernel and other estimators makes the assessment of modes much easier.

5. Multivariate Densities

Histograms can be constructed in the bivariate case and displayed in perspective, but cannot be used to draw contour plots. Kernel methods are well-suited

for this purpose. Given a set of data $(x_1, y_1), \dots, (x_n, y_n)$, the bivariate normal kernel estimator is given by

$$\hat{f}_K(x, y) = \frac{1}{n} \sum_{i=1}^n \phi(x|x_i, h) \phi(y|y_i, h)$$

A different smoothing parameter can be used in each variable.

For three and four-dimensional data, the kernel estimator can be used to prepare advanced visualization renderings of the data. Scott (1992) provides an extension set of examples.

6. Other Materials

A number of other monographs exist that delve into various aspects of nonparametric approaches. A thorough survey of all topics by top researchers is contained in Schimek (2000). Early works popularizing these ideas include Tapia and Thompson (1978) and Silverman (1986). Wand and Jones (1994) and Simonoff (1996) focus on kernel-methods. Fan and Gijbels (1996) and Loader (1999) focus on local polynomial methods. General smoothing ideas are discussed by Härdle (1991) and Bowman and Azzalini (1997). Many other excellent references may be followed in these books.

See also: Probability: Formal

Bibliography

- Bowman A W, Azzalini A 1997 *Applied Smoothing Techniques for Data Analysis: The Kernel Approach with S-Plus*. Clarendon Press, Oxford
- Fan J, Gijbels I 1996 *Local Polynomial Modelling and Its Applications*. Chapman and Hall, London
- Freedman D, Diaconis P 1981 On the histogram as a density estimator: L_2 theory. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete* **57**: 453–76
- Härdle W 1991 *Smoothing Techniques: With Implementation in S*. Springer, New York
- Hebl M, Xu J 2000 In: Lane D (ed.) Rice Virtual Lab in Statistics. <http://www.ruf.rice.edu/~lane/rvls.html>
- Keating J P, Scott D W 1999 A primer on density estimation for the great home run race of '98 *Stats* **25**: 16–22
- Loader C 1999 *Local Regression and Likelihood*. Springer, New York
- Parzen E 1962 On estimation of probability density function and mode. *Annals of Mathematical Statistics* **33**: 1065–76
- Rosenblatt M 1956 Remarks on some nonparametric estimates of a density function. *Ann. Math. Statist.* **27**: 832–837
- Rudemo M 1982 Empirical choice of histogram and kernel density estimators. *Scandinavian Journal of Statistics* **9**: 65–78
- Schimek M G (ed.) 2000 *Smoothing and Regression: Approaches, Computation, and Application*. John Wiley & Sons, New York
- Scott D W 1979 On optimal and data-based histograms. *Biometrika* **66**: 605–10

- Scott D W 1992 *Multivariate Density Estimation: Theory, Practice, and Visualization*. John Wiley & Sons, New York
- Silverman B W 1986 *Density Estimation for Statistics and Data Analysis*. Chapman and Hall, London
- Simonoff J S 1996 *Smoothing Methods in Statistics*. Springer, New York
- Tapia R A, Thompson J R 1978 *Nonparametric Probability Density Estimation*. Johns Hopkins University Press, Baltimore, MD
- Terrell G R, Scott D W 1985 Oversmoothed non-parametric density estimates. *Journal of the American Statistical Association* **80**: 209–14
- Wand M P 1997 Dated-based choice of histogram bin width histogram. *The American Statistician* **51**: 59–64
- Wand M P, Jones M C 1994 *Kernel Smoothing*. Chapman and Hall, New York
- Westergaard H 1968 *Contributions to the History of Statistics*. Agathon, New York

D. W. Scott

Probability: Formal

Probability theory is a branch of mathematics that has evolved from the investigation of social, behavioral, and physical phenomena that are influenced by randomness and uncertainty. For much of its early life, probability theory dealt almost exclusively with gambling games, and, even though there were important contributions made by such distinguished mathematicians as Pierre de Fermat, Blaise Pascal, and Pierre-Simon de Laplace, the field lacked respectability and failed to attract sustained attention.

One cause of the slow development of probability theory in its early days was the lack of a widely accepted foundation. Unlike the geometry of Euclid, or even the analytical investigations of Newton and Leibnitz, the theory of probability seemed to be eternally tied to the modeling process. In this respect, probability theory had greater kinship with the theories of heat or elasticity than with the pristine worlds of geometry or algebra.

Over time, foundations for probability were proposed by a number of deep thinking individuals including von Mises, de Finetti, and Keynes, but the approach that has come to be the most widely accepted is the one that was advanced by Andrey Nikolayevich Kolmogorov (1933).

Kolmogorov's approach to the foundations of probability theory developed naturally from the theory of integration that was introduced by Henri Lebesgue and others during the first two of decades of the twentieth century. By leaning on the newly developed theory of integration, Kolmogorov demonstrated that probability theory could be simply viewed as another branch of mathematics. After Kolmogorov's work, probability theory had the same relationship to its applications that one finds for the

theory of differential equations. As a consequence, the stage was set for a long and productive mathematical development.

To be sure, there are some philosophical and behavioral issues that are not well addressed by Kolmogorov's bare-bones foundations, but over the years, Kolmogorov's approach has been found to be adequate for most purposes. The Kolmogorov axioms are remarkably succinct, yet they have all the power that is needed to capture the physical, social, and behavioral intuition that a practical theory of probability must address.

1. Kolmogorov's Axiomatic Foundations

Central to Kolmogorov's foundation for probability theory was his introduction of a triple $(\Omega, \mathcal{F}, P)$ that is now called a probability space. The triple's first element, the 'sample space' Ω , is only required to be a set, and on the intuitive level, one can think of Ω as the set of all possible outcomes of an experiment. For example, in an experiment where one rolls a traditional six-faced die, then one can take $\Omega = \{1, 2, 3, 4, 5, 6\}$.

The second element of Kolmogorov's probability space is a collection $\mathcal{F}$ of subsets of Ω that satisfy three basic consistency properties that will be described shortly. On the intuitive level, one can think of the elements of $\mathcal{F}$ as 'events' that may occur as a consequence of the experiment described by $(\Omega, \mathcal{F}, P)$. Thus, to continue with the example of rolling a die, the set $A = \{1, 3, 5\} \subset \Omega$ would correspond to the event that one rolls an odd number.

Two of the three consistency properties that Kolmogorov imposes on the $\mathcal{F}$ are quite trivial. First, $\mathcal{F}$ is required to contain Ω . Second, $\mathcal{F}$ must be closed under complementation; so, for example, if $A \in \mathcal{F}$ then $A^c \in \mathcal{F}$ where $A^c = \{\omega: \omega \in \Omega \text{ and } \omega \notin A\}$. This third condition is only a bit more elaborate; the collection $\mathcal{F}$ must be closed under countable unions. Thus, if $A_k \in \mathcal{F}$ for $k = 1, 2, \dots$, then one requires that the union $A_1 \cup A_2 \cup \dots$ of all the events in the countable set $\{A_k: 1 \leq k < \infty\}$ must again be an element of $\mathcal{F}$.

The most interesting element of Kolmogorov's triple $(\Omega, \mathcal{F}, P)$ is the 'probability measure' P . Formally, P is just a function that assigns a real number to each of the elements of $\mathcal{F}$, and, naturally enough, one thinks of $P(A)$ as the probability of the event A . Thus, one can specify a probability model for the outcome of rolling a fair die, by 'defining' P for $A \subset \Omega$ by $P(A) = \frac{1}{6}|A|$, where $|A|$ denotes the number of elements of the set A .

For sample spaces Ω that are not finite, more care is required in the specification of the probability measures P . To deal with general Ω , Kolmogorov restricted his attention to those P that satisfy three basic axioms:

Axiom 1. For all $A \in \mathcal{F}$, one has $P(A) \geq 0$.

Axiom 2. $P(\Omega) = 1$.

Axiom 3. For any countable collection $\{A_k \in \mathcal{F}: 1 \leq k < \infty\}$ with the property that $A_j \cap A_k = \emptyset$ for all $j \neq k$, one has

$$P\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} P(A_i)$$

Axioms 1 and 2 just reflect the intuitive view that $P(A)$ measures the frequency with which A occurs in an imagined sequence of repetitions of the experiment $(\Omega, \mathcal{F}, P)$. For most mathematicians, axiom 3 also just reflects the most basic intuition about the way probabilities should behave. Nevertheless, axiom 3 is more subtle because it deals with an infinite collection of sets, and in some ways, such collections are outside of our direct experience. This has lead some philosophers to examine the possibility of avoiding Kolmogorov's third axiom, and over the years various attempts have been made to replace Kolmogorov's third axiom with the simpler assumption of 'finite additivity.'

2. Random Variables

In most applications of probability theory, the triple $(\Omega, \mathcal{F}, P)$ that gives life to probability as a rigorous mathematical subject is almost invisible. In practice, builders of probability models take advantage of various shortcuts that have the effect of keeping the probability triple at a respectful distance. The most important of these shortcuts is the notion of a 'random variable.'

On the intuitive level, a random variable is just a number that depends on a chance-driven experiment. More formally, a random variable X is a function from Ω to the real numbers with the property that $\{\omega: X(\omega) \leq t\} \in \mathcal{F}$ for all t . What drives this definition is that one inevitably wants to talk about the probability that X is less than t , and for such talk to make sense under Kolmogorov's framework, the set $\{\omega: X(\omega) \leq t\}$ must be an element of $\mathcal{F}$. Random variables make the modeler's job easier by providing a language that relates directly to the entities that are of interest in a probability model.

3. The Distribution Function and Related Quantities

There are several ways to specify the basic probabilistic properties of a random variable, but the most fundamental description is given by the 'distribution function' of X , which is defined by $F(t) = P(X \leq t)$. Knowledge of the distribution function tells one everything that there is to know about the probability theory of the random variable. Sometimes it even tells too much.

The knowledge one has of a random variable is often limited, and in such cases it may be useful to

focus on just part of the information that would be provided by the distribution function. One common way to specify such partial information is to use the ‘median’ or the ‘quantiles’ of the random variable X . The median is defined to be a number m for which one has $P(X < m) \leq 1/2 \leq P(X \leq m)$. Roughly speaking, the median m splits the set of outcomes of X into two halves so that the top half and the bottom half each have probability that is close to one-half.

The quantile x_p does a similar job. It splits the possible values of X into disjoint sets, so that one has $P(X < x_p) \leq p \leq P(X \leq x_p)$. When $p = 1/2$, the quantile x_p reduces to the median, and, when $p = 1/4$ or $p = 3/4$, then x_p is called the ‘lower quartile,’ or the ‘upper quartile,’ respectively.

4. Mass Functions and Densities

If a random variable X only takes values from a finite or countable set S , then X is called a ‘discrete’ random variable. In such cases, one can also give a complete description of the probabilistic properties of X by specification of the ‘probability mass function,’ which is defined by setting $f(x) = P(\{x\})$ for all $x \in S$. Here, one should note that the probability mass function f , permits one to recapture the distribution function by the relation

$$F(t) = \sum_{x \leq t, x \in S} f(x) \quad \text{for all } t$$

There are many important random variables whose values are not confined to any countable set, but for some of these random variables, there is still a description of their distribution functions that is roughly analogous that provided for discrete random variables. A random variable X is said to be ‘absolutely continuous’ provided that there exist a function f such that the distribution function of X has a representation of the form

$$F(t) = \int_0^t f(x) dx \quad \text{for all } t$$

The function f in this representation is called the ‘density’ of X , and many random variables of practical importance have densities. The most famous of these are the ‘standard normal’ (or, ‘standard Gaussian’) random variables that have the density

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2} \quad \text{for all } -\infty < x < \infty$$

which has a key role in the Central Limit Theorem that will be discussed shortly.

Much of the theory of probability can be developed with just the notion of discrete or absolutely con-

tinuous random variables, but there are natural random variables that do not fit into either class. For this reason, the distribution function remains the fundamental tool for describing the probabilities associated with a random variable; it alone is universal.

5. Mathematical Expectation: A Fundamental Concept

If X is a discrete random variable, then its ‘mathematical expectation’ (or just expectation, for short) is defined by

$$E(X) = \sum_{x \in S} x f(x)$$

where f is the probability mass function of X . To illustrate this definition, one can take X to be the outcome of rolling a fair die, so that $f(x) = 1/6$ for all $x \in S = \{1, 2, \dots, 6\}$ and

$$E(X) = 1 \cdot \frac{1}{6} + 2 \cdot \frac{1}{6} + \dots + 6 \cdot \frac{1}{6} = \frac{7}{2}$$

In a way that is parallel—yet not perfectly so—the expectation of an absolutely continuous random variable X is defined by the integral

$$E(X) = \int_{-\infty}^{\infty} x f(x) dx$$

Here one needs to work a bit harder to illustrate the definition. Consider, for example, an experiment where one chooses a number at random out of the unit interval $[0, 1]$. From the intuitive meaning of the experiment, one has $P(A) = a$ for $A = [0, a]$, and, from this relationship, one can calculate that the density function for X is given by $f(x) = 1$ for $x \in [0, 1]$ and by $f(x) = 0$ otherwise. For the density f , one therefore finds that

$$E(X) = \int_{-\infty}^{\infty} x f(x) dx = \int_0^1 x \cdot 1 dx = \frac{1}{2}$$

From this formula, one sees that the expected value of a number chosen at random from the unit interval is equal to one-half, and for many people this assertion is perfectly reasonable. Still, one should note that the probability that X equals one-half is in fact equal to zero, so the probabilistic use of the word ‘expected value’ differs modestly from the day-to-day usage.

The probability distribution function and the expectation operation provide almost all of the language that is needed to describe the probability theory of an individual random variables. To be sure, there are

several further notions of importance, but these may all be expressed in terms of the distribution or the expectation. For example, the most basic measure of dispersion for the random variable X is its ‘variance,’ which is defined in terms of the expectation by the formula

$$\text{Var}(X) = E(X - \mu)^2 \quad \text{where } \mu = E(X)$$

Finally, the ‘standard deviation’ of X , which is defined to be the square root of $\text{Var}(X)$, provides a useful measure of scale for problems involving X .

6. Introducing Independence

The world of probability theory does not stop with the description of a single random variable. In fact, it becomes rich and useful only when one considers collections of random variables—especially collections of ‘independent’ random variables.

Two events A and B in $\mathcal{F}$ are said to be independent provided that they satisfy the identity,

$$P(A \cap B) = P(A) P(B)$$

What one hopes to capture with this definition is the notion that the occurrence of A has no influence on the occurrence of B —and vice versa. Nevertheless, the quantities that appear in the defining formula of independence are purely mathematical constructs, and any proof of independence ultimately boils down to the proof of the defining identity in a concrete model $(\Omega, \mathcal{F}, P)$.

Consider, for example, the experiment of rolling two fair dice, one red and one blue. The sample space Ω for this experiment may be taken to be the 36 pairs (j, k) with $1 \leq j \leq 6$ and $1 \leq k \leq 6$, where one thinks of j and k as the number rolled on the red and blue die, respectively. Here, for any $A \subset \Omega$ one can set $P(A) = |A|/|\Omega|$ in order to obtain a model for a pair of fair dice. Under this probability model, there are many pairs of events that one can prove to be independent. In particular, one can prove that the event of rolling an even number on the blue die is independent of rolling an odd number on the red die. More instructively, one can also prove that the event of rolling an even number on the blue die is independent of the parity of the sum of the two dice.

7. Extending Independence

The concept of independence can be extended to random variables by defining X and Y to be independent provided that the events $\{X \leq s\}$ and $\{Y \leq t\}$ are independent for all real s and t . One easy

consequence of this definition is that for any pair of monotone functions ϕ and ψ the random variables $\phi(X)$ and $\psi(Y)$ are independent whenever X and Y are independent. This mapping property reconciles nicely with the intuitive assertion: if X and Y have no influence on each other, then neither should $\phi(X)$ and $\psi(Y)$ have any influence on each other.

The importance of independence for probability theory will be underscored by the theorems of Sect. 8, but first the notion of independence must be extended to cover more than just pairs of random variables. For a finite collection of n random variables $X_1, X_2, \dots, X_n$, the condition for ‘independence’ is simply that

$$\begin{aligned} P(X_1 \leq t_1, X_2 \leq t_2, \dots, X_n \leq t_n) \\ = P(X_1 \leq t_1) P(X_2 \leq t_2) \cdots P(X_n \leq t_n) \end{aligned}$$

for all real values $t_1, t_2, \dots, t_n$. Finally, an infinite collection of random variables $\{X_s: s \in S\}$ is said to be independent provided that every finite subset of the collection is independent.

8. The Law of Large Numbers

Any mathematical theory that hopes to reflect real-world random phenomena must provide some rigorous interpretation of the intuitive ‘law of averages.’ Kolmogorov’s theory provides many such results, the most important of which is given in the following theorem.

Theorem 1 (Law of Large Numbers). Suppose that $\{X_i: 1 \leq i < \infty\}$ is a sequence of independent random variables with a common distribution function $\mathcal{F}(\cdot)$, so $P(X_i \leq t) = \mathcal{F}(t)$ for all $1 \leq i < \infty$ and all real t . If the expectation $\mu = E(X_1)$ is well-defined and finite, then the random variables

$$\frac{1}{n} \{X_1 + X_2 + \cdots + X_n\}$$

converge to their common mean μ with probability one. More precisely, if one lets

$$A = \left\{ \omega: \lim_{n \rightarrow \infty} \frac{1}{n} \{X_1(\omega) + X_2(\omega) + \cdots + X_n(\omega)\} = \mu \right\}$$

then the event A satisfies $P(A) = 1$.

9. The Central Limit Theorem

The second great theorem of probability theory is the famous Central Limit Theorem. Although it is not tied as tightly to the meaning of probability as the Law of

Large Numbers, the Central Limit Theorem is key to many of the practical applications of probability theory. In particular, it often provides a theoretical interpretation for the ‘bell curve’ that emerges in countless empirical investigations.

Theorem 2 (Central Limit Theorem). Suppose that $\{X_i; 1 \leq i < \infty\}$ is a sequence of independent random variables with a common distribution function $\mathcal{F}$. If these random variables have a finite variance $\text{Var}(X_i) = \sigma^2$, then

$$\lim_{n \rightarrow \infty} P \left(\frac{1}{\sigma\sqrt{n}} \{X_1 + X_2 + \cdots + X_n - n\mu\} \leq x \right) \\ = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^x e^{-u^2/2} du$$

10. Stochastic Processes

The most fundamental results of probability theory address the behavior of sums of independent random variables, but many applications of probability theory lead to sequences of random variables $\{X_n; 1 \leq n < \infty\}$ that may not be independent. Such sequences are called ‘stochastic processes,’ and they are of many different types.

The simplest nontrivial stochastic process is the ‘Markov chain,’ which is used to model random phenomena where X_{n+1} depends on X_n , but, given X_n , the value of X_{n+1} does not depend on the rest of the past $X_{n-1}, X_{n-2}, \dots, X_1$. To construct such a process, one most often begins with an $n \times n$ matrix $T = \{p_{ij}\}$ with entries that satisfy $0 \leq p_{ij} \leq 1$ and with row sums that satisfy $p_{i1} + p_{i2} + \cdots + p_{in} = 1$. One then generates the Markov chain by making sequential selections from the set $\mathcal{S} = \{1, 2, \dots, n\}$ in accordance with the rows of the ‘transition matrix’ T . Specifically, if $X_n = i$, then X_{n+1} is obtained by choosing an element of $\mathcal{S}$ in accordance with the probabilities (p_{ij}) given by the i th row of T .

A second group of stochastic processes that is of considerable practical importance is the set of ‘martingales.’ Roughly speaking, these are stochastic processes that have the property that the expectation of X_{n+1} given the values of $X_n, X_{n-1}, \dots, X_1$ is just equal to X_n . One reason that martingales are important is that they provide a model for the fortune of an individual who gambles on a fair game. Such gambling games are relatively unimportant in themselves, but many economic and financial questions can be reframed as such games. As a consequence, the theory of martingales has become an essential tool in the pricing of stock options and other derivative securities.

The theory of stochastic processes is not confined to just those sequences $\{X_n; 1 \leq n < \infty\}$ with the discrete

index set $\{1 \leq n < \infty\}$, and, in fact, almost any set $\mathcal{S}$ can serve as the index set. When one takes $\mathcal{S}$ to be the set of non-negative real numbers, the index is often interpreted as time, and in this case one speaks of continuous time stochastic processes. The most important of these are the ‘Poisson process’ and ‘Brownian motion.’ Brownian motion is arguably the most important stochastic process.

11. Directions for Further Reading

For a generation, Feller (1968) has served as an inspiring introduction to probability theory. The text only assumes a modest background in calculus and linear algebra, yet it goes quite far. The text of Dudley (1989) is addressed to more mathematically sophisticated readers, but it contains much material that is accessible to readers at all levels. In particular, Dudley (1989) contains many informative historical notes with careful references to the original sources.

For an introduction to the theory of stochastic processes, the text by Çinlar (1975) is recommended, and for an easy introduction to Brownian motion, martingales, and their applications in finance, one can consult Steele (2000).

For background on the early development of probability theory, the books of David (1962) and Stigler (1986) are useful, and the article by Doob (1994) helps make the link to the twenty-first century. Finally, for direct contact with the master, anyone does well to read Kolmogorov (1933).

See also: Probability and Chance: Philosophical Aspects; Probability Density Estimation; Probability: Interpretations

Bibliography

- Çinlar E 1975 *Introduction to Stochastic Processes*. Prentice-Hall, Englewood Cliffs, NJ
- David F N 1962 *Games, Gods, and Gambling: The Origins and History of Probability from the Earliest Times to the Newtonian Era*. Griffin, London
- Doob J L 1994 The development of rigor in mathematical probability (1900–1950). In: Pier J-P (ed.) *Development of Mathematics 1900–1950*. Birkhauser-Verlag, Basel, Switzerland
- Dudley R M 1989 *Real Analysis and Probability*. Wadsworth-Brooks/Cole, Pacific Grove, CA
- Feller W 1968 *An Introduction to Probability and Its Applications*, 3rd edn. Wiley, New York, Vol. 1
- Kolmogorov A N 1933 *Grundbegriffe der Wahrscheinlichkeitsrechnung*. Springer-Verlag, Berlin (trans. Morrison N 1956 *Foundations of the Theory of Probability*. Chelsea, New York)

Steele J M 2000 *Stochastic Calculus and Financial Applications*. Springer, New York

Stigler S M 1986 *The History of Statistics: The Measurement of Uncertainty before 1900*. Harvard University Press, Cambridge, MA

J. M. Steele

Probability: Interpretations

1. Qualitative and Quantitative Probability

1.1 Introduction

It is entirely unremarkable that the commonplace word ‘probability’ and its cognates display a wide variety of distinct meanings in ordinary language. It is used as a classificatory concept, as in ‘It is *probable* that it will snow this winter’ or ‘It is *improbable* that a fair coin will land heads 10 times in succession.’ It is used as a (qualitative) comparative relation, as in ‘It is *more probable* that someone will be killed by lightning than by a falling meteorite next year.’ And, of course, it is used as a quantitative concept throughout science, e.g., Mendel’s law, ‘The probability is 0.75 that the offspring of a cross between simple hybrids will show the dominant trait.’ What is worth noting, however, is that despite longstanding controversies over the legitimate interpretation of ‘probability,’ these diverse usages can nonetheless be linked easily to a few formal theories of mathematical probability.

The received formal theory of quantitative probability, due to Kolmogorov, is elegant. Mathematical probability is a (real-valued) function, P , with values between 0 and 1 (inclusive), defined over a field of events closed under the Boolean operations of union, intersection, and complementation. The sure event Ω is certain, $P(\Omega) = 1$, and P is additive over unions of disjoint events. The qualitative formal theory of probability uses a binary relation $\prec$ also defined on an algebra of events: $E \prec F$ means that event F is more probable than event E . The qualitative formal theory is presented through axioms that mimic the quantitative one.

The outline for this article is as follows. Sections 1.2 and 1.3 review relevant details of the received, formal (mathematical) theories of quantitative and qualitative probability, and some basic connections between them. (This section includes some technical results that the reader might bypass initially, to return to on a second reading.) Section 2 surveys several interpretations of the quantitative theory, including subjective probability (2.1), logical probability (2.2), objective probability (2.3), and an approach (2.4) here

labeled the Fisher–Kyburg theory, relating to Fisher’s enigmatic fiducial probability. Additional readings are suggested in Sect. 3.

1.2 The Formal Theory of Quantitative Probability

Begin with Kolmogorov’s (1956) mathematical theory for quantitative probability, dating from the 1930s. A measure space $\langle \Omega, \mathcal{F}, P \rangle$ is triple comprised by:

(1) a set of possibilities, Ω , whose elements, ω , are points of the space. A field $\mathcal{F} = \{E \text{ of subsets of } \Omega\}$. Elements of $\mathcal{F}$ are the events of the space. It is assumed that $\Omega \in \mathcal{F}$. Being a field, $\mathcal{F}$ is closed under finite unions, intersections, and complementation with respect to Ω . When $\mathcal{F}$ is infinite, usually it is required

(2) that $\mathcal{F}$ is a σ -field, closed under countable unions and intersections.

(3) A probability function P over $\mathcal{F}$ satisfying these axioms:

$$(3.1) \quad 0 \leq P(E) \leq 1$$

$$(3.2) \quad P(\Omega) = 1$$

$$(3.3) \quad P(E \cup F) = P(E) + P(F), \text{ whenever } E \cap F = \emptyset.$$

(finite additivity)

$$(3.4) \quad P(\cup E_i) = \sum_i P(E_i) \text{ whenever } E_i \cap E_j = \emptyset \text{ for } i \neq j. \text{ (}\sigma\text{-additivity)}$$

Moreover, when $P(F) > 0$, a conditional probability given F , $P(\cdot | F)$, is well defined over $\mathcal{F}$ by the equation

$$(4) \quad P(\cdot | F)P(F) = P(\cdot \cap F).$$

Kolmogorov’s mathematical theory provides a basis for each of the two nonquantitative uses of probability, illustrated above.

The classificatory concept of probability may be explicated as:

‘ E is *probable*’ provided that $P(E) > c > 0.5$, for some value c , determined by the context of the assertion. Then, ‘ E is *improbable*’ means that the complementary event E^c is probable.

The (qualitative) comparative concept of probability may be explicated as:

$$\text{‘}E \text{ is more probable than } F\text{’ provided that } P(E) > P(F)$$

1.3 The Formal Theory of Qualitative Probability

A mathematical theory of qualitative probability is readily obtained from the quantitative one. The binary relation $\prec$ over events is said to be a (qualitative) comparative probability relation if it satisfies the following three axioms that parallel the first three axioms of Kolmogorov’s theory.

(5.1) $\prec$ is an acyclic, transitive relation over $\mathcal{F}$. That is, if $E \prec F$ and $F \prec G$ then $E \prec G$, and it is never the case

that $E < F$. The qualitative relation that E and F are equiprobable events, denoted $E \approx F$, is defined by the condition that neither $E < F$ nor $F < E$. Then $\approx$ is transitive and reflexive, i.e., $E \approx E$ always obtains.

(5.2) $\emptyset < \Omega$ and for no event E , $E < \emptyset$.

(5.3) Provided that $E \cap (F \cup G) = \emptyset$, $F < G$ if and only if $(E \cup F) < (E \cup G)$.

Say that a quantitative probability P agrees with the comparative probability relation $<$ on the condition that, $P(E) < P(F)$ if and only if $E < F$. DeFinetti (1931) asked whether every comparative probability relation $<$ has an agreeing quantitative probability. Kraft et al. (1959) established that the answer is no, even when $\mathcal{F}$ is a finite space. However, they also showed that for finite spaces, the following (stronger) additivity condition is both necessary and sufficient for an agreeing quantitative probability to exist.

(5.4) Let $(E_1, \dots, E_m)$ and $(F_1, \dots, F_m)$ be two sequences of events from $\mathcal{F}$ such that, for each point ω in Ω , the same number of events from each sequence contain ω . If it is not the case that $F_i < E_i$ ($i = 1, \dots, m-1$), then it is not the case that $E_m < F_m$.

Aside: It is interesting to note that the very same condition, expressed for a quantitative probability, serves as a necessary and sufficient condition for extending the domain of a quantitative probability, Horn and Tarski (1948).

A different relationship between quantitative and qualitative probabilities is also worth consideration. Say that a quantitative probability P almost agrees with the comparative probability relation $<$ on the condition that $P(E) < P(F)$ only if $E < F$. That is, quantitative strict inequalities entail the corresponding qualitative (strict) inequality. However, when P almost agrees with $<$, it may be that $P(E) = P(F)$ and $E < F$. Kraft et al. (1959) provide an example of a finite space equipped with a qualitative probability relation that admits a continuum of almost agreeing quantitative probabilities but no agreeing quantitative probability. For finite spaces, the analog to (5.4) is the following necessary and sufficient condition for a qualitative probability to admit an almost agreeing quantitative probability.

(5.5) Let $(E_1, \dots, E_m)$ and $(F_1, \dots, F_m)$ be two sequences of events from $\mathcal{F}$ such that, for each point ω in Ω , fewer events from the first sequence contain ω than do events from the second sequence. If it is not the case that $F_i < E_i$ ($i = 1, \dots, m-1$), then $E_m < F_m$.

On infinite spaces, the situation is more complex and more interesting. The following illustrates how a qualitative probability may make finer discriminations than any quantitative probability, i.e., the qualitative relation $<$ is non-Archimedean.

Example: Let $\mathcal{F}$ be the set of all subsets of the positive integers $\{1, 2, \dots\} = \Omega$. Let $<$ be a qualitative probability subject to the following two constraints:

(a) Each integer, i.e., each point is as probable as any other. That is, for each pair of integers i and j , $i \approx j$.

(b) The qualitative probability relation respects proper subsets. That is, whenever $E \subset F$ (but $E \neq F$) then $E < F$.

(Aside: To establish the existence of a qualitative probability satisfying these constraints requires an appeal to the Axiom of Choice.) Condition (a) ensures that no countably additive probability agrees with $<$, since any agreeing probability must assign each point of Ω probability 0. Condition (b) insures that no finitely additive probability agrees with $<$, because, by the fact just noted, $P\{i, j\} = P\{i\} + P\{j\} = 0$. But by (b) $\{i\} < \{i, j\}$. There exists an almost agreeing (uniform) finitely additive probability for $<$, where $P\{i\} = 0$ for each integer i . Hence, this qualitative probability ordering is finer than any quantitative probability can be.

2. Interpretations of Quantitative Probability

2.1 Subjective Probability

The contemporary theory of subjective (or personal) probability has its roots in the writings by Ramsey (1926) and deFinetti (1937), from the 1920s and 1930s. Savage (1954) gives what to many is a definitive account of this line of reasoning. The subjective interpretation of the mathematical theory treats probability as a (rational) agent's assessment of uncertainty about a truth-value bearing proposition, E . (Here, equate the proposition asserting E and the event E of the mathematical theory.) The agent's uncertainty about E is expressed as her/his degree of belief in E , and under appropriate assumptions may be measured by the agent's disposition to wager on E (see *Bayesian Statistics*).

The so-called 'Dutch Book' argument offers a prudential account of why the rational agent who is willing to engage in betting (under conditions made explicit, below) will have degrees of belief that conform to the formal probability calculus. Suppose that betting is in monetary units.

Definitions:

A bet on E , at odds of $r:(1-r)$, with total stake $S > 0$ pays off to the agent with a gain of $(1-r)S$ in case E obtains, and with a loss of $-r(S)$ in case E fails to obtain.

A bet against E , at odds of $r:(1-r)$, with total stake $S > 0$ pays off to the agent with a loss of $-(1-r)S$ in case E obtains, and with a gain of $r(S)$ in case E fails to obtain.

The agent's betting odds of $r:(1-r)$ for E are fair just in case the agent is indifferent between betting on and betting against E at those odds, at least for moderate size stakes.

A (finite) set of bets constitutes a Dutch Book against the agent if the agent suffers a net loss from these bets regardless which state of affairs obtains.

The agent's fair betting odds are coherent if no Dutch Book is possible against them.

An elementary version of the 'Book' argument is captured in the following result:

The Dutch Book Theorem (deFinetti and Ramsey): Suppose that the rational agent offers fair odds for each event E in the space $\mathcal{F}$ and for moderate size stakes is willing to accept finite combinations of bets made at her/his fair odds. Then the agent is coherent, that is, immune to having a Dutch Book made if and only if her/his fair odds satisfy the axioms of mathematical (finitely additive) probability, (3.1)–(3.3).

Conditional probability, too, can be made the subject of the Dutch Book criterion by introducing called-off bets, e.g., a bet on E given F , which is annulled if F fails to occur.

Definition: A called-off bet on E given F , at odds of $r:(1-r)$, with total stake $S > 0$ pays off to the agent with a gain of $(1-r)S$ in case $E \cap F$ obtains, pays off with a loss of $-r(S)$ in case F obtains but E fails to obtain, and results in no change (0 payoff) in case F fails to obtain.

The concept of a called-off bet against E given F is likewise defined in parallel with the respective concepts for betting against an event, as is the condition of fair called-off odds. When called-off bets are included among those to be wagered on, the following improved result ensues.

Dutch Book Theorem for conditional probability:

Assume that the agent offers fair odds and fair called-off odds, and for moderate size stakes is willing to accept finite combinations of wagers at these fair odds. Then the agent is coherent if and only if the fair odds satisfy the axioms of mathematical (finitely additive) probability (3.1)–(3.3), and the definition of conditional probability, (4).

- *Remark:* The reader is alerted to the sad fact that when $P(F) = 0$, condition (4) does not establish that $P(\cdot|F)$ is a probability over $\mathcal{F}$, i.e., it may fail axioms (3.1)–(3.3). Hence, the Dutch Book argument for conditional probability does *not* establish that, given an event F of probability 0, the coherent agent's called-off fair odds given F constitute a probability.

- *Remark:* Some writers use a dynamic interpretation of the conditional probability, so that $P(\cdot|F)$ designates the agents subsequent degrees of belief upon learning that event F obtains. The Dutch Book argument for conditional probability does not underwrite this dynamic interpretation. It is part of the static constraints on a rational agent's degrees of belief at a fixed time. The called-off bets may, however, serve as a model of the agent's hypothetical reasoning at the fixed time.

Shimony (1955) introduces a more restrictive standard of coherence, which he calls strict coherence. An agent's fair odds are strictly coherent provided that there is no possibility of making a 'book' against her/him where the agent stands no chance of winning on net, but risks losing for at least one point of Ω .

Shimony establishes that an agent's fair odds are strictly coherent if and only if they are coherent and only the impossible event $\emptyset$ receives probability 0. Despite the desirability of having one's odds be strictly coherent, the resulting standard precludes continuous probability distributions, unfortunately.

Savage (1954) gives an elegant version of personal probability in which the agent's probability is reduced to her/his rational preference over acts. In the briefest outline, Savage's axioms include the assumption that the agent's preferences reveal a qualitative probability, as follows. Let b and w be prizes and assume that the agent prefers receiving the better prize b to the worse one w . For an event E , let the *wager-on- E* be the act that rewards the agent with prize b if E occurs and with w if E fails. Thus, these wagers are generalization of the bets used in deFinetti's Dutch Book argument. Savage postulates that the agent strictly prefers the *wager-on- F* to the *wager-on- E* if and only if the qualitative probability relation $E \prec F$ obtains. Savage introduces additional postulates to insure that exactly one (finitely additive) quantitative probability agrees with this qualitative probability. That quantitative probability is offered as the agent's personal probability over events (see *Decision Theory: Bayesian*).

2.2 Logical (or Classical) Probability

Dating to Laplace's (1951) eighteenth-century work, at least, is the view that quantitative probability is determined once the space $\langle \Omega, \mathcal{F} \rangle$ is fixed. In its simplest version, LaPlace's *Principle of Insufficient Reason* states that the points of Ω determine equiprobable states. However, this principle leads quickly to a variety of inconsistencies, which are worth noting.

For one, LaPlace's principle makes the probability of an event depend upon which partition of states is used to describe it. A die's landing with the 1-spot up is made equiprobable with its not landing that way when Ω is just the binary partition; in which case, by LaPlace's principle, the probability is $\frac{1}{2}$ for each outcome. However, if the state space is partitioned into six points depending upon how many spots land facing upwards, LaPlace's principle yields a probability of $1/6$ for the outcome with 1-spot showing uppermost. We may continue to refine the state space, to include the information that the sum of the visible side faces on the die (as seen from a given perspective) is either less than, or is equal to, or is greater than the face showing uppermost. Then there are 14 distinct possibilities, but only one of these has a 1-spot uppermost. Then, by LaPlace's principle, the probability of a 1-spot uppermost is $1/14$.

When the state space is continuous, the problem of applying LaPlace's principle grows more serious as a transformation to an equivalent random variable may lead to a different probability. For example, if Ω is the set of real numbers in the unit interval $[0,1]$ and $\mathcal{F}$ is

the space of Borel measurable subsets of the identity variable $X(\omega) = \omega$, then Laplace's principle yields a uniform distribution over X . However, if we transform the space so that $\Omega' = [0, 1]$ with $\omega' = \omega^3$ and define the equivalent random variable $Y(\omega') = \omega'$, then Laplace's principle yields a uniform distribution when applied to the equivalent random variable Y . But a uniform probability distribution over X is evidently inconsistent with a uniform probability distribution over Y .

In an early-twentieth-century variant of Laplace's idea, Keynes (1921) strove to revitalize the principle by applying it to a formalized language in the fashion of Russell and Whitehead's *Principia Mathematica*. Moreover, Keynes's theory has the added subtlety that takes qualitative probability as primitive, and that relation is only partially defined: Not all propositions can be compared even by a qualitative probability relation in Keynes's system. Following Keynes, Carnap (1950) developed quantitative probability for a logical calculus. At first, Carnap sought to determine a single, well-defined quantitative probability for the monadic predicate calculus. But soon enough his approach, Carnap and Jeffrey (1971), became one best described as subjectivist, favoring symmetry conditions for atomic propositions similar to deFinetti's concept of exchangeability. Papers by Scott and Krauss (1966) and Gaifman and Snir (1982) continue this tradition.

A more productive alternative than appeal to a formalized language to revive Laplace's principle is the use of mathematical symmetries as a basis for fixing the necessary equiprobable states. Most notable is Jeffrey's (1961) theory of *Invariance*, dating from the 1930s. Jeffrey's sophisticated Bayesian treatment appeals to transformational invariants of parametric statistical models of the observed data to determine a 'prior' probability for the parameter. For example, a uniform 'prior' probability is the invariant distribution for a location parameter, e.g., for the mean μ of data from a normal $N[\mu, 1]$ distribution. Fraser (1966) uses mathematical groups to determine the necessary invariances for fixing a probability, e.g., Haar measure. Each of these theories provides a *local*, not a *global*, interpretation of probability. That is, these theories provide necessary conditions for probability defined with respect to a specific inference problem. When multiple inference problems are considered at once, e.g., when the same theoretical quantity is a location parameter for one data set and a scale parameter for a second data set, difficulties ensue. These theories then generate what must be seen as either inconsistent or non-unique solutions. Furthermore, the invariant probabilities are not always countably additive. So called 'improper' priors, e.g., a uniform distribution over a real-valued parameter, correspond to a purely finitely additive probability by assigning equal probability to each unit interval for the parameter, which intervals constitute the elements of a countable partition. Dawid et al. (1973) provide an important litany

of 'paradoxes' for these theories, involving both transformations of parameters and multiple, simultaneous inferences.

Yet another approach for reviving Laplace's principle is found in the original work of E. T. Jaynes (1983), who uses entropy as a guide to uncertainty. His principle of *Maximum Entropy* dictates that, subject to a set of constraints (i.e., expectations for bounded random variables), a rational agent should fix probability over the space by choosing that probability that simultaneously satisfies the constraints while maximizing entropy. (For discrete spaces this requires choosing the probability $P = \{p_i; i = 1, \dots\}$ to maximize $-\sum_i \log(p_i) p_i$ subject to the constraints.) For example, when Ω is finite and the constraint set is vacuous, the uniform distribution over Ω maximizes entropy. Unfortunately, the Maximum Entropy principle inherits the problems of Laplace's principle, as it is sensitive both to repartitioning the space Ω and, in the continuous case, equivalent transformations.

2.3 Objective Probability

2.3.1 Limiting and finite frequency interpretations. Since Venn's (1876) writings in the nineteenth century, at least, it has been popular to attempt to link probability with limiting frequency in infinite sequences of repeated trials, thereby underwriting an objective interpretation. Von Mises' (1931, 1939) work in the early twentieth century offered a mathematical program for this approach. For von Mises, probability applies to a collective, which is a suitably random sequence of repeated trials. For example, a (hypothetically infinite) sequence of repeated flips with a particular coin that can land either only heads-up or tails-up forms a collective and supports the fact that the objective probability of heads up is $\frac{1}{2}$ provided that:

(a) the limiting frequency in the sequence of the outcome heads up is $\frac{1}{2}$.

(b) the sequence of outcomes is random, i.e., this limiting frequency for heads up is invariant over selections of infinite subsequences.

Several important mathematical problems attend this approach. As is well known, the set of events that have a limiting frequency in a given sequence may fail to form a field (let alone a σ -field). Moreover, when Ω is infinite, even when limiting frequencies do exist they may fail to be σ -additive. However, the formal aspect of von Mises' program that has been given most scrutiny is the theory of randomness.

Evidently, for a sequence to pass muster as random in von Mises's sense, not all infinite subsequences are relevant. That is, assuming the limiting frequencies of the binary events {heads up, tails up} exist and are not either 0 or 1, trivially, there will be, e.g., the infinite subsequence consisting of outcomes with only heads

up, whose limiting frequency for that event is 1, not $\frac{1}{2}$. To sidestep this reduction of the randomness concept, von Mises proposed that only certain families of subsequences are relevant to objective randomness. These subsequences are determined by combinations of place selection rules, e.g., choose every k^{th} trial, and tests for after-effects, e.g., consider the subsequence of outcomes preceded by two-consecutive 'heads up' outcomes. One intuition von Mises offered for these restrictions is that the random sequence should be shown to be immune to sure profits from any feasible gambling system that might be used against it. Wald (1937) showed that a formally consistent theory was possible if the tests for randomness were limited to a countable infinity of rules. Church (1940) suggested using the denumerable class of recursive functions, in an attempt to make the concept of a random sequence language invariant. However, Ville (1939) showed that collectives based on denumerably many selection rules fail to validate some of the basic 'strong law' theorems of Kolmogorov's theory, e.g., the law of the iterated logarithm fails. Also, by appeal to martingale theory, Ville showed that such collectives fail von Mises's intuitive criterion for randomness: they do not preclude successful gambling systems.

Reichenbach (1938, 1949) proposed a rival frequency theory that avoided von Mises's randomness conditions, even permitting its application to finite sequences. In doing so, however, the resulting theory depends ever more heavily on epistemic considerations for its warranted application. For example, a (sequence) that alternates successively between heads up on odd numbered flips and tails up on even numbered flips has a limiting frequency of $\frac{1}{2}$ for each outcome. But, given that fact about the order of outcomes, we would be hard pressed to find a good use for this 'objective' probability when making inferences about such a process.

A more interesting finite-frequency interpretation is based on Kolmogorov's (1963, 1968, 1984) idea for interpreting randomness as informational complexity in finite sequences. A proper account of this approach is beyond the scope of this entry. For first-rate discussions see Fine (1973) and van Lambalgen (1987).

2.3.2 Chance and propensity interpretation. The philosophers Braithwaite (1953), Hacking (1965) and Levi (1980) propose and defend an objective interpretation of probability based on the idea that an objective chance is a theoretical, quantitative disposition for a process or kind of trial to respond when activated. That is, consider a process or kind of trial described as flipping this coin by hand. The chance of its landing heads up on such a process is taken to be an unobserved, quantitative disposition for the coin to respond when flipped. The propensity interpretation is thin, on purpose, in that chances satisfy the mathematical theory of probability by stipulation!

What is particularly novel in these approaches about making chance a theoretical concept is that probability is *not* reduced to observed frequencies, contrary to the frequency interpretation. Rather, the meaning of the theoretical term *chance* is given by rules for testing chance claims (Braithwaite) or by rules linking chance with subjective probability (Hacking and Levi). In the latter case, central to the meaning of chance is a so-called *Direct Inference* principle. Put simply, that principle asserts: the personal probability P for a specific outcome O of the kind of the trial occurring is p , given the information that the chance is p for outcome O on that kind of trial. To wit: $P(\text{coin lands heads up on a toss} \mid \text{chance is } p \text{ that the coin lands heads up on a toss}) = p$.

This treatment of chance as a theoretical term, with its meaning supplied by rules for application of the concept, accords with the Pragmatic turn in philosophy as proposed by Peirce (1932) in the nineteenth century, though Peirce himself favored a frequency interpretation.

2.4 The Fisher–Kyburg Interpretation

Fisher (1956) and Kyburg (1961) introduce an original semantics for probability, focused on the single-case assertion, e.g. (*) The probability is $\frac{1}{2}$ that the next flip of this coin lands heads up.

For this theory, probability is fundamentally epistemological, grounded on both knowledge and ignorance of class-frequencies. The statistician who asserts (*) is required to have:

(a) *knowledge* of frequencies of outcomes of heads in the population of flips with this coin, a population that includes the next flip.

(b) *ignorance* of the statistics of outcomes in a smaller statistical population of coin flips that includes the next flip.

In this sense, their theory takes instances of direct inference (based on frequency information about populations) as the definition of probability.

For both Fisher and Kyburg, this interpretation of probability underwrites an important statistical analysis, e.g., Fisher's fiducial probability. For example, let the random variable X take values from a population with known frequencies approximated by a normal distribution, $N(\mu, 1)$, but with parameter μ unknown. Then the pivotal quantity $V = (\mu - X)$ has a known distribution in this population. Its values are distributed according to a normal $N(0, 1)$ distribution. Suppose $X = 7$ is observed on a particular trial. The Fisher–Kyburg theory then asserts that on this same trial, given $X = 7$, the statistician's probability distribution for V is $N(0, 1)$. This claim rests on the assertion of ignorance on the part of the statistician: the statistician knows of no smaller population that includes this trial with statistics for V different from those of the $N(0, 1)$ population. The upshot is a

fiducial $N(7, 1)$ probability distribution for μ , given $X = 7$. When a random variable X has a continuous, parametric distribution $F(X, \theta)$, then F itself may be treated as the pivotal quantity, having a known uniform $U[0, 1]$ distribution. Fiducial inference about F , given the observation that $X = x$, yields probability for the unknown parameter satisfying Fisher's enigmatic formula: $P(\theta|X) = -\partial F / \partial \theta|_{X=x}$. Kyburg's theory has the added generality over Fisher's in that he uses interval-valued frequency information, leading to a theory with lower and upper probability for single-case assertions. Despite the considerable novelty of this approach, it is challenged severely by nuisance parameters. That is, this interpretation of probability provides only a limited account of marginal probability, as its theory of conditional probability is not well defined, at least not as required by Bayes's theorem.

3. Additional Readings

Nagel's (1936, 1939) essays on quantitative probability are excellent reviews, still, of philosophical issues relating to rival interpretations. In particular, one important issue not addressed here is whether objective probability requires a commitment to indeterminism in science. Nagel makes clear why there is no conflict between objective probability and deterministic laws in science. A more elementary presentation of rival interpretations of quantitative probability is found in Kyburg's (1970) textbook.

Fine's (1973) book gives a good overview of complexity theories and randomness, relating in particular to Kolmogorov's theory from the 1960s.

Fishburn's (1968) article offers a wealth of insights on both philosophical and formal questions relating qualitative and quantitative concepts of probability. The bibliography there is particularly noteworthy. Also valuable is the discussion of comparative probability given in Walley and Fine (1979).

See also: Probabilistic Thinking; Probability and Chance: Philosophical Aspects; Probability Density Estimation; Probability: Formal

Bibliography

- Braithwaite R B 1953 *Scientific Explanation*. Cambridge University Press, Cambridge, UK
 Carnap R 1950 *Logical Foundations of Probability*. University of Chicago, Chicago
 Carnap R, Jeffrey R (eds.) 1971 *Studies in Inductive Logic and Probability*. University of California Press, Berkeley, CA, Vol. 1
 Church A 1940 On the concept of a random sequence. *Bulletin of the American Mathematical Society* **46**: 130–5
 Dawid P, Stone M, Zidek J V 1973 Marginalization paradoxes

- in Bayesian and structural inference. *Journal of the Royal Statistical Society* **35**: 189–233 (with discussion)
 deFinetti B 1931 Sul significato soggettivo della probabilit . *Fund Math* **17**: 298–329
 deFinetti B 1937 *Foresight: Its Logical Laws, Its Subjective Sources*. Trans. in Kyburg and Smokler (eds.) 1964
 Fine T 1973 *Theories of Probability*. Academic Press, New York
 Fishburn P C 1986 The axioms of subjective probability. *Statistical Science* **1**: 335–58 (with discussion)
 Fisher R 1956 *Statistical Methods and Scientific Inference*. Hafner Press, New York
 Fraser D A S 1968 *The Structure of Inference*. John Wiley, New York
 Gaifman H, Snir M 1982 Probabilities over rich languages, testing and randomness. *Journal of Symbolic Logic* **47**: 495–548
 Hacking I 1965 *Logic of Statistical Inference*. Cambridge University Press, Cambridge, UK
 Horn A, Tarski A 1948 Measures in Boolean algebras. *Transactions of the American Mathematical Society* **64**: 467–97
 Jaynes E T 1983 In: Rosenkrantz R (ed.) *Papers on Probability, Statistics, and Statistical Physics*. D. Reidel Publishing, Dordrecht, The Netherlands
 Jeffreys H 1961 *Theory of Probability*, 3rd edn. Oxford University Press, London
 Keynes J M 1921 *Treatise on Probability*. Macmillan, London
 Kolmogorov A N 1956 *Foundations of Probability*. Chelsea Publishing, New York
 Kolmogorov A N 1963 On tables of random numbers. *Sankhya* **25**: 369–76
 Kolmogorov A N 1968 Three approaches to the definition of the concept of 'amount of information'. *Sel. Transl. Math. Stat. and Prob* **7**
 Kolmogorov A N 1984 The logical basis for information theory and probability theory. *IEEE Trans. Inf. Theory*, **IT 14**: 662–4
 Kraft C, Pratt J, Seidenberg A 1959 Intuitive probability on finite sets. *Ann. Math. Stat.* **30**: 408–19
 Kyburg H E 1961 *Probability and the Logic of Rational Belief*. Wesleyan University Press, Middleton, CT
 Kyburg H E 1970 *Probability and Inductive Logic*. Collier–Macmillan, London
 Kyburg H E, Smokler H E (eds.) 1964 *Studies in Subjective Probability*. John Wiley, New York
 LaPlace P S 1951 *A Philosophical Essay on Probabilities*. Dover, New York
 Levi I 1980 *The Enterprise of Knowledge*. MIT Press, Cambridge, MA
 van Lambalgen M 1987 *Random Sequences*. University of Amsterdam, Amsterdam
 Mises R von 1931 *Wahrscheinlichkeitsrechnung und ihre Anwendungen in der Statistik und theoretischer Physik*. Deuticke, Leipzig
 Mises R von 1939 *Probability, Statistics, and Truth*. Macmillan, New York (Translation of [1928] *Wahrscheinlichkeit, Statistik, und Wahrheit*. Wien.)
 Nagel E 1939 *Principles of the Theory of Probability*. University of Chicago Press, Chicago
 Nagel E, Margenau L, Ducasse C J, Wilks S S 1936 The meaning of probability. *Journal of the American Statistical Association* **31**(193): 10–30
 Peirce C S 1932, 1936 *Collected Papers*. Harvard University Press, Cambridge, MA, Vols. 2 and 6
 Ramsey F P 1926 Truth and probability. In: Braithwaite R B (ed.) *The Foundations of Mathematics 1950*. Humanities Press, New York

- Reichenbach H 1938 *Experience and Prediction*. University of Chicago Press, Chicago
- Reichenbach H 1949 *Theory of Probability*. University of California Press, Berkeley, CA
- Savage L J 1954 *The Foundations of Statistics*. John Wiley, New York
- Scott D, Krauss P 1966 Assigning probabilities to logical formulas. In: Hintikka J, Suppes P (eds.) *Aspects of Inductive Logic*. North-Holland, Amsterdam
- Shimony A 1955 Coherence and the axioms of confirmation. *Journal of Symbolic Logic* 20: 644–60
- Venn J 1876 *The Logic of Chance*, 2nd edn. Macmillan, London
- Ville J 1939 *Etude critique de la notion de collectif*. Gauthiers-Villars, Paris
- Wald A 1937 Die Widerspruchsfreiheit des Kollektivsbegriffs der Wahrscheinlichkeitsrechnung. *Ergebnisse eins mathematisches Kolloquiums* 8: 38–72
- Walley P, Fine T 1979 Varieties of modal (classificatory) and comparative probability. *Synthese* 41: 321–74

T. Seidenfeld

with the question of whether this ethos fitted the spirit of entrepreneurial capitalism. Yet in the 1920 preface to his collected papers on world religions, Weber's interest was rationalization processes, including the rise of capitalism. These changes had gone furthest in the West, and Weber wondered why.

If these two stories about sociology's questions carry a moral, the lesson is not that persons nowadays regarded as sociology's founders did not know what they were up to. The message amounts to a methodological command on meritorious problems. Durkheim grasped that seemingly small questions become pertinent when taken as subproblems of larger ones, and Weber understood that it is wise to divide mammoth problems into more modest ones, and those in their turn into even more tractable questions. Durkheim and Weber entered sociology's hall of fame because they pointed towards a bunch of smaller questions falling under one and the same heading. The subtitle of *Le suicide* runs *Étude de sociologie*, lifting a phenomenon from psychology into that science which studies society—and the extreme case of leaving it.

Problem Selection in the Social Sciences: Methodology

Social methodology textbooks state techniques of proper empirical analysis. For instance, they tell how partial coefficients of association are to be computed, adding that these coefficients provide a more severe test of multifactor explanations than bivariate associations. Books on theory building in sociology set standards for determining whether certain hypotheses are better than others. Logically unfalsifiable ones are out, future propositions should provide more information than current ones. Any science starts with problems. Are there precepts for selecting problems too? Yes, and this entry draws upon principles intimated by sociologists (Lenski 1966, p. 20, Blau and Duncan 1967, pp. 9–10, 194–9, 402), proposed by philosophers of science (Bunge 1967, pp. 165–221), and applied by sociologists to sociology (Merton 1959, Ultee 1980). These standards are elucidated by cases from sociology's past and present.

1. Overarching Problems and Subproblems

Durkheim 'was not interested in suicide at all,' so says Collins' 1994 textbook on sociological traditions. But then, why did Durkheim dub his 1897 study *Le suicide*? Weber in 1910 grumbled that his critics understood him to explain the rise of capitalism fully from the ethos of protestantism, and retorted that he only dealt

2. Problem Structures and Their Dynamics

For Durkheim, a society's suicide rate tells about its cohesion: the higher the rate, the less cohesive a society. Weber observed that a society's inhabitants do not always use their economic freedoms, for instance those of enterprise and labor, and supplemented the problem of how these rights became included in a society's code of law, with the problem of the strength of a person's motivation to use these liberties. Both questions were part of the question of why entrepreneurial capitalism arose in the West. In its turn, the question of the rise of capitalism, including bankers' as well as entrepreneurial capitalism, was a subquestion of the even larger question of the speed of rationalization processes in the West during the last couple of centuries.

In Durkheim's *oeuvre*, the question of a society's degree of cohesion also comprises long-term trends in the prevalence of economic exchanges when comparing ancient Israel, the Roman Empire, and medieval Burgundy, and the degree to which assemblies of hunters and gatherers are religious in nature. Weber's question of rationalization led to less grand ones on the rise of science and the bureaucratization of states. Any methodology of problem selection should contain the rule that questions should not be judged on their own, but according to their place in a structure of overarching problems and subproblems. However small a problem seems, it has more appeal if it is part of an overarching problem. And however towering a question appears, it grows less scary if it is broken down into subquestions.

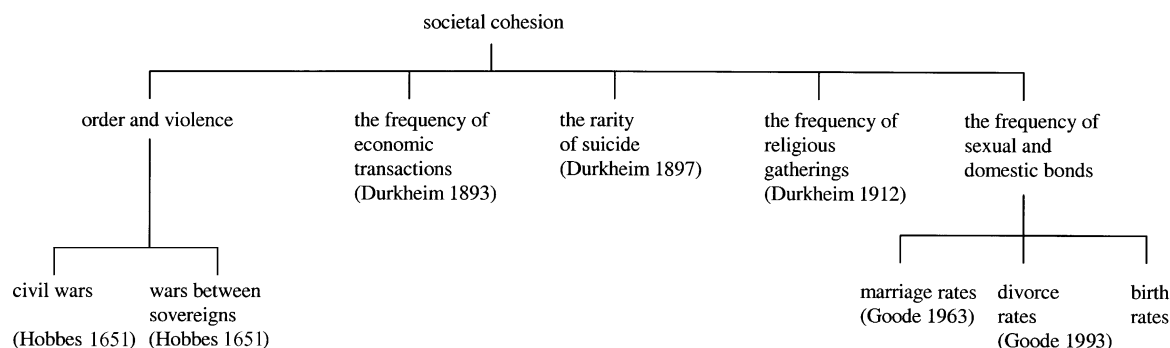


Figure 1
The Durkheimian problem structure and its dynamics

2.1 The Durkheimian Problem Structure

Figure 1 depicts the Durkheimian problem structure. The novelty of *Le suicide* was not simply that Durkheim turned the question of suicide into a subquestion of the larger one of societal cohesion. Hobbes (1651) took order and its opposite violence as the master problem of those studying the body politic. For him, civil wars and wars between sovereigns were part of the problem of order and violence. Next to it, Durkheim placed the problem of the presence and absence of ties between a society's members. The counterpart of the question 'why do people not always live in peace?' is not only 'why do they sometimes use violence?' It also is 'why do they sometimes live relatively isolated from each other?' For Durkheim, suicide was the supreme example of severing the ties between persons and the society in which they live.

Apart from *Le suicide*'s question, Fig. 1 encompasses the questions guiding Durkheim's two other main studies. *De la division du travail social* from 1893 focuses on the frequency of economic transactions between a society's members at various points in history. It narrows this question down to that of the degree to which cooperative law (civil law regulating economic contracts by restitution) is more widespread than criminal law (penal law involving repressive sanctions). *Les formes élémentaires de la vie religieuse* from 1912 notes that Australian aborigines disperse when hunting and gathering their food, and asks to what extent their leisurely meetings celebrate a totem and turn them into members of one society. Durkheim's questions involve a shift from questions on economic ties, by way of questions on co-living, to questions on religious ties, yet they always pertain to societal cohesion.

Given the Durkheimian problem structure, later scholars can raise questions on types of attachment to which Durkheim paid limited or no attention. Examples are questions on changes in the rates of marriage, birth and divorce in industrial societies. A

wedding creates a sexual and domestic bond between two members of the opposite sex, divorce dissolves it, and a higher birth rate amounts to more parent-child relations in a society. Goode (1963, 1993) focuses on the making and undoing of sexual and domestic ties. These monographs amount to a new branch of the Durkheimian problem structure.

2.2 The Weberian Problem Structure

The Weberian problem structure and its dynamics are given in Fig. 2. The overarching problem is that of rationalization, and Merton (1938) pursued the subproblem of the rise of science in the West, and Blau (1974) that of bureaucratization. The art historian Gombrich (1960) studied questions on the skills of painters to furnish a faithful representation.

The relation between problems as depicted in Figs. 1 and 2, are logical in nature. An overarching problem, usually by way of uncontroversial auxiliary assumptions, implies subproblems. It is another, and less important, matter whether a person raising a subproblem of the overarching problem of an earlier scholar, studied under that master or read a book by him or her. Merton and Blau took the lead from Weber, Gombrich was not influenced by him. Weber invokes Heinrich Wölfflin once when discussing the question of representational skills, Gombrich quotes Wölfflin frequently.

As stated, the step from an overarching question to a subquestion involves largely undisputed assumptions. Few would argue that capitalism is less efficient than slavery. Yet this proposition was belied by Fogel and Engerman's (1974) test for the American South shortly before slavery was abolished. A scholar working on a question can make progress by replacing these auxiliary assumptions by better approximations. Weber did not do so, but Durkheim did. In countries prescribing suicide under certain circumstances, suicide is not an instance of low cohesion, but of

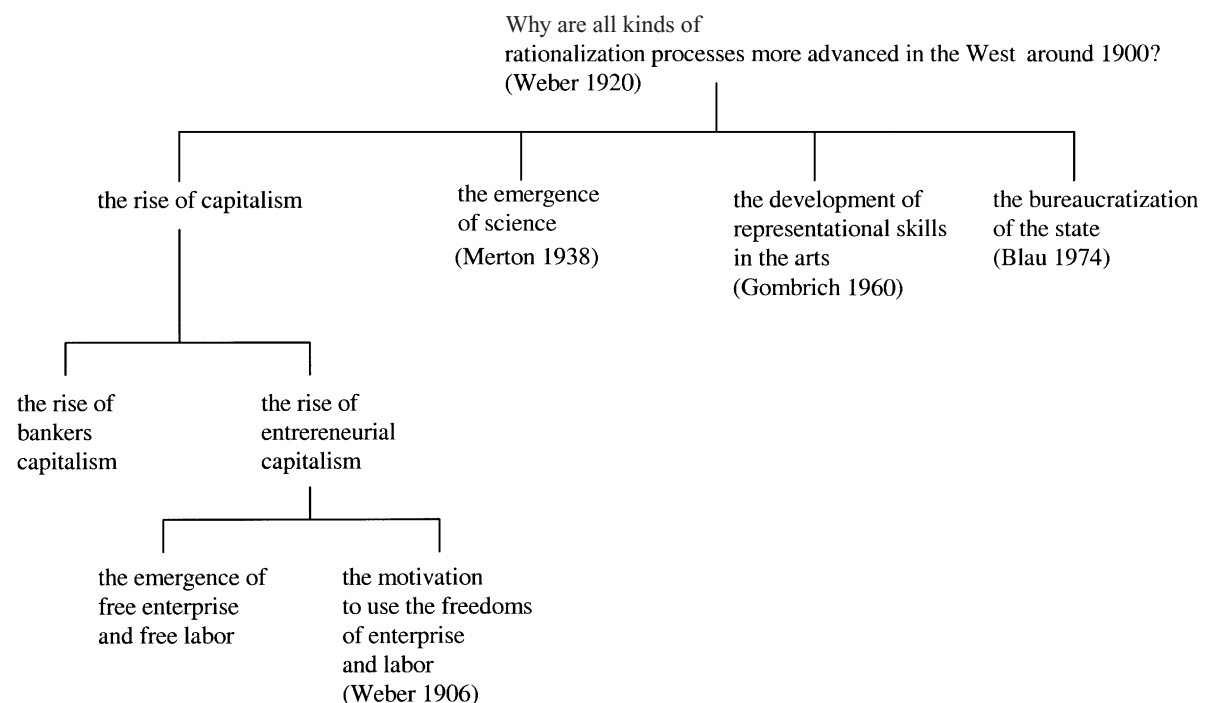


Figure 2
The Weberian problem structure and its dynamics

strong solidarity. In that case, Durkheim spoke of altruistic suicide.

Behind Weber's distinction between bankers' and entrepreneurial capitalism lurks evidence speaking against the hypothesis that the rise of capitalism is always accompanied by the growth of banking. Protestant businessmen did not take out loans. Their religion put a premium on frugality, and contributed to reinvestment of profits in their enterprise. The background of Weber's distinction between laws granting people economic freedoms, and the use of these freedoms by a society's members, forms the observation that people sometimes do not use their rights, as well as the hypothesis that they exercise these rights only if they are motivated to do so. Classical economics assumed that every person had this drive, Weber held that protestantism strengthened it. A wise rule is to state the background of a question as fully as seems necessary for bringing out a problem structure.

2.3 The Weberian Problem Structure Radically Revised

The division of an overarching problem into sub-problem may drastically alter. Studies by Smith (1776), Elias (1939), Samuelson (1954), Van den Doel (1979), and De Swaan (1988) imply the revision of the

Weberian problem in Fig. 3. Weber never made clear what state bureaucracies were doing. States possess the legitimate monopoly on the means of power, Elias added that states wield the legitimate monopoly on taxation. Were all taxes spent on armies, police and prisons? This obviously is not the case. Here the hypothesis is useful that markets make for the optimal production of individual goods (Smith), but not for the optimal production of collective goods (Samuelson). The production of collective goods approaches optimality in a state with universal suffrage and a bureaucracy which executes its laws (van den Doel).

De Swaan listed collective bads and laws contributing to their elimination and to the production of collective goods. Firstly, poverty is a collective bad. Poor people are more likely to trespass the law, and crimes form a general threat to life and property. Following De Swaan, this danger is alleviated by state insurance against disability and unemployment. Contagious diseases amount to a second collective bad. If local authorities provide each household with piped water and an outlet on the sewers, public health improves. Thirdly, compulsory schooling lowers the percentage of children who cannot read and write, and makes for a more optimal functioning of markets and bureaucracies. Weber's conception of the state is saved by the assumption that armies, police and prisons

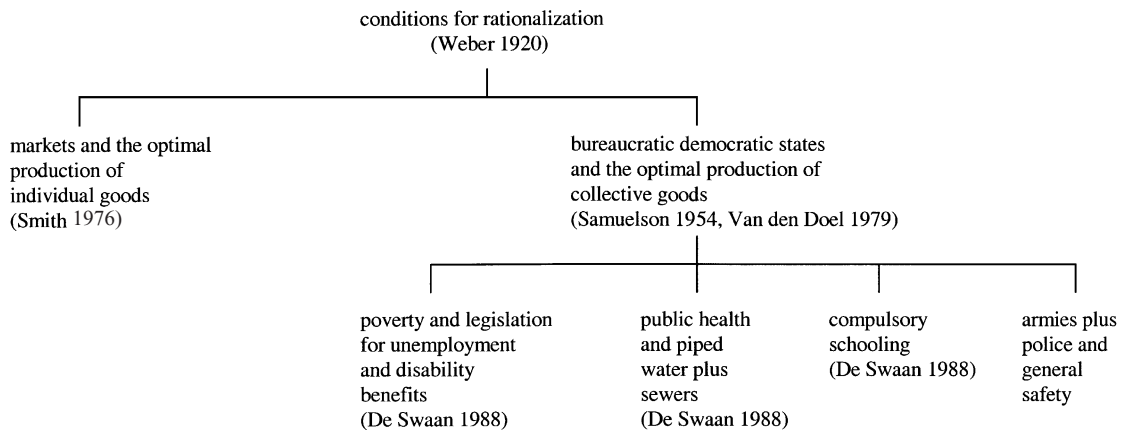


Figure 3
The Weberian problem structure radically revised

avert collective bads. Armies lessen a threat to a state's independence, police patrols contribute to safety on the streets, prisons deter.

2.4 Problem Structures and Theories

As the saying goes, a question well put is a question half answered. If it is clear that some problem is part of a larger problem, and if a solution for the overarching problem is at hand, a possible solution to the smaller problem is available too. But then, if an answer to a particular smaller question of a larger one turns out to be false, the opportunities to disestablish a solution to another subproblem of the larger one increase. Additionally, the notion that problem structures exist and have a dynamic, makes the project of sociology have more continuity than suggested by Boudon (1979) and Gove (1995). They accepted that sociology's questions to outsiders show a bewildering variety, and bypassed the possibility that one and the same larger question may harbor a host of seemingly disparate questions.

3. Contradictions and Issues

One way of formulating a problem in a more useful way is to ascertain its place in a problem structure. Are there much injunctions for problem choice? Sociologists debating sociology have argued about the primacy of theory vs. the supremacy of research. In contrast, Popper (1962) asserted that every science, including sociology, most fruitfully starts with contradictions between general hypotheses and specific research findings.

Contradictions—or, to use Kuhn's 1962 terms, puzzles and anomalies—contribute to sociology's growth. For Merton's (1973) sociology of science, the

prevalence of priority disputes in a field was surprising in several respects. Its explanation was not the egotism of the scientists having made discoveries, since their part in the disputes was small. Nor was it to be accounted for by the increasing competition between scientists, since the frequency and intensity of priority disputes in the West decreased over time. A rule of method runs: single out contradictions.

Incompatibilities between hypotheses also invite new research and theorizing. They sometimes are called issues, but that term also refers to false alternatives. An example is the question of whether intelligence is determined by social or by genetic factors. Intermediate positions are possible, the list of social factors may be incomplete, and the location of intelligence on genes may be unknown. Popular science states issues in a metaphorical way: are we born with a brain hardwired with rules for language, or is the brain a big sponge, capable of creating rules for language by mopping up what we hear?

The phrasing of a contradiction problem sometimes contains seeds for its solution. For instance, Huizinga (1941) asked how it was possible that painting and other forms of culture flourished in the Netherlands during the seventeenth century, so shortly after this society became independent from Spain. His answer was that Dutch culture blossomed because the Netherlands was a new state. In states which have recently become independent, the upper social strata have risen from the lower and the more open a country's elite, the more its culture will flower. Trevor-Roper (1976) wondered how it was possible that an emperor as preoccupied with expanding his territory as Charles V (1500–58), furthered the art of painting so much. In the end Trevor-Roper surmised that Charles V commissioned paintings to the extent that, and precisely because, they legitimated his political power.

Apart from the rule for theory building to mobilize the information in a contradiction problem, there is the precept to turn the problem around. Rather than asking why the number of unruly passengers on airlines is increasing, it might be worthwhile to ponder why this number did not increase much more strongly until now. A more academic example is as follows. Marxism noted that a lot of workers refrain from joining unions and strikes, and raised the question of how workers come to have a false consciousness. Olson (1965) held that unions and strikes produce collective rather than private goods, applied the *prisoners' dilemma* from game theory in mathematics predicted that rational workers would not strike, and raised the question of why people participate in social movements. However, it should be clear that any theory with some substance to it should explain that some people join labor unions and certain others do not.

4. *Strings of Questions*

The terms problem and question sometimes are used interchangeably, referring to queries of any specificity. However, everyday usage of expressions like object of investigation, task, question and problem, suggest that some inquiries are more focused than others. What does the articulation of problems involve?

4.1 *Problem Articulation*

If it is maintained that sociology is about societies (and not about individuals), it sometimes is said that the object of research has been stipulated. Yet no single feature of this unit has been singled out for explanation. If this property is taken to be, say, the 'functioning' of a society, the object of inquiry turns into what has been called a task. Yet the circumscription of this aspect remains limited, since the term does not focus any particular way of functioning. If a distinction is made between optimal and suboptimal functioning, the study involves a question, with the term question being taken in a narrow sense. Of course, other types of functioning may be distinguished too. Questions in their turn do not allude to possible causes. If research on some sample of societies determines the extent to which less stability in a society's environment goes together with a less optimal functioning of this society, the matter at hand is articulated so fully that the problem now is whether some logically falsifiable hypothesis is true. Putting a question mark behind a falsifiable proposition amounts to posing a problem in some narrow sense of that word.

The example of a string of increasingly articulated queries just presented neglects any difficulty in ob-

taining measurements for the degree of stability in a society's environment, and the extent to which the functioning of a society approaches optimality. It has been held that questions should refer to directly observable phenomena only, and therefore in the human sciences to outward behavior and not to mental states. This rule is a leftover from a wrongheaded empiricist philosophy of science. All terms to some extent are theoretical, and any proposition whatsoever, whether about acts or thoughts of persons, or processes and situations within of societies, never can be definitely proved.

4.2 *From Descriptive to Explanatory Questions*

There are strings of questions of another kind. A researcher's early questions often are descriptive: how many of the people living in 1972 Britain and born to working-class parents had moved up to the higher classes? A follow-up question might be whether this probability to climb equals the chance to descend from the upper to the lower classes. A yet later question presupposes the answers to both earlier questions, and seeks to ascertain whether upward and downward mobility were as prevalent in 1949 as in 1972. By raising additional descriptive questions, narratives become richer.

At some point researchers leave descriptive questions behind them, to raise explanatory questions. The explanatory question following up on the previous string of descriptive questions ran: why during the 1970s was upward mobility less widespread in Britain than in Sweden, and more widespread than in France? The obvious answer to that question was that left-wing governments make for more upward mobility (Heath 1981, Goldthorpe 1987).

4.3 *From Superficial to Deeper Explanatory Questions*

Some explanations 'scratch the surface only,' others delve deeper. Why would particular ideologies make for more upward mobility? Just as individuals, governments cannot always attain their goals. By which means then do social democrats attain the goal of more upward mobility? Does Sweden have more upward mobility than Britain, because social democrats make children branch out at a later age to schools which demand more or less of pupils? Or is this the case because social democrats pursue a policy of full employment? And then, if the second question fares better in empirical research, why would there be more upwardly mobile persons in a society without unemployed persons?

There is no obvious point at which the hunt for deeper explanations should stop. Some theoreticians

in effect defend that 'the buck stops here.' No one nowadays seriously defends propositions about societies being logically independent from propositions about individuals. However, whether rock-bottom explanations of the features of societies should always refer to the motives and aversions of a society's members, as well as the opportunities they have and the barriers they face, is controversial. In the discussion on 'methodological individualism' (O'Neill 1973) the borderline between methodological rules and substantive requirements was violated. Perhaps it is wise to state this debate no longer in terms of theories explaining societal phenomena, but to settle it by recognizing that shorthand statements of sociology's questions refer to societal features only, and that these questions, if they are stated fully, always refer to the persons making up a society's population, that is, to *individuals* and *societies* (Ultee 1997).

4.4 From Meagre to Thick Descriptive Questions

Some descriptive questions—and their ensuing explanatory questions can be meagre, whereas others aim for lavishness. Thus, it often has been ascertained that no society ever has been or ever will be completely egalitarian. As Lenski (1966, p. 20) maintained, in the debate between the conservative and the radical tradition in sociology, it is possible to make headway by transforming problems stated in a categorical way into problems phrased in variable terms. It may indeed be the case that full equality never has been observed in any society, but is the lowest degree of inequality observed until now the lowest that is possible? And do long-standing differences between highly industrialized countries in their ruling ideology explain differences in their degree of income inequality? Questions about equality that demand 'yes' or 'no', are less precise than questions about degree of inequality.

Following Lenski, progress also can be made by breaking down the compound concepts contained in problems into their constituent elements. Thus the question of whether power—in the singular—determines privilege—in the singular—will not do. Rather, the question is to what extent differences in the consumption of a certain good in a particular society depend on the control of various resources. When answering this question, Lenski distinguished the skills of hunters and gatherers, the size of gardens in horticultural societies and of fields in agrarian societies, the possession of metal weapons, the private ownership of machines, the technical skills of the people designing machines, the range of legitimate command accruing to a political position, and the right to vote in elections. Lenski also suggested that the change from agriculture to industry was accompanied by the shift, from the right to vote in infrequent political elections for only the propertied classes, to universal suffrage in periodic elections. By breaking

down compound problems into their constituents, thick descriptions and detailed explanations are obtained.

4.5 Erroneous and Well-stated Questions

If explanatory questions are logically prior to descriptive questions, explanatory questions may be wrong. Answers to questions can turn out false, questions themselves may be mistaken. The question of why in the 1970s social mobility was more widespread in France than in Sweden is erroneous, since in that decade mobility occurred more often in Sweden. The Olson question, of exactly which selective incentive makes persons protest against the hunting of seals, assumes that some selective incentive is always at work, which may not be the case. This example suggests that if a rock-bottom explanation slips into a question, that problem turns into a leading question.

Methodological rules which stipulate conditions banishing erroneous questions, and guaranteeing well-stated ones, seem doomed. It is wise to review the presupposed description, but checking has to stop somewhere. In this respect it is interesting that the answer to an explanatory question always implies that the original question may not have been fully right. When Durkheim asked why protestantism goes together with more suicide than catholicism, he devised the hypothesis that more tightly knit religions show a lower suicide rate. To derive what he wished to explain, Durkheim assumed that catholics are more integrated than protestants. He knew that this assumption was off the mark for England, since the Anglican Church was a state church. Yet in this way he could account for England as the deviant case to the regularity he wished to explain. Malewski (1964) showed with examples from social psychology that every explanation of a regularity predicts conditions under which exceptions occur to the regularity being explained.

Duncan held that the question of factors influencing the chances that the members of a society move up or down poses the issue poorly. The prime factor is the father's occupation. However, the proposition that the lower the occupational prestige of a person's father, the higher this person's chances of upward mobility, forms a truism. After all, the hypothesis thrives on bottom- and ceiling-effects. To test more interesting hypotheses, according to Duncan, questions on change in a person's life should not be stated in terms of some difference score, but as questions on the strength of the relation between a person's earlier and later score, and on the factors influencing that relation.

4.6 Undertheorized Questions

Schütz (1933) raised the question of why individuals understand one another, and did so with a certain

urgency by phrasing his question as 'how it is possible that people understand each other?' This query sounds like a contradiction, but although casual observation suggests that some people in some places and at some times indeed do understand one another, no particular hypothesis has been derived from a specific theory that holds that it is impossible for human beings to understand each other. Thus the theory rendering the question compelling is lacking. The Parsonian problem of how order is possible (Parsons 1937) is undertheorized in a similar way. Any derivation of the hypothesis that society is violent, from the principle that persons follow their interests, involves additional assumptions. These can be wrong, which makes the question of how order is possible less pressing than it sounds. Indeed, a full statement of the question 'how is order possible?' implies the conditions under which order occurs, and those under which it does not.

Recent research addresses the question of whether the association observed in a cross-sectional sample of a society's inhabitants between their year of birth and a particular characteristic, is to be accounted for as an age, cohort, or period effect. As methodologists point out, this question is undertheorized too. The available hypotheses do not predict some age effect, some cohort effect and some period effect. Those hypotheses are so vague that their conjunction is almost, if not fully, logically unfalsifiable. The hypotheses at hand have more substance: if people marry, grow up in lean times, and live through a war, then their values will be less post-materialist (Inglehart 1977). Those variables can be independently ascertained, making statistical models identifiable.

5. Conclusion

Ziman (1987) argued that the question of how scientists choose their problems is erroneous, since most researchers do not have a choice. Their problems are pre-given by the organization which employs them. This observation does not do away with the question of problem choice, but makes the question of the formal rules for problem choice by the higher authorities of these organizations more pressing. The methods outlined above should not only be taken as principles for researchers who have a choice, but also as precepts for institutes run on a budget, and for foundations which award subsidies.

In everyday life researchers consider some problems difficult, others significant, and yet others exciting. These notions are difficult to put into formal rules. However, a more difficult problem might be taken as an explanatory question to which several solutions have been proposed, none of which worked or led to closer approximations. Thus, difficult problems are specific kinds of Popperian contradictions. Also, a structure consisting of overarching problems and subquestions, goes some way towards formalizing the

intuition that a problem is significant. Perhaps a problem area is called exciting, if, after many a failed attempt, the solution of an overarching problem seems near, because progress is being made on subproblems.

See also: Generalization: Conceptions in the Social Sciences; Normative Aspects of Social and Behavioral Science; Prisoner's Dilemma, One-shot and Iterated; Quantification in the History of the Social Sciences; Theory: Conceptions in the Social Sciences; Traditions in Sociology

Bibliography

- Agassi J 1975 *Science in Flux*. D Reidel, Dordrecht
- Blau P M 1974 *On the Nature of Organizations*. Wiley, New York
- Blau P M, Duncan O D 1967 *The American Occupational Structure*. Wiley, New York
- Boudon R 1979 *La logique du social*. Hachette, Paris [1981 *The Logic of Social Action*. Routledge and Kegan Paul, London]
- Bunge M 1967 *Scientific Research I, The Search for System*. Springer, Berlin
- Collins R 1994 *Four Sociological Traditions*. Oxford University Press, New York
- De Swaan A 1988 *In Care of the State*. Polity Press, Cambridge, UK
- Durkheim É 1893 *De la division du travail social*. F Alcan, Paris [1933 *The Division of Labour in Society*. Macmillan, New York]
- Durkheim É 1897 *Le suicide, étude de sociologie*. Alcan, Paris [1951 *Suicide. A Study in Sociology*. Free Press, Glencoe, IL]
- Durkheim É 1912 *Les formes élémentaires de la vie religieuse*. Alcan, Paris [1995 *The Elementary Forms of Religious Life*. Free Press, New York]
- Elias N 1939 *Über den Prozess der Zivilisation*. Hans zum Falken, Basel [1994 *The Civilizing Process*. Basil Blackwell, Oxford, UK]
- Fogel R W, Engerman S L 1974 *Time on the Cross. The Economics of American Negro Slavery*, 1st edn. Little, Brown, Boston
- Goldthorpe J H 1987 *Social Mobility and Class Structure in Modern Britain*, 2nd edn. Clarendon Press, Oxford, UK
- Gombrich E H 1960 *Art and Illusion, A Study in the Psychology of Pictorial Representation*. Phaidon, New York
- Goode W J 1963 *World Revolution and Family Patterns*. Free Press, Glencoe, New York
- Goode W J 1993 *World Changes in Divorce Patterns*. Yale University Press, New Haven, CT
- Gove W R 1995 Is sociology the integrative discipline in the study of human behavior? *Social Forces* 72: 1197–206
- Heath A 1981 *Social Mobility*. Croom Helm, London
- Hobbes T 1651/1968 *Leviathan*. [Penguin, Harmondsworth, London]
- Huizinga J 1941 *Nederland's beschaving in de 17^{de} eeuw*. Tjeenk Willink, Groningen [1969 *Dutch Civilization in the 17th Century*. Harper and Row, New York]
- Inglehart R 1977 *The Silent Revolution*. Princeton University Press, Princeton, NJ
- Kuhn T S 1962 *The Structure of Scientific Revolutions*. University of Chicago Press, Chicago
- Lenski G E 1966 *Power and Privilege*. McGraw-Hill, New York

- Malewski A 1964 Generality levels of theories concerning human behavior. In: Zetterberg H L, Lorenz G A (eds.) *Symposium on Theoretical Sociology*. Bedminster, Totowa
- Merton R K 1938 *Science, Technology and Society in Seventeenth-Century England*. Saint Catherine Press, Bruges
- Merton R K 1959 Notes on problem finding in sociology. In: Merton R K, Bloom L, Cottrell L S Jr (eds.) *Sociology Today: Problems and Prospects*. Basic New York
- Merton R K 1973 *The Sociology of Science*. University of Chicago Press, Chicago
- Olson M Jr 1965 *The Logic of Collective Action*. Harvard University Press, Cambridge, MA
- O'Neill J (ed.) 1973 *Modes of Individualism and Collectivism*. Heinemann Educational, London
- Parsons T 1937 *The Structure of Social Action*. McGraw-Hill, New York
- Popper K R 1962 Die Logik der Sozialwissenschaften. *Kölner Zeitschrift für Soziologie und Sozialpsychologie* 14: 233–48. [1976 The logic of the social sciences. In: Adorno T W, Albert H, Dahrendorf R, Habermas J, Pilot H, Popper K R. (eds.) *The Positivist Dispute in German Sociology*. Harper and Row, New York]
- Schütz A 1933 *Der sinnhafte Aufbau der sozialen Welt*. Springer, Wien [1972 *The Phenomenology of the Social World*. Heinemann, London]
- Samuelson P 1954 The pure theory of public expenditure. *Review of Economics and Statistics* 36: 387–9
- Smith A 1776/1970 *The Wealth of Nations*. Penguin, Harmondsworth; Whitestone, Dublin
- Trevor-Roper H 1976 *Princes and Artists*. Thames and Hudson, London
- Ultee W C 1980 *Fortschritt und Stagnation in der Soziologie*. Luchterhand, Darmstadt, Germany
- Ultee W 1997 Do rational choice approaches have problems? *European Sociological Review* 12: 167–79
- Ultee W 1997 Bringing individuals back into sociology. In: Blossfeld H P, Prein G (eds.) *Rational Choice Theory and Large-scale Data Analysis*. Westview Press, Boulder, CO
- Van den Doel H 1979 *Democracy and Welfare Economics*. Cambridge University Press, Cambridge
- Weber M 1922–3 *Gesammelte Aufsätze zur Religionssoziologie*. Mohr, Tübingen [1930 *The Protestant Ethic and the Spirit of Capitalism*. Allen and Unwin, London]
- Winckelmann J (ed.) 1978 *Max Weber. Die protestantische Ethik II. Kritiken und Antikritiken*. Gütersloh, Mohn, Germany
- Ziman J M 1987 The problem of 'Problem Choice'. *Minerva* 25: 92–106
- Zuckerman H 1978 Theory choice and problem choice in science. In: Gaston J (ed.) *Sociology of Science*, 1st edn. Jossey-Bass, San Francisco, CA

W. C. Ultee

Problem Solving and Reasoning: Case-based

Case-based reasoning (CBR) is a paradigm of *artificial intelligence* and *cognitive science* that models the reasoning process as primarily memory based. Case-based reasoners solve new problems by retrieving stored 'cases' describing similar prior *problem-solving* episodes and adapting their solutions to fit new needs.

As case-based reasoners process each problem, they learn by storing records of their problem solving as new cases to guide future reasoning. CBR research studies the CBR process both as a model of human cognition and as an approach to building intelligent systems. Principles from CBR research serve as a foundation for applied computer systems for tasks such as supporting human decision making, aiding human learning, and facilitating access to electronic information repositories.

1. Case-based Reasoning as Cognitive Model

Initial research on CBR was motivated by observations of how humans remember, reason, and learn from experiences. People who encounter surprising events are often reminded of similar past events and use these reminders to help them understand and respond to the new situations (Schank 1982). For example, a doctor whose patient has an unusual reaction to a medicine might be reminded of a previous patient who had similar reaction. Remembering the treatment used for the prior patient suggests a candidate treatment for the new patient; remembering the outcome of the previous treatment helps to predict the new outcome and warns of potential pitfalls to avoid.

Psychological studies provide evidence supporting human use of CBR for a wide range of problem-solving tasks such as learning programming, performing mathematical problem solving, explaining anomalous events, and decision-making (see Leake 1998) for examples of this work). Another use of CBR is for classification or *interpretation*. For example, in *legal reasoning* in the American legal system, prior cases are the basis for legal judgments: lawyers build arguments about how to interpret rules of law by comparing and contrasting precedent cases to the current situation (see Ashley and Rissland (1988) and *Legal Reasoning Models*).

Research on CBR as a cognitive model studies the knowledge and processes that underlie human CBR, addressing topics such as *memory organization*, *similarity assessment*, and *analogical mapping* (Kolodner 1993, Leake 1998). Computational models developed in this work illuminate functional constraints on reasoning processes and provide testable hypotheses about human reasoning for psychological investigation.

2. Steps in Case-based Problem-Solving

Computer CBR systems have been developed for numerous tasks including planning (Hammond 1989, Veloso 1994), design (Pu and Maher 1997), and explanation (Schank et al. 1994). Regardless of the specific reasoning task, any CBR system retrieves relevant prior cases, adapts their lessons to fit the new situation, evaluates and applies the result, and retains

the results of processing as new cases to guide future reasoning. Thus, computer models of CBR must account for four main processing steps: *retrieval*, *adaptation*, *evaluation/application*, and *storage*.

2.1 Retrieval

In order to facilitate retrieval of relevant cases, stored cases are organized by *indices* characterizing circumstances in which the cases are likely to be relevant. For example, in a medical CBR system to suggest treatments for illnesses, treatment cases could be indexed by the diseases they treat and the characteristics of patients (e.g., age) that tend to determine whether a given treatment is appropriate. Given a new problem to solve, a case-based reasoner first performs *situation assessment* to analyze the input problem and describe it in terms of the indices used to organize cases. Retrieval algorithms use the description to retrieve the cases in memory expected to be most relevant. For example, nearest-neighbor retrieval algorithms select the most similar prior cases according to a metric that weights features by their relative importances. Alternatively, case libraries may be organized into discrimination networks to enable retrievals based on only a small number of feature comparisons. After candidate cases have been retrieved, a finer-grained similarity assessment process is often used to identify the best candidates and to determine differences requiring adaptation

2.2 Adaptation

The case adaptation step revises the lessons of the retrieved case (or revises and combines the lessons of multiple retrieved cases) to generate an appropriate new solution. For example, a doctor might have to adopt a prior treatment by reducing the dosage if the prior patient was an adult and the current patient is a child, or might combine elements of prior treatments to address multiple complaints. Computer CBR systems often use rule-based reasoning to determine how to transform prior solutions. In interpretive CBR, the case-based reasoner does not modify the retrieved case, but instead develops justifications for why the prior interpretation applies (or does not apply) to the new situation.

2.3 Evaluation/Application

Tentative results of case adaptation are evaluated to assess the appropriateness of the new solution and to determine whether additional changes are needed, possibly resulting in a cycle of repeated adaptation and evaluation. Once a solution is judged acceptable, it is applied to the real problem. If the solution fails, it may be repaired, reevaluated, and reapplied.

2.4 Storage

Each processing episode of a CBR system, whether resulting in success or failure, is stored to guide future reasoning. Cases containing successful solutions suggest solutions for similar future problems, while cases storing failures provide warnings of problems to avoid and information on how to avoid them. The storage step packages information about the problem, solution and its results in a new case. It also assigns indices for the case and places it in the case library (also called the 'case base') for future use.

3. Research Issues

3.1 Retrieval and Similarity Assessment

Effective case retrieval depends on appropriate retrieval algorithms, well-organized case bases, and indices that are useful for the current task. In case-based problem solving, cases are indexed by information about the problems they solve. For example, a case-based planning system may index plan cases by the goals they achieve and constraints they satisfy. Indices should be predictive of the relevance of cases and should involve features of problems that are likely to be known to the reasoner; both abstract and concrete features may be useful. A crucial research issue in CBR is developing appropriate *indexing vocabularies* for particular reasoning tasks. In order to retrieve useful cases for novel situations, methods are also needed for reformulating or transforming indices to widen search when no directly relevant cases are available. There is also considerable interest in replacing traditional measures of similarity with similarity criteria that directly reflect pragmatic needs by predicting which prior cases will be easiest to adapt to fit the new situation (e.g., Smyth and Keane 1998). Research on retrieval algorithms investigates methods for achieving rapid retrievals from large case bases (e.g., by exploiting parallel algorithms).

3.2 Case Adaptation

Developing case adaptation rules is often difficult because of knowledge acquisition problems: the needed knowledge may not be available or may be difficult to codify. Consequently, improving case adaptation methods is a fundamental challenge for CBR. Research areas include applying machine-learning techniques to automatically generate case adaptation rules; applying CBR to a CBR system's internal processing, to adapt cases based on prior adaptation experiences rather than rules; using a *derivational analogy* process to account for new circumstances by replaying a prior solution-generation process, rather than revising a prior solution (Veloso 1994); and applying artificial intelligence methods

such as constraint-based reasoning to guide the adaptation process. It is important to note that sufficiently powerful case adaptation can enable cases to be used in very novel ways; research on flexible retrieval and adaptation processes has given rise to CBR models of creative problem solving.

3.3 Case-base Maintenance

As CBR systems receive long-term use, keeping the case base accurate becomes an important problem. Likewise, as cases are accumulated, it may be necessary to reorganize or update the case base in order to improve its coverage of important problems and to avoid slow retrievals. This has led to considerable interest in methods for improving case quality and retrieval efficiency by strategic deletion, revision, and addition of cases to a case base. Because cases are increasingly generated by multiple distributed sources (e.g., different offices of an organization), additional issues concern how to update and access distributed case information.

3.4 Integration with Other Methods

CBR systems may be integrated with other systems and approaches—both artificial intelligence systems and other systems—to leverage off the strengths of each. For example, case-based and rule-based reasoning may be integrated with each other, with cases used to guide interpretation of rules or to suggest targets for rule-based reasoning. CBR may also be used for speed-up learning, with initial rule-based problem solving building up a case library that can then be used instead of rules, bypassing costly reasoning from scratch, or to avoid errors by storing exceptions to rules. Conversely, rule-based reasoning may be useful to augment the case library and extend the range of problems that a CBR system can solve.

4. CBR Technology

Applications of CBR include both autonomous and interactive systems. As observed by Kolodner (1993), human users are often adept at evaluating the relevance of prior cases and applying their lessons, but may lack the experience needed to acquire a sufficient set of cases. On the other hand, case adaptation and evaluation are difficult for CBR systems, but practical methods exist for accessing large case libraries. These complementary strengths have led to increasing focus on case-based ‘retrieve and propose’ systems that support human reasoners by providing them with relevant prior cases to aid their decision making (cf. *Decision Support Systems*). For example, case-based ‘help desk’ systems are used widely to aid telephone

support personnel who diagnose customer problems, and commercial CBR ‘shells’, are available to facilitate the development of such applications (Watson 1997). As novel problems and solutions are added to the case base during use, the system builds a sharable *corporate memory*, of cases, aiding corporate *knowledge management*.

CBR is also increasingly applied to *electronic commerce*. For example, case-based support systems are used during on-line purchases of computers, to help buyers configure their systems by presenting similar past configurations as a starting point for them to modify. This helps users find a good match to their needs, even if they do not know all the important features of the product they are ordering.

In addition, the CBR cognitive model and CBR principles are guiding the design of educational systems by, for example, suggesting when useful learning is most likely to occur and how to retrieve examples to support that learning. Likewise, CBR is being applied to accessing hypermedia and on-line repositories of information—providing indexing vocabularies and example-based query interfaces—and is being used to focus information retrieval, in order to extract task-relevant information from large-scale information sources. Another active research area, textual CBR, develops methods for facilitating knowledge acquisition by automatically capturing and indexing cases in textual form. Leake (1996) and Lenz et al. (1998) present projects illustrating these applications areas.

5. Relationship to Other Artificial Intelligence Methods

Much artificial intelligence research focuses on methods for generating new solutions from scratch by chaining together general rules (for example, see *Expert Systems in Medicine*). CBR is an interesting alternative for a number of reasons, including: (a) simplifying knowledge acquisition (experts often find it difficult to provide rules for their actions but can easily provide examples, and CBR systems can be fielded with small case libraries and learn additional cases during use); (b) providing natural explanations of decisions (by accounting for decisions in terms of actual previous examples, rather than in terms of rules that users may not understand or trust); (c) increasing problem-solving efficiency by reusing the results of prior reasoning; and (d) facilitating system development in domains that are hard to codify in terms of general rules (e.g., because the rules would have too many exceptions). In CBR, traditional rule acquisition issues are replaced by issues of capturing and representing reasoning episodes, selecting similarity and indexing criteria, and capturing case adaptation knowledge.

The learning done by CBR systems strikes a middle ground between knowledge-poor and knowledge-rich

learning approaches. Unlike knowledge-poor inductive learning methods, CBR exploits domain knowledge (e.g., many CBR systems use explanations to guide indexing and adaptation). Unlike knowledge-rich methods, such as explanation-based generalization, CBR can still be in the absence of a rule-based domain theory.

In its emphasis on building up knowledge by learning, CBR research is in the same spirit as research on neural networks (see *Connectionist Approaches*). However, CBR contrasts in storing and reasoning from examples rather than storing only learned generalizations. Also, CBR performs 'lazy learning.' Rather than immediately generalizing training instances and discarding the examples, as in, for example, explanation-based and neural network learning, CBR stores cases with minimal processing and processes them further only if and when adaptation is needed to solve new problems. In addition, unlike neural networks, case-based reasoning systems can change behavior immediately in response to new examples: a single correct example of dealing with a situation, and no further training, is all that is needed for the system to process that situation correctly in the future.

6. Conclusion

A case-based problem solver solves new problems by retrieving stored problem cases and adapting their lessons to fit new circumstances; it learns by storing new episodes for future use. This process can provide robust reasoning and learning even in domains that are hard to codify in rule-based form. CBR research has provided computational cognitive models of human information processing, artificial intelligence techniques, and information technology for a wide range of task domains.

The CBR cognitive model is discussed in Kolodner (1993) and Leake (1998). A volume edited by Leake (1996) surveys the field with chapters on issues and applications in tasks such as decision aiding, knowledge navigation, design, diagnosis, and education; Kolodner (1993) provides an in-depth examination of key issues; and a volume edited by Riesbeck and Schank (1989) includes chapters on important early CBR research projects and 'micro-versions', of the computer programs those projects developed. The process of developing CBR applications is described by Watson (1997). A volume edited by Lenz et al. (1998) presents chapters on CBR technology for tasks including supporting electronic commerce and textual CBR.

See also: Case Study; Logic; Problem Solving and Reasoning; Case-based; Problem Solving and Reasoning, Psychology of; Problem Solving: Deduction, Induction, and Analogical Reasoning

Bibliography

- Ashley K D, Rissland E 1988 A case-based approach to modelling legal expertise. *IEEE Expert* 3(3): 70–7
- Hammond, K J 1989 *Case-based Planning: Viewing Planning as a Memory Task*. Academic Press, Boston, MA
- Kolodner, J 1993 *Case-based Reasoning*. Kaufmann, San Mateo, CA
- Leake D B (ed.) 1996 *Case-based Reasoning: Experiences, Lessons, and Future Directions*. AAAI Press, Menlo Park, CA
- Leake D 1998 *Cognition as Case-based Reasoning*. In: Bechtel W, Graham G (eds.) *A Companion to Cognitive Science*. Blackwell, Oxford, UK, pp. 465–76
- Lenz M, Bartsch-Spörl B, Burkhard H D, Wess S (eds.) 1998 *Case-based Reasoning Technology: From Foundations to Applications*. Springer, Berlin
- Pu P, Maher M L (eds.) 1997 *Issues and Applications of Case-based Reasoning in Design*. Erlbaum, Mahwah, NJ
- Riesbeck C K, Schank R C 1989 *Inside Case-based Reasoning*. Erlbaum, Hillsdale, NJ
- Schank, R C 1982 *Dynamic Memory: A Theory of Reminding and Learning in Computers and People*. Cambridge University Press, Cambridge, UK
- Schank R C, Riesbeck C K, Kass A (eds.) 1994 *Inside Case-based Explanation*. Erlbaum, Hillsdale, NJ
- Smyth, B, Keane, M T 1998 Adaptation-guided retrieval: Questioning the similarity assumption in reasoning. *Artificial Intelligence* 102(2): 249–93
- Veloso M M 1994 *Planning and Learning by Analogical Reasoning*. Springer, Berlin
- Watson, I 1997 *Applying Case-based Reasoning: Techniques for Enterprise Systems*. Kaufmann, San Francisco, CA

D. B. Leake

Problem Solving and Reasoning, Psychology of

This article discusses two important aspects of human thinking: how people solve problems and how they use systematic thought to reach conclusions from premises. The two processes are intermingled closely in most everyday human thinking.

Woodworth (1938) wrote:

Two chapters (of 30!) will not be too many for the large topic of thinking.

Woodworth here understates somewhat the importance he assigned to 'higher mental processes,' for another chapter examines reading, and nine others, memory, and learning. In the second half of the twentieth century, the small role of high-level cognition in experimental psychology has been enlarged greatly by treating thinking as symbolic information processing, formulating theories of thinking as non-numerical computer programs, and introducing new observational methods and instruments to capture

increased temporal detail in thought processes (Newell and Simon 1972).

Woodworth (1938) divided his final two chapters ‘according to the historical sources of two streams of experimentation ... One stream arose in the study of animal behavior and went on to human problem solving; the other started with human thinking of the more verbal sort.’ Two branches of the first stream investigated (a) search as the basic problem solving process (trial-and-error, then selective, search; e.g., Morgan 1894, Thorndike 1898), and (b) insight (e.g., Köhler 1917). Little attention was given to expert problem solving.

The second stream of research, on verbal thinking, included concept attainment (e.g., Hull 1920), logical deduction (‘reasoning’), creativity (mostly anecdotal reports on inventors, scientists, and artists), and verbal problem solving, the latter initiated by Selz (1913, 1922) and further developed by Duncker (1935) and De Groot (1946). All of these lines of work were important harbingers of the new information processing psychology, but the connections among them only became apparent later. This article describes today’s knowledge of thought processes within the information processing framework down to, but not including, the neural implementation.

1. Problem Solving

Modern theories of problem solving have gone through three stages. Most early research viewed problem solving as a search process. In the 1960s, attention began to turn to expert problem solving and its dependence on recognition that evoked knowledge from memory. A decade later, it had become evident that solving complex problems usually also required interacting search in several distinct problem spaces.

1.1 Problem Solving as Search

Problem solving is often described as search: moving in a problem space from an initial situation toward a goal situation that satisfies all the conditions for a solution. Problem difficulty was initially thought to vary with the size of the problem space. A space with b choices at each point in the search, and paths s steps in length, would contain b^s different paths.

Choosing a chess move was very difficult, for from a typical current position 10^{20} or more branching paths might lead to stronger or weaker positions, and ultimately to won, lost, or drawn games. As a human being could search only an insignificant fraction of these paths, skill depended on discovering just those that might lead to favorable outcomes. DeGroot (1946), showed that stronger players did not search much more extensively than weaker players, and that even grandmasters rarely looked at more than 100 positions before making a move.

These findings led to a theory of problem solving postulating: (a) an ‘understanding’ process that generates a problem space from the problem description, (b) a selective (heuristic) search process that uses rules of thumb to follow the most promising paths, and (c) a position evaluation function to assess the relative merits of the positions reached. The process can be repeated, re-evaluating the situation after each step. Several chess programs that deliberately limit themselves to searches of the magnitudes that humans can make demonstrate that chess of credible quality can be played in this way, confirming and extending experimental findings about human players (Simon and Schaeffer 1992).

The same scheme applies to problems generally, not just to chess. Methods for generating paths include: generate-and-test (i.e., trial-and-error), hill-climbing (moving toward higher values of an evaluation function), and means-ends analysis. At each step in means-ends analysis: (a) one or more differences are discovered between the current situation and the goal conditions; (b) using past knowledge and experience, an operator is applied that usually eliminates or reduces differences of the kind discovered; (c) the process is repeated until all differences have been removed and the goal attained.

Suppose the goal is to solve an equation for x : $7x + 10 = 4x + 22$. The operators add or subtract the same expression from both sides of the equation or multiply or divide both by the same expression. The differences between the given expression and the solution ($x = N$) are the presence of a stand-alone number on the left side of the equation, an expression in x on the right, or a coefficient other than 1 for x on the left. In the equation above, means-ends analysis would discover an unwanted number, 10, on the left, then an expression in x , $4x$, on the right, and would subtract both from the equation, leaving $3x = 12$. Now it would notice x ’s coefficient, 3, and would divide, obtaining $x = 4$.

1.2 Knowledge for Solving Problems

Some problems are hard even though their problem spaces are small. For example, the Chinese ring puzzle, which requires freeing a metal bar from a series of rings, has an unbranching problem space, so that the solution can be reached simply by never reversing direction. Nevertheless, this problem, with seven rings, is seldom solved in less than half an hour. It has been found that people apply ‘reasonable’ heuristics, e.g., never reconnect a ring that has once been removed. But adherence to this plausible means-ends principle guarantees that the problem will not be solved. Gradually, recognition that rings sometimes have to be reconnected uncovers the solution strategy.

Comparing the problem-solving by experts and novices soon demonstrated that domain-specific knowledge (of which this is a simple example) lay at

the basis of expert skills, connecting problem solving with memory. A large discrimination net in the expert's memory permits sensory recognition of common stimulus patterns in the domain. Thus, a chess grand-master can recognize perhaps 300,000 configurations of pieces that occur repeatedly in games (e.g., 'fork,' 'open file,' 'weak bishop,' etc.). Associated in memory with each such pattern is knowledge of actions that should be explored whenever the pattern is present. Most 'leaps' in thought called intuitive, insightful, or creative are produced by this knowledge-based pattern recognition (Richman et al. 1996).

The expert, then, recognizes familiar patterns, evoking knowledge that guides search to strong moves in a limited time, searching only a little more than the novice, but along productive instead of unproductive paths.

1.3 Multiple Problem Spaces

As empirical studies advanced, it became apparent that problem solving search employs many different spaces, including a space of potential problem spaces. Searching for solutions may also be preceded by searching for effective generators of actions. Problems of scientific discovery require search in spaces of potentially useful instruments, experimental designs, and hypotheses. The Soar program developed by Newell, with Laird and Rosenbloom (Newell 1990), searches in multiple spaces of these kinds.

2. Reasoning

Reasoning encompasses thinking that is orderly or 'logical': for example, the processes of formal logic and of concept attainment (see *Problem Solving: Deduction, Induction, and Analogical Reasoning; Problem Solving and Reasoning: Case-based*). Hence, most problem solving processes can also be called 'reasoning.' In logic, reasoning begins with premises (a situation), and uses rules of inference (actions) to derive conclusions (attain a goal). It may be inductive or deductive.

2.1 Forms of Reasoning

In inductive logic, the premises usually are empirical assertions and the rules of inference produce generalizations of these using terms like 'all,' 'most,' 'probable.' For example, from observation of many white swans, and none of other colors, one may reason inductively to: 'All swans are white,' 'Probably all swans are white,' 'The probability that any swan is white is 99 in 100,' or 'Most swans are white.' At one time or another, all of these sorts of induction have been proposed.

In deductive logic, the initial premises are taken as axioms and the rules of inference are tautologies (e.g., If all A are B and c is an A, then c is a B), so that, in any possible world in which the premises hold, the conclusions must hold.

Reasoning may be used to discover either a concept, an inductive conclusion, or a deductive conclusion; it may also be used to verify such a concept or conclusion, once reached. Discovery is a wholly different process from verification.

Verifying a proof is a deductive process: simply examining each step to assure that it was obtained from previous steps by an application of the rules of inference. It requires no real search. However, discovering a proof is an inductive problem solving process, which involves searching for a path from the premises to the conclusion, using the rules of inference. (The task of the first computer that simulated problem solving, the *Logic Theorist*, was to find, inductively, proofs for theorems in symbolic logic (Newell and Simon 1956)).

In concept attainment, the task is to find, by examining instances, a rule that distinguishes instances from noninstances of the concept: swans from non-swans. This is an inductive task, where the search proceeds in two spaces: a space of possible instances and noninstances and a space of rules (e.g., feature descriptions that can distinguish instances from noninstances). Search in the space of rules generates possible concepts; search in the space of instances generates objects for testing the rules.

2.2 Everyday Reasoning

Much remains to be learned about how human beings reason in their everyday affairs. One attractive hypothesis, only partially substantiated by the evidence, is that they use the same processes that they use for structured problems. Each situation a person encounters evokes from memory a constantly changing array of knowledge, beliefs, and goals, with varying degrees of mutual consistency and inconsistency. Current goals and the problem spaces they evoke trigger search that leads to actions and thence again to changes in the situations and the materials evoked from memory.

The difficulty of understanding such a system largely derives from the size and complexity of memory, with its idiosyncrasies of what it does or doesn't recognize at any moment and does or doesn't know, believe, like, or dislike about the things recognized. The closest approaches to theories of such systems have been simulations (unified theories) of sizable subsets of the cognitive processes (see Richman et al. 1996).

2.3 Reasoning Mechanisms

Sometimes it has been proposed that specific kinds of reasoning employ specialized machinery in the brain,

just as sensory and motor processes have their specialized organs. Chomsky (1976), for example, has claimed that there is a special language facility (not merely language ‘regions,’ but biological structures specialized to processing language). Rips (1994) has made claims for an inherited machinery for logical reasoning. Comparable claims have been made for ‘pictorial’ reasoning, using mental imagery.

The alternative hypothesis does not deny the evidences for brain localization and specialized functions but claims that all these faculties rest upon the same basic associative neural processes and organizations.

The latter alternative has been explored by constructing simulations that are capable of performing these functions without specialized mechanisms. Thus, without any specialization for logic, syllogistic reasoning will be carried out by all systems that employ if-then rules (production systems), because whenever all the conditions (‘ifs’) of a rule are satisfied in memory, the actions (‘then’) will be executed—a generalization of the classical stimulus-response mechanism. Production systems, conjoined with associative memories, provide a completely general architecture for symbolic processing, ‘logical’ or not (Eisenstadt and Simon 1997). They can reason about mental diagrams and can combine verbal with visual reasoning.

Concept attainment, as well as various kinds of verbal learning and expert memory performance, can be achieved by a discrimination net which grows, by exposure to stimuli and feedback of results, a network of tests of stimulus features that classify the stimuli. Linked to the discrimination net, and indexed by it, a large associative (semantic) memory holds information about each of its concepts, permitting the kind of problem solving by recognition that was discussed in Sect. 1.2 (Richman et al. 1996).

The evidence available at the turn of the twenty-first century allows no clear choice between specialized mechanisms for special domains of reasoning vs. regional localizations of functions that all employ production systems capable of reasoning and learning (i.e., acquiring new productions) and discrimination nets for recognition, and share an associative semantic memory of knowledge and beliefs, indexed by the net. Whichever turns out to be the actual architecture, there will remain the important task of finding the neural mechanisms that implement it.

See also: Cognitive Theory: ACT; Cognitive Theory: SOAR; Deductive Reasoning Systems; Problem Solving and Reasoning: Case-based; Problem Solving: Deduction, Induction, and Analogical Reasoning; Scientific Discovery, Computational Models of

Bibliography

Chomsky N 1976 *Reflections on Language*. Temple Smith, London

- De Groot A D 1946 *Het Denken van den Shaker*. North Holland, Amsterdam [trans. 1965 as *Thought and Choice in Chess*. Mouton, The Hague]
- Duncker K 1935 *Zur Psychologie des productiven Denkens*. Springer, Berlin [trans. 1945 as *On Problem-solving*. Psychological Monographs, Vol. 38, no. 5. American Psychological Association, Washington, DC]
- Eisenstadt S A, Simon H A 1997 Logic and thought. *Minds and Machines* 7: 365–85
- Ericsson K A, Simon H A 1993 *Protocol Analysis: Verbal Reports as Data*, rev. edn., MIT Press, Cambridge, MA
- Hull C L 1920 Quantitative aspects of the evolution of concepts: an experimental study. *Psychological Monographs* 28 (whole No. 123)
- Köhler W 1917 *Intelligenzprüfungen an Menschenaffen*. Springer, Berlin
- Morgan C L 1894 *An Introduction to Comparative Psychology*, 1st edn. Scribners, New York
- Newell A 1990 *Unified Theories of Cognition*. Harvard University Press, Cambridge, MA
- Newell A, Simon H A 1956 The logic theory machine. *Transactions on Information Theory*, IT-2(3), 61–79
- Newell A, Simon H A 1972 *Human Problem Solving*. Prentice-Hall, Englewood Cliffs, NJ
- Richman H B, Gobet F, Staszewski J J, Simon H A 1996 Perceptual and memory processes in the acquisition of expert performance: the EPAM model. In: Ericsson K A (ed.) *The Road to Excellence*. Lawrence Erlbaum Associates, Mahwah, NJ
- Rips L J 1994 *The Psychology of Proof*. MIT Press, Cambridge, MA
- Selz O 1913 *Über die Gestze des geordneten Denkverlaufs*. Spemann, Stuttgart
- Selz O 1922 *Zur Psychologie des productiven Denkens und des Irrtums*. Friedrich Cohen, Bonn
- Simon H A, Schaeffer J 1992 The game of chess. In: Aumann R J, Hart S (eds.) *Handbook of Game Theory*. Elsevier, The Netherlands, Vol. 1, pp. 2–17
- Thorndike E L 1898 Animal intelligence: an experimental study of the associative processes in animals. *Psychological Review*, Monograph Supplements, 2 (serial No. 8)
- Woodworth R S 1938 *Experimental Psychology*. Holt, New York

H. A. Simon

Problem Solving: Deduction, Induction, and Analogical Reasoning

1. The Closing of Gaps in Human Problem Solving

Deduction, induction, and analogical reasoning, the three basic forms of logical thought, were derived from natural or everyday situations, where they were originally embedded in commonplace forms of interaction and communication. The differences are explained in the respective paragraphs.

Depending on the context, the German word ‘*schließen*’ (‘conclusion’) may mean to ‘close,’ ‘conclude,’ ‘reason,’ or ‘infer.’ Generally speaking, the term implies the bridging or closing of a gap, be this a *literal* gap between two entities or, in the metaphorical sense, a *logical* gap between a problem and its solution. The closing of such metaphorical gaps by means of reasoning or inference usually entails a series of cognitive ‘steps.’

Problems can be defined in terms of three characteristic stages: (a) an initial state (e.g., the arrangement of the figures on a chess board, an algebra or geometry exercise, or a balance of trade), (b) an end state (e.g., putting the opponent’s king in checkmate, solving the maths exercise, or finding ways of improving the balance of trade), and (c) a well-defined series of cognitive ‘steps’ by which the initial state is transformed to the end state: the solution to the problem (or one of a number of solutions).

They are still to be found in their original, undifferentiated form as IF-THEN relations: in barter, trade, the manufacture of implements, and the planning of collective action, for example. Once raised to the metalevel of pure thought, these forms of reasoning became a subject of philosophical inquiry in the fields of mathematical logic, linguistics, and psychology. Over the course of history they even became independent areas of scientific systematics. It was only at a very late stage, in the time of the great Greek philosophers, that the three types were differentiated, defined, and each traced back to its logical structure. This represented the first step on the path to European or Western thought and the scientific advances that were entailed. And the first to tread this path were the great trio of Ancient Greek thought—Socrates, Plato, and Aristotle.

2. Three Great Greek Philosophers and the Foundations of Logic

In the early Socratic *Dialogues*—preserved in Plato’s writings—the processes of reasoning are dominated by IF-THEN relations. These and other dialogues demonstrate the inevitability of human thought. When we are told that all men are mortal and that Socrates is a man, our cognitive processes preclude us from harboring any doubts that Socrates is mortal. We will return to this aspect. First, it should be noted that this kind of logic is not rooted in words. It was one of Plato’s fundamental theories that the objects perceived by the senses are but imperfect copies of Ideas, the mental representations now known as concepts. Socrates had already used inductive arguments to formulate definitions of concepts by gradually identifying their characteristic properties or features. Regardless of the countless possible variations in the shape perceived as a triangle, for example, the sum of the angles must always be equal to two right angles,

and half the product of baseline and height must always be equal to the area. This is Plato’s concept—or Idea—of triangularity.

By applying the principle of dieresis—the breaking down of a concept into ever more narrowly defined subconcepts—Plato developed a general taxonomy of concepts and a system of concept-subconcept relations. This laid the foundation for the Aristotelian system of classification: first the broad general category (*genus proximum*), then the more specific (*differentia specifica*)—for example, a tree is a plant with branches and leaves or needles.

In his treatise *On Interpretation*, Aristotle stripped away the semantic content of propositions, revealing the comparable structures behind the words. ‘First we must define the terms “noun” and “verb,” then the terms “denial” and “affirmation,” then “proposition” and “sentence.”’ Spoken words are the symbols of mental experience and written words are the symbols of spoken words.’ It was another 2000 years before Wittgenstein again addressed this issue in such clear terms. Aristotle’s approach enabled him to formulate the figures of classical logic, the syllogisms. His theory of syllogistic reasoning drew on Plato’s hierarchical ordering of concepts, and a syllogism was defined as ‘discourse in which, certain things being stated, something other than what is stated follows of necessity from their being so.’ A well-known example is the syllogism: A is true of all Bs. C is an A. Therefore, B is true of C. ‘All humans are mortal. Socrates is a human. Therefore, Socrates is mortal.’ Negation operates on similar lines: ‘No self-luminous bodies are planets. All fixed stars are self-luminous. Therefore, no fixed stars are planets.’ These syllogisms were central to philosophical and logical inquiry throughout the Middle Ages and on into modern times.

Although examples with natural terms were mainly used to analyse and dispute the universality of the logical structures, it was not until 1679–90 that a new level of knowledge was attained in the field of logic. Scholars began to ask which inferences can be made when the premises of an argument are derived from prior experience, and do not allow universally valid conclusions to be drawn. This question was first formulated in clear terms by Gottfried Wilhelm Leibniz. Leibniz envisaged a universal formalized language, a ‘*Mathesis universalis*,’ for the general analysis and description of human knowledge, a kind of algebraic representation of logical reasoning in human thought. The system was to encompass even those cases in which the given circumstances and available data did not allow definitive conclusions to be drawn. Leibniz suggested calculating the degree of probability that a logical conclusion would be reliable. Bolzano (1970) followed this line of thought, postulating that the prediction of a conclusion is governed by the conditions of occurrence in a sequence of events. This approaches the twentieth century definition of probability as the likelihood of an event. These are

examples from an era in which the ground was prepared for a new method for acquiring the precise definition of inductive reasoning.

3. Drawing Conclusions from Insufficient Information: Induction

Inductive reasoning occurs when a conclusion does not follow necessarily from the available information. As such, the truth of the conclusion cannot be guaranteed. Rather, a particular outcome is inferred from data about an observed sample B' , $\{B'\} \subset \{B\}$, where $\{B\}$ means the entire population. That is, on the basis of observations about the sample B' , a prediction is made about the entire population B . For example: Smoking increases the risk of cancer. Mr X smokes. How probable is it that Mr X will develop cancer? As this example illustrates, inferential statistics (t-tests, analyses of variance, and other derived forms) are based on this kind of inductive reasoning.

The problems associated with the use of induction in scientific reasoning have been addressed from both philosophical and the mathematical perspective. From the philosophical point of view, the question arises as to what this process—the drawing of universal conclusions about a given phenomenon on the basis of the probability in a sample—actually entails. Logicians, as we now discuss, have responded in a variety of ways (v. Mises, Reichenbach, Keynes, Jeffrey, and Carnap).

Carnap's influential approach to this issue represents a form of compromise. Carnap calculates probability on the basis of accumulated experiential data, where confirmations of one hypothesis as opposed to another are entered into the equation as weights, basically resulting in a kind of weighted mean.

4. Psychological Assumptions from Bayes' Theorem

In psychological research (artificial or experimental) many attempts have been made to analyse beliefs about probability on the basis of examples. For example, marbles are randomly selected from a jar containing an unknown number of red and black marbles, and study participants asked to predict the next color to be drawn. Participants' predictions are sometimes based on *a priori* probabilities, i.e., unconditional, fixed relative frequencies. In central Europe, for example, the probability that it will rain in July is very high; the probability that it will snow is very low. These predictions would have to be modified for the month of April, and may even have to be reversed for January. Everyday human decision-making is rarely determined by such absolute probabilities however. This can be illustrated by the example of a horse-racing enthusiast's gambling be-

havior. Let us assume that the gambler is familiar with the horses taking part in the race and decides to bet that the favorite, horse E, will win. He goes to the tote and sees that horse E is injured and will not be taking part in the race. Our gambler immediately modifies his predictions, selecting the second favorite from the list of starters and placing a bet on this horse instead. This is not necessarily the horse he had previously predicted would come second, however. He might even decide not to bet on second place. Either way, his decisions are still based on estimates of probability, but as a result of the change in circumstances there are now fewer possible choices available. This points to the conditioned probabilities. The *hypothesis set H* for the decision H (horse E) has been reduced by the conditional probability $p(E|H)$. This can be seen in relation to the probability of the alternative hypothesis.

In calculating the so-called *a posteriori* probability, a comparison is drawn between these two quantities. $p(E|H)$ is weighted with the *a priori* probability:

$$p(E|H) \times p(H)$$

This is then placed in relation to the set of all possible positive or negative outcomes according to probability, resulting in the theorem named after Thomas Bayes:

$$p(H|E) = p(E|H)P(H)/P(E)$$

Many experiments in the field of psychology (see Gluck and Bower 1988) have been conducted to test whether the *a posteriori* probability is indeed as effective in human decision-making behavior as this calculation suggests. While some findings confirm the effectiveness of these components, some significant deviations from the theorem have also been registered. These deviations have provided important psychological insights, however. Edwards and Schmidt (1966), for example, conducted in-depth investigations into the ways people adapt their estimates and decision-making behaviour to the changed circumstances when the given information is modified—H-D Schmidt on the basis of high-risk hits of balls, Edwards using gaming chips and estimates of chance. Both found high levels of interindividual variation in the adaptability of study participants. While some people are flexible with respect to their predictions, many others—the so-called 'conservative estimators'—try to retain their initial predictions for as long as possible.

The findings reported by Kahnemann and Tversky (1973) are also very interesting in this context. They found that study participants largely ignore *a priori* probabilities in their predictions of specified variables, preferring to orientate themselves to conditional probabilities. This result is not counterintuitive, as the

conditional probabilities are the most salient. The so-called Monte Carlo effect is another example of the insensitivity to *a priori* probabilities. Suppose that marbles are drawn with replacement from an urn containing an equal number of red and black marbles. The probability of randomly drawing a marble of either colour is 50 percent. If a red marble has been drawn on a number of consecutive occasions, however, study participants regularly predict that it is more likely that a black marble will be selected next. They overlook the fact that each draw is an independent event, and that the *a priori* probability is independent of the previous draws, i.e., remains constant at 50 percent.

In a fictitious diagnosis experiment, Gluck and Bower (1988) found a high level of correspondence between the estimates made by study participants and the odds calculated using Bayes' formula. On the basis of a medical chart, study participants were asked to judge which of two diseases their (hypothetical) patients were suffering from. The *a priori* probability of one of the diseases was three times higher than that of the other. The conditional probabilities were then varied by the listing of symptoms relevant to the diagnosis. Gluck and Bower found a high degree of correspondence with the Bayes' calculations. When asked to estimate the frequency with which a symptom is associated with one of the two diseases (the rarer one), however, the study participants were unable to make reliable predictions. Hence, the 'diagnosticians' came to a Bayesian conclusion without using Bayesian methods. Gluck and Bower refer to this as implicit knowledge.

5. Implicit Knowledge in Reasoning

The question remains of how such a data compilation functions on the cognitive level. It can be assumed that such processes involve the compression of accessible information and that they are of a non-algebraic character. They constitute abstractions about data sets, leading to a kind of blurred conceptual knowledge. Examination of the effects of quantifiers in logical arguments has provided insights into these processes. The *modus ponens* is quite categorical: If P, then A holds. In other words, P implies ($\rightarrow$) A. Given P then A, hence $P \rightarrow A$. If it rains (P), then the lawn will be wet (A). It has rained, therefore the lawn is wet; $P \rightarrow A$. Is the reverse (not P) also true? That is: $\neg P \rightarrow \neg A$? Not necessarily; the lawn may well have been watered with a sprinkler or a hosepipe. It has been shown that, in more abstract contexts, even adults have difficulties in understanding this so-called *modus tollens* (Wason and Johnson-Laird 1972). But does the *modus ponens* also hold for conclusions beginning with the quantifier 'some'? For instance does it follow from 'some As are Bs, and some Bs are Cs,' that, 'some As are Cs'? Not necessarily. For instance, substituting 'women' for As, 'doctors' for Bs

and 'men' for Cs, we would arrive at the conclusion that some women are men. Alternatively, assuming that some Bs are As and some Cs are Bs, does it follow that some As are Cs? Again, this is not entirely true, as is immediately apparent when terms from natural language are entered into the equation (cf. Anderson). Wason and Johnson-Laird (1972) have investigated the acceptance of such logical arguments in detail. Cosmides and Tooby (cf. Horgan 1995) have attempted to provide an explanation, postulating that such halo effects on techniques acquired over the course of evolution help to reap the maximum benefits. For example, they suggest that the need to detect deception on the part of one's interactants is an essential incentive for the development of human intelligence. The well-known prisoner's dilemma is another such phenomenon. At issue here is whether players trust in the other player having a positive attitude to sharing, or whether they attempt to appropriate the entire winnings at the risk of losing everything. There is no doubt that the maximum-benefit motive plays a role in numerous decision-making processes. In general, however, the evolutionary influences on logically motivated decision-making behavior are located at a deeper level—in strategies which remain stable over the course of evolution and account for both individual and group interests. And not only in the sense of reaping the maximum social benefits. The so-called Monte Carlo effect (see earlier discussion) is but one example. Since time immemorial, humans have experienced that even prolonged natural processes such as weathering, drought, rain, floods, epidemics, disease, etc. eventually come to an end. And it generally holds that the longer these processes have been operating, the more imminent (i.e., the more probable) their conclusion. Such an approach appears to be the heritage of experiences gathered over the thousands of years of hominid history. The fact that humans prefer to extrapolate to the future—as implied by the *modus ponens* with its deterministic IF-THEN relations—is a favorable predisposition, increasing the chances of survival. It is very simple: It is more important to react to potential future events than to past occurrences. As there is only limited storage space in a modular nervous system, some bodies of knowledge become overgeneralized, perhaps as a result of being classified in line with the *modus ponens*. A similar pattern emerges for the weaker inductions of the *modus tollens*: In order to behave appropriately in one's environment, it is less important to know what can be inferred about the antecedent conditions from the absence of an event than vice versa.

6. Reasoning and Analogies

While deductive reasoning, in logic, refers to the *necessary* outcomes of a set of conditions, inductive reasoning is concerned with determining the likelihood of an outcome. Only *possible* conclusions can be

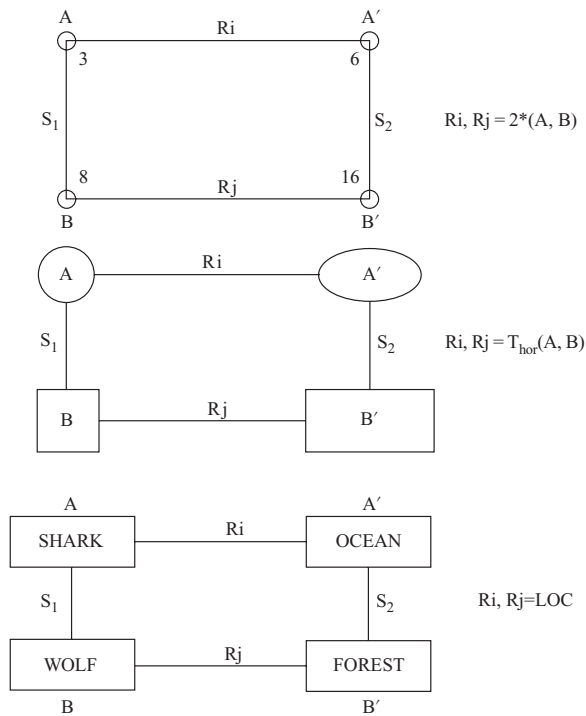


Figure 1
The basic structure of analogical reasoning. A–A': the source area with relation (s) R, between them. B–B': the target area with similar relations R_1 . S_1 and S_2 : mappings between source and target area

drawn, since not all of the conditions influencing or determining the outcome of a given situation are known. Analogical reasoning has a different character; it is basically a more precise form of conjecture. It is based on the mapping of structures or relations valid in one domain of knowledge onto another—less familiar or unknown—domain. The structures to be mapped are selected on the grounds of the similarity between the novel problem and the familiar one. The novel area is defined as the target domain of analogy construction, the familiar area as the source domain. With respect to the truth values of the conclusion, analogical reasoning spans all variants—from the absolute certainty of the *modus ponens* to a *posteriori* probabilities according to Bayes' Theorem, some of which are very weak. Suppose, for example, a second pair of terms is sought for a pair of numbers such as 3:6. Pairs such as 8:16 and 64:128 among many other combinations meet the requirements of the analogy.

As shown in Fig. 1, analogy construction can also be applied to geometric figures. It may be, but does not have to be the same factor of horizontal stretch which is mapped to produce a rectangle from a square or an oval from a circle. The boundaries within which

geometric analogies are accepted are fuzzy, but when these boundaries are overstepped, the analogies are rejected. The same holds for meaningful words in natural languages. Let us imagine that analogous pairs with the same relational correspondence are to be generated for the following pairs of words: (a) collie: dog, (b) walk: run, (c) oak: birch. A good number of satisfactory analogies can be found, for example: (a) rose: flower, (b) speak: shout, (c) carp: trout. In each pair, the first term has all the relevant properties of the second (animal, voice, plant), as well as some of its own. The first pair is characterized by a sub-super-concept relation, the second by the intensity of the characteristic feature (the semantic relation of the comparative). The third pair shares a super-concept with common properties, but each term has specific properties of its own (the semantic relation of coordination). Pairs of terms such as dress and bullet or flea and fog do not allow analogies to be drawn on the basis of common properties, as these are non-existent. Clearly, similarity between the terms is an integral element in the construction of analogies. This similarity is not necessarily rooted in common properties, however. For us pairs of terms such as (a) ship and ocean, (b) hunter and deer, and (c) surgeon and scalpel, for example, do not share common properties. Nevertheless, the relational connections between the concepts are comparable with those found in other pairs of terms. In example (a) the shared relational connection is the location (as in flower and garden), in (b) it is the object of the action (as in butcher and cow), and in (c) the instrument (as in farmer and scythe). More complex relations can also be involved, as in 'learn: know' paired with 'fight: win.' Here, the common semantic relation is the objective or purpose. Such relational connections are the basis for analogy construction: 'The shark is the wolf of the ocean,' for example, or 'May is the Mozart of the calendar' (Erich Kästner). Many such examples are to be found in the Bible, particularly in the parables of the New Testament.

7. The Origins and Original Function of Analogies

Analogies are evidently not isolated examples of mental games, but a constant cognitive law. But what is the point of these transfers between bodies of knowledge? Is it just a case of juggling with combinations of words that match in some way? Or do analogies reflect a deeper (perhaps even social) need of *Homo sapiens*? Viewed from this evolutionary perspective, the question arises of where, exactly, the origins of analogical thought are to be found.

Analogical thought is primarily to be found in language use. Indeed, the key word 'language' puts us on the right track, for the evolutionary origins of

analogy are rooted in the need to communicate, and similes and metaphors are its direct evolutionary source. The cognitive function of analogy is to facilitate discourse; when an interlocutor does not understand a particular term or statement, the speaker attempts to clarify its meaning. This he or she does by referring to common knowledge that is as similar as possible to the term or statement in question. Similes and metaphors are essential ingredients of this procedure. The comparative phrase 'is like' is often added as an explicit reference to the existing similarity (Rumelhart 1979). For example, it is quite possible to imagine the following explanation being given in the course of prehistoric dialog:

Question: 'What's a quagmire?'

Answer: 'It's like a trap in the mud.'

The second term functions as a source of new knowledge. Indeed, the very purpose of this dialog is to impart (or acquire) new knowledge. This is also the case when adults explain things to children: 'The heart is like a motor driving the body,' 'pathologists are like butchers,' or 'rules are like corsets,' for example. Complete analogies can be viewed as extensions of such metaphors—for example, 'Rock in the opera—that's like heavy metal in cathedral,' an executioner in the Salvation Army is like a monk in the artillery,' or 'a June meadow in the mountains is like a bride in her wedding dress.' The list is never-ending. Such examples are no longer explanations or supplements to existing knowledge, but this is the source from which they originated. And, as illustrated by Eichendorff reference, they display originality and creativity of thought: 'It was as if the Sky had silently kissed the Earth, so that she, in the glimmer of blossoms, now had to dream of him.' A somewhat more mundane example can be drawn from one of General Moltke's cadet examinations. When Moltke asked a cadet, 'How wide is the Thames in Paris?', the boy replied, 'Just as wide as the Seine in London, your Excellency.' By inverting the terms, the cadet demonstrated that he was well aware of the existing relations. Analogies can thus also be a source of smiles and approval.

8. Cognitive Procedures in Analogy Construction

Without a doubt, the recognition of structural similarities between the source and target domains plays a pivotal role in the acceptance of proposed analogies, as well as in their completion (e.g., supplying a fourth term when three are given), and creation (suggesting a new and possibly unique second pair of terms). Similarity is a necessary precondition in all three cases.

Different approaches to the definition of conceptual similarity have emerged in the literature definition in terms of (a) shared conceptual properties, as proposed primarily by Tversky *et al.* and (b) structural similarity, as suggested by Gentner and Stevens (1983). A compromise between these two approaches has been

formulated by Bachmann (1989) after Jaccard and Tanimoto.

Tversky defines the similarity between two objects in terms of the relationship between the features common to both structures and the features distinct to each, be they perceptual structures or sets of conceptual properties. As in Tversky's experiment, both components were then aggregated and compared to the estimated values. This resulted in a significant improvement on the correspondence between feature overlap and estimated similarity. The degree of improvement varied from concept, but correspondence calculated on the basis of common features alone was significantly lower.

The weighting on each of the semantic relations in similarity judgments is another matter of interest, and can be inferred from the sequence in which the individual relations are selected and entered into the analogous structure. Participants usually chose the actor and recipient first (e.g., hunt: hunter → game), followed by the instrument (e.g., rifle), the location (e.g., forest) and, often at a comparatively late stage, the objective (e.g., to bag a deer). This sequence of concept selection points to the existence of saliency weightings in the components of the individual concepts. Moreover, these results confirm that a moderate degree of similarity is essential not only for felicitous metaphors, but for creative analogies. Our experiments have shown that particularly characteristic features or relations can have preferential status in both the recognition and construction of analogies (Klix and Bachmann 1992). The originality of an analogy is dependent on factors such as the age and intelligence of its constructor, and analogy completion tasks are often employed in intelligence tests such as the Raven 'progressive matrices,' Sternberg (1977) has linked the factor structure of human intelligence to the solution of analogy tests. Moreover, Gentner and Sternberg have constructed algorithms which realize analogy detection. All of these are based on the comparison of features or semantic relations (Greeno and Riley 1984).

9. Models of Analogy Detection

Following the initial, very simple approaches to the processes of analogy detection, the dominant paradigm is now based on a model devised by Rips *et al.* (1973) which has been modified and extended by Bachmann (1998). According to this model, the following three aspects must be checked before an analogy can be accepted:

(a) Do the terms to be compared share a common feature which is much more salient than others? In train: railway and bus: truck; for example, the means of transport is so prominent that the analogy is immediately accepted.

(b) If this is not the case, the global impression is assessed. This entails a kind of compression of the available information, as discussed above in connection with Gluck and Bower's experiments.

(c) If this global comparison does not result in a decision, the concepts are closely scanned for common features and relations (deliberate checking). If no acceptable similarities are found, the analogy is rejected.

Unfortunately, the question of how much processing time is required for each of these components is as yet unanswered.

10. The Role of Analogy Construction in Scientific Innovation

The relationship between analogical reasoning and intelligence has been investigated in countless experiments and tests, with children and young adults of various ages, using plant, animal, object, and abstract analogies. However, there has been far less investigation of the ways in which analogy construction has impacted on the history of science, and the momentous mental achievements and creative insights which have resulted from this (see *Scientific Reasoning and Discovery, Cognitive Psychology of*).

Many significant examples can be found to illustrate how analogical reasoning has helped to break new ground in the history of science. The origins of analogy-based scientific innovation can be traced back to the early stages of Greek thought, with Archimedes one of its most eminent representatives. The Egyptian shadoof—a long-armed water-raising device used to draw water from wells—was invented long before Archimedes' time, and it was known that levers could be used to amplify physical force. It was left to Archimedes, however, to discover the relationship between the two arms, and thus to establish the principle of the lever. Hence mechanics—an essential precursor of physics—was founded on the basis of a number-pair analogy between the respective distances from the fulcrum and force relations according to the principle of levers. This development was founded on a paradigm of analogical thought, i.e., load arm: effort arm = length A: length B. Using the same method, i.e., in analogy to the principle of the lever, Archimedes was also able to calculate the surface area of any segment of a parabola (Serres 1994).

Eratosthenes is thought to be the first person to have calculated the earth's circumference. People had known how to determine the circumference of a circle using 360° since Babylonian times (influenced by the Sumerian sexagesimal system). Eratosthenes applied this knowledge in the following analogy: The distance between Syene (now Aswan) and Alexandria was approximately 800 km (in today's system of measurement). Once a year, the sun was at its zenith directly over a well in Syene and hence there are no shadows at

that location at that time. At the same time in Alexandria, it cast a shadow from which it could be calculated that the sun was 7° from the zenith. This represents a difference of about $1/50$ of a complete circle, hence the distance between Alexandria and Syene is about $1/50$ of the earth's circumference. By multiplying the distance (800 km) by 50, Eratosthenes calculated the earth's circumference to be approximately 40,000 km.

In his famed *Dialogue on the Two Chief World Systems* (the Ptolemaic system and the Copernican system), Galileo contradicted the Aristotelians of the time. The Aristotelians advocated the Ptolemaic hypothesis, and presented the following argument against the Copernican model: Assuming that Copernicus is right in saying that the earth rotates around its own axis, the circumference of the earth is such that this must happen at high speed—40,000 km in 24 hours (or 920 m in some 2 seconds' free fall). As a test of the Copernican system, one can climb a high tower, drop a stone from the top, and note the position at which it hits the ground. Since the earth is turning while the stone is falling, the stone should be noticeably removed from the foot of the tower when it makes contact with the earth. This is not the case, however. The stone falls perpendicular to the tower and hits the ground at its foot. According to the Aristotelians, this disproved the Copernican system. Using the character of Salviati in the *Dialogue* as his mouthpiece, Galileo refuted this argument with the following analogy: Imagine a sailing boat traveling at high speed. If a stone is dropped from the top of the mast, it will hit the deck at the foot of the mast, since it retains the speed of the boat until it hits the ground. Galileo thus invalidated the anti-Copernican argument by means of an analogy.

The most impressive known example of analogical reasoning in scientific discovery is provided by the equations describing electromagnetism derived by James Clerk Maxwell in 1864. Maxwell's contemporary, Michael Faraday, had demonstrated the existence of electric and magnetic fields, showing that iron filings sprinkled on a thin piece of paper held above a magnet are attracted to its 'lines or force' or its poles. Maxwell investigated the properties of the lines of force described by Faraday and discovered similarities between the behavior of these lines of force and the flow of a liquid in a pipe. The equations for the latter were already known. Maxwell transferred this knowledge to magnetic fields and electric currents, and was thus able to define the behavior of magnetic lines of force at and between poles. Writing on this means of discovery, Maxwell postulated that in order to form an idea of physical laws for which no theory is available, it is necessary to familiarize oneself with the existence of physical analogies. The discovery of the interaction between electric currents and magnetic fields laid the foundations of the theory and practice of electromagnetism, and hence of the dynamo tech-

nology at the close of the nineteenth century. Moreover, it was the precursor of radio and television wave transmission, since one of Maxwell's analogous equations had suggested that radio waves must also exist. Twenty years after Maxwell presented his basic equations, Heinrich Hertz indeed succeeded in producing radio waves under laboratory conditions—another demonstration of how creative thought in the form of analogical reasoning can impact dramatically on the history of science and the fate of humankind.

See also: Analytic Induction; Informal Reasoning, Psychology of; Practical Reasoning; Philosophical Aspects; Problem Solving and Reasoning, Psychology of; Problem Solving (Everyday), Psychology of; Problem Solving (Joint), Psychology of; Scientific Reasoning and Discovery, Cognitive Psychology of

Bibliography

- Bachmann T 1998 Die Ähnlichkeit zwischen analogen Wissensdomänen [The similarity between analogous domains of knowledge]. Doctoral dissertation, Humboldt University, Berlin
- Becker O 1990 *Grundlagen der Mathematik in geschichtlicher Entwicklung* [Foundation of mathematics in historical development]. STW, Freiburg, Germany
- Beyer R 1991 Untersuchungen zum Verstehen und zum Gestalten von Texten [Investigations of the comprehension and organisation of texts]. In: Klix F, Roth E, van der Meer E (eds.) *Kognitive Prozesse und geistige Leistung*. Deutscher Verlag der Wissenschaften, Berlin
- Bolzano B 1970 *Wissenschaftslehre* [Scientific model]. Akademie Verlag, Berlin
- Gentner D, Stevens A L 1983 *Mental Models*. Lawrence Erlbaum, Hillsdale, NJ
- Gluck M A, Bower G H 1988 From conditioning to category learning: An adaptive network model. *Journal of Experimental Psychology: General* **117**: 227–47
- Greeno J G, Riley M S 1984 Prozesse des Verstehens und ihre Entwicklung [Processes of comprehension and their development]. In: Weinter F E, Kluwe R (eds.) *Metakognition, Motivation und Lernen*. Kohlhammer, Stuttgart, Germany, pp. 252–74
- Hoffmann J 1993 *Vorhersage und Erkenntnis* [Prediction and Knowledge]. Hogrefe, Göttingen, Germany
- Horgan G 1995 Die neuen Sozialdarwinisten [The new Social Darwinists]. *Spektrum der Wissenschaft* **12**: 80–90
- Kahnemann D, Tversky J A 1973 On the psychology of prediction. *Psychological Review* **80**: 207–51
- Klix F, Bachmann T 1992 Analogy detection—analogy construction: An approach to similarity in higher order reasoning. *Zeitschrift für Psychologie* **206**: 125–43
- Leibniz G W 1960 *Fragments der Logik* [Fragments of Logic]. Akademie Verlag, Berlin
- Rips L J, Shoben E J, Smith E E 1973 Semantic distance and the verification of semantic relations. *Journal of Verbal Learning and Verbal Behavior* **12**: 1–20
- Rumelhart D E 1979 Some problems with the notion of literal meanings. In: Ortony A (ed.) *Metaphor and Thought*. Cambridge University Press, Cambridge, UK

Schmidt H-D 1966 *Leistungschance, Erfolgserwartung und Entscheidung* [Chance, Expectation and Decision]. Deutscher Verlag der Wissenschaften, Berlin

Serres M 1994 *Elemente einer Geschichte der Wissenschaften* [Elements of a History of Science]. Suhrkamp, Frankfurt, Germany

van der Meer E 1991 Die dynamische Struktur von Ereigniswissen [The dynamic structure of event-related knowledge]. In: Klix F, Roth E, van der Meer E (eds.) *Kognitive Prozesse und geistige Leistung*. Deutscher Verlag der Wissenschaften, Berlin

Wason P C, Johnson-Laird P N 1972 *Psychology and Reasoning*. Harvard University Press, Cambridge, MA

F. Klix

Problem Solving (Everyday), Psychology of

This article reviews developments in the psychology of everyday problem solving. The adjective 'everyday' is interpreted by contrast with the puzzles and academic tasks that have been the main focus of psychological problem solving research. This leads to an emphasis on the relation between planning and action, and between the mental representation of the problem and the physical situation in which problem solving takes place. The role of artifacts in support of everyday problem solving is discussed, and the idea that problem solving can be understood as an adaptation to environmental constraints is reviewed.

1. Background

The psychology of human problem solving was transformed by the work of Newell and Simon (1972) and all subsequent theorizing has used their work as a platform. Newell and Simon (1972) developed a very general abstract conception of problem solving as search through a problem space. A problem is defined by a goal state, a start state, and a set of operators, and the problem solver must achieve the goal state by applying a sequence of operators so as to transform the start state into the goal state. The set of states that are generatively defined by the start state and the operators is called the 'problem space,' and problem solving can be viewed as a search of this problem space for the goal state.

Newell and Simon's early empirical work within this framework concentrated on discovering the heuristics that people use to direct search through the problem space. This led them and their followers initially to

concentrate on well-defined puzzles, with a pre-determined set of operators and a completely specified problem space. This program of research was very effective, leading at least one prominent reviewer (VanLehn 1989) to claim that the psychology of human problem solving had been 'solved' (leaving as a residual mystery the question of what is learned from problem-solving experience).

But however valuable puzzles like the Tower of Hanoi or letter-arithmetic might be for the study of human problem solving, they are clearly not 'everyday' problems. There is no doubt that the problem space theory and search heuristics such as means-ends analysis and hill-climbing are vital to a full understanding of everyday human problem solving. But equally, the salient and manifold differences between many everyday problems and well-defined, self-contained puzzles raise important issues for the problem solver and for the psychologist.

Consider the following examples of everyday problems (from Anderson 1990, p. 193, given by his students as answers to the question 'name a problem that you tried to solve in the last 24 hours').

- Finding a route from a hotel to a restaurant
- Debugging someone else's code
- Getting subjects to sign up for my experiment
- Getting two Diet Cokes from a Coke machine
- Getting my car registered

Perhaps the most obvious special property of puzzles like Tower of Hanoi compared to these problems is that all the knowledge needed for a solution is provided in the problem statement. (This is not quite true of letter-arithmetic, of course, but even then the additional knowledge required, that is, arithmetic, is well-defined and strictly demarcated.) In particular the set of applicable operators is constrained and pre-determined.

An extensive literature has extended the study of human problem solving to 'knowledge-intensive' domains, in which the application of the solver's experience (in particular by choosing operators from long-term memory) is the critical issue. The most extensively studied domains have been academic domains like mathematics, physics, and computer programming (as well as chess, which is self-contained in principle, but reliant on long-term memory in practice), and the most extensively used methodology has been a comparison of expert and novice problem solving. The classic established finding of this research is that experts are better problem solvers than novices not because they are better able to use search heuristics, but rather because their knowledge of the domain allows them to identify and classify problems in terms of the methods that will be useful to solve them, thereby obviating the need for extensive search of the problem space. Experts tend to work forwards from problem statements to solutions, whereas novices tend to work backwards from the goal, searching memory for applicable knowledge (Larkin et al. 1980).

Most everyday problems are even less well-defined than academic problems in terms of the applicable operators. If you want to register your car and you don't know how, the problem is discovering the applicable operators, not merely ordering them. Furthermore, in contrast to puzzles and to academic problems, most everyday problem solving involves interacting with (so as to change) the everyday physical and social environment.

2. The Problem of Everyday Problems

Perhaps the single most important issue that the study of everyday problem solving prioritizes is the nature of the interplay between mental activity and physical activity, or between the mental problem space and the physical setting in which problem solving takes place. Within the traditional domains of problem solving research noted above, this has seldom been treated as a serious issue (although it should be noted that Newell and Simon recognized its importance, and have contributed some important analyses of the relation between external and internal representations, e.g., Simon 1975). Search of a problem space is primarily a mental activity: the current state is mentally represented, operators are considered, and imagined next-states are evaluated. Perhaps hierarchies and sequences of operators are mentally considered ('planning' or 'look-ahead'). Eventually an operator or sequence of operators is chosen, and then enacted. Many computational models of human problem solving gloss any distinction between mental and physical application of an operator. And early AI planning systems, building on Newell and Simon's analyses of search, computed entire sequences of operators from start state to goal state before acting on the world. People are seldom so patient. Instead they iteratively interleave planning and action. Since the early 1980s a body of literature has shown that in many everyday problem-solving situations the specifics of the physical situation and the problem solver's management of the relation between planning, perception, and action are crucial determinants of problem-solving success.

Consider one observation from some very influential empirical studies by the cognitive anthropologist Jean Lave and her colleagues (1984). A supermarket shopper picked up one slab of cheese from a shelf of similar cheeses to be surprised by its price. She guessed that it had been wrongly labeled but wanted to check. She might have used mental division to compute its price per unit weight and then performed the same computation on one of the other slabs (as usual the cheeses varied somewhat in weight and therefore price). Instead she searched the shelf for a cheese of very similar weight and directly compared the prices of the two cheeses.

A similar observation of in-situ problem solving was made by Scribner (1984). She observed dairy workers completing orders for partially full crates of

milk bottles. For example, a packer might receive an order of 1-6, meaning one crate-six bottles. Instead of completing this order by taking a full crate (16 bottles) and removing six bottles, the packer might find a crate has eight bottles in it and add two. This 'nonliteral' strategy involves mental effort, using arithmetic to recode the original order, but saves physical effort. (Scribner found that nonliteral strategies were common among skilled packers, and could almost always be accounted for by the saving of physical effort.)

Both the packer and the shopper are evidently managing a tradeoff between physical and perceptual activities (such as searching a display for relevant items) and mental activities (arithmetic operations). To fully understand their problem-solving behavior it is necessary to understand the opportunities for this tradeoff that are provided by the situation, and the cognitive mechanisms by which the tradeoff is managed.

3. Cognitive Artifacts

The first issue becomes particularly salient when coupled with the observation that much everyday problem solving is supported by 'cognitive artifacts' or tools of thought, ranging from pencil-and-paper through conventional representations like diagrams and graphs to computational systems. How can we understand the role of such artifacts in problem solving?

One theoretical approach is to simply redraw the boundaries of the cognitive agent away from the human problem solver to include the artifacts being used (Hutchins 1995). This approach has led to some insightful descriptions of how various human-machine systems perform their specific tasks (e.g., how a cockpit remembers its speeds). In the work of Zhang and Norman (1994) it has been developed into a general framework for the understanding of distributed cognitive tasks, in which the relation between external and internal representations is the primary focus. Zhang and Norman (1994) argue that when aspects of the representation of a problem can be offloaded into the physical environment, problem solving becomes easier.

A second approach focuses on how the artifact changes the cognitive task of its user. For example, Larkin and Simon (1987) have analyzed how diagrams facilitate certain problem-solving strategies by allowing simple perceptual judgments (such as whether two lines intersect) to replace more demanding mental computations.

Diagrams might be referred to as 'read only' cognitive artifacts. Artifacts that are 'read-write' can have more complex mediating effects on cognition. Consider the use of a simple appointment diary (Payne

1993). Recording an appointment in a diary changes the task of remembering that appointment in several ways. First, the fact that the appointment has been named and written will affect the internal memory trace for the appointment. And of course, the diary can act as an external memory, allowing its user to retrieve the appointment by reading. Furthermore, in conventional paper diaries the act of writing a new appointment leads to the inevitable rehearsal of already noted appointments.

A third approach to understanding cognitive artifacts is to focus on the properties of the artifact itself as an information system, and to ask how readily it allows certain information-processing operations. This approach has been pioneered in the field of human-computer interaction by Green (1989) under the banner 'cognitive dimensions of notations.' Green's enterprise is to define a limited number of widely applicable dimensions on which cognitive artifacts can be compared in terms of the extent to which they support flexible patterns of use. (Green has been particularly concerned with artifacts that are used in design tasks, such as spreadsheets, programming languages, document processors, and musical notations, and with supporting opportunistic design, in which initial drafts are iteratively revised.) One cognitive dimension is 'role expressiveness,' the extent to which a notation exposes meaningful structure. As a simple example, the use of meaningful variable names in a program will tend to increase role expressiveness. Another dimension is 'viscosity,' the extent to which a notation is resistant to local change. Consider a text with section numbers. Insert a section and all subsequent numbers must be changed.

All these approaches are set to continue in various guises. Understanding the cognitive effects of artifacts is one of the most important open problems for the theory of human problem solving.

4. Rational Approaches to Problem Solving

Let us now turn to the second crucial issue raised above, that is, how the tradeoff between mental and physical activity is managed. Scribner (1984) in her analysis of milk crate packing, suggested that packers might be explicitly trading mental and physical costs and benefits. More recently this basic idea has been developed, particularly in the work of John Anderson (1990) on what he has called the rational analysis of cognition. The general idea is to analyze cognition as an optimal adaptation to the probabilistic structure of the environment. Such an analysis asks *why* cognitive behavior is the way it is, and may help structure the subsequent investigation of *how* cognition approximates the optimal solution.

This approach has already thrown some light on the critical characteristics of everyday problem solving.

First, consider the interplay between planning and acting. From a rational perspective, a problem solver should plan before acting to the extent that the estimated benefits of planning outweigh the costs. The benefits of planning will relate to a reduction in the number of actions required and the individual costs of those actions. We might suppose that Scribner's dairy workers would have been less inclined to transform their orders arithmetically if moving bottles around had been somehow much easier. This idea has been tested in a series of experiments by O'Hara and Payne (1998) who showed that clumsy user interfaces to puzzles (i.e., interfaces that made each move more costly in terms of keystrokes or time) produced fewer-move solutions to puzzles than did more efficient interfaces.

The use of cognitive artifacts has also been illuminated by an adaptive analysis. A simple case is the electronic calculator. There are many problems that can be computed either mentally or by calculator. An experiment by Siegler and Lemaire (1997) showed that people are able to make this choice adaptively according to an accurate model of their own capabilities: that is, they reliably choose to do a task mentally when that is indeed the quickest way to do it, and to use the calculator instead when that has real benefits.

Finally, Pirolli and Card (1999) have shown that aspects of people's use of complex computer-based information retrieval systems can be well modeled by equations based on optimal foraging theory. In other words, people behave as if they could compute and use accurate estimates of their current and previous rates of information gain (cf. an animal's intake of energy from a food supply) and of the costs of searching different information resources (cf. an animal's journey to a new food supply).

5. Conclusions

Some theorists have argued that the situational contingencies of everyday problem solving render the abstract cognitive theory of Newell and Simon useless. However, an alternative view is that instead these complexities simply emphasize certain theoretical questions over others, and the required reworking is best considered as an elaboration of aspects of Newell and Simon's framework. This second approach has the advantage of allowing incremental development in our understanding of the psychology of human problem solving.

See also: External Memory, Psychology of; Problem Solving and Reasoning: Case-based; Problem Solving and Reasoning, Psychology of; Problem Solving: Deduction, Induction, and Analogical Reasoning

Bibliography

- Anderson J R 1990 *The Adaptive Character of Thought*. Erlbaum, Hillsdale, NJ
- Green T R G 1989 Cognitive dimensions of notations. In: Sutcliffe A, Macauley L (eds.) *People and Computers V*. Cambridge University Press, Cambridge, UK, pp. 443–60
- Hutchins E 1995 *Cognition in the Wild*. MIT Press, Cambridge, MA
- Larkin J, McDermott J, Simon D P, Simon H A 1980 Expert and novice performance in solving physics problems. *Science* **208**: 1335–42
- Larkin J H, Simon H A 1987 Why a diagram is (sometimes) worth 10,000 words. *Cognitive Science* **11**: 65–100
- Lave J, Murtaugh M, de la Rocha O 1984 The dialectic of arithmetic in grocery shopping. In: Rogoff B, Lave J (eds.) *Everyday Cognition: Its Development in Social Context*. Harvard University Press, Cambridge, MA, pp. 67–94
- Newell A, Simon H A 1972 *Human Problem Solving*. Prentice-Hall, Englewood Cliffs, NJ
- Payne S J 1993 Understanding calendar use. *Human-Computer Interaction* **8**: 83–100
- O'Hara K P, Payne S J 1998 The effects of operator implementation cost on planfulness of problem solving and learning. *Cognitive Psychology* **35**: 34–70
- Pirolli P L, Card S K 1999 Information foraging. *Psychological Review* **106**: 643–75
- Scribner S 1984 Studying working intelligence. In: Rogoff B, Lave J (eds.) *Everyday Cognition: Its Development in Social Context*. Harvard University Press, Cambridge, MA, pp. 9–40
- Siegler R S, Lemaire P 1997 Older and younger adults' strategy choices in multiplication: Testing predictions of ASCM using the choice/no-choice method. *Journal of Experimental Psychology: General* **126**: 71–92
- Simon H A 1975 The functional equivalence of problem solving skills. *Cognitive Psychology* **7**: 268–88
- VanLehn K 1989 Problem solving and cognitive skill acquisition. In: Posner M I (ed.) *Foundations of Cognitive Science*. MIT Press, Cambridge, MA
- Zhang J J, Norman D A 1994 Representations in distributed cognitive tasks. *Cognitive Science* **18**: 87–122

S. J. Payne

Problem Solving (Joint), Psychology of

'Joint problem solving' refers to the situation where two or more people work together to solve a shared problem, where agreement on a single solution may or may not be required. The proverb says 'two heads are better than one,' but why so, or when so, is not explicitly understood yet. Despite this folk knowledge, studies have shown that joint solving does not always result in better performance than that of the most capable solos. Rather, careful studies of the processes have revealed that there are clear cognitive process gains in the joint endeavor. In joint situations, the

participants tended to deepen their conceptual understandings of the topic problem (Miyake 1986), or the subjects generated a greater number of experimental hypotheses and had better chances to duplicate a scientific discovery (Okada and Simon 1997). The reasons for such gains are still under investigation, but research findings suggest that the joint situation allows the participants to naturally externalize their solving processes, and this externalization provides chances for flexible revisions of their trials, leading them to more conceptual understandings and newer ideas.

1. Gains of Joint Processes

Joint problem solving has been extensively studied in social psychology, mainly with an interest in whether people in groups, thus in 'social' situations, would perform better than solos. Hastie (1986) gives an extensive review, where joint cases are compared against solos on tasks like number estimation, vocabulary tests, and arithmetic brainteasers. Hastie summarizes the field and concludes, 'these studies tended to find that group performance exceeded the performance of the average individual, [but] for most of these problems groups did not match the performance of their best members' (p. 150). He also suggests that 'the critical characteristic of performance on the tasks is solution demonstrability, with higher levels of group performance associated with tasks where correct solutions are easily demonstrated and communicated to group members, and relatively low levels of performance associated with difficult-to-demonstrate solutions' (p. 150). His interest was mostly in categorizing group performances on tasks and group membership, not on cognitive aspects of the solving processes.

Miyake (1986) tried to clarify the cognitive gains associated with paired performance. She asked three pairs of subjects to understand how a sewing machine made its stitches and analyzed their verbal protocols. The sewing machine problem has interesting characteristics of having different levels of answers. One answer to the question would just mention that the machine uses two intertwined pieces of thread to fasten the cloth. On hearing this, one can further ask about the exact physical configuration of the seam, which turns out to be a loop-in-loop. Discovering this could open up a new puzzle because it is physically impossible to form a loop-in-loop with two threads without free ends. (At a middle point in sewing, for both upper and lower threads, one end of them is in the spindle or in the bobbin, while the other end is tightly sewn to the cloth.) In fact, while the upper thread comes down into the machine, its loop goes entirely around the bobbin. This is a perfect answer until one starts wondering how, then, the bobbin itself is attached to the machine.

This hierarchical nature of the possible answers to the sewing machine problem made it possible to analyze the protocols in terms of when they reached the intermediate understandings and how and when they noticed the insufficiency of their solutions and started to ask the 'one level deeper' questions. Miyake found that each of her subjects descended this hierarchy according to his or her individualistic formulation of the question. This often resulted in situations where members of the pair worked at different levels at the same time: while one person may worry about the physical configuration of the seam, the other could try to find how the loop-in-loop configuration is possible with two threads without free ends. Because of this disagreement, the subjects were almost always in conflict, which contributed to push each individual to refine his or her own models so that 'the other member would understand my points.' In summary, Miyake argues that the process gain of joint problem solving does not necessarily lie in the performance, but rather in raising the necessity and motivation for deepening and refining each individual's understandings.

In the 1990s, the amount of research on joint problem solving increased, showing more positive results for both pairs and groups. Conditions for successful collaborations were identified where earlier studies focused on degrees of mutual participation, while the more recent studies identified explicit types of talk that contributed to the success (e.g., Teasley's 'transactive' talks, Teasley 1997). A similar style of research has been extended to cover topics of learning and development (e.g., Rogoff 1990). Hatano and Inagaki (1991), for example, analyzed a science classroom where the children selected among the alternative hypotheses provided by the teacher and debated among them. By analyzing the discussion they identified what they named 'the partisan motivation,' shared by the naturally grouped members with the same alternative, that acted as an enhancer for the self-motivated articulation of their explanations.

2. Joint Problem Solving in the Real World

Okada and Simon (1997) took a slightly different course of identifying causes for pairs' success in an experimental scientific 'discovery.' Using a computer-simulated task of making a discovery in molecular genetics, paired subjects were found to be superior to single subjects. This success was associated with higher levels of explanatory activities of the pairs, and further, such explanatory activities were facilitated when paired subjects made explicit requests of each other for explanations and testable hypotheses.

This change occurred with the search into more ecologically valid research in the domain of problem solving. Daily problems often get solved in groups,

and there are pragmatic needs and interests in finding out how groups should be organized to be productive. This led cognitive problem solving research to take an 'in vivo' form (Dunber 1995), where workers and scientists in realistic situations were carefully studied. Dunber observed four laboratories carrying out immunization research and followed their daily scientific activities for a year, with various methods from attending the research meetings and taking notes, reading their lab diaries, to extensive interviews. His findings are diverse, and factors he found positive include frequent exchange of ideas, nonauthoritative membership, positive attitudes toward even negative intermediate results, use of local analogies, etc. In an interview-based study of a real R & D team in a detergent company, Ueda and Niwa (1996) found that having an outsider's view at the right time was crucial for the team to change the course, which led them to produce a new successful product. All such reported factors tend to confirm the folk expectation of what should contribute to high group performance. Yet how we can coordinate these factors to make them all work simply remains to be discovered. Furthermore, the overall conclusion we could draw from these sets of research is that the quality of group performances depends heavily on the organization and activity control of the groups.

3. Joint Problem Solving from a Cognitive Anthropology Perspective

In contrast, cognitive anthropologists have maintained more positive perspectives. Hutchins (1990) observed a team navigating a large ship and found that there was overlapping knowledge shared among the team members that strongly coordinated the teamwork. In the reported case of a 'fix,' or a joint activity of finding the exact location of a ship, at least three roles are involved. The bearing-taker measures the angles between the ship's direction and selected landmarks using a gyroscope, and calls out the number through a tube to the charting house. Then this number is repeated by the bearing timer-recorder who writes down the number into the log book. The number then gets converted to the angle on the device that allows the plotter to make lines on the chart so that they can pinpoint the location of the ship onto the map. Staff start working in this team as the bearing-taker, who after some experience is promoted to the timer-recorder job, who would later move into the plotter position. Thus the plotter knows everything the other two know, and the timer-recorder knows all of what the bearing-taker knows. It is also possible for the lower-rank staffs to closely observe what the staff one rank higher is engaged in. While a cross-sectional, one-shot observation might make this teamwork appear to be highly coordinated across divisions of work, in reality what coordinates this team is the shared

knowledge of a longitudinal, historical nature. Laboratory-oriented, puzzle-based experimental studies of joint problem solving have been quite weak on this aspect.

Hutchins' report maintains one more important observation on how the teamwork is divided and carried out. In the above example of taking a fix, there are three subtasks, each of them producing some intermediate results. Careful inspection reveals that each of them takes quite a different, and redundant form of representation of the intermediate results of a single process. The first outcome of the bearer is by voice, which is then written down in number symbols, which is then again converted into lines on the chart. These different forms, at different times of representation, provide different and redundant forms for people involved to check the process from different perspectives. In this sense this is a real-world documentation of what Hastie called 'demonstrability,' which, in realistic situations, does not just provides the information concerning the correctness of the solution but opens up the possibility for other interpretations and possibly flexible redesigning of the process itself.

4. Toward Further Research into the Mechanisms of Joint Gain

Paired subjects could, in fact, take advantage of externally shared objects in the course of simpler experimental tasks. Shirouzu et al. (1999) asked subjects to indicate two-thirds of three-quarters of the area of a square, folding paper measuring 15 cm × 15 cm. They report that in various conditions people strongly opted (over 90 percent on average) to use external resources to do this arithmetic, either by folding the paper or by marking the edges with their fingers and partitioning, etc. Because this reliance on external resources is quite strong, most of the single subjects did not change the solution method (from folding to calculating, for example) even when they were consecutively asked the second question, i.e., to indicate the area of three-quarters of two-thirds of the same sized paper. The subjects did not change it even after seeing the 'externalized' result to be exactly one-half of the paper. They compared this to pair performances and reported that the pairs are a lot more flexible in choosing different solutions to the second question (70 percent of the pairs changed the solution methods while less than 20 percent of the single subjects did so). By analyzing the solving processes and the protocols of such subjects they found that the flexible pairs were in fact gradually responding to the constantly changing externalized forms of the paper while they worked together, rather than each randomly contributing newer ideas at their turns.

As a future course of joint problem-solving research, new learning studies focus on collaborations. As indicated above, joint situations tend to provide the

participants with verbal expressions and tangible externalized objects that each member can manipulate to concretely suggest newer, or at least different, ideas. In rapidly changing societies, the goal of learning is not only digesting what is taught. Learners are required to acquire skills to learn, to become self-regulated learners. In order to fulfill this goal, it is necessary to give learners more chances to be reflective than currently enforced, to scrutinize upon and revise what they themselves do and how others do similar tasks. Studies on joint problem-solving situations are expected to provide rich resources to help develop effective learning situations (e.g., Bransford et al. 2000).

See also: Group Decision Making, Cognitive Psychology of; Group Decision Making, Social Psychology of; Group Processes in Organizations; Group Processes, Social Psychology of; Group Productivity, Social Psychology of; Joint Action; Problem Solving and Reasoning, Psychology of

Bibliography

- Bransford J D, Brown A L, Cocking R R (eds.) 2000 *How People Learn*, expanded version. National Academy Press, Washington, DC
- Dunbar K 1995 How scientists really reason: Scientific reasoning in real world laboratories. In: Sternberg R J, Davidson J E (eds.) *The Nature of Insight*. Cambridge University Press, New York
- Hastie R 1986 Review essay: Experimental evidence on group accuracy. In: Grofman B, Owen G (eds.) *Decision Research, Vol. 1, Information Pooling and Group Decision Making: Proceedings of the Second Conference on Political Economy*. University of California, Irvine, CA, pp. 129–57
- Hatano G, Inagaki K 1991 Sharing cognition through collective comprehension activity. In: Resnick L B, Levine J M, Teasley S D (eds.) *Perspectives on Socially Shared Cognition*. APA, Washington, DC
- Hutchins E 1990 The technology of team navigation. In: Galegher J, Kraut R E, Egido C (eds.) *Intellectual Teamwork: Social and Technological Foundations of Cooperative Work*. Erlbaum, Hillsdale, NJ
- Miyake N 1986 Constructive interaction and the iterative process of understanding. *Cognitive Science* **10**: 151–77
- Okada T, Simon H A 1997 Collaborative discovery in a scientific domain. *Cognitive Science* **21**(2): 109–46
- Rogoff B 1990 *Apprenticeship in Thinking: Cognitive Development in Social Context*. Oxford University Press, Oxford, UK
- Shirouzu H, Miyake N, Masukawa H 1999 Roles of cognitive externalization for joint problem solving. In: *Proceedings of the 2nd International Conference on Cognitive Science and the 16th Annual Meeting of the Japanese Cognitive Science Society Joint Conference*, pp. 337–42
- Teasley S D 1997 Talking about reasoning: How important is the peer in peer collaboration. In: Resnick L B, Saljo R, Pontecorvo C, Burge B (eds.) *Discourse, Tools, and Reasoning: Essays on Situated Cognition*, NATO ASI Series: Series F: Computer and systems sciences. Springer, Vol. 160, pp. 361–84

N. Miyake

Procedure: Legal Aspects

1. The Subject Matter of Procedure

Procedure addresses the work of adjudicatory processes that are based in courts. While the subject may appear dry (if not boring), the rules of any procedural system are interesting because they must respond to difficult political and social questions.

Traditionally the world of procedure has been divided into three categories—civil, criminal, and administrative. For example, civil cases are those in which no criminal penalties (often but not exclusively incarceration) are sought. Administrative law proceedings are those in which (typically) the government is a party and the case is processed in administrative agencies devoted to a particular subject matter rather than in generalist courts. Courts in turn are defined as institutions that handle civil and criminal cases. However, these lines are blurring. For example, while civil disputes often involve private citizens, in a substantial proportion of civil cases the government is a party, as either plaintiff or defendant. Concerns about overcrowded court dockets have prompted some commentators to recommend that mandates to agencies be enlarged—turning more agencies in essence into specialized courts.

What is shared among civil, criminal, and administrative proceedings is that someone (an individual, a group, or a representative of the government) claims that another has violated a legal norm. In all three categories a third party, empowered by the state, is called upon to determine the validity of the claim—whether a legal right exists ('liability') and what form of response ('remedy') is appropriate. In all three categories, the parties have some degree of autonomy to fashion the contours of the dispute and yet are also limited by conventions and rules. In all three categories, questions exist about the power of decision-makers, the constraints imposed upon them, the kinds of sanctions or remedies that might be imposed, and the ability to seek reconsideration of the decisions rendered. And, in all three categories, the procedures are assessed as to whether they are fair and efficient.

What makes for fair and impartial judgments? Should everyone have a 'right' to an authoritative adjudication, or may courts insist that litigants settle or resolve cases by other means? If trials are permitted, should they be decided by judges (professional adjudicators) or by lay juries? If judges are central, how is their impartiality and independence assured? Does the gender, race, or ethnicity of a decision-maker matter to the perceived or actual fairness of the outcome? Should state subsidies, including free lawyers, be provided for those too poor to litigate to facilitate access to justice?

Of course, some of the questions for proceduralists are more technical, such as: What injuries are grounds for a lawsuit? How does one actually start a case? In

which court? Whom should one sue? Yet responses to even these questions depend on judgments about the role of government, the degree to which individuals should be permitted autonomy as rights seekers, and the harm that litigation itself can inflict on unwilling participants. Therefore, rules of procedure embody, express, and help create values and attitudes about what constitutes fair decision making by, or through, the powers of the state. Procedure is laced with questions of legal policy and political philosophy.

In some countries, those questions are also framed in constitutional terms. For example, the US Constitution deploys the phrase ‘due process of law’ in its Fifth and Fourteenth Amendments, guaranteeing protection against deprivations of life, liberty, or property. Also detailed are rights of jury trials for criminal defendants and for civil damage litigation, as well as rights of counsel for criminal defendants. Other countries expressly address these and yet other rights, such as providing criminal defendants with a right of translation into their own language.

The set of problems facing proceduralists is rich, not only because rules of procedure inevitably implicate basic social and political values. Procedure is also compelling because those rules intersect with all the major problems of contemporary life. Under the umbrella of ‘cases’ can come the full range of social problems—disputes over personal assets, and familial relationships; claims of physical injuries from products like asbestos or tobacco; disagreements about commercial contracts made or breached; regulatory arguments about possible hazards to the environment or the illegal pricing of products; assertions of violations of state or federal statutes providing for civil rights, consumer information, or personal security. In short, Procedure must address the theories, aspirations, and values that animate procedural systems; the kinds of disputes that may prompt litigation; the means by which lawsuits are started and stopped, as well as what happens in between and who can participate; consideration of the identity and characteristics of the participants (parties, lawyers, witnesses, judges and juries); the powers of decision makers; what the structure of courts is, has been, and how it might be changed; and what role courts in turn play in a particular governmental structure or as trans-national institutions.

2. Lawsuits: From Injuries to Litigation

When thinking about Procedure, it is important to remember that not all moments of discord result in the filing of lawsuits, let alone in the kinds of proceedings described in newspapers or the trials that are televised. In fact, individuals and groups disagree a lot without ‘naming’ the problem a conflict. In contrast to popular

assertions, for example, that the US is awash with litigation, scholars of dispute resolution instruct otherwise. When injured, individuals have a variety of responses (from ‘lumping it’ and never seeking redress to ‘claiming’ injury and then ‘blaming’ others); a numerically small percentage of such occurrences results in the actual filing of civil, criminal, or administrative proceedings. How often people file suit, in turn, is related to what forms of redress regulatory and social welfare regimes provide, and to conceptions of what role courts should take in a given polity. The role litigation plays in countries therefore differs.

Moreover, litigation should not be equated with adjudication through trial. In some countries, of cases that are filed, very few conclude by trials. Take the federal courts of the US as an example. Of 100 civil cases filed in the year 2000, about a third will be resolved by judicial rulings in advance of trial; most of the remaining two-thirds will be concluded through settlement. Trials will begin in only three or four cases.

The rate of trial in the US declined during the last decades of the twentieth century, as courts built into their processes a ‘pretrial phase’ that becomes the focus of lawyering and that offers a means of dispute resolution without trial. Many countries have given litigants rights of access to information in advance of trial; ‘discovery’ and ‘disclosure’ are terms of art in the US to refer to such pretrial exchanges. Further, negotiation, mediation, and arbitration (sometimes called ‘alternative dispute resolution’ or ‘ADR’) are concurrent options—or requirements—throughout the course of disputes.

Some procedural systems are styled ‘inquisitorial,’ by which is meant that judges determine a good deal of the shape of the inquiry. Others, described as more ‘adversarial’ and associated with common law countries, authorize parties (or more accurately their lawyers) to determine the scope of a controversy, to develop the information about the events at issue, and to frame legal theories. During the twentieth century, many common law countries revamped their procedural systems—as did civil law regimes—such that the two forms of adjudication are moving closer to each other.

In the US, federal rule makers undertook major reorganization of the procedural system in the 1930s. They spoke about aspirations for a simple, accessible justice system and they crafted a ‘trans-substantive’ set of procedures, by which it is meant that the same rules were to be used, regardless of the kind of case. By the 1970s, some litigants, judges, and lawyers became concerned that the system created decades earlier had generated its own problems. They complained about lawyers exploiting pretrial opportunities for personal gain and to inflict misery on opponents. Critics charged abusive discovery, incivility, and adversarial excess. Some reformers urged lessening reliance on lawyers by reeducating clients to look for resolutions

of disputes without, or with less lawyering. Others proposed rules aimed at constraining lawyers.

Many common law countries turned to judges to take more responsibility for the shape of lawsuits. Judges and other court personnel began to take an active role in scheduling and monitoring the pretrial phases of cases—to move cases along, to urge settlement, to attempt to supervise attorney behavior. Judges came to act as ‘managers,’ working with lawyers to set time frames in which parties may be added or dropped from cases, to resolve disagreements about the relevance of information sought from opponents, and to urge that disputes be ended by mandatory settlement conferences, court-annexed arbitration, and mediation. Procedures once termed ‘extrajudicial’ have been built into rules, and are now regular features of civil and criminal processes.

The term of art—in the US, England, Wales, Canada, and Australia—for such work is ‘case management’ but the content of such work varies by country. In the US, a good deal of the focus is on managing lawyers. In this respect, rules of procedure are connected to rules of professional conduct. Rules of lawyering ethics and rules of procedural systems increasingly overlap as both are aimed at structuring the behavior of attorneys.

Late twentieth-century reforms seek to lower the costs of process and to shorten the time to decision. Some reformers believe that these changes will help lower-income individuals gain access to courts. Critics view the rules as less friendly to small claimants because of limitations on discretion and flexibility. These debates point to a more general feature of procedural rule making. Procedural reforms of any era need to be understood as not ‘only’ about procedure; they are related to the creation of professionalized judiciaries, to country-specific social movements, and to institutionalization of courts as corporate actors within governments. Moreover, rule making is often driven by ‘repeat player’ litigants who have the economic wherewithal and the political interest and insight to seek to shape procedural regimes.

3. *The Many Configurations of Lawsuits*

Given the diversity of problems, not surprisingly the configurations of lawsuits vary. Some cases involve two disputants seeking attention from a single judge, readily envisioned as a pyramidal triangle, with the judge at the top and the two disputants on either side. Other litigations are multi-party conflicts, for which a Calder mobile offers a better visual image; the many parties realign depending on what issue is to be determined in what forum.

Many questions have emerged, including: Should one-on-one cases be paradigmatic of adjudicatory processes? How should aggregate litigations affect

one’s understanding of the purposes and forms of adjudication? The costs of lawyers to individuals make pursuit of remedies expensive; evolving norms of judicial process discourage labor-intensive attention for ‘small’ claims. Further, the paradigmatic car crash, sometimes assumed to involve individual disputants, is often a dispute controlled by two insurance companies rather than by individual litigants. Many commentators urge that such conflicts be turned over to ‘no-fault’ administrative systems that would provide compensation according to schedules and that other ‘small cases’ be resolved through lower cost and lower visibility processes located in administrative agencies.

At the opposite end of the spectrum are large-scale cases, some involving tort law arising from accidents (like airplane crashes) and others related to allegedly defective or injurious products (such as diet drugs, nicotine, and asbestos). Some large-scale cases involve contracts and consumer claims, from shareholders of companies to purchasers of goods; and yet others relate to group-based rights claiming (about issues involving environmental harms to air, food, or water or the fair treatment of women). Such litigation also raises questions of whether administrative models are more appropriate modes of disposition than adjudicatory ones.

Given widespread allegations of injury and hundreds or thousands of lawsuits involving similar or the same parties, courts have devised means, both formal and informal, to process cases *en masse*. Procedural systems respond through mechanisms such as ‘consolidation,’ ‘class actions,’ ‘interpleader,’ ‘bankruptcy,’ ‘multi-district litigation’—all procedures designed to assist in reaching resolutions across categories of claimants.

4. *A Variety of Venues*

Not only do lawsuits vary in size and scale, they also go to an array of places for resolution, both public and private. Once a decision is made to file a lawsuit, a choice often exists about where to file a lawsuit. Some courts have ‘general jurisdiction’—they can handle all kinds of cases. Other courts have ‘limited jurisdiction,’ tied to a particular subject matter or amount in dispute.

In some instances, the jurisdictional grants overlap, and courts may have ‘concurrent jurisdiction.’ When more than one venue is available, common law systems give the power to the parties to decide the court in which to proceed. Upon occasion, defendants argue that plaintiffs have made the wrong selection—either because a particular tribunal lacks authority over the subject matter of the litigation (‘subject matter jurisdiction’) or because the tribunal has no power to compel a particular defendant to appear and defend or

is an inappropriate place to rule on the merits ('personal jurisdiction' or '*forum non conveniens*').

More institutions than courts are available for dispute resolution. Many government agencies are also quasi-courts. In the US for example, the Social Security Administration, the Equal Employment Opportunities Commission, and the Immigration and Naturalization Service process thousands of disputes annually, far in excess of the number decided by the federal court system. Further, court systems in courthouses or in administrative agencies often do not exhaust the set of options. Again using the US as an example, Indian tribes have court systems as well. Given the overlap among courts and the demographic mobility of the citizenry, inter-jurisdictional cooperation among state, federal, and tribal systems has begun to develop.

An increasingly global economy makes international decision-making tribunals ever more important to litigants within the US. One project, based at the American Law Institute, is to develop transnational rules of procedure to which countries or litigants might subscribe voluntarily. Another approach is to create universal jurisdiction for certain kinds of offenses; the International Criminal Court, a product of the Treaty of Rome, exemplifies that effort.

Governments are not the only source of courts. Associations may decide to establish their own 'private' courts and may compel those who contract with them or those employed by them to rely on such tribunals in lieu of proceeding in state-based courts. Further, governmental agencies and courts may also urge litigants to devise means of resolution of disputes without adjudication and may contract for services with 'private' institutions. During the second half of the twentieth century, a number of private providers of dispute resolution have come to the fore. As the market of private providers of dispute resolution services has expanded, some worry that publicly based and financed systems may lose support as those with the resources go elsewhere.

5. Tiers of Decision Making

If cases are decided within formal court or administrative systems (by judges, juries, or hearing officers), losing parties may have a 'right' to appeal to a higher court, which has (often limited) authority to review decisions made below. In many jurisdictions, there are three or four tiers of decision-makers, some working at the trial level, which has 'original' jurisdiction, and others in an appellate system. Some appellate systems are obliged to review lower court rulings when presented with an appeal, whereas others enjoy discretion to decide whether to review cases and if so, which ones. At each stage, parties typically have the option to settle or withdraw.

The concept of 'preclusion' bears on these tiers of decision making. The idea is that at some point, procedural systems prohibit further consideration of a problem—because participants have (knowingly or inadvertently) 'waived' their rights by failing to raise certain claims in a specified time or because participants have made and lost claims along the way. '*Res judicata*,' '*estoppel*,' 'double jeopardy,' and 'statutes of limitations' are some of the legal terms that capture these principles of closure and finality. Yet most procedural systems also have safety hatches—means by which to reopen and reconsider issues because of new information, changing circumstances, or the passage of time. Child custody decisions are typically subject to reopening, and in criminal cases, the availability of '*habeas corpus*' marks a narrow window of reconsideration. Herein lies an exemplary tension of procedural rules, attempting to capture both the impulse toward closure and the recognition of limited knowledge and the imperfections of judgment.

6. The Demography of the Court House

During the twentieth century, the demographics of the participants of the legal system changed. Women gained juridical rights to bring lawsuits, as did peoples of many colors. Further peoples of all colors and both genders gained rights of access to the professions of lawyering and judging. These individuals prompted professional interest in the demographics of the court house to understand more about the racial, ethnic, and gender composition of judges, jurors, lawyers, court personnel, witnesses, and litigants, and to learn whether the processes of decision making are affected by the demographics.

Judiciaries and/or bar associations in countries around the world convened special task forces on the effects of gender, race, and ethnicity in the courts and in the legal profession. These joint bench and bar efforts fall within a long tradition of concern about the administration of justice; judges and lawyers have engaged in studies ranging from the problems of prison litigation and the death penalty to how to handle large numbers of asbestos filings. Proceduralists thus have new knowledge and perspectives from which to consider the processes of justice.

The work on race, ethnicity, and gender in the courts has generated data about the composition of the judiciaries and the bar, as well as about the experiences of male and female lawyers of all colors, and more limited information on litigants. Most of the reports conclude that court systems—like the culture in which they exist—reflect stereotypic assumptions about race, ethnicity, and gender. Women judges and women lawyers are often in dramatic minorities as contrasted with the number of women litigants and, in systems with juries, of women jurors. Women and men

who are lawyers and judges, and peoples of differing colors report different perceptions of the relevance of gender, race, and ethnicity to processes of adjudication. Studies in the US have found that women attorneys believe that gender affects both process and outcome, and describe its effects on their professional lives and the legal system. In contrast, male judges and lawyers are less likely to report gender as relevant to the administration of justice or the legal profession. Studies of minority judges and lawyers have parallel data: people of color report the relevance of color while non-minorities describe its unimportance to the justice system. As a consequence of these projects, topics once considered gender- and race-neutral—from the specifics of jury instructions and courtroom conduct to larger questions about the role of the judge, the nature of lawyering, and the allocation of work among the courts—are being revisited to consider how to make good on the promises of fairness.

7. The Role of Resources

Because providing legal process costs money, takes time, and requires energy, the role of resources is critical to the study of procedure. Rules that treat all individuals as similarly situated belie the reality that some of us are richer than others, some of us have more access to help than do others, and some of us are more likely to be harmed than are others. Procedure must therefore consider how rules affect the costs of lawyering and the impact of rules that shift costs from one side to the other or that provide for government subsidies for lawyers, investigation, expert witnesses, filing fees and other costs associated with litigation. Economic analysis of many areas of law has become commonplace. In this arena, such inquiries consider the incentive structures created by procedural rules, the means by which adversaries can inflict costs on opponents, the mechanisms by which bargaining can occur, how rules about sanctions for failure to settle affect differently those who are 'risk prone' and those who are 'risk averse,' and the impact of attorney fee-shifting rules. Constitutional and political theorists, analyzing the meaning of the 'equal protection' and 'due process' guarantees, consider whether the government has an obligation to make courts readily accessible; if so, for what categories of claims, and how, given economic disparity, to craft rules to implement such principles.

8. The Constant Questions

Despite all the variation in the resources of litigants, the kinds of cases, the people and problems involved, the courts systems, and the procedures themselves, the underlying problems that procedural systems have to solve are the same: Who can seek redress? For what

kinds of injuries? Requesting what remedies? Based on what kind of information? Presented to what decision-maker? With any review or possibility of reconsideration? Obviously these questions do not admit to easy, unvarying answers; ideas about procedure are deeply rooted in the social-political fabric, and the issues central to procedure are value-laden. The questions to be addressed include: Why have process at all? Why care about how a decision is made? If process is required, how much should be provided? Are opportunities to be heard adequate? How formal should a procedure be? How expensive should it be? Who should pay? How free should the parties be to initiate lawsuits? Who gets to litigate and who is foreclosed? And what about the decision-makers—what kinds of information should be required prior to decision? What remedies should courts be able to order? How much power should judges have, and when should that power be constrained? When may decisions be reconsidered, and when should they be considered 'final'? Over and over again, procedure returns to these basic, and central, issues.

See also: Administrative Law; Civil Law; Conflict Sociology; Courts and Adjudication; Criminal Law and Crime Policy; Law: Overview; Laws, Conflict of; Legal Process and Social Science: United States; Parties: Litigants and Claimants; Rules in the Legal Process

Bibliography

- Cover R M 1975 *Justice Accused*. Yale University Press, CT
Cover R M 1986 Violence and the word. *Yale Law Journal* 95: 1601
Damaska M 1997 *Evidence Law, Adrift*. Yale Press, CT
Eisenberg T, Yeazell S C 1980 The ordinary and the extraordinary in institutional litigation. *Harvard Law Review* 93: 463
Fiss O 1978 The forms of justice. *Harvard Law Review* 93: 393
Fiss O 1984 Against settlement. *Yale Law Journal* 93: 1073
Galanter M 1972 Why the 'haves' come out ahead: speculations on the limits of legal change. *Law & Society Review* 9: 95
Galanter M, Cahill M 1994 'Most cases settle': Judicial promotion and regulation of settlement. *Stanford Law Review* 46: 1339
Heinz J P, Laumann E O 1994 *Chicago Lawyers: The Social Structure of the Bar*, rev. edn. Northwestern University Press, Chicago
Hensler D R 1989 Resolving mass toxic torts: myths and realities. *University Illinois Law Review* 89
Hensler D R, Dombrey-Moore B, Giddens B, Gross J, Moller E, Pace N M 2000 *Class Actions Dilemmas: Pursuing Public Goals for Private Gain*. RAND, Santa Monica, CA
Langbein J 1985 The German advantage in civil procedure. *University of Chicago Law Review* 52: 823
Lind E A, Tyler T R 1988 *The Social Psychology of Procedural Justice*. Plenum Press, New York
Menkel-Meadow C 1985 Portia in a different voice: speculations on a women's lawyering process. *Berkeley Women's Law Review* 1: 39

- Ninth Circuit Gender Bias Task Force 1994 The effects of gender: the final report of the ninth Circuit Gender Bias Task Force, republished. *Southern California Law Review* 67: 727
- Purcell E A Jr 2000 *Brandeis and the Progressive Constitution: Erie, the Judicial Power, and the Politics of the Federal Courts in the Twentieth-Century America*. Yale Press, CT
- Resnik J 1982 Managerial judges. *Harvard Law Review* 96: 372
- Resnik J 2000a Trial as error, jurisdiction as injury. *Harvard Law Review* 113: 924
- Resnik J 2000b Money matters: judicial market interventions creating subsidies and awarding fees and costs in individual and aggregate litigation. *University of Pennsylvania Law Review* 148: 2119
- Schuck P 1987 *Agent Orange on Trial*. Harvard Press, NY
- Shavell S 1982 Suit, settlement, and trial: a theoretical analysis under alternative methods for the allocation of legal costs. *Journal Legal Studies* 11: 55
- Subrin S H 1987 How equity conquered common law: the federal rules of civil procedure in historical perspective. *University of Pennsylvania Law Review* 135: 909
- Trubek D M, Sarat A, Felstiner W F, Kritzer H M, Grossman J B 1983 the costs of ordinary litigation. *UCLA Law Review* 31: 72
- Yeazell S 1994 The misunderstood consequences of modern civil process. *Wisconsin Law Review* 631
- Woolf H 1996 *Access to Justice: Final Report to the Lord Chancellor on the Civil Justice Systems in England and Wales*. The Stationery Office, London

J. Resnik

Procrustes Analysis

Procrustes analysis is concerned with matching two, or more, configurations of points. Imagine two maps of a region, each giving the positions of the same towns. The two maps are oriented differently, are of different sizes, differ in their accuracies, and use different map projections. At a first glance it may be difficult to see much resemblance between the two maps. Procrustes analysis addresses such problems, usually taking one of the configurations as a target and fitting the other to it, optimally, by using appropriate transformations. The 'maps' may be derived in many ways and may be in several dimensions but, for the concept of matching to be meaningful, it is important that the configurations refer to two representations of the same things: the same towns, the same cases (in multidimensional scaling), the same variables (in factor analysis), the same products (in marketing), the same landmarks (describing biological organisms).

The name comes from Greek mythology where Procrustes was an innkeeper who had a bed (sometimes two). Travelers were stretched or limbs chopped to make them fit the bed; with two, adjustment was ensured. The pejorative implications are no worse in

Procrustes analysis than in other forms of data-analysis where, misguidedly, data may be fitted into the straightjacket of an unrealistic model.

1. Factor Analysis

The origins of Procrustes analysis (see Hurley and Catell 1962) are in factor analysis, where loadings $\mathbf{X}_1$ of k factors $\mathbf{f}$ occur in the form $\mathbf{X}_1\mathbf{f}$ and so may be replaced by $(\mathbf{X}_1\mathbf{T})(\mathbf{T}^{-1}\mathbf{f})$ showing that they are determined only up to the arbitrary $k \times k$ matrix $\mathbf{T}$. By choosing a suitable $\mathbf{T}$, we may match $\mathbf{X}_1$ 'the traveler' to a previously determined, or hypothetical set of, loadings given in a matrix $\mathbf{X}_2$ 'the bed'; $\mathbf{T}$ represents the 'chopping/stretching transformation. Usually the degree of match is judged by least squares, minimizing $\|\mathbf{X}_1\mathbf{T} - \mathbf{X}_2\|$ the sum of the squared distances between like points of the target and transformed configurations. Without further constraint, this is the multivariate multiplenegression problem with well-known solution

$$\mathbf{T} = (\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbf{X}_2.$$

Two of the most important constraints on $\mathbf{T}$ are:

- (1) $\mathbf{T}$ must be orthogonal (denoted here by $\mathbf{T} = \mathbf{Q}$); and
- (2) $\mathbf{T}$ must be a direction cosine matrix (denoted here by $\mathbf{T} = \mathbf{C}$)

Both choices relate to the interpretation that $\mathbf{f}$ is expressed relative to an orthogonal system of co-ordinate axes and the columns of $\mathbf{T}$ give direction cosines of new axes relative to the original orthogonal axes. When $\mathbf{T} = \mathbf{Q}$, a new set of orthogonal axes is being sought that give a better match to $\mathbf{X}_2$; when $\mathbf{T} = \mathbf{C}$ we imagine $\mathbf{X}_2$ as being referred to oblique axes. The orthogonal Procrustes least-squares solution is $\mathbf{Q} = \mathbf{V}\mathbf{U}'$ where $\mathbf{U}\mathbf{\Sigma}\mathbf{V}'$ is the singular value decomposition of $\mathbf{X}_2'\mathbf{X}_1$. With $\mathbf{C}$, minimization is intrinsically more difficult and is less well defined than in the orthogonal case. This is because there are at least two ways of expressing the position of a point relative to oblique axes and also because there are good arguments for replacing ordinary leastsquares by a weighted form which measures distance in the metric appropriate to the chosen system of oblique axes, giving four variants in which $\mathbf{T}$ is either $\mathbf{C}$, $\mathbf{C}'$, $\mathbf{C}^{-1}$ or $(\mathbf{C}')^{-1}$. The simplest case, $\mathbf{T} = \mathbf{C}$, was first discussed by Mosier (1939) but not finally solved until Browne (1967) gave a practical numerical method for solving the normal equations. Special algorithms have been proposed by Browne and Kristof (1969) and Gruvaeus (1970) for the case $\mathbf{T} = \mathbf{C}^{-1}$. However, all problems may be tackled uniformly by proceeding as if the directions of all but one of the axes are known and then estimating the optimal direction of the remaining axis; cycling through all the

axes iteratively until a stable solution is found for all the directions.

Rather than handling the indeterminacy of factor loadings as above, the modern tendency is to use an internal criterion of simple structure such as Varimax (Kaiser, 1958) that has no reference to an external target $\mathbf{X}_2$. Alternatively, $\mathbf{X}_2$ may be coded in binary form where $x_{ij} = 1$ indicates a desired high loading of the i th variable on the j th factor and $x_{ij} = 0$ indicates a low loading. This latter form remains a type of Procrustes problem but least squares is no longer appropriate and is replaced by choosing $\mathbf{T}$ to maximize the ratio of the sum of squares of the squared high loadings to the sum of squares of the squared low loadings (see Lawley and Maxwell 1971).

2. Variants and Later Developments

The problems outlined above are referred to as two-sets Procrustes problems because two configuration matrices $\mathbf{X}_1$ and $\mathbf{X}_2$ are involved. There are many variants:

- (1) alternative constraints on $\mathbf{T}$ such as $\mathbf{T}$ is a rectangular column, orthonormal, or is a permutation matrix;
- (2) two transformations: minimize $\|\mathbf{X}_1\mathbf{T}_1 - \mathbf{X}_2\mathbf{T}_2\|$;
- (3) isotropic scaling: minimize $\|\rho\mathbf{X}_1\mathbf{T} - \mathbf{X}_2\|$;
- (4) anisotropic scaling: minimize $\|\mathbf{X}_1\mathbf{S}_R\mathbf{T} - \mathbf{X}_2\|$, or $\|\mathbf{X}_1\mathbf{T}\mathbf{S}_R - \mathbf{X}_2\|$, or $\|\mathbf{S}_L\mathbf{X}_1\mathbf{T} - \mathbf{X}_2\|$, or $\|\mathbf{S}_L\mathbf{X}_1\mathbf{S}_R\mathbf{T} - \mathbf{X}_2\|$, or $\|\mathbf{S}_L\mathbf{X}_1\mathbf{T}\mathbf{S}_R - \mathbf{X}_2\|$, or $\|\mathbf{S}_L\mathbf{X}_1\mathbf{S}_{R1}\mathbf{T}\mathbf{S}_{R2} - \mathbf{X}_2\|$ for unknown diagonal $\mathbf{S}_R$ and $\mathbf{S}_L$;
- (5) replace least squares by e.g., L1-norm, inner-product trace ($\mathbf{X}_2\mathbf{X}_1\mathbf{T}$), robust estimation method;
- (6) two-sided Procrustes problems $\|\mathbf{R}\mathbf{X}_1\mathbf{T} - \mathbf{X}_2\|$ where $\mathbf{R}$ and $\mathbf{T}$ are to be estimated, usually under specified constraints;
- (7) weighted forms—minimize trace $\mathbf{N}(\mathbf{X}_1\mathbf{T}_1 - \mathbf{X}_2)'\mathbf{W}(\mathbf{X}_1\mathbf{T}_1 - \mathbf{X}_2\mathbf{T}_2)$ where $\mathbf{N}$ and $\mathbf{W}$ are given matrices, one of which may be ignored;
- (8) handling missing values.

Brief comments on the above list must suffice. It should be clear that $\mathbf{X}_1$ and $\mathbf{X}_2$ no longer represent factor loadings. Isotropic scaling allows for differences in size; the more esoteric versions of anisotropic scaling verge on the more pejorative interpretation of the Procrustes legend: real violence being done to the data. The product of two orthogonal transformation matrices is equivalent to one and the inner product and least-squares criteria then give the same result. Weights $\mathbf{N}$ are useful to cover replication or missing rows of $\mathbf{X}_1$ and need not be diagonal, though they usually are. Note that weights are given matrices whereas anisotropic scaling is expressed in terms of unknown diagonal matrices requiring estimation. However, sometimes, $\mathbf{W}$ may be a function of the unknown $\mathbf{T}$, as with oblique Procrustes analysis in a metric where $\mathbf{W}$ is $\mathbf{C}'\mathbf{C}$ or its inverse.

Changes in orientation are given by an orthogonal matrix $\mathbf{Q}$, which includes rotations and reflections, leaving intact the essential distance and angular properties of the configuration. A $p \times q$ orthonormal matrix $\mathbf{T} = \mathbf{P}$ with q columns represents a projection from p to q dimensions. Projections may be interesting:

- (1) because $\mathbf{X}_2$ has fewer dimensions than $\mathbf{X}_1$;
- (2) a match may be required in some specified number of dimensions, say, $q = 2$.

When $q < p$ we may still match in p dimensions by rotating in the higher dimensional space; e.g., we may rotate three points lying on a line to fit three points lying in a plane. All we have to do is to augment $\mathbf{X}_2$ by $p - q$ zero columns. That $\|\mathbf{X}_1\mathbf{P} - \mathbf{X}_2\|$ can be less than $\|\mathbf{X}_1\mathbf{Q} - \mathbf{X}_2\|$ (where $\mathbf{X}_2$ is augmented by zero columns) has been seen as a justification for preferring projections to rotations but this is not comparing like with like. I believe that projection Procrustes analysis is rarely justified.

Nowadays, applications of Procrustes analysis are more often than not more to do with matching configurations other than factor loadings (Commandeur 1990). For example, $\mathbf{X}_1$ and $\mathbf{X}_2$ may be the two sets of points given by different multidimensional scalings (MDS) of the same data, or they may refer to MDS solutions based on different proximity/distance matrices measured between the same p objects, perhaps based on different variables. Fruitful applications have been found in the food sciences, where $\mathbf{X}_1$ and $\mathbf{X}_2$ are derived from scores given by two judges on traits of several products (e.g., Dijksterhuis and Gower 1991/2). Applications in the biological sciences include the analysis of shape where the configurations refer to the directly measured coordinates of landmark features on organisms (see e.g., Bookstein 1991, Dryden and Mardia 1999). In shape analysis the matching of one configuration to another is often termed registration.

3. Generalizations

Given k configurations $\mathbf{X}_1, \mathbf{X}_2, \mathbf{X}_3, \dots, \mathbf{X}_k$ we may match them in pairs by doing the corresponding two-sets Procrustes analyses, minimizing $s_{ij} = \|\mathbf{X}_i\mathbf{T}_1 - \mathbf{X}_j\mathbf{T}_2\|$ for all pairs (i, j) . Each s_{ij} may then be assigned to a matrix $\mathbf{S}$ and analyzed by some form of MDS (see Gower 1971 for an example from physical anthropology).

Alternatively, we may minimize $\sum_{i < j}^k \|\mathbf{X}_i\mathbf{T}_i\mathbf{X}_j\mathbf{T}_j\|$, which for the least-squares criterion, rejecting the solution $\mathbf{T}_i = 0$, is equivalent to minimizing $\sum_{i=1}^k \|\mathbf{X}_i\mathbf{T}_i\mathbf{G}\|$ where $\mathbf{G} = \frac{1}{k} \sum_{i=1}^k \mathbf{X}_i\mathbf{T}_i$, often termed the group-average. Thus, generalized Procrustes analysis is a three-mode method of analysis. For some information on algorithms, see ten Berge (1977). To help interpretation, each dimension of the analysis may be summarized in an analysis of variance, partitioning

the total into terms for the group average and for departures from the group average. Some dimensions may be identified with noise and some with signal, further subdivided into displayed and undisplayed parts (Dijksterhuis and Gower 1991/2).

The concept of a group average plays an important part in several psychometric methods (see e.g., Individual Differences Scaling) as a useful summary of k data sets. Perhaps, for this reason, it has been suggested that maximizing the group average of Procrustes analysis should be a main objective of Procrustes analysis. For orthogonal $\mathbf{T}$, both approaches are equivalent, but for projections, maximizing the group average is not the same as minimizing the criteria given above. The situation gets very convoluted, especially when the variants listed in Sect. 2 are admitted. This is because when focusing on the group average, the transformations, isotropic, and anisotropic scaling may be regarded as being attached to $\mathbf{G}$ rather than to the $\mathbf{X}_i$, giving two versions of every variant.

The leading eigenvector of the matrix with (i, j) th element trace $(\mathbf{T}_i \mathbf{X}_i \mathbf{X}_j \mathbf{T}_j)$ estimate isotropic scaling factors and also when $\mathbf{T}_i = \mathbf{I}$, this matrix is the basis of the STATIS method. This and similar links between Procrustes analysis and other methods that involve a group average, are likely to be an area for research in the immediate future. Research relating to biological landmark data and shape analysis is currently active and will continue (see Kendall et al. 1999). Other research will focus on algorithms, the effects of estimation criteria and the difficult area of inference.

See also: Configurational Analysis; Factor Analysis and Latent Structure, Confirmatory; Factor Analysis and Latent Structure: IRT and Rasch Models; Factor Analysis and Latent Structure: Overview; Latent Structure and Causal Variables; Multidimensional Scaling in Psychology; Multivariate Analysis: Discrete Variables (Correspondence Models); Multivariate Analysis: Discrete Variables (Overview); Scaling: Correspondence Analysis; Scaling: Multidimensional

Bibliography

- Bookstein F L 1991 *Morphometric Tools for Landmark Data*. Cambridge University Press, Cambridge, UK
- Browne M W 1967 On oblique Procrustes rotation. *Psychometrika* **32**: 125–32
- Browne M W, Kristof W 1969 On oblique Procrustes rotation of a factor matrix to a specified pattern. *Psychometrika* **34**: 237–48
- Commandeur J J F 1991 *Matching Configurations*. DSWO Press, Leiden, The Netherlands
- Dijksterhuis G B, Gower J C 1991/2 The interpretation of generalised Procrustes analysis and allied methods. *Food Quality and Preference* **3**: 67–87
- Dryden I L, Mardia K V 1998 *Statistical Shape Analysis*. Wiley, Chichester, UK

- Gower J C 1971 Statistical methods for comparing different multivariate analyses of the same data. In: Hodson J R, Kendall D G, Tautu P (eds.) *Mathematics in the Archaeological and Historical Sciences*. Edinburgh University Press, Edinburgh, Scotland
- Gruvaeus G T 1970 A general approach to Procrustes pattern rotation. *Psychometrika* **35**: 493–505
- Hurley J R, Catell B B 1962 The Procrustes program: Producing direct rotation to test a hypothesized factor structure. *Behavioural Science* **7**: 258–62
- Kaiser H F 1958 The varimax criterion for analytic rotation in factor analysis. *Psychometrika* **23**: 187–200
- Kendall D G, Barden D, Carne T K, Le H 1999 *Shape and Shape Theory*. Wiley, Chichester, UK
- Lawley D N, Maxwell A E 1971 *Factor Analysis as a Statistical Method*. 2nd edn. Butterworths, London
- Mosier C I 1939 Determining a simple structure when loadings for certain tests are known. *Psychometrika* **4**: 149–62
- Berge J M F 1977 Orthogonal Procrustes rotation for two or more matrices. *Psychometrika* **42**: 267–76

J. G. Gower

Production Systems in Cognitive Psychology

A *production system* consists of a collection of if-then rules that together form an information-processing model of some cognitive task, or range of tasks. Production systems have some special properties that make them highly suited to modeling cognition. From their origins as models of problem solving, production systems have grown to become a major formalism for modeling cognitive skill and aspects of learning. Since the 1990s, they have become increasingly identified with the topic of integrated cognitive architectures.

1. What is a Production System?

A production system is a model of cognitive processing, consisting of a collection of rules (called *production rules*, or just *productions*). Each rule has two parts: a *condition* part and an *action* part. The meaning of the rule is that when the condition holds true, then the action is taken. The basic idea is best illustrated with a nonpsychological example. Consider the following simple production system, with just two rules, describing the behavior of a thermostat in controlling a heating system:

Rule 1: if temperature $< 20^\circ\text{C} \Rightarrow$ turn on heating.

Rule 2: if temperature $> 20^\circ\text{C} \Rightarrow$ turn off heating.

When the room temperature is below 20°C , the condition part of Rule 1 is true, so the thermostat takes the action specified by the rule and turns on the heating. (The rule is said to *fire*.) When the temperature is above 20°C , Rule 2 similarly fires and turns

off the heating. Taken together, the two rules specify a process which describes the behavior of the thermostat.

A production system serving as a cognitive model usually has many more than two rules, perhaps as many as thousands. The production system operates in a cyclic fashion. First, a rule whose conditions are satisfied is identified. Then, that rule is fired, i.e., its action or actions are carried out. Usually, the actions will change the current situation in some way, so that now a different rule has its conditions satisfied, and the cycle repeats. This process is referred to as the *recognize-act cycle*, because the condition parts of the rules can be regarded as forming an active network which monitors the current situation and recognizes when each individual rule is applicable, triggering the act of firing the rule.

1.1 Production System Architecture

A particular production system operates within a definite structure of memories and processes known as its *architecture*. (For clarification of the term 'architecture' used in that sense). The architecture of a production system includes at least two memories: the *production memory*, and the *data memory*.

The production memory holds the rules. It is taken to be a long-term memory, in that its contents persist over time. Indeed, usually it is assumed that a production rule, once acquired, is never lost from memory.

The data memory holds the dynamic information about the task being worked on. (In the research literature, this memory is often described as the 'working memory,' but that terminology is avoided here because of the possible confusion with the psychological notion of working memory, with which it has a close but not straightforward relationship, see *Working Memory, Psychology of*.) The data memory is taken to be a short-term memory whose contents disappear with time, indeed possibly during the processing of a task. It holds the information which is tested by the condition part of production rules. In other words, the conditions of production rules consist of tests or patterns which are matched against the contents of the data memory.

1.2 Conflict Resolution

In a production system with more than a small number of rules, it is almost inevitable that sometimes more than one production rule has its conditions satisfied. Such an occurrence is known as *conflict*. Some production system architectures simply fire, in parallel, all the rules whose conditions are satisfied. Most

architectures, however, insist that a single rule be chosen to fire.

The procedure for selecting a single rule to fire from those that have their conditions satisfied, is known as *conflict resolution*. Different production system architectures use different principles for conflict resolution, which take account of factors such as the recency of the data being matched against, the activation levels of the data, the complexity of the rule conditions, the order in which rules were acquired, and so on.

2. Origins in Problem Solving

Production systems were introduced into psychology by Newell and Simon (1972) in their study of human problem solving. There is a natural connection between production systems and the way that Newell and Simon analyse their data.

Newell and Simon (1972) recorded participants thinking aloud as they solved puzzles in logic, chess, or cryptarithmic (where digits have to be substituted systematically for letters to make a correct arithmetic problem). Such recordings are known as *verbal protocols* (see *Protocol Analysis in Psychology*). Newell and Simon use the protocols to reconstruct the problem-solving steps taken by the participant. They then look for regularities in the participant's behavior. If, for example, whenever the participant assigns a digit to a letter, he or she then scans for another place where the letter is used, it is natural to express that regularity in the form of a production-like rule: 'if digit newly assigned to letter, then scan for occurrence of letter.' With some further attention to detail, a collection of such rules can be treated as a production system, and it then becomes an empirical question as to how well the production system models the participant's problem solving.

3. Properties of Production Systems

A production system specifies a computational process, and is therefore a kind of programming language. However, production systems differ from ordinary computer programming languages in a number of ways.

(a) *Parallel and serial processing*. Production systems exhibit aspects of both parallel and serial processing, corresponding to the two phases of the recognize-act cycle. The recognize phase, i.e., the testing of conditions, occurs in parallel, across possibly hundreds of thousands of rules. In the act phase, a single rule is chosen for firing, so the actions are carried out one at a time.

(b) *Independence of the rules*. Each rule in a production system is potentially independent of all the other rules. Each rule can be written to be self-

contained, expressing an item of knowledge which the problem solver is postulated to know.

(c) *Interruptible and flexible control.* A production system is not forced to follow a fixed control structure, such as a rigid hierarchy of routines and subroutines. Because the processing at each step is determined by what the rule conditions 'recognize' in the data memory, the control can be very flexible. For the same reason, it is interruptible. If relevant information appears in the data memory, the production system can abandon its current line of processing and follow a new one.

These and other properties make production systems well suited to the modeling of cognitive performance and cognitive skill.

4. Models of Cognitive Skill

Production systems have been written to model many different cognitive skills. Although only a few researchers have actually written production system models, their influence, as with other forms of computer modeling, has been far beyond their number. Production systems have now become a medium of choice for theorizing about cognitive skill, especially for skills requiring a large body of knowledge. This section summarizes just a small sample of the published work.

4.1 Written Subtraction

A model of errors in children's written, multicolumn subtraction illustrates how some of the properties of production systems support their application as cognitive models. Young and O'Shea (1981) present a simple production system, of around a dozen rules, for performing subtraction correctly. They proceed to show how small perturbations of the model reproduce many of the errors commonly found in 10- to 12-year-old children's performance.

Several of the most common errors arise by simply omitting one or more rules from the production system. For example, if the rule is omitted that initiates borrowing when the lower digit is larger than the upper, then the model never borrows but instead just subtracts the smaller from the larger digit in each column—a mistake common among children. The implication is that some children have failed to acquire the appropriate rule. Other errors arise by adding erroneous rules, or erroneous versions of rules, to the production system. For example, children who usually subtract correctly but simply copy the lower digit when subtracting from zero can be modeled by adding the corresponding zero-pattern rule to the production system. Although inappropriate for subtraction, that rule is appropriate for addition, suggesting that some children have not acquired sufficiently distinctive conditions for the rule.

The ability to add and remove rules in this fashion and still leave a model which runs and faithfully

reflects children's behavior, depends critically upon several of the properties of production systems mentioned above. Young and O'Shea (1981) discuss the crucial role of the independence of the rules, of the flexible control structure, and of the particular conflict-resolution principles used in the model.

4.2 Reading

Reading is another practically important cognitive skill which has been modeled with production systems. Just and Carpenter (1987) present an ambitious model of the reading and comprehension of single sentences and connected text. Their production system provides an integrated model of the cognitive processes involved in reading, from word encoding and lexical access, through syntax and semantics, to textual understanding. Many aspects of the model are based upon data from participants' eye movements (see *Eye Movements in Reading*).

4.3 Knowledge-intensive Cognitive Skill

Production systems are well suited to building models that encode and deploy large amounts of knowledge. For this reason, they are frequently used in the construction of *expert systems*, a branch of applied artificial intelligence that developed in the 1970s at around the same time as production systems were making their appearance in psychology (see *Expert Systems in Cognitive Science*). On the same grounds, production systems are much used in psychological models of expert skill, such as medical diagnosis, chess playing (see *Chess Expertise, Cognitive Psychology of*), and many other domains of knowledge.

5. Learning and Development

The ability to add rules to a production system, mentioned in Sect. 4.1 in the context of subtraction, raises the possibility of modeling certain aspects of learning and cognitive development by 'growing' a production system, one rule at a time, to correspond to the gradual acquisition of a cognitive skill. Some of the earliest applications of production systems were in the field of learning and development (see Klahr et al. 1987). It continues to be an active area of research. Simon and Halford (1995), for example, report on several computational models of the mechanisms of developmental transition, among which production systems figure prominently.

6. The Emergence of Cognitive Architectures

Since the 1990s, production systems have become increasingly identified with the newly emerging topic of integrated cognitive architectures, of which the

main exemplars are SOAR (Newell 1990) and ACT (Anderson and Lebiere 1998) *Cognitive Theory: SOAR; Cognitive Theory: ACT*).

The primary advantage of locating a production system within a broader, more encompassing theory is that the architectural details of the production system are determined in a principled way, and fixed. In consequence, the architecture of the production system becomes theory-laden, no longer serving just as a convenient and expressive notation but also carrying part of the burden of explanation. This removes extraneous degrees of freedom from the modeler, thereby leading to stronger explanations and theories.

See also: Artificial Intelligence: Connectionist and Symbolic Approaches; Artificial Intelligence in Cognitive Science; Cognitive Theory: ACT; Cognitive Theory: SOAR; Connectionist Approaches; Connectionist Models of Concept Learning; Connectionist Models of Development; Connectionist Models of Language Processing; Discrete State Models of Information Processing; Information Processing Architectures: Fundamental Issues

Bibliography

- Anderson J R, Lebiere C 1998 *The Atomic Components of Thought*. Erlbaum, Mahwah, NJ
- Just M A, Carpenter P A 1987 *The Psychology of Reading and Language Comprehension*. Allyn and Bacon, Boston
- Klahr D, Langley P, Neches R (eds.) 1987 *Production System Models of Learning and Development*. MIT Press, Cambridge, MA
- Newell A 1990 *Unified Theories of Cognition*. Harvard University Press, Cambridge, MA
- Newell A, Simon H A 1972 *Human Problem Solving*. Prentice-Hall, Englewood Cliffs, NJ
- Simon T J, Halford G S (eds.) 1995 *Developing Cognitive Competence: New Approaches to Process Modeling*. Erlbaum, Hillsdale NJ
- Young R M, O'Shea T 1981 Errors in children's subtraction. *Cognitive Science* 5: 153–177

R. M. Young

Professionalization of Social and Behavioral Scientists: United States

This article focuses on social and behavioral scientists as professionals in the United States. It addresses the nature of these professions, the process of professionalization, the role of credentials, patterns in degree production, and the contours of the labor force in social and behavioral science (hereinafter called

social/related science) fields. In this context, professionalization refers to the processes through which individuals become socialized as professionals. Certification refers to the authorization or authentication of credentials. And, labor force refers to that portion of the population (in this case of these professions) that is economically active; that is, either employed or, if unemployed, available to work.

1. The Social and Behavioral Sciences as Professions

The social and behavioral sciences constitute more than a single profession. Disciplines like economics, political science, psychology, and sociology have their own specialized training, expertise, reward structures, associations, networks, and norms. As scientific fields dedicated to the study of human social phenomena, they share features in common. Yet, because of their arenas of knowledge, occupational cultures, collegial interactions, labor markets, and how they are perceived by others, they comprise different professional groups. Moreover, these professions are not the same as the academic profession—although many persons with advanced degrees in the social/related sciences (especially with doctoral degrees) are employed in colleges and universities and, as such, are also academic professionals.

2. The Process of Professionalization

Professionalization involves the development of skills, identities, norms, and values associated with becoming part of a professional group. Through this process, recruits to the social and behavioral sciences acquire both substantive and methodological knowledge and develop understandings of their roles that permit them to function as professionals in these fields. Also, by training newcomers, these professions seek to ensure that the work of their sciences will continue congruent with certain principles and practices.

Research on professionalization dates back to the 1950s. Early research focused on the medical (e.g., Becker et al. 1961, Merton et al. 1957) and legal (e.g., Lortie 1959, Warkov and Zelan 1965) professions. Scholars also began examining graduate education in relation to academic or scientific careers (e.g., Pease 1967, Wright 1964). As these and recent studies (see, e.g., Keith and Moore 1995) indicate, although professionalization continues during the early years of employment, graduate departments are the primary socialization sites. In departments, students learn what is expected and rewarded and, conversely, what constitutes unacceptable practices (e.g., Anderson et al. 1994).

Across scientific disciplines, graduate education aims to develop in students a level of scholarly

proficiency through formal (e.g., courses, qualifying exams, dissertations) and informal (e.g., attendance at scholarly meetings) training. A curriculum that initiates students progressively into more independent work and autonomous thinking helps develop professionals who can both master knowledge and produce original work (Sullivan 1991). Also important is a curriculum that examines the range of roles integral to being a professional scientist (e.g., teaching, engaging in peer review of others' work). Similarly, a curriculum needs to address what it means to be a professional in terms of oral and written communication, ethical conduct, and other career skills. Finally, quality mentoring is key (e.g., Keith and Moore 1995, Anderson et al. 1994). It makes a difference for students' productivity, belief in the norms of appropriate conduct, professional activity, self-confidence, and satisfaction with graduate school.

Through the professionalization process, graduate students also develop shared norms of science that transcend any one discipline. Merton (1973) saw these to be 'universalism'—research should be judged based on scientific merit; 'communalism'—findings should be made public; 'disinterestedness'—research should be done to advance knowledge; and 'organized skepticism'—the scientific community should suspend judgment until the evidence is in. While scholars debate the influence of these norms and the presence of countervailing norms (Zuckerman 1988, Ritzer and Walczak 1986, pp. 225–35), few would dispute that there is a culture of science and that socialization in graduate school aims to engender support for and adherence to its norms.

Professionalization is important as well to the development of expertise, patterns of activity, and values that are discipline-specific. Scientific disciplines differ as to their intellectual tasks, their methods of inquiry, and how their work is done and rewarded (e.g., Biglan 1973, Lodahl and Gordon 1972). For example, chemistry in comparison to sociology has a longstanding tradition of intensive collaboration on more tightly defined problems. Thus, chemistry students are more closely supervised, work in research teams, and are involved in joint publication; sociology students have more intermittent contact with advisors, work more independently, and may or may not engage in collaboration (see Anderson et al. 1994, Babchuk et al. 1999). Sociology students, however, work and publish more collaboratively than do anthropology students, who are in a field more influenced by single-investigator, ethnographic research (Babchuk et al. 1999).

Not unexpectedly, departments vary within and across fields of science in how they structure training programs and the quality of what they do. While successful recruits exit with the same formal credential (a doctoral degree), their actual experiences differentially prepare them to work productively and navigate successful careers. Above and beyond any effects of

department prestige on graduates' careers, the nature of graduate training has far reaching implications for professional development.

3. Credentials and Certification in the Social/Related Sciences

Credentialing generally refers to a system or process that establishes the status, legitimacy, or expertise of persons. With respect to individuals, credentials may be issued in the form of: (a) degrees, diplomas, or other certificates by educational organizations; (b) licenses regulated by states; or (c) certificates by professional associations. Institutions too are credentialed by governmental agencies (e.g., incorporation charters) or private accreditation processes (e.g., regional accreditation associations), which create standards for operation and attest to the quality of what is being offered (Freidson 1986, Chap. 4).

For academic and scientific professions, university-level training is key. In these professions, completion of a program of training culminating in the award of a degree is the most basic credential that attests to expertise. Credentials such as degrees formally bestow on individuals the right to practice a profession or use a professional title (Freidson 1986). The legitimacy of the degree, as Freidson pointed out, relies on the fact that educational institutions themselves are credentialed (i.e., accredited) by the state or, formally or informally, by professional associations.

In the social and behavioral sciences, the advanced degree awarded to an individual constitutes the basic professional credential. The masters degree provides a measure of stature for an individual as a professional in these sciences. While there is a growing market of interest for MA-level social/related scientists (especially outside of the academic workplace), the doctoral degree is still considered to be the full credential.

Psychology is the only social/related science discipline with a well-established process for certifying individuals beyond the advanced degree or for accrediting programs. To engage in clinical practice, psychologists need to be licensed by the states in which they intend to practice. Licensure in psychology requires demonstrating proficiency in designated areas and passing a standard examination created by the Association of State and Provincial Psychology Boards, an alliance of agencies across North America responsible for licensure and certification of psychologists. In terms of accreditation, the American Psychological Association (APA) has a widely used, voluntary process for accrediting doctoral programs in the specialty areas of clinical, counseling, or school psychology programs (or combinations thereof). Some applied or practice associations in other social/related science disciplines are initiating such efforts (e.g., in 1995, the Society for Applied Sociology joined with

the Sociological Practice Association), but none has substantially taken hold.

Despite intermittent calls within some social/related sciences, there has been no strong impulse to further attest to individuals' expertise through additional credentials or certification programs. The American Sociological Association (ASA), for example, established but quickly abandoned a certification program for social psychologists in the early 1960s due to concerns about their certification by APA. In 1986 and 1990, respectively, Ph.D. and MA certification programs were launched by the ASA but generated little interest. In 1992, the ASA's MA Certification Program Committee expressed doubts about continuing the Program, noting that certification of sociologists is handled through receipt of an MA or Ph.D. In 1998, the ASA Council terminated the Association's program.

In the end, the conferral of advanced degrees through the university seems to provide a sufficient credential across fields of science. For each social/related science discipline, only a small proportion of academic departments has the authority to prepare the next generations for careers in their profession. In 1997, approximately 240 of the 4,009 institutions of higher education awarded doctorates. Except for psychology, there are only 110–50 Ph.D.-conferring programs and 90–170 MA-only programs in the social sciences (psychology, with its clinical specialty, has approximately 435 doctoral- and 260 MA-only programs). Thus, while scientific professions generally do not control the market through elaborate mechanisms of certification and accreditation, the size and structure of their educational systems affect the supply of trained professionals.

4. Degree Production

4.1 BA Degrees

Undergraduate majors in social/related science fields constitute the potential pipeline of those who might be attracted to and eligible for becoming social and behavioral scientists. They are also an important 'client' group that creates demand for trained social/related scientists as faculty members in the academic workplace. And, finally, even with only a bachelor's degree, they are a sector of the future workforce that may ultimately be employed in jobs classified as social/related science occupations (ostensibly among those that do not require more advanced expertise or skills).

Compared to mathematical, physical, and biological sciences, undergraduates declare majors in the social/related sciences later in their academic careers. This is primarily attributable to the scarce attention given to the social/related sciences as fields of science

in the elementary and secondary school curricula. Therefore, many undergraduates have their first exposure to these fields as part of fulfilling their general education requirements.

Despite later exposure to these fields, the social and behavioral sciences confer a substantial number of BA degrees. In 1967, even excluding psychology, there were twice as many BA degrees in the social sciences (72,398) as in the next highest field of science and engineering (36,197 engineering BAs). By 1997, the biological/related sciences had more than doubled, but there continued to be more BA degrees conferred in the social sciences than in other fields (see S&ED 2000, Table 5).

The number of BA degrees also varies within the social/related sciences. A comparison of four disciplines—economics, political science, psychology, and sociology—makes the point. As shown in Table 1, in 1967 these fields attracted similar numbers of majors, but by 1997 there was much more dispersion. Most obvious is the substantial increase, almost four-fold, in the number of BA degrees in psychology between 1967 and 1997. Also, there are fluctuations in levels of interest in different disciplines at different points in time. Over the current period—1990 to 1997 (not shown in table), sociology and psychology had the largest proportional increase in BAs (55 and 38 percent, respectively) while political science and economics conferred a declining number of degrees (about –10 and –30 percent, respectively). These data are based on S&ED 2000, Tables 50, 52–4.

4.2 MA and Ph.D. Degrees

As with BA production, both MA and Ph.D. production has generally increased across fields of science since the late 1960s (see S&ED 2000, Tables 12, 19). For most social and behavioral sciences, substantial growth was evident between 1967 and 1972, followed by a dip in the 1980s, and continued growth in the 1990s. At the MA level, the social/related sciences continued to put more MA-trained persons into the labor force than did other science and engineering fields (although there are also a substantial number of MA engineers). At the Ph.D. level, despite growth, there were also differences over time among fields. In the most recent 5-year period, doctoral degree production in the social sciences, biological sciences, and engineering continued to increase, but physical sciences and mathematical/computer sciences declined or remained level.

Within the social/related sciences, there were also variations in Ph.D. production by discipline (see Table 1). Between 1967 and 1997, the number of doctoral degrees awarded in psychology was substantially higher than in other disciplines in absolute numbers and in growth. In 1997, psychology conferred about 3,600 doctoral degrees, about six times more than

Table 1
Degrees awarded in Social/Related Sciences, select years 1967–1997

	Year						
	1967	1972	1977	1982	1987	1992	1997
<i>Bachelor's Degrees</i>							
Economics	13,829	16,408	16,674	21,880	24,402	24,976	17,701
Political Science	19,332	30,677	30,972	31,630	33,876	47,214	37,807
Psychology	19,435	43,421	47,794	41,539	43,195	64,033	74,734
Sociology	17,751	35,630	25,008	16,333	12,359	19,644	24,750
<i>Master's Degrees</i>							
Economics	2,147	2,636	2,662	2,506	2,321	2,564	2,804
Political Science	3,326	5,573	10,162	10,210	9,933	10,767	13,368
Psychology	2,898	5,293	8,320	7,849	8,165	9,852	13,633
Sociology	1,193	1,987	1,851	1,183	970	1,379	1,752
<i>Doctoral Degrees</i>							
Economics	691	1,056	983	940	960	1,051	1,161
Political Science	501	911	879	709	647	804	970
Psychology	1,295	2,279	2,990	3,159	3,173	3,263	3,566
Sociology	331	639	725	568	449	512	598

Source: Data are from the Department of Education/National Center for Education Statistics, Integrated Postsecondary Education Data System Completions Survey (IPEDS), and National Science Foundation/Division of Science Resources Studies (NSF/SRS) Survey of Earned Doctorates (SED). Data are tabulated by NSF/SRS, and published in *Science and Engineering Degrees: 1966–97*. (NSF00–310, NSF/SRS), Tables 50, 52–4 (S&ED 2000)

sociology and three times more than economics (although about 35 percent of these were clinical doctorates). Because of a decline in clinical doctorates in recent years, psychology has grown only modestly over the 1990s (about 10 percent) in comparison to political science and sociology that have both increased their Ph.D. degree production by about 30 percent (based on S&ED 2000, Tables 50, 52–4).

4.3 Women and Minorities and the Ph.D. Degree

4.3.1 Gender. Since the 1960s, there has been a dramatic increase in women receiving Ph.D. degrees across all fields of science and engineering (see S&ED 2000, Tables 19, 23, 26, 35, 49–51). In 1997, 8,921 women received their Ph.D. degrees in contrast to 1,096 women in 1967. These numbers represent a substantial difference in the shares of Ph.Ds. awarded to women (32.8 percent in 1997 and 8.4 percent in 1967). Fields, however, vary considerably in the proportion of Ph.Ds. awarded to women, with the social/related sciences and the biological sciences conferring a much higher proportion than the physical sciences and engineering.

There are notable differences within the social and behavioral sciences as well (see S&ED 2000, Tables 50, 52–4). For example, in sociology and psychology, women received approximately 20 percent of the Ph.Ds. in 1967, and 55 and 66 percent, respectively, in 1997. In comparison, in economics and political

science, even with increasing proportions of women awarded Ph.Ds., in 1997 women still received no more than about 30 percent of these degrees.

4.3.2 Race/Ethnicity. The data on earned doctorates for racial and ethnic minorities were first reported in the 1970s. By 1977, across all fields of science and engineering, the number of persons of color receiving Ph.Ds. (about 750) was quite small (see SED 1997, Table 8). In 1977, proportions of minorities receiving Ph.Ds. ranged from 5 to 7 percent in the life sciences, physical sciences, social/related sciences, and engineering. While the proportions of persons of color receiving Ph.Ds. increased over time, by 1997, they were still quite low—ranging from 12 to 18 percent across fields of science.

Important differences are evident among fields of science in the percentage of doctorates awarded to racial/ethnic minority groups (see SED 1997, Table 8). In 1997, in the physical sciences, Asian-Americans received the highest proportion of Ph.Ds. at over 7 percent and blacks the lowest proportion at under 2 percent. The social sciences included much higher proportions of black and Hispanic Ph.Ds. Overall, across the social/related sciences, somewhat more than 5 percent of the Ph.Ds. were awarded to blacks, slightly less than 5 percent to Hispanics, and slightly less than 4 percent to Asian-Americans.

There are also differences within the social/related sciences (see S&ED 1991, Table 3, SED 1997, Table A-

Table 2

Percentage of scientists with degree training in field of occupation by employment in field of occupation, 1997

Degree	Computer/Math Scientists		Life/Related Scientists		Physical/Related Scientists		Social/Related Scientists		Engineering	
	%	N	%	N	%	N	%	N	%	N
BA	41.5	(722)	10.4	(885)	23.3	(382)	2.6	(2,316)	53.9	(1,380)
MA	51.8	(245)	29.0	(157)	40.3	(115)	21.6	(485)	58.8	(431)
Ph.D.	77.5	(37)	59.9	(163)	60.5	(122)	67.6	(161)	66.6	(98)

Source: National Science Foundation, Scientists and Engineering Statistical Data System (SESTAT 1997): Table A-3

*Number is rounded in thousands of scientists employed at highest degree attained (e.g., 2,316 is 2,315,500 social/related scientists employed with highest degree attained at BA level)

2). For example, in 1997, economics and sociology awarded 28 and 25 percent, respectively, of their Ph.Ds. to racial and ethnic minorities, but varied as to which minorities received Ph.Ds. In economics, 16 percent of the Ph.Ds. were awarded to Asian-Americans and 6 percent to African-Americans; in sociology, almost 10 percent were awarded to African-Americans and 6 percent to Asian-Americans. The overall small number of Ph.Ds. awarded to racial/ethnic minorities signals the need for more intentional outreach to *all* persons of color at earlier stages in education. Despite increases in awarding Ph.Ds., the proportion conferred to racial/ethnic minorities is still below their proportion in the US population.

5. The Social and Behavioral Science Labor Force

As noted in the introduction, the labor force includes those persons who are economically active—either employed, or, if unemployed, available for work. In 1998, according to the Current Population Survey (CPS) conducted by the US Census Bureau, approximately 137 million persons were active in the civilian workforce of the US, including 6 million unemployed persons. Social and behavioral science professionals constitute only a miniscule segment of this overall labor force (approximately one-half of 1 percent) if such professionals are defined as persons having an advanced degrees (an MA or Ph.D.) in a social/related science.

5.1 Employment Status of Social and Behavioral Scientists

The overall employment status of those with advanced degree training in the social and behavioral sciences is quite favorable. According to the National Science Foundation's Scientists and Engineering Statistical Data System (SESTAT), there are some 656,100 social/related science professionals in the labor force (494,100 and 162,000 at the MA and Ph.D. levels, respectively). Employment rates for these individuals

are strong. In 1997, only 0.9 percent of all Ph.Ds. and 1.8 percent of all MAs in the social/related sciences were unemployed (see SESTAT 1997, Table B-2).

Of the 160,500 employed Ph.Ds. with social and behavioral science training, the vast majority (87 percent) was employed full-time (see SESTAT 1997, Tables B-2 and B-5). Also, approximately 117,000 of these Ph.Ds. were employed in science and engineering occupations. This proportion (73 percent) is consistent with other fields of science (see SESTAT 1997, Table A-3). Also, as might be expected, most of these Ph.Ds. (108,000) were working directly in social/related science occupations.

5.2 Social and Behavioral Science Occupations and Professional Training

In terms of standard approaches, the labor force in the social/related sciences can be conceptualized as: (a) those persons *trained* in the social and behavioral sciences, or (b) those persons *working* in social and behavioral science *occupations*. In contrast to a focus on degree training *per se*, classification into occupational categories includes persons trained in other fields who are working in these areas but excludes persons with relevant training who are working in other occupational areas. For example, social/related science occupations would include persons with Ph.Ds. in business who are teaching sociology in a university but would exclude persons with Ph.Ds. in sociology managing a business (see National Science Foundation 1999).

SESTAT classifies persons into occupations in terms of both work and education. Based on the SESTAT surveys, in 1997, approximately 350,000 persons with a BA degree or higher were classified as being in a social/related science occupation. This occupational category includes jobs that vary widely as to the nature of the work and required expertise. As might be expected, most persons classified as working in this occupation received their highest degree in a social/related science (approximately 80 percent across all degree levels). Furthermore, having an advanced degree beyond the BA also matters: the vast majority of persons working in a job classified as a social/

related science occupation have an advanced degree (see SESTAT 1997, Table A-3).

The value of an advanced degree can further be seen by looking at the relationship between BA, MA, or Ph.D. training in a field and working in an occupation classified as being in that same field (i.e., 'in-field' occupation). As shown in Table 2, for all fields of science, persons at successively higher degree levels are more likely to be working in occupations classified as 'in-field.' In the social and behavioral sciences, 2.6 percent of those persons with BAs was working in a social/related science occupation compared to 21.6 percent and 67.6 percent, respectively, of those with MA and Ph.D. degrees. By the Ph.D. level, there is also far less variability across fields of science in the proportion of persons working in 'in-field' occupations.

5.3 Sectors of Employment

Examination of Ph.D.-level professionals in the social and behavioral sciences shows the centrality of the academic workplace in relation to other work sectors. Across scientific fields, more Ph.Ds. were employed by 4-year colleges and universities than by any other employment sector (in 1997, 43.4 percent for all degree fields and 47.4 percent for the social/related sciences). The next highest work sector for all Ph.D. fields was business at 30.5 percent overall, although not as high a proportion of social science/related scientists was employed by the private sector as in other fields (see SESTAT 1997, Table C-2).

While specific disciplines within the social and behavioral sciences followed this overall pattern, there are important differences. As shown in Table 3, in 1997, a substantially greater proportion of sociologists (72 percent) than psychologists (34 percent) were employed by 4-year colleges and universities. In contrast, greater proportions of psychologists and economists were employed in the business sector. Also, for psychology, a larger proportion of Ph.Ds. is self-employed (17 percent), ostensibly due to the sizable number of psychologists who are in clinical practice.

5.4 Women and Minorities in the Social/Related Science Labor Force

5.4.1 Gender. Based on the 1997 SESTAT surveys, of the 160,500 employed Ph.Ds. with social/related science training, 62 percent were male and 38 percent were female (see SESTAT 1997, Table C-2). As could be anticipated from changes in the gender composition of Ph.D. degrees conferred over the years, substantially more employed female Ph.Ds. received their degrees in recent years. Of those who received their degrees 20–4 years ago, approximately 70 percent were male. In contrast, of those who received

Ph.Ds. within the last 5 years, approximately 45 percent were male (see SESTAT 1997, Table C-10). This pattern holds across social/related science disciplines—from psychology that substantially increased in the production of female Ph.Ds. to economics that continued to produce more male Ph.Ds. (see SESTAT 1997, Table C-9).

Males and females with Ph.Ds. in the social/related sciences vary only modestly in terms of sectors of employment. Gender differences within disciplines are more striking than any overall trends. For example, a higher proportion of male than female economists was employed by 4-year colleges and universities (56 versus 49 percent) while a higher proportion of female than male economists was in the government sector (20 versus 12 percent) (see SESTAT 1997, Table C-2).

5.4.2 Race/Ethnicity. Of the 160,500 Ph.Ds. in the social and behavioral sciences employed in the workforce in 1997, the vast majority across social/related science disciplines was white (88 percent). The data on Ph.Ds. of color in the workforce are consistent with what is known about the production of minority Ph.Ds. since the mid-1970s (see SESTAT 1997, Table C-9). Of those who received their Ph.D. degrees 20–4 years ago, 92 percent were white and 2.9, 2.5, and 1.7 percent were black, Asian-American, and Hispanic, respectively. In comparison, of those receiving degrees within the past 5 years, 82 percent were white and 4.4, 9.5, and 3.7 percent were black, Asian-American, and Hispanic, respectively (see SESTAT 1997, Table C-11).

The 1997 data on employment sector reveal an important difference between minority and white Ph.Ds. in the social/related sciences (see SESTAT 1997, Table C-4). Higher proportions of Ph.Ds. of color were employed in 4-year colleges and universities than were their white counterparts (57, 56, and 54 percent, respectively, for blacks, Asian-American, and Hispanic Ph.Ds. compared to 46 percent for white Ph.Ds.). That higher proportions of minority Ph.Ds. are employed in higher education compared to white Ph.Ds. likely reflects strong efforts by the academic sector to diversify its workforce in the context of a very low number of Ph.Ds. of color being trained.

5.5 Current Patterns for New Ph.Ds.

Analysts examining the current employment picture for social/related scientists are variously optimistic, with some being quite favorable (Spalter-Roth et al. 2000) and others being more cautious (Siegfried and Stock 1999). A recent survey of graduates awarded their Ph.D. degrees between July 1996 and August 1997 provides important data. The survey focused on 14 scientific fields, including economics, political science, psychology, and sociology.

Table 3

Employed Social/Related Scientists in selected fields where highest degree attained is Ph.D. by employment sector, 1997

Employment Sector	Social/Related Scientists			
	Economists Percent	Political Scientists Percent	Psychologists Percent	Sociologists Percent
Universities and 4-year colleges	57.1	67.4	34.3	71.7
Other educational	0.6	3.1	5.6	3.9
Private not-for-profit	5.0	5.0	10.1	7.6
Private for-profit	16.7	9.5	21.9	6.2
Self-employed	2.2	3.6	16.9	3.4
Government	13.9	11.3	11.2	7.1
Total	**95.5	100.0	100.0	100.0
Total number*	20,080	15,820	79,320	13,230

Source: National Science Foundation, Survey of Doctoral Recipients (SDR 1997). Data are published in *Characteristics of Doctoral Scientists and Engineers in the United States 1997*, NSF00-308, NSF/SRS, Table 17

*Number is estimated number of persons employed at highest degree attained. Numbers may not add up to 100 percent due to rounding

** Total is not 100 percent because 4.5 percent of economists were classified in 'Other' employment sector (not shown on this table). Also, 0.1 percent of psychologists were so classified

As shown in Table 4, unemployment is generally quite low for newly exiting graduates. With the exception of chemistry and political science at 4.6 and 7 percent, respectively, other disciplines were all below 4 percent. Furthermore, Ph.Ds. in the social/related sciences fared very well. While compared to other scientists, a somewhat higher proportion of social and behavioral scientists were employed part-time (in particular political science and psychology), a lower proportion were in temporary positions. Most striking is the relatively high proportion of social/related science Ph.Ds. employed in the academy, with sociology being the highest at 83 percent. Also, these data suggest a favorable academic market for social/related scientists. Newly minted Ph.Ds. in economics, political science, psychology, and sociology were finding employment in the academy in proportions comparable to, if not higher than, their overall presence in this sector (compare Table 3).

6. Continuing Issues for the Social and Behavioral Science Professions

For at least the next 10 years, the academic market appears to be positive. Despite the complexities of making predictions (see, e.g., Breneman 1988, Smelser and Content 1980, Schuster 1992), Bowen and Sosa (1989) anticipated these trends. In particular, the picture for social/related scientists is consistent with their projections of an undersupply of Ph.Ds. due to such factors as high undergraduate enrollments, stable to diminished graduate enrollments, anticipated fac-

ulty retirements, and increases in nonacademic employment. Bown and Sosa projected more demand than supply through 2007 and believed this pattern would continue, though at a slower rate.

It is important to keep in mind that the academic marketplace is broad and undergoing a restructuring. In reality, it consists of multiple and overlapping markets (e.g., research universities, liberal arts colleges) that have different 'buyers' and 'sellers' (see Smelser and Content 1980, Schuster 1992) and that vary in prestige and mobility. Furthermore, these markets themselves are not fixed. The community college, for example, is currently the fastest growing sector of academic employment. Finally, potential changes in the structure of these markets (e.g., increases in student-faculty ratios, more use of part-time or contractual teaching appointments, increased use of distance learning) could affect Ph.D. shortages.

The market for social/related scientists outside of the academy also merits more attention. For certain disciplines, like economics and psychology, the proportions working in nonacademic settings are sizable. Furthermore, the data indicate that social/related scientists are increasingly located in these settings. A National Academy of Sciences report (1969) observed a slow but evident trend to greater nonacademic employment. Between 1968 and 1997, there was a 12 to 20 percent decrease in the proportion of sociologists, political scientists, economists, and psychologists employed in educational institutions. Consistent with commentators (e.g., Rosenfeld and Jones 1988, Schuster 1992) noting a 'pull' not a 'push' to these work sectors, the steady increases in nonacademic employment seem to occur irrespective of the tightness of academic markets.

Table 4

Employment status of recent Ph.D. graduates

Field of Study	Percent			
	Unemployed	Employed Part-time	Employed in temporary positions	Employed in academe
Biochemistry and Molecular Biology	4.0	2	79	73
Chemistry	4.6	2	58	57
Chemical Engineering	3.2	0.5	26	25
Computer Science	2.4	1	17	43
Earth and Space Sciences	3.9	8	53	53
Economics	2.3	5	19	59
Engineering	2.7	2	24	32
Mathematics	2.4	5	46	70
Microbiology	2.2	1	81	70
Physics	1.8	3	57	45
Physiology	2.7	2	75	82
Political Science	7.0	14	41	78
Psychology	0.6	24	25	37
Sociology	3.4	8	35	83

Source: Prepared by Commission on Professionals in Science and Technology (1998); Employment of recent doctoral graduates in science and engineering: Results of professional society surveys. www.cpst.org/web/site/pages/pubs/pdf_pubs/98data-1.pdf. Data on sociology were updated by the American Sociological Association based on a more complete response

As the social and behavioral sciences move into their second century as professions, the overall environment for these fields is currently strong. The challenges before them, however, are also many. They are at a juncture where their ambitions could enlarge and these professions could grow to address some of the most challenging issues of human behavior and society. How they organize themselves and their work and how they prepare their next generations for contributing knowledge and expertise will, more than any other factors, shape what these professions are and become (see *Infrastructure: Social/Behavioral Research (United States)*).

See also: Education and Employment; Education (Higher) and Gender; Education: Skill Training; Educational Institutions and Society; Educational Institutions, History of; Educational Productivity; History, Professionalization of; Knowledge, Education, and Skills, Geography of; Professionalization/Professions in History; Professions in Organizations; Professions, Sociology of

Bibliography

- Anderson M S, Seashore L K, Jason E 1994 Disciplinary and departmental effects on observations of faculty and graduate school misconduct. *Journal of Higher Education* **65**: 330–50
- Babchuk N, Bruce K, Peters G 1999 Collaboration in sociology and other scientific disciplines. *The American Sociologist* **30**: 5–21
- Becker H S, Geer B, Hughes E C, Strauss A L 1961 *Boys in White: Student Culture in Medical School*. University of Chicago Press, Chicago

- Biglan A 1973 Relationships between subject matter characteristics and the structure and output of university departments. *Journal of Applied Psychology* **57**: 204–13
- Bowen W G, Sosa J A 1989 *Prospects for Faculty in the Arts and Sciences*. Princeton University Press, Princeton, NJ
- Breneman D W 1988 Research on academic labor markets: Past and future. In: Breneman D, Youn T I K (eds.) *Academic Labor Markets and Careers*. Falmer Press, New York
- Commission on Professionals in Science and Technology 1998 Employment of recent doctoral graduates in science and engineering: Results of professional society surveys. [www.cpst.org/web/site/pages/pubs/pdf_pubs/98data-1.pdf]
- Freidson E 1986 *Professional Powers: A Study of the Institutionalization of Formal Knowledge*. University of Chicago Press, Chicago
- Keith B, Moore H A 1995 Training sociologists: An assessment of professional socialization and the emergence of career aspirations. *Teaching Sociology* **23**: 199–214
- Lodahl J B, Gordon G 1972 Structure of scientific fields and the functioning of university graduate departments. *American Sociological Review* **37**: 57–72
- Lortie D C 1959 Laymen to lawmen: Law school, careers, and professional socialization. *Harvard Educational Review* **29**: 352–69
- Merton R K 1973 The normative structure of science. In: Merton R K (ed.) *The Sociology of Science: Theoretical and Empirical Investigations*. University of Chicago Press, Chicago (First published Merton R K 1942 Science and technology in a democratic order. *Journal of Legal and Political Sociology* **1**: 115–26)
- Merton R K, Reader G G, Kendall P L 1957 *The Student-Physician*. Harvard University Press, Cambridge, MA
- National Academy of Sciences, Social Science Research Council 1969 *The Behavioral and Social Sciences: Outlook and Needs*. National Academy of Sciences Press, Washington, DC
- National Science Foundation, Division of Science Resources Studies 1999 Counting the S&E workforce—It's not that easy.

- Issue Brief-NSF-99-344, May 3, 1999. National Science Foundation, Washington, DC
- Pease J 1967 Faculty influence and professional participation of doctoral students. *Sociological Inquiry* 37: 63–70
- Ritzer G, Walczak D 1986 *Working: Conflict and Change*, 3rd edn. Prentice Hall, Englewood Cliffs, NJ
- Rosenfeld R A, Jones J A 1988 Exit and re-entry in higher education. In: Breneman D, Youn T I K (eds.) *Academic Labor Markets and Careers*. Falmer Press, New York
- Schuster J 1992 Academic labor markets. In: Clark B R, Neave G R (eds.) *The Encyclopedia of Higher Education, Vol. 3, Analytical Perspectives*. Pergamon Press, New York
- S&ED 1991 *Science and engineering doctorates: 1960–90. Detailed statistical tables, NSF-91-310 Final*. National Science Foundation, Washington, DC
- S&ED 2000 *Science and engineering degrees: 1966–97. Detailed statistical tables*. National Science Foundation, Washington, DC
- SESTAT 1997 *Characteristics of scientists and engineers in the United States: 1997 Detailed statistical tables*. National Science Foundation, Washington, DC
- Siegfried J J, Stock W A 1999 The labor market for new Ph.D. economists. *Journal of Economic Perspective* 13: 115–34
- Smelser N J, Content R 1980 *The Changing Academic Market: General Trends and a Berkeley Case Study*. University of California Press, Berkeley, CA
- Spalter-Roth R, Thomas J, Levine F J 2000 New doctorates in sociology: Professions inside and outside the academy. *Research Brief*, Vol. 1, no. 1. American Sociological Association, Washington, DC
- Sullivan T A 1991 Making the graduate curriculum explicit. *Teaching Sociology* 19: 408–13
- Survey of Doctoral Recipients 1997 *Characteristics of Doctoral Scientists and Engineers in the United States, 1997 Detailed statistical tables*. National Science Foundation, Washington, DC
- Survey of Earned Doctorates 1997 *Doctorate Recipients from United States Universities: Summary Report 1997*. National Science Foundation, Washington, DC
- Warkov S, Zelan J 1965 *Lawyers in the Making*. Aldine, Chicago
- Wright C R 1964 Success or failure in earning graduate degrees. *Sociology of Education* 38: 73–97
- Zuckerman H 1988 The sociology of science. In: Smelser N J (ed.) *Handbook of Sociology*. Sage, Newbury Park, CA

F. J. Levine

Professionalization/Professions in History

In modern historical scholarship, the professions include the occupations of doctors, lawyers, professors, and engineers, among others. Professions are characterized by special functions and structures, motives, representations, forms of knowledge, and sociocultural styles. Professionalization refers to processes affecting the social and symbolic construction of occupation and status. The history of the professions investigates the ways in which functional groups and status groups are regulated and institutionalized. It shows how meanings, functions, and structures de-

velop in social and temporal context and how professions, in the restricted sense of the term, differ from the status groups and occupations of entrepreneurs, merchants and clerks, skilled manual workers (crafts, *métiers*) and the so-called semi-profession. Professionalization involves the formation of an occupation, on one hand, and interrelated developments regarding the social division of labor, structures of authority, and sociocultural inequality, on the other. Historical research on professionalization concentrates on such issues as the distribution of scarce resources at a particular point in time, the social definition of behavioral prerogatives, and the regulation of central values and functions. It is concerned with the motives, interests, and strategies of actors who either promote or hinder processes of professionalization.

Recent historical research on professionalization has tended to oscillate between individualizing historical perspectives and systematic social science perspectives. The former emphasize the peculiarities of particular professions or the special features of the professions in a particular historical context. The latter concentrate on general features and tendencies. Historical and comparative research shows that, while the professions are often concerned with the same or similar functions, problems, and tasks, these may be interpreted in different ways and institutionalized in different contexts. Similarly, depending on the characteristics of a particular profession and on the particular historical circumstances, the paths, cycles, and types of professionalization and deprofessionalization may vary. By taking these points into consideration, scholars have relativized both teleological understandings of professionalization and static conceptions of the professions.

1. Definitions of Profession and Professionalization

Abstract and temporally undifferentiated concepts of the professions have only a heuristic value in historical research. Therefore, one usually begins with a historical explication of the concept and then asks how professionalization is articulated in a specific context with regard to the relevant dimensions in particular domains. The individual dimensions and processes are more or less closely linked in concrete historical cases. The result is that the professions and professionalism find expression in a graded series of variations. The professions are characterized ideally by the following series of interrelated dimensions. Capabilities and skills are justified scientifically or systematically. Professional knowledge is described in such terms as 'exclusive,' 'more profound,' 'inaccessible to lay persons,' and 'not easily understandable.' It is acquired in special institutions such as universities, professional schools, and internships with professionals. It forms the basis of the superiority of the professional over lay

persons and clients. Also included in this body of knowledge are rules and attitudes which govern its application in a way which promotes trust between the professions and their social environment. These include formal procedures, ritual, titles, a professional habitus of studiousness, collegiality, and altruism, and a general orientation toward the common good. The professions claim for themselves a specific style of earning a living that is characterized by the principle of remuneration 'befitting ones rank' and through specific forms of contract and recompense. The professions demand for themselves higher prestige. They develop specific organizations for collective integration, self-administration, self-control, and for the representation of their interests, such as the 'chartered professions' in England, the disciplinary councils and chambers on the European continent, and, more generally, professional associations of various kinds. The professions strive for 'autonomy.' Viewed historically, this means, first, exiting from relations of patronage with traditional secular and clerical elites and, second, the reduction of external control and dependence on the state, on the rich and powerful, on clients or beneficiaries, and on employers.

Professionalization is a social project, which articulates itself in discourses on science, performance, honor, public welfare, and society and, in the ideology of professionalism. It is intended to secure for professionals a high or mid-range status in the system of social and cultural inequality. The historical literature distinguishes between professionalization that is externally controlled and professionalization that is controlled by the occupational groups themselves. Similarly, it distinguishes between heteronomous and autonomous professions. In the classical research on the 'free professions,' especially in the Anglo-Saxon literature, professionalization is understood largely from the perspective of members of the occupational groups themselves. In recent research, however, the analytical concept of professionalization is used increasingly for professionals who are occupied in state and private administrations and enterprises as well (Burrage and Torstendahl 1990, Perkin 1989, Siegrist 1988). Previously, and especially on the European continent, scholarly interest in professions such as these was often guided by concepts such as 'higher education,' 'meritocracy,' 'bureaucracy,' and 'civil servants.'

The historical connection between occupational role and social status is articulated concretely in concepts such as the 'gentlemanly professions,' the 'learned professions,' or, more simply, the 'professions' in England and the United States; '*professions libérales*,' '*bourgeoisie à talents*,' '*capacités*' in France; '*gelehrte Stände*,' '*höhere Berufe*,' '*bürgerliche Berufe*,' '*mittelständische Berufe*,' '*Akademiker*,' and '*Bildungsbürgertum*' in the German-speaking lands of Europe; '*professioni maggiori*,' '*professioni nobili*,' and '*borghesia umanistica*' in Italy; 'intellectuals' in Western

Europe and '*intelligentsia*' in Eastern Europe (Siegrist 1988, 1996; Freidson 1986; Conze and Kocka 1985; Malatesta 1995; Charle 1996; McClelland et al. 1995). In the twentieth century, professionals express claims to membership in the middle classes through discourses on the 'loss of middle-classness' and the 'threat of proletarianization.' Under National Socialism and Fascism, the meaning of the 'free professions' was hollowed out, as professionals were redefined as militant 'thinking workers,' who took part in the 'struggle of the Volk and the master race.' In the communist world, the concept of the socialist intelligentsia implied the transcendence of class relations, the end of occupational privileges, and the subordination of professionals to the state and the party. Modern western concepts such as 'technocrat' and 'expert' define the role and position of professionals independently of social or economic class.

2. Conceptions and History of Professionalization and Professions

The history of the professions begins in the Middle Ages and in the early modern era with the autonomous corporate colleges of the higher professions (lawyers, physicians, university professors) and the lower professions such as surgeons and notaries (Prest 1986, Betri and Pastore 1997). The various professions have specific orders of knowledge. They enjoy privileges with regard to education, the control of access to the professions, the practice of the professions, and ethical monitoring. The specific way of earning a livelihood distinguishes the higher professions from commercial occupations and handicrafts. The institutional pattern and the mentalities of the old European corporate occupational culture varied according to the degree of the abstraction of knowledge and the social proximity to the nobility, the patriciate, and the urban bourgeoisie. As a rule, the right to practice a profession was limited to a specific territory.

The traditional culture of the professions came under pressure and began to dissolve in the eighteenth and early nineteenth centuries. In the authoritarian and bureaucratic societies of Europe, the reformed professions were ruled under the control of the state. In radical-liberal, democratic, and egalitarian societies such as France in the revolutionary era, the United States, and Switzerland in the mid-nineteenth century, the professions were largely deregulated or were established as free trades (Ramsey 1984; Siegrist 1988, 1996). The general orientation of the different professions varied between (a) loyalty to and dependence on the state, (b) a progressive-liberal worldview, and (c) the desire to defend or revitalize traditional professional cultures and corporate mentalities. Different professions displayed many similarities with regard to the level of education, the orientation toward science and scholarship, and the social position of their

members. Nevertheless the various occupational groups often competed with one another for cultural hegemony, social prestige, and political power.

2.1 Institutionalization of Professional Training

The specific way in which education and training are institutionalized plays a central role in processes of professionalization. The states of the European continent led the way in the creation of theoretically and scientifically based training programs and in the transfer of such programs to (reformed) universities and elite institutions of higher learning such as the French *Grandes Écoles*. On the other hand, it is also true that (complementary) training through practice survived and even retained a central importance (trainees in state agencies, mandatory internships with practitioners, or, for example, when the first professional years serve, in effect, as a sort of apprenticeship). In Anglo-Saxon countries, the transition from 'training on the job' or 'shop culture' to 'school culture' has intensified since the late nineteenth century (Lundgreen 1990). A comparison of the length of time required for professional training in different countries reveals significant differences, even for the same occupation. Some of these differences may be attributed to variations in the contents of training programs, while others depend on factors such as tradition, restrictive state policies affecting the labor market, or the exclusive strategies of the professions themselves.

2.2 The Institutionalization of Professions

Beginning in the eighteenth century, it is important to distinguish between the legal institutionalization of the professions by the state, on one hand, and the formalization and regulation that is intended and implemented by the occupational groups themselves, on the other. With regard to the latter, one must also distinguish among (a) the continual development of a premodern profession, which, in principle, was never dissolved (such as the English 'bar'); (b) the restructuring of professions, which had been temporarily deinstitutionalized (as was the case with physicians and lawyers in the United States); and (c) the formation of new professions by members of functional and occupational groups that first arose in the nineteenth century (especially engineers but also a number of other 'new' professions). All of these involve long processes, which typically culminate in accreditation through the state or the legislation.

In much of Europe, the state or the government or legislation took the initiative to regulate and institutionalize the professions (Siegrist 1996; Ramsey 1984; McClelland et al. 1995). The state, as the guarantor of the legal order, the promoter of culture, and as an

intervening force in society, needed professional services and provided the professions with knowledge, rights, and duties. In this way, the state distinguished the professions symbolically from the occupations and functional roles of business, commerce, and agriculture, which had been deregulated through the abolition of guilds and the liberalization of the trades. Henceforth, the professions included only those vocations that were occupied with central goods and values or with securing property, honor, health, public order, morality, and science. In this new configuration of the professions, the old European conceptions of profession, guild and vocation were combined with the ideas of progress, order, and freedom.

In contrast to liberal and democratic societies, monarchical, bureaucratic, and statist societies tend to define professionals as civil servants and to establish a specific hierarchy of the professions: at the pinnacle of this hierarchy are professionals in the civil service, who are followed by state-appointed professionals, then by highly qualified members of the 'free professions' and finally by those professionals whose training and practice are not prescribed exactly. In liberal societies, the hierarchy is less pronounced, and the free professions stand at the pinnacle (Siegrist 1996). Since the late nineteenth century, professionals have been increasingly employed in private enterprises such as industrial firms, banks, insurance companies, and research laboratories. The occupational and status groups of professionals who are employees develop their self-understanding with reference to the social types of the civil servant, the free professional, and the entrepreneur.

The professional and social types of the nineteenth century may be characterized as follows:

(a) Professionals who are 'civil servants' are to a great extent bound in the service of the king or the state. If necessary, the individual lawyer, physician, engineer, or professor must disregard professional criteria and follow the directives of his superior. Due, however, to the increasing recognition of professional competence and of the indefinite tenure of higher civil servants, specific ideas and forms of professional autonomy also come into play in the judiciary, in the administration, and in higher education. Professions such as the engineers who were educated in the elite technical colleges in France and then entered into the civil service developed a special *esprit de corps* (Schweitzer 1999);

(b) Members of 'state-appointed professions' are holders of public office. Professionals such as these enjoy a kind of protected market, but they are also monitored closely. Political appointment strengthens the authority of the expert in society but limits his autonomy, impedes the formation of a self-defined collective consciousness of the profession, and makes the organization of professional associations more difficult. Examples include the Prussian, Bavarian, and Austrian lawyers well into the late nineteenth

century, the French and Italian court lawyers (*avoués, procuratori*), the Prussian county physicians, and the community physicians of Italy. In the late nineteenth century, some of these state-appointed professions were able to emancipate themselves from state protection, which had come to be viewed as patronizing. Linking discourses on political freedom and the free market with discourses on professional autonomy, they expressed the desire to become 'truly free professions.' Since clients increasingly made similar demands, many states changed the corresponding laws and regulations in the last third of the nineteenth century;

(c) The 'free professions' are characterized by more elaborate forms of training and examination and distinguished titles. Credentials are derived mostly from the state, but the free professional must be able to adapt to one or more segments of the market for professional services. Minimal standards for both professional and ethical behavior are often regulated by law, at least in principle. In practice, such regulations are enforced either by professional courts of honor (disciplinary bodies with the authority of public law or executive boards of associations with informal authority) or by ordinary courts and clients;

(d) In those areas where professional functions are exercised in the form of free trade, the constraints of the market and the increasing authority of science and scholarship encourage informal processes of professionalization. Professionals are themselves motivated to improve their own qualifications. An elite defines professional qualifications, founds an association, and attempts to marginalize or discredit so-called semi-professional practitioners.

2.3 Cognitive and Emotional Orientations and Strategies

In the eighteenth and nineteenth centuries, the contents, forms, and meanings of professional knowledge changed. As was the case with other societal elites, professionals formulated new scientific discourses and new models for professional conduct. The physician achieved greater recognition as a scientifically and practically trained healer and health expert (Huerkamp 1985; Léonard 1981), while the lawyer came to be viewed as a legal expert and 'priest of justice,' who was responsible for the realization of the rule of law and for securing the rights, the property, and the honor of the citizens (Siegrist 1996). The engineer was the chief proponent of scientific and technical progress, and was long considered to be an expert for administration, supervision, and organization (Schweitzer 1999). The university professor and the teacher in the gymnasium or *lycée* were the experts for education, aesthetics, and morality. Finally, the Protestant minister was the educated theologian, whose legitimacy was no longer derived from faith and piety alone.

The degree of the systematization of knowledge increased over the course of the nineteenth century, but the superiority of the professions and the efficiency of professional conduct remained controversial for an extended period of time. This is evident in the debates over the competence of university-trained 'physicians,' practical 'surgeons,' 'herbalists,' 'quacks,' nuns, midwives, and orderlies, and, more generally, over 'professional knowledge' and the knowledge of 'lay' persons. The professions had enduring difficulties in their efforts to prove their greater efficiency and to legitimize their autonomy with reference to science. Therefore, they attempted to support their claims to professional autonomy with supplementary appeals to values and attitudes that they shared with most of their bourgeois clients and employers. The leading concepts in these efforts were independence, responsibility, impartiality, altruism, public service, acting 'in good faith,' and 'delicacy' in the sense of tact and discretion. These values and attitudes were often codified in laws, professional regulations, codes of ethics, and professional manuals. In many cases, such moral standards applied to the whole person of the professional, i.e., not only to his occupational activities but also to his behavior as a private person and citizen as well.

Compulsion and persuasion served to bridge the gap between professional culture and lay culture. Studies that conceive of professionalization as an element in the historical development of disciplinary processes and the growth of class domination emphasize the compulsive character of the social action of professionals. With the spread of medical, legal, technical, moralizing, and educational practices, professional culture as a combination of knowledge and power penetrated all aspects of modern life and thought (Goldstein 1984, 1987). The professions are said to have used their knowledge in order to secure for themselves social and economic advantages. On the other hand, empirical social and cultural historical studies, in which concrete interactions between experts and lay persons are investigated, often conclude that experts were hardly able to realize their cultural projects in a pure form. For quite some time, lay persons avoided contact with professionals by seeking advice with regard to health from pastors or priests and by seeking legal council from union secretaries.

The historical record shows that the professions that were close to the state or involved with legally regulated cases and procedures were, as a rule, more successful in compelling recognition of their competence. In contrast, clients who made a short visit to the doctor's office, to a lawyer, or to a priest were not exposed to experts to the same degree. On the whole, the free professions had to rely more heavily on persuasion. This explains why symbolic performances, clothing, decoration of work spaces, rituals, and rhetoric are such an essential part of professional culture.

2.4 Professions, Markets, and Economic Situation

The rise of the professions in the nineteenth and twentieth centuries was due largely to the increased and increasingly differentiated demand of private and public customers for goods and services. To a degree, professionals were able to stimulate this demand as well. They fulfilled a continually growing need for the explanation of the world, the formulation and attribution of meanings, for edification, and for entertainment. Through their influence, questions of law and property and of health and death were viewed less fatalistically. Many new opportunities arose for them in the construction trade, commerce, industry, banking and insurance, research, higher education, and also in agriculture. The professions profited from a general increase in prosperity and in the cultural needs of the middle classes. In the twentieth century, there was a rising demand for medical, social, and cultural services on the part of lower social strata, which had previously been neglected. Due to developments in the fields of education and insurance in the welfare state, and because of a general rise in income levels, these strata developed into a new mass market. The changing profile of the clients led to a revision of the traditional attitudes and practices of the professions.

The economic situation of the professions has at no time and in no place been homogeneous. The spectrum ranges from the modest conditions of the lower middle class (*petite bourgeoisie*) to those of the upper middle class (*haute bourgeoisie*). Income and wealth vary according to stages in career development. They depend upon individual abilities and performance, on regulations governing fees and honoraria, on conditions in the various markets, and on general economic cycles. In the nineteenth and early twentieth centuries, it was not uncommon for income to be augmented by supplementary, non-professional activities, by doweries and inheritance, by earnings from real estate, or by stocks and bonds. This was typical of the bourgeois and petit bourgeois milieus to which the professionals belonged through birth, marriage, or social relations. In the twentieth century, the income of professionals was still significantly higher than that of workers or than the national average. Due to increasing competition, to increasing recruitment from families of humbler background, and to economic and political crises, professionals have often become convinced that their heyday is behind them and that they are threatened with decline (Jarausch 1990). Under these circumstances, there is often an intensification of conflict within the professions between protectionists and those favoring a more aggressive market orientation.

Beginning in the late nineteenth century, as the model of the free professions continued to spread, all of the professions became increasingly dependent upon the market. In reaction to this, there were many

debates among professionals over the 'overcrowding' of the market and the closure of the various professions. There was a general trend toward lengthening training programs and making them more difficult, in order to stem the influx of young aspirants. The pattern of discrimination against female candidates was widespread. In principle, women had been admitted to higher education since c. 1900, but they were effectively hindered in practicing the classical professions for quite some time (Jacobi 1994). In Central and Eastern Europe, members of religious and ethnic 'minorities,' who had been emancipated in the nineteenth century, still experienced discrimination when they attempted to gain admission to the professions. Due to their supposed 'overrepresentation' in universities and academic occupations, Jews were often made responsible for the worsening of the working conditions and the standard of living among professionals (Jarausch 1990; McClelland et al. 1995). It was not until the second half of the twentieth century that discrimination on the basis of religion, ethnicity, and gender gradually came to be forbidden.

3. Particularism Versus Universalism

Depending on varying perspectives among both contemporaries and historians, the modern history of the professions may be understood in terms of the nationalization of culture and society or in terms of the universalization of knowledge and of the culture of professionalism. In the process of nation-building, members of the professions became elites, which supported the development of national society, politics, science, and culture. The national system provided them with opportunities and assigned rights and duties to them. As functional and cultural elites of the nation, they had a common interest in maintaining the interconnection of power, status, knowledge, and earnings. In both the public arena and in more exclusive spheres, they represented themselves through professional and bourgeois discourses and symbols; these included historical accounts, myths, and *lieux de mémoire* as well as cultural artifacts such as busts, statues, and portraits of famous professionals in public buildings. Professionals have tended to view their own structures, practices, and status as being nationally specific, even in those cases, where, in objective terms, there is a dominant international pattern or style, which is subject only to slight variations and nuances.

Comparative historical research and the history of institutional and cultural transfer has relativized national interpretations and led to significant revisions of national views of history. Such approaches have shown how states and national professions imitate and vary foreign symbol systems and institutions, when they are convinced that this will lead to an improvement in their performance and in their status in international competition (Ramsey 1984; Lundgreen

1990; Siegrist 1996; McClelland et al. 1995). In some cases, occupational patterns that were originally imposed by European empires were maintained after the end of imperial domination. French models for advocates and for engineers were imposed on other lands during the Napoleonic era; but subsequently they were adopted voluntarily, because they were found to be compelling. The Habsburg monarchy influenced the structure of the professions in Southeastern Europe. Prussian prototypes exerted a strong influence in the development of civil service and the professions in northern Central Europe. During the colonial period, the model of the English and French professions diffused worldwide, and they have retained their attractiveness in the postcolonial era as well. The Soviet model of the socialist intelligentsia spread after 1945 to the lands of the Eastern Bloc. In Europe, American patterns have competed with national ones at least since the 1980s.

Since the late nineteenth century, the European or universal character of knowledge and professional culture has often been celebrated at international congresses, but national patterns survived into the twentieth century. In Europe, these have only begun to weaken in recent decades, in the wake of liberalization, the integration of the European market, and globalization. This is a clear indication of the fundamental difference between the professions and other occupations, which were liberalized at an earlier date and more thoroughly.

4. *Historiography and Emphases in Current Research*

Traditional historiography tended to be focused on a single profession. It served primarily to affirm a tradition and to establish an identity, to further the understanding of one's own profession and even to celebrate it, to delimit the profession vis-à-vis other occupations and semi-professions, and to justify the status quo and lend support to demands made in the public arena (Burrage 1990; Siegrist 1996). The oldest genres in the historiography of the professions include works which defend traditions, privileges, and customs in the face of threatening political, social, and legal developments or, alternatively, polemical works by outsiders and civil servants who want to reform the professions. Since the nineteenth century, there has been a proliferation of biographies and memoirs of exemplary professionals, of *festschriften* on the occasion of anniversaries, celebrations, and congresses, of memoranda and surveys with historical aspects, and of learned historical treatises in encyclopedias on general developments and various special aspects of the professions. More ambitious historical monographs on particular professions first appeared in the late nineteenth century. The authors were typically either academic or leading members of the profession

in question with an interest in history. Until recently, academic historiography has only showed a moderate interest in the history of the professions. Prior to the 1960s, both amateur and academic historians were more interested in individual professions than in typical aspects of the various professions and rarely took the approaches and theses of the sociology of the professions into account.

Precursors of the systematic historical comparison (and sociology) of the professions include works which, beginning around 1800, compared local and regional occupational ordinances and occupational conditions with foreign equivalents. This sort of comparison and historical reflection was conducted in connection with the preparation of national laws pertaining to professions. Around 1900, representatives of the new discipline of sociology began to take abstract and systematic approaches to individual professions, to groups of professions, and to specific aspects of processes of professionalization. In the efforts of sociologists to discover the 'essence' of the professions or to develop a 'general theory' of the professions, historical aspects and cultural peculiarities were neglected. In the 1960s, research based on the sociology of class and social conflict and pioneer contributions to international comparison reintroduced a historical perspective to the study of the professions (see the historical overviews on the sociology of professions by Freidson 1986 and Burrage 1990). These works inspired the new historical research on the professions, in which the development of the professions is historicized and recontextualized, without, however, sacrificing the goals of systematizing, conceiving of historical types, and contributing to the formulation of a general theory. Today, the history of the professions is an established and dynamic field of social history, cultural history, the history of the sciences, and the history of education. Historical approaches to gender have shown why professionalization has, until very recently, remained a male phenomenon and how professionalization and gender have entered into specific kinds of symbiotic relationships.

See also: Bureaucracy and Bureaucratization; Bureaucratization and Bureaucracy, History of; Civil Service; Ethical Codes, Professional; Engineering Codes; Expertise, Acquisition of; Legal Professionalism; Medical Profession, The; Professions in Organizations; Professions, Sociology of

Bibliography

- Betri M L, Pastore A (eds.) 1997 *Avvocati, Medici, Ingegneri. Alle Origini Delle Professioni Moderne* (Secoli XVI–XIX). Clueb, Bologna, Italy
- Burrage M, Torstendahl R (eds.) 1990 *Professions in Theory and History*. Sage, London
- Burrage M 1990 The professions in sociology and history. In:

- Burrage M, Torstendahl R (eds.) *Professions in Theory and History*. Sage, London
- Charle C 1996 *Les Intellectuels en Europe au 19e Siècle*. Éditions du Seuil, Paris
- Conze W, Kocka J (eds.) 1985 *Bildungsbürgertum im 19. Jahrhundert* pt. I: Bildungssystem und Professionalisierung in internationalen Vergleichen. Klett-Cotta, Stuttgart, Germany
- Freidson E 1986 *Professional Powers. A Study of the Institutionalization of Formal Knowledge*. University of Chicago Press, Chicago
- Goldstein J 1984 Foucault among the sociologists. The 'disciplines' and the history of the professions. *History and Theory* 23: 170–92
- Goldstein J 1987 *Console and Classify. The French Psychiatric Profession in the Nineteenth Century*. Cambridge University Press, Cambridge, UK
- Huerkamp C 1985 *Der Aufstieg der Ärzte im 19. Jahrhundert. Vom gelehrten Stand zum professionellen Experten*. Vandenhoeck & Ruprecht, Göttingen, Germany
- Jacobi J (ed.) 1994 *Frauen zwischen Familie und Schule. Professionalisierungsstrategien bürgerlicher Frauen im internationalen Vergleich*. Böhlau, Köln, Germany
- Jarausch K H 1990 *The Unfree Professions. German Lawyers, Teachers and Engineers, 1900–1950*. Oxford University Press, New York
- Léonard J 1981 *La Médecine entre les Savoirs et les Pouvoirs. Histoire Intellectuelle et Politique de la Médecine Française au 19e Siècle*. Aubier, Paris
- Lundgreen P 1990 Engineering education in Europa and the USA 1750–1930. The rise to dominance of school culture in the engineering profession. *Annals of Science* 47: 33–75
- Malatesta M 1995 *Society and the Professions in Italy 1860–1914*. Cambridge University Press, Cambridge, UK
- McClelland C, Merl S, Siegrist H (eds.) 1995 *Professionen im modernen Osteuropa/Professions in Modern Eastern Europe*. Duncker & Humblot, Berlin, Germany
- Perkin H 1989 *The Rise of Professional Society in England Since 1880*. Routledge, London
- Prest W R 1986 *The Rise of the Barristers. A Social History of the English Bar 1590–1640*. Clarendon Press, Oxford, UK
- Ramsey M 1984 The politics of professional monopoly in nineteenth-century medicine. The French case and its rivals. In: Geison G L (ed.) *Professions and the French State 1700–1900*. Pennsylvania University Press, Philadelphia
- Schweitzer S 1999 Der Ingenieur. In: Frevert U, Haupt H-G (eds.) *Der Mensch des 19. Jahrhunderts*. Campus, Frankfurt am Main, Germany
- Siegrist H (ed.) 1988 *Bürgerliche Berufe. Zur Sozialgeschichte der freien und akademischen Berufe im internationalen Vergleich*. Vandenhoeck and Ruprecht, Göttingen, Germany
- Siegrist H 1996 *Advokat, Bürger und Staat. Sozialgeschichte der Rechtsanwälte in Deutschland, Italien und der Schweiz (18.–20. Jh.)*. Klostermann, Frankfurt am Main, Germany

H. Siegrist

Professions in Organizations

The study of professions in organizations is about the intersection of two areas of enquiry that have quite distinct histories, namely, the study of professions and the study of organizations, but that converge over the

course of those histories. There have been times when they have been related strongly and other times when they have gone their distinct ways. In fact, the study of professions in organizations has gone through three, overlapping, phases since the mid-twentieth century. The first phase, started in the 1950s and particularly strong in the 1960s, and still apparent today, was that of a critical examination of the role of professionals in bureaucratic organizations. The second phase, begun in the 1970s and again, continuing until today, critically examined the orthodox view of professionals and gave a more power-based, labor market analysis of their position and how it affected organizational labor markets. The third phase, which has its origins in the 1960s and 1970s, and blossomed in the 1980s and 1990s, emphasized the dynamic nature of professionals in organizations and acknowledged the variety of organizational locations. This article outlines these phases, dealing with the central concepts and findings, examining the questions that each one raises, and pointing the way to the next phase of research in this rich field.

1. Professionals and Bureaucracy—1950s Onwards

The origins of the initial discussion of the relationship between professionals and organizations lies in the sociology of the professions and the discussions of defining characteristics of a profession. There are a number of contributions, but one that is often seen as central is that of Greenwood (1957) who suggested that professions have: a basis in systematic theory; prolonged training and certification; recognized authority; community sanction and legitimation; professional culture and codes of ethics. In essence, he came up with a set of traits or components that, when present in their entirety, defined the complete profession. This trait approach was echoed in a variety of ways by Wilensky (1964) and Hickson and Thomas (1969). There were two thrusts to these discussions. The first was to establish the characteristics that make up a profession. The second was to determine empirically which occupations were 'truly' professional and which were 'partly' professional and which 'aspired' to the status of a profession.

Out of these discussions came a relatively agreed upon set of characteristics of a profession that had structural and value aspects. The structural aspects were such things as a full-time occupation, a training and certification system, a professional association, and an ethical code. But, more importantly for the study of professionals and organizations, was the extension of Greenwood's notion of a professional culture into a set of values which encompassed peer reference, vocation, public service, self-regulation, and autonomy. When this work in the sociology of the professions connected with the sociology of organi-

zations it was in terms of the relationship between the values held by professionals and those promulgated by the modern, bureaucratic organization. This relationship was seen primarily as antagonistic (cf. Hall 1968).

Through the 1950s and 1960s the great influence on the sociological study of organizations was Weber (1949). Weber's principal contribution to this field was his characterization of organizations in terms of the authority relations within them (systems of imperative coordination) and the ways in which such systems had evolved historically. Each ideal type of authority, charismatic, traditional, and rational-legal, had its associated organizational form. Weber saw the rational-legal authority system with its organizational form of bureaucracy as the dominant institution of modern society. The authority system is rational because means are designed expressly to achieve certain goals and it is legal because authority is exercised through an office with its associated rules and procedures. For Weber the bureaucratic organization was technically the most efficient form of organization possible.

In terms of Weber's evolutionary history of social systems, bureaucracies represent the final stage in depersonalization. Offices (jobs) are held by experts who are arranged in a hierarchy. Rules and procedures provide predictability and consistency. Information is recorded and stored. Personal and business affairs are separated. There is 'the methodical attainment of a definitely given and practical end by means of an increasingly precise calculation of means.' Starting in the economic sphere, bureaucracy is such a potent method of organizing that it becomes characteristic of all areas of society such as education, government, politics, religion, and so on. Much of organizational analysis of the 1950s, 1960s and 1970s was essentially concerned with establishing, extending, and critiquing Weber's notion of bureaucracy (e.g., Pugh et al. 1969).

As Abbott (1988) in his influential work has pointed out, this 1950s and 1960s view of professions had an underlying assumption of professionalization as a natural process; that is, a regular sequence by which an occupation moved to 'full' professionalism (Wilensky 1964). Essentially, this view, together with 'the search for bureaucracy' produced a unitary view of both professional occupations and organizations. Professionals, it was suggested, were socialized into occupations where the key values were autonomy, peer control, and vocation. For example, Hall (1968) operationalized professional values as the professional organization as a reference, a belief in service to the public, a belief in self-regulation, a sense of calling to the field, and feelings of autonomy. Bureaucracy was operationalized as hierarchy of authority, division of labor, rules, procedures, impersonality, and technical competence. However, his conclusion was 'an assumption of inherent conflict between the professional or the professional group and the employing organization appears to be unwarranted.'

Of course, both sides of this coin, the professional and the organization, could be, and were opened up to scrutiny. On the organizational side of the coin, Scott (1965) had put forward the notion of Autonomous and Heteronomous Professional Organizations. Pugh et al. (1969) distinguished between Full Bureaucracies, Workflow Bureaucracies, and Personnel Bureaucracies as well as Implicitly Structured organizations. The point of these taxonomies and typologies was to open up the possibility of there being organizational forms in which the values espoused by individual professionals would be embraced and protected, thus potentially transforming the nature of the debate beyond an inherent conflict.

On the other side of the coin, the changing nature of professions was being stressed together with their organizational locations. Two issues are important. First, there was a trend of increasing professionalism within bureaucratic organizations. Occupations were arising that claimed professional status, such as social workers, nurses, and managers. These were bureaucratically produced and located occupations. Second, Hastings and Hinings (1970) suggested, from their work on chartered accountants, that there were established professions that carried out tasks that were centrally located in bureaucratic organizations and that, as a result, there was little or no conflict between professional and bureaucratic values in these situations.

This 'opening up' of both organization and profession led to new developments in understanding professionals in organizations. However, this theme of potential conflict between certain forms of organization and professionalism continues today. For example Raelin (1991) has examined the management of professionals in organizational contexts, writing about 'the clash of cultures' when generic managers are responsible for professionals. In the UK, and Europe more generally, there has been a concern with 'The New Public Management' and its organizational impact on professionals employed in healthcare, government at all levels, and education. The concern or thrust is still that there are incompatibilities between professional ways of working and values and certain kinds of organizational principles and practices. It continues to be a fruitful area of study.

2. Professionals in the Labor Market—1970s Onward

The second phase, begun in the 1970s and again, continuing until today, critically examined the orthodox view of professionals and gave a more power-based, labor market analysis of their position and how it affected organizational labor markets. The basis for this was laid down in critiques of the 'professionals in organizations' approach in the previous decades. The previous research questions had been anchored firmly

in issues about the structure and values of professionalism with notions of an underlying process of professionalization arising from the nature of work in modern economies.

The problems that were raised continued the opening up of professional occupations away from a homogeneous view of them. The uneven nature of professionals was stressed; that is, they could not and should not be considered as a unitary set of occupations. Similarly, the rise of the bureaucratic, organizationally based professions both in the context of the state and the private sector was recognized more fully. Abbott (1988) calls this the 'monopoly school' which attributes developments in professionalization 'to a desire for dominance or authority' (p. 15). The occupational/professional project is to control the conditions of work (the labor process and the labor market) and to increase and maintain occupational status and power (Johnson 1972, Larson 1977, Friedson 1986).

The issues raised by this approach 'dislocate' the relationship between professionals and organizations both because it makes our understanding of professional occupations more complex, and also because it shifts the locus of debate away from values and structure to power and control. Indeed, the connection of this critique of the functions of professions is not as strongly related to the theme of professionals in organizations as the theories being critiqued. The emphasis becomes recentered on the nature of a profession and the professional project. In particular, the claim to professional status by an occupational group, for example, social workers, is seen as a way to control entry to the occupation and the demand for membership, and to establish an authoritative position within the labor market.

In a sense, the connection to organizations is primarily through an analysis of the rise of new professions and a continuation of the theme of 'the professionalization of everyone' (Wilensky 1964). Because a key question to be answered is why new occupations make the claim to be professional, that is, the professional project, the examination is, in part, one of organizationally based occupations. Thus, the issue of professionals in organizations is not one of 'traditional,' autonomous professionals coming to terms with the rise of the bureaucratic organization, but the establishment of functionally and bureaucratically derived new occupations as professional. The issue for the new professional in this organizational setting is to acquire authority and control the conditions of their occupation's existence.

Thus, the professional/organizational relationship is not one of adaptation between two different sets of structure and values, but of an occupational group that is functionally defined and organizationally embedded acquiring the legitimacy and dominance that comes from professional status. The professional/organizational relationship is one of power plays,

control of the labor process and fights over legitimacy that are played out both in the organizational setting and the institutional/regulatory setting within a particular society. This approach continues to be part of the discussion of professions in organizations and is to be particularly found in discussions of the healthcare sector (Friedson 1986, Larson 1977) and of the new public management. One of the issues that is raised by this stream of theorizing is that of the 'semiprofessions.' This was an attempt to distinguish between fully professionalized occupations and other occupations that aspired to professional status. These latter occupations tend to be primarily those that are organizationally produced and located and where the individual occupation members do not work as independent practitioners. Johnson (1972) sees this distinction (and the use of other descriptors such as 'quasi,' 'marginal,' or 'limited' professions) as an illustration of power relations between occupational groups. For him, professions retain positions of authority vis-à-vis the semiprofessions, whose claim to professional status is contested. As he puts it, 'the pervasiveness of the ideology (of professionalism) is not indicative of the extent of professionalism' (p. 59).

The theoretical and research issue of competition between groups for professional status continues to be an important topic. Several writers have observed the phenomenon of stratification within the professions. Abbott (1988) observes that interprofessional competition is a fundamental fact of professional life with jurisdictional disputes between occupational groups over claims to particular expertise and knowledge bases. And with more and more occupational groups that lay claim to such expertise and knowledge being strongly based in organizations, then the dynamic between professionals and organizations changes, on both sides.

3. Professionals in Professional Organizations—the 1980s Onwards

The third phase, which has its origins in the 1960s and 1970s, and blossomed in the 1980s and 1990s, emphasized the dynamic nature of professionals in organizations and acknowledged the variety of both professional and organizational locations. The issues of the variety in professional settings had been increasingly raised and, at the level of the professional system, Abbott (1988) has pointed out that there is continuous competition and jurisdictional disputes between professional groups as they dispute claims to expertise. Similarly, on the organizational side, the typologies and taxonomies of organizations both differentiated between types of bureaucracy and introduced new organizational forms.

This discussion has emerged in two different forms. The first is setting the discussion within the changes occurring in national and international economies. As

thinking about 'the new economy' and 'globalization' has developed (Aharoni and Nachum 2000), so two issues have been emphasized which bring together much of the previous theorizing about professions. On the one hand there is the rise of new occupations which claim professional status. For example, in the software development industry there are computer engineers, software engineers, and systems analysts who are highly educated but operate in paraprofessional environments across boundaries of time and space. Financial services have grown rapidly, as has the whole sector of business services, and along with these have come investment analysts, financial planners, and mergers and acquisitions specialists whose professional status is unclear although one sees the same processes as were written about in the 1960s and 1970s, namely those of educational qualifications, professional associations, examinations, and designations and the ideological claim to professional status.

On the other hand there is the rise of the 'knowledge-based organization' (Alvesson 1995) together with the development of concepts such as the flexible organization and the learning organization. The suggestion is that organizations of the new economy are increasingly different from the various kinds of bureaucracy of the old economy. One of the reasons for this is that they are more service-based and that those services require highly educated employees who, in value terms, are much more like the 'old' professions, being committed to autonomy, vocation, and collegiality. These new organizational forms are seen as much more friendly and hospitable to professionals. Also, it is suggested that these forms are spreading into old economy organizations as they respond to the rigors of global competition.

The second form of the discussion is a renewed interest in the professional service firm, for two different reasons, each connected to new organizations and new occupations. Peters (1992) and Alvesson (1995) have suggested that the professional service firm is an example of the flexible, knowledge-based organization; that there is a wealth of experience in these organizations from which the newly developing knowledge-based firms can learn. But, at the same time, with the explosion of knowledge-based services, and, in particular, the tremendous expansion of business services in OECD economies, there is much more competition over professional jurisdictions. If one takes the example of the 'Big 5' accounting firms (formerly the Big 8 and the Big 10), there has been a continuous expansion of services within these firms and more and more integration across services and across nations (Greenwood et al. 1998).

The same processes are at work within other professional jurisdictions and professional service firms. In law, for example, there are strong internationalization pressures. Also, there is much discussion over the multidisciplinary practice, allowing non-lawyers to have a controlling interest in law firms.

In 2000, the Law Society of New South Wales (calling itself 'the most progressive legal jurisdiction in the world') voted to allow multidisciplinary practices in law firms and also to allow them to incorporate.

Part of the process of defining these issues and also a result of them being articulated has been a rise in interest in the professional service firm (PSF). There were classic studies by Smigel (1964) and Montagna (1968) that examined the ways in which professionals operated in the professional service firm. But because of the concerns of theorists with issues of the rise of the bureaucratic organizational form and the response of professionals to that, and the concern with the monopoly critique, there was little or no interest in studying the professional service firm *per se*, even though it was the implicit, professionally hospitable setting contrasted with the bureaucratic form. It is this which has become the focus of much interest and research in the past two decades and in its own way it reflects issues of professionalization and bureaucratization, although the latter has been recast as managerialism and located in the context of knowledge-based organizations and globalization.

The theoretical and empirical location of studies of professional organization has been, primarily, in either the professional association and system or nonprofessional organizations. Very little study had been done of professionals in professional organizations, organizations that can be described as PSF. One important dimension in understanding professions in organizations is the extent to which they are primarily located in professionally dominated organizations or in other kinds of organizations, the so-called bureaucratic professions. For example, practicing lawyers are employed primarily in law firms that are organized as professional partnerships; only a small proportion are employed outside legal practices; for example, as corporate counsel or as lawyers within the public defender justice system. Accountants tend to be split about 50/50 between private practice and employment in private and public sector organizations. However, almost all accountants are trained in private practice. Engineers are employed overwhelmingly in private and public sector organizations with only a small proportion of the profession being employed in engineering consulting firms. And engineers receive their postuniversity professional certification training in a wide variety of settings.

There are three points to be taken from this. First, professionals work in a variety of organizational settings, something that has been ignored in much previous work. Second, there are some professionals who primarily work in bureaucratic settings and experience little or no conflict. Hastings and Hinings (1970) showed that chartered accountants working in industrial settings experienced little role conflict and they argued that part of the activity of accountants is to bureaucratize others. Similarly, the debate over the semiprofessionals such as social workers, teachers,

and nurses centered around the fact that they were located entirely in bureaucratic organizations and did not have independent, autonomous practices. But, third, the arena of professional work that had been ignored in particular was that of the professional service firm in spite of the fact that implicit in many analyses of professionals, their location in independent professional organizational settings was considered to be very important as the archetypal setting for an autonomous professional practitioner.

Much of the work of the 1980s and 1990s has been concerned with understanding the nature of the professionally based, private practice organization (Nelson 1988, Brock et al. 1999). As this work developed it also embraced the theme of bureaucratization although in a different guise, as managerialism.

This initial work was concerned with elaborating the nature of the traditional professional service firm, which Greenwood et al. (1990) typified as the P² form. The P² stands for the two key elements of professionalism and partnership. The PSF is an organization whose core workers are professionals. Traditionally they are from one dominant profession such as accounting or law. And the governance system of the organization is a partnership where the senior members of the professional group jointly and severally take responsibility for the direction of the firm.

There are a number of organizational consequences to these twin defining features. In human resource terms, one is the 'up or out' system or tournament. Young professionals either make it to partner or leave the firm. Organizationally, they are relatively loose structures, representative and collective decision-making with management being something that is done after dealing with clients. Issues of management and methods of organizing are not issues that are seen as important to discuss in the P² firm. The emphasis is on the individual professional and their ability to do good professional work. There are clearly strong echoes in this of the kinds of discussions that have arisen in the 1980s and 1990s about the knowledge-based firm with its employment of highly qualified individuals who are applying new, and sometimes esoteric knowledge, and, it is suggested, require freedom and flexibility to achieve results. In the P² organization, interesting work and the freedom and flexibility to do it are prized.

However, the nature of the professional firm has been changing due to changes in the institutional and market contexts of professional work. Because of the development of the new economy and the rise of services generally and business services in particular, there have been increasing pressures on the established professions such as law, accounting, and medicine. For more than a decade now, Western economies have been in a period of deregulation with the ideology of the market being dominant. Thus, jurisdictional boundaries have blurred and the level of

inter-professional competition has increased. The rapid rise in the numbers of management consultants, financial services advisors, software development engineers, information technology consultants, media and marketing occupations, etc. has changed the dynamics of professional organizations. Similarly, there have been substantial changes in markets. On the one hand the demand for these professional business services has increased markedly, making it one of the most rapidly growing sectors in many economies. But the demands for these services are from increasingly sophisticated clients, operating on a global scale.

It has been suggested that a major consequence of these pressures and changes is that the professional, P² organizational form is in decline (Brock et al. 1999), and that it is being replaced by forms that place a greater emphasis on managerialism (Cooper et al. 1996). Cooper et al. (1996) named this new organizational form in professional firms the Managed Professional Business (MPB). In such an organization professionalism becomes centered on developing lines of business, hierarchies develop, strategic planning is introduced, targets are more binding, and decision-making rests in the hands of management teams rather than the partnership at large. So, the argument goes, the terms and conditions of work in the MPB move towards those of the regular corporation. However, this can only go so far because of the knowledge-based work of professionals (Rose and Hinings 1999).

What this suggests is convergence for professionals in organizations, but of a new kind. It is not convergence around either professionalism or managerialism (bureaucratization), *per se*, but around a combination of these which is organized around the increasing role of knowledge in 'postindustrial,' service-based organizations. As yet, we have not sufficiently tracked these kinds of developments to write authoritatively about them.

4. Conclusion

So what can we conclude about research in this rich field and the status of our current understanding? There are a number of things. A very important one is the continuing theme of the potential conflict between professionalism and bureaucratization and/or managerialism. That is, the notion of a professional encompasses values and cultural statements that are inimical to the forms and cultures of hierarchically organized, profit-driven corporations. However, this continuing theme has been modified in two important ways, both of which require much more study.

The first is the opening up of the idea of the professional. There is still a tendency to work with ideas of 'true,' 'full,' or 'traditional' professions that are seen as archetypal or at least worked with as ideal types. But, much of the original work on the professional-bureaucratic relationship, and, in particu-

lar, the ideas of the monopoly school, suggested that there are a range of professions exhibiting varying degrees of compatibility with bureaucratic working. More research is needed which starts from the *diversity* between professionals, rather than the assumption of homogeneity. There are still incompatibilities between some professional ways of working and their associated values and some organizational principles and practices. But whether those professions that have evolved from bureaucratic organizations have more compatibility with those forms is still very much an open question, and it needs to be connected to issues of claims to professional status and competition between occupational groups for professional status.

The issue that Abbott (1988) sees as central, namely, interprofessional competition and jurisdictional disputes between occupational groups, should be related to the types of organizations within and between which such competition and disputes take place. The healthcare setting is a particularly good one for examining these kinds of issues because of the numbers of professional groups involved.

A second modification to the professional–bureaucratic opposition thesis lies in the variety of organizational settings and their changing nature. While the death of bureaucracy is much over-stated, there is no doubt that new organizational forms are emerging that emphasize knowledge acquisition, use, and dissemination, together with flexibility and autonomy in organizational working. We might argue that the organic, flexible design is the ‘new’ organizational form. But there is a range of established and emerging forms within which there are differing responses to expertise and professionalism.

A further development of these themes comes from our increased understanding of the role and nature of the professional service firm. Within the 1990s more and more interest and research has been centered on, *inter alia*, accounting firms, law firms, management consulting firms, and engineering consulting firms. These are, of course, the archetypal sites for the study of professionals within settings which should be the most accommodating because they are organized around professional work and managed by the professionals themselves. But they are also operating in changing institutional and market environments and face intra- and interprofessional competition. More research needs to be done on the way in which these professional fields are being restructured, new occupational grouping emerging, new organizational forms developing, and the organizational–professional relationship being changed. Part of this is also that there are important differences *within* as well as *between* professionals in terms of organizational setting.

Finally, an important and intriguing topic to be examined is the possible convergence of professionals in organizations and organizations of professionals. That is, on the one hand, many ‘nonprofessional’

organizations are becoming more knowledge-based and are experimenting with more flexible and autonomy-based organizational forms. On the other hand, professional organizations are moving into new markets, facing more competition, changing their knowledge base, and growing in size. In doing this they are becoming more managed, more strategic, and possibly decreasing some aspects of autonomy and flexibility. What they can offer is challenging work with major clients. This suggests a new kind of convergence, combining professionalism and managerialism (bureaucratization), due to the increasing role of knowledge in ‘postindustrial,’ service-based new economies. These changes present a real challenge for research.

See also: Bureaucracy and Bureaucratization; Bureaucratization and Bureaucracy, History of; Organization: Overview; Organizational Culture; Organizational Culture, Anthropology of; Organizations, Sociology of; Professionalization/Professions in History; Professions, Sociology of; Weber, Max (1864–1920)

Bibliography

- Abbott A 1988 *The System of Professions*. University of Chicago Press, Chicago
- Aharoni Y, Nachum L 2000 *Globalization of Services: Some Implications for Theory and Practice*. Routledge, London
- Alvesson M 1995 *Management of Knowledge-intensive Companies*. De Gruyter, New York
- Brock D, Powell M, Hinings C R (eds.) 1999 *Restructuring the Professional Organization: Accounting, Healthcare and Law*. Routledge, London
- Cooper D J, Hinings C R, Greenwood R, Brown J L 1996 Sedimentation and transformation in organizational change: The case of Canadian law firms. *Organization Studies* 17: 623–47
- Friedson E 1986 *Professional Powers*. University of Chicago Press, Chicago
- Greenwood E 1957 Attributes of a profession. *Social Work* 2: 44–55
- Greenwood R, Hinings C R, Brown J L 1990 The P2-Form of strategic management: Corporate practices in the professional partnership. *Academy of Management Journal* 33: 725–55
- Greenwood R, Rose T, Cooper D J, Hinings C R, Brown J L 1998 The global management of professional services: The example of accounting. In: Clegg S, Ibarra E, Bueno L (eds.) *Theories of Management Process: Making Sense Through Difference*. Sage, Thousand Oaks, CA
- Hall R H 1968 Professionalization and bureaucratization. *American Sociological Review* 33: 92–104
- Hastings A, Hinings C R 1970 Role relations and value adaptation: A study of the professional accountant in industry. *Sociology* 4: 353–66
- Hickson D J, Thomas M 1969 Professionalization in Britain. *Sociology* 3: 37–54
- Johnson T 1967 *Professions and Power*. Macmillan, London
- Larson M S 1977 *The Rise of Professionalism*. University of California Press, Berkeley, CA

- Montagna P 1968 Professionalization and bureaucratization in large professional organizations. *American Journal of Sociology* 74: 138–45
- Nelson R 1988 *Partners with Power: The Social Transformation of the Large Law Firm*. University of California Press, Berkeley, CA
- Peters T 1992 *Liberation Management*. Macmillan, London
- Pugh D S, Hickson D J, Hinings C R 1969 An empirical taxonomy of structures of work organizations. *Administrative Science Quarterly* 14: 115–26
- Raelin J 1991 *The Clash of Cultures: Managers Managing Professionals*. Harvard Business School Press, Boston
- Rose T, Hinings C R 1999 Global clients' demands driving change in global business advisory firms. In: Brock D, Powell M, Hinings C R (eds.) *Restructuring the Professional Organization: Accounting, Healthcare and Law*. Routledge, London
- Scott W R 1965 Reactions to supervision in a heteronomous professional organization. *Administrative Science Quarterly* 10: 65–81
- Smigel E O 1964 *The Wall Street Lawyer*. Free Press, New York
- Weber M 1949 *Theory of Economic and Social Organization* (trans. Henderson A). Free Press, Glencoe, IL
- Wilensky H 1964 The professionalization of everyone? *American Journal of Sociology* 70: 137–58

C. R. Hinings

Professions, Sociology of

The sociology of professions is a branch of the sociology of work concerned with the analysis of expert occupations. It analyzes their patterns of organization, types of work, and social status. Because the sociology of professions first arose in the UK and the USA, it took its name and much of its early conceptual apparatus from the pattern of expert work characteristic of those countries. In recent years, however, the field has become more international, a change that is transforming both its empirical base and its theoretical superstructure.

1. Expert Occupations and the Sociology of Professions

The sociology of professions first arose to study the particular social form taken on by expert occupations in England and the USA in the course of the nineteenth century. This form was characterized by a combination of independent individual practice with some type of collective association. It usually involved specialized education, examination or licensure, service fees, and some form of autonomous discipline. By the early twentieth century, such a form was characteristic of medicine, law, accounting, architecture, and various other fields in the Anglophone world.

The first systematic analysis of these 'professions' was by Carr-Saunders and Wilson in 1933, who set

forth a typological analysis of occupations resting on various 'traits' that were taken to characterize the professions. The Carr-Saunders and Wilson analysis had a hard time with the military and the clergy, which were certainly 'professions' in the English sense of the term. Moreover, it made no effective contact with the quite different organization of expert work on the continent.

The French, for example, have no word with the denotation of the Anglo-American 'profession.' The French do distinguish *professions libérales*, but these include only experts in free individual practice, not all experts of a certain type. *Profession* itself has much broader purview than the English 'profession,' as does the parallel *métier*. Moreover, the organization of French experts has historically been far more state-dependent than has that of Anglo-American ones. Nor was France the only problem. The word and concept of profession were no closer to the German culture of expertism than to the French. Indeed, the characteristic pattern for organizing expert workers on the continent has more often been through quasi-civil service arrangements than through autonomous 'professional' ones. The sociology of professions thus began as a study of a particular kind of expert occupation. Only gradually have historical and comparative study made the limiting character of that approach clear.

2. Historical Development of the Field

From the Carr-Saunders book of the 1930s up into the 1970s, typology remained the dominant strategy of the sociology of professions. Millerson (1964) presented a comprehensive summary of all the various qualities that had been used to define 'professionhood.' The typological approach eventually strangled inquiry, as investigators worried whether or not various groups really were professions. Indeed, a substantial literature grew up in this period on the 'semiprofessions,' usually felt to include teaching, librarianship, nursing, and social work. These fit the typological model in some ways (advanced education, licensing) but missed it in others (independent practice). The so-called semiprofessions were all largely female occupations, a fact that would later become more emphasized, but engineering, although largely male, shared many of the same characteristics. Within typological thinking, salaried employment of professionals presented a problem. A small literature in the 1950s and 1960s considered this problem under the heading of 'professionals in bureaucracies,' viewing such professionals as being under strong role strain because of the conflicting demands of profession and organization.

The typological model enabled the sociology of professions to come to a rapprochement with structural functionalism. In a celebrated essay, Parsons (1954) argued that professions were rationalist (as

opposed to traditional), that they exercised functionally specific authority (as opposed to undifferentiated authority), and that they were universalist (as opposed to particularist). The later Parsonian tradition saw the professions as squarely anchoring the pattern maintenance sectors of crucial social subsystems. The typologists' focus on the organization rather than the content of professional work accorded better with the early Parsons, however, which kept them apart from the later structural functional interest in the professions.

Throughout the early postwar period, students of Everett Hughes and others in the Chicago tradition continued to amass case studies of professions and semiprofessions. Many of these fit the typological model badly. Their dependent nature often mocked the high role structural functionalists assigned to professions. One way around this problem was to argue that professionalization was a characteristic process that many occupations had not yet fully traversed. In the 1960s and 1970s, sociologists of professions argued for a characteristic professionalizing process, a sequence of stages through which all expert occupations developed. (Wilensky in 1964 provided the most famous analysis of such an argument.)

In the later 1960s, the sociology of professions, like many other subfields, turned towards skeptical interpretations. Major reinterpretations came from Johnson (1967), Freidson (1970) and Larson (1977). This new 'professional power' interpretation saw the professions in a different light. Freidson brought a generation of detailed studies of medical workplaces and medical knowledge together with a theory of professional organization to argue that the structures of professional organization were hardly the direct outcomes of rational and functionally specific universalism. The power of doctors over nurses suddenly seemed less a matter of advanced training than an enforced local division of labor. At a much more general level, Johnson recast professionalism as one among a number of modes of occupational control. This argument was carried further by Larson's reinterpretation of professions as collective mobility projects in which expert groups sought rewards through control of certain markets for services.

The rise of the power school coincided with a wide new interest in professions among historians, particularly in the USA. Scattered works had previously considered some major professions—English law and US medicine, for example. But the 1970s and early 1980s brought a flood of works covering various professions in the USA, England, and to a lesser extent France and Germany. These various works covered periods from the early twentieth century as far back as the seventeenth and eighteenth centuries. Although there were (and remain) important lacunae, theorists of professions began to work with a new empirical base. Perhaps more important, the historians turned

to the sociological literature for conceptual guidance and in many cases found little of use to them.

A number of other trends helped recast study of professions in the 1970s and 1980s. Most important was the steady drift of professionals away from independent practice and into corporate settings. On the one hand, this was simply a continuation of the trend towards larger and larger professional firms. In law, accounting, and architecture, sizable firms dated from the first half of the twentieth century, at least in the USA. Legal restrictions generally prevented this development both in the UK and on the continent. On the other hand was the development of multiprofessional firms, a pattern that had begun in architecture and now spread into areas like accounting, finance, and consulting. These largely professional firms were complemented by a much larger body of salaried professionals scattered throughout commercial and nonprofit organizations in the USA. By 1980, few US professions indeed were more than 50 percent made up of independent practitioners.

The rise of salaried professional work in the USA accomplished three things. First, it brought the US professional world much closer to the European pattern. The pattern of large-scale salaried employment by professionals had fascinated the French particularly, who had developed, in the concept of *cadres*, the notion of a technocratic intelligentsia as a new mode of occupational structure (see Boltanski 1982). Second, the new salaried professional work reawakened the interest in the 'professionals in bureaucracies' problem of role conflict. Third, the new pattern led to new concepts of 'deprofessionalization' and 'proletarianization of professionals.' Although the latter term obviously derived from the attempt to develop a Marxist analysis of professions, both concepts attempted to come to terms with the notion—an absurdity under the 'professionalization' concept—that occupations could lose the perquisites and possibly even the form of professions once these had been gained. A wide variety of studies appeared in the 1970s and 1980s examining the degree to which professions and professionals were losing control of work and cultural authority as well as enduring more mundane changes in salaries and tenure.

By the mid-1980s, then, the sociology of professions looked quite different. It had lost nearly all touch with structural functionalism. There was confusion over which professions would develop in which directions at which times. A large new body of research opened to theorizing not only a wide variety of quasi-professional occupations in the USA and England, but also the expert occupations of the continent. In many cases, this new data had extraordinary historical depth (see, e.g., the comparative work in Cocks and Jarausch 1990). In 1988, Abbott attempted to synthesize the entire area by refocusing the sociology of professions on actual professional work (both cultural and practical), turning away from the strong emphasis

on the social structures of professionalism that had become the center of the literature. He envisioned a world of interprofessional competition over jurisdiction of work. Changes flowed into this world both from within professions and from beyond them, recasting some professions, privileging others, destroying yet others. Professions attempted to claim control of work in the workplace, before the public, and within the state, and the course of professions' histories was determined by the ensemble of this melee of interaction. Abbott explicitly attacked the still-continuing typological tradition for eliding large areas of work. For him a profession was anything that competed like one.

Although Larson and others had attempted some international comparisons, Abbott's heavy reliance on European cases, particularly French ones, presaged what became the major development in studies of professions in the 1990s. This was the trend toward examining the relation of professions and the state. Much of the most interesting work came from Europe, for example, Jarausch's (1990) comparative analysis of lawyers, teachers, and engineers in Germany in the first half of the twentieth century. Krause (1994) has made the most general attempt at analyzing relations between the state and professions (and also capital), studying the USA, the UK, France, Germany, and Italy by means of case studies of law, medicine, engineering, and academics in each country. Krause's analysis focuses on the relative decline of professional autonomy and power in all these societies in the years after 1970. Both state and capital have increased their authority over professions.

The relation of state and professions has been underscored as an empirical question by the vast complications of creating standardized professional credentials in the European Union. The future of the sociology of professions will be greatly shaped by the theoretical developments necessary to analyse this process of internationalization of professional life. There is also a 'relations with capital' side to internationalization, encompassed in the largely unstudied complexities of the transnational corporations' use of professional services. There has been some study of 'offshoring' of professional services (e.g., of engineering services in telecommunications), but this particular capitalist attack on professional organization has not been studied in detail. In general, there is a startlingly small amount of work on expert occupations in the third world, with the possible exceptions of lawyers and, of course, the military.

3. *Current Problems*

The central present issue for the sociology of professions, then, is the question of how the professions will be reshaped by the economic reorganization that arose out of the crisis of the 1970s. Many of the recent

transformations have long antecedents, particularly the drift to salaried employment with its questioning of professions' cultural authority. Other trends, however, are new or newly important because of the economic conjuncture. Internationalization has been crucial. By the late 1980s the restrictions on professional firm size began to fall in Europe, largely under the onslaught of US firms, particularly law firms and multiprofessional consulting firms, which were the only firms able to manage some of the huge legal and financial projects characteristic of the new international economy. The new gigantism has transformed relations both within the professions, and, more important, between the professions and the state. It is in the international context, too, that the question of the profession-state relation is most clearly posed.

But there have been important internal effects to rethink as well. The decline of medical power from the pinnacle portrayed by Freidson has given new meaning to the phrase 'proletarianization of professionals.' This decline, driven by reorganizations of insurers and regulators, requires revising general theory about how expertise is organized in society. In that sense, the sociology of professions is again confronted with a functionalist question: how is expertise to be produced, controlled and reproduced in modern societies?

The issue of state and capitalist control is only one side of this question. On the other is the issue of commodification of professional knowledge. It is at present much too early to know the effects of computer-aided design systems, medical diagnosis systems, and other expert systems on the future of professions. But it is clear that new theory is required. It is also clear that commodification will be used to further reduce the status of professionals (see Abbott 1991).

Indeed, the field faces again the boundary question that has so long obsessed it. Earlier in this century, the boundary issue was 'what is a profession?' Given the new structure of employment relations, even in such highly professionalized surroundings as technological research firms and law firms, it is not clear what kinds of workers ought to be the subject of the sociology of professions. If the pattern of associational professionalism that gave rise to the field itself is either dead or transformed, then the field must retheorize its objects of scrutiny. Abbott's test—a profession is an occupation that competes by retheorizing others' work—presupposes fixed and organized occupations of a kind that simply may not exist under modern conditions of employment.

While the main focus of the professions literature has been on organizational and structural levels, some have insisted on a focus at the individual level, on the professionals themselves. The most general theme in this area has been a recurrent attempt to resolve the problem professionals (or expert workers) represent to Marxist theories of work. Originating in Djilas's book, *The New Class* (1957), the theme was continued in

Bell's *Coming of Post-Industrial Society* (1973), with its predictions of a knowledge world of the future. Attempting to synthesize this literature Brint (1994) recognizes the intense pressures brought on older structures of professionalism by modern economies. He sees a decline of 'social trustee professionalism' as professionals become mere private citizens with particular interests. Perkin's (1989) much broader analysis argues that the idea of profession became a model for modern societies generally in the middle fifty years of this century. The ideal of an individual with personal 'knowledge capital,' combined with the notion of just reward, was in Perkin's view fundamental to the conception of the modern welfare state. In that state's fall, argued Perkins, fell the professional ideal as well.

Perkins's argument provides a useful conclusion. In his article on 'professions' in the prior edition of this encyclopedia, Talcott Parsons (1968) closed with the confident assertion that 'the professional complex has already not only come into prominence but has even begun to dominate the contemporary scene in such a way as to render obsolescent the primacy of old issues of political authoritarianism and capitalistic exploitation.' Now that history has shown him wrong, the sociology of professions must develop a theory that can stand the test of major historical change.

See also: Bureaucracy and Bureaucratization; Doctor–Patient Interaction: Psychosocial Aspects; Law, Sociology of; Lawyers; Legal Professionalism; Medical Expertise, Cognitive Psychology of; Medical Profession, The; Medical Sociology; Organizations, Sociology of; Parsons, Talcott (1902–79); Professionalization/Professions in History

Bibliography

- Abbott A 1988 *The System of Professions*. University of Chicago Press, Chicago
- Albott A 1991 The future of professions. *Research in the Sociology of Organizations* 8: 17–42
- Bell D 1973 *The Coming of Post-Industrial Society*. Basic, New York
- Boltanski L 1982 *Les cadres*. Editions de Minuit, Paris
- Brint S 1994 *In an Age of Experts*. Princeton University Press, Princeton, NJ
- Carr-Saunders A P, Wilson P A 1933 *The Professions*. Oxford University Press, Oxford, UK
- Cocks G, Jarausch K H 1990 *German Professions*. Oxford University Press, New York
- Djilas M 1957 *The New Class*. Praeger, New York
- Freidson E 1970 *Profession of Medicine*. Dodd Mead, New York
- Jarausch K H 1990 *The Unfree Professions*. Oxford University Press, New York
- Johnson T J 1967 *Professions and Power*. Macmillan, London
- Krause E A 1994 *Death of the Guilds*. Yale, New Haven, CT
- Larson M S 1977 *The Rise of Professionalism*. University of California Press, Berkeley, CA

- Millerson G 1964 *The Qualifying Associations*. Routledge, London
- Parsons T 1954 The professions and social structure. *Essays in Sociological Theory*. Free Press, New York, pp. 34–49
- Parsons T 1968 Professions. In: Sills D L (ed.) *International Encyclopedia of the Social Sciences*. Macmillan, New York, Vol. 12, pp. 536–47
- Perkin H 1989 *The Rise of Professional Society*. Routledge, London
- Wilensky H 1964 The professionalization of everyone? *American Journal of Sociology* 70: 137–58

A. Abbott

Program Evaluation

Evaluations are, in a broad sense, concerned with the effectiveness of programs. While commonsense evaluation has a very long history, evaluation research which relies on scientific methods is a young discipline but has grown massively in recent years. Program evaluation has made important contributions to various social domains; for example, to education, health, economics, and the environment.

This article gives an introduction to program evaluation by dealing with the following three questions: What is program evaluation? What standards can be set to enhance quality and fairness of evaluation? How to conduct a program evaluation?

1. What is Program Evaluation?

1.1 Definition, Distinction from other Activities, Users, and Evaluators

Evaluation, broadly construed, has a very long history. For example, in ancient Rome, tax policies were altered in response to observed fluctuations in revenues, and many thousands of years ago, hunters developed complicated procedures for fishing and trapping game. In recent years, commonsense program evaluation has evolved into evaluation research which relies on scientific research methods. The article focuses on the latter type of evaluation (for introductory texts on evaluation research see, e.g., Berk and Rossi 1998, Fink 1995, Pawson and Tilley 1997, Rossi and Freeman 1993). In the article, the terms evaluation, program evaluation, and evaluation research are used synonymously. Concerns for the first large-scale national evaluations of social programs in the USA had its origins in the War of Poverty, and the way of monitoring programs is still evolving rapidly. Evaluation has proven its usefulness to the extent that

most federal and state agencies require evaluations of new programs.

Programs are the objects of evaluation, sometimes also called projects or interventions. Program evaluation derives from the idea that social programs should have demonstrable merits and explicit aims by which success or failure may be empirically judged; for example, literacy programs for high-school students should lead to measurable improvements in their reading skills. Introductory texts on evaluation describe programs as systematic activities that are provided on a continuing basis to achieve preplanned purposes; for example, a national program to train teachers to work in rural and underserved areas, a campaign to vaccinate all children in a school district (e.g., Fink 1995, JCSEE 1994).

Evaluation systematically investigates characteristics and merits of programs. It is the purpose of evaluation to provide information on the effectiveness of programs so as to optimize outcomes, quality, and efficiency. Program evaluations are conducted by applying many of the same methods that social researchers rely on to gather reliable and valid evidence. These include formulating the study questions, designing the evaluation, setting standards to assess the program impact, ongoing monitoring of the program functioning, collecting and analyzing data, assessing program effectiveness relative to the costs, and reporting results. For examples of evaluation in various fields of clinical and applied psychology see, for example, Bickman et al. (1995), Gueron and Pauly (1991), Hope and Foster (1992), and Spiel (2001).

While few authors (e.g., Scriven 1991) take the view that theories are a luxury for the evaluator, since they are not even essential for explanations, the majority of researchers called for injections of theory into every aspect of evaluation design and analysis (e.g., Pawson and Tilley 1997, Berk and Rossi 1998). This implies that program evaluation capitalizes on existing theory and empirical generalizations from the social sciences, and is, therefore, after all, applied research.

By relying on the scientific method, program evaluation differs from e.g., newspaper reporting and social commentary. Program evaluation differs from the social sciences in its goals and audience.

Evaluation goals are defined by the client; that is the individual, group, or organization that commissions the evaluator(s), for example, policymakers. However, often programs and clients do not have clear and consistent goals that can be evaluated for effectiveness. Therefore, to meet evaluation standards, program evaluation should be conducted by professional evaluators who are experts in social sciences, using a broad repertoire of concepts and methods, but who have also expertise in project and staff management, and have at least basic knowledge of the evaluated domain.

While impetus for evaluation rests in policy, information from evaluations is used by the *stakeholders*, all individuals or groups that may be involved

in or affected by a program evaluation. There are at least five groups (Fink 1995): (a) federal, state, and local government, for example, a city, the National Institute of Education; (b) program providers, for example, the healthcare services, the curriculum department in a school; (c) policymakers, for example, a subcommittee on human services of the state; (d) program funders, for example, the Office of Education; (e) social researchers, for example, in government, public agencies, and universities. Therefore, program evaluation is also an interdisciplinary task where evaluators and stakeholders work together (see Cousins and Earl 1992).

1.2 Key Concepts and Assignment Criteria

Evaluation can and should be done in all phases of a program. *Baseline data* collected before the start of the program are used to describe the current situation; for example, the need for an intervention. In addition, baseline data are used to monitor and explain changes. For example, if the medical emergency services attempt to reorganize their service to better care for children, they may collect interim data to answer questions such as: In what cases did the emergency medical services fail to meet the needs of critically ill children? What are the differences to the needs of adults? What is the current performance of medical staff?

Before the start of the program a *prospective evaluation* can be employed to determine the program's potential of realisation, its effectiveness, and impact; that is, the scope of its effects.

In *formative evaluation*, interim data are collected after the start of a program but before its conclusion. It is the purpose of formative evaluation to describe the progress of the program and, if necessary, to modify and optimize the program design.

A *process evaluation* is concerned with the extent to which planned activities are executed and, therefore, is nearly always useful. For example, process evaluation can accompany implementation of a community-oriented curriculum in medical schools.

Outcome evaluation deals with the question of whether programs achieve their goals. However, as mentioned before, goals are often stated vaguely and broadly and, therefore, *program effectiveness* is difficult to evaluate. In practice, the concept of effectiveness must always address the issue 'compared to what?' Often it is difficult to distinguish program effects from chance variation and from other forces affecting the outcome. Berk and Rossi (1998) distinguish between *marginal effectiveness*, *relative effectiveness*, and *cost effectiveness*. Marginal effectiveness concerns the dosage of an intervention measure. For example, to prove the assumption that a better student-teacher ratio would improve student performance, the class size has to be reduced. In this case,

two or more different student–teacher ratios are the doses of intervention and their outcomes are compared.

Relative effectiveness is evaluated by contrasting two and more programs, or a program and the absence of the program. For example, lecture-based learning can be contrasted with problem-based learning in their effects on students' practical performance. Cost-effectiveness considers effectiveness in monetary units. Evaluation usually is concerned with the costs of programs and their relationship to effects and benefits. For example, in the evaluation of a physical education program, one can ask how much gain in monetary units can be expected from the reduced use of health services.

However, the entire impact of a program, for example, the extent of its influence in other settings, is very difficult, maybe impossible to assess. Therefore, *summative evaluations* are applied, that is, historical reviews of programs that are performed after the programs have been in operation for some period of time.

2. What Standards Can be Set to Enhance Quality and Fairness of Evaluation?

Because of the increasing number of evaluations and the high impact that decisions based on results of evaluations may have, there has been since the early 1970s an ongoing discussion about guidelines for effective evaluations. In the absence of any clear definition of what constitutes a reasonable program evaluation, the Joint Committee on Standards for Educational Evaluation (JCSEE 1994), in cooperation with evaluation and research specialists, compiled knowledge about program evaluation and established a systematic progress for developing, testing, and publishing evaluation standards. The Joint Committee began its work in 1975. In 1989 the progress for developing standards was adopted by the American National Standards Institute (ANSI), and became available worldwide.

Originally, the standards were developed to guide the design, implementation, and assessment of evaluations of educational programs, projects, and materials. However, in the new *Program Evaluation Standards*, published in 1994, new illustrations were included that feature applications of the standards in such other settings as medicine, nursing, the military, business, law, government, and social service agencies (JCSEE 1994). The standards are published for people who commission and conduct evaluations as well as for people who use the results of evaluations.

The Joint Committee defined an *evaluation standard* as a principle mutually agreed on by people engaged in the professional practice of evaluation, that, if met, will enhance the quality and fairness of an evaluation (JCSEE 1994). The standards provide advice on how

to judge the adequacy of evaluation activities, but they do not present specific criteria for such judgments. The reason is that there is no such thing as a routine evaluation and the standards encourage the use of a variety of evaluation methods.

The 30 standards are organized around the four important attributes of sound and fair program evaluation: *utility*, *feasibility*, *propriety*, and *accuracy* (JCSEE 1994).

Utility Standards guide evaluations so that they will be informative, timely, and influential. They require evaluators to keep in any phase of the evaluation the stakeholders' needs in mind. Seven standards are included in this category; for example, Stakeholder Identification, Evaluator Credibility, Report Clarity, and Evaluation Impact.

Feasibility Standards call for evaluations to be realistic, prudent, diplomatic, and economical. They recognize that evaluations usually are conducted in natural settings. The three standards included in this category are Practical Procedures, Political Viability, and Cost Effectiveness.

Propriety Standards are intended to facilitate protection of the rights of individuals affected by an evaluation. They urge evaluators to act lawfully, scrupulously, and ethically. In this category eight standards are included; for example, Service Orientation, Formal Agreements, Disclosure of Findings, Fiscal Responsibility.

Accuracy Standards are intended to ensure that an evaluation will reveal and transmit accurate information about the program's merits. They determine whether an evaluation has produced sound information. The 12 accuracy standards include Program Documentation, Valid Information, Reliable Information, Justified Conclusions, and Metaevaluation.

Validity refers to the degree to which an information or measure assesses what it claims to assess, while *reliability* refers to the degree to which an information is free from 'measurement error'; that is, the exactness of the information. *Meta-evaluation* is evaluation of an evaluation.

For each standard the Joint Committee summarized guidelines for application and common errors. In addition, one or more illustrations of the standard's application are given (JCSEE 1994).

However, the Joint Committee acknowledges that standards are not all equally applicable in every evaluation. Professional evaluators must select to identify those that are most applicable in a particular context.

3. How to Conduct a Program Evaluation?

Program evaluation typically consists of a chronological sequence of activities or tasks. The main tasks or activities are (see, e.g., Fink 1995, JCSEE 1994): Posing questions about the program and defining evaluation

goals, setting standards of effectiveness, designing the evaluation, collecting and analyzing information, reporting the results. In addition, an evaluation plan must be worked out that specifies each of the tasks to be accomplished and the personnel, time, and resources needed.

3.1 Posing Evaluation Questions and Setting Standards of Effectiveness

Based on the program's goals and objectives, evaluation questions are formulated. While goals are often relatively general, e.g., to improve the organization of an institution, objectives refer to the specific purposes of a program; for example, to produce a guidebook about how to review the literature. Evaluation questions may also concern the magnitude, duration, and distribution of effects.

Because program evaluations should provide convincing evidence concerning the effectiveness of a program, specific effectiveness criteria are set. Because these *standards of effectiveness* determine the evaluation design, they should be set prior to any evaluation activity. Information about standards came from the literature, from experts, normative data, and statistical analyses (see, e.g., Donabedian 1982).

Consider, for example, the program goal to teach students how to review the literature (see Fink 1995). Then the evaluation question is, whether students learned to do this. The standard of effectiveness can set, for example, (a) that 90 percent of students learn to review the literature or, (b) that a statistically significant difference in learning is observed between students in Schools A and B, with students in School A participating in a new program and students in School B serving as controls. The decision for either standard (a) or (b) can have consequences for the evaluation design, the selection of participants, and the statistical analyses of the data.

3.2 Designing the Evaluation

An *evaluation design* is a structure created to produce an unbiased appraisal of a program's benefits. The decision for an evaluation design depends on the evaluation questions and the standards of effectiveness, but also on the resources available and on the degree of precision needed. Given the variety of research designs there is no single 'best way' to proceed (for an overview see, e.g., Cook and Campbell 1979, von Eye and Spiel 1996).

Essential criteria for evaluation designs are internal and external validity. *Internal validity* is attained when the evaluator can decide whether a finding is due to the program and cannot be caused by some other factors or biases. *Randomized experiments* are the method of choice to reach this criterion; for example, when participants are by chance assigned either to program A, to program B, or to no program. *External validity*

concerns the generalizability of findings; for example, to participants of the program in other places and at other times.

3.3 Collecting Information

For collecting information an arsenal of methods is available; for example, questionnaires, performance tests, various kinds of interviews, record reviews, observations, etc. (see, e.g., Kosecoff and Fink 1982, Rosenbaum 1995). When selecting the methods of data collection, the evaluator must consider the evaluation questions, the available technical and financial resources, and the time needed to develop a new instrument. A 'good' measure is, in common terms, one that is likely to measure accurately what it is supposed to measure (Berk and Rossi 1998). In methodological terms, a good measure is reliable and valid.

Mostly, more than one measure is used to obtain the relevant information. For example, when health-related behavior of 14-year-old students is the variable of interest, possible methods of data collection include surveys of students, surveys of parents and teachers, observations of students, and reviews of health and medical records (Fink 1995).

3.4 Analyzing and Reporting Data

Program evaluators and social scientists employ statistical methods to analyze and summarize data. However, evaluators have to keep in mind that statistical significance does not always imply practical significance or educational meaningfulness. Standards of effectiveness that have been set in advance of the evaluation (see above) should guarantee that analyses of the data provide also evidence about practical effectiveness of the program (see, e.g., Fink 1995).

To combine results of evaluations that address the same questions or goals, meta-analyses can be conducted. A *meta-analysis* is an analysis of evaluation data from several studies. It is a statistical procedure which attempts to provide estimates of program impact by combining the results from a number of distinct evaluations to increase statistical power and the generalizability of results (see, e.g., Rosenthal 1991).

Reporting results is the last task for the evaluator. The evaluation report answers the evaluation questions, describes how the answers were obtained, and translates the findings into conclusions and recommendations about the program.

4. Concluding Remarks

The demands for sound program evaluation continue to grow in a way that some authors (Pawson and Tilley 1997) fear may become mandatory that everything

needs evaluating. Therefore, it is essential that quality control procedures and standards be introduced for all facets and tasks of evaluation and that the evaluation itself is evaluated against pertinent standards. However, there is also the risk of oversimplifying and only using a checklist of standards. There is no fixed recipe of how to run a successful evaluation.

See also: Curriculum Evaluation; Educational Assessment: Major Developments; Educational Evaluation: Overview; Industrial–Organizational Psychology: Science and Practice; Job Design and Evaluation: Organizational Aspects; Performance Evaluation in Work Settings; Psychological Treatment, Effectiveness of

Bibliography

- Berk R A, Rossi P H 1998 *Thinking about Program Evaluation*. Sage, Thousand Oaks, CA
- Bickman L, Guthrie A R, Foster M, Lambert E W, Summerfelt W T, Breda C, Heflinger C 1995 *Managed Care in Mental Health: The Fort Bragg Experiment*. Plenum, New York
- Cook T, Campbell D 1979 *Quasi-experimentation*. McGraw-Hill, New York
- Cousins J B, Earl L M 1992 The case for participatory evaluation. *Educational Evaluation and Policy Analysis* **14**: 397–418
- Donabedian A 1982 *The Definition of Quality and Approaches to its Assessment*. Health Administration Press, Ann Arbor, MI
- Fink A 1995 *Evaluation for Education & Psychology*. Sage, Thousand Oaks, CA
- Gueron J, Pauly E 1991 *From Welfare to Work*. Sage, New York
- Hope T, Foster J 1992 Conflicting forces: Changing the dynamics of crime and community on 'problem' estates. *British Journal of Criminology* **32**: 488–504
- Joint Committee on Standards for Educational Evaluation (JCSEE) 1994 *The Program Evaluation Standards*, 2nd edn. Sage, Thousand Oaks, CA
- Kosecoff J, Fink A 1982 *Evaluation Basics*. Sage, Newbury Park, CA
- Pawson R, Tilley N 1997 *Realistic Evaluation*. Sage, London
- Rosenbaum P R 1995 *Observational Studies*. Springer-Verlag, New York
- Rosenthal R 1991 Meta-analysis: A review. *Psychosomatic Medicine* **53**: 247–71
- Rossi P H, Freeman H E 1993 *Evaluation. A systematic approach*, 5th edn. Sage, Newbury Park, CA
- Scriven M 1991 *Evaluation Thesaurus*, 4th edn. Sage, Newbury Park, CA
- Spiel C (ed.) 2001 *Evaluation universitärer Lehre—zwischen Qualitätsmanagement und Selbstzweck* [Evaluating University Teaching—Between Quality Management and End in Itself]. Waxmann, Münster, Germany
- von Eye A, Spiel C 1996 Research methodology: Human development. In: Tuijman A C (ed.) *International Encyclopedia of Adult Education and Training. Concepts, Theories, and Methods*, 2nd edn. Sect. 1. Pergamon Press, New York

C. Spiel

Progress: History of the Concept

The idea of progress was integral to the project of creating social sciences in eighteenth-century Europe, and the relationship has continued since. During the nineteenth century, the idea that 'civilization has moved, is moving, and will move in a desirable direction' (Bury 1932) was the historical framework and central problematic of social science theory and practice. The idea of progress as inevitable collapsed in the vicissitudes of the twentieth century, although in mitigated forms, it continued, waxing and waning with historical expectations. One component, the idea of the progress of the social sciences, figured prominently as the ideology of the social science disciplines.

1. Eighteenth-century Origins

The idea of progress and the idea of social science were centrally involved in each other's origin. Both emerged as part of the historicist understanding of history as a realm of human construction, propelled ever forward in time and taking new qualitative forms. Modern society was thus decisively different from its feudal and ancient forerunners, engaged upon a novel historical course. Replacing earlier cyclical and supernatural views of historical time, progress instantiated in human history the teleological movement that governed the Christian drama of redemption. Progress was a universal journey, and the array of human differences sharpened for European thinkers the progressive character of their own society and the stages by which they had reached modernity (Meek 1976). The emergence of the social sciences was itself a lynchpin of progress. Science was understood as the most advanced form of reason and the extension of the progress of science into the realm of government and morals—barely begun—was central to the expectation of civilizational advance. By the same token, the task of the social sciences was to understand the forces that were propelling the modern European world and to guide it into the future.

The theories of progress propounded in the eighteenth century drew upon, and were often coincident with, the early formulations of social science. Associationist psychology explained the ability of human mental capacities to improve over time, while the studies of historians and economists and the observations of travelers set the trajectory of progress toward greater complexity and refinement: with increasing social differentiation and intercourse, came improving manners, morals, knowledge, and rationality. The guiding role of providence in this process was still visible, but it was understood to work through nature. Particularly in the UK and France, natural principles assured the existence and continuation of progress.

Several major theories of progress emerged in this context. Scottish 'philosophical historians' like Adam Smith and Adam Ferguson developed four- or five-stage theories of history in which the economic mode of production at each stage shaped political organization, customs, and moral temper. The expansion of population, the division of labor, and technological advance nonetheless emerged within a complex social process, as history was often the result of unintended consequences. In contrast, the *philosophes* of the French Enlightenment like Condorcet believed reason to be the key factor in historical progress. The rational acts of enlightened individuals, and particularly of utilitarian science, moved history forward, while ignorance and war held it back. In Germany, Kant and Herder made historical progress the work of philosophical and moral reason. History was the progressive fulfillment of humanity's moral and spiritual destiny, realized through the unique development of each historical people.

Emerging in contested political arenas, the early theories of progress staked out ideological positions of longstanding influence in politics and the social sciences. Designedly, Smith's view lent itself to a liberal presumption against state action; Condorcet's view, to a justification of the authority of social scientists in the guidance of society; and Kant's view, to a liberalism subordinated to idealist ends. If the reliance on progressive reason assigned all evil to its reactionary enemies, the Scottish view recognized that progress involved secondary losses. Smith's critique of the diminished life of industrial labor began a longstanding discourse that made progress itself responsible for the problems of modern society.

2. The Nineteenth-century Triumph of Progress

The French Revolution raised the stakes for both progress and social science as the millennial hopes it aroused were projected into the secular future. Progress became the dominant view of history among the educated classes of Europe and North America, while contesting versions of progress proliferated. Responding to rapid capitalist development and the continuing possibility of political and social revolution, theorists of political economy, sociology, and anthropology recast the eighteenth-century theories of progress and sought to specify the factors that controlled the course of history.

The classical political economy that developed from Smith focused on the natural laws of economic growth and, of equal importance, on how the expanded product would be distributed. The question of whether the average workman would be better off, despite growing inequality, became urgent as egalitarian, utopian versions of progress appeared. The Reverend Thomas Malthus argued that food supply would keep

wages at subsistence level. His political economy, informed by an Anglican reading of history, supported a cautious view of progress, incorporating cyclical movement and the necessity of suffering. The theories of David Ricardo were often pressed into this mold, though most practicing economists, including Ricardo, maintained a more optimistic scenario of growth based on capital accumulation. By mid-century, recognition of the role of technology in increasing productivity shifted the weight of economic opinion toward more optimistic predictions of growth (Berg 1990).

American exceptionalists and socialist radicals transformed this liberal political economy in opposite directions. Henry C. Carey argued that the uncultivated American continent would abrogate altogether the action of Malthusian law, vindicating America's utopian historical course (Ross 1991). Socialists, charging the capitalist economy with the degradation of workmen and the destruction of community, made capitalism an intermediate stage, to be superseded by socialism. In Karl Marx's formulation, history progressed by a process of historical contradiction between the forces and social relations of production: capitalism produced within itself the revolutionary proletariat that would destroy it and the technological basis of abundance that would support socialism (Cohen 1978).

August Comte, inheriting Condorcet's identification of historical progress with progress of the human mind, made social science its central agent. Civilization moved by necessary law from the theological to the metaphysical to the ultimate stage, the scientific, still struggling to be born. Although this trajectory was inevitable, action in accord with historical law could ease the transition. A new science of society, sociology, would objectively examine the world as it is and, based thereon, set out a plan for the rational ordering of society, opinion, and feeling (Lenzer 1975). Herbert Spencer's version of sociology, influenced by liberal political economy and Scottish philosophical history, removed the central planning role from sociology, but gave it the same power to discern the laws of historical progress to which successful social action must conform. More complacent about his own era, Spencer believed that the industrial age would encourage moral, intellectual, and social improvement. Spencer turned historical progress into a process of social evolution, drawing on biological notions of adaptation to the environment, the inheritance of acquired characteristics, and an embryological model of differentiation to form a cosmic law of evolution (Peel 1971).

Charles Darwin's *The Origin of Species* (1859) reinforced both the idea of progress and the use of biological analogy. Despite the radically nonprogressivist implications of natural selection, Darwin's theory, with its suggestion of the continuity of animal and human life, was often read in this era of faith in

progress—even by Darwin himself—as a proof of the progressive character of both evolution and history. The contemporaneous recognition of the long age of the earth created a space for ‘prehistory,’ solidifying the uniquely advanced position of European civilization (Segal 2000). Social scientists developed a ‘comparative method’ that sorted peoples and races on an evolutionary grid, assigning them to stages of a single evolutionary process by Euro-centered hierarchical standards, although that gradient was always subject to Romantic subversion (Bowler 1989).

3. *The Problem of Progress, 1890–1914*

During the late nineteenth and early twentieth century, the social sciences formed specialized, university-based disciplines, in part to grapple with the accelerating problems of modernity. As industrialization, democracy, and imperialism transformed society, Enlightenment hopes for modernity took chastened forms. At the same time, more stringent conceptions of science outlawed deterministic laws of history like those of Comte and Spencer as metaphysical constructions and called attention to the moral valuation inherent in the concept. By the early twentieth century, the rediscovery of Mendelian genetics attenuated the link between evolutionary biology and historical change, and critics, notably anthropologist Franz Boas, attacked the comparative method and its assumption of a single line of evolutionary advance. Progress no longer appeared to leading social scientists as an inevitable, universal process, guaranteed by natural or historical law. Yet the question of what form, and how beneficial a form, modern society would take remained central as did many nineteenth-century terms of discussion.

In economics, marginalism shifted the focus of attention from economic growth to allocation. Those who joined the mixed neoclassical discourse that resulted grounded capitalism in laws analogous to the laws of physical nature and retained optimistic assumptions of economic growth and its central role in social advance, but those assumptions no longer framed their specialized and dehistoricized discourse. Other economists who sought a qualitative change in capitalism worked within a self-consciously historical or institutional framework, like the socialist Thorstein Veblen, who asked in 1899 ‘Why Is Economics Not an Evolutionary Science?’ (Ross 1991).

Sociologists narrowed their focus to the tendencies of modern society. In Germany, where antimodernist attitudes flourished, sociologists were most pessimistic. For Ferdinand Tönnies and Max Weber, traditional community (*Gemeinschaft*) and modern capitalist society (*Gesellschaft*) marked polar modes of social organization. The impersonality and loss of spirituality of modern society were inevitable con-

sequences of its organizing principles. Weber, more than Tönnies, saw the positive advantages of instrumental rationality, yet termed the framework it built for modern society an ‘iron cage’ (Liebersohn 1988).

Emile Durkheim, like American sociologists, retained the more hopeful nineteenth-century formulation that placed the present troubled society in transition to a more harmonious modernity. Durkheim (1933, [1893]) argued that a new kind of functional and normative social integration was forming; the anomie of individuals thrown outside the regulative norms of society by the upheavals of social transformation would subside as modern society reached equilibrium. Edward A. Ross posited a transition from the natural social control of face-to-face community to the more fragile, ‘artificial’ controls of modern society, similar to the German model of *Gemeinschaft/Gesellschaft*, but he too expected that a harmonious adjustment would be reached (Ross 1991).

It was primarily in thinking about their own role in modern society that social scientists returned to, and reaffirmed, the central role of science in progress. In his presidential address to the American Psychological Association in 1899, John Dewey expressed both the anxieties of modernity and the solution favored by his colleagues. ‘With tremendous increase in control of nature,’ he noted, ‘we find the actual realization of ends, the enjoyment of values growing unassured and precarious.’ But, he added ‘the entire problem is one of the development of science and of its application to life.’ When the social sciences are further developed, ‘we can anticipate no other outcome than increasing control in the ethical sphere’ comparable to ‘the revolution that has taken place in the control of physical nature’ (Dewey 1976, 1: 149–50). That science itself progressed, that it produced increasing control over nature and social life, and that such control worked to the benefit of humanity remained unquestioned. This faith served as the ideological basis for the social science disciplines.

4. *Progress Dethroned, 1914–1945*

The most severe challenges to the idea of progress occurred in the decades spanned by the two world wars: economic depression, totalitarian politics, and wars that unleashed mass destruction made possible by modern science and technology. At the same time, the intellectual and cultural changes that had undermined the idea of progress in the late nineteenth century gathered force after World War I. A sharp sense of historical discontinuity, breaking off the new world of mass production, mass consumption, and mass politics from the past, also broke the sense of cumulative gain in civilization on which progress depended. Historians mark World War I as the end of progress both as an unquestioned public faith and as

an idea that social scientists could presuppose or hope to demonstrate.

What remained open was the possibility of limited lines of progress, as human constructions, and the hope, still carried by the Marxist and liberal traditions and in the USA by an exceptionalist historical faith, that limited the gains worked toward the larger progress of history. Thus the American sociologist William F. Ogburn explained that social scientists had dropped the term progress and were instead studying 'social change,' as a 'term free from dogmatic or moral implications.' Social change is caused by inventions, Ogburn argued, but inventions occur unevenly, creating 'lags' until society 'adjusts.' In his analysis, however, adjustments can always be made and society was always better off when they were. Progress was not so much abandoned as forced underground (Ogburn 1930).

The point at which it remained still vibrantly alive, especially in the USA, was the faith in science. The new understanding of science itself suggested that if science did not mirror reality, it could more readily reconstruct reality to human purposes (Porter 1994). While Ogburn was moderate in his hopes for prediction and control, other American social scientists of the 1920s, like the behaviorist psychologist John B. Watson and the political scientist Charles Merriam, expressed utopian faith in science and technology (Ross 1991). A more sober tone emerged throughout the West in the 1930s, but the possibility of historical progress through science remained a central feature of American social science theory and ideology.

5. The Return of Progress, 1945–1970

In the decades after World War II, the vision of a progressive modern society guided by science gained energy and urgency from the defeat of fascism, the disintegration of colonial empires, and the threat of communism. As the strongest power to emerge from the war, the USA embraced that vision with renewed self-confidence. In what was experienced as the 'American Moment' of an 'American Century,' the USA already appeared to stand at the summit of world history ready to embody the values its exceptionalist history promised.

This static notion of historical time was mirrored in social science theories of functionalist, cybernetic, and equilibrium systems that nonetheless incorporated the idea of progress. The sense of progress achieved was embodied most visibly in modernization theory, which measured the distance the underdeveloped world still needed to traverse to achieve the American or Western norm (Latham 1999).

Modernization theory also marked one of the many points in these decades at which the social scientists' ideology of progress through social science entered into the substance of social science theory. Thus modernization was defined as a 'process of social

change that seeks to govern itself by rational policy planning' for which social scientists were 'indispensable' (Lerner 1968). Talcott Parsons suggested that academic intellectuals and the practical professionals they trained had become 'the most important single component in the structure of modern societies,' displacing from leadership the older political and newer capitalist class (Parsons 1968).

The possibility of progress, along with the liberal and Marxist traditions in which it was imbedded, revived in Western Europe with postwar reconstruction, although with far greater circumspection. In Germany, Jürgen Habermas blended native Marxist traditions with American theories of pragmatism and democracy to project not a utopian future, but an ideal against which modern society might measure its defects (A. Giddens on Habermas, in Skinner 1985). The dehumanizing character of modern society remained a powerful, and sometimes dominant, theme in German sociology.

The idea of progress and its drawbacks also played a conspicuous role in the development of the social sciences in the non-Western world. As an arm of secular, Western-style development, the social sciences had been since the nineteenth century a major site of debate over whether Western or native paths to modernity should be chosen. After World War II, the developing world also served as a major source of critical revision of Western conceptions of modernization, recasting modernization as economic dependency, or redefining tradition and modernity as mutually transformative rather than diametrically opposed processes.

6. Postmodernism and Progress

Political conflicts in Europe and the USA in the 1960s shattered the American Moment, opening the way for longstanding discontents with modernity and reconsideration of the idea of progress (Salomon 1955). The most radical challenge—the postmodern critique of modernity—emerged in the home of the rationalist Enlightenment. Francois Lyotard attacked the idea of progress, with its universalist and rationalist premises, as the basis of the most oppressive, totalitarian impulses of modern Western society. Michel Foucault pointed specifically to the professions and social sciences as the source of the knowledge/power that disciplined the unruly masses of modern society in the name of enlightened rationality. The pluralistic and libertarian values that animated this critique gained power from the increasing reach of Western capitalism into every area of the world and the proliferation of multi-ethnic societies. Globalization in the post-modern reading was producing not a single narrative of universal progress, but a pluralistic, hybrid world of many historical narratives. However, as the term post-modern suggests this vision can be read as a liberationist narrative of progress, drawn from Romantic,

rather than Enlightenment, strains of modernity (McGowan 1991).

This pluralistic vision had some influence on the social sciences as a support of feminism and multiculturalism within the USA and as a counter to Americanization elsewhere. The liberal, Western-centered idea of progress revived, however, on the strength of the free-market neoclassical paradigm in economics, with its presumption of capitalist-led progress, and a revival of the narrative of American exceptionalism after the fall of communism. Globalization in that context defined progress as the Americanization of the world.

The most fundamental aspect of the postmodern critique of progress, however, was its critique of scientific reason. Here, postmodern critics joined broader currents of philosophical criticism of positivism and analysis of the linguistic and social construction of knowledge. Although Thomas Kuhn argued that *The Structure of Scientific Revolutions* (1962) did not destroy the idea of scientific progress, it did radically remove the sciences from a universalistic narrative of the progressive discovery of truth and resituate them in a social-historical world of shifting purposes and uncertain directions (B. Barnes on Kuhn, in Skinner 1985). Most social scientists in the USA have not paid attention to the philosophical debates around this work. Indeed, in those segments of social science identified most closely with mathematics and natural science, like economics, older positivist understandings of science are especially powerful (Backhouse 1997). However, the context of critique has weakened the self-confidence, and with it, the ideological construction of the progress of social science, among other segments of the social science disciplines. While the eighteenth-century formulation of historicism as the story of progress gave the social sciences their charter, the critical historicism of the late twentieth century called for a reformulation of both the historical assumptions and the self-understanding of the social science disciplines.

See also: Comte, Auguste (1798–1857); Differentiation: Social; Enlightenment; Evolution, History of; Evolutionism, Including Social Darwinism; Functionalism, History of; Kuhn, Thomas S (1922–96); Modernity; Modernity: History of the Concept; Modernization and Modernity in History; Modernization, Sociological Theories of; Postmodernism: Philosophical Aspects; Rationalism; Spencer, Herbert (1820–1903)

Bibliography

- Backhouse R E 1997 *Truth and Progress in Economic Knowledge*. Edward Elgar, Cheltenham, UK
 Berg M 1990 Progress and providence in early nineteenth-century political economy. *Social History* 15: 365–75

- Bowler P J 1989 *The Invention of Progress: The Victorians and the Past*. Blackwell, Oxford, UK
 Bury J B 1932 *The Idea of Progress*. Macmillan, New York
 Cohen G A 1978 *Karl Marx's Theory of History*. Princeton University Press, Princeton NJ
 Dewey J 1976 Psychology and social practice. In: Boydston J (ed.) *John Dewey, The Middle Works*. Southern Illinois University Press, Carbondale, IL, Vol. 1
 Durkheim E 1933 [1893] *The Division of Labor in Society*. Macmillan, New York
International Encyclopedia of the Social Sciences 1968 Macmillan and The Free Press, New York
 Latham M E 1999 *Modernization as Ideology: American Social Science and 'Nation-building' in the Kennedy Era*. University of North Carolina Press, Chapel Hill, NC
 Lenzer G 1975 *Auguste Comte and Positivism*. Harper & Row, New York
 Lerner D 1968 Modernization: Social aspects. In: *International Encyclopedia of the Social Sciences*. Macmillan and the Free Press, New York
 Liebersohn H 1988 *Fate and Utopia in German Sociology, 1870–1923*. MIT Press, Cambridge, MA
 McGowan J 1991 *Postmodernism and its Critics*. Cornell University Press, Ithaca, NY
 Meek R L 1976 *Social Science and the Ignoble Savage*. Cambridge University Press, Cambridge, UK
 Ogburn W F 1930 Change, Social. In: *Encyclopaedia of the Social Sciences*, 3: 330–4. Macmillan, New York
 Parsons T 1968 Professions. In: *International Encyclopedia of the Social Sciences*. Macmillan and the Free Press, New York
 Peel J D Y 1971 *Herbert Spencer: The Evolution of a Sociologist*. Basic Books, New York
 Porter T M 1994 The death of the object: *Fin de siècle* philosophy of physics. In: Ross D (ed.) *Modernist Impulses in the Human Sciences 1870–1930*. Johns Hopkins University Press, Baltimore, MD
 Ross D 1991 *The Origins of American Social Science*. Cambridge University Press, Cambridge, UK
 Salomon A 1955 *The Tyranny of Progress*. Noonday Press, New York
 Segal D A 2000 'Western civ' and the staging of history in American higher education. *American Historical Review* 105: 770–805
 Skinner Q 1985 *The Return of Grand Theory in the Human Sciences*. Cambridge University Press, Cambridge, UK

D. Ross

Progressive Education Internationally

Progressive education (*éducation nouvelle*, *Reform-paedagogik*) is an international movement which originated in Western Europe and North America by 1900 in different national contexts. It can be seen as a reaction to the decisive educational fact of the nineteenth century that, with the introduction of compulsory schooling, education had been achieved for everyone in the younger generation, even the inhabitants of the smallest rural villages, even the inhabitants of the working-class areas of growing big cities.

Education thus institutionalized as a system for the masses showed its deficiencies. Progressive education tried to overcome these defects while maintaining education for all.

1. *Origins and International Spread*

Progressive education started in the UK with the 'new schools' such as Abbotsholme, founded by Cecil Reddie in 1889; Bedales, founded by John H. Badley in 1893; and King Alfred's School, set up by a group of parents in northwest London in 1898. A generation later, Alexander S. Neill, a former teacher at King Alfred's, who had been in contact with child-oriented education in Homer Lane's self-governed 'Little Commonwealth,' and who had grounded his experiences in psychoanalytical theory, founded Summerhill in 1921 as an experiment of antiauthoritarian education.

Progressive education in the USA is linked with the practical and theoretical activities of educators, such as the Columbia University teacher trainer William H. Kilpatrick with his famous project method; the school superintendent of Winnetka, Illinois, Carleton Washburne, who successfully modeled the schools of his district on progressive principles; the Harvard psychologist William James, who was influential on progressive pre- and in-service teacher training; Helen Parkhurst, who in her Dalton Plan developed an institutional model of individualized learning; Jane Addams of Hull House, Chicago, who introduced progressivism into social work; and above all John Dewey (see *Dewey, John (1859–1952)*), who gave progressive education a sound foundation in pragmatist philosophy.

In the French-speaking countries of Europe the French primary schoolteacher Célestin Freinet, the inventor of the methodology of the *école moderne*, a prominent feature of which is 'the printing press in school,' the Swiss cofounder of the Geneva International Bureau of Education, Adolphe Ferrière, with his idea of the *école active*, where the children were to make their own textbooks as a result of their learning processes, and the Belgian medical doctor Ovid Decroly, the founder of the reform school L'Eremitage in Brussels, are key figures of the *éducation nouvelle*. In Italy, the most prominent representative of progressive education is the preschool educator Maria Montessori, who through her specifically designed didactic materials introduced the idea of self-directed learning into early childhood education. In Germany, the basic patterns of the English new schools were institutionalized by Hermann Lietz and Paul Geheeb in their 'rural education centers'; the Munich school superintendent Georg Kerschensteiner brought learning and practical work more closely together in his 'working school' concept; and the director of the Hamburg art gallery, Alfred Lichtwark, inspired the art education movement. In the Scan-

dinavian countries, the Swedish teacher Ellen Key published the 1900 programmatic pamphlet of progressive education 'The century of the child.'

With the school Lev Tolstoy had erected at his estate in Yasnaya Polyana, progressive education was prepared in the USSR. Stanislav T. Šackih, who had transferred the Anglo-American idea of the settlements into summer camps for children even before the Soviet revolution, built the bridge between progressive and Marxist ideas in the education of the early USSR. The main propagator of melting progressivism into a Marxist–Leninist perspective of education was Pavel P. Blonsky, with his program of the production school. For realizing these efforts in educational policies, even more influential was Lenin's wife, Nadeszhda Krupskaya. The Polish physician and orphanage educator, Janusz Korczak, sealed his view of the dignity of the child by accompanying the Jewish children entrusted to him into the Holocaust.

By the end of World War I, progressive education had grown from local and individual initiatives into national movements and even, despite the nationalist cleavages the war had evoked, into an international network. The founding of the American Progressive Education Association in 1919 and of the international New Education Fellowship in 1921 mark the beginning of that new phase of the movement.

Progressive education now spread from its countries of origin into other parts of the world. Theoretical forerunners, like Tomer Tanimoto and Kanjiro Higuchi, transferred progressive ideas to Japan which were later on institutionalized in experimental schools. Small group education, self-directed learning, and the working school were tried out by educators like Heiji Oikawa, Kishie Tezuka, and Taneichi Kitazama. Progressive education with its more individualist though socially responsible approach became an oppositional force against the nationalist imperialism that led Japan into World War II. Brazil is a good example of the adaptation of new education to the Latin American context. It was especially the 'three cardinals'—Lourenço Filho, who made efforts for a psychologically grounded theory of instruction, Fernando de Azevedo, who was envisaging the structural reform of the Brazilian education system, and Anísio Teixeira, who was striving for a democratic educational policy—that were defending educational progressivism against the conservative Catholic educational theory of the country. Later on, these progressive impulses merged with elements of the 'educação popular.'

On the African continent a mass education system in the 1920s was not yet developed. Consequently there was only a scarce influence of progressive education on the educational activities of the African countries. The school farms and school workshops of different missionary societies linked primary schooling for students and adult literacy programs to manual and agricultural activities. The Phelps–Stokes Com-

mission, which in 1919 evaluated the education situation in the British territories of Africa recommended a practical orientation toward basic education. The skills taught at school should be directly useful in family and community life. In Kenya in 1925 there was an attempt to train teachers as community development workers.

On the Australian continent progressive education was 'very much a series of side eddies, and not the main stream of education' (Connell 1995, p. 351), but teacher educators such as G. S. Browne of the University of Melbourne had acquainted their students with the Dalton Plan or the project method, or the educational philosophy of Dewey.

A quantitative content analysis of educational journals proved that by 1930 progressive education concepts had become the leading educational semantics worldwide; progressive education had indeed developed into an international movement (Schriewer et al. 1998).

2. *Core Elements and Comprehensive Principles*

The intertwining educational philosophies, political semantics, pedagogical programs, theories, projects, foundations of institutions, and teaching methodologies which progressive education consists of are not easily summarized. In spite of Cremin's verdict concerning a capsule definition of progressive education—'None exists, and none ever will' (Cremin 1961, p. x)—the common structural origin and the mutual international criticism of progressive educators, which distinguished between that which was pluralistically proven and that which was blinded by ideology, allow for pointing out some core elements or comprehensive principles and practical approaches of progressive education. These can be reconstructed by looking at the classic historical accounts of Cremin (1961), Boyd and Rawson (1965), Mialaret et al. (1976), or Roehrs (1998b), which together with Roehrs and Lenhart (1995) at the same time contain biographical and work-related references to the aforementioned progressive educators:

(a) Progressive education stood for the democratization of the industrial society by opening up chances of participation on an educational basis. This referred to the induction into technical-industrial processes of production and their social structuring as well as to the authorization to the exercise of political rights. Progressive education brought (sub)cultural group identities into focus. The concept of enabling the individual to cooperate included social feedback on the individual possibilities of action that were opened up by education. The interrelations of society as a whole were thus changed via the spheres of proximity of everyday life. An expression of this is the community-oriented approach of progressive educational endeavors. Education is conceived as an aid for everyone to cope with life.

(b) Progressive education is distinguished by its orientation toward the child's individuality. This means a child-centeredness that provides the educands with enough time to find and do things themselves while playing and learning. In life philosophy's metaphor of the 'opening of the eye,' the pedagogical effort is to let the children themselves experience and recognize 'things.'

Progressive education's intellectuality includes a new relationship to the body. The model is a simple, healthy life-style: there is no cult of athletics, but no neglect of the body either. The concept of human nature is rather pragmatic or positivist—Mialaret et al. (1976, p. 12), following Freinet, 'mechanize the functions (of the body) with the aim to free the vigor of life for the use of the higher activities of the mind.' It does not rule out an ultimate religious bond, but does not put it to the fore either.

(c) As a classical-modern reform movement, progressive education is (social and behavioral) science-oriented. It holds that a scientific foundation for and advice on practical education are both possible and necessary. The concept of science here is pluralistic, including different paradigmatic orientations. Progressive education above all refers to psychology, especially developmental psychology as it was constituted in empirical research and theoretical construction, for example in the works of Piaget (e.g., 1963). Next it refers to a sociology as it was put in concrete forms by J. Dewey (e.g., 1966) or G. H. Mead (e.g., 1967; see *Mead, George Herbert (1863–1931)*), for instance, with a pragmatic intention. Finally, it refers to education as an academic discipline as well, as this has, in its idealist-humanistic variant, become almost an accompanying theory of the German pedagogical reform projects.

(d) Progressive education at its heart is a school reform movement. The evolution of school as an educational institution is not completely to be reversed; still, teaching and learning at school is to be brought closer to the learning processes taking place in the family, workshop, workplace, and community. At the same time, the roles of those who encounter each other at school are changing. There is a friendly relationship between teachers and pupils, based on partnership and companionship, and the opportunities of involvement stretch as far as to the rights of participation ('government') in school communities. Progressive education disseminated into further educational fields apart from school: parallel to school and partly supported by the reform movement, kindergarten education, adult education, and social work were irreversibly institutionalized as fields of pedagogical action in their own right.

(e) As heritage of the philanthropist education of the Enlightenment, progressive education is concerned with a free relationship within educational interaction. This is especially manifest in teaching. The reformers invent, or at least accentuate newly, 'the teaching

dialogue, small group teaching, the various sorts of projects, (the) use of teaching aids, (they emphasize) self-governed activity..., (they aim at) a plurality of teaching methods, that (cause) as an “education by peers in the classroom” ...a change in the direction of the activity’ (Roehrs 1998, p. 319). The change of direction finds expression especially in those didactic models that transfer the planning of work and learning at least partly to the pupils. While acquiring and applying a basic knowledge on methodology the children prove their self-governing abilities.

3. *Progressive Education in Practice—Two Examples*

What internal reform of the school on progressive lines means in detail can be illustrated by the famous project method. This approach was theoretically explained by Kilpatrick (1918) and in the following decades practically spelled out worldwide in many variations (cf. Kliebisch and Sommer 1997). For Kilpatrick, a project was a ‘wholehearted purposeful activity proceeding in a social environment’ (Kilpatrick 1918, p. 330). This formal definition covers different project types which the educationist structures according to learning objectives and content matter, but he also gives hints to classify them into the categories of individual and group projects. A project has four phases: ‘purposing, planning, executing, and judging.’ The learners take all steps autonomously; the teacher only interferes if they are asked for advice or if the learning process is threatened by total failure. Later on, Kilpatrick embedded the project method into his philosophy of the school and of teaching procedures:

(1) Count that the school is properly a place of living. Work for the finest quality of living that teacher and pupils can together devise, reaching out as much as feasible into the community... (2) Let the principles of interest and purposeful activity set the predominant pattern... (3) Run everything on a basis of democracy and respect for personality... (4) Work in season and out, depending on pupil maturity and development, for the finest and most inclusive aims of life, these to be increasingly exemplified in everything done in school and out.’ (Kilpatrick 1959, p. 429)

Progressive education in its linkage to social work can be illustrated by the example of Jane Addams. Although she wrote several books and contributed articles to academic journals (e.g., the *American Journal of Sociology*), the focus of her activities was not the production of academic theory but practical political, educational, and social work efforts. These activities had their institutional basis in Hull House, a community center which Addams had set up in a Chicago slum area in 1889 after the model of the English ‘settlements.’ For 46 years the house under her leadership has been a center that has intervened in local educational policies to improve the educational

situation for immigrant children, organized its own educational activities—such as a Shakespeare Club or a Working People’s Science Club—provided meals for the poor, and improved hygiene standards by establishing public baths. Production and consumption cooperatives were operating; for some ‘residents’ common households were organized. Hull House built the first public playground in Chicago, and coinitiated the Juvenile Protection Association, which, through social activities and political pressure (e.g., on shop owners who sold alcoholic beverages to minors), tried to prevent youth delinquency. Addams and her friends succeeded in having a youth court set up that was no ordinary criminal court but acted *loco parentis*. The house was active in reducing child labor, and it fought against the ‘white slave trade,’ the ‘import’ of young women for prostitution. Hull House supported the trade unions, backing up the laborers in conflict situations, such as the big Pullman strike of 1894. The activities of Hull House finally contributed to the differentiation of sociology and social work theory and—against Addams’s intentions—to transform social work into a profession (Eberhart 1995).

4. *History or Ongoing Movement*

Historians of education dispute the question whether progressive education was a historical movement in its time or whether it is a continuous ongoing stream of theories and activities. US theorists such as Sol Cohen (1995) tend to see it as history (and the US school reform efforts of the 1990s even as a countermovement to restore the moral-intellectual discourse of education), while the German contribution of Oelkers (1996) doubts whether there was an epoch of progressive education at all, and that of Roehrs (1998b) conceives of it as a continuing story. The overt history of progressive education in the USA ended in 1955 when the Progressive Education Association was dissolved. In fascist Italy and Germany progressive education with its liberty-oriented and democratic tendencies was politically banned, and some of its German proponents, such as Kurt Hahn, the director of the famous Salem School, were driven out of their home country. A few others, however, such as Andresen, Lietz’s successor in the leadership of the rural education centers, tried to establish links to Nazi ideology, thus enabling several measures of progressive education to be abused by totalitarian educational practice. In the USSR progressive education was suppressed by Stalin’s dictatorship from the early 1930s. He even threatened Krupskaya with using another woman to ‘replace’ her as Lenin’s widow, unless she stopped publishing about polytechnic education.

Conversely, the New Education Fellowship, as the organization of the progressivists, made important contributions to the establishment of UNESCO (United Nations Educational, Scientific, and Cultural

Organization) within the United Nations (UN) system after World War II. Many progressive education institutions have continued their activities into the present. More important, perhaps, is the fact that since the 1950s progressive ideas and practices have trickled down into the day-to-day activities of education worldwide. In this reduced form the movement is still a reservoir of innovation, and this is reflected in the alternative schools that have been established from the 1970s onward. The progressive movement has even made contributions toward meeting the most recent educational challenges, such as peace education, multicultural education, and human rights education. These must now be transformed creatively to build the 'four pillars of education' enumerated by the International Commission on Education for the Twenty-first Century: 'learning to know,' 'learning to do,' 'learning to live together, to live with others,' 'learning to be' (UNESCO 1996, pp. 86–96).

Bibliography

- Boyd W, Rawson W 1965 *The Story of the New Education*. Heinemann, London
- Cohen S 1995 The influence of progressive education on school reform in the USA. In: Roehrs H, Lenhart V (eds.) *Progressive Education Across the Continents*. Lang, Frankfurt, Germany, pp. 331–2
- Connell W 1995 The influence of progressive education on Australian schools. In: Roehrs H, Lenhart V (eds.) *Progressive Education Across the Continents*. Lang, Frankfurt, Germany, pp. 351–6
- Cremin L A 1961 *The Transformation of the School. Progressivism in American Education 1876–1957*. Knopf, New York
- Dewey J 1966 *Democracy and Education. An Introduction to the Philosophy of Education*. Free Press, New York
- Eberhart C 1995 *Jane Addams (1860–1935)*. Schaeuble, Rheinfelden, Germany
- Kilpatrick W H 1918 The project method. *Teachers College Record* 19: pp. 319–35
- Kilpatrick W H 1959 *Philosophy of Education*. Macmillan, New York
- Kliebisch U, Sommer P 1997 *Projektarbeit. Konzeption und Beispiele*. Schneider, Baltmannsweiler, Germany
- Mead G H 1967 *Mind, Self, and Society*. University of Chicago Press, Chicago
- Mialaret G, Chenon-Thivet M, Fabre A, François-Unger C, Gal R, Gloton R, Monod G, Seclet-Riou F 1976 *Education nouvelle et monde moderne*. Presses Universitaires de France, Vendôme, France
- Oelkers J 1996 *Reformpädagogik. Eine kritische Dogmengeschichte*. Juventa, Weinheim, Germany
- Piaget J 1963 *La naissance de l'intelligence chez l'enfant*. Delachaux Niestle, Neuchâtel, Switzerland
- Roehrs H 1998a *Reformpädagogik und innere Bildungsreform. Gesammelte Schriften* 12. Deutscher Studienverlag, Weinheim, Germany
- Roehrs H 1998b *Die Reformpädagogik. Ursprung und Verlauf unter internationalem Aspekt*. Deutscher Studienverlag, Weinheim, Germany

- Roehrs H, Lenhart V (eds.) 1995 *Progressive Education Across the Continents*. Lang, Frankfurt, Germany
- Schriewer J, Henze J, Wichmann J, Knost P, Taubert J, Barucha S 1998 Konstruktion von Internationalität: Referenzhorizonte pädagogischen Wissens im Wandel gesellschaftlicher Systeme (Spanien, Sowjetunion/Russland, China). In: Kaelble H, Schriewer J (eds.) *Gesellschaften im Vergleich. Forschungen aus Sozial- und Geisteswissenschaften*. Lang, Frankfurt, Germany, pp. 151–258
- UNESCO 1996 *Learning: The Treasure Within 1996. Report to UNESCO of the International Commission on Education for the Twenty-first Century*. UNESCO, Paris

V. Lenhart

Projective Methods in Psychology

Projective techniques is a term that encompasses any test or procedure designed to increase insight into individuals by allowing them to respond freely to ambiguous stimuli. The underlying assumption is that, when faced with unstructured or ambiguous stimuli, people will reveal aspects of their personality in their attempt to structure the material. In other words, the vague quality of the stimulus evokes free fantasy play and triggers projection on the part of the subject while simultaneously revealing aspects of psychological functioning in the nature of their responses. Many experts believe that everything said or written in response to a stimulus has dynamic meaning and that people reveal conflicts, anxieties, attitudes, beliefs, and needs in their attempt to structure or complete an incomplete stimulus. Projective techniques tend to have brief, general instructions that, when coupled with ambiguous stimuli, tend to elicit less defensive responses than more structured testing. It should be noted that a concern about conflicts, anxieties, attitudes, beliefs, and needs, and the usual attempts to blend these into a picture of a person, is most closely associated with one approach to personality, the psychoanalytic approach. A preference for psychodynamic theory is not required in order to use projective techniques, but it is frequently found. Perhaps this is because one of the hallmarks of the psychoanalytic approach is the prominent role assigned to unconscious processes, and projective techniques are designed with the intention, satisfied or not, of eliciting unconscious material. (see *Psychoanalysis in Clinical Psychology*).

1. Historical Overview

Historically, the turn of the twentieth century was a time when psychological testing was applied predominantly for the study of intelligence. The trend was

toward finding data or norms descriptive of groups of people (nomothetic data). At this time, personality measures tended to examine only a single trait at a time (e.g., introversion/extroversion, flexibility). The scope began to broaden about two decades later when interest surfaced in tests that yielded specific data about an individual (idiographic data). In the early 1920s, projective testing emerged when Rorschach ([1921], 1942) developed his now famous inkblots in an attempt to aid in the diagnosis of schizophrenia.

The notion of projection with these tests gave birth to the progression of projective techniques throughout the rest of the twentieth century and the belief that ambiguous stimuli can be used to gain insight into the dynamics of the respondent. Frank (1939) coined the term 'projective methods' to describe a variety of techniques used by clinicians with ambiguous stimuli that pull for projection. He postulated a projective hypothesis, namely that the world is structured according to internal needs and external demands, so that ambiguous external stimuli will elicit structuring based on internal factors. The 'Projective Movement' grew tremendously during the 1940s and 1950s, as clinicians moved away from the nomothetic approach to studying their patients toward the more idiographic approach. More recently, because of psychometric considerations, and possibly because of a reduction in the number of clinicians trained in psychodynamic approaches, the use of projective techniques has waned, leaving an open question and considerable controversy over the value of the projective approach.

Perhaps the most commonly used projective techniques are the Rorschach, the Thematic Apperception Test (TAT), figure drawings, and sentence completion tests.

1.1 Rorschach

The use of inkblots was not a new concept in the 1920s. In fact, a popular game among children at the time was Blotto, in which children used bought or handmade blots of their own. Rorschach's clinical interest in the use of blots was sparked in 1911, while working with Eugen Bleuler, his professor and supervisor, who coined the term schizophrenia and instigated the need to differentiate schizophrenia from other forms of psychosis. Rorschach believed that the process of perception, as a respondent struggles to make sense of an ambiguous stimulus, reveals the threshold of each respondent to assimilate or integrate the new stimulus with existing mental images. Rorschach happened to observe in his work that schizophrenic patients tended to respond differently in Blotto than patients with other forms of psychosis or dementia, and decided to study this phenomenon more systematically. By 1921, he had administered the test often enough to believe in its usefulness as a diagnostic aid in identifying schizophrenia, but noted

unexpectedly that increased frequencies of certain kinds of responses were consistent with certain personality traits or styles. Rorschach died in 1922, but his successors saw the potential of his tool as an idiographic, projective measure and worked over the next two decades to develop a standardized scoring system for responses to the inkblots.

Samuel Beck was a psychology graduate student who attempted, shortly after Rorschach's death, to do a standardization study as a dissertation topic. He began his work at the Institute for Child Guidance in New York and later took positions at both Boston Psychopathic Hospital and Harvard Medical School. Over the years he gained extensive experience with the use of the inkblots and became acutely aware of the need for more intensive work toward standardization.

Around the same time, Bruno Klopfer came to Zurich to work as a technician at the Zurich Psychotechnic Institute where Rorschach inkblots were routinely used in the screening of candidates for employment positions. After working there for several years, he emigrated to the USA in 1934 and began working at Columbia University. The psychology graduate students at Columbia, who were frustrated by the lack of formal training available in the use of inkblots, were eager to have him share his experience with them. Klopfer began giving informal seminars to students and, in his work with these students, began to develop new scoring codes that seemed necessary, especially those for unusual locations or shading responses. Thus, by 1936, Klopfer was working hard to expand the existing coding system.

At the same time, Beck was busy writing papers that criticized the existing scoring systems for their lack of empirical soundness. He approached Klopfer's attempts to expand the system with criticism. Over the next 20 years, the arguments continued and different scoring systems were attempted. Despite this controversy over scoring, the method of using the Rorschach continued to burgeon. By 1957, five vastly different scoring systems had emerged.

John Exner began to study the difficulty with using inkblots in 1968. Between 1968 and 1972, he found that most clinicians were scoring the tests in a haphazard manner. Many did not score it at all, and those who did tended to use features of different scoring systems. Only a small minority used one system to score an entire test. Exner set out to develop a system that integrated features of all the available scoring systems that had empirically defensible data, and a system that was not wedded to any single personality theory. Thus, he developed the comprehensive system that is still used by many clinicians and that includes the successful work and research that accumulated from the 1920s onward. The system developed by Exner (1993) is probably the most widely used at the beginning of the twenty-first century, but many clinicians continue to treat the Rorschach purely as a projective device, rather than as a psychometric

instrument, and simply interpret the content of the responses according to their understanding of personality. It should be noted that there are several sets of ambiguous stimuli, usually inkblots, that are in use, but the Rorschach is by far the most popular.

1.2 Thematic Apperception Test

In 1935, Morgan and Murray introduced the Thematic Apperception Test (TAT; Murray 1943), a series of provocative pictures used to explore the respondent's dynamics. The stimuli are more highly structured than inkblots and respondents are required to give more meaningfully organized verbal responses. Respondents are asked to describe the scene in the picture, the apparent emotional content and thoughts of the people involved, what led up to the scene, and what the outcome will be. This test was developed on the premise shared by other projective techniques that people reveal something of their own personality (i.e., their motivational and emotional condition), as well as their cognitive structure, when ascribing thoughts and feelings to ambiguously drawn characters. Just as with the Rorschach, many sets of pictures have been developed for use as an apperceptive approach, but the TAT remains the most widely used.

Perhaps the most popular scoring method for this test is that devised by David McClelland (McClelland et al. 1953). McClelland's work was devoted to measuring achievement motivation and he scored the TAT responses across seven dimensions: (a) stated need for achievement (someone in the story expresses a desire to achieve a goal); (b) instrumental activity (whether the outcome of the instrumental activity is successful, doubtful, or unsuccessful); (c) anticipatory goal status (whether the person in the story anticipates achieving the goal or being frustrated by it); (d) obstacles or blocks (whether the goal in the story is blocked by an obstacle); (e) nurturant press (whether there are nurturing forces aiding the characters in goal achievement); (f) affective states (emotions present associated with achieving or being frustrated by the goal); and (g) achievement themes (when achievement imagery is the focus of the story).

TAT responses have been studied and scored in a variety of ways, with McClelland's approach being only one example. By far the easiest and most commonly used method among clinicians is simply to read the responses, look for common themes, treat the stories as psychologically meaningful communications, and consider them as projective tools for understanding the patient's underlying conflicts and needs.

1.3 Figure Drawings

Figure drawings are a series of drawings by the respondent that the clinician uses to understand nonverbal expressions of the self. Respondents are

asked to draw a full person (gender not specified ahead of time), and then to draw a person of the opposite sex of the first drawing. The absolute and relative size of the persons are examined as well as details such as the position on the page, the quality of the lines, the sequence of parts drawn, the stance, position of the arms, and grounding effects. Clinicians have often found these drawings helpful in understanding the self-perception, motor expression, anxieties, and needs of the respondents. Variations of this task ask the respondent to produce other drawings, including drawings of objects such as a house or a tree. In general, regardless of what is drawn, the research with these approaches has not been very encouraging (Swenson 1968).

1.4 Sentence Completion

Sentence completion is another method for gathering information about personality dynamics. Respondents may be asked, orally or in writing, to complete sentences that are vague and ambiguous (e.g., What worries me ... ; Women ... ; My mother ...). In addition, there have been standardized written sentence completion tests developed by authors such as Rotter (Rotter and Rafferty 1950). As with the TAT and other apperceptive measures, the value of the task often is dependent on the particular set of items that is employed, and the more closely the items approximate the target being studied, the more valuable the responses can be shown to be.

2. Psychometric Considerations

A psychological test or assessment device is subject to evaluation on standard dimensions and, regardless of how compelling a rationale may be, must demonstrate that it meets the usual psychometric criteria before it can be considered to be useful. Foremost among these criteria are reliability and validity (for a more detailed explication of these critical concepts, see Wiggins 1973).

2.1 Reliability

Reliability refers to the extent to which a test provides a consistent measure of what it is measuring. If a test lacks reliability it is akin to using a rubber ruler, providing different measurements each time it is applied, and there is little credence that can be given to any of those measurements. One approach to the measurement of reliability is through retest, with the expectation that a test given on two occasions will produce the same response on both testings. A second approach is internal consistency, which assesses the extent to which every item in a test is measuring the same characteristic.

Although most projective techniques employ many items, the items often are designed to be responsive to different dimensions, and so a projective test may often be seen as a collection of disparate single items rather than a collection of redundant items. Unfortunately, in measurement, redundancy is often desirable, and it is well demonstrated that single items are subject to far more unreliability than collections of items, in which errors may cancel each other out. If there are indeed collections of single items, it can be argued that the test really is a collection of tests, each of which measures a single dimension (and therefore requires a separate assessment of reliability and validity), and does so in a way that reduces the reliability of each. A proponent of these tests would counter that the object being measured is personality, which contains several dimensions, and that the items should converge to produce a clear picture of the multidimensionality of the respondent. Whether or not this is so is a matter of validity.

2.2 Validity

Reliability is necessary but not sufficient for a test. An unreliable test cannot be valid, but a reliable test may not necessarily be valid. Validity is an assessment of the extent to which a test measures what it is intended to measure. This leads to a very interesting problem in assessing the validity of projective techniques. Most projective techniques purport to measure many different aspects of the person, and it is necessary to indicate the validity of every inference that a test draws. Thus, rather than asking, for example, whether the Rorschach is valid, it is necessary to ask, for each inference drawn from the Rorschach, whether or not that inference is valid. One statistical estimate cannot represent the validity of the test, and the general question cannot be answered (Stricker and Gold 1999).

It is not unusual for a research project to compare the validity of projective techniques with more objective, self-report, measures of personality (e.g., McClelland et al. 1989). The usual finding of such studies is that the two approaches do not correlate with each other. Some people have interpreted this as indicating a failure in the validity of the projective approach, a conclusion that assumes the validity of objective tests. Others have determined that this lack of correspondence indicates that each approach is valid, but that they are measuring different aspects of the person. This view suggests that objective tests are superior at measuring objective criteria, such as behavior, whereas projective approaches are superior at measuring more covert, unconscious aspects of the person (Hiller et al. 1999). This interpretation is consistent with the theoretical position that usually underlies each of the approaches to measurement.

It also is reasonable to ask whether the incorporation of a projective approach to a test battery adds anything to the process of assessment (incremen-

tal validity). If the conclusion about measuring different aspects of the person is accepted, then projective approaches do contribute incremental validity insofar as the assessor is interested in less overt aspects of the respondent. Given the length of time necessary for the administration and interpretation of projective techniques, the question of cost effectiveness also can be raised. The answer to this question probably is dependent on the individual circumstance. It should be clear, however, that validity, incremental validity, and cost effectiveness are separate questions, and the answer to one does not determine the answer to the others.

Finally, it should be noted that the subjectivity of the projective approach places great stress on the skill of the examiner, and any question of validity is equally a function of the ability of the approach to elicit material and the ability of the examiner to interpret it.

3. Concluding Comments

It can be concluded, with some certainty, that the original promise of the projective hypothesis (Frank 1939) has not been sustained, and that there are many limitations on the conclusions that can be drawn from ambiguous stimuli. However, there are instances in which such stimuli can be used to produce valid inferences, and inferences that cannot be drawn from other approaches to measurement. Whether these inferences are useful depends on the theoretical approach and goals of the examiner, whose skillfulness also must be considered in reaching a conclusion about the utility of projective methods. Because projective techniques do show the promise of providing valid information that is different than that produced by objective measures, it is a path worth pursuing. However, any pursuit will have to be marked by attention to psychometric considerations, a neglect of which has led to the excessive promise of past attempts.

See also: Bleuler, Eugen (1857–1939); Content Validity; Measurement Theory: History and Philosophy; Mental and Behavioral Disorders, Diagnosis and Classification of; Personality Assessment; Personality Theories; Personality Theory and Psychopathology; Psychoanalysis in Clinical Psychology; Psychology of the Real Self: Psychoanalytic Perspectives; Psychometrics; Reliability: Measurement; Unconscious: History of the Concept

Bibliography

- Beck SJ 1950 *Rorschach's Test. I. Basic Processes*, 2nd edn. Grune and Stratton, New York
- Exner J 1993 *The Rorschach: A Comprehensive System*, 3rd edn. Wiley, New York, Vol. 1

- Frank L K 1939 Projective methods for the study of personality. *Journal of Psychology* 8: 389–413
- Hiller J B, Rosenthal R, Bornstein R F, Berry D T R, Brunell-Neuleib S 1999 A comparative meta-analysis of Rorschach and MMPI validity. *Psychological Assessment* 11: 278–96
- Klopfer B, Ainsworth M D, Klopfer W G, Holt R R 1954 *Developments in the Rorschach Technique. Vol. 1, Technique and Theory*. World, New York
- McClelland D C, Atkinson J, Clark R, Lowell E 1953 *The Achievement Motive*. Appleton-Century-Crofts, New York
- McClelland D C, Koestner R, Weinberger J 1989 How do self-attributed and implicit motives differ? *Psychological Review* 96: 690–702
- Morgan C D, Murray H A 1935 A method for investigating fantasies: The Thematic Apperception Test. *Archives of Neurology and Psychiatry* 34: 289–306
- Murray H A 1943 *Thematic Apperception Test Manual*. Harvard University Press, Cambridge, MA
- Rorschach H [1921] 1942 *Psychodiagnostics: A Diagnostic Test Based on Perception*, 4th edn. Grune and Stratton, New York
- Rotter J B, Rafferty J E 1950 *The Rotter Incomplete Sentences Test*. Psychological Corporation, New York
- Stricker G, Gold J R 1999 The Rorschach: Toward a nomothetically based, idiographically applicable configurational model. *Psychological Assessment* 11: 240–50
- Swenson C H 1968 Empirical evaluations of human figure drawings. *Psychological Bulletin* 70: 20–44
- Wiggins J S 1973 *Personality and Prediction Principles of Personality Assessment*. Addison-Wesley, Reading, MA

G. Stricker and K. Somary

Pronouns

Noun referents can be substituted by personal pronouns. In most contexts the pronoun is coreferential with and interchangeable with the noun to which it refers. Other pronouns determine nouns, some have the capability of doing both. In English, the personal pronouns substitute, the articles determine, and the demonstratives do both. These, together with indefinites, negatives, locatives, interrogatives, and relatives are grouped together as non-personal pronouns.

The word *tree* refers to an object in the real world (exophoric use), but *it* can have meaning only in referring to some noun such as *tree* by being coreferential with it (endophoric use). Within the linguistic context *it* can replace *tree* or any other noun of the same class—but not, for example, *man* or *trees*, which require special pronominal forms in English.

Pronoun sets can also be distinguished on the basis of case and emphatic differences. Not all pronouns are in fact interchangeable with nouns, such as the reflexives, the logophorics, and the sequentials. Pronouns may be morphologically free or bound, that is, they may be separate words or affixes.

Instead of treating pronouns as part of the lexicon, Everett (1996) argues for considering them to be composed of phi-features, such as case, gender, number, tense, and person. These phi-features are stored in the lexicon, their phonological shapes are given by postlexical or precompiled spell-out rules. (see *Clitics, Linguistics of; Morphology in Linguistics; Word, Linguistics of*.)

1. Personal Pronouns

The personal pronouns can be studied under the parameters of person, number, and gender distinctions, which vary considerably from one language to another.

1.1 Person

In the person parameter, the third persons (*he/she/it/they*) are usually used endophorically, whereas the first and second persons (*I, we, you*) are frequently used exophorically.

The only language with person but no number distinctions reported to date is Mura Pirahã, an unclassified Amerindian language (Brazil). First, second, and third constitute the only formal pronoun distinction this language has. One language has been found which appears to distinguish in the pronominal system two persons only: first and nonfirst. In the *Balante* variety spoken in Guinea-Bissau, *N*-(homorganic nasal) is used for 1sg, *ha-* for 2sg or 3sg, and *be-* for plural in any of the three persons.

Many languages have pronouns that combine 1 + 2 persons; that is, speaker and hearer, into an *inclusive* first person. This is in contrast to the 1 + 1 or first person *exclusive* category, in which the hearer is excluded explicitly. Some languages combine other categories as well, such as 1 + 3—that is, speaker plus one person not present; 2 + 3—grouping together one hearer and one person not present; and 3 + 3—grouping together two persons not present. In addition to binary combinations, the tertiary combination of 1 + 2 + 3 is sometimes used, where everybody is included explicitly (particularly the ancestors).

1.2 Number

The person distinctions are usually combined with number distinctions, singular and plural being the most common. When person categories are combined, making the smallest possible number two, or even three, it is more sensible to talk about a minimum-plural distinction. The most complex person-number combinations so far reported are found in Ghomala'

Table 1
Ghomala' subject pronouns

Person		Minimum	1pl	2pl	3pl
Mono	1	gǎ	pyǎ		
	2	o		po	
	3	e			wap
Double	1 + 2	pu		pə	
	1 + 3	pyǎ	pyayú		pyəapu
	2 + 3	poé		poayú	poapu
	3 + 3	pué			wap
Triple	1 + 2 + 3	pəayú	pəayú	pəayú	wəkpampu

(Bameleke, Cameroon). The plural categories have hierarchical order: first person comes before second, which comes before third.

Sursurunga (New Guinea) has what is probably the most complex number system in operation with mono, dual, triple, and quadruple, as well as four-person distinctions (Hutchisson 1986).

1.3 Gender

Gender distinctions are basically of two kinds: natural gender, such as masculine, feminine, and neuter (or animate/inanimate, human/non-human, for example), and grammatical gender having to do with noun classes. The latter is common in African and some Asian languages. Normally, gender distinctions are restricted to the third person, but may be found in other persons as well, particularly in possessive function.

1.4 Honorific Use

Many languages have special uses of pronominal forms for showing respect or social distance. Instead of the second singular, the plural may be used. Completely different pronouns can also be found. For Bangkok Thai, 17 different forms for first, 19 for second, and 10 for third person are in use, distinguished by such categories as power and status, kinship, friendship, occupation, ethnic-religious groups, age, sex, genealogical distance, solidarity, formality, presence of certain persons, emotional manifestations, and combinations of all these (Palakornkul 1975).

1.5 Special Meanings

There may also be unusual uses of person distinctions. Exclusiveness may be an added dimension to any of

the personal pronouns, though there are no special forms to mark the distinction. In several Amerindian languages the 1sg has the exclusive meaning of 'only I (speaker), but not anybody else,' whereas the second means 'only you (hearer) but not I' (speaker). The only pronoun which can be used inclusively is the 1pl form.

Other pronouns can have meanings which are not apparent from examining the system. In Kaingang (Je, Brazil) whenever a family unit is involved, the 3pl fem. form is used after the name of the man; the 3pl masc. is reserved for a group of men only (see *Classifiers, Linguistics of; Grammatical Number; Language and Gender; Language and Social Class; Numeral Systems; Politeness and Language*).

2. Nonpersonal

Nonpersonal pronouns are used in the third person only, but may vary in number and gender.

2.1 Definite

Demonstratives and articles are included in the series of definites. A distinction is often found according to where the determined noun is placed in relation to the speaker and/or the hearer, which gives a two- or three-way distinction, as when the locative adverbs are used pronominally.

In English:

- (a) *the* man is anaphoric: the already mentioned;
- (b) *this* man is close to speaker (often anaphoric);
- (c) *that* man is far from speaker (often anaphoric);
- (d) *this* makes me wonder if ...

This and *that* may be used as substitutes (exophorically) if the object is in sight (and in that case not necessarily anaphorically) or to replace arguments and other parts of discourse. In German the articles may substitute for the noun.

In Didinga (Nilo-Saharan, Sudan) separate forms exist for information that is generally known as against

exclusive knowledge of the speaker; or whether the pronoun refers to a happening long ago or recently, or whether the person was referred to recently or much earlier.

2.2 Indefinite

A number of languages also have indefinite pronouns, referring to no one in particular, but rather to everyone or certain unspecified individuals, such as *they* in English or *on* in French. The second person, singular or plural, might be used in this way also: *you don't beat children!*

Among the indefinite forms can be classed such words as *a, certain, other, such, few, some, many, any*, and others. Some languages class the distributives among the indefinite, such as *all, every, each, both, half, part of, some, another, other, either*, and others.

2.3 Negative

Certain forms are entirely negative. They are often subsumed under indefinite. In English they include *none, nobody, nothing, no, neither*. In many languages the negation is not expressed in a pronoun (or noun) but rather in the verbal expression.

2.4 Interrogative

In all languages, people ask questions with the aid of interrogative substitutes or determiners. Some of these are simple, others are constructed from simple forms. In English, the simple forms are *how? what? when? where? which? who? whom? whose? why?*, constructed forms *how many? whatever?*, and others.

Some languages form interrogatives from demonstratives, such as:

Fulfulde (Cameroon)

- (a) dewtere nde'e 'this book' (demonstrative)
- (b) dewtere ndeye 'which book?' (interrogative)
- (c) gorko o'o 'this man' (demonstrative)
- (d) gorko oyo 'which man?' (interrogative)

2.5 Relative

Relative constructions usually contain a relative pronoun, which may be a demonstrative or an interrogative, or an entirely different form, restricted in use to relative constructions. English: 'Here is the one *who* could help.' In Konzimé (Cameroon), the pronoun comes at the end of a relative construction and can be:

- either a special relative pronoun which varies according to the noun classes; or

- the demonstrative *ní* which is invariable.

In some languages the relative pronouns also indicate differences in tense and form part of the verbal paradigm:

Denya (Ekoid, Cameroon)

- (a) à-wá 'he killed' – á-wà-né 'he who killed'
- (b) à-wà-nè 'he kills' – á-wá-nè 'he who kills'

2.6 Pronominally Used

A number of adverbs have pronominal use, especially the indicators of space and time. The locative adverbs indicate at least three different localities: close to speaker, close to hearer (but far from speaker), and far from both, as in English *here—there—yonder*. There may also be other types of information, such as *inside, outside, underneath, above, before, after* that could be counted among them.

Nouns can also be used pronominally, especially (but not only) as honorifics. In Brazilian Portuguese, *a gente* 'the people' is commonly used instead of *nós* 'we'. (see *Adverbial Clauses; Anaphora; Grammaticalization; Negation, Linguistics of; Quantifiers, in Linguistics; Relative Clauses*).

3. Set Distinctions

In noun phrases and clauses, pronouns may be marked for any case distinction found in the language. Case indicators including circumstantial information may be fused to the pronouns causing different sets. English distinguishes *I—me, he—him, she—her, we—us, they—them*.

Yaouré (Ivory Coast) shows pronoun fusion with postpositions (Hopkins 1986):

i + lē → *jíé* '2sg + of → of you sg'

ká + lē → *ké* '2pl + of → of you pl'

3.1 Genitive

In genitive constructions the possessor can be replaced by a pronoun. The possessed item often takes the mark of the clause level function, whereas the mark on the possessor identifies it as such by the genitive case. The agreement rules of the associative noun phrase can be quite complex and varied.

English:

- (a) John's book – *his* book (agrees with John)
- (b) Mary's book – *her* book (agrees with Mary)

French:

- (a) le livre de 'the book of
Jean – *son* livre John – his (masc) book'
(agrees with book)

Table 2
Kabiye *subject + mode*

	Singular	Plural	Singular	Plural
1	mòn	dí	máá	díí
2	ñ	é	nn	éé
3	è	pà	éé	páá

Table 3
Rendille (Kenya) *subject + aspect*
(verb root: táx-, -e marks perfective, -a imperfective)

	Perfective	Imperfective	Person markers
1sg	táxe	táxa	
1pl	táxne	táxna	-n
2sg	táxte	táxta	-t
2pl	táxten	táxtan	-t..-n

- (b) la case de ‘the hut of
Jean – *sa* case John – his (fem) hut’
(agrees with hut)
- (c) les livres de ‘the books of
Jean – *ses* livres John – his (pl) books’
(agrees with books)

African class languages show a wealth of pronominal forms in all the cases. In the genitive, agreement holds in first and second person as well. The agreement may be with the possessor (as in English) or with the possessed item (as in French), or with both.

There may be different kinds of genitives, such as alienable and inalienable possession. The differences may be marked in the pronoun sets:

Lendu (Zaire)

(a)	<i>mà</i>	<i>ndzá</i>	<i>búkù</i>	'my book'	alienable/ possessive
	my	of-cl.	book		
(b)	<i>nì</i>	<i>dhí</i>	<i>búkù</i>	'your book'	
	your	of-cl.	book		
(c)	<i>mā</i>	<i>djō</i>		'my brother'	inalienable/ qualificative
	my	brother			
(d)	<i>n'ī</i>	<i>djō</i>		'your brother'	
	your	brother			

In the alienable construction, the tone of the possessor pronoun is different from the inalienable construction, and the alienable particle 'of' (which changes in form according to noun class) is absent in the inalienable construction.

3.2 Noun Phrase Substitution

The whole possessive noun phrase can be substituted by a pronoun or a nominalized or adjectivized pronoun, which functions on clause level and may carry case markers, as in German:

- (a) Mutters Garten – *ihr* Garten – *ihrer* – der *ihrige*
(b) *mein* Garten – *meiner* – der *meinige*

Table 4
Attie (Ivory Coast) *subject + mode + aspect*

	Perfective	Imperfective	Negative Perfective	Subjunctive
1sg	mě	mẽ	mẽ	mẽ
2sg	bù	bò	bù	bú
3sg	ò	wò	ò	ó

Table 5
Mundani (Cameroon) emphatic pronouns (simplified)

Person	1. Degree	2. Degree	Authoritative
1sg	mmʒ	mmʒ mbɔŋ	mmʒ mbɔŋ
1pl	bàá	bàá mbɔŋ	bàá tà mbɔŋ

3.3 Subject + Distinctions

Subject pronouns may contain, in addition to person, number, and gender information, a case marker, combined with verbal information such as mode—as illustrated from Kabiye (Togo), or aspect as illustrated from Rendille (Kenya), or a combination of these, as illustrated from Attié (Ivory Coast).

The Kabiye irrealis pronouns seem to be a fusion of the first series plus a suffix *-ka*, which is seen in its full form when following nouns.

The Attie imperfective can be reconstructed as a fusion of the perfective plus a suffix *-a*. In the negative perfective series, only two forms are different from the perfective: the 1sg and 2pl forms have high instead of low tone. Nevertheless, the whole set constitutes a new series. The subjunctive forms all have high tone.

3.4 *Emphatic*

Emphatic pronouns can be found in all languages, especially in subject function. These are usually focus forms, which can also be differentiated by degree. All of these differences can result in different pronoun series.

In Bafut (Grassfields, Cameroon), in addition to the location an insistence distinction adds member to the demonstrative set. These distinctions go from 0, the 'normal' form, to 3, where the people are actually shouting at each other. Since Bafut is a class language, the whole set includes corresponding forms for each class; class 5 forms are presented here. The location distinctions are: *nú* 'close to speaker'—*ní* 'close to hearer'—*nyá* 'far from both,' and all have corresponding forms:

- (a) nìbá nû 'this (here) wing' (0)
 (b) nìbá núà '(no,) *this* (here) wing!' (1)
 (c) nìbá núlà '(no, no,) *this (here)* wing!' (2)
 (d) nìbá núànuù '(of course not!) *this (here)* wing!' (3)

Table 6
Kobon SS/DS pronouns

Person	SS			DS		
	Singular	Dual	Plural	Singular	Dual	Plural
1	-em	-ul	-un	-nō	-lo	-no
2	-mön	-mil	-mim	-ö	-lö	-be ~ -pe
3	-öm	-mil	-öm	-ö	-lö	-ö

(see *Grammatical Agreement; Possession (Linguistic); Serial Verb Constructions; Sign Language; Tense, Aspect, and Mood, Linguistics of*).

4. Noninterchangeable

Some pronouns are not interchangeable with the nouns with which they are coreferent. *John hurt himself* cannot be replaced by **John hurt John* without changing the meaning.

4.1 Reflexive and Reciprocal

Reflexive pronouns are used to show that the agent and the patient of an event are coreferent:

- (a) John_i hurt himself_i
- (b) John_i and Mary_k hurt themselves_{i+k}
- (c) John_i and Mary_k hurt each other
/one another_{i <-> k}

If more than one participant is involved, and in the course of the action they exchange roles, the construction is reciprocal.

The relationship can hold between the subject (or agent) and the possessor:

- (d) John_i watched his (own)_i son reflexive
- (e) John_i watched his_k son not coreferent

4.2 Logophoric

Logophoric indicators, usually pronouns, are found in many African languages. Their primary use is in indirect reported speech. They express coreference between the quoted speaker and any participant in the quote content, but primarily the subject or agent.

Tupuri (Cameroon)

- (a) Pɔ_i rin Jan_k gá sɛ_i le'ɛ
Paul_i said to John_k that LOG_i fell
'Paul said to John that he (Paul) fell'
- (b) Pɔ_i rin Jan_k gá a_{k,i} le'ɛ
Paul_i said to John_k that he_{k,i} fell
'Paul said to John that he (John or someone else) fell'

In all the languages that have such a logophoric indicator, the relationship holds between the quoted

speaker and the subject of the quote content. Some languages use the same particle to show coreference between the speaker and other functions in the quote content, such as direct object, indirect object, or possessor.

The logophoric indicator is normally used in the third person, always singular, but sometimes plural as well. Some languages also use it in the second person, and in Lélé (Chad) it is used for all three persons, singular and plural. In languages with nominal classes, the marker may be distinguished according to the class of the noun to which it refers (Wiesemann 1986).

Logophorics are used with certain verbs related to speaking, especially the equivalent of 'say'. Verbs of thinking and feeling may also permit their use. In some languages, the logophoric pronouns can also be used with any verb, in which case the clause is one of intention or cause, which can be interpreted as truncated indirect speech.

Babungo (Grassfields, Cameroon)

- ɲwá nyiŋ lāā Kí vǝŋ sáN yí mē
3s_i flee that NEG 3p_k beat LOG_i not
'He fled so they would not beat him'

4.3 Sequential

In a sequence of clauses one may be the principal and the others subordinate to it. This relationship can be made explicit by the form of the subject pronoun. Any dependent clause in which the subject is coreferential with that of the principal one, the same subject (SS) is marked by a special sequential form of the pronoun. Or the converse may be true: any subject that is not coreferent with that of the main clause, the different subject (DS) is marked in some way. This relationship is also called switch-reference (Munro 1979).

Bafut (Grassfields, Cameroon)

- (a) á-ghèè ndá ŋ-kw érs fōrə
3s_i-go house SS_i-take rat
'... then he went home and caught a rat.'
- (b) á-ghèè ndá ŋ-kw érs fōrə
3s_i-go house gwa 3s_k-take rat
'... then he went home and gwa caught a rat'

In Bafut, the SS–DS difference is restricted to the third person subject. In Kobon (Papua New Guinea) all the persons have a different form for SS and for DS, as well as a 'normal' pronoun form. The actual use of such forms can be illustrated by:

kobon (Papua New Guinea)

- (a) kabö mjanu lau-ö in-a
stone that heat-DS3sg hot-3sg
'he heated the stones and they were hot'
(b) yad kaj pak-em ram ud ar-nab-in
1sg pig strike-SS2sg house take go-FUT-1sg
'I will kill a pig and take it home.'

Reflexives, Logophorics, and Sequentials can all be marked in the same language, either by the same markers or by different ones. In Mundani (Grassfields, Momo, Cameroon) the sequential pronoun is a homorganic nasal *N* in imperative and realis mode, and *e* in irrealis mode, the logophoric pronoun is *yè*, and the reflexive *vi*. They work together to keep the participants in a text straight (Parker 1986).

See also: Generative Syntax; Language and Gender; Linguistics: Incorporation; Serial Verb Constructions; Word Classes and Parts of Speech; Word, Linguistics of

Bibliography

- Delisle G 1972 Three universal personal pronoun features. *Stanford University Working Papers in Language Universals* 9: 181–3
- Everett D L 1996 *Why There Are No Clitics. An Alternative Perspective on Pronominal Allomorphy*. The summer Institute of Linguistics and The University of Texas, Arlington, TX
- Hopkins E B 1986 Pronouns and pronoun fusion in Yaouré. In: Wieseemann U (ed.) *Pronominal Systems*. Gunter Narr Verlag, Tübingen, Germany
- Hutchisson D 1986 Sursurunga pronouns and the special uses of quadral number. In: Wieseemann U (ed.) *Pronominal Systems*. Gunter Narr Verlag, Tübingen, Germany
- Ingram D 1978 Typology and Universals of Personal Pronouns. In: Greenberg J H (ed.) *Universals of Human Language*. Stanford University Press, Stanford, CA
- Munro P (ed.) 1979 *Studies in Switch Reference*. University of California, Los Angeles, CA
- Palakornkul A 1975 A sociolinguistic study of pronominal usage in spoken Bangkok Thai. *Linguistics* 165: 11–41
- Parker E 1986 Mundani pronouns. In: Wieseemann U (ed.) *Pronominal Systems*. Gunter Narr Verlag, Tübingen, Germany
- Wieseemann U 1986 Grammaticalized coreference. In: Wieseemann U (ed.) *Pronominal Systems*. Gunter Narr Verlag, Tübingen, Germany

U. Wieseemann

Property: Legal Aspects

Property encompasses the governance of resources—real property (land), natural resources, and personal property (chattels and intangibles, including copyrights, patents, stocks, claims of right). Natural law theory, beginning with Aristotle and perhaps best articulated in the writings of John Locke (1698), posits that 'life, liberty, and property' exist naturally as a

matter of fundamental principles of justice, irrespective of the state. Echoing Locke's formulation, the United States Constitution provides that a person may not 'be deprived of life, liberty, or property, without due process of law; nor shall private property be taken for public use, without just compensation.' By contrast, legal positivists, such as Jeremy Bentham (1789), assert that property rights arise only to the extent that they are formally recognized by the state: 'Property and law are born together, and die together. Before laws were made there was no property; take away laws, and property ceases.' Both views give content to what we understand as property. Legal positivism captures the practical reality that legislatures and courts define property rights in most societies, and that property rights have practical meaning only to the extent that they are enforceable. Nonetheless, natural law (and other philosophical frameworks) inform the content of property rules by providing the justification for state-created rights.

1. Philosophical Justifications for and Critiques of Private Property Rights

Waldron (1988) suggests that property is best viewed as a 'system of rules governing access to and control of material resources.' These rules are often characterized as a bundle of rights—in the case of land, the rights to possess (or occupy), to exclude, to use, to transfer, to enjoy fruits, and to alter or destroy—which can be allocated in different ways. Particular bundles represent different conceptions of property, which derive from one or more philosophical justification.

1.1 Labor Theory

John Locke (1698) offered a strong natural rights justification for private property, which remains a central pillar of property theory today. Beginning with the proposition that all humans possess property in their own 'person,' Locke argued that

[t]he 'labour' of his body and the 'work' of his hands, we may say, are properly his. Whatsoever, then, he removes out of the stat[e] that Nature hath provided and left it in, he hath mixed his labor with it and joined to it something th[at] is hi[s] own and thereby makes it his property. It being by him removed from the common state Nature placed it in, it hath by this labour something annexed to it that excludes the common right of other men. For this 'labour' being the unquestionable property of the labourer, no man but he can have a right to what that is once joined to, at least where there is enough and as good left in common for others.

Locke's theory is widely viewed as justifying first possession as a basis for acquiring property rights (Rose 1985, Epstein 1979). While useful as a basis for establishing rights in land and raw materials in a world of abundant unclaimed resources, Locke's theory significantly qualifies the laborer's claim in a world of scarcity—requiring 'enough and as good left in com-

mon for others.' Waldron (1988), Rose (1987), Ryan (1984), Becket (1977), and Schlatter (1951) elaborate and critique Locke's justification for private property rights.

1.2 Libertarianism

Drawing upon Lockean principles, libertarians emphasize individual freedom to control and alienate their private property. Nozick (1974) asserts that the state should play no role in distributing or redistributing property acquired properly, apart from respecting the voluntary transfers of property owners. Epstein (1985) argues for a broad interpretation of the government's obligation to compensate owners of private property for regulation and other governmental decisions interfering with private property rights.

1.3 Utilitarianism

First articulated by Jeremy Bentham (1789), utilitarianism judges the allocation of resources and the design of governance institutions on the basis of the utility or individual satisfaction that they produce. In applying this theory to property rights, utilitarian scholars have focused principally upon the inefficiency that can result from incomplete or shared resource ownership. In a now classic work, Hardin (1968) describes the 'tragedy of the commons,' whereby regulated access to a common pasture (or any other common resource such as air, water, or roads) results in overuse. Hardin suggested that private property or government regulation is needed to govern the commons. Economists generally prefer private property institutions (Posner 1998), although the efficacy of private property depends critically upon the costs of creating, monitoring, and enforcing such rights (Lueck 1995, Anderson and Hill 1975, Demsetz 1967). Government regulation often entails direct costs as well as inefficiencies attributable to rent-seeking by those interests affected by regulation. Although private property regimes often ameliorate the tragedy of the commons, shared governance institutions can maximize utility in a range of circumstances depending upon the cohesion and history of the community affected and the nature of the resources in question (Libecap 1994, Ellickson 1991, Ostrom 1990, Rose 1986).

1.4 Personhood

The personhood justification for private property, derived from Kant's *Philosophy of Law* and Hegel's *Philosophy of Right*, has been elaborated in modern legal discourse in the work of Radin (1982, 1993). 'The premise underlying the personhood perspective is that to achieve proper development—to be a person—an individual needs some control over resources in the

external environment. The necessary assurances of control take the form of property rights' (Radin 1982). The personhood justification for private property emphasizes the extent to which property is personal as opposed to fungible: the justification is strongest where an object or idea is closely intertwined with an individual's personal identity and weakest where the 'thing' is valued by the individual at its market worth. The personhood justification for private property is limited and highly subjective (Schnably 1993).

1.5 Distributive Justice

Theories of distributive justice seek to allocate society's resources on the basis of just principles. The process of determining such principles is the focus of considerable debate. Many theories incorporate utilitarian and Lockean principles. In perhaps the most influential work on distributive justice of the past century, Rawls (1971) offers an 'ideal contractarian' theory of distributive shares in which a just allocation of benefits and burdens of social life is determined by what rational persons would choose from behind a 'veil of ignorance,' which prevents them from knowing what abilities, desires, parentage, or social stratum they would occupy. Rawls concludes that people behind such a veil would adopt what he calls the 'difference principle': 'primary goods'—not only wealth, income, and opportunity, but also bases of self-respect—would be distributed to the maximal advantage of a representative member of the least advantaged social class. Rawls (1993) emphasizes that the difference principle does not call for simple egalitarianism but rather measures to assure that 'the basic needs of all citizens can be met so that they can take part in political and social life.'

1.6 Civic Republicanism

Republicanism rejects the view that individuals are wholly autonomous and that the public interest is just the aggregation of individual preferences. Republicanism stresses civic virtue and the role of participation in public discourse as essential to the realization of the common good. Civic republicanism relies upon the propertied independence of its citizens. Such ownership of capital affords citizens the competence to engage in meaningful discourse and protects against the subversion of democracy by a wealthy oligarchy or an aggressive state (Simon 1991).

1.7 Marxist Theory

Karl Marx viewed human history as an evolutionary process of class struggle leading, ultimately, to a classless society (communism) in which the state will wither away. In the capitalist phase of this evolution, private property provides the basis for the most important social relation: ownership of the means of

production. According to Marx, the class owning the means of production (bourgeoisie) acquires the property (capacity to work) of the labor class (proletariat) through the employment relation, thereby obtaining exclusive control of and hence power over the worker class. Such subordination, exploitation, and alienation sow the seeds of revolution, leading to the establishment of a proletarian dictatorship, socialism (productive resources owned by workers through participatory institutions), and eventually communism, a classless society without exploitation or inequalities in which property in land and rights of inheritance are abolished (Marx and Engels 1848).

1.8 Ecological Theories

As understanding of the connections between technology, land development, and the environment has deepened, many environmental scholars have increasingly viewed private property rights more skeptically (Sax 1983). Traditional reformist theories—which are anthropocentric in nature and draw heavily upon utilitarian and other traditional philosophical frameworks—favor internalization of the adverse impacts of development (Menell and Stewart 1994), Ellickson (1977, 1973). Building upon Leopold's (1949) call for a land ethic which seeks to 'preserve the integrity, stability, and beauty to the biotic community,' Freyfogle (1996), Sprankling (1996), and Sax (1993) propose a shift away from private property toward greater regulation and coordination of land and natural resource development. At the other extreme, Anderson and Leal (1991), drawing upon property rights and public choice literatures, advocate complete reliance on private property and free market principles as the best institutional choice to address substantially all environmental problems. Menell (1992) highlights the institutional limitations of both extremes and urges careful attention to the strengths and limitations of the broad spectrum of institutional alternatives in governing resources.

1.9 Pluralist and Hybrid Theories of Property Rights

The insights of the many philosophical perspectives have led some scholars to synthesize multiple strands into new and comprehensive theories for justifying and evaluating property. Munzer (1990), for example, constructs a theory of property built upon principles of utility and efficiency, justice and equality, and rewards based on labor. Simon (1991) synthesizes civic republicanism and socialism to advocate an intermediate form of property emphasizing enterprise and community-based democracy (cf. Hansmann 1996), offering a utilitarian explanation for worker owned and managed firms and organization).

2. The Emergence of Private Property Rights

Although no single philosophical theory provides a complete explanation for private property rights, utilitarianism has emerged as the dominant framework for understanding property rights systems. In a seminal article, Demsetz (1967) posited that private property regimes develop to internalize externalities. As the benefits of internalizing externalities rise (e.g., where a common resource becomes more valuable) or the cost of internalizing externalities fall (e.g., as technology reduces the costs of fencing or monitoring property rights), private property systems emerge to promote efficient management of the resource. As support for this thesis, Demsetz cites anthropological evidence suggesting that native Americans established private hunting territories following the arrival of Europeans and the development of the fur trade. By spurring the hunting of beavers, the fur trade threatened the productivity of this renewable resource, thereby creating social pressure to institute property rules that internalize the effects of resource exploitation.

Demsetz's theory explains the emergence of private property institutions across a broad range of contexts, including air (sulfur oxide emissions in the United States), water allocation, fishing, telecommunications spectrum, wetlands, roads (tolls), trademarks (causes of action for diluting famous marks), and publicity (commercial exploitation of a person's image, likeness, or voice). Nonetheless, by focusing exclusively upon private property institutions, Demsetz's theory overlooks much of the richness of property institutions and the complex social processes that affect their evolution. Libecap (1994), Ostrom (1990), Ellickson (1991), Rose (1986), and others contribute to the explication of a broad spectrum of property institutions lying between the poles of pure private property and the commons where informal and shared governance emerge and succeed in allocating resources.

The emergence and nature of property rights also reflect the political economy driving the legislative process and government agencies that administer property rights (Buchanan and Tullock 1962). Interest groups tend to form in those areas where government policies create concentrated benefits or impose concentrated costs (Olson 1965). Laws and agencies creating or affecting property rights (e.g., through regulation) attract substantial rent-seeking activity. Thus, the design of such property regimes often reflect political, as well as economic, considerations.

3. Evolution of Jurisprudential and Analytical Conceptions of Property

Although property rights are as old as human civilization, the modern jurisprudential conception of property traces its roots to the agrarian and early mercantilist economies that developed during the

decline of feudalism. Sir William Blackstone (1765), author of the first widely accepted legal treatise on English law, regarded property as an absolute right

which consists in the free use, enjoyment, and disposal of all acquisitions, without any control or diminution, save only by the laws of the land ...

So great moreover is the regard of the law for private property, that it will not authorize the least violation of it; no, not even for the general good of the whole community ...

Although some have questioned Blackstone's rigor and cogency (Burns 1985) few doubt his influence upon the jurisprudence of property in England and America. Maine (1861) reinforced the superiority of private property institutions through his historical account of the progress of civilization from primitive forms of shared ownership among kinship groups (status) to individual ownership and alienability of property interests (contract).

The industrial revolution and the shift from an agrarian economy to urban and other community structures changed the ways in which land and other resources were owned and exploited. The desire to restrict land through contract pushed courts to liberalize constraints on non-possessory interests, presaging the disintegration of the unitary bundle of sticks. Property came to be seen increasingly as a malleable concept that could be tailored to serve utilitarian purposes.

In the early twentieth century, Hohfeld (1913) developed a detailed descriptive framework for capturing the richness of legal rights reconceptualizing property as *relationships* among people relative to things. Recognizing four primary legal entitlements—rights, privileges, powers, and immunities—Hohfeld postulated four basic legal relations ('correlatives'): 'right-duty,' 'privilege-absence of right,' 'power-liability,' and 'immunity-disability.' Hohfeld provided the vocabulary and structure for representing the complexity of property relations—bundles of rights (and relations)—as well as the conceptual menu from which legislatures and courts could craft property rights.

Blackstone's conception of property as an inviolate, unified bundle of rights came under increasing assault with the proliferation of zoning and other public land use controls upon private property rights and the expansion of the role of government at all levels following the New Deal in the USA. Property rights, like much of the law governing the legal relations, came increasingly within the purview of political institutions, and accordingly gave way to the instrumentalist (utilitarian) nature of politics.

In a seminal article, Coase (1960) developed a new, instrumentalist framework for analyzing property rights. In much the way that a frictionless surface enables physicists to isolate and study the laws of physics, Coase hypothesized a world of zero transaction costs (costless bargaining and enforcement of

agreements), which provided valuable insights into the interplay of legal and market institutions. Coase demonstrated that in a world of costless transactions, an efficient allocation of resources will result regardless of how property rights are initially allocated. The desire of conflicting property owners to maximize their joint profit will lead them to bargain to an efficient allocation of resources. The initial allocation of entitlements will affect the distribution of wealth, but not the total level of wealth.

As with the frictionless surface, Coase recognized that costless transactions do not exist in the real world. His article fueled a deluge of scholarship exploring the implications of costly transactions for the allocation and design of property rights. Building upon Coase's construct, Calabresi and Melamed (1972) developed a framework that divides the analysis of property rights into two inquiries: (1) how should property rights be allocated; and (2) how should such entitlements be enforced—by a property rule (enabling property owners to enjoin interference with their entitlement and requiring consent for violation or transfer of the entitlement), a liability rule (allowing property owners court-determined damages for invasion of their property right, as in the case of eminent domain), or by a rule of inalienability (preventing transfer of a property right, as with prohibitions on the sale of body parts).

Calabresi and Melamed propose that entitlements and enforcement rules be based on economic efficiency, distributional goals, and other considerations of justice. Where transactions are costly, entitlements should be allocated so that the party who is best able to minimize harm (the least cost avoider) bears responsibility for the harm in order to promote economic efficiency. With regard to the enforcement of entitlements, Calabresi and Melamed demonstrate that the use of liability rules can minimize the costs of making an erroneous decision about the allocation of entitlements where transactions are costly. Subsequent scholars have refined and extended the framework and applied it in a range of contexts (Symposium 1997, Kaplow and Shavell 1996, Merges 1996, Radin 1996, Krier and Schwab 1995, Rose-Ackerman 1985, Polinsky 1980, Ellickson 1977, 1973).

In a provocative essay published in the early 1980s, Grey (1980) argued that the field of property 'ceases to be an important category in legal and political theory' as a result of the 'disintegration' of the conception of property. In his view, the dividing of property ownership into a 'bundle of rights' and the development of differing conceptions of property across disciplines has 'fragmented' property theory 'into a set of discontinuous usages.' Numerous scholars have questioned Grey's claims on a range of bases (Munzer 1990), yet the observation that provided the point of departure for Grey's essay, the 'disintegration' of property, resonates today. Property has lost the simplicity of Blackstone's conception—unitary, ex-

clusive, and omnipotent control—but a new conception of property integrating the many institutional structures governing resources, the cultural contingency of governance regimes, and the dynamism of governance institutions has emerged.

4. Property as a Triadic Relation

Menell (2001) and Dwyer and Menell (1998) present property as a dynamic system comprising three components: resources, institutions, and culture. Resources may be governed by background legal institutions (such as common law rules, or in civil law regimes, codes), social norms, markets (contractual agreements supported by the background legal institutions), political institutions (e.g., zoning), or some combination thereof. At any point in time, the mix of governance rules and institutions reflects the nature of the resources governed and the society or culture (knowledge, values, cohesion, technology) in which the resources are situated. Thus, in smaller, more cohesive social settings, we might observe more commonly particular resources are governed through social norms or other common property institutions whereas the same resources might be governed through formal private property institutions in a larger, more heterogeneous community. In this way, the triadic framework captures the economic and cultural dependency of resource governance as well as the richness of property regimes.

The triadic conception of property also incorporates the dynamic nature of resource governance. As reflected in Fig. 1, the framework incorporates causal connections and feedback mechanisms. Changes in the society (such as the value attributed to the protection of endangered species or technological advances reducing the costs of monitoring private property rights) generate changes in the design and mix of governance institutions. Such changes, by affecting the resource base, may generate further reactions within the overall system. The system may

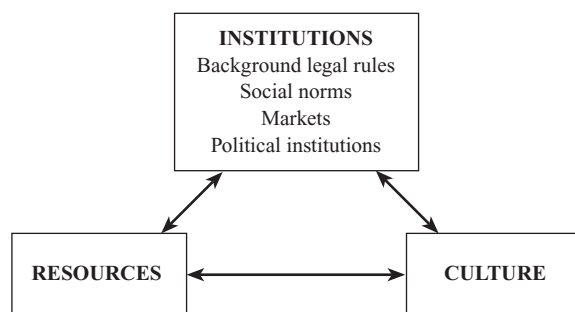


Figure 1
Property

also be affected by external shocks, such as charges in the larger political atmosphere. Demsetz's (1967) positive account represents a special case of this dynamic process, but the triadic relation generalizes the dynamism of institutional evolution to account for a broader mix of governance institutions.

5. New Frontiers

Notwithstanding the settlement of substantially all significant land masses on the planet, the frontiers of property continue to expand and evolve. Among the most important resources in today's economy are intangibles—ideas, genetic code, cyberspace, telecommunication spectrum (Merges et al. 2000). The internet has significantly expanded opportunities to distribute expressive works, but has also threatened the ability of copyright owners to control their works. This new medium has reshaped perceptions about rights to reproduce and distribute expressive works. Advances in bioscience have spawned disputes over rights and control of human genetic code, cell lines, and indigenous plants. And even in the more conventional realm of land management, increasing pressures on natural resources, greater knowledge of ecosystem effects, and changes in attitudes toward the environment continue to alter the balance among private property rights, public ownership, and regulation (Dwyer and Menell 1998, Menell and Stewart 1994). Public rights in artistic treasures, historical documents, architectural works, archeological discoveries, and scholarly research present a further set of challenges for a system of governance built significantly upon a foundation of private property rights (Sax 1999). Just as the advent of rights in government benefits expanded our conception of property four decades ago (Reich 1964), advances in technology and changing social values continue to expand and alter our conception of property.

See also: Intellectual Property: Legal Aspects; Legal Positivism; Natural Law

Bibliography

- American Law Institute 1936 *Restatement of the Law of Property*. American Law Institute Publishers, St Paul, Minnesota
- Anderson T L, Hill P J 1975 The evolution of property rights: A study of the American West. *Journal of Law and Economics* **12**: 163–79
- Anderson T L, Leal D R 1991 *Free Market Environmentalism*. Pacific Research Institute for Public Policy, San Francisco
- Bentham J 1789 *An Introduction to the Principles of Morals and Legislation*.

- Becker L 1977 *Property Rights: Philosophic Foundations*. Routledge & Kegan Paul, London
- Blackstone W 1765 *Commentaries*.
- Buchanan J, Tullock G 1962 *The Calculus of Consent: Logical Foundations of Constitutional Democracy*. University of Michigan Press, Ann Arbor
- Burns R P 1985 Blackstone's theory of the 'absolute' right of property. *University of Cincinnati Law Review* **54**: 67–86
- Calabresi G, Melamed A D 1972 Property rules, liability rules, and inalienability: One view of the Cathedral. *Harvard Law Review* **85**: 1089–128
- Coase R 1960 The problem of social cost. *Journal of Law and Economics* **3**: 1–28
- Demsetz H 1967 Toward a theory of property rights. *American Economic Review Papers and Proceedings* **57**: 347–58
- Dwyer J P, Menell P S 1998 *Property Law and Policy: A Comparative Institutional Perspective*. Foundation Press, Westbury, NY
- Ellickson R C 1973 Alternatives to zoning: Covenants, nuisance rules, and fines in land use controls. *University of Chicago Law Review* **40**: 681–781
- Ellickson R C 1977 Suburban growth controls: An economic and legal analysis. *Yale Law Journal*. **86**: 385–511
- Ellickson R C 1991 *Order Without Law: How Neighbors Settle Disputes*. Harvard University Press, Cambridge, MA
- Ellickson R C 1993 Property in land. *Yale Law Journal* **102**: 1315–1400
- Epstein R A 1979 Possession as the root of title. *Georgia Law Review* **12**: 1221–43
- Epstein R A 1985 *Taking: Private Property and the Power of Eminent Domain*. Harvard University Press, Cambridge, MA
- Freyfogle E T 1996 Ethics, community, and private land. *Ecology Law Quarterly* **23**: 631–61
- Grey T 1980 The disintegration of property. *Nomos* **22**: 69–85
- Hansmann H 1996 *The Ownership of Enterprise*. Harvard University Press, Cambridge, MA
- Hardin G 1968 The tragedy of the commons. *Science* **162**: 1243–8
- Hohfeld W N 1913 Fundamental legal conceptions as applied in judicial reasoning. *Yale Law Journal* **23**: 16–59
- Kaplow L, Shavell S 1996 Property rules versus liability rules: An economic analysis. *Harvard Law Review* **109**: 713–90
- Krier J, Schwab 1995 Property rules and liability rules: The cathedral in another light. *New York University Law Review* **70**: 440–83
- Leopold A 1949 *A Sand County Almanac*. Oxford University Press, New York
- Libecap G 1994 The conditions for successful collective action. *Journal of Theoretical Politics* **6**: 569–98
- Locke J 1698 *Second Treatise of Government*.
- Lueck D 1995 The rule of first possession and the design of the law. *Journal of Law and Economics* **38**: 393–436
- Maine H S 1861 *Ancient Law*.
- Marx K, Engels F 1848 *The Communist Manifesto*.
- Menell P S 1992 Institutional fantasylands: From scientific management to free market environmentalism. *Harvard Journal of Law and Public Policy* **15**: 489–510
- Menell P S 2001 Property as a triadic relation. (manuscript in progress)
- Menell P S, Stewart R B 1994 *Environmental Law and Policy*. Little, Brown, Boston, MA
- Merges R P 1996 Contracting in liability rules: Intellectual property and collective rights organizations. *California Law Review* **84**: 1293–393
- Merges R P, Menell P S, Lemley M A 2000 *Intellectual Property in the New Technological Age*, 2nd edn. Aspen Law & Business, New York
- Munzer S R 1990 *A Theory of Property*. Cambridge University Press, Cambridge, UK
- Nozick R 1974 *Anarchy, State, and Utopia*. Blackwell, Oxford, UK
- Olson M Jr 1965 *The Logic of Collective Action: Public Goods and the Theory of Groups*. Harvard University Press, Cambridge, MA
- Ostrom E 1990 *Governing the Commons: The Evolution of Institutions for Collective Action*. Cambridge University Press, Cambridge, UK
- Polinsky A M 1980 Resolving nuisance disputes: The simple economics of injunctive and damage remedies. *Stanford Law Review* **32**: 1075–112
- Posner R 1998 *Economic Analysis of Law*, 5th edn. Aspen Law & Business
- Radin M J 1982 Property and Personhood. *Stanford Law Review* **34**: 957–1015
- Radin M J 1993 *Reinterpreting Property*. University of Chicago Press, Chicago
- Radin M J 1996 *Contested Commodities*. University of Chicago Press, Chicago
- Rawls J 1971 *A Theory of Justice*. Harvard University Press, Cambridge, MA
- Rawls J 1993 *Political Liberalism*. Harvard University Press, Cambridge, MA
- Reich R 1964 The new property. *Yale Law Journal* **73**: 733–87
- Rieser A 1997 Property rights and ecosystem management in US fisheries: Contracting for the commons? *Ecology Law Quarterly* **24**: 813–32
- Rose C 1985 Possession as the origin of property. *University of Chicago Law Review* **52**: 73–88
- Rose C 1986 The comedy of the commons: Custom, commerce, and inherently public property. *University of Chicago Law Review* **53**: 711–81
- Rose C 1987 'Enough, and as good' of what? *Northwestern University Law Review* **81**: 417–42
- Rose C 1998 The several futures of property: Of cyberspace and folk tales, emission trades and ecosystems. *Minnesota Law Review* **83**: 129–82
- Rose-Ackerman S 1985 Inalienability and the theory of property rights. *Columbia Law Review* **85**: 931–69
- Ryan A 1984 *Property and Political Theory*. Blackwell, Oxford, UK
- Sax J L 1983 Some thoughts on the decline of private property. *Washington Law Review* **58**: 481–96
- Sax J L 1993 Property rights and the economy of nature: Understanding *Lucas v. South Carolina Coastal Council*. *Stanford Law Review* **45**: 1433–55
- Sax J L 1999 *Playing Darts with a Rembrandt: Public and Private Rights in Cultural Treasures*. University of Michigan Press, Ann Arbor
- Simon W H 1991 Social republican property. *UCLA Law Review* **38**: 1335–413
- Schlatter R 1951 *Private Property: The History of an Idea*. George Allen & Unwin
- Schnably S J 1993 Property and pragmatism: A critique of Radin's theory of property and personhood. *Stanford Law Review* **45**: 347–407
- Sprankling J P 1996 The antiwilderness bias in American property law. *University of Chicago Law Review* **63**: 519–89

Symposium 1997 Property rules, liability rules, and inalienability: A twenty-five year retrospective. *Yale Law Journal* 106: 2083–213

Waldron J 1988 *The Right to Property*. Clarendon Press, Oxford, UK

P. S. Menell

Property: Legal Aspects of Intergenerational Transmission

1. Wills and Succession Throughout History

Property exists essentially in two forms: movable (including intangible) and immovable (land). The transmission of that property at death generally occurs in one of three ways: by will, by customary or legal rules of succession, or by a combination of the two. An heir's right of succession by law or custom is an ancient and powerful right rooted in kinship ties. It was, for example, an attribute of early European feudal life, allowing family members to share that portion of the land vacated by a kinsman's death (see Chester 1982). By the twelfth century in England and many parts of Western Europe, lands were not partitioned, but descended to the eldest son of the landowner under the concept of primogeniture.

The right of a beneficiary to take property under a will is dependent on the legal power or right of the decedent to transmit this property at death to that beneficiary. This right was recognized relatively recently in Western thought and has often been regarded, at least by the positivist tradition, as having been created for society's convenience. Natural rights advocates by contrast might consider the right to bequeath an aboriginal attribute of property itself (see Chester 1998b).

In Western society, the legitimization of the right of bequest symbolized the shift from feudal to individualistic conceptions of property. Unknown to the early law of Greece, India, Egypt, Babylon, and Israel, the will was by no means a universal institution. The testamentary conception built upon by Europeans had its roots in Roman law and was resurrected after feudalism by ecclesiastical courts applying Romano-canonical law (see McMurray 1919).

In England the customary right of testamentary disposition was recognized in very early times, but land—the primary source of wealth—could not be willed until after the Norman invasion and the establishment of feudal tenures. Finally, in 1540, the Statute of Wills declared the right of an individual to leave real property to whomever the individual pleased by means of a duly executed will. This statute

confirmed a customary right long recognized for personal property (Chester 1982).

Once established, the right of an individual to leave property outside the family had to be reined in by providing protection from disinheritance for certain family members. For example, in England, primogeniture protected the eldest sons of landed decedents until well into the nineteenth century. Common law rights of dower and curtesy for surviving wives and husbands provided them with interests in the land of the decedent (see Chester 1982). Even though English women of the sixteenth, seventeenth and eighteenth centuries could not will property or own it in their own name, it could descend through them. However, married women's assets were also protected to some extent by courts of equity by means of trusts and other devices, especially toward the end of that period.

Children who were omitted from wills were not protected directly, except those benefiting under primogeniture. However, the wills of their parents were strictly scrutinized to ensure that they had not been left out unintentionally. One English rule derived from Roman law even stated that to cut off an heir 'without a shilling' raised a presumption that the testator was legally incompetent to make a will. Thus even in England, the family's right to inherited property was protected not only when the decedent died without a will (intestate) but to some extent when the decedent did leave such a will (see Chester 1982).

The entrenchment of the right of testation in England was associated with the movement from feudalism, with its emphasis on family and kinship, to free-market capitalism, with its emphasis on the individual. By the mid-nineteenth century, the adult male, Anglo-American owner of both realty and moveable property had nearly unlimited power of testamentary disposition. In contrast, countries such as France, Spain, and Germany on the European continent made the move to capitalism with a right of testation that, even when recognized, was much more sharply curtailed than that in Anglo-America by the requirement of fixed shares for surviving spouse and children.

In the eighteenth and nineteenth centuries, the principal difference between the Continental and Anglo-American systems of succession was the different attitude each displayed toward the right of family members to succeed to the decedent's property. Although the Spanish principle of mandatory heirship appeared in theory to be a significant curb on the Anglo-American concept of testamentary freedom, in practice the differences were not so marked. This reality stems from two separate phenomena: (a) in the majority of instances free testamentary power among the Anglo-Americans was exercised in favor of those who would be forced heirs under a mandatory system such as that in Spain; and (b) even where testamentary power appears absolutely undiluted, far more people

die without wills than with them, resulting in the passage of their property to those who, under Spanish law, for example, would be designated forced heirs (see McKnight 1996).

Spanish lawyers of the eighteenth and nineteenth centuries were trained in the Germanic tradition, which accepted only an extremely narrow testamentary power. Thus, while Anglo-American lawyers thought of both testate and intestate succession, the basic distinction for the Spanish was between those who were forced heirs and those who were not.

What is more, married women in the Spanish system not only had a much more extensive right to property than did those in Anglo-America, but they also had will-making power, which they did not have in Anglo-America. American women, for example, did not gain the power of will-making until the passage of the Married Women's Property Acts of the late nineteenth century.

In Spain the surviving spouse was also entitled to half of the gains of the marriage under what is today termed the community property system. Most of the property that was the decedent's alone passed to descendant forced heirs, despite any contrary testamentary provisions, and if there were no descendants, ascendants took as forced heirs, although in smaller shares than descendants. If the decedent died without a will, the estate passed to the forced heirs, or, if none, to other heirs as dictated by law (see McKnight 1996).

French law also maintained forced heirship under the concept of legitim. During the French Revolution, this concept was viewed as a means of eliminating the feudal principle of primogeniture in order to break up feudal estates. There was some relaxation in the strict adherence to forced heirship by the time of the Napoleonic Code, which announced the general French view of succession that is held even today. Under the Code, half of the testator's estate (after marital property has been distributed to the surviving spouse) was freely disposable if there was one child; one-third if there were two children, and one-fourth if there were three or more children (see McKnight 1996).

In general, since the early fifteenth century, the difference between Continental and Anglo-American views of succession, seen most strikingly in the Hispanic tradition, is that the Spanish adhered to the traditional Germanic tribal concept of familial right to property, whereas the Anglo-Americans took an individualistic view of who should succeed to the decedent's property (see McKnight 1996). By the mid-nineteenth century, this English individualistic notion had triumphed throughout the USA, with the exception of Louisiana which followed the French law of legitim. The Anglo-American system was, however, tempered by the concept of community property in states with a Spanish tradition and by dower and curtesy (mentioned above) in the others. The former

gave each spouse an immediate half interest in any property acquired during the marriage, while dower provided a surviving wife with a life interest in one-third of her dead husband's real property, and curtesy gave a surviving husband a life interest in all of the wife's real property.

2. Trusts

A particularly ingenious device for the intergenerational transfer of wealth was developed in Anglo-American courts of equity. Called the trust, it has sometimes been copied in other legal systems, for example that of modern Japan (see *International Law Digests* 1999). This form of property holding enabled the decedent to pass legal title to a third party who held the property for the benefit of its intended beneficiary. This third party trustee was subject to the equitable interest of the beneficiary, who generally received the income from the property for a period of time up to life. At the end of the designated period, the equitable title of that beneficiary would cease, transferring absolute ownership to that beneficiary or another chosen by the decedent.

The temporary division of legal and equitable title to property in trust allows, among other things, professional management of trust property assets. It also allows decedents to delay the actual vesting of full title to the property for a period of time up to and including a generation or more, during which period a beneficiary may receive a managed stream of income. Trusts in the USA that pay income to a decedent's children during their lives, and upon their deaths pay the principal to the grandchildren are called generation-skipping trusts. Estate tax advantages of these trusts (achieved because the children never received full title) were cut back by the Tax Reform Act of 1986 and subsequent amendments (which imposed a generation skipping transfer tax (see Johnson and Eller 1998, Dukeminier and Johanson, 1995).

3. Wills and Succession Among Nations in the Twentieth Century

The USA took freedom of testation one step further than did England and the countries of the British Commonwealth. While in varying degrees these nations developed flexible systems of family maintenance to protect surviving spouses and children (at least adult, competent children) from disinheritance, the USA maintained the right of testator to disinherit his or her children completely, even those children who were minors and/or had need. Thus, while surviving spouses in the USA are protected either by the elective share (in 'common law' states following the English tradition) or community property laws (in

those states following the Spanish tradition), children are theoretically left to the whims of unrestrained individualism.

Although this disparity has been attacked, it has also been noted that flexible concepts such as lack of testator competence, undue influence, and 'unintentionally' omitted children indirectly protect surviving children to some extent in most 'common law' states (see Chester 1997, 1998a).

That the USA's focus on the individual rights of the decedent, even to the exclusion of children and other descendants, is unusual in the world can also be seen by examining the systems of the Soviet Union (before its dissolution) and China. Traditionally, the Soviet law of inheritance did not differ substantially from that of the civil law countries of Western Europe (see Dukeminier and Johanson 1995). Although there was an attempt by the Bolsheviks to change this in their takeover of power in 1918, it failed. Nearly unique in the modern history of humankind, a 1918 Soviet Union law read: 'inheritance, testate and intestate, is abolished. Upon the death of the owner his property (moveable and immovable) becomes the property of [the state]' (1 *Sob. Uzak.*, RSFSR, No. 34, item 456, Apr. 26, 1918). This attempt to abolish inheritance proved so unpopular with the people that inheritance was reestablished within four years (see Foster-Simons 1985).

At present, inheritance law in Russia and the republics of the former Soviet Union generally follows the civil law views of Western Europe. In this way, it is similar to the law in modern China (see Dukeminier and Johanson 1995). In that country, if first order heirs (spouse, children and parents) inherit, then no second-order heirs (brothers, sisters, and grandparents) will inherit. Following a kind of community property system, all property in China acquired during marriage is joint property, unless the marital partners have executed a contract to the contrary (see *International Law Digests* 1999). Since much of China is still rural and nonindustrial, custom plays an important part in the actual distribution of estates (see Foster-Simons 1985), just as it does for example, in tribal Africa. This distribution depends partly on the law and the terms of any will and partly on the social and economic facts relevant to the intestate, or the testator and his heirs (see *International Law Digests* 1999).

4. Free Testation and Capitalism

It is tempting to view the relatively free testation of the USA as coincident with that country's status as the exemplar of capitalism. Except for the elective share (typically one-third) provided for surviving spouses, testation in America is predicated on protecting the free will of the testator. While England and the Commonwealth countries are certainly capitalistic,

their laws show, on the whole, more of a socialist's regard for protection of society (and family) than do those of the USA. While family protection is achieved in these countries by reposing a great deal of discretion in probate judges to ensure family maintenance (as distinguished from the use of fixed shares in the Continental system), this still provides a strong check on the whims of the testator. The governmental intrusion into testation is, of course, even pronounced on the continent where fixed shares for the family are the norm.

In its 1987 decision in *Hodel vs. Irving* (481 US 1987), the USA Supreme Court even seemed ready to protect *constitutionally* the right of an individual to transmit property at his death in any manner that he or she chooses. While that decision has been criticized and not extended beyond its facts (which involved government confiscation of noneconomic land parcels descending at the death of American Indians), it symbolizes America's fierce protection of individualism, even after the individual's death (Chester 1995, 1998b). Interestingly, the occasion for that decision was the government's attempt to save Indians from the disastrous economic effects of the Anglo-American individual property system which had been forced on the various tribes. Without imposition of the notion of individually held property in the first place, functioning Indian tribes, much like the tribes in sub-Saharan Africa, might have continued to treat property, which is chiefly in the form of land, as a family, and by extension community asset, which was never the individual's to transmit at death in the first place.

5. Implications of Inheritance and Taxation Policy for Equal Opportunity and the Distribution of Wealth

Allowing large amounts of capital to pass between generations without significant taxation is problematic for those countries such as the USA which profess a belief in equal opportunity. Taxation of estates and gifts is widely seen as an intrusion on the freedom to transmit property at death—a freedom that Americans particularly seem to cherish. While the American estate and gift tax has high nominal rates, its high exemption levels and numerous loopholes, including a deduction of up to 100 percent for all property left to a surviving spouse, may make it in effect a 'voluntary tax' (see Cooper 1979).

Countries such as France and Spain have both an inheritance tax and an annual tax on accumulated wealth. Perhaps not coincidentally, these countries do not have so strong a tradition as does the USA of allowing testators to do almost anything they want to do with their money at death.

Commentators (e.g., see Ascher 1990, Chester 1976, 1982, 1998a, 1998b) have noted that the policies of free wealth transmission and equality of opportunity

coexist very awkwardly in the USA. Although legal systems cannot control differences in natural ability or the luck of being born into a family that values, e.g., culture and education, they *can* curb that form of luck represented by inheritance. If equality of opportunity is truly a key goal, then children fortunate enough to have been reared and educated by wealthy parents need not be allowed the additional luck of inheriting their parents' property. Strong and effective taxation of inheritances would contribute significantly to this goal of equal opportunity.

Whatever the logic of this point of view, in countries like the USA there appears to be little public support for more effective taxation on postmortem (or even lifetime) transmission of wealth to heirs. In fact, at the beginning of this twenty-first century bills in Congress relating to the estate tax generally call for its softening or its abolition. Those supporting such bills argue that low or nonexistent taxation of inherited wealth is good for the economy because it preserves large blocks of investment capital in private rather than government hands. However, it might also be said that untrammelled inheritance saps the initiative both of those who receive it and of those who cannot possibly compete on equal terms with those who do (see Chester 1976, 1982).

6. The Future

In looking toward the future of intergenerational wealth transmission in the twenty-first century, one might begin by considering the changing nature of the family itself. Particularly in advanced capitalist nations such as the USA, the family is showing signs of disintegration. Traditional heterosexual marriages often end in divorce and several forms of non-traditional relationships are being legitimized. Since advanced capitalist nations offer many forms of wealth that are obtainable by individuals, there is no longer a critical need for children to succeed to a parent's land. Thus, the strict family protection practiced by the inheritance systems of many capitalist nations may seem outdated to some.

US law, unlike that of many nations, does not directly protect children from disinheritance by their parents. However, other nations may not desire to aggravate by such a legal stance the atomization of family and society toward which many of them are already heading. Their common judgment may be that, even in the twenty-first century, some homage to the family, traditional or otherwise, still needs to be paid in the field of intergenerational wealth transfer. Whether the restraints on individualism that such a judgment entails will become the norm as this century progresses remains an open question.

See also: Children and the Law; Family Law; Gender and the Law; Intergenerational Justice; Property: Legal Aspects

Bibliography

- Ascher M L 1990 Curtailing inherited wealth. *Michigan Law Review* **89**: 69–151
- Chester R 1976 Inheritance and wealth in a just society. *Rutgers Law Review* **30**: 62–101
- Chester R 1982 *Inheritance, Wealth, and Society*. Indiana University Press, Bloomington, IN
- Chester R 1995 Is the right to devise property constitutionally protected?—The strange case of *Hodel v. Irving*. *Southwestern University Law Review* **24**: 1195–213
- Chester R 1997 Should American children be protected against disinheritance? *Real Property, Probate, and Trust Journal* **32**: 405–53
- Chester R 1998a Disinheritance and the American child. *Utah Law Review* 1998 : 1–35
- Chester R 1998b Inheritance in American legal thought. In: Miller Jr. R K, McNamee S J (eds.) *Inheritance and Wealth in America*. Plenum, New York, pp. 23–43
- Cooper G 1979 *A Voluntary Tax?* Brookings Institution, Washington, DC
- Dukeminier J, Johanson S M 1995 *Wills, Trusts, and Estates*, 5th edn. Little, Brown, Boston
- Foster-Simons F 1985 The development of inheritance law in the Soviet Union and the People's Republic of China. *American Journal of Comparative Law* **32**: 33–62
- International Law Digests* 1999 Martindale-Hubbell, Summit, NJ
- Johnson B W, Eller M B 1998 Federal taxation of inheritance and wealth transfers. In: Miller R K, McNamee, S J (eds.) *Inheritance and Wealth in America*. Plenum, New York, pp. 61–90
- McKnight J W 1996 Spanish legitim in the USA—its survival and decline. *American Journal of Comparative Law* **44**: 75–107
- McMurray O K 1919 Liberty of testation and some modern limitations thereon. *Illinois Law Review* **14**: 96–123
- Rheinstein M, Glendon M A 1995 Inheritance and succession. In: *The New Encyclopedia Britannica (Macropaedia)*. Encyclopedia Britannica, Inc., Chicago, IL, Vol. 21, pp. 638–47

R. Chester

Property, Philosophy of

The philosophy of property concerns itself with the distribution of benefits, entitlements, and responsibilities associated with goods in a society. Most major Western political philosophers have constructed theories of property rights and property ownership including the limits and justification of property (e.g., Plato, Aristotle, Hume, Locke, Mill, Marx). Philosophical inquiry into property rights is closely associated with distributive justice, the role of the state, and the fundamental tensions between contemporary socialist and liberal political theory. Philosophical interest in property and property rights involves conceptual analysis of what property, own-

ership, and property rights are, as well as critical evaluation of the justifications given for property systems. In this entry the meaning of property, property rights, and ownership are addressed, before providing an examination of the contexts (historical and intellectual) within which property theories arise, an overview of competing approaches to justifying property and, finally, a discussion of current (and potential) debates concerning property in political philosophy and philosophy of law.

1. The Term 'Property'

Property is probably best understood relationally. Property rights involve relations among people concerning benefits and responsibilities which often are (but need not be) associated with particular objects in the world. In ordinary speech, however, the term 'property' is often used to describe the objects over which property relations are held themselves. Although it is common to speak of property as concerning objects in the world, philosophical discussions of property concern themselves with property understood as social, legal, and moral relations with regard to objects, benefits, responsibilities, and their control, within given legal and social contexts. The law of property concerns the legally recognized distribution of rights, responsibilities, and benefits associated with objects of property (both physical and immaterial). Social and moral relations of property, however, are not bound by legally recognized understandings of the limits and extent of property rights.

The range of legally recognized property rights and responsibilities vary from jurisdiction to jurisdiction. Similarly, social and moral property rights vary between cultures and across history. It can be difficult to identify whether a particular claim to a benefit or attribution of responsibility is best characterized as a property right or some other form of legally or socially recognized benefit or responsibility. In jurisprudence, clear distinctions between property, contract, and tort bases for legal entitlements or legally protected interests may be impossible to make (see *Property: Legal Aspects*). In this way, philosophical discussions of property are bound inevitably with broader discussions of distributive justice, social justice and with normative discussions of the role of the state in protecting or promoting certain kinds of interests.

The extent and limits on property rights vary along three dimensions: the range of objects over which property rights can be held (for example, whether water, air, or titles can be held as property); the range of specific property rights over an object that are recognized (for example, whether property can be bequeathed on death or whether property rights can be leased to others); and the entities that can bear specific property rights (for example, individuals and corporations, communities, the state). In capitalist

states, there is a presumption that anything could be conceived of as property and that anything which is property could be alienated, although there may be good moral grounds for the state to limit property rights in the public interest (for example, prohibitions against the sale of human beings, or prohibitions against the purchase of insurance against some legal responsibilities like legal culpability for criminal conduct). In pre-capitalist and socialist states there may be limits to the kinds of property which can be privately owned (for example there may be state or collective ownership of land, factories, or other means of production), with some presumption of a state interest in limiting private property rights for the purpose of redistribution of benefits and burdens through, for example, taxation, control over prices, limits on rights of sale, and control over wages.

1.1 Property as a 'Bundle of Rights'

In the conceptual analysis of property, it is generally understood as a term which refers to a bundle of rights, or rights and responsibilities. In liberal democratic theory and jurisprudence, specification of the bundle starts with the full array of rights and responsibilities enjoyed by a private owner of a thing. Minimally these are taken to include Hohfeldian rights of exclusion, rights of use, the power to alienate property and immunity from expropriation (Hohfeld 1919). Theorists have provided a number of ways of enumerating the different aspects (rights and responsibilities) of ownership, most are now heavily influenced by Honoré's (1961) list of the 11 'incidents' which together constitute 'full, liberal ownership.' These incidents are the rights of use, possession, management, income, capital, security, absence of term, residuary, and the responsibilities of prohibition of harmful use and liability to execution. Honoré's analysis applies most readily to property in material objects, but is intended also to apply to immaterial property such as future production, control over labor and copyright. The terms 'thing' and 'object over which property rights are held,' then, are not meant to exclude these forms of property.

Honoré's analysis allows that the constitutive rights which together make up ownership could be identified differently (that finer distinctions could be made, for example between the right to capital through sale and through consumption of property). He also recognizes that for any given object over which property rights may be held (for example, a car) the range of property rights which a person must hold to be deemed the owner of that object may differ from what is required to own another kind of object (for example, a pharmaceutical company). Other theorists provide distinct analyses of property which are less extensive, but draw on Honoré's list: for example, Becker (1977)

and Snare (1972) (see *Property Rights; Property: Legal Aspects*).

1.2 'Ownership'

Within liberal capitalism, philosophical discussion of property is often understood as a discussion of 'ownership'. Ownership can be seen as a kind of legal trump with regard to other property rights, but not necessarily with regard to other kinds of moral or social claim.

Many writers seeking to analyze the concepts of property and ownership as social and legal institutions have focused on the rights an owner has regarding a thing, with which nonowners may not interfere. 'Ownership' is the relation one is in (characteristically in a private property system), when one has 'the greatest possible interest' in a thing (Honoré 1961). That is it is the title given to the person who has the widest range of property rights over the thing. It may be more correct to say that the owner is the person who holds the most socially or legally significant property rights over a particular kind of thing. The specific property rights which are seen as most significant will vary with the kind of thing owned and the possible uses of those rights: rights of control over production and rights to income from profits will be relevant in identifying the owner of an industrial business, while rights over control of access and use may be more significant in identifying the owner of a retail space; rights of possession and use may be more important in identifying the owner of personal property (such as a car). In legal contexts, the person who is deemed to be an 'owner' has a stronger legal claim than any other holders of property rights over the thing. Nonetheless, there are many cases where it is unhelpful to try to determine who has rights of ownership, as property rights may be separated and limited in a huge number of ways. Further, the terms 'owner' and 'ownership' most naturally arise in discussion of private property. In the justification of property theories where relative merits of private, communal, state or mixed systems of property are to be assessed, it is frequently more helpful to refer to the range of property relations recognized within the theory, rather than attempting to specify relations of ownership.

1.3 Property Systems

Property systems, or the legal and social institutions concerning property a society are frequently divided into three kinds: private, collective, and common property systems. Waldron (1988) describes property systems as systems of 'rules governing access to and control of material resources' and distinguishes between private, collective, and common property

systems in terms of the different 'organizing idea' around which each kind of system is framed. 'In a system of private property, the rules governing access to and control of resources are organized around the idea that resources are on the whole separate objects each assigned and therefore belonging to some particular individual.' Collective property systems are characterized by the organizing idea or a 'social rule that the use of material resources in particular cases is to be determined by reference to the collective interests of society as a whole.' In common property systems, the organizing idea of the rules of allocation of rights to resources is the idea that 'each resource is in principle available for the use of every member alike' (Waldron 1988). For Waldron, kinds of property systems are distinguished according to their organizing ideas, yet Waldron's concept of an organizing idea for property systems is not meant to be reductive. Most property systems in actual societies contain elements of each system: some resources, e.g., sunlight, are understood as common property; some, e.g., state-run utilities, are held as collective property and some resources, e.g., clothing, are held privately.

2. The Context of Debates Surrounding Property

Within political philosophy debate about property rights and the justification of property are closely connected with debate about the legitimacy of state authority, national sovereignty, individual rights against the state and justice. Historically the shift from feudalism to representative democracies in Europe coincided with debate about the source and nature of the justification of property rights. Broadly speaking, feudal modes of production were agricultural. In feudal regimes all property in land was held by the sovereign and rights to use of that land were allocated in accordance with the sovereign will. The sovereign then levied taxes or required some form of payment for the use of the land and had ultimate authority to determine the use of the land and the income generated from the working of that land. Land obtained through conquest or treaty similarly came under the sovereign power. With the rise of representative democracies, the scope for exercising sovereign rights over land and determining taxes and other levies on the use of the land became constrained by the popular will and claims of the citizenry. What remained, however, was the idea that ultimate authority over grants of land, the use of land, and taxation rested with the sovereign, whether that be an individual monarch or the sovereign authority of a parliamentary government.

2.1 Property, Patriarchy, and God

An early debate about the source of legitimate property rights and its relation to political authority was

between Filmer (1680) and Locke (1690). Filmer had argued that the king's sovereignty could be traced directly back to God's gift of the earth and its bounty to Adam. The patriarchal lineage of the king gave him and no other person ownership of the world, and authority over all those living therein. Locke's refutation of Filmer rested on three propositions: the fundamental moral equality of all men; the gift by God of the earth and its fruits to all men in common (and not exclusively to Adam and his descendants); and that individual appropriation of property from the common could be legitimate, in the absence of the universal consent of all the other common owners. Filmer's argument tied the ownership of land and the political authority of the king together through Adam's sovereignty. Locke's argument separated the two, as he argued that legitimate property rights could be held in the state of nature (outside of political societies). However, he argues that political societies would not be entered into with the consent of the citizens unless the sovereign authority protected the pre-political property rights. Thus, the legitimacy of the state depends, in part, on its protection of property rights.

2.2 Property and Liberalism

Within liberal political philosophy more generally, the recognition, enforcement, and protection of private property rights is viewed as a function of the state (Rawls 1971), and as a source of its legitimacy. Private property is understood by many liberals as providing a realm within which individuals can exercise personal liberty, free from interference by others. If something is owned by a person, others must not interfere with the harmless exercise of the property rights held over the thing. Some liberal theorists, particularly those who emphasize the significance of 'negative freedom' argue that expansion of personal liberty depends on private possessions through which liberty can be exercised. Other liberals, however, recognize that the unequal distribution of private property rights limits the equal enjoyment of liberty and therefore argue that property rights must be limited to protect against gross inequalities in freedom (see *Freedom/Liberty: Impact on the Social Sciences*).

2.3 Equality and Justice

Because systems of property distribute goods among individuals within a state, discussion of property is closely linked with concerns about distributive justice and equality. Significant inequalities between individuals or groups which result from the unfettered exercise of property rights can be addressed within the justificatory structure of a property system (placing

limits on what can be owned or how property rights can be exercised) or through separate appeals to equality or justice within a broader political theory. Rawls (1971) provides the 'difference principle' as part of his two principles of justice to protect against significant inequalities which might emerge through the various political institutions of his just state. The difference principle can be applied to any form of inequality within the state; applied to property it holds that differences in property rights are justifiable only if they are of the greatest benefit to those who are least advantaged. This would allow some inequalities within a property system, such as those which are required to provide sufficient incentive for some people to engage in risky, dangerous, or socially stigmatized activities where those activities benefit the least advantaged. It would also require that there be some decent minimum level of welfare for all members of the society and would limit the extent of difference between the most and least advantaged (see *Justice: Philosophical Aspects; Equality: Philosophical Aspects*).

3. Property and Justification

Although conceptual analysis of property, property rights, and ownership have concerned philosophers and students of jurisprudence, by far the greater philosophical interest in property concerns its justification. The array of property rights identified by Honoré as jointly constitutive of full ownership could be justified as a whole or separately. Some justifications of property are for mixed systems; so some specific property rights are justified by appeal to one kind of grounding while others may be justified by appeal to another grounding.

Proudhon (1890) famously argued that no private property could be justified. In writing '*La propriété c'est le vol*' (property is theft), he argued that any plausible justification of private property would both demand equality and would, thereby, prove self-defeating. On Proudhon's view, whatever justified one person's exclusive right to a thing, equally justified a similar right for all others. Given the scarcity or at least physical limits on potential objects of property, exclusive rights to property entailed unequal rights, hence one person's exercise of private property rights constituted theft of another's (potential or actual) property.

Despite Proudhon's skepticism that any adequate justification of property could be developed, many political theorists have attempted to provide normative justifications for property. The justification of property rights has taken a number of (sometimes overlapping) forms, these include property rights justified by natural law; property rights justified by acquisition; property rights grounded in the desert of the laborer; property rights justified by appeal to

utility; and property rights as necessary for the development of personality.

3.1 Natural Law

Locke (1690) offers both a natural law theory of property and a labor theory of property in the Second Treatise. First he argued that the earth and its bounty were given to mankind in common by God, such that each individual had the right to take some of the common bounty of nature to provide for life, preservation and comfort. The individual comes to own what is taken from the common, through the exercise of his natural liberty, the labor of his body. In the state of nature, property rights are limited: fruits taken from the common are for the individual's use and must not be allowed to waste in a person's possession. Further, the appropriator must leave 'enough and as good' bounty in the common for others. In the transformation from the state of nature to the civil state, property rules are formalized and the state takes on the responsibility of enforcing the property rights of citizens. Macpherson (1962) and Tully (1980) offer two interpretations of Locke's argument: Macpherson arguing that Locke's argument is essentially possessive individualist, so that the natural law grounding of property is de-emphasized and the protection of exclusive private property is emphasized. Tully, in contrast, argues that natural law limits to property extend into the political state, thus providing significant limitations to individual acquisition.

3.2 Acquisition

Nozick (1974) offers a contemporary libertarian justification for property drawing on Locke's argument, with three significant differences from Locke's argument. For Nozick, the earth and its products are initially unowned. Each individual comes to have legitimate private property rights in those things which he or she has legitimately acquired. Without the premises of common ownership or the aim of legitimate acquisition being the life, preservation and comfort of the acquirer, Nozick's argument justifies largely unrestricted property rights and unlimited property acquisition. Nozick's sole limitation of property rights is what he calls his 'lockean proviso' property rights can be limited if 'the position of others no longer at liberty to use the thing is thereby worsened' (Nozick 1974, p. 178). The situation of others is not worsened, however, by the restriction of liberty which they experience as a result of the thing becoming privately owned. Moreover, 'worsening' the situation of others through acquisition requires compensation, it does not curtail acquisition of the thing in question. Nozick is notoriously vague about which acts constitute acts of legitimate initial acquisition of

unowned objects. Nozick's view is further criticized for failing to provide an argument for the unilateral imposition of duties on all other individuals which follows from an act of acquisition (see *Libertarianism*).

3.3 Labor and Desert

Contemporary arguments which justify limited property rights grounded in labor and desert can be found in Becker (1977) and more recently Munzer (1990). On both accounts, a person's (morally legitimate) labor to create, make available, or to improve something which was unowned grounds a claim to property. The person deserves some recompense for the labor expended, and this will often justify property rights in the improved thing or in the surplus value so generated. For Munzer, the laborers' use of their bodies in the production of something or in the provision of a service gives them the basis for a property claim on the product or on wages. These claims are *prima facie* claims, they are qualified by the needs or rights of others, and by scarcity. Munzer recognizes that the work involved in production is a social activity, and thus argues that payment of wages should be commensurate with desert. It is unclear, however, how desert is to be measured against labor, given the great differences between individuals in capacity, talent, and opportunity.

3.4 Utility and Efficiency

Hume (1740) and Mill (1891) offer utilitarian justifications for property. A system of property rights which distinguishes what is one person's from another's is required for the efficient use of the earth's resources. What the extent, limit, and scope of property rights should be, on a utilitarian account, is a matter of expediency and frequently involves balancing stability of expectations against efficient production and distribution. Hume favored custom and a fixed prescription of property rights, even if that approach allowed for considerable inefficiency. For Hume, inheritance and rights to alienate (sell, bequeath, or give) property acted as incentives for individuals to develop useful habits and accomplishments. Mill, by contrast, argued that the law of property ought to be used by the state to promote greater efficiency and a greater level of communal production of the goods needed within a society. Mill favored private property, but with limited rights of exclusive access and enjoyment of land so as to promote efficient exploitation. Mill recognized that property rights formed a bundle; each property right associated with ownership of a particular kind of property needed to be justified independently on the grounds of utility.

More contemporary utilitarian justifications of property concern themselves primarily with questions of efficiency in allocation of property rights. Economists in particular favor property systems grounded in efficiency, and argue for (somewhat moderated) capitalist relations of production and exchange as the most efficient way to respond to individual wants and needs. Philosophers emphasizing utility, rather than efficiency, are likely to justify mixed property systems (some public as well as some private property) which are roughly egalitarian, as the principle of utility can be directed to make interpersonal comparisons of individual preference satisfaction (see *Economics and Ethics*).

3.5 Property and Personality

Hegel (1821) argued that property is justified as an expression of human freedom. It is important to note, however, that Hegel does not ground property in natural freedom as natural law theorists, like Locke, sought to do. Property relations involve both a relationship between material objects and persons and relations among persons. In taking up an unowned object as property, the acquirer expresses his or her personality upon the object and gives the object value. In the absence of human actions of acquisition and use, material objects have no value and are useless. A human being can only develop—as a free moral agent—in the context of a community of other persons. To be a person is necessarily communal, it depends on recognition by others (Ryan 1984, p. 121). Each individual's will can only be recognized by others through the individual's impression of it on objects of property. Private property, then, is essential for personality. Ownership involves the impression of one's personality on what is possessed, to the exclusion of others. One's will or personality is embodied in the object. Waldron argues that Hegel's account is a general-right based justification for private property: that private property is something which it is important for every person to have, so that the poverty and propertylessness of some individuals is something a property system ought to protect against (Waldron 1988). Property rights justified on Hegel's account included possession, use, and alienation.

Radin (1993) offers a contemporary personality justification of property rights. Radin does not provide a fully developed argument for private property, but rather uses her argument to describe and evaluate existing property relations. Radin argues that at least some private property is required for the development of personhood, and that some things are so closely tied to personhood that they ought not to be commodified as property at all. Radin emphasizes a distinction between property relations based on personal attachment, property constitutive of personality as opposed to the more readily commodified, 'fungible' relations

of property. Radin's view is not entirely Hegelian, as it is strongly influenced by more traditional liberal conceptions of freedom and individualism, yet it is clearly distinguishable from utilitarian emphasis on economic efficiency and Lockean labor justifications of property.

3.6 Marxist Critiques

Most contemporary theorists writing on property, write against the backdrop of the Marxist critique of private property and its centrality to capitalist modes of production. Marx (1971) argued that private property systems presuppose and reproduce inequalities in power between those who have amassed sufficient wealth to become owners of capital (the means of production of goods) and those who are forced to sell their labor power in order to purchase those goods needed for life. Given the equal needs of all humans for the material necessities of life, workers who sell their labor power are open to exploitation by those who own capital. The capitalist pays the worker at the rate required for the subsistence and reproduction of labor, but sells the product produced through the worker's labor at a price which exceeds the cost of the labor and the cost of the resources required to produce the product. The value added by the laborer's efforts (surplus value) is taken by the owner of capital as profit. The degree to which workers are exploited depends on how much surplus value is appropriated by the capitalist. Further, capitalist modes of production involve various forms of alienation including the alienation of people from nature, of workers from the products of their collective labor, and of workers from their own labor power. This alienation, for Marx, is a distortion of human relations. The supposed expansion of freedom offered by private property which some liberal and libertarian theorists argue to defend private ownership, must be reconciled with the alienation and loss of freedom suffered by the propertyless under capitalism (Cohen 1981) (see *Analytical Marxism*).

4. Ongoing Debates About Property

Current debates about property rights raise both fundamental questions about the relationship between property rights and state sovereignty and contextual questions about the appropriate limitation or expansion of property rights. One source of this discussion is response to the demand for recognition of indigenous people's prior ownership of colonized lands in the New World and Australasia. Tully (1995), for example, argues that proper recognition of indigenous peoples, their forms of government, and their legal systems, including their property systems is required for the current legitimacy of states such as

Canada, Australia, and the USA. This area of debate ties the justification of property systems to state legitimacy as well as demanding an account of international relations between states and justice in recognition of other states' property systems.

Another current area of debate concerns conflicting interests in protection of property rights and protection of environmental values. Environmental ethicists have argued for greater limitation of property rights where those rights risk serious damage to the environment and have also argued for increasing the responsibilities of owners for long term damage to environmental values (Dodds 1994). Debate about whether 'pollution taxes' on factories that produce effluent unfairly constrain property rights or can act as an effective deterrent against pollution are among the issues concerning property and the environment.

A third area of considerable current debate about property involves the range of entities which are properly understood as property and as commodities, including ideas, human body parts, reproductive capacities, organisms, and electronic commodities. Radin (1996) challenges the ready extension of property rights allowing for sale and purchase of human tissue, embryos, sexual services, and other forms of property which she views as incompatible or threatening to personality. Other theorists argue that whether property rights are formally recognized over these things or not is irrelevant as there exists a market for these 'goods' and that will determine what is property, regardless of legal or moral attempts at constraining property (e.g., see *Market and Nonmarket Allocation*). Finally, with the growth of electronic commerce, a new field of property debate is also opening up as there will inevitably follow a need for legal determinations about property relations involving purely electronic commodities.

See also: Contractarianism; Economics and Ethics; Equality: Philosophical Aspects; Freedom/Liberty: Impact on the Social Sciences; Intergenerational Justice; Justice: Philosophical Aspects; Law and Morality: A Continental-European Perspective; Law and Morality: An Analytic Perspective; Libertarianism; Market and Nonmarket Allocation; Property: Legal Aspects; Property Rights; Rights; Welfare: Philosophical Aspects

Bibliography

- Becker L C 1977 *Property Rights: Philosophic Foundations*. Routledge & Kegan Paul, Henley, UK and New York
- Cohen G 1981 Illusions about private property and proletarian unfreedom. In: Mepham J, Hillel Ruben D (eds.) *Issues in Marxist Philosophy*. Harvester Press, Brighton, UK, Vol. 4, pp. 223–42
- Dickenson D 1997 *Property, Women and Politics*. Polity Press, Cambridge, UK

- Dodds S 1994 Property Rights and the Environment. In: Cosgrove L, Evans D G, Yenken D (eds.) *Environmental Values, Knowledge and Action*. Melbourne University Press, Melbourne, Australia, pp. 47–58
- Filmer R 1680 Patriarcha or the natural power of kings. In: Laslett P (ed.) *Patriarcha and Other Political Works of Sir Robert Filmer*. Blackwell, Oxford, UK
- Harris J W 1996 *Property and Justice*. Clarendon Press, Oxford, UK
- Hegel G W F 1821 *Grundlinien der Philosophie des Rechts*. Berlin [1942 Knox T M (trans. with notes) *Hegel's Philosophy of Right*. Oxford University Press, Oxford, UK]
- Hohfeld W N 1919 *Fundamental Legal Conceptions as Applied in Judicial Reasoning*. [reprinted 1978. Greenwood Press, West Port, CT]
- Honoré A M 1961 Ownership. In: Guest A G (ed.) *Oxford Essays in Jurisprudence (First Series)*. Oxford University Press, London
- Hume D 1740 *Treatise Concerning Human Nature* [1888 Selby-Bigge L A (ed.) Clarendon Press, Oxford, UK]
- Locke J 1690 *Two Treatises of Government* [1960 Laslett P (ed.) Cambridge University Press, Cambridge, UK]
- Macpherson C 1962 *The Political Theory of Possessive Individualism: Hobbes to Locke*. Clarendon Press, Oxford, UK
- Marx K 1971 *Critique of the Gotha Programme*. Progress Publishers, Moscow
- Mill J S 1891 *Principles of Political Economy*. Routledge, London
- Munzer S R 1990 *A Theory of Property*. Cambridge University Press, Cambridge, UK
- Nozick R 1974 *Anarchy, State, and Utopia*. Basic Books, New York
- Proudhon P-J 1890 Qu'est ce que la propriété? [1970 *What is Property?* Tucker B R (trans.) Dover Publications, New York]
- Radin M J 1993 *Reinterpreting Property*. University of Chicago Press, Chicago and London
- Radin M J 1996 *Contested Commodities: The Trouble with Trade in Sex, Children, Body Parts, and Other Things*. Harvard University Press, Cambridge, MA and London
- Rawls J 1971 *A Theory of Justice*. Harvard University Press, Cambridge, MA
- Ryan A 1984 *Property and Political Theory*. Blackwell, Oxford, UK
- Snare F 1972 The concept of property. *American Philosophical Quarterly* 9: 200–6
- Tully J 1980 *A Discourse on Property: John Locke and his Adversaries*. Cambridge University Press, Cambridge, UK
- Tully J 1995 *Strange Multiplicity: Constitutionalism in an Age of Diversity*. Cambridge University Press, Cambridge, UK
- Waldron J 1988 *The Right to Private Property*. Clarendon Press, Oxford, UK

S. Dodds

Property Rights

1. The Nature of Property Rights

A property right is the right to use a particular item of property in a particular way. An item of property can be used in more than one way, as long as those

different uses do not exclude each other. Therefore, there can be more than one property right attached to one item of property.

There are various types of property over which property rights can be exercised: water (e.g., extracting water from a river), intellectual property (e.g., authorship of a book), moveable property ('chattels'), and immovable property (land and buildings). For this article (property rights in connection with urban studies and planning), we restrict ourselves to immovable property (alternatively called landed property, real estate, or real property).

It is customary to classify the various types of rights in immovable property which can be attached to one item:

- (a) according to the type of use (e.g., easements);
- (b) according to the division of the space (i.e., special purpose boundaries within the general purpose boundaries of the item of property, such as an office in an office block);
- (c) according to the time (temporary, such as usufructs or leases, or in perpetuity).

There are some other classifications of property rights which cut across those given above:

- (a) superior/inferior (the owner of the superior right is entitled to acquire the inferior rights under certain circumstances);
- (b) alienable/inalienable (the owner of an alienable right is entitled to transfer it to someone else of their own choosing);
- (c) material/contractual (this distinction is derived from Roman law. Material rights, or 'rights in rem,' are attached to the property irrespective of the (legal) person exercising the right and can therefore be freely traded. Contractual rights, or 'rights in persona,' are attached to the property but may be exercised only by the (legal) person specifically named in connection with that right: they cannot, therefore, be traded).

2. Separate Property Rights

It follows from this that it is not an item of property which is owned or exchanged but rights in property and, because there can be many different rights attached to one item, it is possible that one right is owned separately from other rights in the same item (e.g., owning a lease to use a building 'owned' by another).

Where the property right may be traded (and this does not apply to rights which are inalienable and 'rights in persona'), this will have a market value equal to the 'present worth of rights to future benefits arising from ownership' (Jaffe and Sirmans 1989). The owner of more than one right over an item of property might have the choice between selling those rights to one person (keeping the bundle of rights intact) or to several persons (splitting or decomposing the rights). If the total value is greater undivided than divided, the

seller will probably decide not to split the rights. Conversely, if the rights over one item are divided and if the total value of all the rights would be greater if they were amalgamated (there is a 'marriage value'), then one person will try to gain possession of as many rights as possible.

3. The Right(s) of Property Ownership

Because it is a right in property rather than property itself which is owned, it is possible to conceive of a situation in which rights over one item of property were so dispersed that no one person could direct or manage the way in which the item was used. This is not found often: there are inferior and superior rights and there are temporary rights which revert to the owner of perpetual rights. Clearly, it is found to be socially (and economically) desirable that one (legal) person should be able to act as owner of the property (and for this purpose it does not matter whether the legal person is a private or a public entity).

In countries where the law on property is still strongly influenced by the feudal past, property ownership is regarded as being ownership of a full 'bundle of property rights' (fee simple, in Anglo-Saxon legal terms): the starting point is many rights in one item of property, which have been bundled and are in possession of one person. In countries where the Napoleonic legal code shapes the present system, the starting point is that property ownership is complete and that partial rights can then be split off from it. For example, in Dutch law, which is based on that code, ownership is defined as 'the most comprehensive right that a person can have over an object' (Civil Code, article 5.1).

The particular set of rights which constitute 'property ownership' varies from society to society. Nevertheless, it has been suggested that ownership of a property item always includes:

- the right to exclude others from that property;
- the right to dispose of (to alienate) that property;
- the right to derive income from that property.

4. Restrictions on the Exercise of Property Rights

There are many restrictions on the exercise of property rights, most of them accepted voluntarily by the owner when entering into a contract (e.g., someone acquires a building lease from the owner of the fee simple). In addition, there are restrictions imposed by the public administration. These latter can be politically very contentious, when the right(s) of property ownership are constrained.

The owner of an item of property might wish that, 'It is my land and I can do what I like with it' but in practice the legislature always restricts the right of

ownership. Again we take the example of Dutch law (Civil Code, article 5.1) where the definition of ownership quoted above is followed by:

- ‘the owner is entitled freely to use the object and to exclude others from that object, as long as this use does not infringe the rights of others, nor the constraints imposed by written and unwritten law’;
- ‘the owner of the object is owner also of the benefits arising from that object, apart from any rights that others may have.’

A restriction on property rights commonly found is expropriation (the state may acquire the ownership against the wishes of the owner). Another restriction is land-use planning by a public authority (see Sect. 7). It is interesting to note that although, under the US Constitution (5th amendment), ‘nor shall private property be taken for public use, without just compensation,’ the courts have ruled that public actions which restrict the rights of property owners, as long as this does not take 100 percent of the economic value of the property, falls under ‘police powers’ which allow regulation in the interests of ‘public health, safety and morals.’ Such ‘taking’ does not need to be compensated.

5. Property Rights as a Social Creation

In order that a property right can be exercised, it is necessary to delineate the right (define the limits of the property item and the limits of the particular property right), to register who is entitled to dispose of the right, and to determine how the right can be enforced or defended. Establishing those regulations is a matter for the legislature, enforcing them a matter for the courts. It follows from this that property rights are a ‘social creation’: there can be no property rights without agreement on what they entail and without the possibility of enforcing them legally. It is not that something is protected because it is a property right, but that something becomes a property right by being protected (Bromley 1998, p. 25).

From this it follows that property rights can be changed by act of the legislature. This can take the form of restricting an existing right. Introducing a law on noise zones provides an example: henceforth, the owner of land may build upon it only if noise levels on that land are lower than a certain level. Property rights can also be changed by identifying a use which was previously included in a wider right, and making of it a separate right. A well known example is the nationalization of development rights in Britain. The right to develop property was part of the right to own property: in 1947 the British parliament separated the right to development from the right to ownership, nationalized the right to development (which became ownership of the state), and compensated owners of the ownership right for the loss of the right to development. Yet another way of changing property

rights is to establish private rights over something which previously could not be owned. It has been suggested, for example (and see Sect. 7), that the owner of a right to use land in a particular location should have also the right to clean air at that location, whereby that is a right which could be traded.

6. Property Rights Regimes

The statement that property rights are a social creation does not mean that they can be created only by the formal legislature (the law-making institutions). Many societies had property rights before they had such a legislature. The Dutch, for example, set up water boards with legislative powers over rights to use land and water long before they had a general purpose government at any level.

Moreover, a society has often developed its formal legal systems in close connection with its system of property rights: the wish to define and enforce property rights has influenced the legal system. Alternatively, a society may reform its legal system deliberately in order to change the existing system of property rights (as happened after the French Revolution and in the United States after the declaration of independence).

It is for such reasons that it is common to talk about a society having a ‘property rights regime’ (or a land regime), an amalgam of law and custom as distinct from the formal law on property. A regime can be characterised by the way in which it regulates three aspects of property rights:

- who may exercise the right (and may exclude others from exercising it);
- how the right may be used (including restrictions on that use); and,
- how the right may be transferred.

(It will be seen that these are the same three aspects used to define the minimum set of rights that constitute property ownership—see Sect. 1.) Two sorts of regime have already been mentioned: the Anglo-Saxon and the Napoleonic. Others have distinguished in addition (‘legal families’ in Europe, see Newman and Thornley 1996): the Germanic, the Scandinavian, and the East European regimes.

It is sometimes claimed that there are social laws which explain the creation of property rights regimes and changes in them (that is, the regimes do not arise accidentally). These laws have the form of a ‘positive’ theory which can be tested empirically. In an influential article, Ellickson (1992/3) put forward an ‘efficiency thesis’ that ‘land rules within a close-knit group evolve as to minimise its members’ costs.’ The costs he identifies as ‘deadweight’ (net external costs of individual actions) and transaction costs. A change in land rules is efficient when it reduces the sum of these two sets of costs.

7. The Politics of Property Rights

Because a property rights regime is a social creation, it can then be evaluated socially, with a view to changing it in some desired way. That is not new: property rights have always attracted intense political attention.

One strand in the political argument may be called the dyadic view of property rights: private versus public. At one level, the issue is: who should own property rights (in particular, who should own land), the individual (or the family unit) or some collectivity such as the state? Some regard private ownership of property as a bulwark against state despotism, at the other extreme are those who claim (with Proudhon, 1809–1865) that ‘property is theft’ and that it should be owned collectively.

At a more refined level, the issue in most present-day Western societies is: to what extent and in what ways should the state intervene in the exercise of private property rights? Most Western societies agree that the private exercise of property rights cannot be unconstrained (see Sect. 4) and even the most fervent defenders of private property rights usually want the state (or some collectivity) to own land and build roads upon it if this gives access to their land.

Another strand in the political argument about property rights regimes can be called triadic. Then there are not two parties—a private person versus the state—but three—one person who, in the exercise of their property rights can affect another person, and the state as an arbiter in that private/private relationship. Most property law and custom is, in fact, civil law: a collectivity regulates the relationships between (private) legal persons in the use of their property.

Much of the current debate about property rights and land-use planning can be interpreted in terms of dyadic/triadic. To the extent that land-use planning by a public body is justified in terms of welfare economics, it is usually the arguments of Pigou (1932) which are invoked. The market in which private persons exchange property rights has certain characteristics which mean that a welfare optimum (in a Paretian sense) will not necessarily be reached. There are ‘market failures.’ This is seen as a problem, which is resolved when a public body intervenes in the market for property rights. Examples include: restricting the right to use a plot of land in the way that the land owner wants, expropriating the ownership right, public ownership to provide ‘public goods.’

In 1960, Coase argued that market failures arise mainly because of transaction costs and because property rights are not clearly delineated. Examples of this latter are the right to clean air (see Sect. 5) and the right to graze cattle on common land. Incomplete separation of rights creates common property, i.e., in the public domain (Barzel 1898) and, if these rights are valued, resources are spent on capturing them. This can lead to rent-seeking and ‘common-property dissipation’ (of which the best known description is

Hardin’s (1968) ‘tragedy of the commons’). According to this argument, the problem of market failures can be tackled not only by government intervention to restrict property rights (Pigovian measures: moreover, these can make matters worse by creating ‘public failures,’ see Wolf 1979) but alternatively by public measures to enable markets in property rights to function well, to reduce transaction costs and to delineate exclusive property rights.

The economic argument is that helping the market in private property rights to work better will reduce risks, which will stimulate investment, which will achieve a better allocation of resources than through public land-use planning. The political argument is that a property rights regime should be created in which—in dyadic terms—most property rights are held privately and in which—in triadic terms—the role of the public administration is to regulate private transactions in property rights in ways described above. If this is not possible (because, for example transaction costs are high and cannot be reduced, as is the case with roads in urban areas) then ‘Pigovian’ measures can be justified. This approach to land-use planning is called the ‘property rights model’ (see e.g., Lai 1997).

See also: Land Rights and Gender; Land Tenure; Property: Legal Aspects; Property, Philosophy of

Bibliography

- Barzel Y 1898 *Economic Analysis of Property Rights*. Cambridge University Press, Cambridge, UK
- Bromley D W 1998 Rousseau’s revenge: The demise of the freehold estate. In: Jacobs H M (ed.) *Who Owns America?* The University of Wisconsin Press, Madison, WI, pp. 19–28
- Coase R H 1960 The problem of social cost. *Journal of Law and Economics* 3: 1, 1–44
- Ellickson R C 1992/3 Property in land. *Yale Law Journal* 102: 1315–400
- Hardin G 1968 The tragedy of the commons. *Science* 162: 1243–8
- Jaffe A J, Sirmans C F 1989 *Fundamentals of Real Estate Investment*. Prentice Hall, Englewood Cliffs, NJ
- Lai L W C 1997 Property rights justification for planning and a theory of zoning. In: Diamond D, Massam B H (eds.) *Progress in Planning* 48: 161–246
- Newman P, Thornley A 1996 *Urban Planning in Europe*. Routledge, London
- Pigou A C 1932 *The Economics of Welfare*. 4th edn. Macmillan, London
- Wolf C J 1979 A theory of non-market failure, framework for implementation analysis. *Journal of Law and Economics* 22: 107–39

D. B. Needham

Prophetism

Prophetism may be defined as a religious current or movement promoted by the activities of a prophet. A prophet in turn may be defined as a leader of or key participant in a movement or group who proclaims a religious mission on the basis of personal revelation. While many prophets have been founders of religious movements they are not necessarily so but may be significant figures within a revitalization of a religious group or simply play a central role in the religious life of a group or community by pronouncing upon religious matters through revelatory insights or experiences. Neither is the foretelling of future events a necessary component of prophecy but is not uncommonly associated with it. A prophet may or may not convey a message from the gods or be the mouthpiece of divinity. They may express the word of God or the gods or may simply convey religious truths to their followers, which are not claimed to stem from such supernatural sources. Buss (1980) puts the point usefully in the following terms: 'The role of a prophet or seer can be characterised as the receiving and transmitting of communications not available to ordinary conscious sensitivity which are held to come from a source (the divine) or through a form of perception transcending normal spatio-temporal limitations.' Prophets may or may not claim spiritual or supernatural powers themselves and may or may not receive revelation through the experience of states of trance or possession.

What is essential to the notion of prophecy, then, is the relatively emotional proclamation of a religious message that has been revealed personally in a non-routine and nonmundane manner. Prophets are called to their mission and compelled to follow it whether voluntarily or, as is not uncommon, very reluctantly.

The category of prophet as a religious specialist or type of religious figure is one which owes much to Weber who gives it a central role in his sociology of religion (see *Religion, Sociology of*). For Weber a prophet is 'a purely individual bearer of charisma, who by virtue of his mission proclaims a religious doctrine or divine commandment' (1965). Weber distinguishes the prophet from the category of priest. The priest commands religious authority on the basis of a position within a sacred tradition or religious organization and dispenses salvation by virtue of holding an office while the prophet commands religious authority on the basis of personal charisma and recognition of a personal calling to a mission through revelation. The term 'charisma' is used by Weber to refer generally to extraordinary powers essentially of a magical or non-mundane nature (see *Charisma and Charismatic*). The priest for Weber is thus the upholder and guardian of tradition while the prophet is a religious innovator or at least a radical revitalising force (see *Religious Specialists*).

Weber is careful to distinguish the prophet from a number of other types of religious specialist. While the authority and following of prophets may be based upon magical charisma they are not primarily magicians or workers of miracles. The prophet is concerned primarily with doctrine and commandment. It is this that distinguishes the prophet from the mystagogue whose following is based purely upon attribution of magical powers. Yet it is rare that prophets are not expected to give proof of their particular gifts and such proof has more often than not taken the form of possession of magical abilities. Reflecting the fact that charisma may be either inherent or acquired through undergoing extraordinary experiences or practices, these capacities may be inherent in the person or may be acquired through, e.g., ascetic regimes or cultivation of ecstatic states.

Again, while prophets engage, in many instances, in divinatory activities, as was the case, e.g., with the Biblical *nebiim* (singular *nabi*), they may be distinguished from diviners on the basis that it is normal for the latter to be economically recompensed for their services. Prophets provide their spiritual revelations gratuitously and, indeed, often unbidden. To do otherwise would be to undermine their distinctive claim to religious and spiritual influence.

Likewise, Weber distinguishes the prophet from related or similar religious specialists such as legislators, arbitrators, social reformers, orators and teachers on the grounds that lacking in the case of all of them is the proclamation of a religious truth on the basis of personal revelation. Later usages of the term 'prophet' are often less precise and looser than Weber's but in general at the root of them is this specific emphasis upon a purely personal revelatory impulse and calling.

Weber distinguished between two fundamentally different types of prophecy. In the 'ethical' or 'emissary' type, exemplified most clearly by figures such as Zoroaster or Muhammad, the prophet is an instrument for the proclamation of the message or commandments of a god. The doctrines proclaimed by such a prophet take the form of ethical duties. In the 'exemplary' type, most clearly exemplified by the Buddha, the prophet proclaims a soteriological message by personal example and demonstration of the means or path to salvation which others are free to follow to their personal benefit if they wish or ignore to their personal detriment.

In both cases the prophetic revelation acts to promote a meaningful understanding of life in relation to the whole of reality to which conduct must be brought into alignment if salvation is to be attained. It involves the construction of a cosmos as a meaningfully ordered totality. As Parsons (1965) says 'the prophet is above all the agent of the process of breakthrough to a higher, in the sense of more rationalised and systematised, cultural order, an order at the level of religious ethics, which in turn has

implications for the nature of the society in which it becomes institutionalised.' While there is such a stress in Weber's characterization of the prophet upon the construction of a more rationalized and systematized meaningful order, prophecy, Parsons is careful to point out, is not fundamentally an intellectual or theological enterprise but frequently takes emotional and sometimes even violent forms.

Weber's most extensive analysis of specific prophetic activity and prophetic movements occurs in his study of the development of Judaism, in his *Ancient Judaism* (1952), a development in which the Biblical prophets are given a crucial role (see *Judaism*). Weber traces the emergence of a new type of prophet towards the end of the period of the confederacy that he distinguishes from the earlier ecstasies or *nebiim* and the *roeh* or oracle givers. These new prophets saw themselves as called upon to transmit the will of Yahweh to the people and to the rulers. Early prophets of this type did not pronounce unless solicited but progressively they spoke independently and increasingly critically of the political authorities. From the time of Elijah onwards prophets spoke as much to the masses as to the rulers culminating in a situation in which the prophets became the intellectual leaders of the opposition against all that monarchy was perceived to stand for—centralization of power and government, heavy taxation, and social injustice.

Not that Weber considered social justice to be their primary concern. It was the destiny of the state and of the people that motivated their essentially religious yet inextricably political message. It was their purely religious concerns with fulfilling the commandments of Yahweh that brought upon them the opposition of the authorities. For Weber the Biblical prophets represented the values and ethos of the presettlement confederacy.

Classical Biblical prophecy in Weber's analysis is characterized by independence, spontaneity, and disregard for popularity among any sector of Israelite society. It was indifferent to economic, material, or other reward in contrast to other types dependent for their livelihood upon prophecy and prediction. They were at best tolerated and not infrequently persecuted.

While not primarily diviners and seers the classical Biblical prophets nevertheless stressed the dire fate that Israel would suffer if it did not keep the terms of its covenant with God. Such was the precarious situation of the Israelite states in Palestine that gloomy predictions were not uncommonly fulfilled. The religious conduct of the Israelite people was linked intimately by the activities of the prophets to their material and political fate and to that of the state.

The conditions which promoted the development and impact of prophetic currents of this type were those of military and political insecurity and unpredictability. Weber considers that new religious insights rarely occur in the main centers of civilization but rather in areas on their periphery which are subject

to their influence and pressure. Those who live in such regions repeatedly are given reason to question the meaning of events that affect them. Unlike the inhabitants of long stable, politically secure, and settled regions who tend to take the world for granted, those outside these areas constantly are puzzled by the turn of events; 'the possibility of questioning the meaning of the world presupposes the capacity to be astonished about the course of events' (Weber 1952).

Subsequent research on Biblical prophecy has often been critical of aspects of Weber's analysis and by implication of his analysis of prophecy and prophetism in general. Berger (1963) and Turner (1974) have questioned whether the Biblical prophets were the independent force based upon pure charismatic authority and motivated entirely by religious concerns that Weber portrays them as being. In fact, the classical Biblical prophets are more likely, according to some recent research, to have occupied official positions in the cult of Yahweh. The conclusion that writers such as Turner draw from this, that it requires us to reconsider Weber's understanding of prophetic charisma as an autonomous and personal force and to recognize that it arises from the incumbency of positions in an established social institution is, perhaps, rather extreme. It is likely that there were several different types of prophet in ancient Israelite society. Emmet (1956), who considers that Weber equates prophecy with charismatic authority which she finds too restrictive, usefully differentiates a number of types of prophet. First, there is the institutional prophet who fulfils a traditional and established role. Second, corresponding to Weber's charismatic prophecy, there is the leader of a movement which seeks radical change. Finally, there are prophets who do not attract a personal following but who act as independent critics, advocates of reform or who provide moral leadership. While the prophets associated with the Temple in Jerusalem, whose function was an institutionalized one within the Temple cult, were perhaps more numerous, the classical Biblical prophets with which Weber was concerned may be considered to be of a different type. Whether they enjoyed greater autonomy and independence stemming from personal charisma is still, however, open to question.

Weber's stress upon the charismatic authority of the classical Biblical prophets has been challenged by Petersen (1979) on the grounds that there is no evidence that they were leaders with a group of followers, which he claims is implied by Weber's notion of charismatic authority. Neither is it useful, in his view, to characterize Biblical prophets as office holders. The dichotomy between charismatic leader and office holder is, in any case, he argues, far too strictly drawn. We are led, therefore, to question whether the Israelite prophets could have provided the breakthrough that Weber attributes to them. In a later work Petersen (1981) argues that we may get a better

understanding of the Biblical prophets if we see them as having a specific role. He applies role theory in an analysis of their position in Israelite society based upon a close examination of the use of a variety of terms in Hebrew used to refer to such figures. The results show that there was more than one type of prophetic role in this society. Some held established positions in the official cult and operated also by offering a personal service for a fee. Others, termed central morality prophets, pronounced as individuals at times of crisis but in order to uphold established values and institutions. Still others, peripheral prophets, operated outside institutional contexts and established values.

Also doubtful of the simple dichotomy of functionary vs. autonomous charismatic is Lang (1983). The prophetic tradition, he argues, was largely an institution of the land-owning nobility that enabled them to express publicly their political views. He attempts to show that the prophets were largely from a land-owning, noble background. While they were not chosen by formal procedures but simply emerged, they nevertheless, according to Lang, had to learn how to be an acceptable prophet and to observe the rules governing appropriate behavior for prophets. They were closer, however, to Emmet's third type, in Lang's view, and not wholly established functionaries within the cult. The contrast between charisma and tradition in discussions of prophetism is indeed often overstated. Lang points out that they are often bound up with one another in complex ways such that prophetic charisma can be seen as a force for renewal which emerges within traditional structures at certain times. It is a force that changes and reforms without destroying.

While the Weberian view of Biblical prophecy has, in recent decades, been questioned by some, it has for others stood the test of time and deeper scrutiny rather more robustly. Williams' analysis of sources (1969), e.g., supports the view that many of the classical prophets were independent of the official cult. A detailed study by Wilson (1980) is alert to the complexities and differences between periods and regions. He contrasts the premonarchical Ephraite tradition of the North with the monarchical period in both the Northern and Southern kingdoms and also different periods and circumstances in the South. In the premonarchical North, prophetic behavior was expected to conform to familiar patterns regarding the stereotypical linguistic style in which pronouncements were made and to customary behavioral patterns. This clearly suggests that these prophets occupied a defined position and role within the central religious institutions of the society. In the monarchical period prophets seem to have become figures of less central significance but rather more peripheral to the established religious institutions and rather more concerned with producing change. In Judah, however, political circumstances and alliances differentially

affected the type of role and significance that prophets had. Some, standing outside the established cult, were more critical of the status quo while others exercised a more supportive function. This reminds us that prophetism, at least of this type and in these circumstances, may not simply refer to a movement or development led by a prophetic figure but to a dynamic situation of struggle to define or lay claim to the accepted interpretation of a religious tradition or interpretation. Lang (1983) points out that, like many prophets in various traditions and cultures, Jeremiah engaged in conflict with rival prophets opposed to his policy of accommodation with the Babylonians and who advocated total resistance.

Such debates highlight the extent to which Biblical prophecy, and it is an observation applicable to all prophecy, is associated with crises which, while differing extensively in kind, generally call forth a questioning of the status quo and of established institutions and patterns of power and authority. This point has been made by Carroll (1979) in the case of ancient Palestine and Weber's analysis. The religious-political factions that tend to emerge in such circumstances may understandably become aligned with pro-regime and anti-regime types of prophecy.

Not that this necessarily means that prophetism is really a form of politics under the guise of religion. As Albrektson (1972) notes, Weber may well be correct in attributing to the Biblical prophets motives which were primarily religious. Circumstances often made their pronouncements political in tone but at root they were essentially utopian and idealistic rather than pragmatic or designed to stir up political opposition for its own sake.

The political element in prophetism is a factor that comes to the fore most notably in studies of those forms of prophetism that can be described as millennial. It is in the context of the quite extensive studies of such movements that prophetism has been most systematically examined (see *Millennialism*). Not all prophetic movements are millennial and not all millennial movements are prophetic, though most are. It is prophetic millennial movements, however, that have received most attention (see, e.g., Wilson 1975). The role of the prophet in millennial movements varies. Such movements may be initiated by the visions, revelations, and activities of a prophet. Once underway they may call forth a prophet who focuses relatively disordered and uncoordinated activities of participants and who gives coherent expression to previously vague, inchoate, and diverse ideas. The appearance of the prophet transforms millennial ferment into a millennial movement.

The role of the prophet in the millennial movement generally has been found to reflect Weber's emphasis upon prophetism being associated with social and political crisis or radical social change. On the other hand, it is rarely a complete break with the past and more often a process of bridging past and present, the

traditional and the new. The prophetic message in such movements tends to reinterpret old understandings in ways appropriate for an altered situation while at the same time adding new elements and opening up the imagination to new possibilities. The prophet gives concrete expression to what are often inarticulate feelings and ideas and channels them into a program of action promising redemption from current ills. The prophet helps those who are confused by circumstances of change, uncertainty, and instability to comprehend and grasp the meaning of their situation by offering insight derived from special revelation which through understanding points the way forward for remedying what is unsatisfactory. Prophetic movements frequently are found in circumstances of disaffection in which a population is subject to foreign or colonial rule or have little comprehension of the mechanisms and structures of power or how to operate within them in order to ameliorate their situation. The prophet provides the missing knowledge and means of finding solutions. It is significant that many prophetic leaders of such movements are more knowledgeable than their followers as a result of wider experience through travel, limited education, or access to information. The prophet, by providing a focus, is often able to bring together otherwise fragmented communities and diverse groups previously unable to operate in concert in the pursuit of common aims. What is promised, however, is not simply material but also usually, and more significantly, moral and emotional redemption through the recovery of lost personal and cultural dignity and sense of self-worth.

See also: Charisma and Charismatic; Charisma: Social Aspects of; Millennialism; Monasticism: West and East; Religion: Definition and Explanation; Religion, History of; Religious Organizations; Religious Specialists; Weber, Max (1864–1920); Weberian Social Thought, History of

Bibliography

- Albrektson B 1972 Prophecy and politics in the Old Testament. In: Biezais H (ed.) *The Myth of the State*. Almqvist and Wiksell, Stockholm, Sweden
- Berger P 1963 Charisma and religious innovation: the social location of Israelite prophecy. *American Sociological Review* **28**: 940–50
- Buss M J 1980 The social psychology of prophecy. In: Emerton J A (ed.) *Prophecy*. W. de Gruyter, Berlin
- Carroll R P 1979 *When Prophecy Failed: Cognitive Dissonance in the Prophetic Traditions of the Old Testament*. Seabury, New York
- Emmet D 1956 Prophets and their societies. *Journal of the Royal Anthropological Institute* **86**: 13–23
- Lang B 1983 *Monotheism and the Prophetic Minority: An Essay in Biblical History and Sociology*. Almond Press, Sheffield, UK

- Parsons T 1965 Introduction. In: Weber M (ed.) *The Sociology of Religion*. Methuen, London
- Petersen D L 1979 Max Weber and the sociological study of ancient Israel. In: Johnson H M (ed.) *Religious Change and Continuity*. Jossey Bass, San Francisco
- Petersen D L 1981 The Role of Israel's Prophets. *Journal for the Study of the Old Testament*, Supplement Series, **17**
- Turner B S 1974 *Weber and Islam*. Routledge, London
- Weber M 1952 *Ancient Judaism*. Free Press, Glencoe, IL
- Weber M 1965 *The Sociology of Religion*. Methuen, London
- Williams J G 1969 The social location of Israelite prophecy. *Journal of the American Academy of Religion* **37**: 153–65
- Wilson B R 1975 *Magic and the Millennium*. Paladin, St Albans, UK
- Wilson R 1980 *Prophecy and Society in Ancient Israel*. Fortress Press, Philadelphia, PA

M. Hamilton

Prophylaxis in Psychiatry

In psychiatry, prophylactic drug therapy includes the use of medication for both prevention and maintenance. For the sake of clarity prevention is measures used in order to prevent the occurrence of a first illness-episode in a group of high-risk individuals (primary prevention), or the occurrence of subsequent illness episodes following the occurrence of a first one (secondary prevention). While secondary prevention is used in a variety of psychiatric disorders, primary prevention, especially for the psychotic disorders, is still under investigation. In order to familiarize the reader with both the concept of primary and secondary prevention, the following paragraphs will discuss a variety of strategies suggested for implementation of primary and secondary prevention focusing on their use in schizophrenic illness.

In the absence of current curative treatments in psychiatry, the use of maintenance therapy—which is associated with long-term control of symptoms and illness-episodes—is especially important. Some of the current drug-prophylactic-maintenance strategies will be described that are used to control abnormal behavior and symptoms in three major psychotic disorders: schizophrenia, major depression, and bipolar illness. Since a knowledgeable choice of a proper maintenance drug treatment necessitates a set of predictors for the success or failure of a given drug, this issue will be exemplified below while discussing maintenance treatment for bipolar illness.

Drug therapy, whether preventive or maintenance, depends on patient compliance. This in turn depends on several variables and certain measures can be taken to maximize compliance. This issue will be discussed as well.

Finally, future directions in prophylactic medication treatment for psychiatric disorders will be discussed focusing on strategies for primary prevention.

1. *Primary Prevention in Schizophrenia*

There is as yet no consensus as to the nature of a subgroup of individuals carrying a high risk for the development of schizophrenia. Research designs used toward this aim involve both retrospective and prospective methods.

Prospective research projects mainly apply long-term follow-up of selected high-risk groups of subjects. Such prospective strategies assume that biological and psychological data obtained from subjects who will later develop the disease will help to identify markers associated with this disease. This, it is hoped, will enable a knowledgeable prediction as to who will develop a schizophrenic illness. One example of such a prospective approach is the follow-up of offspring of schizophrenic patients or their family members (Mednick et al. 1987, Weintraub et al. 1987, Asarnow et al. 1988, Erlenmeyer-Kimling et al. 1993). Another is the examination of psychomotor and neuropsychological variables in large cohorts of individuals while identifying variables associated with later development of psychotic illness and schizophrenia. Yet, another prospective strategy is the study of a variety of factors and early symptoms associated with a transition phase—that is, the phase leading from ‘normal’ into a psychotic state—in order to identify relevant risk factors. All the above strategies require long-term follow-up. Interestingly, there is also a suggestion for a short-term prospective strategy employed by Yung et al. (1998) aiming at identifying a very high risk group of subjects for the development of psychosis. This strategy utilizes a 6–12 months follow-up of selected subgroups of individuals demonstrating high risk for a manifestation of a schizophrenic episode.

In general, retrospective and prospective studies have identified a host of markers increasing the risk for future psychosis including a family history of schizophrenia, distressed family functioning, early separation and institutional rearing, neurobehavioral deficits, delivery complications, maternal exposure to influenza during the second trimester of pregnancy, along with certain behavioral and personality patterns. The latter are mainly reported by teachers and caregivers. Also, a variety of biological findings were identified, all of which still need to be verified in further research. An example of a biological finding is that of small amygdala-hippocampal complex or thalami in people at high risk of developing schizophrenia (Lawrie et al. 1999). Another finding is that of increased plasma homovanillic acid that is negatively correlated with performance in arithmetic tasks in a group of subjects

demonstrating prodromal phase of schizophrenia (Shimushi et al. 2000). All the markers discussed above should be taken at the moment with caution until further research is done. Also, it is our opinion that the use of certain combinations of the above markers for the initiation of antipsychotic treatment in high-risk individuals—although suggested by some (Phillips et al. 1999)—is not recommended at this time as routine treatment.

2. *Maintenance Treatment in Schizophrenia*

Maintenance therapy in schizophrenia is provided by long-term treatment. Based on accumulated experience, it was suggested that schizophrenic patients could be looked at as two groups. One in which maintenance therapy seems to play a ‘suppressive role,’ and its discontinuation leads to a rapid re-appearance of psychotic symptoms, and another group in which the treatment has more of a ‘prophylactic’ role. In this last group discontinuation of treatment leads to a more gradual and remote in time relapse of symptoms (see Gelenberg et al. 1991).

It is estimated that about 10 percent of remitted medication free schizophrenic patients will relapse in any given month. That means that after 6 months roughly about half of these patients will relapse whereas after 12 months about 70 percent will do so and only a small portion of these patients will not relapse at the end of 2 years.

Several strategies of maintenance treatment are suggested for these patients. The first consists of a continuous maintenance treatment. This strategy applies relatively lower doses of antipsychotic treatment as compared with that used for acute exacerbation.

The second maintenance strategy includes ‘drug holidays,’ that is, treatment periods interrupted by drug free periods. This strategy was suggested in order to lower the incidence of side effects while retaining drug efficiency. However, such a strategy is not recommended nowadays since research studies could not support its effectiveness. A third strategy suggests ‘intermittent’ treatment or ‘targeted dose’ for chronic schizophrenic patients. In this case, patients are treated during the acute episode only, while monitored for the appearance of prodromal symptoms (first symptoms anticipating the appearance of an exacerbation). Once prodromal symptoms appear, pharmacological treatment is initiated.

In general pharmacological treatment in schizophrenia includes the use of a group of drugs collectively called ‘antipsychotic drugs.’ These drugs have the ability to block neurons using the neurotransmitter dopamine to transmit messages to other neurons.

In the 1980s, a second generation of drugs came into use. The prototype was clozapine, a drug that besides its ability to block some types of dopaminergic

neurons can also exert a blocking effect on certain serotonergic neurons—(labeled as 5HT-2 neurons) in the central nervous system. This drug was shown to have a lower side-effect profile compared to the more traditional drugs, especially it had minimal extrapyramidal signs (a set of abnormal motoric signs that interfere with daily activities) and it did not elevate plasma levels of prolactin (the lactating hormone). However, clozapine also has several distressing side effects including the lowering of white blood cell count that may expose these subjects to severe infections. Clozapine seems to affect not only the positive illness symptoms but also the negative symptoms of schizophrenia (a set of symptoms characterized by impaired volition and motivation, apathy and paucity of speech) and it may also improve some neurocognitive impairments demonstrated in this disease (see Kane et al. 1988). Beside clozapine, other second-generation antipsychotic drugs were also developed of which mainly olanzapine and risperidone are in use. However, there is still a question whether all these second generation antipsychotic drugs or only some of them can mimic clozapine's effect and if so to what extent.

3. Maintenance in Bipolar Illness

This illness course is characterized by the appearance of manic episodes with or without the occurrence of depressive episodes. It is estimated that the majority of patients who had a manic episode will have one or more recurrences. Since both manic and depressive episodes may be associated with progressive deterioration of interpersonal and occupational functioning and may also carry the risk of harming oneself or others, an adequate maintenance drug treatment for bipolar illness is of utmost importance. Several mood stabilizers—that is, drugs designed for maintenance of the euthymic (normal) state while preventing the appearance of both manic and depressed episodes—are now available. The first mood stabilizer to be used was lithium, a drug still considered by many to be the gold standard treatment for bipolar illness. Other commonly used mood stabilizers are valproic acid and its derivatives, and carbamazepine whereas newer drugs such as the anticonvulsants lamotrigine, gabapentin, and topiramate are still under study. This article will mainly address data published for lithium.

Three open trials suggested that lithium may be effective in the prophylaxis of bipolar disorders (Baastrup 1964, Hartigan 1963, Baastrup and Schou 1967). During the 1970s, a series of double blind studies were conducted to examine the possible efficacy of lithium in the prophylaxis of bipolar disorder. Goodwin (1994) summarized these studies calculating the recurrence rates over a one year period for placebo (a nonactive treatment) to be 81 percent, but only 34 percent for lithium. These reports included two types of studies. Prospective studies treating patients with

either lithium or placebo and discontinuation studies where one group of patients continues with lithium treatment for a given period of time, whereas the other discontinues treatment for a similar time period.

From the 1980s and on more prospective and discontinuation studies were conducted. These studies basically confirmed lithium's efficacy as maintenance treatment for bipolar disorder. They also suggested that there may be an advantage in keeping maintenance lithium blood levels at about 0.6 meq/L as it may lead to less side effects and increase compliance (Coppen et al. 1983) and that lithium plasma levels of 0.8–1.0 meq/L seem to be associated with less relapse (Gelenberg et al. 1989). In general, although many authors still suggest that lithium is an efficient drug for bipolar illness, especially in the classic bipolar patient (Grof 1999), some authors now point to the fact that there is a group of patients who are less responsive to this drug or that demonstrate a decreased response throughout the years.

One important issue associated with long-term maintenance therapy is the finding of predictors for a beneficial, or lack of, response for a given drug. Several predictors of lithium response were offered. Grof et al. (1983a, 1983b, 1999) suggested a better response to lithium treatment in a group of patients demonstrating a course of illness with periods of euthymia (normal range of mood), followed by mania, depression, and then euthymia. These authors also suggested that the absence of rapid cycling (four and more illness episodes per year), the absence of personality disorder, and the diagnosis of primary bipolar disorder (i.e., not secondary to street drugs or organic cause) predict a good prophylactic response to lithium. Faedda et al. (1991) in agreement with Grof suggested that mania-depression-euthymia course is associated with a good response to lithium. Others suggested that good response to lithium in relatives and a family history of mood disorders carry a good response. Aagaard and Vestegaard (1990) suggested that bipolar patients tend to demonstrate decreased response to lithium treatment when they have a lengthy episode, they are young or of female gender. Abou-Soleh and Coppen (1990) suggested that positive prophylactic response to lithium over 6–12 months predicts further good response in the future. Patients with dysphoric and mixed mania (manic episode associated with some depressive symptomatology) were suggested to respond less to lithium treatment and some studies suggest this is also true for rapid cycling bipolar disorder. Hopefully, a knowledgeable use of predictors may help the physician to choose selected subgroups of patients, which will benefit from lithium maintenance treatment.

4. Maintenance in Major Depression

Depression is a serious illness. It carries high suicidal risk. It is estimated that the recurrence of major depression is between 70–85 percent. Two types of

studies report the efficacy of antidepressants to prevent relapse in the first year after remission from a depressive episode. The first is that of a cross-over design, whereas the second is that of a continuation design. In the first design patients are first brought into remission with a given medication and are then maintained on this medication for a period of 2–3 months. Patients who respond are then entered into a double blind design for up to another year, in which they receive randomly either the same antidepressant or placebo. In the continuation design, acutely depressed patients are assigned randomly to either active drug or placebo. Patients who respond continue with the same treatment. Eighteen studies compared a group of antidepressants collectively called tertiary amine tricyclics (this name describes the common molecular structure characterizing this group of drugs), finding that while 50 percent relapsed on placebo only 23 percent relapsed on the active drugs. Three studies compared another group of newer antidepressants collectively called selective serotonin reuptake inhibitors (SSRIs) (this name describes the hypothesized mechanism of action of this group of drugs) finding that the relapse rate for the active drug was between 13–26 percent whereas for placebo it was 46–57 percent (see Janicak et al. 1997).

In general, it is agreed that prevention of relapse in major depressive illness is of major importance and every effort must be taken to address this. Contrary to previous reports that supported the use of low doses in the prophylactic treatment of major depression, several authors now recommend that the full dose be used, that is, a similar dose to that used for the acute treatment of a major depressive episode.

In general, it is recommended to continue maintenance treatment after a first episode of major depression for about 6 months and then to taper the drug gradually. If another episode occurs the recommendation is to continue treatment for several years and after a third episode a lifetime maintenance treatment is recommended.

5. *The Issue of Compliance and its Importance in Prophylactic Drug Treatment*

Compliance to drug treatment is defined as adherence to an agreed treatment schedule between a physician and the patient. It is estimated that no more than 50 percent of patients fully comply with their treatment and we suspect that the magnitude of non-compliance may be even greater in the case of long-term maintenance treatment. Various factors seem to influence compliance including the nature of the patient-physician relationship, the patient's set of beliefs and expectations as well as the occurrence of side effects. Issues relevant to compliance in psychiatric disorders are: (a) the occurrence of delusions and severe cognitive distortions about the index drug, its effect and

side effects, (b) the patient's cognitive functioning and motivation, which may be impaired in certain psychiatric disorders, and (c) the patient's general attitude toward mental illness. Especially important in this context is the level of stigma felt by the patient; the greater the stigma, the less compliant the patient will be.

Several remedies can be offered to enhance compliance and they should be tailored according to the patient's condition and level of understanding. First, an empathic approach, one that builds basic trust with the patient and their family is of utmost importance. Second, whenever possible, the patient with the guidance of their physician should be an active participant in decisions made for the treatment plan. This does not mean that the patient decides or guides such decision making, but that they are presented with several treatment alternatives and an educated decision is agreed upon with the guidance of the physician. Third, the patient should be educated as to side effects and ways to cope with them once they appear (i.e., lower the drug by a certain degree and contact your physician). Treatment diaries and involvement of family members whenever necessary can also be of help. Since maintenance treatment is a long-term treatment, it is our opinion that a responsible physician should renew such a drug treatment 'agreement' at least every several months, discussing it with the patient and perhaps with relevant family and community members if needed. The physician should also actively ask in each meeting with the patient about side effects as well as examine whether there is any change in the patient's set of beliefs about the drug and its action that may need renewed education.

6. *Future Directions for Primary Prevention of Psychiatric Disorders*

Genetic epidemiological studies have shown that genes play a major role in several psychiatric disorders including schizophrenia, bipolar illness, and major depression. It is now believed that these disorders are not related to a single gene, but rather are determined polygenetically. In general, it seems that gene-environment interactions are involved in the causation of many psychiatric disorders. Interestingly, the vulnerability to environmental influences may also be genetically determined (see Owen et al. 2000 for review).

The human genome project, now completed, is expected to contribute to the finding of genes responsible for the susceptibility, occurrence, and maintenance of these disorders. The sorting out of genetic variations in the population, which are related to these psychiatric disorders, may open new avenues to genetic medical interventions prior to the manifestation of these diseases. Such interventions may alter the course or even prevent these diseases from

occurring. Although this project is still far off, one cannot ignore it at the dawn of this new era, which started with the recent completion of the decoding of the human genome.

See also: Antidepressant Drugs; Bipolar Disorder (Including Hypomania and Mania); Depression; Drugs and Behavior, Psychiatry of; Patient Adherence to Health Care Regimens; Psychological Treatment, Effectiveness of; Psychological Treatments, Empirically Supported; Public Psychiatry and Public Mental Health Systems; Schizophrenia; Schizophrenia and Bipolar Disorder: Genetic Aspects; Schizophrenia, Treatment of

Bibliography

- Aagaard J, Vestegaard P 1990 Predictors of outcome in prophylactic lithium treatment: a two-year prospective study. *Journal of Affective Disorders* **18**: 259–66
- Abou-Saleh M T, Coppen A J 1990 Predictors of long term outcome of mood disorder on prophylactic lithium. *Lithium* **1**: 27–35
- Asarnow J R 1988 Children at risk for schizophrenia: Converging lines of evidence. *Schizophrenia Bulletin* **14**: 613–31
- Baastrop P C 1964 The use of lithium in manic depressive psychosis. *Comprehensive Psychiatry* **5**: 396–408
- Baastrop P C, Schou M 1967 Lithium as a prophylactic agent. *Archives of General Psychiatry* **16**: 162–72
- Coppen A, Abou-Saleh M, Milln P, Bailey J, Wood K 1983 Decreasing lithium dosage reduces morbidity and side effects during prophylaxis. *Journal of Affective Disorders* **5**: 353–62
- Erlenmeyer-Kimling L, Cornblatt B A, Rock D, Roberts S, Bell M, West A 1993 The New York High Risk Project: Anhedonia, attentional deviance and psychopathology. *Schizophrenia Bulletin* **19**: 141–53
- Faedda G L, Baldessarini R J, Tohen M, Strakowski S M, Waternaux C 1991 Episode sequence in bipolar disorder and response to lithium treatment. *American Journal of Psychiatry* **148**: 1237–9
- Gelenberg A J, Bassuk E L, Schoonover S C (eds.) 1991 *The Practitioner's Guide to Psychoactive Drugs*, 3rd edn. Plenum, New York, pp. 170–6
- Gelenberg A L, Kane J M, Keller M B, Lavori P, Rosenbaum J F, Cole K, Lavelle J 1989 Comparison of standard and low serum levels of lithium for maintenance treatment of bipolar disorder. *New England Journal of Medicine* **321**: 1489–93
- Goodwin G M 1994 Recurrence of mania after lithium withdrawal. *British Journal of Psychiatry* **164**: 149–52
- Grof P 1999 Excellent lithium responders people whose lives have been changed by lithium use. In: Birch W J, Galliccio V S, Becker K W (eds.) *Lithium: 50 Years of Psychopharmacology*. Weidner Publishing Group, Cheshire, CT
- Grof P, Alda M, Grof E, Fox D, Cameron P 1999 The challenge of predicting response to stabilizing lithium treatment. *British Journal of Psychiatry* **163**(21): 16–19
- Grof P, Hux M, Grof E, Arato M 1983a Prediction of response to stabilizing lithium treatment. *Pharmacopsychiatry* **16**: 195–200
- Grof P, Lane J, Hux M 1983b Predicting the response of long term lithium treatment: Interim results of a prospective study, in the prediction of lithium response. In: Dufour H, Pringuey D, Milech T (eds.) *Proceedings from Symposium*

- in Marseille, France; June 2–3, 1980. France University, Marseille, pp. 43–53
- Hartigan G P 1963 The use of lithium salts in affective disorders. *British Journal of Psychiatry* **109**: 810
- Janicak P G, Davis J M, Preskorn S H, Ayd F J 1997 *Principles and Practice of Psychopharmacotherapy*, 2nd edn. Williams & Wilkins, Baltimore, MD, pp. 295–301
- Kane J, Honigfeld G, Singer J, Meltzer H 1988 Clozaril collaborative study group: clozaril for the treatment-resistant schizophrenic: a double-blind comparison with chlorpromazine. *Archives of General Psychiatry* **45**: 789–96
- Lawrie S M, Whalley H, Kestelman J N, Abukmeil S S, Byrne M, Hodges A, Rimmington J E, Best J J, Owens D G, Johnstone E C 1999 Magnetic resonance imaging of brain in people at high risk of developing schizophrenia. *Lancet* **353**(9146): 30–3
- Mednick S A, Parnas J, Schulsinger F 1987 The Copenhagen High-Risk Project, 1962–1986. *Schizophrenia Bulletin* **13**(3): 485–95
- Owen M J, Cardo A G, O'Donovan M C 2000 Psychiatric genetics: Back to the future. *Molecular Psychiatry* **5**: 22–31
- Phillips L, Yung A, Hearn N, McFarlane C, Hallgren M, McGorry P 1999 Preventive mental health care: accessing the target population. *Australia and New Zealand Journal of Psychiatry* **33**: 912–17
- Shimushi T, Kurachi M, Kurokawa K, Yotsutsugi T, Uehara T, Itch H, Saitoh O 2000 Plasma homovanillic acid in the prodromal phase of schizophrenia. *Biological Psychiatry* **47**: 428–33
- Weintraub S 1987 Risk factors in schizophrenia: The Stony Brook High-Risk Project. *Schizophrenia Bulletin* **13**(3): 439–50
- Yung A R, Phillips L J, McGorry P D, McFarlane C A, Francey S, Harrigan S, Patton G C, Jackson H J 1998 Prediction of psychosis. A step towards indicated prevention of schizophrenia. *British Journal of Psychiatry* **172**(Suppl.): 14–20

J. Levine and S. Gershon

Proportional Representation

As a standard for making political evaluations, 'proportional representation' implies that members of a group (R_g) participate in a governmental entity in numbers that bear the same ratio to the total membership of that body (R) as persons identifying with the group (P_g) have to the total population governed (P)—thus, $R_g/R = P_g/P$. The group in question may be supporters of a political party, adherents of an ideological or policy tendency, or persons sharing an identity based on race, ethnicity, religion, language, or gender. 'Proportional representation' also, and more commonly, refers to methods of electing representatives designed to achieve, or at least approximate, the ideal of proportionality in converting citizens' votes into legislative seats. In this article 'proportional representation' (PR) will indicate electoral systems, while the standard at which they aim will be referred to as 'proportionality.' (For related articles, see *Electoral Systems*; *Parliamentary Government*; *Political Representation*; *Voting: Tactical*; *Voting: Turnout*.)

1. PR Electoral Systems

There are two basic types of PR electoral systems—list PR and the single-transferable vote (STV). All PR systems require multimember districts—constituencies from which voters elect more than one representative. Thus, except in the trivial sense that the winner should have more votes than any loser, proportional representation does not apply to elections that choose a single winner, such as for executive posts or legislatures based on single-member districts (SMDs). The degree to which a PR electoral scheme can approximate perfect proportionality depends on the number of representatives elected per constituency, or district magnitude (M). In the limiting case, the entire nation constitutes a single constituency, as in Israel and The Netherlands, where $M = 120$ and 150, respectively.

1.1 List PR

In the pure form of list PR, the closed list, political parties designate before the election a rank-ordered list of M (or fewer) candidates. Electors vote for a party, not for individual candidates. Parties winning seats select representatives from their lists according to the predetermined ranking.

The number of seats each party wins is determined by a formula that employs one of two basic mechanisms—quotas or divisors. In quota (or largest remainder) formulas, parties receive an initial allocation of seats equal to the number of quotas achieved in the vote for their list. The quota is determined by dividing total votes by M (Hare or natural quota), $M+1$ (Droop quota), or $M+2$ (Imperiale quota). Any remaining seats are assigned according to largest remainders—i.e., by the size of parties' leftover votes after subtracting the quotas they have used. In divisor (or highest average) methods, seats are assigned sequentially. Each seat goes to the party which, at that point in the allocation, has the largest quotient after its votes are divided by a number that varies with the seats the party has already received (S). Well-known divisors include $S+1$ (1, 2, 3, ...), the *d'Hondt* formula, and $2S+1$ (1, 3, 5, ...), the *Sainte-Laguë* formula.

Formulas inspire political controversy, because different quotas and divisors favor large, medium, or small parties. Nevertheless, district magnitude has a more powerful effect on parties' fortunes; small parties generally do better as M rises. Some list PR systems incorporate another element to exclude small parties—a threshold of votes (e.g., five percent) that a party must achieve to win seats.

List PR systems are employed for national parliamentary elections in most countries of Latin America and continental Europe.

1.2 The Single-transferable Vote

STV, also known as the quota-preferential system, depends on a preferential ballot, in which voters rank individual candidates from most to least preferred. A quota of votes, Q , is established, usually by dividing the total number of voters by $M+1$. All candidates receiving Q or more first preferences are elected. To determine the remaining winners, two types of transfers are carried out. In downward transfers, surplus votes (those above Q) received by winning candidates are reallocated to voters' second (or, if necessary, subsequent) choices. In upward transfers, the candidate ranking last in votes is eliminated, and ballots giving first preferences to him or her are reassigned to the voters' second (or subsequent) choices. Transfers are iterated until M candidates have achieved the quota.

Perhaps because of its complexity, both in balloting and selecting winners, STV is employed in national elections by just a few countries—Ireland, Malta, and (for the Senate only) Australia.

1.3 Parties and Candidates in PR Voting Systems

In its pure form, list PR assigns control over the fate of individual candidates to parties' internal decision processes, whereas STV reserves that power for voters. In addition, list PR forces voters to choose just one party (or sometimes an electoral alliance of parties), whereas STV allows voters to move back and forth across parties while ranking candidates. Thus, list PR strengthens party organization and promotes party discipline in the legislature, whereas STV is conducive to intraparty electoral competition and individualistic behavior among legislators.

In practice, however, these differences may be blurred. Many list PR countries use open lists, in which electors are permitted (or, as in Finland, required) to indicate preferences for candidates as well as a party. In the *panachage* system of Switzerland and Luxembourg, electors may even distribute votes for candidates across different party lists. Conversely, in STV systems, unless district magnitude is kept low (on the order of three to five representatives per constituency), the task of ranking numerous candidates becomes so difficult that most voters will depend on instructions from parties. Thus, if it imposes excessive cognitive demands, a system intended to maximize the autonomous power of individual voters may end up functioning rather like list PR.

2. Mixed-member Proportional

In an effort to combine the virtues of proportionality and local representation, Germany after World War II pioneered the mixed-member proportional (MMP)

electoral system. In MMP, a share of representatives (half in Germany) are elected from single-member districts, and the remainder are elected from party lists. Electors cast two votes—one for an individual constituency representative and one for a party. Seats that parties win from constituencies are subtracted from their list allocations to make overall delegations proportional to party votes. This crucial compensatory feature makes MMP a truly proportional system.

3. *Semi-proportional Systems*

Another mixed type, easily confused with mixed-member proportional, is the parallel system. Like MMP, parallel systems elect some legislators from single-member constituencies and some from PR lists. Parallel systems diverge from MMP by not compensating for constituency victories in allocating seats to party lists. In effect, the legislature is chosen by an amalgamation of two separate electoral systems, only one of which aims at proportionality. If, as is usually the case, seats chosen by PR comprise no more than half the assembly, then the parallel system is appropriately classified as semiproportional. Japan and Russia adopted parallel systems in the 1990s.

The term semiproportional is also applied to electoral systems that are intended to achieve a degree of proportionality, but do not employ PR formulas to do so. These systems use multimember districts (of low magnitude), voting for individual candidates, and the plurality formula. The opportunity for proportionality arises from their balloting rules. The limited vote requires electors to cast fewer than M votes. In the extreme version case, the single-nontransferable vote (SNTV), electors may vote for only one candidate, even though M will be elected. Japan employed SNTV until 1993. Cumulative voting systems give electors M votes, but allow them to cast multiple votes for a favored candidate or candidates. Limited voting and cumulative voting, which have similar underlying strategic properties, enable well-coordinated minorities to win seats. However, because parties must avoid dispersing votes among too many candidates, these systems tend to reduce competition, and such competition as does occur tends to be within parties or voting blocs, causing internal conflict and factionalism.

4. *Effects and Evaluations*

The obvious intended benefit of PR systems is to establish what advocates consider an essential condition of electoral fairness—a close correspondence between parties' votes and seats. The use of a PR formula, however, does not preclude significant deviations from proportionality when thresholds exist or district magnitudes are low, especially if parties have

not yet adapted strategically to the parameters of the electoral system.

Although the immediate effect of list PR is to promote proportionality for political parties rather than underlying population groups, most parties will compete for votes of minority groups by placing members in electable list positions. If they do not, groups large enough to surpass the threshold can win representation by organizing their own parties. Consequently, list PR systems usually offer improved representation to minorities. The same incentives apply to electing women. Where they have become politically active and ambitious, women generally win more legislative seats under list PR than under single-member district (SMD) systems.

Because PR is less likely than SMD to waste votes cast for smaller parties, it usually results in multiparty parliaments (see *Party Systems*). The association between PR and multipartyism must be regarded as permissive rather than automatic. Within the constraints imposed by the electoral framework (district magnitude, threshold, and formula), the number of parties depends on the number of societal cleavages and on the willingness and ability of politicians to exploit them by organizing parties (see *Cleavages: Political*). Proportionality and multipartyism together make one-party majorities rare, so parliaments elected by PR typically form coalition governments or minority governments that must get support from other parties to pass legislation.

Critics of PR usually concede its representational advantages, but they contend that it produces a fragmented party system, unstable coalition governments, and excessive influence by extremist parties. Their favorite negative examples include Italy (before 1993 reforms), Israel, and Weimar Germany. In response, defenders of PR recommend 'moderate PR' with significant thresholds or districts of intermediate magnitude, rather than the pure forms those three cases exemplify. They defend the need for coalitions under PR as fostering consensus politics and stable policies, if not always stable cabinets.

5. *Historical Development*

Because PR methods are non-obvious and complex, their development stands as one of the key intellectual achievements in the history of democracy. The Marquis de Mirabeau enunciated the principle of proportionality in 1789, but methods designed to achieve it were not refined until well into the nineteenth century. In 1834 Victor Considérant devised a list PR system that Antoine Morin of Switzerland developed further in 1862. In 1885 an international conference on proportional representation in Antwerp recommended the list PR method proposed in 1882 by the Belgian Victor D'Hondt. In 1899, Belgium adopted D'Hondt's system and thus became the first country to use list

PR. By 1920 most countries of western Europe had adopted versions of PR (Carstairs 1980).

Beside the intellectual and moral appeal of proportionality, two practical motives spurred the turn to PR in Europe. In societies divided along linguistic or religious lines, such as Switzerland, Belgium, and The Netherlands, PR offered all groups a stake in democracy (see *Consociationalism*). For countries about to expand the franchise, PR assured ruling conservative parties that they would not be shut out of parliament, as might have occurred under existing electoral systems once previously disenfranchised citizens became a majority of voters. These motives also influenced later decisions to adopt PR, including the notable case of South Africa in 1994.

Perhaps because of their continuous history of elections predating the development of PR, the UK and the USA (as well as some countries they influenced) have clung to the older, simpler method of single-member districts and plurality rule (see *First Past the Post, in Politics*). Proponents of PR in English-speaking countries have usually favored STV, reflecting its invention by the Englishman Thomas Hare in 1857 and John Stuart Mill's enthusiastic advocacy. That tradition was broken in 1993 when New Zealand adopted MMP.

Agreement on a list PR system was often a key ingredient in democratizations during the late twentieth century, as in Portugal, Spain, South Africa, and Indonesia. However, in the 1990s the global trend shifted to mixed electoral systems, both the semi-proportional parallel version and fully proportional MMP. By 1997, of the world's 211 independent states or semiautonomous territories, 75 (35 percent) used PR systems (list, MMP, or STV) to elect the principal chamber of the national legislature. Countries using PR comprised 46 percent of states respecting political rights and civil liberties, and 59 percent of well-established democracies (Reynolds and Reilly 1997). In many other nations, PR is used for elections to upper chambers, regional and local councils, or supranational authorities, such as the European Parliament.

See also: Electoral Systems; Party Systems; Political Representation; Voting: Tactical

Bibliography

- Carstairs A M 1980 *A Short History of Electoral Systems in Western Europe*. Allen & Unwin, London
 Cox G W 1997 *Making Votes Count*. Cambridge University Press, Cambridge, UK
 Lijphart A 1994 *Electoral Systems and Party Systems*. Oxford University Press, Oxford, UK
 Reynolds A, Reilly B 1997 *The International IDEA Handbook of Electoral System Design*. International Institute for Democracy and Electoral Assistance, Stockholm

Taagepera R, Shugart M S 1989 *Seats and Votes*. Yale University Press, New Haven, CT

J. H. Nagel

Propositional Representations in Psychology

In cognitive psychology the term 'proposition' refers to a truth-valued unit of meaning that is represented symbolically in the human cognitive system. Propositions have two important components. First, they consist of semantic information structures that represent aspects of experience (e.g., events, states), real and hypothetical situations in the world, and abstract conceptual knowledge. Second, they include truth value and modality information that represents an attitude or belief on the part of a speaker, writer, or any holder of the proposition, with respect to the truth or falsity of the semantic relationships that are represented by the proposition. Propositions function as a semantic base for natural language where they represent types of semantic information that are distinguished in natural languages; and they function as a base for reasoning, serving as truth-valued and quantified predicates to which logical reasoning operations may be applied. This article examines propositional representations from a psychological point of view, the functions of propositions in the human cognitive system, models of propositional representations, and a closely related discourse analysis and model-building technique in cognitive science known as 'propositional analysis.'

1. Propositional Representation in Cognitive Systems

1.1 Propositions as Cognitive Representations

Like all cognitive representations, propositions are a form of semantic representation. As such, they may be used to represent concrete situations of cognition and action in the world, or they may represent more general or abstract conceptual knowledge that is a part of one's long-term store of knowledge about the world and how to act in it. However, in contrast to propositions, systems of cognitive representation are usually thought of as large networks of concepts and relations that reflect a person's comprehension and knowledge. Such networks are often referred to as 'mental models,' 'situation models,' 'conceptual or semantic networks,' 'schemata,' 'knowledge frames,' or in terms of other similar constructs (see Rumelhart and Norman 1988 for a review).

For example, when a visual scene is interpreted, it is represented semantically in the cognitive system as a

network in which objects are identified and represented (as nodes), their properties are represented by links to attributes, and their relationships to each other are represented by particular types of semantic links. To talk about a scene requires the construction of propositions. These consist of truth-valued assertions about information from this network. The validity of these assertions can be evaluated and they can consequently serve as a basis for discussion, inferences, and reasoning. A major task for a theory of cognitive representation is to identify the types of concepts and relations that are used by humans to interpret and represent their physical world as it is interpreted and described through natural language.

An essential property of systems of cognitive representation is their relationship to the particular situation or situations in the world that they represent. Thus, systems of cognitive representation must support evaluations of their validity as representations of (real or hypothetical) situations in the world, or in relation to bodies of shared abstract or concrete knowledge (e.g., in a scientific field). This semiotic property of cognitive representations underlies the truth value and modality information present in propositions, since it is information about asserted truth conditions that permits discussion and reasoning about perceived situations in the world. Consequently, the study of propositional representations in natural language processing, and of reasoning and inference through natural language, has provided psychologists with models of semantic representation that have had a large impact on theories of cognitive representation in general (see Frederiksen 1975, 1986, Kintsch 1974, Rumelhart and Norman 1975, van Dijk and Kintsch 1983).

1.2 The Influence of Other Disciplines on Cognitive Theories of Propositional Representation

Propositions are closely associated with the production and comprehension of natural language, where they represent semantic information that is asserted by a speaker or writer and explicitly encoded in language structures and discourse. Researchers studying language comprehension have shown that in addition to interpreting the propositions encoded in lexical and syntactic structures, understanding discourse also involves construction of inferred propositions. Inferences are used to complete, elaborate, and connect local propositional information into a coherent text representation, and relate a text's meaning to prior knowledge and to its situational context (Denhière and Rossi 1991, Graesser and Bower 1990, Kintsch 1998). The construction of propositional representations, whether through interpretation of natural language or through reasoning about situations in the world, is a fundamental aspect of the cognitive processes that enable us to understand and interpret the world. Thus, propositions are a fundamental aspect of the human

cognitive system where they function in a variety of ways by providing the basic representational units of meaning in cognition and language.

Cognitive theories of propositions as units of semantic representation have been influenced by work from many fields including: linguistics (case grammar, linguistic semantics, text linguistics, pragmatics, systemic linguistics); logic; model theory; situation semantics; philosophy of language (speech act theory, semiotics); and computer science (artificial intelligence and language understanding, conceptual graph and semantic network representations, frame theory, scene analysis, and computational linguistics). Other relevant disciplines include ethnography of communication (conversational inference); cognitive sociology (conversational analysis, ethnomethodology); and cognitive linguistics. For further reading on work in computer science, see Brachman and Levesque 1985; for a multidisciplinary overview, see van Dijk 1997. Fauconnier (1997) discusses recent work in cognitive linguistics and related fields.

1.3 The Role of Propositions in the Architecture of the Cognitive System

The architecture of the cognitive system consists of its general properties as an information-processing system (e.g., Simon 1996 and other papers in the same volume). These include systems for: processing input information; constructing and storing internal symbolic codes; interpreting input information; accessing, using, and modifying stored prior knowledge; operating on internal representations (e.g., to understand, reason, and solve problems); and producing appropriate actions and results (including language) as output. The cognitive architecture that supports natural language processing capabilities consists of a stratified semantic processing system in which propositions are intermediate semantic units that are involved in producing and comprehending natural language in relation to its situational context and to the participants' long-term store of conceptual and relational knowledge. Figure 1 identifies the major elements of this stratified processing system and the central role that propositions play in comprehending and producing natural language discourse. The major levels of representation and processing are:

Natural language discourse: phonetic or lexical patterns; word/morpheme sequences; syntactic structures; linguistic forms signaling discourse cohesion; linguistic and extralinguistic markers (e.g., gestures, stress, intonation) that relate discourse constituents to events, states, and representations in the spatio-temporal environment and signal the situational cohesion of the discourse; and (if the discourse is interactive) speech acts and conversational structures.

Propositional representations: representations of concepts and relational semantic information explicitly encoded in natural language expressions (the text

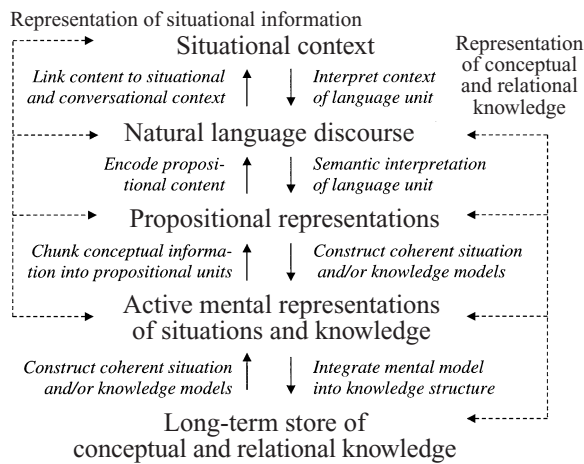


Figure 1

Levels of processing in the production and comprehension of situated discourse

base); inferred propositions that elaborate and connect local propositions in the text and resolve anaphoric references; coherence inferences; macropropositions that summarize content; propositions generated by logical rules to represent the logical structure of arguments; and propositions that resolve extralinguistic references to the situational context.

Active mental representations of situations and knowledge: conceptual networks constructed to link propositional information into a contextually appropriate connected and coherent cognitive representation—a currently active mental model, situation model, or knowledge structure that is constructed based on the text propositions, situational information, and prior knowledge.

In a stratified semantic processing system, these multiple representations are constructed in a situational context that includes a physical environment (of things, states, actions, events, processes, sequences, etc.) and a representational environment (of symbols, language, icons, images, texts, etc.). They also reflect an individual's use of prior conceptual and relational knowledge that is potentially relevant to the situation and content of the discourse.

2. Cognitive and Communicative Functions of Propositions

2.1 The Semantic Content of Linguistic Expressions

Propositions represent the semantic content of linguistic expressions. For example, in 'The airplane was rising swiftly,' the proposition would represent the structure of an event involving an action (rising), an object undergoing change as a result of the action (the airplane), and an attribute of the action (swiftly). It

would also specify how the airplane is determined (definite) and quantified (singular), and the truth assertion about the proposition (positive truth value). To see how this can become more complex, in 'Laura felt the airplane rising swiftly,' the previous proposition would have to be embedded so as to reflect its semantic role as the thematic content of a process (feel) of a patient or experiencer (Laura), and this proposition would have its truth value specified and objects (Laura) quantified. As can be seen in Fig. 1, producing utterances in a language involves the specification of propositions that 'chunk' conceptual and semantic information taken from situation or knowledge models into propositional units so that the information can be encoded morphologically and syntactically in a particular language.

2.2 Units of Meaning for Comprehension, Inference, Reasoning, Thinking, and Learning Processes

In comprehension, the end product is a mental representation of a situation and knowledge. The direction of processing can be data-driven (when the flow of processing is from semantic interpretation of language units, to propositional inferences, to the construction of a situation model), or it can be knowledge-driven (when prior conceptual knowledge or knowledge frames are applied to guide and constrain the construction of a situation model that incorporates propositional information from a text base). Processes of inference and logical reasoning can operate on propositional predicates, or they may be based on mental models that are constructed to facilitate a more inductive style of reasoning based on concrete examples (Johnson-Laird 1996). Thinking processes involve the construction, activation, and modification of situation models and propositions, and learning involves the incorporation of propositions and information from active situation and knowledge models into a more permanent knowledge structure.

2.3 Building Blocks for Systems of Cognitive Representation

Propositions provide the building blocks for systems of cognitive representation. If knowledge is to be communicated through language or treated as an object of logical reasoning, then it must at some point be propositionalized. Consequently, there should be a close relationship between propositions and systems of cognitive representation. Thus, cognitive psychologists have adopted the strategy of building models of knowledge structure to incorporate semantic relationships that are present in the propositional semantic structures that are expressed in natural language (Frederiksen and Breuleux 1990, Rumelhart and Norman 1988).

2.4 The Semantic Representation of Situations and Knowledge

The propositional information that we construct to describe situations through language is similar to the semantic structures we use to interpret and construct cognitive representations of situations and knowledge. Thus, a primary function of propositions and situation models is to represent and continually update our representations of knowledge and of the currently experienced situational contexts in which we are participating.

2.5 Communication, Conversational Inference, and Situated Cognition

In social environments of collaborative action, reasoning, and problem solving, the main channel of communication is conversational dialog. Participation in such task-oriented dialogs demands conversational inferences by the participants, and these are based on comprehension of the propositional content of the speech as it occurs within the situation and with reference to relevant domain knowledge. Communication and collaborative action in such environments depends on comprehension of the propositions asserted by the participants in the conversation. Thus, propositional representations play a central role in such situations of conversational interaction and socially situated cognition (see Koschmann 1999 for an example).

3. Models of Propositional Representations

3.1 Types of Propositions

Propositions represent a variety of types of semantic information that may be expressed in language. Types of propositions have been defined in terms of the particular semantic relations and concepts that are involved in a particular type of situation that is being represented. A listing of some of the types of propositions that have been identified can help convey the richness of propositional representations.

Events: Events are structures involving an action that produces a change, its causes, results, goals, instruments, and other information specifying the participants and objects effected, the internal structure of an event, and properties of the action, e.g., 'The man walked from the top of the hill to the cemetery gate.'

Processes: Processes identify a process and the objects or persons who experience or are involved in the process; they include cognitive processes having thematic content (e.g., know, believe) and represent the structure of a complex process and its properties,

e.g., 'The war raged.' 'John knows Sally.' 'Laura felt the airplane rising gradually.'

States: States identify objects and specify properties of the objects (e.g., attributes, parts, set relations) as well as their determination and quantification, e.g., 'The painting is beautiful.' 'All of the visitors were tourists.' 'The painting is on the wall.'

Abstract relations: Abstract relations identify abstract concepts (corresponding to propositions) and their properties, e.g., 'This theory is complex.' 'The diagnosis was not obvious.'

Identities: Identity relations link objects or propositions into identity sets, e.g., 'The American President is George W. Bush.'

Algebraic relations: Order and equivalence relations are involved in comparative statements and apply to variables that represent values identified in other propositions, e.g., 'John is taller than Mary.' 'His temperature is normal.'

Functions: Functions represent operations defined on variables that return values, e.g., 'Ottawa is close to Montreal.' 'Montreal is closer to Toronto than to New York.' 'The mean height of the team members is six feet.'

Dependency relations: Dependency relations make one proposition depend on another, such as by causal, conditional, and logical implication relations, e.g., 'He won because he had trained intensively.' 'If you finish your supper, then you can have dessert.'

Conjoint relations: Conjoint relations, such as 'and,' alternating 'or' and exclusive 'or,' specify intersecting, alternating, or mutually exclusive sets of objects or propositions.

Tense and aspect components: These specify temporal properties of propositions, e.g., 'He was attending the opera.'

Modality and truth value components: These specify the asserted truth conditions of propositions, e.g., 'He may not have been at the Opera.'

Determination and quantification components: These specify the determination and quantification of concepts, e.g., objects, and of abstract concepts, e.g., 'Six of the men volunteered,' 'All people are mortal.'

3.2 Predicate-argument vs. Conceptual Graph Structures

Two sorts of formalisms have been used to represent propositions: predicate-argument notation, and conceptual graphs (Sowa 1984). Predicate-argument formalisms use a list notation in which a proposition is represented as a predicate (e.g., event) followed by a list of its arguments (the components of the event: the action, the object of the action, the result, etc.). Information about the determination and quantification of objects may be embedded in its argument slot. For example, consider the following sentence taken

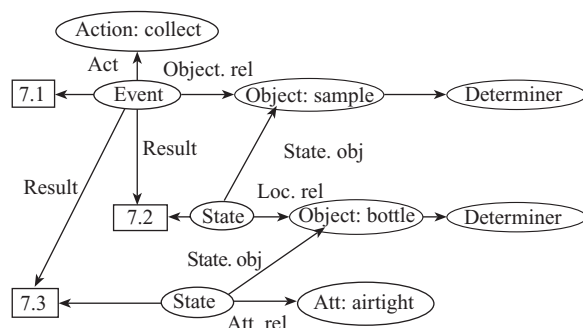


Figure 2

Conceptual graph of 'Collect the samples in an airtight bottle' (adapted from Frederiksen and Breuleux 1990)

from a procedural text in chemistry: 'Collect the samples in an airtight bottle.' An event proposition for this sentence might be as follows (in predicate notation):

7.1 Event: Act (*collect*), Object (*samples*: Def, Plural), Result (7.2, 7.3), Truth.Value (Positive);

The result consists of pointers to two embedded state propositions representing the location of the samples after the action has taken place and a property of the bottle:

7.2 State: Object (*samples*: Def, Plural), Locative. Relation (Object (*bottle*: Token, Singular)), Truth.Value (Positive);

7.3 State: Object (*bottle*: Token, Singular), Attribute.Relation (*airtight*), Truth.Value (Positive)

The alternative conceptual graph notation for this example is given in Fig. 2. Conceptual graphs are equivalent but have the advantage of helping to visualize the semantic relationships being represented by the propositions. This notation is particularly useful in developing models of knowledge structures in particular domains being studied.

3.3 Propositions as Canonical Frames or Semantic Grammars

If propositions are constructed in interpreting and producing language, and in making inferences, how can we explain how they are generated? Two types of explanation have been considered: the canonical frame explanation, and the semantic grammar explanation. It is possible to think of propositions as a vast collection of frame structures which contain 'slots' (or variables) that can be 'bound to' particular conceptual information.

If we replace the arguments in the event proposition by variables, we have an event frame for the action collect which can be applied to many different situations by 'binding' the variables to particular 'values':

'Collect the butterflies in the jar' (Y = butterflies, Z = a state representing the butterflies' location in the jar).

Event: Act (*collect*), Object (Y: Determiner, Quantifier), Result (Z), Truth.Value (Positive);

A conceptual frame theory requires a very large store of such frames in memory and processes to apply them to new linguistic or situational data.

Others have argued that representing propositions as a set of frames is insufficiently powerful to represent the full range of propositional representations that can be constructed. It is argued that like natural language structures, propositional structures reflect rules of formation, and propositional representations need to be defined in terms of knowledge representation languages or semantic grammars. Semantic grammars provide a way of precisely defining a language of propositional representation, and provide rules to guide the generation of propositions and semantic parsing of linguistic expressions. Both canonical frames and semantic grammars appear to be plausible as cognitive mechanisms for the construction of propositional representations (see Brachman and Levesque 1985, Frederiksen 1986, Rumelhart and Norman 1988 for discussions of both perspectives).

4. Propositional Analysis

Theories and models of propositional representation have provided a principled basis for the development of precise methods for analyzing the semantic content of texts as representations of knowledge, developing models of expert domain knowledge, and analyzing the cognitive representations and processes underlying cognitive processes in discourse comprehension and production, inference, reasoning, and other cognitive processes that are reflected in verbal protocols. Propositional analysis is essentially a method for analyzing the propositional content of texts, verbal protocols such as text recalls, responses to questions, reasoning, or talk-aloud protocols, and even task-oriented dialog in collaborative situations of learning or performance (see Frederiksen 1986 and Kintsch 1998).

Propositional analysis normally begins with transcription and segmentation of textual material. Linguistic analysis of the text segments may be undertaken in terms of the linguistic structures previously cited. Semantic parsing rules or propositional frames are applied to specify the propositions explicitly encoded in the sentences. The generation of propositions can be guided by computer tools that enable the application of semantic parsing rules or apply frames to analyze the propositional content of text segments. Inferred propositions may be added to the model to fill recognized gaps in the text information or argument structure, relate the text to aspects of the context or task, summarize the macrostructure of the text, provide responses to questions, etc. Inference analysis can

be based on data from subjects who verbally recall, discuss, or analyze a text. Models of situation and knowledge models may be constructed applying conceptual network formalisms to the propositions. Finally, texts can be analyzed in terms of high-level conceptual frame structures such as underlie procedural texts, problem-solving or reasoning protocols, argument structures, narratives, and story plot structures.

See also: Cognitive Psychology: History; Cognitive Psychology: Overview; Critical Thinking, Cognitive Psychology of; Imagery versus Propositional Reasoning; Informal Reasoning, Psychology of; Logic and Cognition, Psychology of; Scientific Reasoning and Discovery, Cognitive Psychology of

Bibliography

- Brachman R J, Levesque H J 1985 *Readings in Knowledge Representation*. Morgan Kaufmann, Los Altos, CA
- Denhière G, Rossi J-P 1991 *Text and Text Processing*. North-Holland, Amsterdam
- Fauconnier G 1997 *Mappings in Thought and Language*. Cambridge University Press, Cambridge, UK
- Frederiksen C H 1975 Representing logical and semantic structure of knowledge acquired from discourse. *Cognitive Psychology* 7: 371–458
- Frederiksen C H 1986 Cognitive models and discourse analysis. In: Cooper C R, Greenbaum S G (eds.) *Studying Writing: Linguistic Approaches*. Sage, Beverly Hills, CA, pp. 227–67
- Frederiksen C H, Breuleux A 1990 Monitoring cognitive processing in semantically complex domains. In: Frederiksen N, Glaser R, Lesgold A, Shafro M G (eds.) *Diagnostic Monitoring of Skill and Knowledge Acquisition*. Erlbaum, Hillsdale, NJ, pp. 351–91
- Graesser A C, Bower G H (eds.) 1990 *The Psychology of Learning and Motivation: Inferences and Text Comprehension*. Academic Press, New York, Vol. 25
- Johnson-Laird P N 1996 The process of deduction. In: Steier D S, Mitchell T M (eds.) *Mind Matters: A Tribute to Allen Newell*. Erlbaum, Mahwah, NJ, pp. 363–99
- Kintsch W 1974 *The Representation of Meaning in Memory*. Erlbaum, Mahwah, NJ
- Kintsch W 1998 *Comprehension: A Paradigm for Cognition*. Cambridge University Press, Cambridge, UK
- Koschmann T (ed.) 1999 Edge of many circles: making meaning of. *Discourse Processes* 27(2): 103–17
- Rumelhart D E, Norman D A 1975 *Explorations in Cognition*. W H Freeman, San Francisco
- Rumelhart D E, Norman D A 1988 Representation in memory. In: Atkinson R C, Herrnstein R J, Lindzey G, Luce R D (eds.) *Stevens' Handbook of Experimental Psychology*, 2nd edn. Wiley, New York, pp. 511–87
- Simon H A 1996 The patterned matter that is mind. In: Steier D S, Mitchell T M (eds.) *Mind Matters: A Tribute to Allen Newell*. Erlbaum, Mahwah, NJ, pp. 407–31
- Sowa J F 1984 *Conceptual Structures: Information Processing in Mind and Machine*. Addison-Wesley, Reading, MA
- Van Dijk T A (ed.) 1997 *Discourse as Structure and Process*. Sage, London

Van Dijk T A, Kintsch W 1983 *Strategies of Discourse Comprehension*. Academic Press, New York

C. H. Frederiksen

Prosecution

In its broadest sense, 'prosecution' signifies an aggrieved party's initiation and pursuit of a legal case through the court system. It begins with the filing of a complaint document and continues through trial, awarding of remedies, and appeal. While this term technically applies to both civil and criminal matters, it is most commonly used in the criminal law context. In a criminal case, the term 'prosecution' may refer to the process of handling a dispute in criminal court or to the aggrieved party who files the complaint in criminal court.

This article addresses both definitions of criminal prosecution using comparative data to illuminate differences and similarities among prosecution systems throughout the world. Prosecutors have been called the 'gate-keepers of criminal justice' because without their actions, judicial intervention and sanctions could not occur (di Federico 1998). However, a study of prosecution should not be limited to assessments of the prosecutor's role in individual cases. One must instead consider the ways in which countries deal with competing needs and limited resources when they establish and operate their prosecution systems. Because prosecution contributes substantially to the definition and implementation of criminal policy, any comprehensive examination of prosecutors' functions must include an analysis of their relationship to the state.

This article considers six of the most salient features of prosecution systems: (a) the degree to which prosecution is handled by public officials or private parties; (b) the degree to which the prosecutor's office is centralized or decentralized; (c) the political accountability of the prosecutor; (d) the prosecutor's relationship to and with the police; (e) the prosecutor's filing standards; and (f) the prosecutor's role in court. The content of this essay will be limited to the role of prosecution in democratic countries; for a discussion of criminal courts in nondemocratic nations, see *Trials: Cultural; Nazi Law; Socialist Law; Political Trials; War Crimes Tribunals*.

1. Public or Private Prosecution

In many legal systems, criminal law was once a branch of what scholars now call 'torts,' (see *Torts in Law*) and prosecution was initiated and pursued by the injured parties. With the rise of the modern state, crime became a distinct offense against the state itself, and responsibility for prosecution was assumed by the

state (Langbein 1974). This trend emerged earlier and more sharply in Europe than in England, where crimes were defined as offenses against the state, the system of privately initiated and financed prosecution remained in place well into the nineteenth century.

1.1 Prosecutors as Public Officials

On the European continent, public prosecutors play a much larger role in the criminal process than they do in the United States and England. Continental prosecutors control the investigation of crime, often having direct supervisory powers over the police, and typically oversee the execution and disposition of almost all criminal arrests. This tradition derives largely from the institution of the *ministère publique* from the Napoleonic Code. Furthermore, some civil law countries combine the role of prosecutor and judge, allowing officials to move back and forth between the two positions at different points in their careers and sometimes within the same case. A number of non-western countries that have based their criminal justice systems on German or French models (such as Japan) also permit role flexibility to varying degrees (see also *Trials: Cultural; Judges* for the relationship between prosecutors and judges). Recent developments in France and Italy, however, suggest the use of more clearly defined boundaries between criminal justice system actors.

Common law countries reject this expansive conception of the prosecutor's office and formally prohibit any degree of fluidity in roles, regarding such interchangeability as in conflict with due process and separation of powers principles (see also *Due Process in Law; Civil Rights*). In common law countries such as the United States, the public prosecutor unilaterally (with very few exceptions) makes filing and case-processing decisions. The investigative work is typically left to the police and judging is the domain of elected or appointed members of the judiciary. (Of course, in actual practice, the discretion with which each of these actors is vested allows some flexibility in task performance; see also *Plea Bargaining* and *Policing* for examples.) In England and Wales, the newly developed Crown Prosecution Service, comprised exclusively of professional prosecutors under the leadership of the Director of Public Prosecutions and the Attorney General, handles the prosecution of most crimes. Like their American counterparts, British judges and police officers function independently of prosecutors, although they do coordinate their activities.

1.2 Elements of Private Prosecution

Since the mid-nineteenth century, public prosecutors everywhere have become more powerful and inde-

pendent. However, since the 1970s, with the growth of the victims' rights movement in many countries, legislatures have begun to provide victims with a greater voice in criminal proceedings. The degree of victim empowerment varies tremendously among countries. For example, Belgian law recognizes that certain crimes (i.e., defamation) require a formal complaint by the victim in order for the public prosecutor to initiate a criminal case. Even in that event, however, the public prosecutor is not obligated to file charges. In Germany, a victim may act as a private prosecutor if the public prosecutor does not file charges, or he may join the public prosecution as a joint plaintiff and support the prosecutor's case. A French crime victim who feels ignored by the authorities may file a civil case based on the criminal behavior. This action automatically initiates a criminal case, thereby forcing an otherwise reluctant public prosecutor to become involved.

American crime victims generally have no formal right to participate directly in criminal cases; their role is limited to testifying as witnesses and to receiving restitution as part of the final case disposition. Nonetheless, victim enthusiasm or willingness to participate (or lack thereof) may greatly influence the prosecutor's decision to proceed with a case; cooperative victims are not a prerequisite to a successful prosecution (given the expansive laws of evidence), but they often bolster an otherwise shaky case. Victims of crime in the United States can also file a civil lawsuit against the defendant based on the same conduct at issue in the criminal case, although the remedies available remain limited to civil damages. Interestingly, the role of private interests in prosecution has made a comeback in the United States, as scholars have recently debated the legitimacy of allowing private victims to finance investigations or prosecutions which are conducted by public officials (Kennedy 1997).

2. Centralized or Decentralized Office

Centralization or decentralization of oversight affects all aspects of the prosecutorial function, from the hiring and training of staff, to the implementation of policies for filing and processing cases, to levels of responsiveness to public opinion. Scholars can distinguish between those highly centralized systems that place prosecutors in an national civil service hierarchy led by the Ministry of Justice, and those decentralized systems that consider prosecutors to be locally selected (and often elected) political officials. Overall these structural and institutional features are likely to be closely related to the larger structure of government and to make a difference (Damaska 1981). Needless to say, variations exist between these extremes, and even within each category, according to hiring practices or filing control.

At one extreme is the American experience. The vast majority of criminal cases—both serious and minor—in the United States are filed and processed at the state level. However, ‘State Control’ is something of a misnomer, because most American state prosecutors are county officials who enforce state law. The chief prosecutor of each office (usually called the district attorney) is elected by the people, while the deputies or assistants may be either career civil servants or employees of a specific administration. State attorneys general have little control over their behavior or filing practices, handling mostly training and appeals (Flemming et al. 1992). Each county operates its own hiring system and typically recruits graduates from local law schools. There is little advance training for the job; on-the-job training occurs locally, within the office or in short courses run by the state. As a result, each office sets its priorities according to the demands of its electorate and according to the interests of its leaders; this can produce great diversity in policies among prosecutors’ offices, even within the same state (Utz 1979, Feeley 1979).

At the other end of the spectrum lie countries like France and Japan, where the procuracy is tightly controlled and pyramidal in shape. For example, at the apex of the Japanese system sits the Supreme Public Prosecutors Office, followed by eight High Public Prosecutors Offices and then fifty District Public Prosecutors Offices. At the base are 452 Local Public Prosecutors Offices, located in major cities throughout Japan. The Local Offices prosecute only relatively minor cases, while the District Offices handle all serious crimes. Moreover, Japanese prosecutors are an elite subset of the population; they have graduated from top law schools, passed an extremely rigorous bar exam, and endured two years of training at the Supreme Court’s Legal Training Institute. Because they are overseen by the Supreme Court, Japanese prosecutors at all levels have close association with judges; indeed, prosecutors can transfer to judicial work if they like and vice versa. As a result of their training and the hierarchy within which they work, and perhaps their close association with the judiciary, Japanese prosecutors value consistency in case-handling to a degree which other nations’ prosecutors find almost inconceivable (Johnson 1996, 1998).

Centralization also affects the extent to which prosecutors can be influenced by public opinion. Officials in decentralized offices are often vulnerable to intense local political pressures and media criticism. In contrast, where public prosecutors’ offices form part of a civil service hierarchy headed by the Ministry of Justice, prosecutors may consider themselves more immune from public opinion. They contend that this immunity is crucial in order for the prosecutor to remain objective and impartial; it allows him to more easily dismiss cases for insufficient evidence, or to prosecute popular or famous citizens, without fear of

retaliation on election day. These claims to immunity should not be taken at face value, however; some modern prosecutors in Italy, Spain, France, and Portugal will actively court the media to generate public interest and support for their cases. Prosecutors in civil service countries are also much less likely than American prosecutors to leave government work for private practice; hence, they are less concerned about keeping good relations with the local defense bar. However, the existence of tight governmental controls provides a different sort of pressure: at least one scholar has noted that Japanese prosecutors have been severely criticized for failing to prosecute corrupt politicians (Johnson 1996).

3. *Accountability to the Public*

As the previous comments regarding public opinion and corrupt politicians reveal, prosecutors in most countries must answer to someone for their unpopular decisions. As discussed above, some countries have responded to this issue by providing victims with the right to initiate private prosecutions after the public prosecutor declines to file charges. While there is not much data on the degree to which these provisions are used by private citizens, the existence of these policies serves (at least in theory) as a check on the prosecutor’s monopoly over charging decisions.

Where the prosecutor is an elected official, as in the United States, accountability is likely to come from another source: re-election pressures. A prosecutor who wishes to be re-elected must present herself as an ethical, fair, and supremely competent attorney and administrator who is responsive to the demands of the local community (Flemming et al. 1992). She is thus not likely to file charges for technical violations about which the local community cares little (such as prostitution in some places), and she may press for rigorous prosecution and sentencing for those crimes which inspire publicity, political lobbying, or protest (such as drunk driving or domestic violence in the latter part of the twentieth century).

Where the prosecutor is a career civil servant, accountability is usually to his superiors or to the Ministry of Justice. Decisions not to file charges will be reviewed by prosecutorial superiors, who tend to be conservative in matters of discretion not to prosecute (Damaska 1981). Japanese prosecutors, for example, must seek the approval of two or three superiors before making any charge decision (Johnson 1996, 1998). Yet some civil servant prosecutors have found another source of leverage; as noted above, prosecutors in Latin European countries often attempt to use media coverage to pressure office superiors into filing or pursuing controversial criminal cases. Hierarchical control is not unproblematic, though, particularly when it affects a prosecutor’s career advance-

ment. France, for example, enacted reforms in 1993 and 1994 to limit the discretionary powers of the Minister of Justice, with regard to both his hierarchical supervision of prosecutorial activities and his decisions concerning the status, promotion, and discipline of public prosecutors (di Federico 1998).

Assessments of accountability must acknowledge context: prosecutors are, *inter alia*, employees of large organizations. As such, they must advance the interests of the organization in order to advance their own careers. Hence, prosecution policy and behavior are likely to follow bureaucratic dictates or public demands, despite (or within) the constraints provided by formal law.

4. Relationship to and with Police

Is a prosecutor a justice official or a member of law enforcement? While Langbein (1974) reports that English justices of the peace served as both investigators and prosecutors in the fourteenth and fifteenth centuries, most countries today have formally separated the roles of police and prosecutor. Nonetheless, prosecutors vary in the degree to which they control or participate in the investigative phase of a case. A prosecutor may bring a high level of control to police activities by formally supervising a police investigation (before or after bringing charges), by interviewing witnesses, or by attending the execution of search warrants. In contrast, a prosecutor may manifest low levels of control by simply filing charges in response to documents and evidence collected and submitted by investigating police officers.

In code-based countries with strongly centralized prosecution agencies, high control is the norm. Police may be statutorily required to notify prosecutors of an ongoing investigation or the detention of suspects, and the prosecutor can directly interrogate suspects or collect evidence. While prosecutors often delegate these decisions to police officers, the existence of prosecutorial investigative authority may serve as a useful check on police power (Johnson 1996, Frase 1990, 1999).

Low levels of control characterize common law countries like the United States; police typically handle the investigation and arrest phases of a criminal matter and notify the prosecutor only after arrest. While federal prosecutors often become far more involved in the investigatory phases, coordinating wire-taps, search warrants, and witness interviews, most state prosecutors do not interrogate suspects or assist in the collection of evidence (Leo 1996). Hence, police often control the flow of information on which prosecutors must rely in order to decide whether and what crimes to charge. This has led some commentators to conclude that prosecutors do little more than ratify police decisions (Feeley and Lazerson 1983), rendering the arrest itself far more significant than the prosecutor's

filing decision. The same tendency has been observed in Canada (Grosman 1969).

The division of labor between police and prosecutors is not always clear-cut or unproblematic, as police behavior can influence both the onset and eventual disposition of a criminal case (Skolnick 1975). These tensions are manifest in the British experience. The police were Britain's public prosecutors from 1829 to 1986, due to the absence of a suitable alternative; in this time period, they developed a vested interest in controlling all stages of criminal cases—from investigation to arrest to prosecution (Fionda 1995). The Crown Prosecution Service was created to separate police and prosecutorial functions in order to remedy a perceived lack of police objectivity in prosecutorial duties, such as the filing and disposing of cases. However, in the fifteen years since the creation of the CPS, police officers have continued to fight for control over filing decisions and have been accused of manipulating files and witnesses to influence prosecutors' options (Nduka-Eze 1995). (See *Policing* for more about police discretion.)

5. Filing Standards and Guidelines

With respect to filing policies, scholars are principally concerned with whether a nation imposes a program of mandatory or compulsory prosecution on the one hand, or officially allows its prosecutors discretion in filing charges on the other. Nonetheless, whether a country actually requires its prosecutors to file charges is often hard to ascertain; the official codes may dictate one practice while informal policies allow others. For example, the American criminal prosecution system gives vast discretion to prosecutors; many continental systems, in contrast, maintain a formal policy of 'full prosecution' and explicitly make no provision for individual discretion. However, most American systems have placed some restrictions on their prosecutors' discretionary powers, while most continental systems permit discretion in hidden (or not-so-hidden) forms. This paradox has led observers to conclude that 'the rule of compulsory criminal prosecution ... may foster a wider use of discretion without responsibility if no control is at work' (Guarneri and Pederzolo 2001).

Mandatory prosecution, which requires a prosecutor to bring a charge whenever there are sufficient grounds to suspect a person of having committed an offense, is a more common feature of civil law countries, whose provisions and requirements are strictly controlled by the criminal codes. Civil law countries typically choose to address undesired consequences by amending their codes, rather than openly permitting the prosecution to use discretion in enforcement. For example, Germany, Sweden, and other countries have handled their increased crime problem by legislatively downgrading or decriminalizing certain offenses, and by making

others subject to administrative (rather than criminal) sanctions (Sessar 1979, Feeley 1979). However, Germany also recognizes the 'principle of opportunity' for petty offenses, which gives the prosecutor discretion to drop the charge, to issue a penal order (which allows a person to pay a fine in a minor case), or to bring the case for summary trial in order to expedite proceedings (Herrmann 1992, Felstiner 1979).

In France, the opportunity principle works somewhat differently. While the public prosecutor is not obligated to bring a prosecution, once she initiates a prosecution she no longer has a choice regarding whether to pursue the case (Frase 1999). In Italy, compulsory prosecution exists without the concurrent opportunity principle, but discretion reigns nonetheless (Gaurneri 1996, Corso 1993). When a prosecutor does not wish to file a case, she may neglect to investigate and ask the judge to declare the evidence insufficient, or she may allow the statute of limitation to lapse.

The Italian prosecutor's manipulation of formal criteria to justify a non-filing decision raises the issue of discretion: in countries which officially allow some discretion, are prosecutors subject to any formal rules regarding its exercise? For example, Israeli and British prosecutors must prosecute when the evidence is sufficient, unless they find that prosecution would not be in the public interest (Harnon and Stein 1999, Fionda 1995). Notably, British prosecutors may be using this public interest criterion as an informal presumption against prosecution in relatively minor cases, especially where the defendant has no previous criminal record (Fionda 1995).

In the United States, where there is no official 'public interest' component to the filing decision, prosecutors have an enormous amount of discretion; they need to justify filing decisions only to their superiors, as victims have no legal rights to intervene. As a result, prosecutors allow informal social criteria, including victim characteristics, to influence their charging decisions. Cases involving victims who are related to defendants, or who are otherwise 'undesirable' people (such as prostitutes or gang members) trigger larger societal issues, and thus appear less distinctively 'criminal' to American juries. Because these cases pose conviction problems, prosecutors concerned about their personal 'win rates' are reluctant to file them (Frohmann 1997, Vera Institute of Justice 1981). American prosecutors have also been accused of overcharging cases, filing multiple charges in the complaint with an eye towards plea bargaining later. Overcharging gives the prosecutor leverage in the plea bargaining process, but it has been called a coercive, corrupting, and deceptive technique to get the defendant to agree to virtually the same sentence he would receive after trial (Alschuler 1968) (see *Plea Bargaining*).

Discussion of filing policy necessarily leads to consideration of alternatives to prosecution, because

many cases which are not technically prosecuted are also not fully dismissed. Rather, they are redirected into other channels, alternatively called diversion, penal orders, or summary procedures. These programs arose in response to demands to keep minor offenders and unimportant cases out of the already overworked court systems. However, they have all been accused of 'widening the net' of criminal social control, as prosecutors are more inclined to divert a minor offender into this type of alternative program than to drop his case altogether.

6. Roles in Court

The final point of comparison concerns the prosecutor's role during the trial of a criminal case. Her role will largely be determined by the type of trial system in which she works (adversarial, inquisitorial, or some mixture thereof; see *Trials: Cultural*). She may be an active adversarial participant striving to win a conviction, or a more passive, impartial actor charged with ensuring that justice results.

Where the trial depends largely on the contents of a written dossier (in countries such as Japan, Argentina, France, and Germany), the prosecutor is generally more passive; she submits the dossier containing the investigative reports to the judge (or to the panel of judges) before trial, but she generally relinquishes control over questioning of witnesses. In many of these systems, the prosecutor is obligated to ask the judge for an acquittal following trial if the evidence merits this result, and the prosecutor typically has the right to appeal both the outcome of the case and the sentence.

In systems that require oral testimony (including Italy, Great Britain, Israel, and the United States), the prosecutor is the principal player in the courtroom. He may sit near the factfinders or judges to establish rapport; he controls the majority of the witnesses and the presentation of evidence; he argues his case both first and last, providing both the initial opening statement and the final closing argument. For this reason, prosecutors in oral-trial systems feel personally invested in the outcome of a case and often place tremendous pressure on themselves to win convictions (Heilbroner 1990, Baker 1999). They generally have no right to appeal acquittals.

A final consideration is the prosecutors' input into the sentencing decision. Prosecutors are often said to have tremendous influence over sentencing based on their charging and plea bargaining policies. Because each crime is associated with a specific minimum and maximum sentence, the prosecutor can control the type of sentence a defendant will receive by filing or making him plead guilty to certain charges. This power is exacerbated by mandatory sentencing guidelines; by eradicating the judge's discretion to sentence appropriately after conviction, these guidelines give

the prosecutor the ability to control the sentence through his charge-bargaining practices. Furthermore, some nations, such as Great Britain and Germany, limit the ability of certain courts to impose certain punishments; hence, the mode of trial pre-determines the maximum sentencing powers of the court in a given case. The prosecutor may thus indirectly control sentencing by filing the case in a particular type of court.

Prosecutors might also influence sentencing directly, although nations differ concerning the degree to which they allow prosecutors to make sentencing recommendations after trial. In many European countries, prosecutors can assist judges in navigating complex sentencing formulas, while in the United States the sentencing hearing is regarded as another phase of the adversarial process in which the prosecutor may present both witnesses and arguments. In Great Britain, the prosecutor has no legal authority to recommend an appropriate sentence; indeed, some commentators believe such a power would be an unconstitutional infringement on judges' power (Fionda 1995). However, British prosecutors have a duty to correct a judge who is in danger of imposing an illegal sentence, and the British Attorney General, like prosecutors in South Africa and the United States, may appeal any sentence imposed by the trial court on the grounds of undue leniency or statutory error.

7. Conclusion

Given the vast array of criminal prosecution systems existing at the close of the twentieth century, it no longer seems useful to rely on the broad categories of adversarial and inquisitorial systems to illustrate similarities and differences. A number of axes exist that better describe the elements of various prosecution systems; to ignore these dimensions in favor of simple dichotomies provides us with too superficial a view of the prosecutor's role in court, in public life, and in political affairs. Moreover, the prosecutor is not an isolated actor: her responsibilities, relationships, and roles depend on social, legal, and political contexts. In order to truly understand the role of prosecution from a social scientific perspective, we must commit ourselves to understanding these environments and to accepting the cultural differences which exist among the world's nations.

See also: Common Law; Courts and Adjudication; Criminal Justice, Sociology of; Criminal Law and Crime Policy; Death Penalty; Integration: Social; Law: Economics of its Public Enforcement; Plea Bargaining; Police, Sociology of; Policing; Punishment: Social and Legal Aspects; Trials: Cultural

Bibliography

- Alschuler A W 1968 The prosecutor's role in plea bargaining. *University of Chicago Law Review* 36: 50–112
- Baker M 1999 *D.A.: Prosecutors in Their Own Words*. Simon and Schuster, New York
- Corso P 1993 Italy. In: Van Den Wyngaert C (ed.) *Criminal Procedure Systems in the European Community*. Butterworths, London, pp. 223–60
- Damaska M 1981 The reality of prosecutorial discretion: Comments on a German monograph. *American Journal of Comparative Law* 29: 119
- di Federico G 1998 Prosecutorial independence and the democratic requirement of accountability in Italy. *British Journal of Criminology* 38: 371–87
- Feeley M M 1979 *The Process is the Punishment*. Russell Sage, New York
- Feeley M M, Lazerson M H 1983 Police–prosecutor relationships: An interorganizational perspective. In: Boyum K, Mather L (eds.) *Empirical Theories about Courts*. Longman Press, New York, pp. 216–43
- Felstiner W L F 1979 Plea contracts in West Germany. *Law and Society Review* 13: 309–26
- Fionda J 1995 *Public Prosecutors and Discretion: A Comparative Study*. Clarendon Press, Oxford
- Flemming R B, Nardulli P F, Eisenstein J 1992 *The Craft of Justice: Politics and Work in Court Communities*. University of Pennsylvania Press, Philadelphia, PA
- Frase R S 1990 Comparative criminal justice as a guide to American law reform: How do the French do it, how can we find out, and why should we care? *California Law Review* 8: 539–686
- Frase R S 1999 France. In: Bradley C M (ed.) *Criminal Procedure: A Worldwide Study*. Carolina Academic, Durham, NC, pp. 143–84
- Frohmann L 1997 Convictability and discordant locales: Reproducing race, class, and gender ideologies in prosecutorial decisionmaking. *Law and Society Review* 31: 531
- Goldstein A, Marcus M 1977 The myth of judicial supervision in three 'inquisitorial' systems: France, Italy, and Germany. *Yale Law Journal* 87: 240–83
- Goldstein A, Marcus M 1977 Comment on continental criminal procedure. *Yale Law Journal* 87: 1570–7
- Grosman B 1969 *The Prosecutor: An Inquiry into the Exercise of Discretion*. University of Toronto Press, Toronto, Canada
- Guarneri C 1996 Prosecution in two civil law countries: France and Italy. In: Nelken D (ed.) *Comparing Legal Cultures*. Aldershot, Brookfield, VT
- Guarneri C, Pederzolo P 2001 *The Power of the Judge*. Oxford University Press, Oxford, UK
- Harnon E, Stein A 1999 Israel. In: Bradley C M (ed.) *Criminal Procedure: A Worldwide Study*. Carolina Academic, Durham, NC, pp. 217–44
- Heilbroner D 1990 *Rough Justice: Days and Nights of a Young D.A.* Pantheon Books, New York
- Herrmann J 1992 Bargaining Justice—A bargain for German criminal justice. *University of Pittsburgh Law Review* 53: 755–76
- Johnson D 1996 The Japanese way of justice. Doctoral Dissertation. University of California at Berkeley, Berkeley, CA
- Johnson D 1998 The organization of prosecution and the possibility of order. *Law and Society Review* 32: 247–308
- Kennedy J 1997 Private financing of criminal prosecutions and

- the differing protections of liberty and equality in the criminal justice system. *Hastings Constitutional Law Quarterly* 24: 665
- Langbein J H 1974 *Prosecuting Crime in the Renaissance: England, Germany, France*. Harvard University Press, Cambridge, MA
- Langbein J H, Weinreb L L 1977 Continental criminal procedure: 'myth' and reality. *Yale Law Journal* 87: 1549-69
- Leo R 1996 Inside the interrogation room. *Journal of Criminal Law and Criminology* 86(2): 266-303
- Nduka-Eze C 1995 The CPS and independence from the police. *New Law Journal* 1843-4
- Sessar K 1979 Prosecutorial discretion in Germany. In: Macdonald W F (ed.) *The Prosecutor*. Sage, Beverly Hills, CA, pp. 255-76
- Sigler J 1979 The prosecutor: a comparative functional analysis. In: Macdonald W F (ed.) *The Prosecutor*. Sage, Beverly Hills, CA, pp. 53-74
- Skolnick J 1975 *Justice Without Trial: Law Enforcement in Democratic Society* (2nd edn.). Wiley, New York
- Utz P 1979 Two models of prosecutorial professionalism. In: Macdonald W F (ed.) *The Prosecutor*. Sage, Beverly Hills, CA, pp. 99-124
- Vera Institute of Justice 1981 *Felony Arrests: Their Prosecution and Disposition and New York City's Courts*. Longman, New York
- Weigand T 1999 Germany. In: Bradley C M (ed.) *Criminal Procedure: A Worldwide Study*. Carolina Academic Press, Durham, NC, pp. 187-211

K. Levine and M. Feeley

Prosocial Behavior and Empathy: Developmental Processes

Research on helping others in distress took off with the brutal murder of a young woman in Queens, New York in the mid-1960s, which was witnessed by 30 people from their apartment windows. The initial research on why witnesses do not help (Latane and Darley 1970) focused on the context and found that the presence of other bystanders can interfere with a person helping by activating certain assumptions—'diffusion of responsibility' (I'm sure someone has called the police) or 'pluralistic ignorance' (no one is doing anything, so it can't be an emergency). There was also research on when people do help, which showed among other things that the tendency of bystanders to help increases with age.

Recent research has focused more on prosocial motivation and emotion. It highlights empathic emotion, sympathy, and guilt. This article is concerned with developmental processes in empathic emotions and their contribution to helping. Also covered is the role of cognition in arousing, developing, and shaping empathic emotion and enlarging the individual's perspective on prosocial behavior. Finally, the article deals with the contribution of socialization especially by parents and peers, and gender differences.

1. Definition of Empathy

Empathy has been defined in two ways: (a) the cognitive awareness of another person's internal states (thoughts, feelings, perceptions, intentions); (b) the vicarious affective response to another person. This chapter deals with affective empathy, because of its motive properties (see Ickes 1997 for cognitive empathy).

Affective empathy was initially defined as a match between the observer's and the model's feeling. It is now defined as it is here, more in terms of the processes that make an observer's feelings congruent with the model's situation not his own (Hoffman 1978, 2000). There may be a match, but not necessarily as when one feels empathic anger on seeing someone attacked even when the victim feels sad or disappointed rather than angry. Most of the research on affective empathy as a prosocial motive focuses on empathic distress because prosocial moral action usually involves helping someone in discomfort, pain, danger, poverty, and the like.

2. Modes of Empathic Arousal

Five empathy-arousal modes have been identified (Hoffman 1978). Three are primitive, automatic, and preverbal. (a) Mimicry, which has two steps: one spontaneously imitates another's facial, vocal, or postural expressions of feeling, producing changes in one's own facial or postural musculature; these changes trigger afferent feedback to the brain, which results in feelings that resemble the other's feelings. (b) Classical conditioning, in which empathic distress becomes a conditioned response to distress in others through observing others in distress at the same time that one is in actual distress. (c) Direct association of cues from others or their situation with one's own painful past experience. The empathy aroused by these three modes is passive, involuntary, based on the pull of surface cues, and requires little cognitive processing. Nevertheless, these modes are important because (a) they show that a person's distress can indeed result from someone else's painful experience; (b) they make empathy arousal possible in the early preverbal years of childhood; and (c) they give empathy an involuntary dimension throughout life.

Two empathy-arousing modes are higher order cognitive. (a) Verbally mediated association: language communicates the other's distress and connects the other's situation to one's own painful past experience. (b) Perspective-taking: one feels something of the victim's distress by imagining how one would feel in the victim's place and/or trying to imagine directly how the victim feels by using information one has about him or her. These higher order cognitive modes may be drawn out over time and subject to voluntary control, but they may also be triggered immediately on witnessing another's distress. What they contribute to

empathy is scope; they also allow one to empathize with others who are not present.

Multiple modes are important because they enable observers to respond empathically to whatever distress cues are available (Hoffman 2000). Cues from the victim's face, voice, or posture can be picked up through mimicry, situational cues, through conditioning or association. If a victim expresses distress verbally or in writing, or someone else describes the victim's situation, one can empathize through verbally mediated association or perspective-taking. Empathic distress is thus a multidetermined hence reliable human response, which fits the argument that it survived natural selection and has a hereditary component (Hoffman 1981, Zahn-Waxler et al. 1992).

3. Developmental Stages

Mature empathy has a metacognitive dimension: one knows one's feeling of distress results from another's plight and how the other presumably feels. One thus has a sense of oneself and others as separate beings with independent inner states (that are only partly reflected in outward behavior), separate identities, and separate life conditions. Before 6 or 7 years children can empathize but without this metacognitive dimension. This suggests that empathic affect develops along with cognitive self/other development. Five stages of empathy development have been hypothesized (Hoffman 1978, 2000).

(a) Global empathic distress: newborns show a vigorous, agitated distress response to another infant's cry. (b) Egocentric empathic distress: one responds to another's distress as though oneself were in distress. This occurs when children can be empathically aroused by the three preverbal modes but still lack a clear self-/other distinction. (c) Quasiegocentric empathic distress: one knows the other is in actual distress but confuses the other's inner states with one's own and tries to help by doing for the other what would comfort oneself. (d) Veridical empathic distress: one's feeling is closer to the other's because one realizes the other has inner states independent of one's own. (e) Empathy for another's experience beyond the immediate situation (e.g., chronic illness, economic hardship, deprivation): this high empathy level is possible when children realize that others have an identity and a generally sad or happy life condition; and empathy with an entire group (homeless; Oklahoma City bombing victims).

4. Causal Attribution's Shaping of Empathic Affect

Humans spontaneously attribute causality to events (Weiner 1985) and they presumably do this when witnessing someone in distress. If they blame the victim, empathic distress is of course reduced. Other-

wise, depending on the attribution, empathic affect may be shaped into four empathy-based moral affects (Hoffman 2000). (a) Sympathetic distress, when the cause is beyond the victim's control (illness, accident, loss). (b) Empathic anger, when someone else is the cause. (c) Empathic feeling of injustice, when a discrepancy exists between the victim's character and the victim's fate (a good person fares badly). (d) Empathy-based guilt over inaction, when one does not help or one's efforts to help fail and the victim's distress continues; here one's empathic distress combines with blaming oneself not for causing the victim's distress but allowing it to continue.

5. Empathy as a Prosocial Motive

Empathy correlates positively with helping. There is also experimental evidence that arousal of empathic distress leads to helping; and that observers help more quickly the greater the victims' pain and the higher the observers' empathic distress. Furthermore, observers' empathic distress becomes less intense and they feel better if they help, but not if they don't help or if despite their best efforts and through no fault of their own, the victim's distress is not alleviated, which implies that the main objective of empathy-based helping is to alleviate the victim's distress. There is also evidence that empathy reduces both aggressive and manipulative (Machiavellian) behavior. It thus seems clear that empathic distress is a prosocial motive. See reviews by Hoffman (1978, 2000) and Eisenberg and Miller (1987).

Empathy has limitations, however (Hoffman 1984, 2000). (a) Egoistic motives may intervene and one may not help, even at the price of feeling guilty afterwards. (b) Empathic distress increases with the level of victims' distress, but empathic distress can become aversive enough to shift one's attention from the victim's to one's own personal distress (empathic overarousal). (c) People empathize with almost anyone in distress (the victims in research are usually strangers), but they empathize more with kin, in-group members, friends, and people who share their personal concerns (familiarity bias)—which may not be a problem except when intergroup rivalry contributes to intense empathic anger (hatred, ethnic cleansing) toward outgroup members. (d) People empathize more with victims who are physically present (here-and-now bias), probably because most empathy-arousing processes require immediate situational and personal cues.

6. Socialization by Parents and Peers

The role of parent as socializer has three facets: model, disciplinarian, nurturer. Each contributes to children's empathy and prosocial behavior, although discipline has been the most studied (Hoffman 2000). The

discipline method most likely to contribute to empathy, helping, and guilt over harming others is a type of reasoning called induction, often used by educated middle-class American parents when children harm or are about to harm someone. Inductions direct the child's attention to the victim's distress and may thus engage the empathy-arousing modes noted above. They also highlight the child's role in causing the victim's distress, which may lead to empathy-based transgression guilt. Power-assertive disciplines (physical force, threats, commands) are associated with low empathy, low helping, and low guilt, although some power assertion may be needed at times to get children to attend and process an induction's message.

Parents' behavior outside discipline encounters may provide prosocial models that reinforce children's empathic proclivity and legitimize helping (Eisenberg and Fabes 1998, Hoffman 2000). Examples are expressing compassion toward people in difficult straights such as the homeless; linking a television protagonist's feelings or situations with the child's own experience; pointing out similarities among all humans such as sadness due to separation and loss. Parent models may thus reinforce children's empathic dispositions especially when they are bystanders and someone needs help. Parent models may also make children more receptive to inductions, and, along with inductions, make children more receptive to claims of their peers (Hoffman 2000).

Piaget suggested in the 1930s that interaction with peers (not parents, who are too powerful) fosters prosocial moral development in children (Piaget 1932). Cognitive-developmental psychologists have recently elaborated this view (Turiel 1998), briefly as follows: children 'co-construct' and discover moral norms in the give-and-take of peer interaction (disputes, arguments, negotiations). These peer conflicts put pressure on children to take different points of view and to consider and coordinate the needs and rights of self and others (for example, regarding claims to ownership and possession of objects). This is a compelling hypothesis that merits research. Hoffman (2000) suggests that the peer interaction processes involved can serve to add the finishing touches to the prosocial outcomes of having inductive, nurturant, prosocial models as parents.

7. The Role of Cognition

Though we deal with empathic affect, cognition's importance is evident in two empathy arousal processes (verbally mediated association and perspective-taking): in the key role of self/other differentiation in empathy development; in the production of empathic anger and empathy-based guilt through causal attribution; and in children processing the information in inductions. Beyond that, cognitive development enables humans to form images, represent people and

events, and imagine themselves in another's place; and because represented people and events can evoke affect, victims need not be present for empathy to be aroused in observers. Empathy can be aroused when observers imagine or read about victims, discuss or argue about economic or political issues that involve victims, or make moral judgments about hypothetical moral dilemmas (see *Moral Reasoning in Psychology*) in which there are victims. See Hoffman (1987, 2000) for discussion of the importance of cognition, including the cognitive dimension of moral principles, in empathy's development and prosocial functioning.

8. Gender Differences

Girls are socialized more than boys to be kind and feel responsible for others' well-being. Girls are thus expected, increasingly with age, to be more empathic and prosocial and less likely to harm others. The research supports these expectations but in varying degrees (Eisenberg and Fabes 1998). Girls show more kindness and consideration for others and, to a lesser extent, are more likely to help and share. They are less likely to harm others and more apt to feel guilty when they do. They are also more empathic than boys, but this is clear only with self-report, especially questionnaire measures, in which it is obvious what is being assessed and responses are under the subject's control. It is not clear with unobtrusive naturalistic observations or physiological indices of empathy. The empathy findings may thus reflect subjects' conceptions of what boys and girls are supposed to be like. Whether they reflect gender differences in actual behavior is a matter for future research to decide.

See also: Adulthood: Prosocial Behavior and Empathy; Altruism and Prosocial Behavior, Sociology of; Antisocial Behavior in Childhood and Adolescence; Moral Reasoning in Psychology; Social Cognition and Affect, Psychology of; Social Support, Psychology of

Bibliography

- Eisenberg N, Fabes R A 1998 Prosocial development. In: Damon W (ed.) *Handbook of Child Psychology: Social, Emotional, and Personality Development*, 5th edn. Wiley, New York, vol. 3
- Eisenberg N, Miller P A 1987 The relation of empathy to prosocial and related behaviors. *Psychological Bulletin* **101**: 91–119
- Hoffman M L 1978 Empathy, its development and prosocial implications. In: Keasey C B (ed.) *Nebraska Symposium on Motivation*, University of Nebraska Press, Lincoln, NV. Vol. 25, pp. 169–218
- Hoffman M L 1981 Is altruism part of human nature? *Journal of Personality and Social Psychology* **40**: 121–37
- Hoffman M L 1984 Empathy, its limitations, and its role in a comprehensive moral theory. In: Kurtines W, Gewirtz J (eds.)

- Morality, Moral Behavior, and Moral Development*. John Wiley, New York, pp. 283–302
- Hoffman M L 1987 The contribution of empathy to justice and moral judgment. In: Eisenberg N, Strayer J (eds.) *Empathy and its Development*. Cambridge University Press, New York, pp. 47–80
- Hoffman M L 2000 *Empathy and its Development: Implications for Caring and Justice*. Cambridge University Press, New York
- Ickes W 1997 *Empathic Accuracy*. Guilford Press, New York
- Latane B, Darley J M 1970 *The Unresponsive Bystander: Why Doesn't He Help?* Appleton-Century Crofts, New York
- Lennon R, Eisenberg N 1987 Gender and age differences in empathy and sympathy. In: Eisenberg N, Strayer J (eds.) *Empathy and its Development*. Cambridge University Press, New York, pp. 195–217
- Piaget J 1932 *The Moral Judgment of the Child*. Harcourt, New York
- Turiel E 1998 The development of morality. In: Damon W (ed.) *Handbook of Child Psychology: Social, Emotional, and Personality Development*. Wiley, New York, Vol. 3
- Weiner B 1985 'Spontaneous' causal thinking. *Psychological Bulletin* 97: 74–84
- Zahn-Waxler C, Emde R N, Robinson J L 1992 The development of empathy in twins. *Developmental Psychology* 28: 1038–47

M. L. Hoffman

Prosopagnosia

Prosopagnosia is a neuropsychological condition involving loss of ability to recognize familiar faces. Prosopagnosic patients instead rely on voice, context, name, or sometimes clothing or gait to achieve recognition of people they know. The problem in face recognition is not due to blindness (the person with prosopagnosia can still see) or general intellectual impairment (people are still recognized from nonfacial cues). Yet even the most familiar faces may go unrecognized: famous people, friends, family, and the person's own face when seen in a mirror.

Face recognition impairments are socially disabling, and most patients develop strategies for tackling the problem. They may note carefully what clothes someone is wearing, ask relatives always to wear a particular distinctive item, or become adept at initiating or maintaining a conversation whilst they work out who they are talking to. These strategies are not completely effective. One prosopagnosic patient lost a legal action when he discussed his case with his opponent's lawyer—he thought it was his own advocate because he was wearing similar court clothes.

Prosopagnosia should not be confused with other types of impairment that can compromise face recognition (Young 1992). In prosopagnosia, recognition from other cues than the face is usually possible. In contrast, cases of inability to recognize people from face, voice, and name have been described—these

seem to reflect loss of semantic information about the identities of individuals. There are also reports of problems of name retrieval (anomia), in which familiar people are successfully recognized and appropriate semantic information is accessed, but their names cannot be recalled. Face recognition impairments therefore reflect damage to an underlying system which can break down in different ways.

Although reports of problems affecting face recognition after brain injury can be traced back to the nineteenth century, prosopagnosia was identified as a distinct neuropsychological problem by Bodamer in 1947 (Ellis and Florence 1990). Whilst it is generally considered a rare deficit, there are now several hundred case descriptions in the literature. These include cases where brain injury early or late in life has led to inability to recognize previously familiar faces, and developmental cases where the disorder is present from birth. People with prosopagnosia know when they are looking at a face, and can describe its features, but the loss of any sense of overt recognition is often complete, with no feeling of familiarity. In contrast, other aspects of face processing, such as the ability to interpret facial expressions, or to match views of unfamiliar faces, can remain remarkably intact in some (though not all) cases. The ability to recognize everyday objects other than faces may also remain good, and many prosopagnosic patients are able to read without difficulty. In one of the cases of early acquired deficit, a child who had remained unable to recognize any faces since suffering meningitis and subsequent complications in infancy was nonetheless able to learn to read (Young and Ellis 1989); this underlines the implication that reading and face recognition depend on different types of visual analysis.

Deficits commonly associated with prosopagnosia include a visual field defect in the left upper quadrant, achromatopsia (loss of color vision), and topographical disorders (inability to find one's way about). These are useful clinical pointers, but they are thought to be due to anatomical proximity of otherwise unrelated processes rather than having any direct functional significance; cases of prosopagnosia without each of these associated deficits have been described, and they have also each been reported in the absence of prosopagnosia.

Studies of prosopagnosia initially concentrated on the location of the lesions responsible, and to what extent the problem is specific to face recognition. More recent work has asked whether there exist different underlying forms of deficit, and has explored the implications of findings of covert recognition.

1. Lesions Causing Prosopagnosia

The underlying pathology involves lesions affecting ventromedial regions of occipitotemporal cortex; these include the lingual, fusiform, and parahippocampal

gyri, and more anterior parts of the temporal lobes. Functional imaging studies of normal observers have confirmed the importance of these regions to face perception (Kanwisher et al. 1997). Bilateral lesions are usually present in the relatively small number of cases that have come to autopsy (Damasio et al. 1982), but cases involving unilateral lesions of the right cerebral hemisphere have been reported in several computerized tomography (CT) and magnetic resonance imaging (MRI) studies, and the functional imaging findings with normal people also emphasize the importance of the right hemisphere.

2. A Face-specific Deficit?

Prosopagnosia involves a severe impairment of face recognition, yet a number of prosopagnosic patients can read and can recognize many everyday objects, so the deficit does not compromise all aspects of visual recognition. But is it only face recognition that is impaired?

In most cases, the answer is clearly 'no.' So, for example, recognition of familiar buildings, flowers, types of car, or animal species may all pose difficulties. Like faces, these are all visual categories with a high degree of interitem similarity.

This has led to the hypothesis that what is lost in prosopagnosia is the ability to recognize the individual members of categories that contain several items of similar appearance. Ability to identify the general category to which the items belong (face, building, flower, etc.) is preserved, but within-category recognition is defective. Although attractive, this idea has been undermined by (rare) reports of highly circumscribed impairments. One of De Renzi's (1986) cases could find his own belongings when they were mixed in with similar objects, identify his own handwriting, pick out a Siamese cat among photographs of other cats, recognize his own car without reading the number plate, and sort domestic coins from foreign coins. In all of these tasks he was therefore able to identify individual members of visual categories with high interitem similarity, yet he could not recognize the faces of relatives and close friends. Just as strikingly, the deficit can affect mainly *human* faces: when McNeil and Warrington's (1993) prosopagnosic patient took up farming, he was able to learn to recognize his sheep, and correctly identified several of them from photographs of their faces!

Although they are exceptionally rare, such cases imply that the possibility of face-specific deficits must be taken seriously. However, it may be important to reconsider the underlying issue. The presumption behind the interest in face-specific deficits has been that these might indicate whether the brain has an evolved neural substrate for the important social task of face recognition. But this rationale would only be

valid if the organization of the evolved substrate was so rigid that it could not become involved in other important visual discriminations. Studies of the properties of visual cells in the primate brain (Logothetis and Pauls 1995) and acquired expertise in humans (Diamond and Carey 1986) suggest, in contrast, a more plastic organization. It is possible that much of our visual expertise involves extending the sphere of involvement of cells that would otherwise become dedicated to face recognition.

3. Different Forms of Prosopagnosia

There has been much discussion as to whether prosopagnosia can have more than one different underlying cause. Often, a distinction is made between cases that seem to reflect defective face perception and those that are more like a loss of face memories. When face perception is defective, patients may be poor at telling male from female faces, or matching photographs of unfamiliar faces, whereas patients with a face memory problem may remain able to perform such tasks fairly normally.

For many years, this debate proved inconclusive because it was conceivable that the so-called perceptual cases simply reflected a more severe form of the same impairment. However, studies of ability to form mental images of premorbidly familiar faces have shown that someone suffering from the perceptual form of prosopagnosia can perform better than someone with a loss of face memories (Young et al. 1994). A plausible interpretation is that this happens because we need the same representation to conjure up a mental image of someone's face that we need to recognize that face. In perceptually based cases of prosopagnosia, these stored representations of familiar faces could remain intact, but a seen face could not access them because it is not perceived sufficiently accurately; imaging faces which cannot be recognized would then remain possible. In memory-based cases, stored memories cannot be used for imagery or for recognition—either because they are themselves damaged or because access to the store is defective.

4. Covert Recognition

The interpretation of some cases of prosopagnosia as a domain-specific impairment of face memory is strengthened by findings of covert recognition, which show clear parallels to priming effects found more generally in amnesia. Prosopagnosic patients usually fail all tests of overt recognition of familiar faces. They cannot name the face, give the person's occupation or other biographical details, or even state whether a face is that of a familiar person (all faces seem unfamiliar).

Surprisingly, however, there is substantial evidence of covert recognition from physiological and behavioral measures (Bruyer 1991).

Bauer (1984) measured skin conductance whilst a prosopagnosic patient, LF, viewed a familiar face and listened to a list of five names. When the name belonged to the face LF was looking at, there was a greater skin conductance change than when someone else's name was read out. Yet if LF was asked to choose which name in the list was correct for the face, his performance was at chance level. Hence, there was a marked discrepancy between LF's inability to identify the face overtly and the relatively good recognition shown using the indirect skin conductance measure.

A number of other indices of nonconscious, covert recognition in prosopagnosia have also been developed. Matching of familiar faces is better than for unfamiliar faces, priming has been found from familiar faces onto the recognition of name targets, and learning of correct face-name pairings is better than learning of incorrect pairings (Bruyer 1991). All of these effects can be demonstrated when overt recognition of the faces is at chance level.

Not all prosopagnosic patients show covert recognition, however. As might be expected, when there is evidence of substantial impairment of face perception, covert as well as overt performance can fall to chance level.

There is a parallel between preserved covert recognition abilities and those aspects of recognition that operate automatically for normal people. We cannot look at a familiar face and decide not to recognize it—the mechanisms responsible for visual recognition are not open to conscious control in this way. It seems that some of these automatic aspects of recognition continue to function in some cases of prosopagnosia. A simulation of this pattern can be made by halving some of the connection strengths in a computer model of the architecture of the recognition system (Young and Burton 1999). The network is then no longer able to classify face inputs as familiar, yet it continues to display priming effects. This helps us to understand how covert responses can be preserved when there is no overt discrimination, but it does not solve the more difficult issue of what is involved in being aware of recognizing a face.

Despite the extensive range of covert effects demonstrated in the laboratory, prosopagnosic patients do not act as if they recognize faces in everyday life. Instead, most are acutely conscious of their problems in face recognition. However, some patients lack insight into their disability and show unawareness of impairment (anosognosia). Case SP was very poor at recognizing familiar faces, yet she maintained that she recognized faces 'as well as before,' even when directly confronted with her failure to recognize photographs of familiar faces (Young et al. 1990). In contrast, SP showed adequate insight into other physical and

cognitive problems produced by her illness—her lack of insight into her face recognition impairment involved a deficit-specific anosognosia. Such deficit-specific anosognosias may reflect impairment to mechanisms used for monitoring performance in everyday life; we can all make occasional errors of recognition, so we constantly monitor our own performance to correct these errors quickly and avoid embarrassment whenever possible.

Remarkably, the deficits of overt recognition noted in prosopagnosia need not be absolute. Sergent and Poncet (1990) found that their patient, PV, could achieve overt recognition of some faces if several members of the same semantic category were presented together. This only happened when PV could determine the category herself; otherwise she continued to fail to recognize the faces overtly even when the occupational category was pointed out to her. Sergent and Poncet (1990) suggested that the simultaneous presentation of several members of the same category may have temporarily raised their activation above the threshold at which recognition could be achieved.

Such findings in prosopagnosia show that the boundary between awareness and lack of awareness is not as completely impassable as it seems to the patients' everyday experience. However, the circumstances under which this has been found to happen are at present very limited, and it has not yet proved possible to find a way to provide significant remedial assistance in real life.

See also: Face Recognition Models; Face Recognition: Psychological and Neural Aspects

Bibliography

- Bauer R M 1984 Autonomic recognition of names and faces in prosopagnosia: A neuropsychological application of the guilty knowledge test. *Neuropsychologia* **22**: 457–69
- Bruyer R 1991 Covert face recognition in prosopagnosia: A review. *Brain and Cognition* **15**: 223–35
- Damasio A R, Damasio H, Van Hoesen G W 1982 Prosopagnosia: Anatomic basis and behavioral mechanisms. *Neurology* **32**: 331–41
- De Renzi E 1986 Current issues in prosopagnosia. In: Ellis H D, Jeeves M A, Newcombe F, Young A (eds.) *Aspects of Face Processing*. Martinus Nijhoff, Dordrecht, The Netherlands, pp. 243–52
- Diamond R, Carey S 1986 Why faces are and are not special: An effect of expertise. *Journal of Experimental Psychology: General* **115**: 107–17
- Ellis H D, Florence M 1990 Bodamer's (1947) paper on prosopagnosia. *Cognitive Neuropsychology* **7**: 81–105
- Kanwisher N, McDermott J, Chun M M 1997 The fusiform face area: A module in human extrastriate cortex specialized for face perception. *Journal of Neuroscience* **17**: 4302–11
- Logothetis N K, Pauls J 1995 Psychophysical and physiological evidence for viewer-centred object representations in the primate. *Cerebral Cortex* **5**: 270–88

- McNeil J E, Warrington E K 1993 Prosopagnosia: A face specific disorder. *Quarterly Journal of Experimental Psychology, Series A* **46**: 1–10
- Sergent J, Poncet M 1990 From covert to overt recognition of faces in a prosopagnosic patient. *Brain* **113**: 989–1004
- Young A W 1992 Face recognition impairments. *Philosophical Transactions of the Royal Society, London, Series B* **335**: 47–54
- Young A W, Burton A M 1999 Simulating face recognition: Implications for modelling cognition. *Cognitive Neuropsychology* **16**: 1–48
- Young A W, De Haan E H F, Newcombe F 1990 Unawareness of impaired face recognition. *Brain and Cognition* **14**: 1–18
- Young A W, Ellis H D 1989 Childhood prosopagnosia. *Brain and Cognition* **9**: 16–47
- Young A W, Humphreys G W, Riddoch M J, Hellawell D J, De Haan E H F 1994 Recognition impairments and face imagery. *Neuropsychologia* **32**: 693–702

A. W. Young

Prosopography (Collective Biography)

Prosopography, or collective biography, has developed considerably since the 1960s, expanding in the fields of modern and contemporary history, after the tool was invented and applied primarily in ancient and medieval history. The method's principle is simple: defining a population on the basis of one or several criteria and designing a relevant biographical questionnaire containing a range of variables or criteria which serve to describe it in terms of social, private, public and/or cultural, ideological or political dynamics, depending on the population under scrutiny and the questionnaire that is being used. In the words of Lawrence Stone: 'Prosopography is the investigation of the common background characteristics of a group of actors in history by means of a collective study of their lives (Stone 1971, p. 46) Once the documentation has been compiled, which is the most time-consuming aspect of the work, the data can be evaluated with the help of a number of techniques, both quantitative and qualitative, including manual counting or computerized processing, statistical tables or factor analyses, depending on the scope and sophistication of the questionnaire and the sample.

To appreciate the success of this method and demonstrate its extension to different branches of history, several historiographic references are cited in Sect. 1. To gain insight into its merits and limitations, it is necessary to analyze a few examples of its application, taken primarily from recent periods see Sect. 2 where other social history methods can be used, permitting us to weigh the advantages and disadvantages of both alternatives. An attempt is made to define the heretofore untapped potential of the method in the area of comparative studies in Sect. 3. This possibility has been the least explored to date.

1. Historiographic References

Ancient history scholars claim to be the fathers of prosopography (Chastagnol 1970, Nicolet 1970). The term was conspicuously used in the modern sense by T. Mommsen in the announcement of the *Prosopographia Imperii Romani*, published in 1897 (Groag and Stein 1897, Lalouette 1999). This serial was imitated for other periods of ancient history throughout the twentieth century, both in collective works and in individual monographs. Prime examples are *The Prosopography of the Later Roman Empire* by A. H. M. Jones and J. R. Martindale (1971–92), and *Prosopographie chrétienne du Bas-Empire*, of which two volumes have appeared, edited by André Mandouze (1982) and Charles Pietri and Luce Pietri (1982). Countless monographs contain prosopographies of elites of antiquity. One of the earliest is Jean Kirchner's *Prosopographia attica* (1901), followed by Paul Poralla's 1913 work on the Lacedaemonians and the classics by Friedrich Münzer (1920), Ronald Syme (1939) and Claude Nicolet (1966), etc. The application of prosopography to medieval times started to develop between the world wars. Studies focus on administrative personnel and the entourage of the sovereigns of the principal Western European states. Since the 1960s, the body of literature has become quite copious, as various surveys (Bulst 1986, Griffiths 1986) reveal. The same goes for modern history, where the initiator was Sir Lewis Namier (1929, 1930), who studied English political elites. His works inaugurated an ongoing series of studies on the House of Commons from the Middle Ages to 1832 (Cruikshanks 1986). French and Italian investigation has concentrated on diverse selections of officials taken from public administration and the judiciary as well as on ecclesiastical, intellectual, financial and commercial elites (see Charle et al. 1980, Nagle 1986).

In the area of contemporary history, work began during the second half of the 1960s. Most of it was inspired by the perspective on elites introduced by sociologists, particularly American, influenced by the theses of Pareto and Mosca beginning in the 1950s. In France, one of the issues that stimulated research of this kind was the debate on the French Revolution opened by the theses of F. Furet and D. Richet (Richet 1969, discussed by Vovelle 1974), proposing an alternative to the Marxist interpretation then dominant in France, founded on prosopographic study of elites before and after the upheaval. This is the background of the investigation by L. Bergeron and G. Chaussinand-Nogaret on notables of the Napoleonic Era (1979). Similar works on prominent figures of the nineteenth century and the bourgeoisie since the 1960s have relied on the prosopographic method. This approach was inspired by Ernest Labrousse (Tudesq 1964) (for an overall survey, see Charle 1992). French sociology of education, influenced by the works of Pierre Bourdieu (Bourdieu and Passeron 1970,

Bourdieu 1989), has also been the source of numerous intellectual and academic prosopographies (Karady 1972, 1973). In Britain, Germany, and the USA and more recently in Italy, Spain, and Switzerland, collective biographies of political, economic, religious, intellectual, artistic, and militant elites have proliferated. They are an outgrowth of great national historiographic debates on subjects such as the foundation of political rifts in the British parliament, the evolution and function of English nobility, the distinguishing characteristics of German, Swiss, and Italian bourgeoisies (Kocka and Frevert 1988, Tanner 1990, Augustine 1994, Malatesta 1995), the formation and reproduction of American elites (Baltzell 1958, Jeher 1973), and so on. In contemporary history, the method has also been applied to larger groups than elites, which are the sole focus of prosopography in the spheres of ancient, medieval and modern history because records on other segments of the population are not available. The history of women, the middle classes, lower classes and even marginals have also availed themselves of collective biography in recent decades thanks to abundant printed sources, oral history and new assessments of large biographic files which already existed (Dreyfus et al. 1996, Omnès 1997).

2. *Merits and Limitations*

The proliferation of studies in contemporary history and historical sociology using the prosopographic method is due to intellectual trends and techniques developed since the 1970s. The concomitant decline of global statistical methods and of the intellectual impact of Marxism, which featured social classes with huge populations, and the desire, typical of societies focused more and more on the individual, to come to grips with individual experience and the diversity of social patterns have inclined historians of recent times to emphasize social microhistory, which gives priority to prosopography or collective biography. Finally, the widespread advance of computerization has facilitated complex processing of bigger and bigger databases (Stone 1971, Millet 1985). Whereas in former times the lack of mass documentation required patient gathering of data on the individual level before any collective generalizations could be made, in recent periods resorting to prosopography has been construed as a change of perspective from traditional social history based on serial documentation.

Thanks to this technique, since the 1970s historians have increasingly set their sights on gaining intimate knowledge of small problem-related constituencies. Thus, S. W. Serman's study of officers of the Second Republic and the Second Empire traced the long-lasting confrontation in nineteenth century France between the aristocratic ideal and the rising democratic tide (Serman 1978). Investigations of elites

(high-ranking personalities of the July Monarchy (Tudesq 1964, Girard et al. 1976), the business leaders of the Second Empire (Plessis 1985, Barjot et al. 1991–1999), ruling circles of the Third Republic (Estèbe 1982, Charle 1987), and deputies to the national assemblies in Paris and Frankfurt in 1848–49 (Best 1990)) were attempts to understand the evolution of models of reproduction dominant in one regime toward those prevailing in other systems, the disparity between myths justifying the social order and sociological mechanisms, the possible social impact of political revolutions, reasons for the unique French road to industrialization, the peculiar German path to an incomplete parliamentary system, and the relations between English entrepreneurs and previously dominant elites (Berghoff 1991). In short, these collective biographies allow us to revamp our answers to some major questions as well as to conduct an in-depth reassessment of cherished tenets of social history without confining ourselves to the preconceived structure of serial and quantifiable sources. In these new approaches, groups define themselves with reference to relational characteristics or their reciprocal images or even their capacity to impose an image of themselves on others and on the majority of their members. The notion of construction of the object thus becomes decisive in this microhistory based on collective biographies.

This is precisely the point which since 1971 has triggered premature criticism of prosopography, summed up by Lawrence Stone with reference to modern history studies inspired by Sir Lewis Namier: the biases of official sources on which the biographies rely lead to a partial view of reality, the delimitation of populations is largely arbitrary, and criteria applied for the biographies are often restrictive. All this contributes to an elitist, cynical and conformist perspective on the spheres of leadership and their relations with society at large (Stone 1971). These criticisms are not valid unless the historians confuse the method with its objective and forget that they can never consider more than a fraction of reality as determined by the sources which they select and the limitations of their biographical questionnaires. They are less likely to fall into these traps for recent periods, where the findings of collective biographies can be confronted with other sources, where partial prosopographies can be compared and merged, and where the diversity of groups readily available for investigation goes well beyond the scope of the elites to which earlier studies were confined.

The prosopographical historian navigates permanently between two reefs: the undefined biography of individuals (with the risk of losing sight of the collective dimension) and the opposite: the expansion of prolific samples to the scale of the entire society (with the risk of reducing the questionnaire to its lowest common denominator). To escape from this dilemma, collective investigation by means of large

research teams, adopting common principles, has been applied with varying success in different countries. Examples include the ongoing study of the British House of Commons (Namier and Brooke 1964), the study of mayors in France (Agulhon et al. 1986), the investigation of members of parliament of the French Third Republic (Corbin and Mayeur 1995), the collective biographies of legislators by the Zentrum für Historische Sozialforschung in Cologne (Schröder et al. 2000, Best and Cotta 2000). These surveys require a long-term institutional financial commitment, which has become more and more difficult to find since profitability criteria were introduced in academic research. They also assume long-term involvement of loyal research personnel, which is less and less compatible with normal levels of staff turnover. This is why the principle of coordinated monographs, being more flexible and decentralized, appears to be more realistic in the contemporary university environment. It also overcomes the need to sacrifice the originality of each individual's work in the anonymity of a hierarchic endeavor. However, to avoid the risk of dissipation or non-comparability of partial results, it posits a minimum of harmonization and entails meetings to fine-tune the common questionnaires, coding systems and evaluations. In exploiting collective biography, one is not only forced to gauge the conditions of sample validity. One must also recognize the limits of the categories applied *a priori* to the data. Since the responses to the questions which are included involve social stakes, the process of interpretation of prosopographic data is characterized, more so than in other methods, by historians who consciously or unconsciously put their mark on the results at all stages of the study: sampling, data gathering, coding, and processing.

3. Toward Comparative Prosopography

In France and increasingly in most other developed countries, nearly all groups that lend themselves to collective biography have been adopted by a biographer: aristocracies, dignitaries, urban bourgeoisies, administrative elites, financial, commercial and entrepreneurial elites, intellectual and university elites, artists, middle-class professionals (physicians, jurists, journalists, professors), students, officeholders at all levels, labor leaders and militant feminists, even marginals, etc. Is a decline of the method of collective biography imminent? Will it be stopped in the name of the unyielding law of diminishing returns and as a result of the resurgence of historiographic topics which are less amenable to this approach (cultural history, history of memory, history of collective sensibilities)? Three considerations allow us to reject this premature diagnosis of a decline in the advantages of a method which has been so fruitful in the area of recent social history.

First, apart from elites, other social groups which are beginning to revise their social history thanks to insiders' collective biographies (middle classes and, increasingly, lower classes, not only through the eyes of their militant elites, but also in the form of life stories of privileged witnesses) (Ozouf et al. 1992, Pudal 1991, Gribaudi 1987) are far from being thoroughly investigated. Application of the method to these new fields requires new reflection on the composition of samples (which can no longer be exhaustive) and on the relations between individual paths and morphology: the internal diversity of the configurations is directly proportional to the size of the target population. This again raises the issue of sample representativeness, which had disappeared when investigators employing prosopography of elites resorted to exhaustive polls. All these groups are transitory, while elites, by definition at the top of the heap, are target groups; hence characteristics of the individuals of which such groups are composed cannot be viewed in isolation, but must be understood as vectors of multiple strategies.

Second, European historiographers are acquainted with the adjustments required to enable them to merge complementary studies to arrive at a homogeneous scrutiny of comparable groups from different societies. Thus, the comparative research coordinated by Jürgen Kocka on European bourgeoisies has demonstrated that not all of the difficulties in analyzing this group at an international level are due to the unequal advancement of studies; above all they arise from the collateral persistence of divergent investigations—some inspired by classical methods of social history, particularly in Germanic and Anglo-Saxon countries, others, particularly in France, primarily involving research based on collective biographies, which result in different approaches to different problems, and at times to inappropriate comparisons (Kocka and Frevert 1988, Charle 1990).

Collective biography, therefore, still has a nearly unexploited field at its disposal: that of comparative prosopography. The principal, eternal objection to comparative historical study is the necessity, given the abundance of material, to conduct secondhand research, using monographs prepared by other scholars, with the compounded risks of hasty generalization of conclusions taken out of their unique context and of the permanent lack of correspondence in the compared data, particularly in social history, on account of the recurring coding problem. Despite these obstacles, German researchers have already forged ahead, and certain French historians and specialists in England have followed suit since the 1990s (Best 1990, Berghoff and Möller 1994, Charle 1994, Siegrist 1995, Cassis 1997, Ruggiu 1997). Collective biography is an extraordinarily demanding methodology in terms of research time and rigorous organization of data. Comparative prosopography multiplies the time it takes to assemble the data by at least two—for two

countries—or more, if one wants to compare a greater number of cases. The problem is compounded by the extra time needed to design the research and evaluate the data. In order to test the initial hypothesis of feasibility and fruitfulness of a comparative project, it is convenient to start in a domain where one of the two countries involved already has a critical mass of usable data allowing researchers to define an analogous scope of investigation in the other selected country, which must possess a comparable, evident interest in the chosen elite. This explains the fact that comparative prosopographies deal primarily with the best-documented and most frequently studied elites (in big business, politics, higher education and high society), as in the cases mentioned previously. Collective biographies at a national level have permitted us to better understand the internal rifts in the various groups which have been studied as well as their social and generational dynamics. In certain favorable cases, it has been possible to relate these rifts to political, ideological or religious stances or apply them to better understand certain failures or successes. Comparative prosopography should allow us to go even further in this interpretive analysis by relativizing correlations which were taken for granted in a given national or social setting. They should account for common trends which transcend borders and the peculiarities of one moment in time, of one milieu or one nation. For instance, the generally accepted notion that the expansion of higher education translates into an opening of university teaching positions to a wider segment of society is disproved if one compares the collective biographies of the professors of the University of Berlin (Arts and Sciences) with their counterparts at the Sorbonne at the turn of the nineteenth and twentieth centuries. The major divergence accounts for the misunderstandings between the two academic communities even when they attempt to cooperate, and goes a long way toward explaining their widely varying relations with other elites during political crises (Charle 1994).

The goal of comparative prosopography, as with any kind of prosopography, is to employ collective biographies to portray the real social behavior of the institutions or the milieus in which the individual subjects of the study interact (Genet and Lottes 1996). The main danger in applying comparative prosopography is therefore, as with many erudite ventures, that one will discover nothing but well-established or generally acknowledged truths in each country concerned. One usually arrives at this impasse or this disappointing result when one deals with samples that are too vast to lend themselves to controlled comparisons which would allow the study to qualify any dominant feature or put it into perspective. In light of this danger it is essential to make a particularly judicious choice of variables to be compared. We face a series of common methodological issues which transcend historical periods.

Their resolution would allow us to launch a wider debate on the possible foundations of a comparative sociocultural history of European society.

See also: Antiquity, History of; Biographical Methodology; Psychological Perspectives; Biography and Society; Biography: Historical; Comparative History; Ethnoarchaeology; Historiography and Historical Thought: Current Trends; History: Forms of Presentation, Discourses, and Functions; History: Theories and Methods; Memory: Collaborative; Middle Ages, The; Quantification in History; Social History; Social Inequality in History (Stratification and Classes)

Bibliography

- Agulhon M et al. 1986 *Les Maires en France du Consulat à nos jours*. Publications de la Sorbonne, Paris
- Augustine D L 1994 *Patricians and Parvenus. Wealth and High Society in Wilhelmine Germany*. Berg, Oxford
- Autrand F (ed.) 1986 *Prosopographie et genèse de l'État moderne*. L'École normale supérieure des jeunes filles, Paris
- Baltzell E D 1958 *Philadelphia Gentlemen. The Making of a National Upper Class*. Free Press, Glencoe, IL
- Barjot D (ed. in chief) 1991–1999 *Les patrons du Second Empire*. Cenomane, Picard, Le Mans, Paris
- Bergeron L, Chaussinand-Nogaret G 1979 *Les Masses de Granit*. EHESS, Paris
- Berghoff H 1991 *Englische Unternehmer 1870–1914. Eine Kollektivbiographie führender Wirtschaftsbürger in Birmingham, Bristol und Manchester*. Vandenhoeck & Ruprecht, Göttingen, Germany
- Berghoff H, Möller R 1994 Tired pioneers and dynamic newcomers? A comparative essay on English and German entrepreneurial history, 1870–1914. *Economic History Review* XLVII 2: 262–87
- Best H 1990 *Die Männer von Bildung und Besitz. Struktur und Handeln parlamentarischer Führungsgruppen in Deutschland und Frankreich, 1848/49*. Droste Verlag, Düsseldorf, Germany
- Best H, Cotta M (eds.) 2000 *Parliamentary representatives in Europe, 1848–2000. Legislative Recruitment and Careers in Eleven European Countries*. Oxford University Press, Oxford, New York
- Bourdieu P, Passeron J-C 1970 *La reproduction*. Minuit, Paris
- Bourdieu P 1989 *La Noblesse d'État, grandes écoles et esprit de corps*. Paris, Minuit Paris
- Bulst N, Genet J-Ph (eds.) 1986 *Medieval Lives and the Historian*. Studies in Medieval Prosopography, Kalamazoo, MI
- Bulst N 1986 La recherche prosopographique récente en Allemagne (1250–1650). Essai d'un bilan. In: Autrand F (ed.). *Prosopographie et genèse de l'État moderne*. L'École normale Supérieure des jeunes filles. Paris, pp. 35–52
- Cassidy Y 1997 *Big Business: the European Experience in the Twentieth Century*. Oxford University Press, Oxford, New York
- Charle C et al. 1980 *Prosopographie des élites françaises (XVIè–XXè siècles). Guide de recherche*. CNRS, Institut d'histoire moderne et contemporaine, Paris
- Charle C 1987 *Les Élités de la République (1880–1900)*. Fayard, Paris
- Charle C 1990 A la recherche des bourgeoisies européennes. *Le Mouvement Social* 153: 91–7

- Charle C 1992 The present state of research on the social history of elites and the bourgeoisie: a critical assessment on middle class and elites *Contemporary European History* 1(1): 99–112
- Charle C 1994 *La République des universitaires (1870–1940)*. Le Seuil, Paris
- Charle C 1996 Du bon usage de la biographie sociale comparée. In: Dreyfus M, Pennetier C, Viet-Depaule N (eds.) *La part des militants, Biographie et mouvement ouvrier: Autour du Maïtron, Dictionnaire biographique du mouvement ouvrier français*. Editions de l'Atelier, Paris, pp. 51–72
- Chastagnol A 1970 La prosopographie méthode de recherche sur l'histoire du Bas Empire. *Annales ESC* 25(5): 1229–35
- Corbin A, Mayeur J-M (eds.) 1995 *Les Immortels du Sénat*. Publications de la Sorbonne, Paris
- Cruikshanks E 1986 Multi-biographical analysis as an approach to parliamentary history. In: Autrand F (ed.) *Prosopographie et genèse de l'État moderne*. L'École normale supérieure de jeunes filles, Paris, pp. 335–44
- Dreyfus M, Pennetier C, Viet-Depaule N (eds.) 1996 *La part des militants, Biographie et mouvement ouvrier: Autour du Maïtron, Dictionnaire biographique du mouvement ouvrier français*. Editions de l'Atelier, Paris
- Estèbe J 1982 *Les Ministres de la République (1871–1914)*. Presses de la Fondation Nationale des Sciences Politiques, Paris
- Genet J-Ph, Lottes G (eds.) 1996 *L'État moderne et les élites XIII^e–XVIII^e siècles. Apports et limites de la méthode prosopographique*. Publications de la Sorbonne, Paris
- Girard L, Serman W, Gossez R 1976 *La Chambre des députés en 1837–39*. Publications de la Sorbonne, Paris
- Gribaudo M 1987 *Itinéraires ouvriers. Espaces et groupes sociaux à Turin au début du X^e siècle*. Editions de l'EHESS, Paris
- Griffiths R 1986 Bureaucracy and the English State in the Later Middle Ages. In: Autrand F (ed.) *Prosopographie et genèse de l'État moderne*. L'École normale supérieure des jeunes filles, Paris, pp. 52–65
- Groag E, Stein A (eds.) 1897 *Prosopographia Imperii Romani Saec I. II III Pars I*. Walter de Gruyter, Berlin and Leipzig, Germany
- Jaher F C (ed.) 1973 *The Rich, the Well born, and the Powerful. Elites and Upper Classes in History*. University of Illinois Press, Urbana, IL
- Jones A H M, Martindale J R (eds.) 1971, 1980, 1992 *The Prosopography of the Later Roman Empire*. Cambridge University Press, Cambridge
- Karady V 1972 Normaliens et autres enseignants à la Belle Époque. Note sur l'origine sociale et la réussite dans une profession intellectuelle. *Revue française de sociologie* 13(1): 35–58
- Karady V 1973 L'expansion universitaire et l'évolution des inégalités devant la carrière d'enseignant au début de la III^e République. *Revue française de sociologie* 14(4): 443–70
- Kocka J, Frevert U (eds.) 1988 *Bürgertum im 19. Jahrhundert*. Deutsche Taschenbuch Verlag, Munich, Germany
- Lalouette J 1999 De l'exemple à la série: histoire de la prosopographie. *Siècles, cahiers d'histoire des entreprises et des communautés* 10: 3–20
- Malatesta M (ed.) 1995 *Society and the Professions in Italy, 1860–1914*. Cambridge University Press, Cambridge
- Mandouze A (ed.) 1982 *Prosopographie chrétienne du Bas-Empire 1 Prosopographie de l'Afrique chrétienne (303–533)*. Editions du CNRS, Paris
- Millet H (ed.) 1985 *Informatique et prosopographie*. Editions du CNRS, Paris
- Münzer F 1920 *Römische Adelsparteien und Adelsfamilien*. J. B. Metzlersche Verlagsbuchhandlung, Stuttgart, Germany
- Nagle J 1986 Prosopographie et histoire de l'État: la France moderne XVI^e–XVIII^e siècles. In: Autrand F (ed.) *Prosopographie et genèse de l'État moderne*. L'École normale supérieure des jeunes filles, Paris
- Namier L 1929 *The Structure of Politics at the Accession of George III*. Macmillan, London
- Namier L 1930 *England in the Age of the American Revolution*. Macmillan, London
- Namier L, Brooke J 1964 *The House of Commons, 1754–1790*. Published for the History of Parliament Trust by her Majesty's Stationery Office, London
- Nicolet C 1966 *L'Ordre équestre à l'époque républicaine (312–43 avant J.-C.)*. De Boccard, Paris
- Nicolet C 1970 Prosopographie et histoire sociale: Rome et l'Italie à l'époque républicaine. *Annales ESC* 25(5): 1209–28
- Omnès C 1997 *Ouvrières parisiennes. Marchés du travail et trajectoires professionnelles au 20^e siècle*. Éditions de l'EHESS, Paris
- Ozouf J et al. 1992 *La République des instituteurs*. Gallimard, Le Seuil, Paris
- Pietri C, Pietri L (eds.) 1982 *Prosopographie chrétienne du Bas-Empire. 2 Prosopographie de l'Italie chrétienne (303–533)*. Ecole française de Rome, Rome
- Plessis A 1985 *Régents et gouverneurs de la Banque de France (1852–1870)*. Droz, Geneva, Switzerland
- Poralla P 1913 *Prosopographie der Lakedaimonier bis auf die Zeit Alexanders des Grossen*. Kommissions-Verlag von Max, Breslau, Germany
- La prosopographie, problèmes et méthodes 1988. In: *Mélanges de l'École française de Rome, Moyen Age, Temps modernes* 100 (1): 5–284.
- Pudal B 1991 *Prendre parti, pour une sociologie historique du PCF*. Presses de la Fondation Nationale des Sciences Politiques, Paris
- Richet D 1969 Autour des origines idéologiques lointaines de la Révolution française. *Annales ESC* 24(1): 1–23
- Ruggiu F-J 1997 *Les élites des villes moyennes en France et en Angleterre XVII^e–XVIII^e siècles*. L'Harmattan, Paris
- Schröder W-H, Weege W, Zech M 2000 Historische Parlamentarismus-, Eliten- und Biographieforschung, Forschung und Service am Zentrum für Historische Sozialforschung, Köln, Zentrum für Historische Sozialforschung. *Historische Sozialforschung Supplemente/Beiheft* 11
- Serman S-W 1978 *Le corps des officiers sous la Deuxième République et le Second Empire*. Atelier de reproduction des thèses de Lille III, Lille, France
- Siegrist H 1995 *Advokat, Staat und Bürger*. V. Klostermann, Frankfurt, Germany
- Stone L 1971 Prosopography. *Daedalus* 100(1): 46–79
- Syme R 1939 *The Roman Revolution*. Clarendon Press, Oxford
- Tanner A 1990 Aristokratie und Bürgertum in der Schweiz im 19. Jahrhundert: Verbürgerlichung der 'Herren' und aristokratische Tendenzen im Bürgertum. In: Brändli S et al. (ed.) 1990 *Schweiz im Wandel. Studien zur neueren Gesellschaftsgeschichte*. Helbing & Lichtenhahn, Basel & Frankfurt, pp. 209–28
- Tudesq A-J 1964 *Les grands notables en France (1840–49). Étude historique d'une psychologie sociale*. Presses Universitaires de France, Paris
- Vovelle M 1974 L'élite ou le mensonge des mots. *Annales ESC* 29(1): 49–72

Kirchner J 1901 *Prosopographia attica*. Reimer, Berlin, Germany

C. Charle

Prospective Memory, Psychology of

Prospective memory refers to remembering to perform an action in the future such as remembering to pick up a loaf of bread on the way home, to give a friend a message, to send a birthday card to your aunt, to take your medication, and to pick someone up at airport at 4 p.m. Ellis (1996, p. 1) defines it somewhat more precisely as the 'realization of delayed intentions and associated actions.' Prospective memory can be contrasted with retrospective memory, which is memory for past events.

Examples of retrospective memory include remembering the name of your eighth grade teacher, the causes of the World War I, the formula for the area of a circle, what you had for lunch yesterday, and a list of words learned in an experiment 15 minutes earlier. In thinking about the distinction between prospective and retrospective memory, it is worthwhile considering that many of the same processes may be involved in retrospective and prospective memory tasks. It is interesting to note that all prospective memories are to some extent retrospective in the sense that you have to have retrospective memory for the action that is to be performed and when it is to be performed in order to remember to perform it on time. For example, in order to remember to give a friend a message, you have to remember the message and to whom it is to be given. In addition, however, successful prospective memory requires that you actually remember to do this when you see your friend, and this is the prospective component of the memory. One fairly common error that is made in this situation is that upon seeing your friend, you remember that you have a message to give but forget the message (a retrospective memory problem). Another is that you forget to give the message even though you have good memory for it (a prospective memory failure).

1. Contrasts with Retrospective Memory

When studying retrospective memory (at least explicit forms of retrospective memory) in the laboratory, the experimenter presents some materials for learning. After a delay of some duration, the experimenter puts participants in a retrieval mode and asks them to recollect intentionally as they perform a designated criterial task such as free or cued recall. By contrast, when studying prospective memory, the experimenter asks participants to perform an action or several actions at a later point in time (either at a particular

time or when a certain event occurs; see Sect. 2.2 for examples). The interest is in whether participants remember to perform the task at the appropriate time/event and without an external agent (e.g., the experimenter) prompting participants to remember at the critical period. A key difference then between retrospective and prospective memory is that while the process of remembering is initiated by the experimenter in retrospective memory, it is initiated by the participant in prospective memory.

This self-cueing feature of prospective memory, along with the fact that retrieval needs to occur sometime in the future, places a different and perhaps more complex combination of demands (than typical retrospective memory tasks) on our cognitive systems. Indeed, many researchers have argued that prospective memory involves much more than memory processes (Dobbs and Reeves 1996, Ellis 1996, Burgess and Shallice 1997, Marsh et al. 1998b). They make the case that plan creation, plan modification, monitoring or rehearsal of the conditions for performing the action, and output monitoring, as well as individual factors such as working memory resources and personality variables, are all essential to a complete understanding of prospective memory.

Despite the differences just mentioned, it is likely that prospective and retrospective memory share similar memory processes and, accordingly, our vast understanding in the retrospective memory arena will be useful for understanding prospective memory. For example, given that many prospective memory tasks involve realizing a connection between an event (seeing a friend) and an intended action (giving a message), our basic understanding of associative memory processes should prove useful in understanding prospective memory. Also, as in retrospective memory, it seems that the level of processing of the target event is critical for good prospective memory as is the overlap between processing of the target event at encoding and retrieval (McDaniel et al. 1998).

2. Research in Prospective Memory

2.1 Historical Overview

Nearly all memory research conducted thus far has focused on retrospective memory, and this seems inconsistent with the nature of our everyday memory demands. Our lives are replete with prospective memory demands such as fulfilling an appointment, giving a message, remembering to flip the pancakes at the appropriate time, and returning a library book. To get some sense of the prevalence of prospective memory in everyday life, for several years G. O. Einstein asked students in his classes to write down the last thing they remembered forgetting. Interestingly, 62 percent of college-age students and 54 percent of older students (60 years and older) listed responses that were prospective memory failures. Given its prevalence in

everyday living, there is a real gap in our scientific understanding of prospective memory. It is difficult to overemphasize this gap as Kvavilashvili and Ellis (1996) in a recent review found only 45 prospective memory papers published over the prior 20 years, and only a small percentage of these papers consisted of rigorous experimental work in major journals. However, as evidenced by the inclusion of prospective memory in this volume, there now appears to be strong interest in the area.

2.2 Experimental Paradigms

Initial research in prospective memory required prospective memory responses to be initiated outside of the laboratory. For instance, subjects were asked to return postcards (e.g., Meacham and Singer 1977) or make telephone calls at specified times (Moscovitch 1982). Although these paradigms were valuable in producing the first investigations of factors thought to influence prospective remembering (like a person's age), they did not allow control of extraneous factors that affect prospective remembering. For instance, in these experiments there may have been variation in the degree to which subjects used external aids such as calendars to help complete the prospective memory task.

Empirical and theoretical work in prospective memory has increased significantly with the introduction of laboratory paradigms that bring experimental control and precision to the study of prospective memory. One influential paradigm, developed by Einstein and McDaniel (1990), requires that the subject be busily engaged in a primary task that serves as a cover task for the prospective memory component. During the instructions for the primary task, subjects are also instructed that they are to remember on their own to perform another action (e.g., press a special function key on the keyboard) at some designated point after they begin the primary task.

Sometimes the designation is an event embedded in the primary task (e.g., a particular word, a particular type of question) and other times the designation is a period of time that has elapsed (e.g., after 5 minutes). In this paradigm, the retrospective memory demands are designed to be simple and well remembered so that the major focus is whether participants remember to perform the intended action at the appropriate point. Other paradigms that focus on planning involve having participants remember to perform multiple sets of activities, each with their own constraints (e.g., Kliegel et al. 2000).

2.3 Principal Findings and Theoretical Perspectives

The general theoretical orientation guiding prospective memory research has turned on potential differ-

ences between prospective memory processes and retrospective memory processes. Unique features of prospective memory seem possible at encoding, storage, and retrieval stages of memory.

Encoding refers to rehearsal and elaboration of information so as to enter or commit that information to memory. People may use different encoding strategies for prospective memory tasks than they use for retrospective memory tasks. Consistent with this possibility, Koriati et al. (1990) found that actions encoded to be executed later are better recalled than actions that are encoded for a later recall test. It is likely that additional encoding processes found in prospective memory tasks are related to the requirement to remember to perform the intended action. Most prominently, during encoding of a prospective memory intention people often construct plans for how and when the intended activity will be accomplished (Ellis 1996, Marsh et al. 1998b). These plans may also anticipate the conditions or arrange the conditions under which the intended action must be retrieved, information that may not typically be considered when engaged in encoding for a retrospective memory task.

Only a few studies have examined the planning processes in prospective memory, but it is already clear that individuals vary in their prospective memory planning in several ways. People vary in the extent to which they use external aids like notes or lists to keep track of their plans (Marsh et al. 1998b). The complexity of the plans also varies and for older adults especially, more complex plans seem to be related to the effectiveness with which the intention is eventually executed (Kliegel et al. 2000).

Storage refers to maintaining information in memory during the retention interval. It would be adaptive if the storage of prospective memory were somewhat special so that the intended action would be remembered at the time it needed to be performed. Evidence to date suggests there may indeed be several ways in which prospective memory storage is different from retrospective memory storage. First, it may be that prospective memories (intended actions) enjoy greater and more persistent activation in memory (Goschke and Kuhl 1993, Marsh et al. 1998a).

Theoretically, memories in a more activated state more easily come to mind when one comes across cues in the environment that are related to those memories. It would also be adaptive if these memories for intentions would be deactivated after the activity was performed, so that they would not come to mind later. Preliminary findings suggest that such deactivation may occur (Marsh et al. 1998a), although for older adults under demanding conditions it appears that the memory for the intention may persist so that they inappropriately repeat their performance of the intended action (Einstein et al. 1998).

Another special storage feature of prospective memory suggested in one study is that prospective mem-

ories may actually become stronger as the retention (storage) time lengthens (Hicks et al. 2000). This is in dramatic contrast to the well-known finding that retrospective memory often declines with increased retention intervals. Additional research is needed to explain this provocative finding that prospective memories become better as the retention time increases. One clear possibility is that people remind themselves about the intended activity during the retention interval, especially as the duration increases between the formation of the initial intention and the point at which the intention can be performed. Reminders would presumably refresh the memory of the intended activity, thereby increasing the memorability of the intended action and its likelihood of being successfully performed at the appropriate time (Guynn et al. 1998).

Retrieval refers to accessing information in long-term memory so that it can be brought into awareness and used in a conscious fashion. As noted earlier in this article, a critical feature of prospective remembering is that retrieval must be initiated by the individual performing the prospective memory task. By contrast, retrieval in a retrospective memory task is initiated by somebody else (e.g., an experimenter) who prompts the individual to retrieve the target information. Accordingly, retrieval processes in prospective memory are thought to be somewhat different from those that are characteristic of retrospective memory retrieval.

Cognitive psychologists are not in agreement about how prospective memory retrieval operates. Consistent with the above analysis, one view proposes that prospective memory tasks are very high in self-initiated retrieval (Craik 1986) and therefore require a high level of cognitive resources to support retrieval. Another view proposes that initiating an intended action is not so much a memory process as it is an attentional process in which the person monitors the environment for signals that indicate that it is appropriate to perform the action (e.g., Burgess and Shallice 1997). By these views, prospective memory tasks should be especially difficult for older adults because of declines in cognitive or attentional resources that accompany aging. In support of this expectation, age-related decrements have been reported in time-based prospective memory tasks (Einstein et al. 1995, Park et al. 1997), tasks for which the intended action must be performed after a certain period of time has elapsed.

However, in some studies, older and younger subjects have surprisingly demonstrated equivalent prospective memory performance in event-based prospective memory tasks (Einstein and McDaniel 1990, Einstein et al. 1995). In an event-based task an external event signals that it is appropriate to perform the intended action. This event (or cue) might be the appearance of a friend to whom you have to give a message or a particular word in an experiment to

which you have to give a special response. The absence of age differences in event-based prospective memory suggests that prospective memory retrieval can be a relatively automatic process. One idea is that the environmental cue may automatically stimulate the retrieval from memory of the intended action, after which the action can be performed (McDaniel et al. 1998).

More research is needed to understand completely prospective memory retrieval. Based on the literature it seems likely that prospective memory tasks may vary such that those with very few environmental cues (such as time-based tasks) may rely heavily on self-initiated retrieval and attentional processes, whereas prospective memory tasks with salient environmental cues may be supported by more automatic retrieval mechanisms. In addition, future work in prospective memory will no doubt increasingly focus on the planning processes that are involved in prospective memory (e.g., Dobbs and Reeves 1996, Burgess and Shallice 1997, Kliegel et al. 2000) and the potential special storage dynamics of prospective memories.

See also: Citizenship: Political; Elaboration in Memory; Incidental versus Intentional Memory; Memory Retrieval; Recognition Memory, Psychology of

Bibliography

- Burgess P W, Shallice T 1997 The relationship between prospective and retrospective memory: neuropsychological evidence. In: Conway M A (ed.) *Cognitive Models of Memory*. 1st MIT Press edn., MIT Press, Cambridge, MA, pp. 247–72
- Craik F I M 1986 A functional account of age differences in memory. In: Klix F, Hagendorf H (eds.) *Human Memory and Cognitive Capabilities: Mechanisms and Performances*. Elsevier, Amsterdam, pp. 909–22
- Dobbs A R, Reeves M B 1996 Prospective memory: more than memory. In: Brandimonte M, Einstein G O, McDaniel M A (eds.) *Prospective Memory: Theory and Applications*. L. Erlbaum, Mahwah, NJ, pp. 199–225
- Einstein G O, McDaniel M A 1990 Normal aging and prospective memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition* **16**: 717–26
- Einstein G O, McDaniel M A, Smith R E, Shaw P 1998 Habitual prospective memory and aging: remembering intentions and forgetting actions. *Psychological Science* **9**: 284–8
- Einstein G O, Richardson S L, Guynn M J, Cunfer A R 1995 Aging and prospective memory: examining the influences of self-initiated retrieval processes. *Journal of Experimental Psychology: Learning, Memory, and Cognition* **21**: 996–1007
- Ellis J A 1996 Prospective memory or the realization of delayed intentions: a conceptual framework for research. In: Brandimonte M, Einstein G O, McDaniel M A (eds.) *Prospective Memory: Theory and Applications*. L. Erlbaum, Mahwah, NJ, pp. 1–22
- Goschke T, Kuhl J 1993 Representation of intentions: persisting activation in memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition* **19**: 1211–26
- Hicks J L, Marsh R L, Russell E J 2000 The properties of

- retention intervals and their affect on retaining prospective memories. *Journal of Experimental Psychology: Learning, Memory and Cognition* **26**: 1160–9
- Kliegel M, McDaniel M A, Einstein G O 2000 Plan formation, retention, and execution in prospective memory: a new approach and age-related effects. *Memory and Cognition* **28**: 1041–49
- Koriat A, Ben-Zur H, Nussbaum A 1990 Encoding information for future actions: memory for to-be-performed tasks versus memory to to-be-recalled tasks. *Memory and Cognition* **18**: 568–78
- Kvavilashvili L, Ellis J 1996 Varieties of intention: some distinctions and classifications. In: Brandimonte M, Einstein G O, McDaniel M A (eds.) *Prospective Memory: Theory and Applications*. Erlbaum, Mahwah, NJ, pp. 23–51
- Marsh R L, Hicks J L, Bink M L 1998a Activation of completed, uncompleted, and partially completed intentions. *Journal of Experimental Psychology: Learning, Memory, and Cognition* **24**: 350–61
- Marsh R, Hicks J L, Landau J D 1998b An investigation of everyday prospective memory. *Memory and Cognition* **26**: 633–43
- McDaniel M A, Robinson-Riegler B, Einstein G O 1998 Prospective remembering: perceptually driven or conceptually driven processes? *Memory and Cognition* **26**: 121–34
- Meacham J, Singer J 1977 Incentive effects in prospective remembering. *Journal of Psychology* **97**: 191–7
- Moscovitch M 1982 A neuropsychological approach to memory and perception in normal and pathological aging. In: Craik F I M, Treub S (eds.) *Aging and Cognitive Processes*. Plenum, New York, pp. 55–78
- Park D C, Hertzog C, Kidder D P, Morell R W, Mayhorn C B 1997 Effect of age on event-based and time-based prospective memory. *Psychology and Aging* **12**: 314–27

M. A. McDaniel and G. O. Einstein

Prostitution

Prostitution is simultaneously a sexual and an economic institution, and it is also highly gendered. The majority of prostitutes are female, and an even larger majority of customers are male. While men's prostitute use is widely tolerated, female prostitution is popularly viewed as a form of social and sexual deviance, and mainstream social scientists have traditionally reproduced such attitudes in their research on the topic. Feminist theorists, by contrast, have long been concerned to explore parallels between marriage, prostitution, slavery, and wage labor, as well as the sexual, political, and economic relations that underpin these institutions. They do not speak as one on the subject, however, and the division between feminists who are concerned with prostitution as a sexual institution and feminists who approach it first and foremost as a form of economic activity is particularly sharp. This article provides a brief overview of prostitution in the contemporary world and highlights the theoretical and actual problems it poses for gender scholarship.

1. Prostitution: A Stigmatized and Criminalized Activity

A profoundly negative stigma is almost universally attached to prostitute women. Religious thinking on men's prostitute use varies, but there is no major world religion that actively sanctions female prostitution and, in secular societies, 'scientific' thinking has done little to displace traditional attitudes. Mainstream medical, psychological, psychoanalytic, and sociological research on the topic has generally assumed that while men's prostitute use is based on natural, biologically determined sexual drives, women who prostitute are somehow abnormal, unnatural, a threat to public health and order. Prostitution law varies from country to country and even within individual nation states, but typically enshrines this kind of stigma by treating female prostitutes as a distinct class of persons, separate from other workers and/or women in terms of their rights to protection, privacy, and/or self-determination (Walkowitz 1980, Bindman 1997).

Law enforcement almost invariably focuses upon female prostitution, ignoring men's prostitute use (often also male prostitution), and is generally informed by one of two basic models: prohibition/abolition, or regulation/registration. The prohibitionist model has historically led to an emphasis on controlling public manifestations of prostitution through the criminalization of (usually street working) prostitute women. The regulation/registration model, which legalizes prostitution providing it takes place in licensed brothels or designated geographical zones, has been associated with a variety of civil rights violations, such as requirements for prostitute women to register with the police and/or other authorities, have compulsory health checks, and so on. In some countries, there have been moves to deregulate prostitution either explicitly or by default, a shift which is taking place largely for reasons of 'financial exigency, prosecutor indifference, court deadlocks, and resistance to the overreach of the criminal law among some social sectors' (Davis 1993, p. 8).

2. The Economic Significance of Prostitution

Because female prostitution is ideologically and legally construed as a form of deviance, states do not officially recognize or regulate prostitution as an economic sector like any other. It is therefore extremely difficult to obtain accurate data on the size and earnings of the sex trade. However, research indicates that sex commerce is a significant feature of economic life in many nations, regardless of their overall level of economic development. Studies suggest that, for example, in London, prostitution generates around £200 million annually (Matthews 1997); that in New Zealand, almost one in every 150 women aged between 18 and 40 is employed in some form of sex work (Lichtenstein 1999); and that in Thailand, the Philippines, Malaysia,

and Indonesia, the sex sector accounts for between 2 percent and 14 percent of gross domestic product (Lim 1998). Research further suggests that a fairly consistent set of factors underpin prostitution labor markets. Gender discriminatory social practices are structural 'push' factors in economically developed and underdeveloped nations alike (Davis 1993).

As part of an unofficial and often criminalized economic sector, prostitution is frequently connected with corruption, organized crime, and drug abuse, but it is often simultaneously integrated into mainstream economic structures, such as the entertainment and the tourism industries (Lim 1998). This draws attention to the increasingly international nature of the sex industry.

3. *Global Politics and Markets and Prostitution*

Prostitution is affected by global politics and market trends in three main ways. First, there are countries in which the development of the sex sector has been directly shaped by the military and economic interests of external powers. Thailand, for example, was used for 'Rest and Recreation' by the US military in the 1950s, 1960s, and 1970s, and a highly sexualized entertainment sector developed to meet demand from legions of young men. Western financial institutions and businesses soon recognized that such an entertainment industry was also suited to serve one particular segment of the long haul tourist market, and economic initiatives based on a 1975 World Bank report 'led to what is routinely described today as a \$4-billion-a-year business involving fraternal relationships among airlines, tour operators, and the masters of the sex industry' (Bishop and Robinson 1998, p. 9). Similar developments have occurred elsewhere in South East Asia, as a result of either US or Japanese military or economic interventions in the region.

Sex commerce in poorer countries has also been affected by a set of linkages between international debt, price fluctuations in global commodity markets, economic development policy, and prostitution. Since the 1970s, world financial institutions have encouraged indebted nations to respond to economic crisis by developing tourism and/or 'nontraditional' export industries such as gold, diamonds, and timber. One side effect of such development policies is the creation of highly concentrated, effective demand for prostitution: affluent tourists seeking 'entertainment' and predominantly male, migrant workers in isolated mining and logging regions with cash to spend on 'recreation.' Meanwhile, structural adjustment measures have expanded the prostitution labor market. It is women and children who have been most adversely affected by the sudden introduction of cash economies into what were formerly subsistence economies, currency devaluations and concomitant falls in the price of labor, and the redirection of subsidies away from social spending and basic commodities towards debt

servicing. In such climates, prostitution has often become the best or only economic alternative for large numbers of women and teenagers (Chant and McIlwaine 1995, Clift and Carter 1999, Kempadoo 1999). Third, the sex trade is international in the sense that many of those working in prostitution have either voluntarily migrated in order to prostitute or have been trafficked by third parties for that purpose. Migrant prostitutes are especially vulnerable to abuse and exploitation by third parties and, as both illegal migrants and prostitute women, particularly likely to be denied legal protection and civil rights (Bindman 1997).

Ideological, as well as economic and political factors underpin the globalization of sex commerce. Racist discourses, which typically construct the 'Other' as sexually debased and/or exotic, have long fueled desire for 'racially' Other prostitutes and continue to shape demand today (Shrage 1994).

4. *The Diversity of Prostitution*

Prostitution varies enormously in terms of its social organization and the power relations it involves. In virtually every country of the world, hierarchies exist within both female and male prostitution (West 1992, Aggleton 1999). At their apex are independent, self-employed adult prostitutes who exercise a relatively high level of control in transactions with clients and whose earnings are relatively high. At their base are individuals who gain little or no financial reward from their prostitution and exercise little, if any, control over it (including control over condom use). This may be because they are physically forced into a given work regime by a third party pimp or brothel-keeper, but it may also be because extreme economic and social marginalization, and/or youth and/or drug addiction renders them horribly powerless within transactions.

Between the two extremes are prostitutes who either work independently or enter into some form of employment relation with a third party. The degree of control they exercise over whether, when, how often, and on what terms they prostitute varies according to a range of factors, including their level of economic desperation; the contractual form of the prostitute-client exchanges they enter into; and the specific legal, institutional, social, political, and ideological context in which they prostitute. There are, for example, settings in which prostitution law so heavily penalizes independent prostitution that it effectively operates as a pressure on prostitutes to enter and remain in third party controlled prostitution no matter how exploitative the third party may be. Equally, in contexts where certain social groups (for instance, women, children, homosexuals, migrants, particular 'castes' or racialized minorities) are generally devalued, and/or denied full juridical subjecthood, and/or denied independent access to welfare support, their

vulnerability to prostitution and to third party exploitation and abuse within it is increased. It should also be noted that people come to prostitution as individuals of a certain age and with particular personal histories and experience, and this leaves some more open to exploitation and at higher risk of violence and sexually transmitted disease than others (O'Connell Davidson 1998).

5. *Conceptualizing Prostitution*

One approach to the task of conceptualizing prostitution is to examine how it simultaneously resembles and differs from other social institutions. In prostitution, as in the institution of marriage, the sexual and the economic are conjoined. Unlike marriage, however, prostitution does not construct long term, mutual, or diffuse relations of dependency between individuals and their kin. Instead, sexual interaction between prostitute and client takes place on the understanding that some immediate economic benefit will accrue to either the prostitute or a third party as a result, and the client's obligations are thus discharged by payment in cash or kind. To this extent, relations within prostitution resemble market relations.

However, prostitution does not involve any ordinary commodity exchange. The client does not wish to purchase the disembodied means to an orgasm, a 'thing' that can be detached from the prostitute's person. Rather clients part with money and/or other material benefits in order to secure certain powers of sexual command over the person of the prostitute. Thus, although prostitution is often thought of as the exchange of *sex* or *sexual services* for money and/or other benefits, it is better conceptualized as an institution which allows certain powers over one person's body to be exercised by another. In this sense, prostitution has something in common with slavery, wage labor, and relations between buyers and self-employed sellers of services. All are institutions through which certain powers of command over the person are transferred from one individual to another, conferring a right upon the client/slave holder/employer/buyer to require that the prostitute/slave/worker/seller acts in ways that she/he would not otherwise choose to act.

Yet the relationship between prostitute and client differs from that between slave and slave holder or worker and employer in crucial respects. Not only is it generally brief and transitory, but also, clients enter into it as consumers, not producers. Clients do not depend upon prostitutes for their economic well being in the way that slave-holders and employers depend upon slaves and workers for their material existence. And even where the prostitute is self-employed, the prostitute-client exchange is made unlike other exchanges between independent sellers and purchasers of

services by the social meanings that are popularly attached to sexuality.

In monetary economies, sexual and economic relations are typically imagined, explained, and justified in very different ways. The products of human labor, human labor power, sometimes even human beings themselves, have been and are imagined as commodities to be bought and sold on the basis of rational economic calculation and/or in pursuit of status and social prestige. Human sexual interaction, by contrast, is generally regulated and given meaning through reference to premarket or nonmarket ideas, such as those pertaining to honor, shame and duty, and/or romantic love, and/or recreation, pleasure, and desire. Prostitution thus occupies a troubled and troubling space between two quite different symbolic domains. It does not readily fit into popularly understood categories of 'sex' or 'work,' and this tension has been central to much feminist debate on the subject.

6. *Feminist Debates on Prostitution*

Though sharing a common concern with the welfare of prostitute women, feminists are often deeply divided on the question of whether to view prostitution as a sexual or an economic institution. For 'radical' feminists, who foreground the sexual domination of women by men in their analyses of women's political subordination, prostitution is the unambiguous embodiment of male oppression. In this view, as a sexual institution, prostitution reduces women to bought objects, allows men temporary, but direct, control over the prostitute, and increases their existing social control over all women by affirming their masculinity and patriarchal rights of access to women's bodies. For such commentators, prostitution is always and necessarily damaging to women, there can be no distinction between 'forced' and 'free choice' prostitution, and in tolerating, regulating, or legalizing prostitution, states permit the repeated violation of human rights to dignity and sexual autonomy (Barry 1995, Jeffreys 1997).

Feminists who adopt a sex workers' rights perspective generally reject the idea that prostitution is intrinsically or essentially degrading, and view links between prostitution and patriarchal domination as contingent, rather than necessary. Treating prostitution as an economic institution, they make a strong distinction between 'free choice' prostitution by adults and all forms of forced and child prostitution. While they believe the latter should be outlawed, they hold the former to be a type of paid work, a job like any other. Since sex workers' rights feminists view free choice prostitution as a mutual voluntary exchange, they see state actions which criminalize or otherwise penalize those adults who make an individual choice to enter prostitution as a denial of human rights to self-determination (see Nagel 1997).

Because hierarchies exist within prostitution in terms of earnings, working conditions, vulnerability to violence and exploitation, and autonomy as regards specifying, limiting, and retracting from contracts with clients, these two very different interpretations of prostitution can each be partially supported empirically. Radical feminists point to research which indicates that around the world, the average age of entry into prostitution is very low (often under 18 years); that it is often precipitated by the experience of rape, sexual assault, and/or incest; and that the experience of prostitution can be extremely damaging psychologically, as well as associated with drug and solvent abuse, various forms of sexual and physical violence, and suicide (Hoigard and Finstad 1992, Kelly et al. 1995). Radical feminism's emphasis on a link between prostitution and systems of patriarchal domination is supported by the strong historical and contemporary association between the military and organized prostitution, while the persistence of highly abusive and slavery-like practices within prostitution and the significant presence of children aged between 12 and 18 in mainstream prostitution, even where prostitution is legally regulated rather than proscribed, lends further credence to a view of prostitution as a hugely problematic institution (Ennew 1986, Enloe 1989).

Sex workers' rights feminists point out that slavery-like practices and child labor persist in other industries, and that historically, such abuses have been most successfully addressed through collective struggle for workers' political representation, international conventions guaranteeing minimum rights to various groups of workers/persons, national employment laws outlawing abusive practices, and so on. They cite research which suggests that prohibitionist legislation increases, rather than reduces, prostitutes' vulnerability to violence, third party coercion, and extortion, and emphasize empirical studies in which adult women and men in prostitution describe themselves as perfectly content with their chosen occupation (Bindman 1997, Kempadoo and Doezema 1998).

Though certain feminist commentators remain rigidly wedded to a view of prostitution as either work or sexual violence, many theorists and activists (including some feminist abolitionists and sex workers' rights feminists) place a growing emphasis on the need to develop analyses of prostitution which can embrace its diversity and its particularity as both a sexual and an economic institution (see, for example, Shrager 1994, Chapkis 1997, O'Connell Davidson 1998).

See also: Feminist Theory: Radical Lesbian; Heterosexism and Homophobia; Male Dominance; Poverty and Gender in Affluent Nations; Poverty and Gender in Developing Nations; Rape and Sexual Coercion; Sexual Harassment: Social and Psychological Issues; Sexuality and Gender; Transsexuality, Transvestism, and Transgender

Bibliography

- Aggleton P (ed.) 1999 *Men Who Sell Sex*. UCL, London
- Barry K 1995 *The Prostitution of Sexuality*. New York University Press, New York
- Bindman J 1997 *Redefining Prostitution as Sex Work on the International Agenda*. Anti-slavery International, London
- Bishop R, Robinson L 1998 *Night Market: Sexual Cultures and the Thai Economic Miracle*. Routledge, London
- Clift S, Carter S 1999 *Tourism and Sex: Culture, Commerce and Coercion*. Pinter, London
- Chant S, McIlwaine C 1995 *Women of a Lesser Cost: Female Labour, Foreign Exchange & Philippine Development*. Pluto, London
- Chapkin W 1997 *Live Sex Acts*. Cassell, London
- Davis N (ed.) 1993 *Prostitution: An International Handbook on Trends, Problems and Policies*. Greenwood, Westport, CT
- Enloe C 1989 *Bananas, Beaches, Bases*. Pandora, London
- Ennew J 1986 *The Sexual Exploitation of Children*. Polity Press, Cambridge, UK
- Hoigard C, Finstad L 1992 *Backstreets: Prostitution, Money and Love*. Polity, Cambridge, UK
- Jeffreys S 1997 *The Idea of Prostitution*. Spinifex, Melbourne, Australia
- Kelly L, Wingfield R, Burton S, Regan L 1995 *Splintered Lives: Sexual Exploitation of Children in the Context of Children's Rights and Child Protection*. Barnardo's, Ilford, UK
- Kempadoo K (ed.) 1999 *Sun, Sex and Gold: Tourism and Sex Work in the Caribbean*. Rowman and Littlefield, Boulder, CO
- Kempadoo K, Doezema J (eds.) 1998 *Global Sex Workers: Rights, Resistance and Redefinition*. Routledge, New York
- Lichtenstein B 1999 Reframing 'Eve' in the AIDS era: The pursuit of legitimacy by New Zealand sex workers. *Sexuality and Culture*. 2: 37-60
- Lim L 1998 *The Sex Sector: The Economic and Social Bases of Prostitution in Southeast Asia*. International Labor Organization, Geneva, Switzerland
- Matthews R 1997 *Prostitution in London: An Audit*. Middlesex University, London
- Nagel J (ed.) 1997 *Whores and Other Feminists*. Routledge, London
- O'Connell Davidson J 1998 *Prostitution, Power and Freedom*. Polity Press, Cambridge, UK
- Shrage L 1994 *Moral Dilemmas of Feminism*. Routledge, London
- Walkowitz J 1980 *Prostitution and Victorian Society: Women, Class and the State*. Cambridge University Press, Cambridge, UK
- West D 1992 *Male Prostitution*. Duckworth, London

J. O'Connell Davidson

Protein Synthesis and Memory

A role for proteins in memory formation was invoked by early neuroscientists in their search for the 'engram', an envisioned physical manifestation of stored memory. It was recognized that measured electrical activities were too transient to endure for a

lifetime, and chemically stable macromolecules such as proteins seemed more likely candidates, perhaps as metabolically stable bricks that comprised the structures of altered synapses (Gerard 1953). It eventually became evident that proteins are constantly being produced and broken down, necessitating more complex models of the molecular basis for memory storage than a simple accretion of structural molecules.

The nature of protein synthesis in the brain became of particular interest for behavioral scientists as a result of demonstrations in a number of vertebrate and invertebrate species that chemical agents that inhibit protein synthesis in the brain block the formation of long-term memory (LTM) of a variety of training tasks, yet do not block acquisition of the task (Agranoff et al. 1999). More recently, analogous effects of protein synthesis inhibition have been reported in a reductionist model of learning and memory, termed long-term potentiation (LTP) (see *Long-term Potentiation (Hippocampus)*).

For the purposes of this article, learning and short-term memory (STM) formation are defined by experimental measures of improved performance of an adaptive response observed during or shortly after a training session (acquisition), while LTM formation is evidenced by retention of the learned response in a subsequent training session, typically administered some days or weeks later. An experimental advantage of using an interventive agent such as a protein synthesis blocker in studying learning and memory is at the same time its great disadvantage: while it can establish the timing of an inferred molecular mechanism taking place in an otherwise intact animal, it does not specify cells or even brain regions critical for LTM formation, nor does it identify specific proteins on which memory formation is presumed to be dependent. Such proteins could hypothetically be identified by complementary correlative studies, for example, by examining the nature of newly synthesized proteins on the basis of their incorporation of radio-labeled precursor amino acids during the period of LTM formation. Together, the interventive and correlative approaches have proven useful in establishing the molecular bases of other complex cellular processes.

1. Proteins Defined

Proteins are biological macromolecules consisting of a linear array of 20 component L-amino acids. The amino acids are joined in chemically stable peptide linkages that are not easily disrupted by thermal, chemical or mechanical stresses encountered by living cells. A protein typically consists of several hundred amino acids in a precise sequence (termed its primary structure), much like a very long word written in a 20-letter alphabet. The secondary and tertiary structures

of a protein refer to how the polypeptide chain is folded into helices, ribbons and sheets. This overall conformation of proteins is largely dictated by the chemical properties of the individual amino acids in its primary structure.

Proteins play both structural and functional roles in cells. Membranes, consisting of lipid bilayers in which proteins are embedded, define a neuron and its intracellular organelles, such as the nucleus, mitochondria, synaptic vesicles, and endoplasmic reticulum. Proteins may be primarily structural, adding integrity to the membrane, or serving as a bridge to other membranes or to one of a group of fibrillar proteins including tubulin, actin and neurofilament proteins, known collectively as the cytoskeleton. Functional roles of proteins are exemplified by the enzymes, which are generally soluble proteins, present in a globular conformation. Hundreds of enzymes have been characterized and shown each to be a chemical catalyst that accelerates the rate of a specific biochemical reaction, without itself being consumed. Structural and functional proteins together catalyze and otherwise mediate the business of the cell (Voet and Voet 1995), expressing its genetic phenotype, including its appearance, its metabolism and its repertoire of responses. Knowledge of the primary structure of a protein is useful in determining its relation to other known proteins, in some instances yielding information about its evolutionary lineage as well as providing insight into its function. Short amino acid sequences within proteins (domains) identify probable catalytic active sites, the number of times a protein chain spans a membrane and the probable site of a protein's binding to other known proteins. There are also short amino acid sequences that identify the nature and likely sites of 'decoration' of proteins with small chemical groups following their synthesis. In addition to acting as catalysts, proteins can mediate recognition of a variety of small molecules that are intracellular chemical messengers. Membrane proteins on the cell surface may also act as receptors that recognize a variety of small molecules, including extracellular neuromodulators, hormones, neurotransmitters and environmental chemical signals (odorants) (see *Neurotransmitters*; Siegel et al. 1999). Proteins can recognize other proteins, even xenobiotic (novel) ones, as in the immune response. While proteins are clearly the arbiters of all cellular activities, not all cellular actions depend on ongoing protein synthesis, as will be seen.

1.1 Proteins are Synthesized on Ribosomes from Messenger RNA

The primary structure of a protein ultimately is determined genetically by linearly coded triplet deoxy-nucleotides in DNA in the cell nucleus, which are

transcribed to a complementary sequence of nucleotide triplets in the messenger RNA (mRNA) for the protein. The newly synthesized mRNA is *translated* ('decoded') into the new protein on polyribosomes in a cellular region near the nucleus, the RNA triplets establishing the primary structure of the new protein. The entire process, beginning with the availability of DNA template for transcription and ending with protein synthesis, generally takes minutes to hours (Voet and Voet 1995). Antibiotic inhibitors of protein synthesis such as puromycin and acetoxycycloheximide are structurally unrelated, and block translation by different molecular mechanisms, yet each has been shown to block LTM formation, confirming that their behavioral effects are indeed the result of their inhibition of protein synthesis rather than of some unknown side effects of the agents. Inhibition of RNA synthesis has also been found to block LTM formation.

Upon release from the polysomes, newly synthesized proteins may be subjected to *post-translational modification*, typically the addition of stable, small chemical substituents, such as fatty acid, isoprenoid, carbohydrate, and phosphate or methyl groups. These modifications are catalyzed by specific enzymes, and may also require cofactors such as calcium ions or cosubstrates such as ATP. Phosphorylation of a protein by ATP requires a protein kinase, although some proteins can autophosphorylate themselves. As a consequence of such modifications, a protein may be tagged with a chemical 'zip code' that designates its intracellular destination. For example, a lipid substituent will target a protein to a membrane site, while phosphorylation could direct a protein to a nuclear site. The overall conformation of proteins can be affected by these post-translational changes. A newly formed protein may be inactive until it is cleaved by a specific protein-splitting enzyme that removes a built-in inhibitory sequence.

Many proteins are part of intracellular messenger systems that act sequentially to produce multiplicative cascades, resulting ultimately in a cellular response, such as hyperpolarization, depolarization, contraction, or secretion. Unlike the ribosomal synthesis of proteins, in which the translation of proteins is measured in tens of minutes, post-translational modification can occur rapidly, even in milliseconds, and unlike typical protein synthesis, is not restricted to the perinuclear region of the cell. Post-translational modification of proteins can occur at any time in the life of a protein, and can even target proteins for degradation.

1.2 Brain Proteins are in a Dynamic State of Formation and Degradation

The thousands of protein species present in a brain cell are in a metabolically dynamic state, i.e., they are

continuously being synthesized by ribosomes and are then broken down into their component amino acids by the action of enzymes called proteases. The released amino acids may be metabolized or can be reincorporated into new proteins. Turnover studies employing radiolabeled amino acid precursors show that a given cellular protein in a specific cell type is turning over with a characteristic half-life of minutes to weeks. Each protein is present at a characteristic steady state level, i.e., the rate of synthesis of the protein is equal to its rate of degradation. Even though most neurons are formed early in life and remain in a postmitotic state through adulthood, their component structures are nevertheless in a dynamic metabolic equilibrium at a molecular level. A neuron may thus maintain an unchanging microscopic appearance throughout life, yet, with the exception of DNA, every component molecule, including all of its neuronal proteins, is constantly being replaced. There is considerable evidence that the cell's morphology and other cellular modifications of long duration are maintained by continual feedback to nuclear DNA, signaling initiation of transcription of mRNAs for constitutive proteins as they are depleted, or the induction of novel proteins, as occurs in cellular differentiation during development and in nerve regeneration. Thus, changes in synaptic connections might be similarly maintained by such transcriptional feedback loops in LTM storage. Translational and post-translational feedback loops that mediate memory formation have been proposed (Lisman and Fallon 1999).

1.3 Temporal Considerations in Establishing the Role of Protein Synthesis in Memory Formation

Since acquisition of a new behavior can occur within seconds, it is immediately apparent that it cannot be either directly or indirectly mediated by the relatively slow process of *de novo* protein synthesis. Acquisition could, however, involve a sustained post-translational modification at the synapse, which can occur within milliseconds and at any cellular locus. As noted, the macromolecular process of protein synthesis requires initiation of transcription in the nucleus followed by ribosomal translation, and takes minutes to hours, a time frame compatible with LTM, but not with STM formation. Additional time may be required for a putative chemical message from the presumed site of STM formation in the synaptic region (see *Memory: Synaptic Mechanisms*) to be transported to the nucleus in the neuronal cell body and to initiate transcription and translation. Still more time may be required for newly synthesized proteins to be transported from the cell body to a synaptic destination, where they could mediate long-term neuroplastic changes.

Rates of protein flow away from the cell body (anterograde axonal transport) and from the dendrites and axons toward the cell body (retrograde axonal

transport) were established in neurons with long processes. The amount of time required for trafficking of a chemical signal from a synapse to the nucleus and of a newly synthesized protein back to the synapse hypothetically could be extrapolated from known rates of axonal transport, together with knowledge of the morphology of the neuroplastic neurons hypothesized to be mediating the process. Ribosomal protein synthesis at a distance from the nucleus (Steward 1997) could shorten the calculated transport times. As will be addressed in the following section, the advent of antibiotic agents that selectively block protein synthesis offered the opportunity to study these complex issues experimentally.

2. Agents that Inhibit Protein Synthesis Block Formation of Long-term Memory

In the early 1960's, the excitement over the identification of the genetic code spurred renewed speculation and experimentation directed at the possible molecular nature of the engram. In 1962 Flexner and co-workers reported that intracerebral injection of the antibiotic puromycin, an aminonucleoside that prematurely cleaves the growing peptide chain from ribosomes, blocked brain protein synthesis in the mouse brain and could also block formation of memory of a Y-maze shock avoidance task (Flexner et al. 1962). This was followed by studies in the goldfish in 1964 (reviewed in Agranoff et al. 1999), and subsequently in a variety of other vertebrates and invertebrates.

2.1 Effects of Protein Synthesis Blockers on LTM Formation in Experimental Animals

Injection of puromycin into the cranial fluid (intracranially; IC) surrounding the goldfish brain produced a profound block in brain protein synthesis for several hours. Studies on acquisition (STM) and on LTM formation of a light-coupled-to-shock shuttlebox task revealed that a block in brain protein synthesis had no effect on acquisition of avoidance behavior during an initial multitrial training session, but eliminated retention of the improved performance that would otherwise have been seen in a second session administered a week later. The depth and duration of the inhibition of brain protein synthesis and the effectiveness of the block of LTM formation correlated well. Injection of the blocker into the fish body cavity rather than IC was without effect on LTM formation. IC injection of acetoxycycloheximide, another inhibitor of protein synthesis, was equally effective in blocking LTM formation.

The behavioral effects of inhibition of protein synthesis differ from that of a number of neurodisruptive agents and treatments, such as electroconvulsive

shock, in that animals exhibit normal behavior during the action of the interventive agent. This permitted an exploration of STM decay. It was known that fish injected IC with puromycin before the first training session would show increasing avoidance scores, comparable to those seen in uninjected control fish; yet in a second training session a week later, such fish performed at a naïve level, i.e., had formed no LTM of the first session. Groups of fish trained under conditions of blocked protein synthesis (and that nevertheless demonstrated normal acquisition, as noted above) were then retrained at various times following the initial session. It was found that performance gradually dropped to the level of naïve fish over a 2–3 day period, providing evidence for the decay of STM for this task. Block of protein synthesis immediately after training also blocked LTM formation, whereas if animals were returned to their home tanks and injected with puromycin a few hours later, there was no detectable effect on LTM. This period of LTM susceptibility was comparable in its duration to that seen with electroconvulsive shock and other disruptive agents administered after training. When fish were detained in the training apparatus following the first training session, the consolidation time was prolonged, suggesting that the onset of LTM formation required an 'environmental trigger.' This phenomenon, in which continued exposure of an experimental subject to its training environment delays the onset of LTM formation, has also been termed 'detention.'

Behavioral studies with numerous other species, including young chicks and rats (for review, see Davis and Squire 1984), support the general hypothesis that LTM formation requires ongoing protein synthesis, while STM formation does not. Among vertebrates, studies in *Hermessenda* and in *Aplysia* (Bailey et al. 1992) support this hypothesis (see Agranoff et al. 1999). *Drosophila* have proven useful in behavioral studies, with genetic mutants in which selected proteins are mutated, deleted or overexpressed, implicating various neurotransmitter-related and second messenger system proteins in memory formation (see Agranoff et al. 1999). Effects of protein blockers on human memory have not been reported, but the possibility did not escape the imagination of a novelist (MacDonald 1968).

2.2 Protein Synthesis Inhibition in Long-term Potentiation, a Reductionist Model of Learning and Memory

Attempts to identify specific proteins, the synthesis of which are crucial to LTM formation in the brain, have been beset with many experimental problems, not the least of which is knowledge of the precise anatomical sites of the presumed altered molecular events underlying an observed altered behavioral performance

interpreted as LTM formation. LTP (see *Long-term Potentiation (Hippocampus)*) can be induced and studied *ex vivo* in localized tissue preparations such as brain hippocampal slices, and has been proposed as a cellular model of LTM formation. Blockers of protein synthesis have been shown to prevent LTP formation, a result taken by some as its validation as a model of LTM formation. Numerous pharmacological, biochemical, molecular, and genetic studies have been performed to implicate one or another structural protein, enzyme, receptor, or intracellular messenger pathway in the development of LTP, as a result of which over 100 known molecules have been implicated in LTP formation, including intracellular messengers, vesicular trafficking proteins, proteases, kinases, etc. (Sanes and Lichtman 1999). Eventually it may prove to be the case that the ongoing synthesis of one or only a few of these proteins is indeed necessary and sufficient for LTP formation, and can account for the observed block of LTM formation by protein synthesis inhibitors. It would seem more likely that the synthesis of many such proteins is necessary for the complex process of LTM formation to proceed. This embarrassment of riches in candidate compounds continues to grow as additional previously undiscovered cellular proteins become implicated in neuroplasticity, in that their selective deletion or inactivation prevents LTP or LTM formation. Further information regarding the temporal sequence of the observed molecular events accompanying LTP may prove useful in establishing whether they are indeed causal or epiphenomenal in its formation.

2.3 Protein Nutrition and Mental Function

Half of the 20 amino acids of which proteins are composed can be synthesized in the human body, while the remaining 10, termed 'essential amino acids' (EAAs), must come from the diet, principally in the form of ingested proteins that are degraded to their component amino acids in the gut and absorbed into the bloodstream. Eventually some will cross the blood/brain barrier, which has efficient EAA uptake systems. Also, as the brain's own proteins are broken down, the released amino acids are conserved and become a rich source of EAAs for new brain protein. Thus, under conditions of dietary EAA deficiency in adults, as occurs in starvation, brain protein metabolism is relatively protected. This is not true for the developing brain in infancy and early childhood, periods during which the brain must grow. For this, the brain relies on an exogenous source of EAAs, without which permanent cognitive deficits can ensue, as in kwashiorkor, a protein deficiency disease of children encountered in impoverished tropical countries. A blood amino acid imbalance is seen in the genetic defect phenylketonuria in which there is a high

blood level of the amino acid phenylalanine, while blood levels of the other EAAs are normal. This imbalance overloads and blocks the amino acid transport system at the blood/brain barrier and effectively prevents entry of the other EAAs, eventually resulting in mental retardation. If infants with this genetic defect are raised on a diet deficient in phenylalanine, the brain develops normally, and as they mature, the toxic effect of dietary phenylalanine is much reduced.

3. Summary

It is apparent from the foregoing that in a number of animal species, disruption of protein synthesis in the brain for a period of hours has no effect on acquisition of a behavioral task, but blocks the formation of LTM. Since all brain proteins turn over metabolically, it is presumed that positive feedback loops are induced during LTM consolidation that involve transcription and/or translation of new protein, but how and where these loops function and are distributed in the complex structure of the brain is not yet known. Selective inhibition or elimination of various proteins by genetic means has been shown to block memory formation, but the salience of each new finding is dampened by the increasing number of identified genetic defects or deletions that can be shown to lead to a defect in memory and learning. The synthesis of many proteins each may prove necessary (but not sufficient) for LTM formation.

STM formation, as defined in human and primate learning and memory studies, may be measured in fractions of seconds, with lifetimes that are compatible with the durations of electrophysiological measurements of neuronal activity. Other forms of STM formation, primarily in lower vertebrates and invertebrates, may be long-lived (seconds to hours or even days) as indicated in Sect. 2.1. Intermediate stages in STM formation have been proposed in studies on taste aversion in the chick (Gibbs and Ng 1979). It is possible that the longer lasting forms of STM reside in synaptic regions and are mediated by post-translational feedback loops between proteins, and thus would not be affected by the protein synthesis blockers, as has been observed. It must be remembered that the various molecular, pharmacological, behavioral, and electrophysiological interpretations of memory formation are not mutually exclusive and in many instances may reflect observation of the same neurobiological phenomenon from dissimilar vantage points.

See also: Learning and Memory: Computational Models; Learning and Memory, Neural Basis of; Long-term Potentiation and Depression (Cortex); Long-term Potentiation (Hippocampus); Memory,

Consolidation of; Memory: Synaptic Mechanisms; Short-term Memory: Psychological and Neural Aspects; Working Memory, Neural Basis of

Bibliography

- Agranoff B W, Cotman C W, Uhler M D 1999 Learning and memory. In: Siegel G J, Agranoff B W, Albers R W, Fisher S K, Uhler M D (eds.) *Basic Neurochemistry: Molecular, Cellular and Medical Aspects*, 6th edn. Lippincott-Williams & Wilkins, Philadelphia/New York, Chap. 50
- Bailey C H, Montarolo P, Chen M, Kandel E R, Schacher S 1992 Inhibitors of protein and RNA synthesis block structural changes that accompany long-term heterosynaptic plasticity in Aplysia. *Neuron* **9**: 749–58
- Davis H P, Squire L R 1984 Protein synthesis and memory: A review. *Psychological Bulletin* **96**: 518–59
- Flexner J B, Flexner L B, Stellar E, de la Haba G, Roberts R B 1962 Inhibition of protein synthesis in brain and learning and memory following puromycin. *Journal of Neurochemistry* **9**: 595–605
- Gerard R W 1953 What is memory? *Scientific American* **189**: 118–25
- Gibbs M E, Ng K T 1979 Neuronal depolarization and the inhibition of short-term memory formation. *Physiological Behavior* **23**: 369–75
- Lisman J E, Fallon J R 1999 What maintains memories? *Science* **283**: 339–40
- MacDonald J D 1968 *The Girl in the Plain Brown Wrapper*. Fawcett, Greenwich, CT
- Sanes J R, Lichtman J W 1999 Can molecules explain long-term potentiation? *Nature Neuroscience* **2**: 597–604
- Siegel G J, Agranoff B W, Albers R W, Fisher S K, Uhler M D (eds.) 1999 *Basic Neurochemistry: Molecular, Cellular and Medical Aspects*, 6th edn. Lippincott-Williams & Wilkins, Philadelphia/New York
- Steward O 1997 mRNA localization in neurons: A multipurpose mechanism? *Neuron* **18**: 9–12
- Voet D, Voet J G 1995 *Biochemistry*, 2nd edn. Wiley, New York

B. W. Agranoff

Protestantism and Gender

At the beginning of the twenty-first century, the term 'Protestant' is used to designate any group of Christians that does not acknowledge direct loyalty and connection either to the Roman Catholic hierarchy, or to the many catholic churches rooted in Euro-Asian cultures known collectively as Eastern Orthodox Christianity. In contemporary discourse the term 'Protestant' has come to apply to a broad range of spiritual communities. While this discussion of Protestantism and gender focuses on older Protestant 'denominations,' it should be remembered that by the late nineteenth century, Protestantism had expanded and changed greatly. By then, the majority of Prot-

estants were chiefly in movements whose historical development occurred outside of Europe, particularly in the Northern Hemisphere of the so-called 'New World.' Beginning in the nineteenth century, new forms spontaneously generated indigenous Christianities emerged in the Western Hemisphere, and by the early twentieth century, on the continents of Latin America, Africa, and Asia. The newer indigenous movements, identifiable as Christian by their emphasis on personal conversion, by the centrality of Jesus to their doctrine and practice, and by their fidelity to those texts broadly recognized as Christian scriptures (often called 'The Bible'), nevertheless bear little resemblance to the movements and cultures of Protestantism that had their roots in European history. Such Protestants now constitute the fastest growing edge of Christianity. It is too early to describe how such faith-communities will affect the future of gender relations, or reshape the historical directions of the original Protestantisms described here. And as will become clear, all Protestantisms, including these newer forms, are racked with conflict over gender issues, including the roles of women in churches and the nature of Christian attitudes toward sexuality. Today, strong forces within Christianity seek to reverse the clear directions toward gender equality and greater parity between men and women that have slowly shaped most Protestant belief and practice.

The origins of Protestantism are much debated. Deep resistance to the existing western Catholic sacramental system, loosely organized through the leadership of the bishops of Rome, appeared in late medieval Christianity. Much of this resistance presaged issues raised in the Protestant Reformation. However, Protestant beginnings are usually dated from the work of the German Reformer, Martin Luther, early in the sixteenth century. The four main trajectories that shaped the social formations of early Protestantism became differentiated and broadly identifiable 'traditions' of Protestant Christian belief and practice. The earliest, emerging under Luther's leadership, prospered chiefly in regions of Central Europe and quickly moved northward into Scandinavia. Almost simultaneously, a second major tradition emerged from Swiss Reform movements led by Zwingli, and was quickly joined by French-speaking groups under the leadership of John Calvin. These churches came to be known as 'Reformed.'

A third Continental stream of Protestantism emerged from the even more ardent and diverse dissenters within agrarian and peasant communities. These reforming groups were known in the first eras of Protestantism as Anabaptists. They were antihierarchical and egalitarian, and would eventually profoundly shape Protestant culture. However, initially the Anabaptist movements and communities generated hostility both from Lutheran and Reformed groups and from the Roman Catholic Church. The tensions generated by religious conflict on the Con-

tinient also reached the UK and also sparked a prolonged English Reformation struggle. The fourth Protestant tradition emerged from the settlement that created the Church of England and Anglicanism, which claimed to be a 'bridge' between Catholic and Protestant spirituality. It needs to be remembered that for three centuries Europe experienced a prolonged period of political, economic, and cultural turmoil as Christianity changed dramatically. This period is sometimes referred to as 'the Wars of Religion.'

By the late eighteenth century these realignments and newly-reordered ecclesial communities stabilized into clear cut 'denominations' with differences in doctrine and ritual practice that became matters of ongoing public dispute in cultures where Christianity predominated. Importantly, doctrinal differences were paralleled to differing normative views of how reforming Christian ecclesiastical culture and institutions should be related to political order and state power. Such differences are important to the topic of gender because Protestantist groups that were opposed to religious establishment and church support of state power would also become more egalitarian about issues of gender. But where the new Protestantisms became aligned with state power, they joined in suppressing dissent from their doctrine and practice as Roman Catholics frequently did during the Counter Reformation. As a result Anabaptist and Reformed Protestants migrated to seek refuge in any area of Europe or the New World of the Americas where they could find toleration or freedom from persecution.

Through these historical dynamics, Protestant reform seeded and reseeded itself, both in Europe and by the eighteenth century in North America. Migrations of many language groups also created hybrid Protestant churches that were Lutheran and Reformed. Lutheranism came to speak all the northern European languages, and 'Reformed Churches' from Switzerland, France, and the Netherlands dramatically flourished in the UK where the Church of England, or Anglicanism, had many rivals. Early European settlements in the Western Hemisphere often also brought multiple 'denominations' of the same Reformation traditions to North America. Dutch Reformed groups and Scottish and British Reformed communities created several separate denominations. Later European colonialism carried the four European-generated forms of Protestantism to areas of Africa, Asia and to parts of the sections of the Western Hemisphere previously colonized by Roman Catholic nations.

Reformed and 'Anabaptists,' sometimes called left-wing Protestants, in particular migrated in search of religious freedom and to escape state religious coercion. These traditions especially prospered in North America. Their emphasis on 'believers' communal equality' in the task of overseeing church life, in doctrinal discernment, and in scriptural interpretation, fit well with the democratizing tendencies of immigrant life. As a result, North American Prot-

estantism took on a dynamic life of its own that quickly produced new 'schisms of reform' that rebounded back to Europe. For example, Episcopalians, Anglicans in the USA, quickly generated a major division led by John Wesley that became known as 'Methodism.' Anabaptist or 'free church' movements that especially emphasized the priesthood of all believers and local autonomy of believing congregations came forth in great profusion, often spontaneously. Protestant denominationalism came to have indescribable diversity. Many such groups came to be known as Baptists. Baptists included forms of 'slave Christianity' that enabled African-Americans in slavery to generate a religious culture independent of the slave owners' 'white Christianity' imposed on them and designed to keep them in their place. Methodism and white Baptist practice spread in rural and frontier communities throughout the northern hemisphere. Even in the twenty-first century, the two largest Protestant churches in North America are Black and white Baptist denominations. Cultural historians now recognize that gender was never a theorized social reality until the modern period of western cultural history. This means that an accurate assessment of Protestantism's impact on gender cannot be uncovered through exercises in intellectual history alone. More recent methods of historical-material historiography enable us to see that the major impacts on gender relations within Protestantism were actually the result of changes in the religious practice initiated by the new Protestant churches. Slowly these changes produced concrete redirections in gender relations and, more slowly, reshaped how Protestants theorized gender.

Not surprisingly, early Protestantism continued what may be called 'traditionalist' Christian notions of gender. After the high Middle Ages, Catholic (now Roman Catholic) Christianity had made the moral theology of Thomas Aquinas its formal moral-theological standard. Thomistic thinking formalized Catholic moral teaching so that it came to presume two normative sources of moral knowledge: revelation and reason. Though what was meant by either revelation or reason differed in specific theological accounts, western Christians came to assume that both of these sources contributed to a proper knowledge of gender. What the resulting 'traditionalist Christian' view of gender presumed, roughly following Aristotelian biology, was that males and females differed by nature (science or reason) and divine will (revelation). Such thinking supported notions widely held in popular culture that made deep challenges to traditional gender relations difficult. Throughout Christian history women's roles varied and were frequently changed by resistance to existing traditionalisms. Well before the modern advent of women's organized efforts at political change, forms of what today are called Christian feminism existed and should be read as expressions by women of refusal of domination. Contemporary critical historiography confirms these continuous chal-

allenges but given the strength of patriarchal control, the point of visible institutional strain was never reached.

Even so, the full relevance of the Reformation to gender change is frequently missed because only formal teaching on gender is noted. The alternatives in communal form, ritual, and teaching in the Protestant traditions under discussion here were simultaneously shifts both in ecclesial/social forms and in Christian communal life. It was these new social formations and forms of ritualized communal practice that slowly changed gender relations in churches and in the broader culture.

The closing of the clergy and monastic-centered penitential systems, including the break with clerical celibacy brought the Christian parish, or in left-wing parlance, the Christian congregation, into the center of Protestant Christian faith practice. As a result, the primary units of Protestant spirituality soon became the local community (parish or congregation) and especially the family. In particular the patterns used to socialize persons as Protestant Christians dramatically altered gender relations. The Protestant family was theorized as an arena where both men and women, albeit initially in varied ways, were called to be agents of the formation of Christian communal life. Men were chiefly responsible for the public and the most authoritative ecclesial roles, but men and women alike were teachers of the faith, leaders of prayer, and witnesses to divine presence.

Protestantism particularly flourished in tandem with the invention of the printing press, which made Luther's, Calvin's and the Anabaptists' formal emphasis on scripture as the central theological source for Christians an actualizable reality. In all Protestant communities the locus of prayer, ritual practice, and theoretical (i.e. theological) formation shifted to the congregation and the family. None of the new Christian duties articulated by the Reformers were considered more important than that of transmitting the new 'knowledge of faith,' whether as scriptural literacy, or as newer catechism (doctrine). While none of this resulted in immediate transformations in female social roles because of the deeply gendered nature of the sex role systems, the new practices increased women's direct participation in matters of faith discernment. They also situated women within the central sacral systems of Protestant life. Prior to the Reformation, in spite of women's creativity and resistance, a lack of formal clerical power and women's remoteness from male-monopolized institutionalized hierarchies of ecclesial power prevailed. Pre-reformation Christian women certainly had created powerful gender alternatives, in particular monastic institutions for women. These enabled some women ('nuns') to live in an alternative social space where intellectual, artistic, and spiritual gifts could be cultivated. The impact of such efforts was severely marginalized structurally, however, and could be

easily manipulated by male hierarchies. Historians have also recognized that some initial Reformational changes had negative impacts on women's roles and status. For example, Luther's closing of nunneries and his and Calvin's insistence on clerical marriage, especially when gender dislocations led to arranged marriages between priests and nuns, robbed women of crucial existing social space. However, the need for alternatives also led the Protestant movements to create new gendered sub-cultures for women, usually in the form of Deaconess Orders or catechetical offices that enabled women to work with women and children in initiating them into 'the fullness of faith.'

The trajectory of changes in gender practice and, with the rise of modernity, gender theory, within these Protestant movements cannot be traced in detail. However, by the late nineteenth and early twentieth centuries, when gender began to become a contested political issue and calls for gender justice emerged widely in public discourse, calls for justice for women were voiced in all but the most 'traditionalist' Protestant Christian communities. It is erroneous, as some 'secular' historians as well as conservative Christians have done, to imagine that demands for change in gender roles occurred primarily outside of, and apart from, the life and practice of Protestant communities. Christian feminisms fed public political feminism and Christian anti-feminist resistance testifies to the militant presence of feminist agitation within Protestantism throughout the modern period. Feminism is perceived as secular in origins only because contemporary historians tend to discount the role of religion in positive social change and because feminist discourse within Protestantism was initially less confrontative than later feminist political rhetoric. But in Protestant cultures by the late nineteenth century, women's energies were powerful within the churches and pressure for change escalated simultaneously with the rising public efforts of women for political change. It is not too much to say that by the end of the nineteenth century Protestant churches were thoroughly feminized and women had gained considerable financial and organizational power. Within Christianity, antifeminism should be read as a symptom of increased pressure for change from women. A review of historical gains for women make clear that change came fastest in nations where Protestantism was strong but where the churches did not aspire to state-enforced establishment.

Even as many male church leaders reasserted Christian gender traditionalism against the claims of Christian feminisms, gender parity slowly increased in Protestant communities, and on-going feminist contentions did not disappear. On the contrary, by the twentieth century, contentions over gender exploded into nearly all Christian communities. Because feminist pressures were more muted in Anglican and Lutheran traditions until well into the century, and because controversies over ordination when they did

erupt were especially strident there, public awareness of church gender debates became widespread only at that historical moment. However, nearly all nations where Protestantism has had a clear religio-cultural role were the first to adopt explicit legal and moral norms of equality between men and women both within churches and in the wider society. Furthermore, because the congregation and the family have remained the cultic center of Protestant life and practice, active patriarchal misogyny is rarely overtly expressed within it. Protestant pastoral practice is invariably more egalitarian than official theologies would suggest so women are more likely to be pedestilized than criticized in Protestant rhetoric.

The ways in which clergy-lay ritual practice and power were realigned and the patterns by which Protestants groups related to state political power are also crucial to the pace of gender change. Where Protestant denominations replicated Roman Catholic efforts at church-state establishment as in the Church of England, in sectors of Lutheranism and in some Reformed churches such as those in early New England, USA, or predemocratic South Africa, Protestant leaders sought to reassure Christian rulers that the churches would officially conform to 'traditional values' and to the stability of social order. Efforts in support of keeping the Christian family 'traditional' were frequent among male leaders, though much more common in established churches than elsewhere. Gender change came much more rapidly where Reformed churches broke with theories of the divine right of kings and endorsed democratic political theory and where the 'free churches,' actively refusing religious establishment, sought to actualize a theology of universal priesthood. In fact, it was these free-church Protestants who demanded, on theological grounds, that religious toleration become the public stance of the just state. Left-wing Protestants militantly insisted that church-state separation is a matter of constitutional or political 'right' and in the USA they were successful.

Today it is often these Protestant movements, the first to ordain women and the strongest critics of identification with state power, who have become the most conservative on gender. This is because in the late twentieth century the locus of Protestant traditionalism in relation to gender change shifted dramatically. Until recently, theological progressivism had slowly carried the day in the older Protestantisms under discussion here. Over time, newer forms of knowledge prevailed and the historical-critical or hermeneutic sciences prevailed. The main dynamic of Protestant theological development was towards recognition that the 'two' sources of religious and moral knowledge—revelation and reason—supported the central Protestant historical/theological emphasis on 'continuing reformation.' Such 'theological liberalism,' as conservatives called it, was comprehended by most Protestants as the heart of their faith-claims.

Even when Protestant theological polemics included charges that gender justice was 'heresy' the trend was toward equality. In recent debates, however, the peculiar Protestant theological reaction of Biblicism has begun to alter older historical directions that embraced historical-critical readings of scripture.

In the late nineteenth century Protestant reaction to change began to take the form of what is called, somewhat misleadingly, 'biblical literalism.' So-called literalists do not seek to replicate ancient sex role patterns, but they project current traditionalist views to suggest that the Bible does not support gender justice. Such change, they contend, is impious because women's 'place' is forever set by what 'the Bible' teaches. This refusal of all truth except that found in 'the Bible' produced what in modern parlance is known as 'Fundamentalism.' Fundamentalism is a reactionary historical dynamic intrinsic to Protestantism as a text-based religion. Today many claim that other major religious movements are also becoming 'fundamentalist,' but such charges should be approached cautiously. Fundamentalism was generated by the particularities of Protestant inter-religious conflict. The refusal of gender justice within Protestant Christianity has parallels in other contemporary religions that accounts for the glib extension of the term. However, such usage is seriously misleading as to what is occurring within other religion. Even when reactionary Protestants and the Papacy attempt to enlist conservatives within Judaism, Islam, and even the great Asian religions against gender justice movements, Christian fundamentalist justifications for resisting such justice are not readily generalizable to other groups. Political and economic dynamics profoundly different than those occurring in older Protestant cultures are taking place elsewhere. For example, many former progressive left-wing Protestants in North America, are today for the first time, strongly aligned with global capitalist political power, whereas Islamic religious reaction is often aligned with anti-colonialism and resistance to capitalist 'modernizing' dynamics. Failure to notice such disparities may pit justice-seeking activists within the religions against each other.

In spite of the rising tide of fundamentalist anti-feminist attack within Protestant cultures and communities the struggles for justice continue in all of them. However, today even some moderate Protestants portray debates on gender justice as 'trivial.' Arguments about women's ordination, Christian sexual ethics, including Christian positions on gay, lesbian and transsexual lifestyles, and debates about reproductive choice and abortion are not understood, as earlier Protestant thought understood them, as aspects of the Protestant theological struggle towards genuine moral discernment and responsibility in a complex world. Most Protestants continue to espouse 'continuous and on-going reformation,' as the proper theological norm of Christian ecclesial and communal

life but the new fundamentalism has been chillingly successful in obscuring the connection between progressive Protestant theological principles and these particular issues. This is why in spite of the historical directions outlined here the question of whether continued movement toward equality in male and female relations among Protestants remain salient. Progressive moral change is always difficult to sustain both within religious communities and in the wider society. The fundamentalism now aligned with conservative political and economic power enjoys considerable success in part because of better political mobilization on the right. Even reputable media-spokespersons and scholars, who surely should know better, occasionally mislead in their interpretation of these contemporary controversies. For example, until the last decades of the twentieth century most Protestant churches from the traditions discussed here, emphatically embraced family planning and took no role in efforts to criminalize abortion. In ongoing debates legal abortion came to be embraced on various theological and moral grounds by major Protestant churches. However, the Christian Right has now succeeded in efforts to construct an 'unbroken' Christian anti-abortion tradition out of a pastiche of traditionalist Biblical pronatalism, Roman Catholic teaching, and post-sonogram technical imagery. To be sure, a few Protestant moral-theologians have been strongly anti-abortion on Biblicist grounds, but until recent efforts to 'reconstruct' Protestant history, it is not accurate to portray Protestants as anything other than committed to family planning and as moderately supportive of safe and legal abortion. Today, no reputable theological spokesperson would criticize family planning or embrace Roman papal doctrines that sex is chiefly for procreation, but fewer and fewer Protestant leaders recall the active debates that led to the embrace of reproductive choice for women in their communities. As a result the slow but dramatic march of Protestantism toward inclusion of women in 'the priesthood of all believers' may be reversed and the historical vocation of protestant theological reform rendered unrecognizable.

See also: Accidents, Normal; Buddhism and Gender; Catholicism and Gender; Christianity: Evangelical, Revivalist, and Pentecostal; Christianity: Liberal; Christianity Origins: Primitive and 'Western' History; European Religion; Feminist Theology; Fundamentalism (New Christian Right); Islam and Gender; Judaism and Gender; Reformation and Confessionalization; Religion and Gender; Religion, History of; Religiosity: Modern; Women's Religiosity

Bibliography

- Connery J R 1977 *Abortion, the Development of the Roman Catholic Perspective*. Loyola University Press, Chicago
- Coward H, Maguire D C (eds.) 2000 *Visions of a New Earth: Religious Perspectives on Population, Consumption, and Ecology*. State University of New York Press, Albany, NY
- Dombrowski D A, Deltete R 2000 *A Brief, Liberal, Catholic Defense of Abortion*. University of Illinois Press, Urbana, IL
- Feldman D M 1998 *Birth Control in Jewish Law: Marital Relations, Contraception, and Abortion as Set Forth in the Classic Texts of Jewish Law*. J. Aronson, Northvale, NJ
- Harrison B W 1983 *Our Right to Choose: Toward a New Ethic of Abortion*. Beacon Press, Boston
- Harrison B W 1985 In: Robb C S (ed.) *Making the Connections: Essays in Feminist Social Ethics*. Beacon Press, Boston
- Harrison B W (ed.) 1995 Rhetorics, rituals, and conflicts over women's reproductive power. *Journal of Feminist Studies in Religion*. Scholars Press 11(2) Special issue
- Gilson A B 1999 *The Battle for America's Families: a Feminist Response to the Religious Right*. Pilgrim Press, Cleveland, OH
- Gordon L 1990 *Woman's Body, Woman's Right: Birth Control in America*. Penguin Books, New York
- Kertzer D I 1993 *Sacrificed for Honor: Italian Infant Abandonment and the Politics of Reproductive Control*. Beacon Press, Boston
- Macdonald L O 2000 *A Unique and Glorious Mission: Women and Presbyterianism in Scotland 1830 to 1930*. J. Donald Publishers, Edinburgh, UK
- Maguire D C (ed.) 2001 *Sacred Choices: The Right to Contraception and Abortion in Ten World Religions*. Fortress Press, Minneapolis, MN
- Petchesky R P 1990 *Abortion and Woman's Choice: The State, Sexuality, and Reproductive Freedom*. Northeastern University Press, Boston
- Steffen L H 1994 *Life/Choice: The Theory of Just Abortion*. Pilgrim Press, Cleveland, OH
- Taub N 1994 *International Conference on Population and Development*. American Society of International Law, Washington, DC
- Walker W, Norris R A, Lotz D W, Handy R T 1986 *A History of the Christian Church*, 4th edn. T. & T. Clark, Edinburgh, UK

B. W. Harrison

Protocol Analysis in Psychology

Protocol analysis is a rigorous methodology for eliciting verbal reports of thought sequences as a valid source of data on thinking. When psychology emerged as a scientific discipline at the end of the nineteenth century, the majority of psychologists were interested in consciousness. They sought to examine the structure and elements of their thoughts and subjective experiences through introspective analysis. Within a few decades, the introspective method was discredited, and introspective descriptions were rejected as scientific evidence. Psychology turned away from studies of individuals' descriptions of their thoughts and first-hand experiences, and focused on observations of how other adults performed tasks in the laboratory. The rejection of introspection made the study of thinking come to a virtual halt until the 1950s, when the

technical innovations such as the computer led to the emergence of cognitive and information processing theories of psychological phenomena. In the new research approach to the study of thought processes, subjects were asked to 'think aloud,' leading to a new type verbal report of thinking that differed from the earlier introspective methods and became the core method of protocol analysis. The next section describes the fundamental problems of introspection and general solutions offered by protocol analysis.

1. Problems with Introspective Descriptions of Thinking and Proposed General Solutions

Adults describe their subjective experience, such as dreams, memories, and mental slips, on a regular basis, so why are observations of one's own experiences and thinking problematic as psychological data? The fundamental problem with descriptions of inner experience is that they are private and known only to the person who experienced them. There are no methods to verify the accuracy of these experiences independently. Consequently, protocol analysis is restricted to situations where the scientist has other types of information about the participant's thinking that can be used to validate the verbal reports.

The other major objection to introspection is that the mere act of observation of one's thoughts and experiences changes them. For example, when someone examines a single visual image appearing during thinking to search for specific associated details, this analysis tends to make the desired details emerge and thus changes the content and structure of the original image. This extended analysis also interrupts the normal flow of visual images, so it is impossible just to resume the thought process where it was interrupted. To minimize this type of interference in protocol analysis, research participants are instructed to think aloud while problem solving so they can concentrate on the experimental task, and let verbalization of thought be secondary.

2. Historical Background

Self-observation of thinking has a long history, which helps explain the controversies surrounding introspection, as well as the development of less controversial methods for eliciting verbalizations of thinking. The first systematic attempts to understand the structure of thinking by observing one's own spontaneous thoughts have been traced back at least as far as the Greeks. Based on observation of his thoughts, Aristotle argued that thinking corresponds to a sequence of thoughts (see Fig. 1), where the brief transition periods between consecutive thoughts (see the arrows in Fig. 1) do not contain any reportable information.

Aristotle's general account of thinking as a thought sequence has never been seriously challenged (Ericsson and Crutcher 1991). The controversies about introspection have all involved attempts to go beyond the immediately observable aspects. For example, Aristotle tried to infer the unobservable processes that control the thought sequence (see the arrows in Fig. 2). By examining his memory of which thoughts tended to follow each other, Aristotle concluded that previously experienced associations between thoughts were the primary determining factor. Philosophers in the seventeenth, eighteenth and nineteenth centuries (Ericsson and Crutcher 1991) asked questions that required more involved inferences and introspective analyses. For example, can every thought be described as an amalgam of memories? In order to evaluate these claims empirically, philosophers would typically relax and daydream, allowing their thoughts to wander. Once in a while they would try to freeze an image and inspect it carefully to assess its sensory modality and components. Based on such detailed observations, the famous British philosopher David Hume came to the surprising conclusion that thoughts were as detailed as original perceptions—we cannot have a visual image a printed page of text without imaging every letter on the page at the same time.

By the nineteenth century, philosophers generally rejected Hume's claim and agreed that the process of analyzing a visual image changes its original appearance. For example, when a person has an image of a rose and then decides to inspect its petals, the whole mental image changes. Consequently, the process of examining mental images is fundamentally different from observing solid objects in the real world. A real rose can be looked at repeatedly from different directions without being changed, even its petals can be inspected with a looking glass without any change. In contrast, extended introspective analysis of a mental image changes its content and structure and cannot therefore be used to study thoughts as they emerge during spontaneous thinking.

In the beginning of the twentieth century, psychologists at the University of Würzburg tried to minimize this form of reactivity by training introspective observers. These observers were given thought-provoking questions and asked to respond as fast as possible. Only after their answers had been given would the observers start to recall the thoughts that they had had while answering the questions. The observers were trying to recall the most basic elements of their thoughts and visual images. Most reported thoughts consisted of visual and auditory images, but some subjects claimed to have experienced thoughts without any corresponding imagery—imageless thoughts. The alleged existence of imageless thoughts had far-reaching theoretical implications and led to a heated exchange between the psychologists who claimed to have observed them and others who argued that these reports were artifacts of inappropriate

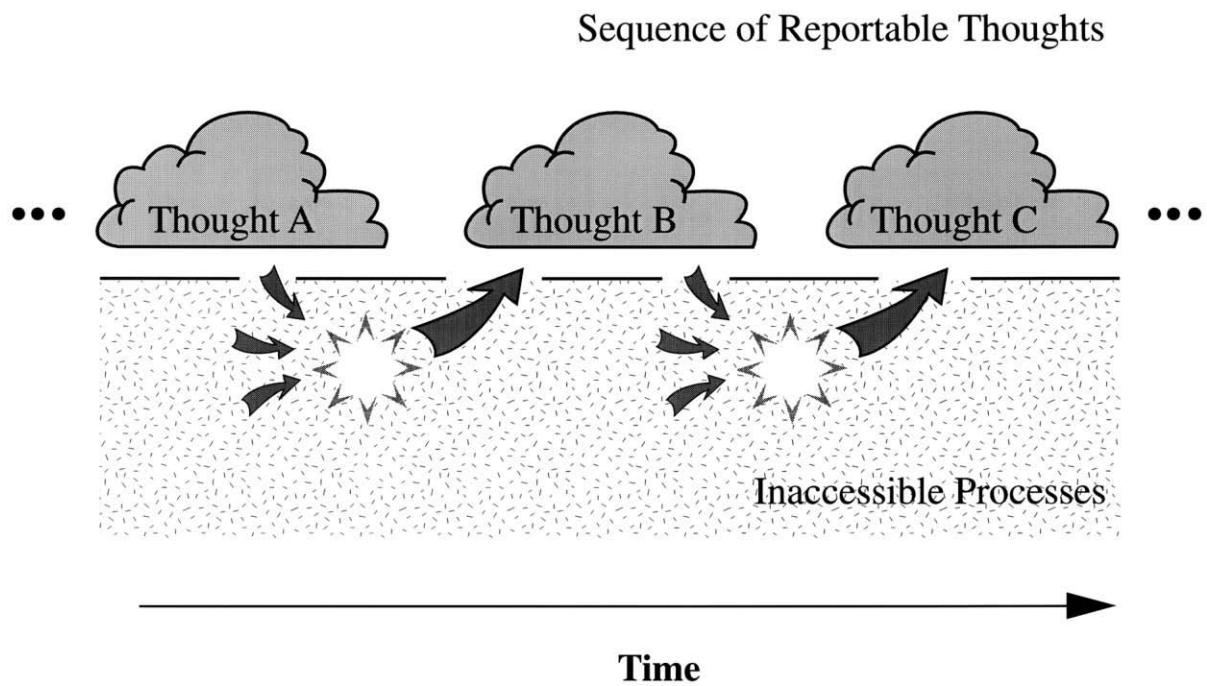


Figure 1

An illustration of Aristotle's view of thinking as a sequence of states with reportable thoughts, while the processing responsible for transitions between consecutive thoughts remains inaccessible

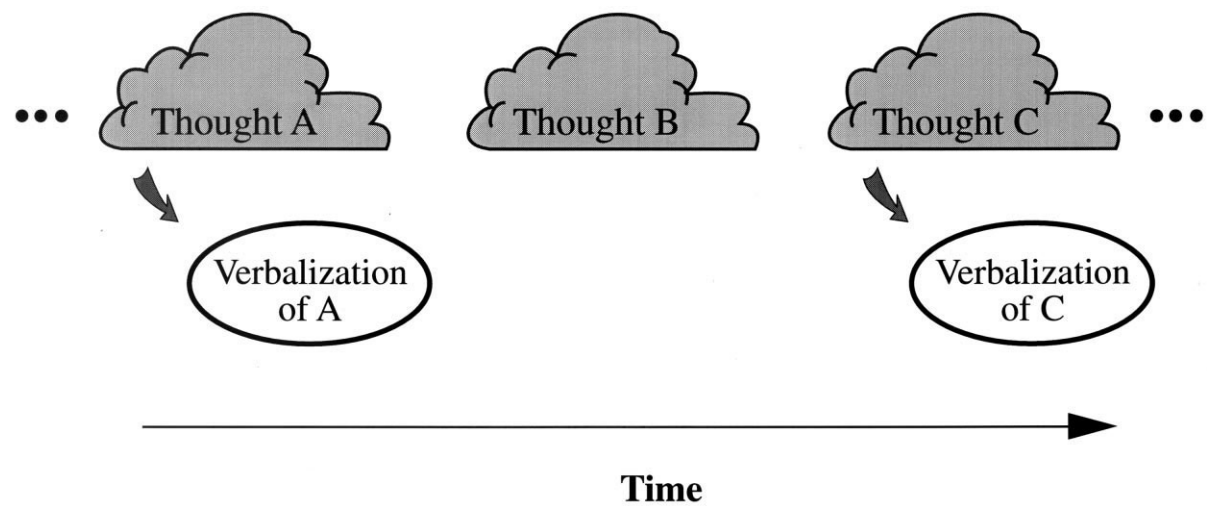


Figure 2

An illustration of a covert sequence of thoughts and the associated overt verbalizations of most thoughts passing through attention during the performance of a task

reporting methods and theoretical bias of the observers. A devastating conclusion arose from this controversy: the existence of imageless thoughts could

not be resolved empirically. This finding raised fundamental doubts about analytic introspection as a scientific method.

Reacting to the methodological and theoretical problems of the introspective method, psychologists redirected their research from conscious experience and thinking and towards observable behavior, and focused on basic learning processes shared by adults, children, and animals. For example, some of the tasks involved learning the path from a starting location to a goal location in an arbitrary maze, where the benefits of prior experience and knowledge would be minimal. Similar learning tasks required participants to memorize 'meaningless' materials, such as nonsense syllables like XOK and ZUT, where the memory performance was measured objectively by scoring the sequence of recalled syllables for accuracy. The hypothesis that memory for nonsense syllables was based only on basic associations, and was later rejected in studies by cognitive psychologists showing that during memorization subjects reported thinking of words that were similar to the studied nonsense syllable, such as 'cozy' for KOZ and 'payment' for PYM (Prytulak 1971). When subjects were able to associate the nonsense syllable with a word, the speed and accuracy of memorization improved (Montague 1972), which provides validation for the reported thoughts.

The cognitive revolution in the 1960s renewed interest in higher-level cognitive processes and how thinking allows individuals to generate solutions to novel tasks. Cognitive theories describe how individuals are able to apply acquired knowledge and procedures to novel problems, such as mental multiplication of any combination of two two-digit numbers. Information processing theories (Newell and Simon 1972) proposed computational models that could reproduce the observable aspects of human performance on well-defined tasks through the application of explicit procedures.

One of the principle methods of the information processing approach is task analysis. Task analysis specifies the range of alternative procedures that people could use, in light of their prior knowledge of facts and procedures, to generate correct answers to a task. Let me illustrate how task analysis can be applied to mental multiplication. Most adults have only limited mathematical knowledge: they know their multiplication table and only the standard 'pencil and paper' procedure taught in school for solving multiplication problems. Accordingly, one can predict that they will solve a specific problem such as 36×24 by first calculating $4 \times 36 = 144$ then adding $20 \times 36 = 720$. More sophisticated subjects may recognize that 24×36 is equivalent to $(30 + 6) \times (30 - 6)$ and use the formula $(a + b) \times (a - b) = a^2 - b^2$, thus calculating 36×24 as $30^2 - 6^2 = 900 - 36 = 864$. The choice of alternative procedures participants use to generate the answer can be inferred by the time needed and verbal reports of their thoughts during problem solving.

In conclusion, the theoretical and methodological controversies about verbal reports have never cast doubt people's ability to recall part of their thought

sequences. The controversies have centered around efforts to go beyond the sequence of thoughts (see Fig. 1), to analyze their detailed structure through introspection, and to infer the processes controlling the generation of new thoughts. In fact, all major theoretical frameworks concerned with thinking have advocated the use of verbally reported sequences of thoughts (Ericsson and Crutcher 1991). For example, the behaviorist John B. Watson pioneered the use of 'think aloud,' and the gestalt psychologist Karl Duncker established it as a major method.

3. Protocol-analysis Methodology

The central assumption of protocol analysis is that it is possible to instruct subjects to verbalize their thoughts in a manner that doesn't alter the sequence of thoughts mediating the completion of a task, and can therefore be accepted as valid data on thinking.

3.1 Elicitation of Nonreactive Verbal Reports of Thinking

Based on their theoretical analysis, Ericsson and Simon (1993) argued that the closest connection between thinking and verbal reports is found when subjects verbalize thoughts generated during task completion (see Fig. 2). When subjects are asked to think aloud, some of their verbalizations seem to correspond to merely vocalizing 'inner speech,' which would otherwise have remained inaudible. Nonverbal thoughts can also be often given verbal expression by brief labels and referents. For example, when one subject was asked to think aloud while mentally multiplying 36 by 24 on two test occasions one week apart, the following protocols were obtained:

(a) OK, 36 times 24, um, 4 times 6 is 24, 4, carry the 2, 4 times 3 is 12, 14, 144, 0, 2 times 6 is 12, 2, carry the 1, 2 times 3 is 6, 7, 720, 720, 144 plus 720, so it would be 4, 6, 864.

(b) 36 times 24, 4, carry the—no wait, 4, carry the 2, 14, 144, 0, 36 times 2 is, 12, 6, 72, 720 plus 144, 4, uh, uh, 6, 8, uh, 864.

In these two examples, the reported thoughts are not introspectively analyzed into their perceptual or imagery components, but merely verbally expressed and referenced, such as 'carry the 1,' '36,' and '144 plus 720.' Similarly, subjects are not asked to describe or explain how they solve these problems. Instead, they are asked to remain focused on solving the problem and merely to give verbal expression to those thoughts that emerge in attention while generating the solution under normal (silent) conditions.

If the act of verbalizing subjects' thought processes does not change the sequence of thoughts, then subjects' task performance should not change as a result of thinking aloud. In a comprehensive review of

dozens of studies, Ericsson and Simon (1993) found no evidence that the sequences of thoughts (accuracy of performance) were changed when subjects thought aloud as they completed the tasks, compared with subjects who completed the same tasks silently. However, some studies showed that think-aloud subjects would take somewhat longer to complete the tasks—presumably due to the additional time required to produce the overt verbalization of the thoughts.

The same theoretical framework can also explain why other types of verbal report procedures consistently change cognitive processes. For example, when subjects are instructed to explain or carefully describe their thoughts, they are not able merely to verbalize each thought as it emerges, they must engage in additional cognitive processes to generate the thoughts corresponding to the required explanations and descriptions. This additional cognitive activity changes the sequence of mediating thoughts. Instructions to explain and describe the content of thought are reliably associated with changes in ability to solve problems correctly (Ericsson and Simon 1993).

In sum, after brief training in giving verbal reports, subjects can think aloud without any systematic changes to their thought process (see Ericsson and Simon 1993 for detailed instructions and associated warm-up tasks recommended for laboratory research). This means that subjects must already possess the necessary skills for verbalization of thoughts.

3.2 Validity of Verbalized Information from Think Aloud

When adults are able to perform tasks while thinking aloud without sacrificing accuracy and speed, the verbalized information would almost have to reflect some aspect of the regular cognitive processes. By analyzing the information expressed as verbalized thoughts, it is possible to assess the validity of the verbalized information. In protocol analysis, the verbalized thoughts are compared with intermediate results generated by different strategies that are specified in a task analysis (Ericsson and Simon 1993). The sequence of thoughts verbalized while multiplying 24 by 36 mentally (reproduced above) agrees with the sequence of intermediate thoughts specified by one, and only one, of the possible strategies for calculating the answer.

Even when subjects think aloud with its close connection between thoughts and reports, correspondence between verbalized thoughts and intermediate products predicted from the task analysis is not perfect (see Fig. 2). The lack of one-to-one correspondence is due primarily to the fact that not all thoughts which pass through attention are verbalized and some processing steps (thoughts) may be short-circuited with acquired skill. However, there is persuasive

evidence of validity for the thoughts that are verbalized (Ericsson and Simon 1993). Even if a highly skilled participant's think-aloud report in the multiplication task only consisted of '144' and '720,' the reported information would still be sufficient to reject many alternative strategies, because these strategies did not involve generating both of the reported intermediate products. The general finding that a task analysis can identify *a priori* the specific intermediate products that are later verbalized by subjects during their problem solutions, provides the strongest evidence that concurrent verbalization reflects the processes that mediate the actual generation of the correct answer.

More generally, verbal reports are only one indicator of the thought processes that occur during problem solving. Other indicators include reaction times (RTs), error rates, patterns of brain activation, and sequences of eye fixations. Given that each kind of empirical indicator can be separately recorded and analyzed, it is possible to compare the results of such independent data analyses. In their review, Ericsson and Simon (1993) found that longer RTs were associated with verbal reports of a larger number of intermediate thoughts than those corresponding to shorter RTs. Furthermore, there seemed to be close correspondence between subjects' thoughts and what information that they looked at—when subjects verbalized thoughts about objects in the environment they very frequently looked at them.

Finally, the validity of verbally reported thought sequences depends on the time interval between the occurrence of a thought and its verbal report, where the highest validity is observed for concurrent, think-aloud verbalizations. For tasks with relatively short response latencies (less than 5–10 s), subjects are able to recall their sequences of thoughts accurately immediately after the completion of the task, and the validity of this type of retrospective reports remains very high. However, for cognitive processes of longer duration, the problems of accurate recall of prior thoughts increases, with a corresponding decrease in validity of the verbal reports.

3.3 Explanations of Validity Problems with Other Types of Verbal Reports

Protocol analysis specifies the constrained conditions necessary for valid, nonreactive verbalizations of thinking. Many of the problems with verbally reported information obtained by other methods can be explained as violations of protocol analysis methodology. The first problem arises when the investigators ask subjects to give more information than is contained by the their recalled thought sequences. For example, some investigators ask subjects why they responded in a certain manner. Sometimes subjects will have deliberated and their recalled thoughts will

provide a sufficient answer, but typically the subjects need to go beyond any retrievable memory of their processes to give an answer. As subjects can only access the end products of their cognitive processes during perception and memory retrieval, they cannot report why only one of several logically possible thoughts entered their attention, so they must confabulate to answer such questions. In support of this argument, Nisbett and Wilson (1977) found that subjects' responses to 'why' questions were in many circumstances as inaccurate as those given by other subjects who merely observed and tried to explain another subject's behavior.

Second, scientists are interested primarily in the general strategies and methods subjects use to solve a broad class of problems in a domain, such as mathematics or text comprehension. They often ask subjects to describe their general methods after solving a long series of different tasks, which often leads to misleading summaries or after the fact reconstructions of what subjects think they must have done. In the rare cases when subjects have deliberately and consistently applied a single general strategy to solving the problems, they can answer such requests easily by recalling their thought sequence from any of the completed tasks. However, subjects typically employ multiple strategies, and their strategy choices may change during the course of an experimental session. Under such circumstances subjects would have great difficulty describing a single strategy that they used consistently throughout the experiment, thus their reports of such a strategy would be poorly related to their averaged performance. Hence, general strategy descriptions are usually inadequate, even when reports after each trial are valid descriptions of thought sequences.

Similar problems have been encountered in interviews of experts (Hoffman 1992). When experts are asked to describe their general methods in professional activities, they have difficulties, and there is often poor correspondence between the described methods and their observed behavior. However, when experts perform challenging representative problems from their domain of expertise while thinking aloud, their verbalizations consistently reveal a valid and informative account of their thought sequences during complex thinking and planning (Ericsson and Lehmann 1996). In sum, to obtain the most valid and complete trace of thought processes, scientists should strive to elicit those verbalizations that most directly reflect the subjects' spontaneous thoughts generated while they perform tasks that are representative of the studied phenomenon.

4. Range of Application

Protocol analysis has emerged as one of the principal methods for studying thinking in cognitive psychology (Crutcher 1994), cognitive science (Simon and Kaplan

1989), and behavior analysis (Austin and Delaney 1998). As further evidence of its validity, protocol analysis now plays a central role in applied settings, such as in the design of surveys and interviews (Sudman et al. 1996) and user testing of computer software (Henderson et al. 1995). Finally, several interesting adaptations of verbal-report methodology are emerging in the study of text comprehension (Pressley and Afflerbach 1995) and education (Renkl 1997).

See also: Behavior Analysis, Applied; Cognitive Science: Overview; Critical Thinking, Cognitive Psychology of; Education: Skill Training; Educational Assessment: Major Developments; Ergonomics, Cognitive Psychology of; Introspection: History of the Concept; Job Design and Evaluation: Organizational Aspects; Language and Thought: The Modern Whorfian Hypothesis; Organizational Behavior, Psychology of

Bibliography

- Austin J, Delaney P F 1998 Protocol analysis as a tool for behavior analysis. *Analysis of Verbal Behavior* **15**: 41–56
- Crutcher R J 1994 Telling what we know: The use of verbal report methodologies in psychological research. *Psychological Science* **5**: 241–44
- Ericsson K A, Crutcher R J 1991 Introspection and verbal reports on cognitive processes—two approaches to the study of thinking: A response to Howe. *New Ideas in Psychology* **9**: 57–71
- Ericsson K A, Lehmann A C 1996 Expert and exceptional performance: Evidence on maximal adaptations on task constraints. *Annual Review of Psychology* **47**: 273–305
- Ericsson K A, Simon H A 1993 *Protocol Analysis: Verbal Reports as Data*. MIT Press, Cambridge, MA
- Henderson R D, Smith M C, Podd J, Varela-Alvarez H 1995 A comparison of the four prominent user-based methods for evaluating the usability of computer software. *Ergonomics* **38**: 2030–44
- Hoffman R R (ed.) 1992 *The Psychology of Expertise: Cognitive Research and Empirical AI*. Springer-Verlag, New York
- Montague W E 1972 Elaborative strategies in verbal learning and memory. In: Bower G (ed.) *The Psychology of Learning and Motivation*. Academic Press, New York, Vol. 6, pp. 225–302
- Newell A, Simon H A 1972 *Human Problem Solving*. Prentice-Hall, Englewood Cliffs, NJ
- Nisbett R E, Wilson T D 1977 Telling more than we can know: Verbal reports on mental processes. *Psychological Review* **84**: 231–59
- Pressley M, Afflerbach P 1995 *Verbal Protocols of Reading: The Nature of Constructively Responsive Reading*. L Erlbaum Associates, Hillsdale, NJ
- Prytulak L S 1971 Natural language mediation. *Cognitive Psychology* **2**: 1–56
- Renkl A 1997 Learning from worked-out examples: A study on individual differences. *Cognitive Science* **21**: 1–29
- Simon H A, Kaplan C A 1989 Foundations of cognitive science. In: Posner M I (ed.) *Foundations of Cognitive Science*. MIT Press, Cambridge, MA, pp. 1–47

Sudman S, Bradburn N M, Schwarz N (eds.) 1996 *Thinking about Answers: The Application of Cognitive Processes to Survey Methodology*. Jossey-Bass, San Francisco, CA

K. A. Ericsson

Psychiatric Assessment: Negative Symptoms

Negative symptoms comprise, among others, flattening of affect, reduction of speech, and social withdrawal. They occur in different psychiatric diseases, but especially in schizophrenia, where they are traditionally seen as an important part of the symptomatology. In the last decade, they have evoked a special research interest due to the fact that novel/atypical antipsychotics have been developed with an improved therapeutic efficacy in negative symptoms of schizophrenia.

1. Definition of Negative Symptoms

Negative symptoms in schizophrenia comprise broadly of blunted affect, apathy, attentional impairment, poverty of thought or speech, incoherence, loosening of association, thought blocking, and social withdrawal. However, there is no unanimous agreement on the concept of negative symptoms in schizophrenia: some authors adhere to a broad interpretation (Meltzer and Zureick 1989, Rösler and Hengesch 1990), while others prefer a narrow approach (Crow 1985, Strauss et al. 1974, Wing 1989). The controversy centers on incoherence, loosening of association, and thought-blocking, since these are also attributes of positive symptoms. This disparity in the selection of items also manifests itself in the outcome of patient testing with different rating scales. Different authors use different concepts and definitions for the type and number of symptoms implicated (Rösler and Hengesch 1990).

These problems of definition can also be identified in the rating scales used for measuring negative symptoms.

2. History of the Term 'Negative Symptoms'

The origins of the positive/negative distinction are usually attributed to Jackson (1887), the British neurologist who drew the distinction in order to delineate primary from secondary neurological phenomena. Accordingly, positive symptoms are defined as the presence of a behavioral function that is otherwise not normally present in an individual. Negative symptoms are defined as the absence of a behavioral

function that is usually present in the normal individual. Negative symptoms all appear to be characterized by the loss of an important cognitive or behavioral function, while positive symptoms represent an exaggeration or distortion of a normal function.

3. Standardized Assessment and Dimensional Differentiation of Negative Symptoms

To improve the reliability of diagnostic assessment of negative symptoms, several scales have been developed, mostly in the last two decades (SANS: 1981, PANSS: 1987, SAPS: 1991).

Standardization of the assessment of negative symptoms through rating scales had a positive effect on reliability. For example, the interclass correlations for the Scale for the Assessment of Negative Symptoms (SANS) and the Scale for the Assessment of Positive Symptoms (SAPS) complex global ratings range between 0.70 and 0.88. Principally, the negative symptoms can be rated with as much reliability as positive symptoms.

Rating scales applied in the past decade include, among others, the anergia score of the Brief Psychiatric Rating Scale (BPRS), the Manchester Scale and Crow's modification of this scale, the negative score of the Positive and Negative Scale (PANSS), the SANS, and Abrams and Taylor's Rating Scale for Emotional Blunting. Nowadays, the SANS and the PANSS are the most frequently used rating scales.

In nearly all studies in which a broad spectrum of negative symptoms was investigated by a factor analytical approach, two or three dimensions of negative symptoms were found. This emphasizes the inadequacy of a unidimensional concept of negative symptoms in the complex phenomenology of this domain (Meltzer and Zureick 1989).

4. Symptomatic Overlap with Parkinsonism and Depression

The differentiation between depression and negative symptoms is difficult on the basis of both clinical phenomenology and rating scales, and therapeutic effects on depressive symptoms might be misinterpreted as successful treatment of negative symptoms and vice versa. Gaebel (1989) reported a correlation of $r = 0.34$ – 0.57 at different measurement points for the BPRS scores depression and anergia. In a follow-up study on schizophrenic patients, the correlation between the scores for depression and retardation/apathy in the Inpatient Multidimensional Psychiatric Scale (IMPS) amounted to $r = 0.35$ (Möller and von Zerssen 1986).

Another problem in the interpretation of rating scale data from studies in this field is that the negative symptoms score correlates positively with the drug-

induced Parkinson syndrome. Furthermore, there is a correlation between negative symptoms and hyperkinetic neurological symptoms such as akathisia and dyskinesia. One could therefore conclude that drugs without extrapyramidal side effects in general may appear to be more advantageous in treating negative symptoms.

5. Negative Symptoms in Schizophrenia

Both Kraepelin and Bleuler described a core of schizophrenic symptoms corresponding to the modern concept of negative symptoms (Table 1). Later, this position was gradually displaced by an emphasis on productive features as being pathognomonic of schizophrenia, in fact, as being a prerequisite for its diagnosis. This is exemplified in the diagnostic criteria for schizophrenia put forward by Schneider 1959, Feighner et al. (1972), in the research diagnostic criteria (Spitzer et al. 1978), and in the *Diagnostic and Statistical Manual DSM-III*, *DSM-III-R*, and *DSM-IV* criteria.

The need to incorporate both negative and positive features into a model of schizophrenia was brought out by the factor analytic studies of the 1960s and in particular by the research of Strauss et al. (1974), who adapted Jackson's (1887) terminology for classifying different profiles in schizophrenia.

Whereas Crow's (1980b) model viewed the positive and negative symptoms as annulated facets of schizophrenia that may reflect separate pathological processes, Andreasen (1982) proposed instead a typological model. They described positive and negative symptoms as opposing features that characterize different subtypes.

Which symptoms of schizophrenia are considered positive or negative varies across different authors. Nearly all authors include flat affect and poverty of speech among the negative symptoms, and hallucinations and delusions among the positive symptoms. Most regard anhedonia, apathy, avolition, or abulia

as negative symptoms. Thought disorder, bizarre behavior, and inappropriate affect, however, are variably classified as positive or negative.

6. Negative Symptoms in other Psychiatric/Mental Disorders

Anergia, anhedonia, poverty of speech, and affective flattening, which occur as part of the negative symptoms in schizophrenia, were also observed in patients severely ill with depressive disorder. Although negative symptoms of schizophrenia may be different from depressive symptoms in depression, they do overlap, thus making negative symptom complexes and psychomotoric retardation, lack of energy, anhedonia, and other symptoms of depression significantly associated. On the other hand, it was observed that a distinct hypokinetic syndrome involving anhedonia, depressed mood, and cognitive impairment existed not only in negative schizophrenia and retarded depression, but also in Parkinson's disease, suggesting that reduced dopamine turnover might play a role in the pathophysiology of each.

7. Primary and Secondary Negative Symptoms

The differentiation between primary and secondary negative symptoms, as proposed by Carpenter et al. (1985), seems of great theoretical and clinical relevance. While primary negative symptoms are the consequence of a moribogenic deficit state, secondary negative symptoms are the consequence of:

- (a) positive symptoms (e.g., social withdrawal because of paranoid ideas);
- (b) extrapyramidal effects (e.g., motor retardation as an indicator of neuroleptogenic akinesia);
- (c) depressive symptoms (e.g., postpsychotic or pharmacogenic depression); or
- (d) social understimulation (e.g., hospitalization).

Table 1
Relationship between the concept of basic symptoms and the concept of negative symptoms (Klosterkötter 1990)

Kraepelin (1904)	Bleuler (1911)	Wing and Brown (1970)	Strauss et al. (1974)
Cognitive decline ('Stultification,' 'Dementia')	Association disorder	Schizophrenic thought disorders Slowness of thought Poverty of speech	Certain kinds of formal thought disorders, such as blocking
Flatness of affect	Affective disorder	Flatness of affect	Blunting of affect
Loss of intentional power	Autism Ambivalence Abulia Lack of initiative	Social withdrawal Low motivation	
Loss of energy	Lack of interest	Lack of initiative Underactivity	Apathy

According to this distinction, the treatment of negative symptoms has to be differentiated concerning the different causes.

Angst et al. (1989) underlined the necessity of discriminating negative symptoms according to different stages in the course of schizophrenia, i.e., the prepsychotic period, the productive psychotic period (characterized by a mixture of positive and negative symptoms, whereby negative symptoms may be hidden at this stage), and the postpsychotic period (characterized by residuum or defect with or without positive symptoms, postpsychotic endogenous, or reactive depression, postremissive exhaustion).

Crow (1980a) postulated the stability of a specific subtype of schizophrenia with positive symptoms (type I) and a subtype with negative symptoms (type II), explaining this by different biological causes. Most other authors stressed the point that negative and positive symptoms can be combined cross-sectionally or longitudinally and cannot be taken as indicators for a specific biological or phenomenological subtype of schizophrenia.

8. Relationship between Positive and Negative Symptoms in Schizophrenia

The positive/negative distinction may represent different manifestations of the common underlying process. It may also suggest the existence of different pathophysiologies in schizophrenia. Hints about the inter-relationships of these processes may be gleaned from the intercorrelations among their final common expressions as symptomatic phenomenology. An initial effort by Andreasen (1982) showed that the positive and negative symptoms clusters were significantly correlated within clusters and inversely correlated between clusters, suggesting a bipolar construct with positive and negative symptoms at the opposite ends of a continuum. This was replicated by another group. However, later investigations showed little or no relationship between positive and negative symptoms, strongly suggesting that they are independent processes within schizophrenia.

Since both positive and negative symptoms are frequent in schizophrenia, they are likely not to be completely independent (Möller et al. 1995). However, some findings have been controversial. For example, Rosen et al. (1984) failed to find a positive/negative symptom correlation.

9. Course of Negative Symptoms in Schizophrenia

Findings on the course of positive and negative symptoms over time have aroused considerable controversy. In acute and/or first episode cases, negative symptoms are sparse and linked tightly to the presence of other manifestations of psychoses, especially to

positive symptoms (Möller and von Zerssen 1986). In the North American Risperidone Study (Marder and Meibach 1994), in which risperidone was compared to haloperidol, both drugs improved positive and negative symptoms significantly. Similar results were obtained in clinical trials on other novel/atypical neuroleptics such as olanzapine, sertindole, amisulpride, etc. Clearer results may be obtained if more sophisticated designs are used for psychopharmacological trials on negative symptoms. A European collaboration group therefore made some suggestions (Möller et al. 1994), focusing on patients with predominantly negative symptoms (Table 2).

During the acute phase, negative symptoms can often be unstable. For example, McCreddie (1989) studied schizophrenic patients suffering a first episode at three points in time (admission, one- and two-year follow-up). The percentages of patients with positive and negative symptoms were 96 and 23 percent at admission, 22 and 40 percent at the one-year follow-up, and 21 and 25 percent at the two-year follow-up, respectively. Other studies with acute patients have shown that negative symptoms are more stable over time. Ragin et al. (1989) found an increase in the level of poverty of speech between index admission and seven months after discharge. Wagman et al. (1987) found that the anergia score of the BPRS was significantly correlated across four months of repeated measures in patients with deficit symptoms, but not in patients without deficit symptoms.

Overall, negative symptoms appear to be more stable in subacute/subchronic patient samples. Pogue-Geile and Harrow (1985) recorded high and stable levels of anhedonia one and three years after episodes of schizophrenia. Biehl et al. (1986) found that negative symptoms had significant retest stability between years two and five in a sample from the International Pilot Study of Schizophrenia.

In a follow-up study of schizophrenia carried out in Munich (Möller and von Zerssen 1986), patients were rated using the IMPS at admission, discharge, and at a five-year follow-up. A decrease was observed in the mean score of positive symptoms and also in syndrome retardation as an indicator of negative symptoms. The correlation for the retardation score between discharge and follow-up was 0.36. In a similar 15 years follow-up study on first hospitalized patients with schizophrenia, a decrease of negative symptoms was found between admission and discharge, and an increase between discharge and follow-up (Möller et al. in press).

Many authors came to the conclusion that, in contrast to the views of Crow (1980b), the negative subtype and the positive subtype of schizophrenia are not stable conditions during the short- or long-term course of the disease. Most authors stressed the point that negative and positive symptoms can be combined cross-sectionally or longitudinally, and cannot be taken as indicators of a specific biological or phenomenological subtype.

Table 2

Suggestions of the working group on negative symptoms in schizophrenia (Möller et al. 1994)

Aim: evaluating treatment effects on negative symptoms in schizophrenia

Patient selection.

- Positive symptoms not dominating the actual clinical picture, e.g. PANSS negative type
- Duration of negative symptoms > 6 months
- Stable condition of the schizophrenic illness > 6 months
- Flat affect and poverty of speech as core symptom of negative symptoms
- No/low depression score

Design: double-blind comparison to placebo or active drug

Efficacy parameters: BPRS or SANS or PANSS

Other scales: depression scales, EPS scales

Statistical analyses

- End-point comparison
- Interaction with productive symptoms, depression, EPS

BPRS: Brief Psychiatric Rating Scale; EPS: extrapyramidal symptoms; PANSS: Positive and Negative Syndrome Scale; SANS: Scale for the Assessment of Negative Symptoms

10. *Biological Correlates of Negative Symptoms in Schizophrenia*

The most prominent biochemical finding in schizophrenic patients with negative symptoms appears to be the reduction in central dopaminergic, serotonergic, and noradrenergic activity, and, if at all, a general common pattern can be described. This decrease in amine activity tends to be associated with structural brain abnormalities, i.e., cortical atrophy or enlarged ventricles. There are indications that typical neuroleptics reduce at least those negative symptoms of schizophrenia that are secondary to positive symptoms when the positive symptoms are effectively treated. However, negative symptoms of schizophrenia that occur independently of positive symptoms may also be reduced, especially through treatment with atypical antipsychotics or with antidepressants. The efficacy of drug treatment of negative symptoms seems to be related to their pharmacological profile, i.e., their interference with dopaminergic, noradrenergic, and serotonergic receptor systems. As to atypical antipsychotics, among others the 5-HT_{2A} blockade was hypothesized as being a relevant mechanism for the efficacy of atypical antipsychotics in treating negative symptoms.

As mentioned above, there is a considerable overlap in the symptomatology of psychiatric disorders with respect to negative symptoms. Also, different types of negative symptoms in schizophrenia have to be considered. In addition, the lack of homogeneity in the definition and assessment of negative symptoms in schizophrenia can be a problem. All these difficulties have consequences for research and explain some of the inconsistencies in biochemical and pharmacologic schizophrenia research (Meltzer and Zureick 1989). The results of biochemical investigations as well as of positron emission tomography and magnetic resonance imaging studies on negative symptoms, warrant

further research efforts in this field to find, among others, a rationale of neuroleptic drug treatment in negative symptoms of schizophrenia.

11. *Treatment of Negative Symptoms in Schizophrenia*

There has been a strong debate in the literature on whether neuroleptics can really reduce negative symptoms, and especially whether they can reduce primary negative symptoms and not just secondary negative symptoms. An earlier review found evidence that some neuroleptics, especially some atypical neuroleptics, can reduce negative symptoms. However, since most of the trials were carried out in patients with negative and positive symptoms during acute exacerbations of schizophrenic psychoses, the question arises whether only those negative symptoms were reduced that were secondary to positive symptoms. In addition, the overlap between negative symptoms of schizophrenia and extrapyramidal side effects and depression has to be taken into account. As an alternative hypothesis to the direct effect of a drug on negative symptoms, the possibility that a drug reduces depression, akinesia, sedation, etc., and in this way also reduces negative symptoms, must be considered seriously.

One possibility of differentiating between direct and indirect drug effects on negative symptoms is to use statistical procedures. Formerly, only two studies had used an elementary statistical approach to account for the possibility of indirect effects on negative symptoms. Meltzer (1985) used a covariance analysis and found that the neuroleptic effect on negative symptoms could not be explained via a neuroleptic effect on positive symptoms alone. Van Kammen et al. (1987) used a regression analysis which showed that

the neuroleptic effect on negative symptoms could be explained by the effect on positive symptoms.

More complex statistical approaches were not used until the mid-1990s. Möller et al. (1995) applied a complex statistical model, a path analysis, to the data of the North American Risperidone Study (Marder and Meibach 1994), including effects on positive, negative, and extrapyramidal symptoms to determine whether risperidone had a more potent direct effect on negative symptoms than haloperidol. The aim was to determine whether a (direct) effect on negative symptoms remains after the effects which are possibly caused through the reduction of positive and extrapyramidal symptoms have been excluded. The estimate from the path model showed that even after statistical control for indirect effects on secondary negative symptoms, risperidone improved negative symptoms to a greater extent than haloperidol.

See also: Depression; Differential Diagnosis in Psychiatry; Mental and Behavioral Disorders, Diagnosis and Classification of; Nosology in Psychiatry; Schizophrenia; Schizophrenia: Neuroscience Perspective; Symptom Awareness and Interpretation

Bibliography

- Andreasen NC 1982 Negative vs. positive schizophrenia. Definition and validation. *Archives of General Psychiatry* **39**: 789–794
- Angst J, Stassen H H, Woggon B 1989 Effect of neuroleptics on positive and negative symptoms and the deficit state. *Psychopharmacology* **99**: S41–S46
- Bandelow B, Muller P, Gaebel W, Kopcke W, Linden M, Muller-Spahn F, Pietzcker A, Reischies F M, Tegeler J 1990 Depressive syndromes in schizophrenic patients after discharge from hospital. *European Archives of Psychiatry and Clinical Neuroscience* **240**: 113–120
- Biehl H, Maurer K, Schubart C, Krumm B, Jung E 1986 Prediction of outcome and utilization of medical services in a prospective study of first onset schizophrenics. Results of a prospective 5-year follow-up study. *European Archives of Psychiatry and Clinical Neuroscience* **236**: 139–147
- Carpenter W T, Heinrichs D W, Alphas L D 1985 Treatment of negative symptoms. *Schizophrenia Bulletin* **11**: 440–452
- Crow T J 1980a Molecular pathology of schizophrenia: more than one disease process? *British Medical Journal* **280**: 66–8
- Crow T J 1980b Positive and negative schizophrenic symptoms and the role of dopamine. *British Journal of Psychiatry* **137**: 383–6
- Crow T J 1985 The two-syndrome concept: origins and current status. *Schizophrenia Bulletin* **11**: 471–86
- Feighner J P, Woodruff-R A, Winokur G, Robins E, Guze S B, Munoz R 1972 Diagnostic criteria for use in psychiatric research. *Archives of General Psychiatry* **26**: 57–63
- Gaebel W 1989 Indikatoren und Prädiktoren schizophrener Krankheitsstadien und Verlaufsausgänge [Indicators and predictors of stages in schizophrenic illness and course outcomes]. Thesis, Free University of Berlin
- Gaebel W 1990 Erfassung und Differenzierung schizophrener Minussymptomatik mit objektiven verhaltensanalytischen Methoden [Recording and differentiation of schizophrenic negative symptoms with objective behavior analytical methods]. In: Möller H J, Pelzer E (eds.) *Neuere Ansätze zur Diagnostik und Therapie schizophrener Minussymptomatik* [Newer approaches to the diagnosis and treatment of schizophrenic minus symptoms]. Springer, Berlin, Heidelberg, New York, pp. 79–92
- Jackson J H 1887 Remarks on evolution and dissolution of the nervous system. *Journal of Mental Science* **23**: 25–48
- Klosterkötter J 1990 Minussymptomatik und kognitive Basis-symptome [Minus symptoms and basic cognitive symptoms]. In: Möller H J, Pelzer E (eds.) *Neuere Ansätze zur Diagnostik und Therapie schizophrener Minussymptomatik* [Newer approaches to the diagnosis and treatment of schizophrenic minus symptoms]. Springer, Berlin, Heidelberg, New York, pp. 15–24
- Marder S R, Meibach R C 1994 Risperidone in the treatment of schizophrenia. *American Journal of Psychiatry* **151**: 825–35
- McCreadie R G 1989 The Scottish first episode schizophrenia study. VII. Two-year follow-up. *Acta Psychiatrica Scandinavica* **80**: 597–602
- Meltzer H Y 1985 Dopamine and negative symptoms in schizophrenia: Critique of the type I-II hypothesis. In: Alpert M (ed.) *Controversies in schizophrenia*. Guilford, New York, pp. 110–36
- Meltzer H Y, Zureick J 1989 Negative symptoms in schizophrenia: A target for new drug development. In: Dahl S G, Gram L F (eds.) *Clinical pharmacology in psychiatry*. Springer-Verlag, Berlin, pp. 68–77
- Möller H J, von Zerssen D 1986 *Der Verlauf schizophrener Psychosen unter den gegenwärtigen Behandlungsbedingungen* [The Course of Schizophrenic Psychoses under the Current Treatment Conditions]. Springer-Verlag, Berlin Heidelberg New York
- Möller H J, van Praag H M, Aufdembrinke B, Bailey P, Barnes T R, Beck J, Bentsen H, Eich F X, Farrow L, Fleischhacker W W, Gerlach J, Grafford K, Hentschel B, Hertkorn A, Heylen S, Lecrubier Y, Leonhard J P, McKenna P, Maier W, Pedersen V, Rappard A, Rein W, Ryan J, Sloth Nielsen M, Stieglitz R D, Wegener G, Wilson J 1994 Negative symptoms in schizophrenia: Considerations for clinical trials. *Psychopharmacology* **115**: 221–8
- Möller H J, Müller H, Borison R L, Schooler N R, Chouinard G 1995 A path-analytical approach to differentiate between direct and indirect drug effects on negative symptoms in schizophrenic patients. A re-evaluation of the North American risperidone study. *European Archives of Psychiatry and Clinical Neuroscience* **245**: 45–9
- Möller H J, Bottlender R, Groß A, Hoff P, Wittmann J, Wegner U, Strauß A in press The Kraepelinian dichotomy: Preliminary results of a 15-year follow-up study on functional psychoses: Focus on negative symptoms. *Schizophrenia Research*
- Pogue-Geile M F, Harrow M 1985 Negative symptoms in schizophrenia: Their longitudinal course and prognostic importance. *Schizophrenia Bulletin* **11**: 427–39
- Ragin A B, Pogue-Geile M, Oltmanns T F 1989 Poverty of speech in schizophrenia and depression during in-patient and post-hospital periods. *British Journal of Psychiatry* **154**: 52–7
- Rosen W G, Mohs R C, Johns C A, Small N S, Kendler K S, Horvath T B, Davis K L 1984 Positive and negative symptoms in schizophrenia. *Psychiatry Research* **13**: 277–84
- Rösler M, Hengesbach G 1990 Negative Symptom in AMDP-System [Negative symptoms in the AMDP system]. In: Baum

- U, Fähndrich E, Stieglitz R D, Woggon B (eds.) *Veränderungsmessung in Psychiatrie und klinischer Psychologie [Measurement of Changes in Psychiatry and Clinical Psychology]*. Profil, Munich, pp. 329–39
- Schneider K 1959 *Clinical Psychopathology*. Gune & Stratton, New York
- Spitzer R L, Endicott J E, Robins E 1978 *Research Diagnostic Criteria (RDC) for a Selected Group of Functional Disorders*, 3rd edn. New York State Psychiatric Institute, New York
- Strauss J S, Carpenter Jr W T, Bartko J J 1974 The diagnosis and understanding of schizophrenia. Part III. Speculations on the processes that underlie schizophrenic symptoms and signs. *Schizophrenia Bulletin* **11**: 61–9
- van Kammen D P, Hommer D W, Malas K L 1987 Effect of pimozide on positive and negative symptoms in schizophrenic patients: Are negative symptoms state dependent? *Neuropsychobiology* **18**: 113–17
- Wagman A M I, Heinrichs D W, Carpenter Jr W T 1987 Deficit and nondeficit forms of schizophrenia: Neuropsychological evaluation. *Psychiatry Research* **22**: 319–30
- Wing J K 1989 The concept of negative symptoms. *British Journal of Psychiatry* **155**: 10–14

H.-J. Möller

Psychiatric Care, Economics of

A discussion of the economics of mental health care cannot take place without considering the epidemiological characteristics of mental illness and its cost to society.

1. Epidemiology of Mental Illness

1.1 The Burden of Mental Illness

Mental disorders are among the most prevalent and disabling illnesses worldwide. Unipolar major depression is the leading cause of disability globally in 1990, both in high-income and in low- and middle-income countries. It has been estimated that depressive and anxiety disorders account for 20–30 percent of all primary care visits worldwide. Approximately 29.5 percent of the US population is estimated to experience a diagnosable mental or addictive disorder in a 12-month period, which is comparable to the 20 percent rate for cardiovascular disorders (Kessler et al. 1994). Prevalence estimates of schizophrenia from studies in low-income countries vary from 0.8 per thousand reported from rural China to 5.9 per thousand in Calcutta, India. There were about 23 million people with schizophrenia in the world in 1985, three quarters in the less developed countries. Severe and persistent mental illnesses (as defined by diagnosis, disability, and duration) affect about 4 percent of the

population each year. These disorders are long lasting, produce significant level of disability, reduce an individual's ability to function in major social roles, and interfere with the creation and maintenance of social networks (Desjarlais et al. 1995).

These national and international statistics do not reflect the enormous toll of misery from mental disorders. Neuropsychiatric conditions have been ignored for a long time as they are absent from cause of death lists. However, disease priorities change dramatically as measurement of disease burden shifts from simple mortality indicators to indicators that incorporate disability. When disease-burden measurement includes time lived with disability, several of the neuropsychiatric disorders become a leading cause of disease burden worldwide. The World Bank estimates that neuropsychiatric diseases account for 6.8 percent of the Global Burden of Disease (GBD), measured in disability adjusted life years (DALYs), a toll greater than that exacted by tuberculosis, cancer, or heart disease. For adults aged 15–44 in developing economies, neuropsychiatric diseases account for 12 percent of the GBD. When intentional, self-inflicted injuries are added, the total constitutes 15.1 percent of the GBD for women and 16.1 percent for men. Because of the limited mortality consequences, this burden has been previously underestimated (World Bank 1993).

According to a 1999 estimate, neuropsychiatric conditions account for 10 percent of the burden of disease measured in DALYs in low and middle-income countries and 23 percent of DALYs in high-income countries. In comparison, cardiovascular diseases are responsible for 10 percent of DALYs in low and middle-income countries and 18 percent of DALYs in high-income countries. In high-income countries, one out of every four DALYs is lost to a neuropsychiatric condition. Of the ten leading causes of disease burden in young adults (in the 15–44 year age group) four are neuropsychiatric conditions. Alcohol dependence, bipolar disorder, and schizophrenia are among the leading causes of disease burden in this age group in 1998 (World Health Organization 1999).

1.2 The Social and Treatment Costs of Mental Illness

Society pays a high price for mental and addictive disorders—in both direct and indirect costs. Indirect costs (lost productivity, lost earnings due to illness or premature death, accidents, absenteeism) are twice the direct costs of care, adding to the actual economic impact of mental and addictive disorders. The total costs for the United States for 1990 for all mental disorders was \$148 billion (Rice et al. 1990). The direct cost of care for all mental disorders was \$67 billion, which represents 10 percent of the total \$670 billion direct cost of all health care in the United States

in 1990 (Levitt et al. 1991). In comparison, the total cost for all cardiovascular system diseases in US for 1990 was \$159 billion. More recent estimates show the full economic cost of alcohol and drug abuse in USA in 1992 to be \$246 billion. The cost burden from depression is about the same as that from heart disease. In the experience of one Fortune 100 company, mental health was ranked number two in the company's 10 most costly diseases, topped only by ischemic heart disease in terms of direct cost. If the same 10 diseases were ranked according to the full burden of disease, indirect costs drove mental health to the number one position and ischemic heart disease to the number 10 position (NAPHS 1999).

Spending on treatment is concentrated on those people with the most disabling conditions. Although individuals with severe and persistent mental illness represent only 5–10 percent of all the individuals who have mental illness, they account for approximately 40 percent of the specialty care expenditures (Steinwachs et al. 1992). Another estimate shows that 30 percent of the spending on mental health and substance abuse care is accounted for by 5 percent of the users of care (Frank et al. 1994). The total cost of severe and persistent mental illness in 1990 was estimated to be \$74 billion (accounted for by the dollar costs of treatment, lost productivity, shortened lives, and the costs incurred in the criminal justice and social services systems), with \$44 billion of the total being indirect cost (Rice et al. 1990).

However large, the dollar cost of mental illnesses does not account for the enormous emotional cost and pain that these illnesses exert not only on the affected individual but also on their family and on society. Mental disorders are often chronic conditions that create substantial disability. It is estimated that approximately 20–25 percent of the single adult homeless population suffers from some form of severe and persistent mental illness (Jencks 1994). Mental illness is correlated with other costly social problems: unemployment, crime and violence (Steadman et al. 1998), child abuse and neglect (Kelleher et al. 1994), and auto accidents (Rice et al. 1990).

1.3 Utilization of Services

It is not clearly established what proportion of individuals with mental disorder need, or want, treatment. Unlike other physical illnesses, not everyone with diagnosable mental illness perceives a need for treatment, and not all who seek treatment have a diagnosable disorder. Thus the under treatment and the over treatment of mental health problems can coexist within the same system (Regier et al. 1998).

In the United States, of the 30 percent of the population that has a mental illness at some point during the year, only 17.3 percent get some treatment in the health care sector with an additional 7.4 percent getting their only treatment from a human service

agency or a self-help group. Thus, only about 25 percent of those with a diagnosable condition get some form of treatment over a twelve-month period. Low rates of treatment characterize the most severely ill. Only 36 percent of the people diagnosed with bipolar disorder or major depression in a year are treated in any health sector, only 25 percent of those with substance abuse are treated, and 57 percent of the patients with schizophrenia are treated. At the same time 4.5 percent of individuals with no disorder, as assessed with a diagnostic instrument, receive mental health care (Kessler et al. 1997). According to Frank and McGuire (1999) nearly 38 percent of the users and 28 percent of all visits for mental health are not associated with a diagnosable disorder.

Two conclusions particularly relevant to the economic analysis of mental health care can be drawn from the above data on the epidemiology of mental illness and its cost to the society. First, mental and addictive disorders are prevalent and disabling and associated with a variety of social costs. Second, the social costs of mental and addictive disorders are concentrated in the 4 percent of the population that experiences severe and persistent mental illness, making this subgroup of people with mental illness 'undesirable' to insurers, employers, and neighbors.

2. The Market for Mental Health Care Services

Economics is the science and art of the distribution and allocation of scarce resources. The allocation of resources and rationing of scarce goods and services are guided by the market economy prices. The appropriate price for a product is determined by the relationship between *demand* and *supply* for this product. In the context of the market economy, health care in general and mental health care in particular can be conceptualized as a product. The classic relationship between the demand and the supply can be applied to mental health care as well. However, mental health care does not conform to the traditional rules of a marketplace. Several factors account for this. The nature of mental illness, the use of treatment services, and the supply side of the market differ from the general health sector. Plus, there are a greater variety of organizations and professionals with different theoretical orientation that supply mental health services. These include the specialty mental health sector (psychiatrists, psychologists, psychiatric social workers, psychiatric nurses, and counselors), the general medical sector (primary care physicians), self-help groups, and the clergy. Despite efforts to quantify mental health services and make them more objective, there is greater clinical uncertainty compared to the rest of medical care. Many mental disorders affect a person's ability to make decisions in their own best interest. Markets for other goods normally consist of a straightforward transfer of goods between the provider and the consumer. However, in the health sector,

in addition to this arrangement, funds from patients often flow through an intermediate funding body (either the government or an insurance agency). Market economics in mental health care are strongly influenced by *externalities*, that is the value that society derives from the health of its individual members.

Health care markets suffer from some kind of *market failure* but the mental health care market is subject to more failure than most. The most common market failures affecting the health care sector are due to lack of incentive to supply 'public goods,' neglect of externalities, and absence of well informed consumers. Failures in the market can be explained also by *adverse selection* (persons more likely to use coverage choose better coverage) and *moral hazard* (those who are insured will tend to overuse insured services). Evidence suggests that the forces of adverse selection and moral hazard work are at work in mental health (Frank and McGuire 1999). The special effort that is needed to reduce moral hazard and adverse selection in mental health care markets is one of the reasons why mental health services are organized and paid for differently than other types of health care (for review see Sharfstein et al. 1992, 1995).

3. Financing of Mental Health Services

Health care in general and mental health care in particular can be financed by public or private sources, the funds can be managed by public or private entities. Funds can be raised through taxes, mandates, private health insurance, out-of-pocket payments, grant assistance, charitable contributions, and borrowing. Funds can be managed by government, for-profit or not-for-profit private organizations, or consumers. Funds are then used to purchase health services that can be publicly or privately provided. A number of issues related to both public and private sources of financing are predicated on government's goals (allocational, distributional, and economic) and the policies utilized to correct market failures and instabilities.

3.1 Relationship Between Finance and Provision of Mental Health Care

It is important to distinguish between financing and provision of services. There is no absolute connection between the way mental health care is paid for and the way it is delivered. Public finance does not imply public provision. It is possible for services to be publicly financed yet privately provided or vice versa. Three major forms of relationships exist between the providers of mental health care and those who finance it, reimbursement, contract, and integrated approaches. Under the reimbursement approach, providers receive retroactive payments for services that they have supplied. Reimbursement can be found in systems with multiple private and public insurers and

multiple suppliers as in the United States. In low- and middle-income countries reimbursement is rarely combined with public finance with the exception of Chile. Contract involves an agreement between third-party payers (insurers) and health care providers. It is found in social insurance systems. Preferred provider organizations (PPOs) in the United States also use this form. In integrated health care systems the same agency controls the funding and the provision of health services. Medical providers are paid salaries and the budget is the main instrument for allocating resources. Integrated public systems are used in the Nordic countries and in developing countries as well. Health maintenance organizations (HMOs) in the United States are an example of an integrated private system. In countries with national health systems, the purchaser of care and the provider are commonly one and the same, namely, the government. In Europe there has been increasing interest in separating provider and financing functions in national health systems in order to promote competition.

3.2 Inequities in the Financing of Mental Health Care Services

A key characteristic of the mental health care service system is the inequity in financing mental health services compared to the coverage for other disorders. Resources available to mental health care have been eroding over the past decade (Buck and Umland 1997). Discriminatory limits on treatment of mental disorders characterize both the public and private sector. This discrimination against mental illness is based on arbitrary and often artificial distinctions between mental and physical illness.

3.3 Insurance and Managed Care

Health insurance is a system in which the prospective consumers of care make payment to a third party, which in the event of future illness will pay for some or all of the expenses incurred. Health insurance is a mixed source of finance, as it often draws contributions from both employers and employees, and sometimes from government. The existence of risk is the fundamental rationale for insurance. Insurance reduces risk by improving the predictability of adverse events through the pooling of a large number of similar risks. The few who become ill are covered for care by the many who remain well. Thus, risks that are unpredictable for the individual become predictable for the group and can be estimated.

Insurance markets are characterized by inherent instability. Adverse selection occurs in private insurance markets. Private insurance markets attempt to counter adverse selection by excluding the riskier customers and adjusting premiums which undermines the value of insurance altogether. This is one of the main arguments in favor of *public insurance* which can

be made universal and force everyone to share the risks. Cost sharing (deductibles, copayment), limits of benefits, total expenditures limits on policies, all are ways to deal with moral hazard. Despite the use of these mechanisms, the increased cost of health care and mental health care in particular has led to the introduction of other cost-containment mechanisms known as *managed care*. Managed care includes various organizational arrangements, such as HMOs and PPOs in the United States. Managed care organizations integrate the three functions of managing, financing and providing care. Managed care also includes financing mechanisms (capitation and risk-sharing contracts) that have strong incentives to reduce the amount of mental health care provided. The essence of managed care is its ability to ration care without imposing undue financial risk on consumers. Managed care principles can be utilized by public and private financing entities (see Mechanic 1997, 1998).

Managed care mechanisms are intended to affect rationing through establishment of a network of selected providers, contracts with providers that include incentives to limit care, introduction to the concept of medical necessity, and directing individuals to levels of care. Managed care plans also use various strategies to select good (profitable) risks from an insurance pool thus undermining the very existence of insurance and precluding effective risk pooling (Cutler and Zeckhauser 1998). They may prevent or discourage high-cost individuals from joining the plan or distort the services they provide in order to attract good and deter the bad risks. Supply-side management mechanisms to reduce costs include preauthorization requirements, utilization review, case-management for high utilizers, reducing access to specialty care providers, and substituting low cost providers for high cost providers.

Managed care has been shown to reduce overall mental health and substance abuse costs and service utilization. However, there has been concern about how the cost-containment strategies affect the access to care and the quality of care for patients. Cost-containment strategies can lead to excessive denial of care and under utilization of services, thus limiting access to care. Financial incentives to providers to reduce specialty referrals, hospital admissions, and length and amount of treatment can ultimately contribute to lowered quality of care.

Some demand-side cost-containment mechanisms, such as higher deductibles, copayments, visit- and day-limits, used by managed care can be reduced by introducing parity for mental health services with all other medical services. However, the introduction of parity generates incentives for insurance plans to replace the demand side cost-containment mechanisms with such mechanisms on the supply side. According to Frank and McGuire (1998) parity in insurance benefits for mental health care is not sufficient to guarantee equality in access to mental

health services if managed care-rationing devices such as 'medical necessity' denials cannot be controlled. Nevertheless, introducing parity for mental health services is a step towards overcoming the discrimination of mentally ill, reducing the financial barriers to treatment of mental disorders and improved access to services. Parity in private insurance benefits may result in some public-to-private sector cost shifting.

3.4 Private Financing

Private financing can be direct (personal payments) or indirect (payments for health services by employers). Private insurance, out-of-pocket payments for direct purchase of services, medical savings accounts, and charitable contributions are all sources of private financing of health services. Private financing is an essential part of health care financing given government's limited ability to raise enough resources to finance health services. Reliance on private financing is a necessity in many low-income countries.

The combination of price competition and the ownership incentives that the private sector offers may contribute to improved efficiency in the sector as a whole. However, instabilities in private insurance markets necessitate public financing of services or, in some cases, government regulation of private insurance markets. Except for a few countries (Switzerland, the United States) private health insurance is a minor source of finance. In developing countries this is because most of the population cannot afford private insurance. In high-income countries it is because the state has assumed most of the insurance function (Schieber and Maeda 1997).

No industrial country relies exclusively on free markets to produce and allocate health care in general and mental health care in particular. Government intervention is justified, in order to improve the efficiency and equity with which the private market produces and allocates its services. Public regulation of private markets is also essential to ensure effective risk pooling, affordability, informed choice, and continuity of coverage.

3.5 Public Financing

The main argument for public financing of health services is the achievement of universal access. This argument is based on efficiency (arising from market failures of the private market) and equity considerations (financing should be according to the ability to pay, and the distribution according to need). Public financing of health services can be of two types, a national health-service approach and social insurance funds. In both models collective risk pooling is achieved through compulsory taxation. There are a few major differences between the two systems.

National health services are usually financed by general taxes and other public revenue sources and thus they are subject to annual budget processes. Social insurance funds are generally financed through payroll taxes and tend to be more independent of annual political machinations. Public finance accounts for 75 percent of health care financing in high-income countries with the exception of The United States where public finance covers about 50 percent of the overall health care financing. In developing countries public finance is less important and accounts for about 20 percent of the total health care financing with out-of-pocket payments accounting for most of it. The composition of public finance varies in high-income countries. France, Germany and The Netherlands rely on social insurance while Canada, the Nordic countries and the United Kingdom use general taxation. United States operates both kinds of publicly financed systems (Johnson and Musgrove 1997).

3.6 Financing of Services for Severe and Persistent Mental Illness

Direct public services for the most severely ill people is a common feature of health systems that may otherwise be organized to provide and pay for other health care with a range of approaches (Hollingsworth 1992). The public mental health system in the United States, Germany, Canada, UK, and France are quite similar despite vast differences in their approaches to health care financing. The responsibility for mental health care is assigned to sub-national government and tends to rely on local tax funding. The role of local government-provided mental health care tends to emphasize providing care to the poor and disabled. The commonality can be explained partially by the fact that provision of mental health care has a public safety component and, therefore, the locality has a greater interest in assuming more direct control over the delivery of certain forms of mental health care to fulfill its obligation of public protection. Most countries have laws that tend to favor society's interests (externalities) over individual liberties of the mentally ill. This can explain partly the greater role of the government in provision of mental health care, including public hospitals and clinics.

Treatment programs aimed at people with severe and persistent mental illness are concerned not only with the acute care of symptoms but address also housing needs, income support, rehabilitation, social contacts, and social control of the affected individuals. The nature of severe and persistent mental illness requires coordination of services from various providers. Many of the successful innovations in the treatment of this group of people with mental illness involve new ways of organizing and coordinating the various elements of care.

4. Setting Priorities and the Role of the Government

There is little consensus or empirical evidence as to the right level of financing for health care and for mental health care in particular. Policymakers face the perpetual challenge of raising sufficient revenue for the health sector in an equitable and efficient way. Decisions about health spending cannot be isolated from a country's political, social, and economic characteristics. People's values, the responsiveness of the political system to public preferences, and the resources available are all reflected in the pattern of financing of health care. In the context of scarcity, mental health services are seldom given high priority by national governments or by international aid programs. Despite changing public attitudes towards mental illness and trends towards integration of the mentally ill into the mainstream of the health care system, the fact remains that, in terms of public and private policy, mental illness and substance abuse are not treated the same as other illnesses. In most countries today priorities are set in ways that exclude large numbers of people from access to organized care. Ability to pay appears increasingly to be the mechanism rationing access to care. In middle and upper income countries, health-financing policy is driven frequently by the need to increase coordination, reduce fragmentation, and exert better control over total health care costs. Countries in this group are often worried that their level of health spending will threaten economic growth and competitiveness by making their labor force, and therefore goods and services, more expensive. If universal access to affordable and effective mental health care is to be achieved, priorities have to be agreed upon and implemented. Choosing public priorities is how economic reality becomes an integral part of mental health system development and reform.

See also: Health Economics; Health Insurance: Economic and Risk Aspects; Psychology and Economics; Psychopharmacotherapy, Costs of

Bibliography

- Anonymous. Health Care Reform for Americans with Severe Mental Illness: Report of the National Advisory Mental Health Council 1993. *American Journal of Psychiatry* **150**: 1447-65
- Buck JA, Umland B 1997 Covering mental health and substance abuse services. *Health Affairs* **16**: 120-26
- Cutler D, Zeckhauser R 1998 The anatomy of health insurance. In: Newhouse JP, Culyer AY (eds.) *Handbook of Health Economics*. North Holland Press, Amsterdam
- Desjarlais R, Eisenberg L, Good B, Kleinman A 1995 *World Mental Health. Problems and Priorities in Low-Income Countries*. Oxford University Press, New York

- Dorwart R A, Epstein S S 1993 *Privatization and Mental Health Care. A Fragile Balance*. Auburn House, Westport, CT
- Frank R G, McGuire T G 1998 The economic functions of carve-outs in managed care. *American Journal of Managed Care* 4 (suppl.): SP31-SP39
- Frank R G, McGuire T G 1999 Economics and mental health. In: Newhouse J P, Culyer A Y (eds.) *Handbook of Health Economics*. North Holland Press, Amsterdam, pp. 10
- Frank R G, McGuire T G, Regier D A, Manderscheid R, Woodward A 1994 Paying for mental health and substance abuse care. *Health Affairs* 13: 337-42
- Hollingsworth E J 1992 Falling through the cracks: Care of the chronically mentally ill in the United States, Germany, and the United Kingdom. *Journal of Health Politics, Policy and Laws* 17: 899-928
- Jencks C 1994 *The Homeless*. Harvard University Press, Cambridge, MA
- Johnson B, Musgrove P 1997 Government financing of health care. In: Schieber G J (eds.) *Innovations in Health Care Financing*. World Bank, Washington, DC
- Kelleher K, Chaffin M, Hollenberg J, Fischer E 1994 Alcohol and drug disorders among physically abusive and neglectful parents in a community-based sample. *American Journal of Public Health* 84: 1586-90
- Kessler R C, Frank R G, Edlund M, Katz S J, Lin E, Leaf P 1997 Differences in the use of psychiatric outpatient services between the United States and Ontario. *New England Journal of Medicine* 336: 551-7
- Kessler R C, McGonagle K A, Zhao S, Nelson C B, Hughes M, Eshleman S, Wotchen H U, Kendler K S 1994 Lifetime and 12-month prevalence of DSM-III-R Psychiatric Disorders in the United States. *Archives of General Psychiatry* 51: 8-19
- Levitt K R, Lazenby H C, Cowan C A, Letsch S W 1991 National health care expenditures. *Health Care Financing Review* 13: 29-54
- Mechanic D 1997 Muddling through elegantly: finding the proper balance in rationing. *Health Affairs* 16: 83-92
- Mechanic D (ed.) 1998 *Managed Behavioral Health Care: Current Realities and Future Potential*. Jossey-Bass Publ., San Francisco
- NAPHS 1999 Trends in behavioral healthcare systems: a benchmarking report. *1998 Annual Survey Report*. The National Association of Psychiatric Health Systems, Washington, DC
- National Advisory Mental Health Council 1998 *Parity in financing mental health services, managed care effects on cost, access and quality* 1998. US Department of Health and Human Services, NIH Publication 98-4322
- Regier D A, Kaelber C T, Rae D S, Farmer M E, Knauper B, Kessler R C, Norquist G S 1998 Limitations of diagnostic criteria and assessment instruments for mental disorders: Implications for policy. *Archives of General Psychiatry* 55: 109-115
- Rice D et al. 1990 *The Cost of Alcohol and Drug Abuse and Mental Illness*. USGPO, Washington, DC
- Schieber G, Maeda A 1997 A curmudgeon's guide to financing health care in developing countries. In: Schieber G J (eds.) *Innovations in Health Care Financing*. World Bank, Washington, DC
- Sharfstein S S 1991 Prospective cost allocations for the chronic schizophrenic patient. *Schizophrenia Bulletin* 17: 395-400
- Sharfstein S S, Goldman H H, Arana J 1992 The market for mental health care: New rules of reimbursement and new delivery systems. In: Talbott J A, Hales R E, Keill S L (eds.) *Textbook of Administrative Psychiatry*, 1st edn. American Psychiatric Press, Washington, DC pp. 91-115
- Sharfstein S S, Webb W L, Stoline A M 1995 Economics of psychiatry. In: Kaplan H I, Sadock B J (eds.) *Comprehensive Textbook of Psychiatry/VI*, 6th edn. William & Wilkins, pp. 2678-89
- Steadman H J, Mulvey E P, Monohan J, Robbins P C, Appelbaum P S, Grisso T, Roth L H, Silver E 1998 Violence by people discharged from acute psychiatric inpatient facilities and by others in the same neighborhood. *Archives of General Psychiatry* 55: 393-401
- Steinwachs D M, Kasper J D, Skinner E A 1992 Patterns of use and costs among severely mentally ill people. *Health Affairs* 11: 178-85
- World Health Organization 1993 *Evaluation of Recent Changes in the Financing of Health Services*. Report of a WHO Study Group 829. World Health Organization, Geneva, Switzerland
- World Health Organization 1999 *The World Health Report 1999 - Making a Difference*. World Health Organization, Geneva, Switzerland
- World Bank 1993 *World Development Report 1993: Investing in Health*. Oxford University Press, New York

S. S. Sharfstein and S. Tyutyulkova

Psychiatric Disorders: Culture-specific

1. Introduction: Development of the Concept and Definition

The concept of 'culture-bound syndromes' has been the focus of an ongoing debate in the field of transcultural or comparative cultural psychiatry between psychiatric universalists who interpret these conditions as cultural elaborations of universal neuro-psychological or psychopathological phenomena, and cultural relativists who see them as generated and expressive of distinctive features of a particular culture. The 8th edition of Kraepelin's textbook of psychiatry in 1909 was the first handbook describing clinical pictures which were later labeled 'culture-bound syndromes,' namely *latah*, *koro*, and *amok*. In the past the term 'exotic' was used to emphasize the non-Western character of culture-related conditions. It was not anticipated then that a disorder such as *anorexia nervosa* in young women would become common due to sociocultural factors operant in modern Western society, so that it has now been recognized as a Western culture-reactive syndrome (DiNicola 1990). The concept of culture-specific psychiatric disorders was originally introduced into psychiatric literature in the 1950s and 60s by Yap, who also made the first attempt to order what he called 'culture-bound reactive syndromes,' known under a great variety of folk names, in a diagnostic classification schema (Yap 1967).

A list of 168 so-called culture-bound syndromes was compiled in a glossary by Hughes (1985), while the *Diagnostic and Statistical Manual* of the American Psychiatric Association, 4th edition, contains a glossary with 25 entries. These listings also include colloquial and local names for some well-known ubiquitous conditions. The Diagnostic Criteria for Research of the 10th Revision of the *WHO International Classification of Diseases* list 12 examples of 'culture-specific disorders' and suggest diagnostic codes while also stating that these conditions cannot be easily accommodated in established psychiatric classifications. Indeed, attempts at a nosological classification of culturally related syndromes have never been quite satisfactory due to (a) the overlap of diagnostically relevant symptoms, and (b) the fact that the various indigenous terms for similar syndromes are charged with different culture-specific meanings and traditional interpretations that codetermine the illness behavior of the patient and the reaction of the human environment.

On the basis of the data to be presented in this article, culture-specific psychiatric disorders are defined as clusters of symptoms and behaviors that can be plausibly related to specific cultural emphases, or to specific sociocultural stress situations, which are typical of a particular population.

2. Local Folk Idioms of Distress

Local folk idioms of distress for psychologically stressful situations that can occur in any culture are not signifiers of culturally specific disorders although the manifestations are influenced by the local culture. *Susto*, *espanto*, *miedo* are colloquially equivalent Spanish terms implying an experience of fright or scare, widely used among Hispanic populations of the Western hemisphere for a great variety of symptoms (Rubel et al. 1985). The conditions so labeled frequently resemble neurotic or somatoform syndromes but may also be caused by other kinds of disorders, including organic pathology. The Spanish word *nervios* is often used among Hispano-American and Caribbean populations, as is the English term *nerves* among rural Anglo-American groups, to denote chronic dysphoric mood states with various somatic manifestations especially in postmenopausal women under psychosocial stress, as is commonly encountered also in other cultures. The Spanish term *ataque de nervios* is, among Hispanic populations of Central and North America and the Caribbean, applied in reference to women showing a brief episode of behavioral discontrol as immediate reaction to any emotionally traumatizing experience (Oquendo 1994). Middle-aged Korean women, like those in other cultures, may react to stressful life situations and ongoing marital or family problems, with depressed mood, feelings of

anger, and somatizing complaints, typically of a sensation of an epigastric mass. This chronic dysthymia is known among Koreans and Korean-Americans as *hwa-byung*, 'fire disease,' and traditionally treated by female shamans (Lin 1983).

3. Culturally Stereotyped Reactions to Extreme Environmental Conditions

Certain behavioral reactions to extreme environmental conditions are culturally stereotyped although they are not truly culture-specific. This is the case with conditions generally subsumed under the heading 'arctic hysteria,' a term used by outside observers for the often dramatic behavioral reactions shown by indigenous inhabitants of arctic and subarctic areas in stress situations in which the person affected experiences a temporary state of dissociated consciousness induced by acute anxiety (Foulks 1972). Physical deprivation, mineral and vitamin deficiencies, as well as the psychological stresses of surviving in an extreme climate, were adduced to explain these phenomena which have practically disappeared as the traditional Eskimo way of life has largely been abandoned.

Frequently cited in the ethnological literature are *pibloktoq* (Inuit, 'crazy') among Eskimo in Greenland, arctic Canada, Alaska, and Siberia, and *kayak-svimmel* (Danish, 'kayak dizziness') also called *kayak-angst*, among seal-hunting Greenlanders. *Pibloktoq*-type attacks usually affected younger women who after a prodromal phase of anxious and irritated behavior started imitating animal calls, then suddenly became agitated and, screaming loudly, rushed into the snow or even into icy water, albeit not too far off so that they were easily rescued. The motor excitation ended with collapse or sometimes with convulsive movements, followed by deep sleep. There was only a vague recollection of the event which, in traditional culture, was attributed to possession by evil spirits.

Attacks of *kayak-svimmel* occurred among seal hunters experiencing extreme isolation and sensory deprivation when sitting lonely and motionless in the kayak, surrounded by the monotonous whiteness of the arctic sea, waiting for a seal to appear. In this situation the hunters might suddenly fear the kayak was filling with icy water when they had a sensation of coldness in the legs. In a state of mounting anxiety they felt dizzy and unbalanced; were trembling and sweating, paralyzed with panic anxiety and subject to illusionary perceptions. With help coming in time they could reach shore safely but more intense attacks followed on other trips so that they had to give up kayak hunting.

In the past, frequently referred to as a culture-specific condition of the Algonquin Indian tribes of northern Canada, was the '*windigo* (*witiko*) psychosis,' described as an obsessive murderous desire for human

flesh under the delusion of being possessed by, or turning into, the mythical cannibalistic *Windigo* monster with a heart of ice, said to appear during severe subarctic winters. The struggle for survival in the harsh northern climate may have led to isolated instances of famine cannibalism but a critical examination of the many reports of 'windigo psychosis' could not confirm a single case of cannibalistic compulsion (Marano 1982). The term *windigo* is today used among the northern Ojibwa to denote a state of deep depression, without implication of cannibalistic impulses.

4. Syndromes related to Specific Cultural Emphases or Specific Socio-cultural Stress Situations

4.1 Syndromes Related to a Cultural Emphasis on Fertility and Procreation

4.1.1 'Genital shrinking' syndrome. The Malayo-Indonesian term *koro* and the Mandarin Chinese term *suo-yang* (Cantonese *suk-yang*) denote 'shrinking of the penis.' These terms refer to an acute state of panic anxiety prompted by the belief of the affected male person that his penis is shrinking and will disappear, and of the affected female person that her breasts and/or labia are shrinking. The 'genital shrinking' is believed not only to lead to impotence and/or sterility but to death once the organs are completely retracted. This syndrome came to the attention of European psychiatry in the late nineteenth century through reports by colonial physicians working in South East Asia and through the growing interest in traditional Chinese medicine. Chinese traditional medicine interprets *suo-yang* as yin-yang imbalance due to deficiency of the 'warm' male principle yang and excess of the 'cold' female principle yin, so that vitalizing yang remedies are indicated.

For the theorists of psychoanalysis, *koro* served as the concrete example of Oedipal castration anxiety, while some Chinese authors like Yap saw in *suo-yang* the paradigm of a true culture-bound syndrome, assuming that the disorder itself was generated by the suggestive effect of traditional Chinese concepts. However, it has since been ascertained that the typical 'genital shrinking' syndrome also occurs not infrequently among Asian and African populations without Chinese cultural influence. Complaints of genital shrinking are also occasionally verbalized by Western patients but these cases are different from the culturally related 'genital shrinking' syndrome as they are associated with other, usually chronic, neurotic, psychotic, or organic conditions; and the patient's ethnic-cultural group in Western society does not share the belief in the reality of genital disappearance and

ensuing death. *Koro/suo-yang* is the only culturally related syndrome that has occurred in major epidemic outbreaks. Epidemics of 'genital shrinking' have been repeatedly reported in recent decades (Jilek 1986). Outbreaks in Singapore, Thailand, North East India, and South China were precipitated by collective anxieties over perceived threats to the security or even survival of the afflicted population; in Nigeria by fears of magic robbing of the genitals.

4.1.2 'Semen loss' syndrome. Known in the traditional medicine of China as *shen-k'uei*, 'weakness of the kidneys,' in the Ayurvedic medicine of India as *dhat* (from Sanskrit *dhatu*, essential body substance) or *jiryan*, in Sri Lanka as *sukra prameha*, the complaint of 'semen loss' is not uncommon among young men of these countries in which a high premium is placed on reproductive capability and semen is considered a precious elixir of life. The presented symptoms have a hypochondriac quality; they are associated with an anxious and dysphoric mood state and typically attributed to the imagined loss of sperm in urine or to natural nocturnal emissions (Paris 1992). Sperm loss is assumed to cause the symptoms of general weakness, loss of energy, exhaustion, difficulty concentrating, sexual dysfunction, and rather vague somatic complaints. Many of these patients consult practitioners of traditional medicine as well as medical doctors; as immigrants they are also increasingly seen in European and North American clinics.

4.2 Syndromes Related to a Cultural Emphasis on Learnt Dissociation

4.2.1 Latah-type reactions. The peculiar *latah* reactions have been cited by cultural relativists as a paradigm of culturally generated behavior (Kenny 1978), and by biopsychological universalists as cultural elaborations of the neurophysiological startle reflex (Simons 1996). *Latah* reactions were interpreted by Gilles de la Tourette as a variant of his neurological syndrome, and by Kretschmer as a phylogenetic mechanism. The complete *latah* syndrome consists of an initial startle response provoked by a person of certain social status and triggered by a visual, acoustic, or tactile stimulus, the sensitivity to which is both individual and culture-specific. The startle response induces an altered state of consciousness, associated with repeated impulsive utterance of sounds or inarticulate words, often also swearing with sexual or obscene words ('coprolalia'). Characteristic are echopraxia, echomimia, and echolalia, namely the exact imitation of the provoking person's movements, facial and oral expression; even of words and sentences spoken in foreign languages. Automatic

obedience to commands may occur in hypnoid trance and serve the amusement of onlookers. The *latah* syndrome can sometimes culminate in a brief state of cataleptic rigidity. Because of the typical echo symptoms, *latah* reactions are known in French psychiatry as *névroses d'imitation*.

Automatic imitation behavior has also repeatedly been observed in normal indigenous people during first contact with Europeans, as already reported by Charles Darwin from Tierra del Fuego and more recently from remote areas of Melanesia (Gajdusek 1970) and East Africa (Jilek-Aall 1979). The typical *latah* syndrome occurs mostly in females over two geographic zones (Winzeler 1995): the South East Asian zone, encompassing Malaysia, Indonesia (*latah*, 'nervous, ticklish'), Thailand (*bah-tshi*, 'tickle-crazy'), Burma-Myanmar (*yaun*, 'ticklish'), the Philippines (*mali-mali*); the North Eurasian zone, encompassing the Ainu people of Hokkaido and Sakhalin (*imu*, 'startle'), aboriginal peoples in Mongolia and Siberia (Russian, *miryachit*), and the Sami of northernmost Europe ('Lapp panic'). The varied ethnic groups of these two *latah* zones have no culture traits in common other than (a) learnt dissociation which is practiced from early youth on in religious, social, and therapeutic ceremonies, trance rituals, and shamanic seances; (b) learning by imitation and coping by copying the behavior of persons who appear to be more powerful. Sporadic occurrence of *latah* symptoms has been reported among some males in Arabic-Islamic populations of North Africa and Yemen, and among South African Bantu people. While European cases are rarely encountered elsewhere, the endemic and familiar incidence of so-called 'jumping,' a reaction similar to *latah*, among males of French Canadian background in certain rural areas of Maine, New Hampshire, and Quebec, has attracted scientific interest since the 1870s.

4.2.2 Amok-type reactions. The term *amok* is originally derived from the Portuguese-Indian *amuco*, referring to heroic warriors ready to die in battle who are immortalized in Malayan epics, similar to the *berserkr* in ancient Norse sagas. Today, 'running amok'—likewise 'going berserk'—has become a colloquial label for any episode of apparently unprovoked randomly aggressive and homicidal behavior, usually engaged in by attention-craving angry young men in modern Western society. Although such behavior has certain features in common with the *amok* reaction observed among Asian *amok*-runners, it does not represent the specific dissociative syndrome described under this term (Pfeiffer 1994). Typical *amok* reactions are preceded by a dysphoric-tense mood state in the context of perceived or imagined social adversity or slight, loss of prestige or status, interpersonal or socioeconomic problems. A sometimes trivial event then triggers the actual *amok*

reaction. After a phase of meditative brooding an alteration of the state of consciousness ensues in which visual perception changes: the outside world turns into darkness or becomes red, threatening visions may appear, causing feelings of fear or rage. The affected person is suddenly seized by hypermotility and embarks on the *amok* run, a discharge of aggression leading to random destruction and killing, and eventually to suicide or suicide attempt. In case of survival, the attack is terminated by stuporous sleep; afterwards amnesia is claimed for actions committed.

Amok reactions have been presented as specific to the male ethos of Malayo-Indonesian culture. Such reactions also occur among related ethnic-cultural groups in the Philippines and in Laos. However, there are also typical *amok* cases in Papua New Guinea (Burton-Bradley 1975), besides the common non-violent pseudo-*amok*, the psychodramatic 'wild man behavior' of New Guinea highlanders. In historical analysis, *amok* in Southeast Asia appears to have changed from a glorified warrior ethos and means of social protest of the precolonial era, to a dissociative reaction, which in more recent times is mainly found in marginal figures or as an episode during the course of a chronic mental disorder (Murphy 1973).

4.3 Syndromes Related to a Cultural Emphasis on Presenting a Pleasing Physical Appearance: Taijin-Kyôfu Reactions

In contrast to the condition commonly diagnosed in Western psychiatry as social phobia, the Japanese form of anthropophobia, *taijin-kyôfu* ('fear in relation to others'), is not a phobic avoidance of social contacts in order to avoid unpleasant feelings for oneself; rather it is the fear that one's external appearance may be disturbing or offending to the other who as a rule is neither a stranger nor a close relation but a respected acquaintance. Symptoms of *taijin-kyôfu* were already mentioned in the eleventh century Genji saga by Lady Murasaki Shikibu, and detailed clinical descriptions suggest this condition to be typical of Japanese culture (Kimura 1995). However, this type of anthropophobia is also not uncommon in Korea (Lee et al. 1994). The clinical syndrome *taijin-kyôfu* was introduced into psychiatric literature in the 1920s by Shoma Morita as a subtype of the nervous temperament (*shinkeishitsu*) that he considered to be the main indication for his 'Morita Psychotherapy.' The syndrome is not uniform but occurs as one of several phobic anxiety manifestations: as erythrophobia (i.e., fear of blushing in front of others); as eye contact phobia (i.e., fear that others feel bothered when looked at); as body odour phobia (i.e., fear that one's own assumedly offensive body odour will bother others); as dysmorphophobia (i.e., fear of having an extremely unattractive, repulsive face or defective physique); or as fear that one

will grimace or otherwise behave embarrassingly. The sufferers tend to be convinced of their imagined defects to a degree that their phobias appear delusional, and the presence of other psychotic symptoms has to be ruled out in order to delimit *taijin-kyôfu* from a schizophrenic disorder. However, the anthropophobic patient's relation to the human environment is fundamentally different from that seen in chronic schizophrenic psychoses. *Taijin-kyôfu* patients are neither autistic nor paranoid; they suffer from a fear of presenting themselves.

4.4 Syndromes Related to Acculturative Stress

4.4.1 Brain fog. The term 'brain fog,' coined by Nigerian students, refers to a condition first reported by Prince (1960). The brain fog syndrome has since been recognized in a high percentage of Nigerian students at home and abroad, and also in other African countries (Liberia, Ivory Coast, Uganda, Tanzania, Malawi, Swaziland) in students exposed to the acculturative stress of a Western-type education system emphasizing theoretical book knowledge, quite different from the practical know-how and tribal lore acquired through oral traditions by older generations in Africa. It is noteworthy that subnormal intelligence, malnutrition, or physical disease do not account for the symptoms of brain fog which are also seen at educational institutions in Europe and North America among African students in good state of nutrition and general health. The brain fog syndrome is categorically different from the general study stress and examination crises experienced by students from Western or Asian cultures with centuries of literary tradition and education systems already long-based on reading and writing and on memorizing texts for formal examinations. The brain fog syndrome is also clinically different from the common complaints everywhere related to study stress: The mere effort of reading may cause blurred vision and flow of tears. There are also inability to concentrate on reading material and lectures, impairment of retention and memory, daytime fatigue and disturbed night sleep. Characteristic symptoms are the feeling of burning heat on head and body and the sensation of crawling as if caused by insects on, or by worms under, the skin.

4.4.2 Bouffée délirante-type reactions. Clinical findings as well as research data indicate that acute transient psychotic and psychosis-like reactions occurring in African and Afro-Caribbean populations are disorders directly related to culture and culture change, and are encountered with increasing frequency in Africa (Jilek and Jilek-Aall 1970). Such reactions are characterized by intense fear of magical persecution

or retribution and are usually triggered by an event or experience evoking acute anxiety in the context of beliefs in sorcery or witchcraft that persist or even increase under rapid sociocultural changes. *Bouffée délirante*-type reactions are of sudden onset and brief duration, they are associated with paranoid delusions, oneroid (i.e., dream-like) confusion, frequently also visual and auditory hallucinations, and often lead to dramatic and sometimes dangerous acting-out behavior.

Anglophone psychiatrists have described transient psychotic and psychosis-like reactions in Africans under various terms while their francophone colleagues use the diagnostic label *bouffée délirante*, which was introduced into French psychiatry in the 1880s by Magnan and later redefined by French-speaking authors in Africa in accordance with their clinical experience (Collomb 1965). The well-known Nigerian psychiatrist Lambo introduced the English term 'frenzied anxiety' and equated it with *bouffée délirante* (Lambo 1960). As repeatedly stated by Lambo and by other researchers, the main underlying causes for these states are not toxic-organic but sociocultural factors. Among these factors, of paramount importance is the stress exerted by rapid culture change which is leading to the social marginalization of many Africans in the process of 'modernization,' as recently also confirmed in Swaziland (Guinness 1992). Under situations of urbanization and Westernizing acculturative pressures the traditional communalistic society is disintegrating and the supportive kin network is breaking down. The individual then experiences an increasing economic rivalry and social isolation which intensifies the old fears of witchcraft and sorcery, never obliterated by Christianity, while the traditional protective and remedial resources are no longer readily available.

5. Culture-specific Disorders vs. Traditional Ritual Behavior

Culture-specific psychiatric syndromes have to be differentiated from extracurricular but not pathological behavior in the traditional rituals of non-Western cultures. A Western-trained clinician who attaches pathology labels to behaviors and ideas that are considered normal in a non-Western culture, commits a 'eurocentric fallacy.' As a general statement it can be said that the culture-specific disorders are not part of ceremonials in which ritualized altered states of consciousness are induced and practiced for religious or therapeutic purposes. While such institutionalized trance and 'possession' states constitute culture-congenial behavior in certain ceremonials, the culture-specific syndromes described above are considered as abnormality or illness also by the indigenous experts of the culture and are explained in terms of a particular folk etiology which determines traditional therapeutic management.

See also: Body: Anthropological Aspects; Cross-cultural Psychology; Eating Disorders: Anorexia Nervosa, Bulimia Nervosa, and Binge Eating Disorder; Obesity and Eating Disorders: Psychiatric

Bibliography

- Burton-Bradley B G 1975 *Stone Age Crisis: A Psychiatric Appraisal*. Vanderbilt University Press, Nashville, TN
- Collomb H 1965 Bouffées délirantes en psychiatrie africaine. *Psychopathologie Africaine* 1: 167–239
- DiNicola V F 1990 Anorexia multiforme: Self-starvation in historical and cultural context. *Transcultural Psychiatric Research Review* 27: 165–96, 245–86
- Foulks E F 1972 *The Arctic Hysterias of the North Alaskan Eskimo*. Anthropological Studies No. 10. American Anthropological Association, Washington, DC
- Gajdusek D C 1970 Physiological and psychological characteristics of stone age man. *Engineering and Science* 33: 26–33, 56–62
- Guinness E A 1992 Patterns of mental illness in the early stages of urbanization. *British Journal of Psychiatry* 160: Supplement 16: 12–72
- Hughes C C 1985 Glossary of 'culture-bound' or folk psychiatric syndromes. In: Simons R C, Hughes C C (eds.) *The Culture-bound Syndromes: Folk Illnesses of Psychiatric and Anthropological Interest*. Reidel, Dordrecht, pp. 469–505
- Jilek W G 1986 Epidemics of 'genital shrinking' (koro): Historical review and report of a recent outbreak in South China. *Curare* (Heidelberg) 9: 269–82
- Jilek W G, Jilek-Aall L 1970 Transient psychoses in Africans. *Psychiatria Clinica* 3: 337–64
- Jilek-Aall L 1979 *Call Mama Doctor: African Notes of a Young Woman Doctor*. Hancock House, Saanichton, BC
- Kenny M G 1978 Latah: The symbolism of a putative mental disorder. *Culture, Medicine and Psychiatry* 2: 209–31
- Kimura B 1995 *Zwischen Mensch und Mensch: Strukturen japanischer Subjektivität*. Wissenschaftliche Buchgesellschaft, Darmstadt
- Lambo T A 1960 Further neuropsychiatric observations in Nigeria. *British Medical Journal* 2: 1696–1704
- Lee S H, Shin Y C, Oh K S 1994 A clinical study of social phobia for 10 years. *Journal of the Korean Neuropsychiatric Association* 33: 305–12
- Lin K-M 1983 Hwa-byung: A Korean culture-bound syndrome? *American Journal of Psychiatry* 140: 105–7
- Marano L 1982 Windigo-psychosis: The anatomy of an emic-etic confusion. *Current Anthropology* 23: 365–412
- Murphy H B M 1973 History and the evolution of syndromes: The striking case of latah and amok. In: Hammer M, Salzinger K, Sutton S (eds.) *Psychopathology: Contributions from the Biological, Behavioral and Social Sciences*. John Wiley, New York, pp. 33–55
- Oquendo M 1994 Differential diagnosis of ataque de nervios. *American Journal of Orthopsychiatry* 65: 60–5
- Paris J 1992 Dhat: The semen loss anxiety syndrome. *Transcultural Psychiatric Research Review* 29: 109–18
- Pfeiffer W M 1994 *Transkulturelle Psychiatrie*, 2nd edn. Thieme, Stuttgart
- Prince R 1960 The 'brain fog' syndrome in Nigerian students. *Journal of Mental Science* 106: 559–70

- Rubel A J, O'Neill C W, Collado R 1985 The folk illness called susto. In: Simons R C, Hughes C C (eds.) *The Culture-bound Syndromes*. Reidel, Dordrecht, pp. 333–50
- Simons R C 1996 *Boo! Culture, Experience and the Startle Reflex*. Oxford University Press, New York
- Winzeler R L 1995 *Latah in Southeast Asia: The History and Ethnography of a Culture-bound Syndrome*. Cambridge University Press, Cambridge
- Yap P M 1967 Classification of the culture-bound reactive syndromes. *Australian and New Zealand Journal of Psychiatry* 1: 172–9

W. G. Jilek

Psychiatry and Sociology

Sociology and psychiatry have a longstanding relationship which has never been free of tensions and can at least be characterized as ambivalent. On the one hand, psychiatry as a subspecialty of medicine, which has to solve concrete problems in everyday life, has become a subject matter of sociology, which often criticizes psychiatry's concepts and practices. On the other hand, psychiatry has profited from cooperating with sociologists by making use of sociological research methods. The article sets out with describing the relationship between the two disciplines, and then proceeds to discussing the contributions made by sociology towards the understanding of how psychiatry functions ('sociology of psychiatry') before describing the uses of sociology within the field of psychiatry ('sociology in psychiatry').

1. Sociology and Psychiatry—An Uneasy Relationship

Sociology and psychiatry have a common topic: deviant behavior. Sociology, as a scientific discipline which examines the rules by which society functions, must have a genuine interest in phenomena which threaten these rules. Emile Durkheim (1895), one of the founding fathers of sociology, points out in his famous book, *The Rules of the Sociological Method*, that the standards and rules defining what is pathological help to reinforce the norms and values of society. For psychiatry, in turn, a specific form of deviant behavior, called mental illness, is its very 'raison d'être.' In this respect, the relationship between sociology and psychiatry could be regarded as a relationship of equal partners joining their intellectual forces in analyzing a topic of common interest, though using different approaches. However, there is also a different type of relationship between these two disciplines, in so far as one of them, psychiatry as a subspecialty of medicine, has become the subject matter of the other, sociology.

When discussing the relationship between sociology and psychiatry, a profound difference between these two disciplines has to be kept in mind: Psychiatry is not only a scientific discipline—like sociology—but also a practice field, akin to medicine in general, or to education. Sociologists might aim at changing things, but sociology has never been a major player in practical problem-solving, in contrast to psychiatry, which from its very beginning 200 years ago was assigned the societal task of treating the mentally ill or managing (controlling, as sociologists would say) the problem of mental disorders. Van der Geest (1995), in comparing medicine and social sciences (exemplified by cultural anthropology), speaks of two ‘ethnocentrisms,’ i.e., two different cultures, which are dissimilar in three aspects. The culture of medicine is practical and problem-oriented (doctors are supposed to find concrete solutions to concrete problems), whereas the social sciences culture is theoretical; doctors have no time to lose (they have to act promptly before it is ‘too late’), social scientists are not in a hurry to finish their work; finally, doctors measure their success in maintaining and restoring people’s health (they are accountable to people); for social scientists the fulfilment of their task does not lie in an improvement of the lives of the people studied, but in the production of texts about them. The ‘ethnocentrism’ of these two disciplines makes communication difficult and prone to misunderstandings.

Psychiatry’s relationship to sociology is ambivalent. On the one hand the application of sociological research methods to psychiatric problems have undoubtedly contributed to raise psychiatry’s societal visibility and importance—by furthering epidemiological research, which has shown that between one quarter and up to nearly half of the US population suffer from at least one diagnosable mental disorder during a person’s life time (e.g., Robins and Regier 1991, Kessler et al. 1994). On the other hand, sociologists have been, and still are, among the most radical critics of psychiatry, concerning both its concepts and its practice.

One further aspect renders the discussion of the relationship between sociology and psychiatry difficult: Both disciplines are extremely diverse within themselves. Sociology comprises a multitude of different philosophies (e.g., the ‘social constructivist’ and the ‘social realist’ perspectives) and research methods (e.g., the ‘quantitative’ and the ‘qualitative’ approaches). Psychiatry, both as a scientific discipline and a practice of identifying, explaining, and treating behavioral problems, which are deemed to be mental disorders, is even more heterogeneous, since psychological and sociological, as well as biological, approaches live under its roof. There is also a substantial overlap between each of the two disciplines and other fields. Sociology can be broadened to mean social sciences in general, and incorporate social anthropology, social psychology, and other disciplines. Fur-

thermore, fields such as epidemiology, public health, social medicine, health economics, and others, if applied to psychiatric problems, would have to come in. Also, no attempt can be made here to discuss the boundaries of psychotherapy, clinical psychology, and social psychiatry in relation to clinical psychiatry.

For the purpose of this review, which addresses a broad audience, the author—who is a psychiatrist with strong interests in sociology—has found the dichotomy of the relationship between sociology and medicine, proposed by Robert Straus (1957), useful. Following Straus’ distinction between ‘sociology of medicine’ and ‘sociology *in* medicine,’ the ensuing part of the text will be divided into the ‘sociology of psychiatry’ and the ‘sociology *in* psychiatry.’ The first approach discusses sociological contributions toward the understanding of psychiatry as a whole, i.e., of how psychiatry behaves in terms of developing and using concepts and applying them in practice. The second approach deals with the application of sociological methods within the field of psychiatry. While these two approaches are sometimes intertwined, their separation is nevertheless useful for the purpose of orientation. This distinction can, at least implicitly, also be found in two recent reviews of the sociology of mental health and illness, to which the reader of the present review is referred for more detailed discussions of the issues raised here (Pilgrim and Rogers 1999, Busfield 2000). It has to be noted, though, that ‘sociology of mental health and mental illness’ is a broader concept than the two sociologies (*of* and *in* psychiatry) discussed here. In the following two sections these two approaches are illustrated by examples.

2. *Sociology of Psychiatry*

The sociology of deviant behavior is central to sociology. Analyzing how boundaries are drawn between the normal and the abnormal, in other words, how the normal and the abnormal are classified, and analyzing how society reacts to the abnormal is one of the main interests of sociology. Since psychiatry has obtained an important definitional power over the specific type of deviant behavior called mental illness, and since it also has substantial influence on how such behavior is dealt with, it is obvious that psychiatry itself has become a subject matter of sociology. Sociologists have dealt both with psychiatric concepts and with the practice of psychiatry, with a special emphasis on the professional nature of psychiatry, on psychiatric services and the issue of legal vs. medical control of mental disorders.

2.1 *Psychiatric Concepts*

Sociologists have always expressed doubts about the coherence, credibility, and validity of psychiatric

concepts and knowledge. The most profound sociological criticism of psychiatry concerns the concept of mental illness and its classification into discrete diagnostic entities. While such criticism has come forward from many quarters, including the anti-psychiatric and consumer movements as well as scientific psychiatry itself (where some advocates favor dimensional rather than categorical concepts), the specific and original contribution from sociology was the labeling theory in the 1960s (Scheff 1999), a variant of what is called 'social constructivism,' i.e., a theory assuming that reality is a product of human definitions and not something waiting out there to be discovered (Brown 1995). More specifically, using 'symbolic interactionism' as their framework (a theory about how people map out social reality in everyday life), the advocates of 'labeling theory' claimed that the same deviant behavior may either go unnoticed or be labeled as mental illness, depending on the circumstances in which it occurs and is perceived by others. For instance, in a now famous study, Rosenhan (1973) observed that healthy test persons ('Pseudopatients') who presented themselves to psychiatric hospitals and told the doctor that they were hearing voices (saying 'empty,' 'hollow,' and 'thud') were admitted and received a psychiatric diagnosis. Once labeled, so the theory goes, the person's identity and social status is altered significantly. Part of the process is that the labeled person internalizes the new identity ascribed to them and becomes trapped in the new role of a mental patient.

A radical social constructivist perspective creates hostility among doctors, patients, and relatives, since they get the feeling that it denies the reality of suffering connected with mental disorders. While the idea that mental disorders, especially schizophrenia, are caused by labeling is more or less abandoned in recent thinking, labeling is certainly a powerful mechanism when it comes to discussing the role of stigmatization and discrimination (Link et al. 1987). A more moderate form of criticizing psychiatric diagnosis is the notion that through the diagnostic process illnesses are 'socially framed' (Rosenberg 1992). Mirowsky and Ross (1989), while acknowledging that psychological problems are real, criticize that psychiatric diagnosis treats attributes as entities, and that psychiatrists often speak of psychological problems 'as if they are discrete entities, entering the bodies and soul of helpless victims.' They target especially the operationalization of psychiatric diagnosis in modern diagnostic systems, such as the *Diagnostic and Statistical Manual III* (DSM-III) (and onwards) of the American Psychiatric Association (1994). It has to be acknowledged, though, that the construction of DSM-III was partly a reaction to the criticism that psychiatrists were not able to diagnose mental disorders reliably (Rosenhan 1973). However, while operationalization might have increased the reliability of psychiatric diagnosis, the question of validity is still debatable. Brown (1995) has

suggested that instances of medicalization of normal life problems (such as 'late luteal phase dysphoric disorder—LLPDD'; now called 'pre-menstrual dysphoric disorder—PMDD') are mainly a feature of professional 'expansionism.'

2.2 Psychiatry as a Profession

Sociologists have not only dealt with psychiatric concepts but also with the process of professionalization of psychiatry, which is linked closely to both conceptual issues and to the practice of psychiatry. The rise of the psychiatric profession during the nineteenth century has been regarded alternatively as having been due to its functional value for economic order and efficiency under capitalism (Scully 1979), and to its role for maintaining the moral, rather than the economic order (Foucault 1965).

Psychiatrists certainly have difficulties in agreeing with these views, since they have always seen themselves as primarily a helping profession, whose practice is based on scientific advances. There is a different strain of sociological analysis of psychiatry as a profession, going back to the German sociologist Max Weber, who postulated that professionals want to advance their own social status, persuade clients and potential clients about the need for the service they offer, corner the market in that service, and exclude competitors (Freidson 1970, Pilgrim and Rogers 1999). Other sociological approaches to the issue of professionalization include the sociology of deviance (crises in personal lives are reframed as mental illness), symbolic interactionism (regarding the professional power in negotiating the meaning of mental illness), and feminist sociology (men occupy higher-status positions within the more powerful professions).

The concept of power is central to this analysis, and the specific knowledge base of a profession (which can only be obtained by a difficult process of socialization) gives the profession power over its clients, its recruits and related occupational groups. With the emergence of nonpsychiatric mental health professionals, professional power has become a complex issue, since clinical psychologists, community psychiatric nurses, psychotherapists, to name only a few, are also increasingly professionalized and state registered. Also, in the era of managed care, professional autonomy and decision-making power are threatened increasingly by third-party authorization mechanisms.

Finally, those receiving care from professionals start to threaten traditional professional identities. They are themselves becoming more powerful in forming self-help and advocacy groups, and tend to use terms such as 'clients,' 'consumers,' and even 'survivors of psychiatry,' instead of the term 'patients.' These terms have different implications of how the traditional

doctor–patient relationship is redefined. ‘Client,’ unlike ‘patient,’ implies partnership rather than submissivity and passivity. The term ‘consumer’ goes even further in claiming that patients have certain entitlements for service for which they can sue, and the concept of ‘survivor of psychiatry’ denies the right of psychiatry to intervene in a person’s problems. Pilgrim and Rogers (1999) give a detailed account of sociological thinking about these issues, and how the situation has changed since Talcott Parsons (1951) formulated the paradigm of the ‘sick role.’ According to Parsons, the sick role allows people to be exempted temporarily from social responsibilities, but, in exchange, they must accept that they cannot get well on their own, and that they have to look for professional help. Foucault’s (1977) concept of ‘productive power’ provides an interesting sociological framework for the evolving discourse between professionals and the (potential, actual, or ex-) users of their services.

2.3 Psychiatric Practice

The old mental hospital was the mainstay of psychiatry for over 150 years. At the end of this period, in the 1950s, the mental hospital (or the ‘asylum’ as it had been called) became a favorite topic of sociology. There were at least two reasons for this interest. Mental hospitals could be regarded as a ‘small society,’ an experimental situation, so to speak, in order to study sociological phenomena in general (Caudill 1958), and, probably more important, sociologists became critics of the impersonal way in which patients were treated. Goffman (1961), in his famous study based on participant observation, coined the concept of the ‘total institution’ for the mental hospital and similar organizations which disregard the personal needs of their inmates. In a more rigid research design, Wing and Brown (1961, 1970) showed a strong association between the typical poverty of the social environment of the mental hospital and the ‘clinical poverty’ of psychopathological phenomena such as blunted affect and social withdrawal—a view which has been challenged later (Curson et al. 1992). In a different approach, social historians attributed the growth of the asylum during the nineteenth century to more general developments in society, such as the growth of humanitarianism (Jones 1960), of capitalism (Scull 1979), and of surveillance tendencies (Foucault 1977).

Beginning during the 1950s, in countries with large numbers of beds in psychiatric hospitals—such as the USA and the UK—the numbers of psychiatric beds, which had been steadily rising over the first half of the twentieth century, started to fall. While, for instance, in the USA, in the mid-1950s about 600,000 patients were in mental hospitals, this number had come down to below 100,000 some 50 years later. The debate about the reason for this fall is still not resolved.

Psychiatrists attribute it mainly to the introduction of the neuroleptic drugs (which can control delusions and hallucinations). Sociologists have suggested different explanations: That, with the growth of the welfare state, controlling by segregation had become too costly, and that the public’s attitude toward psychiatric patients had changed in the sense of becoming more favorable (for an overview, see Busfield 1986). Furthermore, differences in the extent and kind of implementation of community psychiatry in different countries have been related to different health and social welfare systems (Goodwin 1997).

Forster (2000) examined three sociological theories (political-economy: cost containment; professional dominance: psychiatric ‘expansionism’; and post-structuralism: maintaining order and control) about the process of de-institutionalization, and concludes that none of them explains sufficiently the move away from hospital to community psychiatry. He proposes that the kind and degree of the medicalization of psychiatry has changed profoundly: While mainstream psychiatry is now more medicalized than before, the treatment of less serious mental health problems cannot be monopolized any more by medicine, and is becoming increasingly the practice field of other professions.

Compulsory admission and compulsory treatment are an intrinsic part of psychiatric services. Since the beginning of the twentieth century the legal professions have been involved increasingly in the process of admissions to mental institutions in order to counteract what has been regarded as mismanagement by the medical professions. This has prompted sociologists to analyze the social processes involved. There are basically two sociological positions: Those seeing the legal and the medical systems as two competing institutions (Mechanic 1978), and those who argue that, in practice, they join to become one actor, a ‘treatment-control’ system which ‘takes care’ of residual cases viewed as problematic for society, although the written law distinguishes between treatment and control (Dallaire et al. 2000). In the UK and other countries, a new dimension has been added to this issue with the concept of a ‘community treatment order,’ which would allow compulsory medication for patients living in the community and would let the pendulum swing back to medical control (Pilgrim and Rogers 1999).

Finally, sociologists have also analyzed and criticized specific therapeutic approaches. An example is the classical antipsychotic drugs and the barriers to the professional recognition of their side effects, especially of tardive dyskinesia, as an ‘iatrogenic disease’ (Brown and Funk 1986).

3. Sociology in Psychiatry

Sociologists working within the field of psychiatry are often criticized by their colleagues that they accept

psychiatric concepts naively. Robert Straus (1957) has noted that 'the sociologist *in* medicine risks a good relationship, if he tries to study his colleagues.' Work presented in this section is therefore clearly separated in nature from that discussed in the previous section. In principle, psychiatric diagnosis is taken for granted, or at least not regarded as problematical by these authors.

Typical variables investigated in psychiatric studies using sociological methods include gender, age, social class, and inequality, as well as ethnicity and race (Pilgrim and Rogers 1999). An overlap with the sociology of psychiatry can sometimes not be avoided, especially if questions of disadvantage, stigma, and mental health policy are the topic of a study. Gender is a typical example of such a topic belonging to both sociologies (Busfield 1996).

Sociologists *in* psychiatry have been, and are, active in a large number of fields. They study, or help to study, the distribution of mental disorders in society, and the social factors and processes which may be causal for mental illness or influence its course. They investigate which social factors prevent the early detection, and thus the early treatment, of mental disorders, and which factors influence 'illness behavior' or 'help-seeking behavior' (here conceptual issues such as the professional and lay definition of illness, i.e., aspects of the sociology of psychiatry, are involved), and how effective psychiatric treatments are under different institutional and social conditions. They have introduced new outcome measures for psychiatric interventions, which are less 'illness-centered' and more 'life-centered' (such as quality of life, social functioning, disability, social support, burden on the family) and have promoted the use of these measures to assess the 'reality of life' of persons with mental disorders, including the effects of stigma. Finally, in times of increasing restrictions of funds for health services, sociologists apply these concepts to health policy and health economics.

The epidemiology of mental disorders is discussed elsewhere in this volume (see *Mental Illness, Epidemiology of*). It must be noted though that the epidemiological survey methodology cannot only be used to study the influence of *social* factors, but also of *biological* and *genetic* factors. In fact, before the advent of the molecular biological approach, but also still at the time of writing, many conclusions drawn by psychiatric geneticists were based on epidemiological studies on the distribution of specific mental disorders among the relatives of mentally ill persons.

A few important historical examples of 'sociology in psychiatry' are worth mentioning. During the 1930s, the Chicago school of urban sociology studied the distribution of psychiatric disorders, such as schizophrenia, across the city alongside other forms of deviant behavior, and found higher rates in deprived areas (Faris and Dunham 1939). In the 1950s and 1960s, groups of sociologists and psychiatrists carried

out a number of seminal studies: Cumming and Cumming (1957) investigated public attitudes toward mental illness and tested the efficacy of a large intervention program (which turned out not to be successful), Hollingshead and Redlich (1958), and Langner and Michael (1963) found an association between lower social class and mental illness. Srole et al. (1962) demonstrated the astonishingly high prevalence of mental disorders in metropolitan New York. Star (1955) found that public attitudes towards the mentally ill were stereotyped and quite unfavorable. Finally, Brenner (1973), in a time series analysis covering 117 years, demonstrated that in times of economic recession the yearly rates of admissions to mental hospitals in New York increased.

A few more recent studies and topics applying sociological research methods are described here by way of illustration. Brown and Harris (1978) conducted a detailed sociological analysis on the etiology of depression in women and found that, among women who experienced a major life event, only those who had no close confidant, or in whom other social vulnerability factors were present, became depressed. Concerning the presumed dangerousness of psychiatric patients, Steadman et al. (1998) reported that psychiatric patients who did not abuse drugs or alcohol had the same rates of violence as their non-patient neighbors, and that, if violence occurred, it was directed toward family and friends rather than strangers. A longstanding debate concerns the interpretation of the finding of a higher prevalence of mental disorders in the lower social classes and in deprived city areas, where the 'social causation' and the 'social selection/drift' hypotheses rival for explanation (Fox 1990). Recently Löffler and Häfner's (1999) study in Germany supported the social drift theory, while Fox (1990) had concluded earlier that there was very little, if any, empirical support for social selection/drift processes in serious mental illness. The epidemiological approach to suicide using published statistics has been a topic of much controversy since Durkheim's (1897) seminal study. Douglas (1967) has denied the usefulness of this approach, and suggested that the only scientific topic worthwhile studying is the social meaning of suicide, i.e., the motives of those who kill themselves. A recent review on the sociological literature on suicide is provided by Stack (2000).

Typical problems on the microsocial level which have been studied by sociologists (partly in conjunction with psychiatrists) are: Stigma, quality of life and satisfaction, illness behavior, life events, the emotional climate within the family ('expressed emotion') and its relation to relapse in schizophrenia, social support, coping, compliance and the process of caregiving for the chronic mentally ill, homelessness, and the work environment. Phelan et al. (2000) found that people with psychosis are more feared today than they were in the 1950s. Markowitz (1998) studied the effects of

stigma on the psychological well-being and life satisfaction of persons with mental illness, and Link et al. (1987) looked into the effects of stigma on recovery. Rosenfield (1992) found that the single most powerful influence on the improvement of subjective quality of life of the chronic mentally ill was the improvement of the individual's actual power in terms of economic status. Katschnig et al. (1997) have reviewed the concept of quality of life in relation to mental disorders, and Jackson et al. (2001) have summarized research on predictors of patient satisfaction. Mechanic (1995) relates these issues to the concept of illness behavior, arguing that the disability process is specifically relevant to promoting function and maintaining patients' quality of life, especially in primary care settings, where patient motivation should be increased, helpful attitudes encouraged, coping strategies taught, family members and employers educated, and support provided.

Finally, services research as well as studies which evaluate the functioning of psychiatric services in terms of quality control (Scheid and Greenley 1997) and mental health policy papers (Mechanic 1994) are further examples of sociological involvement in psychiatric issues.

4. Conclusion

As this review has shown, there are numerous points of contact between sociology and psychiatry, and each can learn much from the other. This is especially true at a time when the focus of psychiatry is moving further back into the community, and consumer participation and a 'civil dialogue' are becoming more important. Qualitative research methods, which seek to explain quantitative findings, generate hypotheses and ensure that questionnaires are relevant, are an appropriate tool for accompanying this process (Chard et al. 1997). Also, it might be useful to develop a sociology of the relationship between sociology and psychiatry. This review could serve as modest point of departure for such an undertaking.

See also: Culture as a Determinant of Mental Health; Epidemiology, Social; Health and Illness: Mental Representations in Different Cultures; Mental Health: Community Interventions; Mental Illness, Epidemiology of; Psychiatry, History of; Psychoanalysis: Overview; Public Psychiatry and Public Mental Health Systems; Social Intervention in Psychiatric Disorders

Bibliography

American Psychiatric Association 1994 *Diagnostic and Statistical Manual of Mental Disorders*, Fourth Edition. Washington, DC

- Brenner H 1973 *Mental Illness and the Economy*. Harvard University Press, Cambridge, MA
- Brown G W, Harris T 1978 *The Social Origins of Depression: A Study of Psychiatric Disorder in Women*. The Free Press, New York
- Brown P 1995 Naming and framing: The social construction of diagnosis and illness. *Journal of Health and Social Behavior* (extra issue): 34–52
- Brown P, Funk S C 1986 Tardive dyskinesia: Barriers to the professional recognition of a iatrogenic disease. *Journal of Health and Social Behavior* 27: 116–32
- Busfield J 1986 *Managing Madness*. Hutchinson, London
- Busfield J 1996 *Men, Women and Madness: Understanding Gender and Mental Disorder*. Macmillan, London
- Busfield J 2000 Introduction: Rethinking the sociology of mental health. *Sociology of Health and Illness* 22: 543–58
- Caudill W 1958 *The Mental Hospital as a Small Society*. Harvard University Press, Cambridge, MA
- Chard J A, Lilford R J, Court B V 1997 Qualitative medical sociology: What are its crowning achievements? *Journal of the Royal College of Medicine* 90: 604–9
- Cumming E, Cumming J 1957 *Closed Ranks: An Experiment in Mental Health Education*. Harvard University Press, Cambridge, MA
- Curson D A, Pantelis Chr, Ward J, Barnes Th R E 1992 Institutionalism and schizophrenia 30 years on. *British Journal of Psychiatry* 160: 230–41
- Dallaire D, McCubbin M, Morin P, Coehen D 2000 Civil commitment due to mental illness and dangerousness: The union of law and psychiatry within a treatment-control system. *Sociology of Health and Illness* 22: 679–99
- Douglas J D 1967 *The Social Meanings of Suicide*. Princeton University Press, Princeton, NJ
- Durkheim E 1950 *The Rules of the Sociological Method* [1895 French original]. Free Press of Glencoe, Glencoe, IL
- Durkheim E 1951 *Suicide* [1897 French original]. Free Press of Glencoe, Glencoe, IL
- Faris R E, Dunham H W 1939 *Mental Disorders in Urban Areas*. University of Chicago Press, Chicago, IL
- Forster R 2000 The many faces of deinstitutionalisation—a discussion of sociological interpretations. *Psychiatric Praxis* 27 (Suppl. 2): 39–43
- Foucault M 1965 *Madness and Civilization* [French original 1961]. Random House, New York
- Foucault M 1977 *Discipline and Punish: The Birth of the Prison* [French original 1975]. Pantheon, New York
- Fox J W 1990 Social class, mental illness, and social mobility: The selection-drift hypothesis for serious mental illness. *Journal of Health and Social Behavior* 31: 344–53
- Freidson E 1970 *Profession of Medicine*. Harper & Row, New York
- Goffman E 1961 *Asylums*. Doubleday, New York
- Goodwin S 1997 *Comparative Mental Health Policy: From Institutional to Community Care*. Sage, London
- Hollingshead A B, Redlich F C 1958 *Social Class and Mental Illness*. Wiley, New York
- Jackson J L, Chamberlin J, Kroenke K 2001 Predictors of patient satisfaction. *Social Science and Medicine* 52: 609–20
- Jones K 1960 *Mental Health and Social Policy 1845–1959*. Routledge & Kegan Paul, London
- Katschnig H, Freeman H, Sartorius N 1997 *Quality of Life in Mental Disorders*. Wiley, Chichester, UK
- Kessler R C, McGonagle K A, Zhao S, Nelson C B, Hughes M, Eshleman S, Wittchen H U, Kendler K S 1994 Lifetime and

- 12-month prevalence of DSM-III-R psychiatric disorders in the United States. Results from the National Comorbidity Survey. *Archives of General Psychiatry* **51**: 8–19
- Langner T, Michael S T 1963 *Life Stress and Mental Health*. The Free Press, New York
- Link B, Cullen F 1989 A modified labelling theory approach to mental disorders, an empirical assessment. *American Sociological Review* **54**: 400–23
- Link B, Cullen F, Frank J, Wozniak J 1987 The social rejection of ex-mental patients: Understanding why labels matter. *American Journal of Sociology* **92**: 1461–500
- Löffler W, Häfner H 1999 Ecological pattern of first admitted schizophrenics in two German cities over 25 years. *Social Science and Medicine* **49**: 93–108
- Markowitz F 1998 The effects of stigma on the psychological well-being and life satisfaction of persons with mental illness. *Journal of Health and Social Behavior* **39**: 335–47
- Mechanic D 1978 *Medical Sociology*, 2nd edn. The Free Press, New York
- Mechanic D 1994 Establishing mental health priorities. *Milbank Quarterly* **72**: 501–14
- Mechanic D 1995 Social dimensions of illness behavior. *Social Science and Medicine* **41**: 1207–16
- Mirowsky J, Ross C E 1989 Psychiatric diagnosis as reified measurement. *Journal of Health and Social Behavior* **30**: 11–25
- Parsons T 1951 *The Social System*. Free Press, Glencoe, IL
- Pilgrim D, Rogers A 1999 *A Sociology of Mental Health and Illness*. Open University Press, Philadelphia, PA
- Phelan J C, Link B G, Stueve A, Pescosolido B A 2000 Public conceptions of mental illness in 1950 and 1996: What is mental illness and is it to be feared? *Journal of Health and Social Behavior* **41**: 188–207
- Robins L, Regier D A 1991 *Psychiatric Disorders in America: The Epidemiologic Catchment Area Study*. Free Press, New York
- Rosenberg C E 1992 Introduction. In: Rosenberg C E, Gooden J (eds.) *Framing Diseases. Studies on Cultural History*. Rutgers University Press, New Brunswick, NJ
- Rosenfield S 1992 Factors contributing to the subjective quality of life of the chronic mentally ill. *Journal of Health and Social Behavior* **33**: 299–315
- Rosenhan D 1973 On being sane in insane places. *Science* **178**: 250–58
- Scheff T J 1999 *Being Mentally Ill: A Sociological Theory*, 3rd edn. [1st edn. 1967]. De Gruyter, New York
- Scheid T L, Greenley J R 1997 Evaluations of organizational effectiveness in mental health programs. *Journal of Health and Social Behavior* **38**: 403–26
- Scully A 1979 *Museums of Madness*. Penguin, Harmondsworth, UK
- Srole L, Langner T S, Michael S T, Opler M K, Rennie T A C 1962 *Mental Health in the Metropolis: The Midtown Manhattan Study*. McGraw-Hill, New York
- Stack S 2000 Suicide: A 15-year review of the sociological literature. Part I: Cultural and economic factors; Part II: Modernization and social integration perspectives. *Suicide and Life-Threatening Behaviour* **30**: 145–62, 163–76
- Star S A 1955 *The Public's Ideas About Mental Illness*. National Opinion Research Center, Chicago, IL
- Steadman H J, Mulvey E P, Monahan J, Robbins P C, Appelbaum P S, Grisso T, Roth L H, Silver E 1998 Violence by people discharged from acute psychiatric inpatient facilities and by others in the same neighborhoods. *Archives of General Psychiatry* **55**(5): 393–401

- Straus R 1957 The nature and status of medical sociology. *American Sociological Review* **22**: 200–4
- Van der Geest S 1995 Overcoming ethnocentrism: How social science and medicine relate and should relate to one another. *Social Science and Medicine* **40**: 869–72
- Wing J K, Brown G W 1961 Social treatment of chronic schizophrenia: A comparative survey of three mental hospitals. *British Journal of Psychiatry* **107**: 847–61
- Wing J K, Brown G W 1970 Institutionalism and schizophrenia. A comparative study of three mental hospitals. Cambridge University Press, London

H. Katschnig

Psychiatry: Anthropological Aspects

Psychiatry—if we think for instance of psychoanalysis and of medical anthropology—has had an essential influence on philosophical anthropology, determining humankind in its world and its being-with-others from the perspective of manifold risks and disturbances. However, this article is concerned with exactly the opposite relationship, namely that of anthropology to psychiatry (see Blankenburg 1980a, Kraus 1999, Lang 1982).

What mental illness is in its various manifestations and which methods of investigation and therapies are appropriate for mentally ill patients from the aspect of their humanity, is also a question of anthropology as a comprehensive framework. A particular anthropology determines object theory to a high degree, that is, how mental illness can become objective, as well as the methodological accessibility of this object in psychiatry. Anthropological assumptions do not only determine norms usually applied in psychiatry to differentiate between mental illness and health but also determine ethics in psychiatry and address the relationship between body and soul and problems of free will and of self-determination. Because psychiatry as a science and therapeutic practice often latently implies an anthropology, by which psychiatry is deeply determined, psychiatry cannot avoid reflecting these premises. Above all, this is valid for the question of the adequacy of its methods for the respective object. A particular anthropology, however, does not only contribute to the development of what in psychiatry can become the object and its methodological accessibility, but it can also even prevent this development. Thus, certain results of research can be enabled, but can also be prevented.

As yet, it has not been sufficiently realized that psychiatry can be restricted in its scientific and therapeutic possibilities by its anthropology and can even be endangered by dogmatic and overgeneralized

basic assumptions. Problems of reductionism, however, of nothing-but theories, of the methodological preconditions of experience, are increasingly acknowledged. What Jaspers (1965) said about philosophy is also valid for anthropology. Stressing its significance for psychiatry, it does not only free the inner space for new possibilities of knowledge, but can also open horizons more adequate to a better understanding of the patient. A switching between anthropological and empirical research (Mundt and Spitzer 1999) has decisive significance for innovation and thus for the development of psychopathology as well as of psychiatry in its therapeutic and rehabilitative practice.

Ever since the separation of natural and cultural sciences we have differentiated between anthropology in the sense of natural science and philosophical and cultural anthropology. Whereas scientific or biological anthropology, identical with physical anthropology or human biology, is concerned with the biological nature of humankind, philosophical anthropology makes statements about our personal or spiritual essence. Cultural anthropology, mostly interchangeable with ethnology, is concerned with the question of which information is given by cultural statements about what is variable and constant in human beings. To what extent is it possible to start with a previously given essence of human beings and in which way does the human being as the 'nondetermined animal' (Gehlen 1966) determine what he or she is. This question became one of the main topics of existential philosophy, where this human self-determination became the most important feature of a person's essence.

Immanuel Kant ([1798] 1978) already differentiated between a physical and a pragmatic anthropology. Whereas the first is concerned with what humans are by nature, the second, identical with philosophical anthropology, asks what they can make of themselves. Kant's philosophy became important for psychiatry not only because he developed interesting ideas about mental illness, seeing its general features in a loss of communal sense (*sensus communis*), that is, the loss of the ability of a person to be corrected by others (§ 53, p. 117), but especially because he has influenced psychiatry with respect to his methodology. Showing the dependency of experience on *a priori* ideas and thus overcoming the traditional splitting of subject and object, Kant's philosophy ([1781/87] 1965) stood for an opposition of idealism to realism which maintains the independence of the recognizing subject from the object. This opposition has had a constant effect not only on many philosophical schools of thought but also on psychiatry. With his concept of the transcendental subject he paved the way for Husserl's phenomenology and for existential approaches of philosophy. These new concepts and methods on their part made possible approaches like *Daseinsanalyse* (existential analysis), existential psychiatry, and cognitive approaches. Because above all in psychoses the

comprehension of the world by the experiencing subject has changed, these psychiatric approaches claim a better understanding of the cognitive processes in these patients.

The following authors of a philosophical anthropology have become important for psychiatry: J. G. Herder (1744–1803), emphasizing the human as a being gifted with language; G. W. F. Hegel (1770–1831) with his historical philosophical critique of anthropology; M. Scheler (1874–1928), the founder of modern anthropology in his work *The Place of Man in the Universe*; H. Plessner (1892–1985), who speaks of the eccentric position of human beings, who have only to make out of themselves what they are; and A. Gehlen (1904–67) with his determination of humans as deficient beings because of lack of adjustment to their surroundings and reduction of instinct. At the same time an openness to the world corresponds to this lack. For a psychiatry which calls itself anthropological, the following philosophers were above all influential: S. Kierkegaard (1813–55) who initiated the development of existential philosophy; E. Husserl (1859–1938), the main founder of the phenomenological approach in philosophy; M. Heidegger (1889–1976) with his *Daseins* analytic; J. P. Sartre (1905–80) with existential psychoanalysis; and M. Merleau-Ponty (1908–61) with his phenomenology of the body. But on the other hand there has been an influence of medicine, especially of psychiatry, on philosophical anthropology: here we could mention Heinroth (1773–1843), V. von Weizsäcker (1886–1957), and M. Binswanger (1881–1966).

Multiple influences on psychiatry also arose from biological anthropology (including paleoanthropology, human genetics, research on human bodily constitution), from social anthropology and cultural anthropology, from cognitive and evolutionary anthropology. Social anthropology and role theory (see Kraus 1980) had a strong impact on the development of social psychiatry and antipsychiatry followed by many psychopathological concepts and therapeutic measures.

Cultural anthropology examining the cultural variation of the human being as well as human constants independent of culture and epoch shows very clearly the double significance of the concept of anthropology already mentioned above. Showing the variability of the actual being and behavior of human beings, at the same time it marks the boundaries of anthropology as a universally valid theory of the essence of humankind. Cultural anthropology has cultivated a special approach of a transcultural or cultural psychiatry which investigates the variability and frequency of symptoms of disturbance entities, nowadays defined in glossaries, as well as of these entities themselves in different cultures. It has also led to the question of whether regarding the classical entities of disturbances like schizophrenia and manic-depressive illness we are actually dealing with entities of natural diseases or if

these are entities only coming from concepts of understanding. Transcultural or cross-cultural psychiatry is thus fundamentally involved in the contemporary crisis of nosology in psychiatry.

With the notion of a cognitive anthropology, Emrich (1999) sums up different approaches of a psychological, physiological (Singer 1994, Roth 1994), and philosophical constructivism. In cognitive anthropology the leading idea is the general quality of constructivity ascribed to systems generating consciousness (Roth 1994). The progress of modern evolutionary biology in the future could enable a phylogenetic and ontological understanding of mental disturbances.

Heinroth can be looked at as the founder of an anthropological medicine in its true sense in the context of psychiatry, not to be confused with a medical anthropology. Many anthropological approaches in general medicine, especially those of Weizsäcker and Wyss, have also had effects on psychiatry.

Janzarik (1988) has presented an approach of his own to anthropological psychiatry. In his structural-dynamic concept he sees psychopathological phenomena in their neurobiological (dynamic) foundation as coherently connected with biographical structurally bound significances.

Under the heading of a phenomenological-anthropological psychiatry, different approaches can be summarized which have a phenomenological method as well as an anthropological orientation in common. These psychiatric approaches of an eidetic phenomenology of essence and a constitution-phenomenology as well as of *Dasein* and existence analysis relate to different authors of philosophical anthropology in its broader sense such as E. Husserl, M. Heidegger, J. P. Sartre, and M. Merleau-Ponty.

Even if the understanding of phenomenology and phenomenological in these different approaches is not always the same, the notion of phenomenon aims at that 'which-shows-itself-in-itself' and as such is different from the notion of the symptom which points to something which does not just show itself directly, namely to the illness underlying it. An approach strongly oriented to the phenomenon enables us to describe the original individual subjective experience as well as the respective lifeworld of the patient more adequately and differentiated than otherwise possible.

'Anthropological' signifies on one hand what is common to all people and on the other hand what in contrast to the animal is specifically human. In the context of anthropological psychiatry with this term the relationship of psychiatric disturbances to basic human structures is meant, as for instance consciousness and subjectivity, intersubjectivity and individuality, history and temporality, freedom, corporality and so on. These approaches not only take into consideration human essence because of ethical or humanistic reasons, but above all of methodological

reasons. The aim is to understand psychiatric disturbances as regular variations of these basic structures, that means of human essence itself. This is done not so much to investigate the factual conditions for the development of psychopathological phenomena but to investigate the conditions originating from human essence which made them possible. For instance hallucination can be understood out of the normal act of perception as an uncoupling of spontaneous imaginative acts from a normal interaction with receptivity (Blankenburg 1980a, 1980b).

Binswanger's (1955) approach with 'anthropological proportion' has turned out to be especially fertile in psychiatry. The partly oppositional dimensions of height and breadth in human existence, of self-realization and world-realization, of individuation and relationship to community, of need for continuity and openness to innovation, of tolerance and intolerance of ambiguity, of overidentification and underidentification normally stand in a certain proportion to each other. These proportions are shifted in psychiatric disturbances. Relating to this alteration of the respective proportions some pictures of illness are oppositional. Thus an overidentification with their respective being in melancholics can be opposed to an underidentification in hysterics (Kraus 1996a). Overidentification with social role for example, with environment, etc., can understandably precipitate a melancholic phase in the case of their loss. One speaks of an empty-nest depression if a parent becomes depressed when children leave home. Other examples are a depression in the situation of retirement or of moving house. Hysterics, on the contrary, tend rather not to identify with their respective being or their actual feelings. They are striving for being more and something different than they actually are or to show feelings they have not. The abnormal as well as the normal can in this way be located in a polar field which is structured through ideal norms which oppose one another in a field in which 'deviations' from at least two sides are possible, as Blankenburg (1980a) put it. The pathological element can in this way be traced back to healthiness. On the other hand healthiness shows itself as something which always holds the possibility of becoming pathological by an alteration of proportions.

Such a concept of mental disturbances offers a special possibility of identification with the patient in the therapeutic process. Beyond a purely descriptive psychopathology in the sense of Jaspers (1965) it is thus possible to develop a dynamic phenomenology of psychiatric disturbances which has a dialectic character. Jaspers in his subjective psychopathology emphasized the special experience of the patient as opposed to a mere behavioral description in the sense of an objective psychopathology.

An eidetic essence-oriented phenomenology described above goes a step further in trying to grasp the essence of altered experiences and behaviors in a so-

called eidetic reduction in the sense of Husserl. In a process of free imaginative variation here one tries to get from a fact to the essence of a fact, from a sensual to a categorical presentation. Categorical presentation means for instance to recognize a particular species of tree in an actual tree, or in different kinds of behavior and experiences of hysterics the element of under-identification, of the pretence of being instead of actual being. Such an eidetic reduction can also be carried out in relation to Heidegger's notion of '*In-der-Welt-sein*' or Merleau-Ponty's '*Etre-au-monde*' showing the special relationship of patients to themselves and to the world which is expressed in their experiences and behaviors. A constitutive or transcendental phenomenology in the sense of Husserl is to be differentiated from an eidetic phenomenology. Instead of going from the fact to the essence of a fact as in eidetic phenomenology, constitutive phenomenology examines the constitutive achievements of consciousness, which let us perceive the concrete as it appears to us.

This approach of a constitutive phenomenology and the following philosophy of existence became important for psychiatry because in the perception of psychotic patients—but not only for them—the world and its significances constitute themselves in another way than in healthy people. In psychiatry this has resulted not only in making the experience of objects, for instance classical symptoms, into research topic, but also the preobjective experience, that is, the constitutive processes out of which self and world develop.

The preobjective experience is especially important in the intuitive diagnostics, for instance in the so-called precox feeling. Such preobjective experiences are for instance, the spatialization, temporalization, intersubjectivity, and corporality of a patient. Classical symptoms such as delusion, hallucination, phobic anxieties, and obsessive-compulsive symptoms assessed by a psychopathology oriented to the object, turned out to be constructed on such preobjective experiences.

Phenomenology of constitution, as Husserl put it, recently entered into a manifold relationship with cognitive theory and cognitive neuroscience (Wiggins and Spitzer 1997, Gallagher 1997). Varela (1996) even proposed a neurophenomenology. Several authors of the phenomenological-anthropological approach showed how in the contents of delusion and hallucination vs. the psychotic tendency of dissolution of their world, patients try again to give sense to their new world, even if a distorted one (Feldmann and Schmidt-Degenhardt 1997, Müller-Suur 1980, Kraus 1994a).

Husserl's concept of an intersubjectively constituted lifeworld became of increasing significance for psychiatry, above all for the understanding of schizophrenic modes of existence (Callieri and Castellani 1981, Blankenburg 1979). People with schizophrenia are characterized by a damaged rootedness in the

lifeworld, connections of significances and relevancies becoming confused, which shows itself in extravagance and oddity but also in the delusions of these patients. Blankenburg (1971) speaks of a lack of natural self-evidence in schizophrenics. Whereas schizophrenics often lack an inner standard which would allow them a behavior appropriate to the social norms, melancholics being too adapted to these norms are too strongly fixed to the common sense and to their lifeworld (Blankenburg 1979, Stanghellini 1997, Kraus 1996b). This is already shown premorbidly by their hypernormic behavior which is characterized by a lack of distance to normative expectations (Kraus 1996b). This is also shown by the fact that they find their identity above all in their social roles. For this reason they are especially in danger of becoming depressed in the case of the loss of a social role (Kraus 1996b).

Binswanger has developed his existential analysis above all under the influence of Heidegger's *Dasein* analytic in *Being and Time* (Heidegger 1927), which, against the ontologic intention of the author, Binswanger understood as an anthropology. With existential analysis we are concerned with a hermeneutic of the inner life history which is asked for its underlying individual concept of world which determines the concrete experience and behavior of a person.

Sartre's (1962) existential psychoanalysis has also contributed to the better understanding of certain pathological modes of existence, such as those of hysterics, melancholics, and patients with anxiety and obsessive-compulsive disorders. In this way we can establish certain existential types which can have an impact on the classification of these disorders (Kraus 1994b).

The phenomenological-anthropological orientation in psychiatry has promoted a series of therapeutic measures, for instance the anthropological-integrative psychotherapy of Wyss (1982) and existential analysis in the sense of Boss (1951) and Condrau (1963). But it has also influenced other therapies.

Summing up we can say that all anthropological approaches in psychiatry have as their aim to bring psychiatric disturbances in contact with the whole of human reality to achieve better understanding of those disturbances in a way that is appropriate to the essence of humankind. Considering anthropological prepositions in this way allows us to reflect the fundamentals of psychiatry in research as well as in therapy by which the phenomenological-anthropological approaches contribute to further development.

See also: Anthropology, History of; Cognitive Anthropology; Health and Illness: Mental Representations in Different Cultures; Mental Health and Normality; Mentally Ill: Public Attitudes; Mind–Body Dualism; Psychological Anthropology

Bibliography

- Beck A 1970 The core problem in depression: The cognitive triad. In: Massermann J H (ed.) *Depression. Theories and Therapies*. Grune & Stratton, New York
- Binswanger L 1955 Vom anthropologischen Sinn der Verstiegtheit. In: *Ausgewählte Vorträge und Aufsätze*. Franke, Bern, Germany, Vol. 2, pp. 235–42
- Binswanger L 1992 Drei Formen missglückten Daseins. In: Herzog M (ed.) *Ludwig Binswangers ausgewählte Werke*. Asanger, Heidelberg, Germany, Vol. 1, pp. 1–443
- Blankenburg W 1971 *Der Verlust der natürlichen Selbstverständlichkeit*. Enke, Stuttgart, Germany
- Blankenburg W 1979 Phänomenologische Epoché und Psychopathologie. In: Sprondel W M, Grathoff R (eds.) *Alfred Schütz und die Idee des Alltags in den Sozialwissenschaften*. Enke, Stuttgart, Germany, pp. 125–39
- Blankenburg W 1980a Anthropologisch orientierte Psychiatrie. In: Peters U W (ed.) *Ergebnisse für die Medizin, 2: Psychiatrie. Die Psychologie des 20. Jahrhunderts*. Kindler, Zürich, Vol. 10, pp. 182–97
- Blankenburg W 1980b Phenomenology and psychopathology. *Journal of Phenomenological Psychology* **11**: 50–78
- Boss M 1951 Beitrag zur daseinsanalytischen Fundierung des psychiatrischen Denkens. *Schweiz Archives of Neurology and Psychiatry* **67**(1): 15–19
- Callieri B, Castellani A 1981 On the psychopathology of the life-world. In: Bello A A (ed.) *Analecta Husserliana*. Reidel, Dordrecht, Germany, pp. 172–202
- Condrau G 1963 *Daseinsanalytische Psychotherapie*. Huter, Bern, Germany
- Emrich H M, Schiefenhövel W 1999 Philosophische Anthropologie als Grundlagenwissenschaft der Psychiatrie. In: (ed.) *Psychiatrie der Gegenwart*, 4th edn. Springer, Heidelberg, Germany, Vol. 1
- Feldmann H, Schmidt-Degenhard M 1997 Strukturelle Affinität des Unverständlichen im schizophrenen Wahn. *Nervenarzt* **68**: 226–30
- Gallagher S 1997 Mutual enlightenment: Recent phenomenology in cognitive science. *Journal of Consciousness Studies* **4**(3): 195–214
- Gehlen A 1966 *Der Mensch*, 8th edn. Athenäum, Frankfurt, Germany
- Heidegger M 1963 *Sein und Zeit*. Niemeyer, Tübingen, Germany
- Janzarik W 1988 *Strukturdynamische Grundlagen der Psychiatrie*. Enke, Stuttgart, Germany
- Jaspers K 1965 *Allgemeine Psychopathologie*. Springer, Berlin
- Kant J [1781/87] 1965 *Critique of Pure Reason* (trans. Kemp Smith N). St. Martin's Press, New York
- Kant J [1798] 1978 *Anthropology* (trans. Dowdell V L, rev. and ed. Rudnick H H). Southern Illinois University Press, Carbondale, IL
- Kraus A 1980 Bedeutung und Rezeption der Rollentheorie in der Psychiatrie. In: Peters U H (ed.) *Ergebnisse für die Medizin, 2: Psychiatrie. Die Psychologie des 20. Jahrhunderts*. Kindler, Zürich, Vol. 10
- Kraus A 1994a Phenomenology of the technical delusion in schizophrenics. *Journal of Phenomenological Psychology* **25**(1): 51–69
- Kraus A 1994b Phenomenological and criteriological diagnosis: Different or complementary. In: Sadler J, Wiggings O P, Schwartz M (eds.) *Philosophical Perspectives on Psychiatric Diagnosis and Classification*. Johns Hopkins University Press, Baltimore, MD, pp. 148–62
- Kraus A 1996a Identitätsbildung Melancholischer und Hysterischer. In: Seidler G H (ed.) *Hysterie Heute*. Ferdinand Enke, Stuttgart, Germany, pp. 103–10
- Kraus A 1996b Role performance, identity structure and psychosis in melancholic and manic-depressive patients. In: Mundt C, Goldstein M J, Hahlweg K, Fiedler P (eds.) *Interpersonal Factors in the Origin and Course of Affective Disorders*. Gaskell, London, pp. 31–47
- Kraus A 1999 Phänomenologisch-anthropologische Psychiatrie. In: Helmchen H, Henn F, Lauter H, Sartorius N (eds.) *Grundlagen der Psychiatrie*. Springer, Berlin
- Lang H 1982 Phänomenologisch-anthropologisch orientierte Psychiatrie und Psychologie im französischen Sprachraum. In: Eicke D (ed.) *Tiefenpsychologie, Vol. 4 Individualpsychologie und Analytische Psychologie*. Weinheim, Basel, Switzerland
- Müller-Suur H 1980 *Das Sinn-Problem in der Psychose*. Hofgreve, Göttingen, Germany
- Mundt C, Spitzer M 1999 Psychopathologie heute. In: (ed.) *Psychiatrie der Gegenwart*, 4th edn. Springer, Heidelberg, Germany, Vol. 1, pp. 4–44
- Roth G 1994 Gehirn und Geist. In: Schiefenhövel W, Vogel C, Vollmer G, Opolka U (eds.) *Vom Affen zum Halbgott*. Thieme, Stuttgart, Germany
- Sartre J P 1962 *Das Sein und das Nichts*. Rowohlt, Hamburg, Germany
- Singer W 1994 *Gehirn und Bewusstsein*. Spektrum, Heidelberg, Germany
- Stanghellini G 1997 For an anthropology of vulnerability. *Psychopathology* **30**: 1–11
- Varela F 1996 Neuropsychophenomenology. A methodological remedy for the hard problem. *Journal of Consciousness Studies* **4**: 330–49
- Wiggings O P, Spitzer M 1997 Cognitive science. In: Embree L, Behnke E A, Carr D et al. (eds.) *Encyclopedia of Phenomenology*. Kluwer Academic Publishers, Dordrecht, The Netherlands, pp. 101–4
- Wyss D 1982 *Der Kranke als Partner*. Vandenhoeck und Ruprecht, Göttingen, Germany, Vols. 1 and 2

A. Kraus

Psychiatry, Evidence-based

No country can afford health services to meet all the possible needs of the population, so it is advisable to establish criteria for which services to provide. Two basic criteria are the size of the burden caused by a particular disease, injury or risk factor, and the cost-effectiveness of interventions to deal with it (Bobadilla et al. 1994).

Health services are rationed. Access in most countries is controlled by money, waiting lists, and by the clinician's decision about what should be done. Health services are concerned to avoid the explicit rationing that would follow if they adopted the position outlined above and so they have become concerned with maximizing efficiency. They have done this by limiting

the input of funds, and by using scarcity to encourage improvements to the process of care. They keep asking 'can science inform practice?' an issue beset by many questions and too few answers. Yet it does seem relatively simple. Research is concerned with discovering the right thing to do; clinical audit with ensuring that the right things are done. The link that can close the gap is said to be evidence-based medicine: 'the conscientious, explicit, and judicious use of current best evidence in making decisions about the care of individual patients' (Sackett et al. 1996).

The evidence-based medicine movement seems to have originated from Archibald Cochrane's set of essays 'Effectiveness and Efficiency: Random Reflections on Health Services,' (Cochrane 1998) first published in 1971 and reprinted by BMJ Books in 1998. He argued that no new treatments should be introduced unless they had been shown in randomized controlled trials to be more effective than existing treatments, or equally effective and cheaper; and that one should evaluate all existing therapies, slowly excluding those shown to be ineffective or too dangerous. Cochrane actually meant efficacy when he wrote effectiveness (he acknowledged hating the word efficacy) but it is now usual to distinguish the three terms: efficacy, does the treatment work; effectiveness, does it work in practice; and efficiency, is it a wise expenditure of funds? The popular press is preoccupied with efficacy (can the doctors cure disease X?), not with effectiveness (do they?) or with efficiency (can we afford the cost?). Evidence-based medicine is concerned with identifying the efficacy of treatments and encouraging their implementation. The goal is increased effectiveness and efficiency (Andrews 1999).

Efficacy: 'Does the treatment work?' means does it produce benefit reliably over and above that to be expected from just being in treatment? Randomized placebo-controlled trials are valuable because they control for the uplifting effect of being in treatment (the true placebo effect). They also control for changes in symptoms and disability attributable to spontaneous remission, including the remission that occurs because people with chronic and fluctuating disorders commonly delay consulting until their disorder is particularly severe and can be expected to improve whatever happens. In some disorders the extent of this overall 'placebo response' is very large. In depression, it is two to three times as large as the change related to the present specific therapies, be they drugs or psychotherapy (see the discussion in Enserink 1999, pp. 238–40). That is, people given a sugar tablet will improve, not because of the tablet, but because of these other processes. People taking the active antidepressant tablet will improve more, but only one third of the overall change will be due to the antidepressant. It is for this reason that changes in symptoms with a specific treatment are suspect, unless they are shown to exceed the changes expected from a nonspecific placebo treatment.

With a few significant exceptions, most randomized controlled trials in psychiatry, even multicentre trials, have relatively small numbers of subjects. It is therefore usual to defer judgment until a number of trials have been done in different centers. If they agree that a certain treatment is better than placebo, then it probably is, but as most use different measures of outcome it has been necessary to develop a method for estimating the average degree of improvement with this treatment. Meta-analysis employs standard data analysis techniques to compare the results of studies by converting, if necessary, the results onto a standard metric, called an effect size. Even so, the wide range of outcome measures in use makes comparison difficult, and it is not always adjusted by expressing change in effect size terms. For example, while trials in depression almost always use the Hamilton scale so that results can simply be averaged, a meta-analysis of panic disorder found over 300 different measures in 58 studies, which has to be evidence of social irresponsibility in science. While each investigator will need to use a disorder-specific measure of symptoms, there is some real advantage in all using a generic measure of disability such as the short forms from the RAND medical outcomes study.

Structured systematic reviews are now common. In memory of Cochrane, a center at Oxford has been established to encourage and disseminate systematic reviews of the medical literature. These reviews abide by strict methodological guidelines and identify the level of confidence that the reader can place on any finding. In mental health, the dominant US publication is a book by Nathan and Gorman (1998). Any clinician, asked by a client what other treatments there are, would be obliged ethically to mention the treatments listed in that book as supported by type 1 evidence before offering advice about the treatment to be preferred. The self-help version of that book should be in every mental health clinician's waiting room (Nathan et al. 1999).

Effectiveness: Does the treatment work in practice when applied by the average clinician to the average client? There is a paucity of such data. Such data as there are show that effectiveness is reduced as one moves from experimental practice to routine clinical practice. Too often treatment as usual seems almost ineffective. Thus the problem is not so much defining the treatments that work as overcoming the degradation of benefit that occurs in practice. There are four issues that undermine the potential effectiveness of a treatment: deficits in coverage, deficits in compliance, deficits in clinical competence, and treatment integrity. Epidemiological studies in the USA, Canada, the UK, and Australia show consistently that only a fifth to a half of people who meet criteria for a mental disorder get help for their disorder (Andrews and Henderson 2000). Coverage is therefore low, variably limited by access to services and by the expectation that there is little point in attending. Poor

coverage when access is limited is understandable, but why should there be poor coverage when a national health service provides access? The Australian national mental health survey asked why the burden of depression persisted when access to care was free. Half the people who identified depression as their main complaint (and were disabled as a result) did not consult because they 'had no need for treatment.' There are data on the mental health literacy of Australians that show that the majority of the population think that relaxation and tonics are good for depression, and that medication is not. It is clear why they do not consult their doctors.

Deficits in compliance and treatment integrity are difficult to estimate, simply because auditing of either will produce a temporary change in behavior. There is evidence that asking people if they have taken all their medication is a valid measure of compliance; at least those who say they haven't can be counseled. One advantage of providing patient treatment manuals for psychological treatments is that the manuals contain explicit sheets on which the homework can be recorded and reviewed. Like homework at elementary school, if the clinician doesn't ask, the client doesn't do. The patient treatment manual also serves to ensure treatment integrity. After all, the manual is the curriculum for the treatment program and clinicians have little opportunity but to follow the framework, personalizing it for each patient, and following the sequence in every case. All psychological therapies should be manualized, because psychotherapy is about learning, and most learning programs have text books.

Coverage and compliance concern the system and the consumer. Clinical competence and treatment integrity concern the physician. How do you make sure the physician knows what to do, and then does what should be done. The evidence-based medicine movement has clear ideas about how clinicians should practice. Ideally, clinicians should constantly be asking 'what is the evidence about treatment for this type of patient with this type of disorder?' Searching the Internet they should source systematic reviews, such as those from the Cochrane Collaboration, before deciding which treatment is appropriate. There are three problems. The first is time; busy practitioners, especially those in private practice, have no opportunity to consult sources in this way. The neediness of patients simply consumes the day. The second is that systematic reviews are about people who satisfied the inclusion and exclusion criteria for RCTs, and who are unlike the pregnant woman, for whom information is needed. The third issue is that the methodological formalism required for such reviews may remove the very data or comments that would be helpful. Systematic reviews are essential, but they may need to be supplemented with more clinically relevant material.

Review journals are flourishing. For example, the aim of the British Medical Association's journal *Evidence Based Mental Health* was to print one-page

summaries of research that had met rigorous standards. Clinicians, with no time to do their own literature searches, could rely on the summaries to guide practice. To a large part this happens, but methodological formalism may be limiting the scope and obscuring meaning. In January 2000, the journal's website had a summary of an RCT in which three sessions of CBT and three sessions of psychodynamic therapy were equally effective in relieving sub-syndromal depression. The take-home message for the busy practitioner was unclear. Psychotherapy that requires transference can be completed in three meetings? CBT that teaches how to change thoughts and behaviors can be achieved in three sessions? That it is important to treat mild symptoms that do not constitute a disorder? Which is strange in a country like the UK, in which 70 percent of people with major depression do not see a doctor and most who do are treated inadequately. The review journals are important, but the editors need to think about the burden of the disease and the implications for practice, as well as about the probity of research designs. Complex problems are unlikely to have simple one-page answers.

Clinical practice guidelines are the traditional avenue used to put research findings into practice. Properly written they can combine a continuum of information from good research to good clinical wisdom, differentiating between statements supported by research, and statements, no less valuable, supported by opinion. The steps in preparing clinical practice guidelines are illustrated in Fig. 1. In summary: when there is a clinical issue and there are no existing guidelines, a panel of all stakeholders should be convened to identify the desired health outcome. They should proceed by preparing a systematic review of efficacy of interventions in relation to this outcome, differentiating clearly between interventions supported by the differing levels of the weight of evidence (level 1a: meta-analysis of randomized controlled trials; level 1b: at least one RCT; level 2: at least one controlled or quasi-experimental study; level 3: non-experimental descriptive studies; level 4: expert committee reports or opinions). Each statement in the guideline is qualified traditionally in accord with the quality of the supporting research evidence. Increasingly, the principal recommendations of guidelines require support from level 1a evidence, lower levels of evidence being valuable to fine-tune that principal recommendation for special groups of patients, or atypical clinical situations.

Most clinicians have access to at least half a dozen such guidelines, and many doctors in primary care have mountains of guidelines, so many that their ability to use any is questionable. While most originators claim that regular revisions will be issued, this has not proven to be the case. The problem is that guidelines have largely been written by advocates keen to support a position and, understandably, they are

Clinical Practice Guidelines Flow Chart

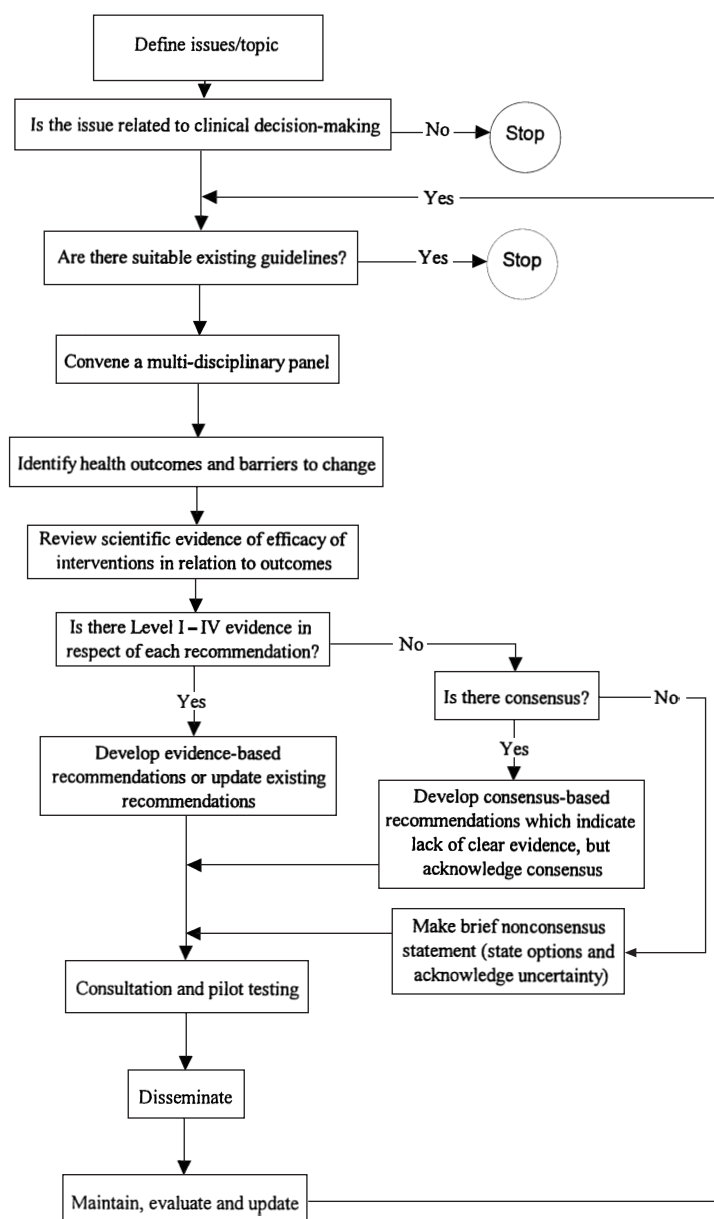


Figure 1
Clinical practice guidelines flow chart

not keen to update the data in terms of new developments. The Royal Australian and New Zealand College of Psychiatrists produced guidelines for 1981–91 (Quality Assurance Project 1983), and in 2001 they at last began working on a revision of some of these. The American Psychiatric Association and the UK Royal College of Psychiatrists have also produced

guidelines for their members. An attempt to produce mental health consensus guidelines that incorporated the views of all stakeholders—the various providers, the consumers and carers, and the public and private funders—appears to have run out of support: who wants to be fair when there is so much to be gained by being partisan? Moreover, when guidelines are prod-

used by independent bodies such as the Agency for Health Care Policy and Research they can run foul of vested interests so that the impact is nullified (cf. breast cancer screening, spinal fusion) (Deyo et al. 1997). Clinical practice guidelines should be a flowering of the scientific method—find out what works and implement it—but too often they are political acts that represent the interests of an advocacy group.

Leaving aside the problem of the accuracy of the information, how is it possible to introduce guidelines that busy practitioners will follow. The Australian National Health and Medical Research Council recently convened a meeting to learn how 14 studies in medicine, surgery, and psychiatry fared as they sought to implement their guidelines. None fared well. Simply publishing the guidelines had no effect at all. The key elements seemed to be to produce guidelines that were simple, short, accurate, and current. Short meant on two sides of a plastic card, even if backed by a source book that contained the science. The content should be sponsored by a respected professional organization and the guidelines should represent a collaboration between all the stakeholders likely to use them, or be affected by them. They gain credence by being published in a reputable journal.

The rules for successful implementation of clinical practice guidelines are not at all clear, but look like a combination of hard work and marketing, done at a local hospital or regional professional organization level. *Getting Research Findings into Practice* (Haines and Donald 1998) is an excellent source book. The process is as follows: involve the key stakeholders and the clinical and management opinion leaders, have an accurate idea of what one wants to change, and a clear idea of the changes to organizational procedures, medical equipment, and clinical skills that will be necessary if implementation is to occur. Provide academic detailing to the clinicians so that all will have a chance to discuss and question the guidelines before implementation starts. Involve the consumers in the preparation of the guidelines and then in their implementation. Put the guidelines on the hospital or group's homepage, and 10 percent of consumers will consult these before attending. Use clinical audit to show that change is occurring and to show improved clinical outcomes. If the change does not improve clinical outcomes, ask whether further implementation is wise. Sustain the whole process by showing an improved cost effectiveness profile and arrange that some of the savings benefit the clinicians who brought this about. Expect that the implementation procedure will cost a considerable amount in staff time; the time costs for staff to attend meetings to produce a simple change to the way acute asthma was managed in one hospital exceeded \$50,000. The costs of implementation were perceived as inordinate even though that money was quickly recouped from the savings that accrued from the improved outcomes once the guidelines were implemented. If originators of the guide-

lines are not prepared to do all this work then they should simply enjoy the process of producing them, for the guidelines will not influence practice. This is not to deny that identifying empirically supported treatments carried out by a profession does not have important professional and political advantages for the profession. Funders, providers, and consumers all like to pretend that efficacy is the same as effectiveness, and lists of empirically supported treatments feed this delusion.

Efficiency: Efficiency or cost effectiveness is about the wise expenditure of scarce health resources. Costing studies are now common; the problem is that most are based on efficacy calculations in which coverage, compliance, and treatment integrity are all believed to be perfect. As effectiveness is much, much less than efficacy, the true costs per unit of benefit are much, much more than those identified in the costing studies. Australia was the first country to require costing data to be submitted when pharmaceutical companies sought to have a new drug placed on the list of subsidized medications. It has rightly been praised for this. Enquiries as to whether the data submitted are cost effective data or cost efficacy data are met with the assurance that the former is the case. As the data are 'commercial—in confidence' they are not amenable to 'freedom of information' enquiries. None are published and one might suspect that most are estimates of cost effectiveness under ideal conditions in which effectiveness is presumed to equal efficacy. As psychological therapies are not subsidized they are not subject to regulation, and randomized controlled trials about efficacy and cost are not required by any country. If we are to be anything more than a medicated society we do need such data.

League tables of the 'cost effectiveness' of medical interventions are beginning to appear. Few include mental health interventions. The median medical intervention cost in a study of 500 life-saving interventions was estimated as \$19,000 per life year saved. In mental disorders, the principal burden comes from the years lived with a disability, not from the years of life lost. Nevertheless, shadow prices under \$30,000 per disability-adjusted life year averted or quality-adjusted life year gained are said to be of interest to health planners. In specialized clinics, where costs and outcomes are known, it appears that drug and psychological therapies for the common anxiety and depressive disorders can be provided at little more than \$5,000 per life year gained. These figures are competitive and it might prove more advantageous if accurately costed league tables of empirically supported therapies in mental disorders were to be prepared by an organization of high repute. I cannot think of a greater stimulus to the implementation of evidence-based psychiatry than some appreciation of the good bargains available. After all, treatment as usual that produces little benefit is very costly to both funder and consumer. The providers benefit in the

short term, making the majority of their income from people who do not get better, but in the long term they become disillusioned by their lack of effectiveness. Evidence-based psychiatry can close the gap between research and practice, but it is not a trivial task. It can improve effectiveness to resemble efficacy and so result in better outcomes for patients, and that is not a trivial achievement. Isn't that what mental health services are for?

See also: Behavior Therapy: Psychiatric Aspects; Clinical Treatment Outcome Research: Control and Comparison Groups; Cognitive and Interpersonal Therapy: Psychiatric Aspects; Health Outcomes, Assessment of; Psychiatric Care, Economics of; Psychological Treatment, Effectiveness of; Psychological Treatments: Randomized Controlled Clinical Trials; Truth and Credibility: Science and the Social Study of Science; Truth, Verification, Verisimilitude, and Evidence: Philosophical Aspects

Bibliography

- Andrews G 1999 Efficacy, effectiveness and efficiency in mental health service delivery. *Australian and New Zealand Journal of Psychiatry* **33**: 316–22
- Andrews G, Henderson S (eds.) 2000 *Unmet Need in Psychiatry*. Cambridge University Press, Cambridge, UK
- Bobadilla J L, Cowley P, Musgrove P, Saxenian H 1994 Design, content and financing of an essential package of health services. In: Murray C J L, Lopez A D (eds.) *Global Comparative Assessments in the Health Sector*. World Health Organization, Geneva, pp. 171–80
- Cochrane A L 1989 *Effectiveness and Efficiency: Random Reflections on Health Services*. British Medical Journal, London
- Deyo R A, Psaty B M, Simon G, Wagner E H, Omenn G S 1997 The messenger under attack: Intimidation of researchers by special-interest groups. *The New England Journal of Medicine*. **336**: 1176–80
- Enserink M 1999 Can the placebo be the cure? *Science* **284**: 238–40
- Haines A, Donald A 1998 *Getting Research Findings into Practice*. BMJ Books, London
- Nathan P E, Gorman J M 1998 *A Guide to Treatments that Work*. Oxford University Press, New York
- Nathan P E, Gorman J M, Salkind N J 1999 *Treating Mental Disorders: A Guide to What Works*. Oxford University Press, New York
- Quality Assurance Project 1983 A treatment outline for depressive disorders. *Australian and New Zealand Journal of Psychiatry* **17**: 129–46
- Sackett D L, Rosenberg W M C, Gray J A M, Haynes R B, Richardson W S 1996 Evidence based medicine: What it is and what it isn't. *British Medical Journal* **312**: 71–2

G. Andrews

Psychiatry, History of

Attempts to heal madness go back in Western medicine at least to Hippocrates. However psychiatry as a medical specialty—that is, a gainful form of medical practice that occupied its practitioners full-time and to the exclusion of other branches of healing—and, hence, as a self-conscious scientific project, was the invention of the late eighteenth century. It had, at the level of ideas, two major prerequisites: a new technique for treating madness, called the moral treatment, believed to effect cure more reliably than anything that had preceded it; and a new institution, the asylum, conceptualized both as the appropriate setting for the cure of madness and as itself an active instrument of cure. While the name eventually given to the specialty in most countries was of German coinage (*Psychiatrie*, first used in 1808 by Johann Christian Reil), the practical initiatives critical in establishing psychiatry came from England and France.

1. Moral Treatment

Psychiatry's founding therapeutic strategy was of combined Anglo-French provenance and as such became the subject of a muted, international priority dispute. cursorily defined, the moral treatment for the cure of insanity meant the use of methods that engaged or operated on the intellect and emotions, as opposed to the traditional methods of bleedings, purges, and emetics applied directly to the lunatic's body. By all accounts, the British cleric and physician Francis Willis had employed moral methods when in 1788–9 he assumed the weighty responsibility of treating the madness of King George III. His contemporary, the French physician Philippe Pinel, duly credited Willis and others with having inspired his own efforts. But when codifying the moral treatment in his *Traité médico-philosophique sur l'aliénation mentale* (1801), Pinel also accused his British colleagues of willfully failing to spell out the method they so enthusiastically recommended. For their part, British reviewers of Pinel's *Traité* dismissed the Frenchman's supposed innovations as pale imitations of British practice.

Wherever priority lay, the moral treatment proved notoriously resistant to precise definition. Contrary to his avowed intention, Pinel never provided an explicit definition in his *Traité*. He instead presented a series of case histories of his own use of the treatment, leaving to his readers the task of inferring its central features. These included:

(a) A preference for 'gentleness' as a therapeutic technique. This choice was rooted in an optimistic Enlightenment belief that insanity never involved a complete annihilation of reason, reducing the mad person to animality, but that a glimmer of 'humanity'

always remained and could be appealed to and reinforced.

(b) The use of 'repression' only as a last resort and then without anger or cruelty. When acting in a repressive capacity, the healer of the insane was expected to exude authority through his physical presence—Willis, for example, was known for his ability to command the obedience of the mad by his piercing gaze.

(c) The grounding of the treatment in an etiology of insanity based on the sensationalist psychology of John Locke and the abbé de Condillac. Regarded as a set of erroneous ideas that had become implanted firmly in the mind, insanity could be cured by a jolting experience that dislodged those ideas and enabled their replacement by sound ones. Hence the theatrical aspects of Pinel's moral treatment: vivid scenes were staged deliberately to impress on mad persons the falsity of their pathological beliefs (Goldstein 1987, Porter 1987).

Clearly, such a therapeutic strategy stamped psychiatry as a humanitarian endeavor and a product of the Enlightenment faith in science. In France, these affiliations were strengthened further by the link between psychiatry and the French Revolution. The Revolutionary government put Pinel in control of the Paris institutions for the insane, and according to a myth that took shape over the course of the nineteenth century, and has only recently been debunked by historians, he inaugurated the moral treatment by literally unchaining the inmates of Bicêtre, an emancipatory gesture that cured them instantaneously. In the USA, the chief proponent of the moral treatment, Benjamin Rush, was so ardent a supporter of the American Revolution that he even figured among the signatories of the Declaration of Independence (Binger 1966). In Britain, ratification of the moral treatment took place in the context of parliamentary investigations into the civil liberties violations entailed in the maltreatment of incarcerated lunatics (Scull 1975).

But if the moral treatment had strong associations with Enlightenment and political liberty, it also lent psychiatry an unscientific taint that proved difficult to eradicate. For, as its proponents liked to point out, it had originated among healers without official educational credentials. Pinel was a bona fide physician but, in keeping with the hostility of the French Revolution toward all privileged corporations, he proudly announced that neither he nor any other member of a medical faculty had devised the moral treatment, that it was the work of a humble layman who served as hands-on caretaker of the insane at Bicêtre. His own contribution was, he said, to have rendered the caretaker's version of the treatment scientific, translating it into the language of sensationalist psychology and demonstrating its efficacy with the new mathematical tool of statistics. Reliance on the moral treatment left Pinel's successors vulnerable to competition from traditional Catholic healers of the

insane, who argued persuasively that medical men had no special skills in administering a noncorporeal remedy. Indeed in France the dispute between alienists and clerics for control of the treatment of insane continued until the end of the nineteenth century, when the anticlerical policies of the Third Republic decided conclusively in favor of the former (Goldstein 1987).

In Britain, it was a layman, the tea merchant William Tuke, who instituted the moral treatment at the York Retreat, the influential Quaker facility for the insane. Tuke was highly vocal about his distrust of medical practitioners pretending to treat insanity with Physick. After a long period of uneasy debate between the models of medical and moral therapy, the official line of British psychiatry became in the 1840s that a judicious combination of both had greater efficacy than the sole use of either. In adopting this position, British psychiatrists shrewdly obviated the protest that clergymen were better equipped than physicians to handle insanity (Scull 1975).

2. The Asylum

The insane asylum is one of a group of institutions, which, in the 1960s, the American sociologist Erving Goffman (1961) labeled 'total' and, in the 1970s, the French philosopher Michel Foucault (1975) described as 'panoptic.' For Goffman, such institutions are characterized by locating the ordinarily compartmentalized activities of sleep, play, and work under a single roof and a single authority and by erecting barriers, both legal and architectural, to social intercourse with the outside world. For Foucault, they are characterized by the incessant visibility of the inmates to the authorities, a visibility redounding to the power of the latter. Both theorists stressed the tight schedule of activities enforced in such institutions. Both noted their proliferation in the West during the late eighteenth and early nineteenth centuries, the period that saw the birth of prison and poorhouse as well as insane asylum.

In itself, the incarceration of deviant populations was no novelty. At least since the 'Great Confinement' of the mid-seventeenth century, western European governments, especially the absolutist monarchies of the continent, had made a concerted effort to establish good order by removing troublemakers from social circulation. The novelty lay rather in an unprecedented confidence in the rehabilitative power of certain forms of incarceration and in a fresh taxonomic sensibility that led to the division and strict segregation of deviants according to type. By contrast, the French *hôpital général* of the age of Louis XIV and its counterparts, the British workhouse and the German *Zuchthaus*, had entertained no therapeutic ambitions and had lumped together madness, idleness, and

debauchery as fully comparable species of unreason (Foucault 1961).

The classical scientific rationale for the asylum was provided in 1819 by Pinel's student J. E. D. Esquirol, who also appropriated the old term 'asylum' (*asile*), meaning a place of refuge, to underscore the benevolent ethos of the new institution. The asylum could cure insanity, said Esquirol, because it temporarily removed mad persons from their habitual surroundings. By dint of this 'isolation,' as he called it, the institution operated as an instance of the moral treatment: the psychic jolt of sensory stimuli arising from a radically altered physical and interpersonal environment dislodged the patient's erroneous ideas, which could then be replaced by rational ones through the expert interventions of the psychiatric doctor. The latter's therapeutic authority was enhanced by the fact that, in the artificial, self-contained world of the asylum, he ruled absolutely (Goldstein 1987).

Like the embrace of the moral treatment, so, too, the wedding of psychiatry to the institutional form of the asylum imposed special professionalization problems on the emergent specialty. In order to survive collectively, psychiatrists needed the support of a governmental or private agency willing to make large capital investments in the construction of asylums and motivated to place the psychiatrist at the top of the asylum's hierarchy of personnel. These goals were first achieved in France, where Esquirol and his disciples joined with moderate liberal politicians to secure passage of the Law of 1838. The Law mandated the establishment of a nation-wide network of asylums headed by psychiatrists, who thereby became salaried functionaries of the French state. In the less centralized British state, psychiatric victory came more gradually. An Act of 1808 providing for the construction of asylums financed by the county rate made no provision for the inclusion of a medical man on the staff. Only with the passage of the Lunatics Act of 1845, which placed institutions for the insane under a permanent national inspectorate composed of lay, legal, and medical commissioners, did the British introduce even an implicit force for the medicalization of the asylum (Jones 1955, Scull 1975).

A combined impulse toward rational state-building and enlightened humanitarianism led the disunited German lands of the first half of the nineteenth century to a spate of new-style, public asylum foundings as well as the medical restructuring of old-style, multi-functional absolutist institutions of incarceration (Goldberg 1999). In the USA, the asylum-building vogue of the same period occurred in two phases: first, emulation of European models in new, privately funded eastern-seaboard institutions; then, goaded by the Jacksonian-era critique that the frenetic competition for wealth and power in democratic America actively bred insanity, the establishment of public asylums by most state legislatures. American asylums were run by medical superintendents—the early name

for psychiatrists in the USA—and were so plentiful that an Association for Medical Superintendents was formed in 1844 (Rothman 1971).

3. Late Nineteenth-century Reconfigurations

The second half of the nineteenth century saw important changes in the institutional and intellectual models that had launched psychiatry. Most dramatically, the prevailing optimism about the therapeutic effect of the asylum gave way to deep pessimism as asylum populations swelled with chronic patients demonstrably impervious to moral treatment (Lantéri-Laura 1972, Scull 1979, Rothman 1971). At about the same time, psychiatrists abandoned their earlier conception of insanity as a curable psychological or psychosomatic disorder and theorized it as an irreversible brain condition and often as a product of 'degeneration.' This degeneration was defined as a pathological departure from the norm initially caused by a noxious environment, poor nutrition, or alcoholism, and subsequently transmitted in the Lamarckian manner through heredity, becoming more severe with each generation. Every European nation had its *fin-de-siècle* theorists of degeneration: Bénédict-Augustin Morel and Valentin Magnan in France, Cesare Lombroso in Italy, Henry Maudsley in Britain, Richard von Krafft-Ebing and Max Nordau in Austria (Pick 1989).

The shift to an organicist conception of insanity was coupled with a shift in psychiatry's international center of gravity from France to Germany. By the mid-nineteenth century, German physicians had thoroughly repudiated an earlier affinity for Romantic, holistic *Naturphilosophie*. Their new taste for positivism was epitomized in psychiatry by Wilhelm Griesinger's textbook, *Pathologie und Therapie der psychischen Krankheiten* (1845) that equated psychiatry with the study of cerebral lesions. Griesinger also epitomized the era's new psychiatric career pattern: he was not an asylum doctor but a university psychiatrist, having been appointed in 1865 to the recently created Chair of Psychiatry and Neurology at the University of Berlin (Decker 1977). Later adopted in other countries, the university psychiatry model increased the attractiveness of the psychiatric career, offering the practitioner a less reclusive lifestyle than was afforded by residence in an asylum. It also changed the intellectual content of the career: psychiatric knowledge was now assumed to be generated less from clinical contact with the insane than from neuroanatomical research in a university laboratory.

A second late nineteenth-century reconfiguration likewise encompassed both psychiatry's scientific theory and its career structure. Asylum psychiatry had worked with a simple binary opposition between sanity and insanity. After 1850, however, psychiatrists began to speak of a large intermediate zone occupied by the *demi-fous*, or half-mad. This conceptual move

eventually led to the articulation of two categories fundamental to the enterprise of twentieth-century psychiatry: the 'psychotic,' whose contact with reality is severely ruptured, and the 'neurotic,' whose minimal maladaptations do not preclude getting on in ordinary society. The former, requiring institutionalization, fit the older model of asylum psychiatry. The latter, who could safely remain in the outside world but would benefit from scheduled visits to the doctor's office, made possible the practice of outpatient psychiatry.

A rough international division of labor governed research on the neuroses and psychoses. The Parisian Jean-Martin Charcot and his many students dominated the study of the less severe forms of mental illness, especially hysteria. Munich-based Emil Kraepelin and his school continued German psychiatric preeminence through their domination of the study of the psychoses. Kraepelin made a psychiatric staple of the diagnostic category, *dementia praecox*, a state of emotional detachment and vacuousness that declared itself in adolescence and carried an invariably bleak prognosis. (The Swiss psychiatrist Eugen Bleuler would in 1911 rename the condition schizophrenia and insist that it might have a favorable outcome.) The Kraepelinian approach held sway in many countries, including the USA, through the first quarter of the twentieth century.

4. *Twentieth-century Trends*

In broad outline, three trends distinguished psychiatry in the twentieth century: the emergence of Freudian psychoanalysis; the discovery and use of psychotropic drugs; and the antipsychiatry movement.

Created by the Viennese neurologist Sigmund Freud as an outpatient treatment for neurotic disorders, psychoanalysis was initially distant from institutional psychiatry. Indeed, Freud's professional overtures to Bleuler and Carl Jung, Bleuler's assistant at the Burghölzli asylum in Zurich, can be read as efforts to obtain a psychiatric imprimatur for the new discipline of psychoanalysis (Freud 1914). Freud never fulfilled this ambition in Europe, but in the less tradition-bound scientific environment of the USA, many professors and medical superintendents welcomed his ideas. Eventually, American psychiatry fully incorporated psychoanalysis. In 1924, when William Alanson White, who had introduced psychoanalytic psychotherapy for schizophrenic patients at St. Elizabeth's Hospital, became president of the American Psychiatric Association, that organization and the American Psychoanalytic Association began holding their annual meetings together. This arrangement lasted until 1974, by which time the Freudian ardor of American psychiatrists had cooled (Barton 1987).

Waning interest in psychoanalysis on the part of American psychiatrists coincided with rising expectations for psychopharmacology on the part of the international psychiatric community. The first psycho-

tropic drug, lithium, was used successfully to manage manic-depression in Australia in 1949. A variety of antipsychotic and antidepressant compounds, most notably the phenothiazines and imipramine, were developed in France and Switzerland in the early 1950s (Baldessarini 1977). The fluoxetine family of antidepressants, whose first member went under the brand name Prozac, was introduced in the USA in the late 1980s and became among the most widely prescribed drugs in the world. Psychopharmacology brought a new wave of self-confidence and therapeutic optimism to psychiatry, promising a relatively safe, speedy, and cost-effective method of alleviating mental illness. It also revived psychiatry's identity as a 'hard' and exact science, the identity cultivated in the second half of the nineteenth century.

Somewhat ironically, then, the reliance on psychotropic drugs was one of the causes of the antipsychiatry movement of the 1960s. Since the beginning of the specialty, psychiatry's incarcerative practices had given rise to sporadic criticism, and some psychiatric abuses, notably the 'euthanasia' program mounted by the Nazi regime during World War II, have only been fully uncovered after the fact (Burleigh 1994). But the antipsychiatry movement of the 1960s, centered on the London-based psychiatrist R. D. Laing, questioned the specialty even more radically and globally. It charged that madness was neither an objective biochemical nor behavioral condition, but a context-dependent label; that Western rationalism was predicated on a denial of its value; and that psychopharmacology robbed patients of the potentially enriching and self-healing experience of full-blown psychosis. Laing's views were supported by the psychiatrist Thomas Szasz in the USA and to some degree by the philosopher Michel Foucault in France. Antipsychiatry gave impetus to the deinstitutionalization of the insane, which occurred in the late 1970s and the 1980s. A less controversial outcome of the movement was the recognition of the patient's right to refuse antipsychotic drugs and electroshock therapy (Sedgwick 1982).

The antipsychiatry movement has now largely spent its force. But as the twenty-first century begins, psychiatry, while among the oldest of the medical specialties, lacks the stability that age might seem to confer. It remains hostage to the mind-body problem, buffeted between psychological and physical definitions of its object and its methods.

See also: Deinstitutionalization, Psychiatry of; Mental Health and Normality; Philosophy, Psychiatry, and Psychology; Psychiatric Care, Economics of; Psychiatry and Sociology; Psychiatry: Anthropological Aspects; Psychoanalysis, History of; Psychoanalysis in Clinical Psychology; Psychoanalysis in Sociology; Psychologism, History of; Psychology: Overview; Psychometrics; Psychophysics; Psychotherapy, History of; Psychiatric Aspects; Social Psychology

Bibliography

- Baldessarini R J 1977 *Chemotherapy in Psychiatry*. Harvard University Press, Cambridge, MA
- Barton W E 1987 *The History and Influence of the American Psychiatric Association*. American Psychiatric Press, Washington, DC
- Binger C 1966 *Revolutionary Doctor: Benjamin Rush, 1746–1813*, 1st edn. Norton, New York
- Burleigh M 1994 *Death and Deliverance: 'Euthanasia' in Germany 1900–1945*. Cambridge University Press, Cambridge, UK
- Decker H S 1977 *Freud in Germany*. International Universities Press, New York
- Foucault M 1961 *Histoire de la Folie*. Plon, Paris [English translation 1965 *Madness and Civilization: A History of Insanity in the Age of Reason*. Random House, New York]
- Foucault M 1975 *Surveiller et Punir: Naissance de la Prison*. Gallimard, Paris [English translation 1977 *Discipline and Punish: The Birth of the Prison*. Pantheon, New York]
- Freud S 1914 On the history of the psycho-analytic movement. *The Standard Edition of the Complete Psychological Works of Sigmund Freud*. Hogarth Press, London, vol. 14
- Goffman E 1961 *Asylums: Essays on the Social Situation of Mental Patients and Other Inmates*, 1st edn. Anchor Books, Garden City, NY
- Goldberg A 1999 *Sex, Religion and the Making of Modern Madness: The Eberbach Asylum and German Society, 1815–1849*. Oxford University Press, New York and Oxford, UK
- Goldstein J 1987 *Console and Classify: The French Psychiatric Profession in the Nineteenth Century*. Cambridge University Press, New York and Cambridge, UK
- Jones K 1955 *Lunacy, Law and Conscience, 1744–1845*. Routledge & Kegan Paul, London
- Lantéri-Laura G 1972 La chronicité dans la psychiatrie française moderne. *Annales ESC* 27: 548–68
- Pick D 1989 *Faces of Degeneration: A European Disorder, c. 1848–1918*. Cambridge University Press, Cambridge, UK
- Porter R 1987 *Mind Forg'd Manacles: A History of Madness in England from the Restoration to the Regency*. Harvard University Press, Cambridge, MA
- Rothman D J 1971 *The Discovery of the Asylum: Social Order and Disorder in the New Republic*, 1st edn. Little, Brown & Co., Boston
- Scully A T 1975 From madness to mental illness: medical men as moral entrepreneurs. *Archives Européennes de Sociologie* 16: 218–51
- Scully A T 1979 *Museums of Madness: The Social Organization of Insanity in Nineteenth-century England*. St. Martin's Press, New York
- Sedgwick P 1982 *Psycho-Politics: Laing, Foucault, Goffman, Szasz and the Future of Mass Psychiatry*, 1st edn. Harper and Row, New York

J. Goldstein

Psychiatry: Informed Consent and Care

Informed consent is a philosophical and legal doctrine that has come to serve as the cornerstone of ethically sound clinical care in Western society. This doctrine requires that individuals truly understand and freely

make decisions regarding their treatment. Informed consent is an ideal that emphasizes respect for the individual and places primacy on autonomous decision-making. Furthermore, it is inherently relational, that is, it occurs in the context of a professional relationship in which the fundamental values of the medical profession find expression. Fulfilling the ideal of informed consent in practice is an elusive aim, however. As suggested by intuition, clinical phenomenology, and empirical evidence, it is particularly challenging when consent is sought from individuals who suffer from mental illness. For this reason, it is important to consider the specific issues arising in informed consent in the context of mental health care. This chapter provides a clinically-oriented overview of informed consent in mental illness by: (a) characterizing the core conceptual elements of informed consent; (b) applying informed consent concepts in mental health clinical practice; and (c) giving a brief overview of key empirical data related to mental illness in clinical consent decisions.

1. Elements of Informed Consent

Informed consent does not occur in isolation or in a moment, nor is it reducible to a signature on a legal document. It is a shared, significant process between people—one of whom faces a series of evolving choices critical to his or her mental and physical health, the other of whom is a professional person entrusted with helping to define these choices astutely and present them truthfully, respectfully, and compassionately. In clinical care, promoting the health and alleviating the suffering of the individual patient will be the basis of any treatment considered in the consent process. Consent in this situation focuses on the patient's personal values, is based on the overall goals for care, and hinges on trust in the clinician–patient relationship.

1.1 Information

The sharing of appropriate information is the first of three core elements of informed consent. The patient should be provided with accurate and balanced information in understandable language. This information should include the anticipated risks and benefits of a given treatment, the procedures involved, and its alternatives and their expected outcomes. Alternatives may include implementation of another treatment or additional treatment, or it may involve no intervention upon the course of illness at all. While every minute detail cannot be reviewed in this discussion, rare but serious risks should be characterized in a manner that conveys their relevance and reality. Furthermore, the omission of information in the disclosure process should not be intended to deceive or mislead the person undertaking the decision.

Ideally information related to a matter requiring consent will be disclosed in a manner that fosters dialogue, allows for clarification of complex issues and the identification of misunderstandings, and keeps opportunities open for subsequent discussion as new issues emerge. In obtaining informed consent for treatment of mild depression, for instance, the clinician should discuss the variable course of depressive illnesses, including the potential for symptom resolution in the absence of treatment and the chance of recurrence without intervention. The clinician should review the rationale for a proposed antidepressant agent given the symptoms of the patient and the patient's attitudes and values. The discussion should cover the medication's known possible side effects, potential adverse effects, and their relative likelihood. The clinician should explicitly anticipate the choices that will be faced by the patient in the future in light of the therapeutic alternatives discussed and the nature of the illness. This discussion should be documented by the clinician as an important element of the care process.

Consent forms that accurately and completely describe a treatment and outline patient rights represent an important tool in the consent process. Consequently, consent forms should be developed carefully in terms of their understandability (e.g., simple but precise language), readability (e.g., size of font), content (e.g., purpose, risks and benefits, rights, responsibilities of the clinician/researcher), and related issues. Consent forms, furthermore, signify the importance of a new decision for the patient, and they give specific focus to the dialogue between the professional and the patient. Even general consent forms such as those for hospitalization or for initiating medication—which may or may not confer legal protection when a medico-legal conflict occurs—have this value. From a legal perspective, notes written by the clinician reflecting the disclosure dialogue with the patient may be far more salient and can carry greater weight.

1.2 Decisional Capacity

Decisional capacity, the second element of informed consent, represents a sophisticated clinical assessment of an individual's global and specific cognitive abilities within a particular context and in relation to a particular choice. It is distinct from 'competence' which is a legal determination, i.e., a formal assessment by a judge or other officer of the court regarding a person's ability to function within a general domain of one's life such as financial or medical decisions (Appelbaum and Gutheil 1991). The presence of a serious symptom (e.g., delusions, memory loss, suicidal ideation, diminished cognitive functioning) or of a serious diagnosis (e.g., schizophrenia, dementia, major depression, developmental disability) does not

mean that an individual lacks decisional capacity. Similarly, being a young adolescent or a very elderly person does not, *a priori*, preclude one from being decisionally capable. In the context of mental illness, decisional capacity assessment may be subtle, complex clinical work that entails constructing a cohesive sense of the individual's strengths and deficits in light of a number of factors. The type of decision to be made, the nature, time course, and severity of the person's symptoms, the patient's situation, and past experiences, personal values, psychological defenses or coping 'style,' and other factors should all be considered.

Decisional capacity itself has four principal components, each of which should be viewed on a continuum and not as 'all or none' phenomena (Appelbaum and Grisso 1988). The first component is the individual's ability to express a preference (i.e., capacity for expression or communication). On the extreme end of the continuum will be severely ill psychiatric patients who are mute or catatonic, or who manifest complete language and thought disorganization, and therefore may be intermittently 'incapable' of communication. In the medical context, a patient who is comatose due to metabolic disturbances, aphasic secondary to a stroke, delirious, or unconscious after a head injury will also lack this most fundamental ability. There are times when this component of decisional capacity is very difficult to assess, such as in the spinal cord injured patient who communicates only through eye blinks, or the dysphoric person with Broca's aphasia who is frustrated by the process of generating words but has the capacity to understand and indicate choices. For these reasons, while the ability to communicate is often obvious, this assessment in many situations will require careful interactions with patients and astute observational and listening skills on the part of the clinician.

The second component of decisional capacity is the individual's ability to understand relevant information (i.e., capacity for comprehension). In talking with the clinician, the patient with this ability will be able to take in factual information regarding the nature of the illness, the need for intervention, the kinds of treatments under consideration, their relative risks and benefits, and the possible consequences of not intervening. Comprehension is influenced by intelligence, educational level, and prior experience. In assessing this dimension, it is critical that the patient state his or her understanding of the situation and the choices at hand in their own words. Otherwise it is difficult to know if the patient is using terms and phrases of the clinicians without truly knowing their meaning.

Third is the individual's ability to think through the decision in a way that is 'rational' and deliberate (i.e., capacity for reasoning or rational thinking). This component of decisional capacity relates to the person's analytical abilities given a specific set of facts or data and, like comprehension, is affected by

intelligence and educational level. For instance, the 'rational' person, all other factors being equal, will consistently choose a medication that has a 75 percent likelihood of treating the illness effectively over one that has a 50 percent chance. Similarly, the 'rational' person will consistently choose a medication that poses fewer risks and promises fewer adverse side effects than another medication. In essence, this dimension of decisional capacity has to do with the ability to use objective information.

Fourth, and often most important, is the person's ability to make meaning of the decision and its consequences within the context of his or her personal history, values, emotions, and life situation (i.e., capacity for appreciation). Deeply dependent upon psychological insight, this is the most sophisticated component of decisional capacity. It can be very difficult to assess clinically. In some individuals with early dementia, for instance, a person's social graces and cognitive skills may be only superficially affected and yet these individuals may be unable to think through the meaning of an important treatment decision and its repercussions. In some patients suffering from either anxiety or affective disorders, the negative cognitive distortions inherent to their illnesses may interfere with their abilities to interpret information within a personal context, although other components of decisional capacity may be intact. Importantly, in people with severe mental illnesses such as schizophrenia, the simple presence of psychotic symptoms such as hallucinations or delusions may *not* interfere with the individual's overall decisional capacity, but the cognitive and insight deficits associated with the illness may interfere with their capacity for appreciation of the situation at hand. Alternatively, even the most gravely ill individuals may have difficulty expressing themselves and an incomplete sense of the factual information but may truly appreciate the severity of the situation and possess a deep and consistent sense of who they would like to make decisions for them.

Decisional capacity is therefore a rich construction, not simply reducible to an individual choosing a predetermined 'right answer' when making a decision of importance. It is the person's ability to gather, assimilate, and use information, to truly see the reasons for and against a given option, and to develop a meaningful decision in the larger context of one's life. The seriousness and complexity of any specific decision will govern how rigorously these criteria are applied. In the context of serious mental illness, a person may be fully capable of deciding whether or not to begin a new psychotropic medication but, on a given day, may be incapable of deciding whether to enroll in a clinical protocol involving a medication-free interval and potential randomization to placebo. The standard for decisional capacity is also robust when a very ill patient refuses treatment that is highly recommended because of the benefit it is likely to bring

to him (see discussion in Sect. 2). This is not to say that patients cannot refuse treatment; when the risks of refusal are great, however, there is a greater responsibility on the caregiver to assure that the patient is genuinely capable of this decision and to manage its consequences.

1.3 Voluntarism

Beyond information disclosure and decisional capacity, the final element of informed consent is voluntarism or autonomy. Autonomy represents the expression of the authentic preferences of the individual in the absence of coercive pressures. Voluntary preferences may be tempered by the realities of the situation such as the presence of a serious diagnosis and the lack of 'curative' treatments for many mental illnesses, but should reflect the person's sustained and genuine wishes as evidenced in their expressed values and prior life decisions. Autonomous wishes may be deeply influenced by social and cultural expectations and psychological history. For instance, a patient who has been in the military may be oriented to hierarchical power relationships and may view a clinician's recommendation as an 'order' to be obeyed. Similarly, a seriously mentally ill person who has been institutionalized involuntarily in the past may see themselves as no longer possessing the right to their own choices. Illness processes may also affect patient preferences as well as the ability to express autonomous wishes. A person who suffers deeply either psychologically or physically may feel desperate and desolate, phenomena that can generate inherent internal coercive pressures. A person who has been the victim of violence may see themselves as relatively powerless. Finally, clinical syndromes that involve apathy, alexithymia, amotivation, or cognitive slowing may render ill individuals less able to identify and express their true wishes.

2. Clinical Application of Informed Consent Concepts

Translation of informed consent concepts to clinical practice is affected by the clinical status and situation of the patient. In emergencies, for instance, there may be little time for information-sharing and mutual decision-making and the standard for consent is less rigorous than in other contexts. Acting beneficently, i.e., to save the life or preserve the health of the patient in an acute situation, is the clinician's duty in such circumstances. This is held to be critical ethically and justifiable legally; it is presumed that in such emergency situations the patient wishes to be helped and that the clinician has permission to act with the patient's well-being and best interests at heart. Often the patient, if alert and conscious, will explicitly assent to treatment in these circumstances, i.e., voluntarily accept the choice proffered by the clinician at the time. If the patient is unconscious or is unable to com-

municate at all, the clinician acts with the 'implied consent' of the patient.

When the patient refuses treatment in an emergent situation, however, a very rigorous standard for consent must be applied because of the likelihood of an adverse outcome and because there are many clinical factors that may be adversely affecting the judgment of the patient. Undiagnosed illness, delusional thinking, severe physical pain, emotional distress, and interpersonal pressure all may undermine the ability of the patient to make a fully informed, autonomous choice in an acute situation. Evaluation, hospitalization, and treatment, even against the apparent wishes of the patient, may be essential in such circumstances. Consider two illustrative patient cases:

Case Example 1. A man diagnosed with a serious mental illness characterized by psychotic and mood symptoms was brought to a psychiatric crisis unit after being found wandering on the street near his home. It was discovered that he had superglued his eyelids shut, and it was imperative that he receive medical attention immediately. He was outwardly calm, did not speak, answered questions through nodding, writing, and hand gestures, but actively refused to get into an ambulance to be taken to the medical emergency room. The clinician on duty that evening knew the patient well and was able to deduce that the patient's peculiar behavior was linked to a recurrent delusion he had when entering a manic period. In this case, the patient's ability to make a fully informed decision was impaired by his delusional belief that if he looked directly at people, they would be instantly destroyed by powerful lasers that had been implanted in his eyes. Once the clinician initiated a conversation about these beliefs with the patient, it was possible to clarify that the patient wanted help and that he was willing to accept treatment, but was terrified that he would hurt people or himself be killed. The patient thus had some decisional ability and could offer assent for emergency care. The clinical team opted to pursue a legal process that would allow involvement of a time-limited treatment guardian due to concerns over the patient's compromised insight and judgment and the severity of his delusions.

Case Example 2. An agitated man demanded discharge from a surgical unit two days after an emergency appendectomy. He had elevated heart rate, blood pressure, and temperature, had signs of abdominal infection, and yet was insistent that he leave the hospital. The medical team was conflicted about granting the patient an 'AMA discharge' ('Against Medical Advice')—the nurses and residents believed this to be the patient's 'right to refuse treatment' but the attending clinician did not agree. On initial psychiatric evaluation, the patient showed irritability and very subtle indications of diminished attention, was oriented except to day of the week, and had no other apparent abnormalities on an abbreviated mental status examination. He was noted to be tremulous.

He was placed on involuntary 'hold' status. Within an hour, the patient was agitated, disoriented, and expressed disorganized thoughts. Additional history was then obtained that the patient drank one to two pints of vodka daily prior to admission. Post-surgical infection in combination with alcohol withdrawal had triggered a significant delirium, an early indication of which, ironically, was his treatment refusal and discharge demand. His most distressing symptoms and disorientation resolved rapidly with treatment, and he withdrew his request to leave the hospital.

When a patient refuses treatment there may be a significant clinical issue that requires close attention, and involuntary treatment may be imperative. Clarification of underlying clinical considerations in such circumstances is the caregiver's primary responsibility—a duty grounded in the bioethics principles of respect for persons, clinical standards of care, fidelity, and beneficence. Careful evaluation of the patient's true strengths and abilities is important so that he or she may exercise choice in every area possible. The intent is not to undermine the patient's autonomy. Rather, it is to empower the clinician only insofar as to help advance the health and well-being of the patient and to assure that the patient is optimally able to express his or her true wishes to accept or decline a given therapeutic approach. In sum, in situations where clinical decisions may have very serious or irreversible health consequences, including death, informed refusal of treatment must fulfill the most stringent standards.

The goal of informed, voluntary decision-making may be more clearly fulfilled in the context of care for chronic illness. Over the course of time, the clinician may come to know the patient—their symptoms, their values and concerns, their responsiveness to treatment interventions, and their personal supports. In turn, the patient may have a greater sense of trust, express themselves with greater openness, and gather more complete information about the illness and potential therapeutic interventions. The patient with a chronic illness also lives out the consequences of side effects, adverse reactions, and benefits of treatments in a manner different from the patient with a time-limited health problem. In this context, the patient may better appreciate how their illness and treatment affect other aspects of their life, and it is possible to come to know their preferences and concerns with greater certainty. Expressing autonomous wishes and deciding upon specific treatment approaches and the overall goals of therapy through mutual understanding and dialogue are more intrinsically possible in the context of chronic illness. The standard for consent in chronic illness, thus, is very rigorous in terms of the patient's level of information, opportunities to enhance decision-making, and creating a circumstance in which patients may indicate their wishes and hopes.

In clinical practice, psychiatric advance directives may be valuable in translating concepts of informed

consent to patient care decisions. Advance directives articulate patient preferences for future care. Advance directives are often formal written papers created and signed by the patient; they may also consist of a body of documentation developed by the patient, family members, and clinicians regarding his or her wishes and values concerning personal health care. Advance directives are helpful in that they serve as a reference for the clinician if the patient becomes decisionally incapable or compromised. For example, knowing that they had a pattern of developing compromised judgment during periods of mania, a patient with bipolar disorder completed an advance directive requesting that treatment be instituted despite their (anticipated) protests. Advance directives also provide evidence of the patient's past personal healthcare preferences that may result in controversial clinical decisions, e.g., when patients refuse treatment based on deeply and consistently held personal convictions such as the refusal of blood transfusions by a patient who is a Jehovah's Witness. Ideally, these issues will not be addressed for the first time in a moment of great crisis. They should be anticipated, discussed, and documented carefully beforehand. Advance directives do not, under any circumstances, replace the wishes expressed by a patient who is currently decisionally capable. Moreover, there should be opportunities to adapt, or even reverse, earlier decisions as one cannot always tell prospectively how one will respond in a crisis situation. For these reasons, advance directives may guide but can never supplant prudent, reflective clinical decision-making.

In surrogate decision-making, healthcare choices are made for the decisionally impaired patient by another person. This person may be a spouse, sibling, adult child, friend, or a court-appointed guardian. The surrogate decision-maker has the responsibility of serving the best interests and well-being of the patient while seeking to make decisions that are faithful to the personal values and prior wishes of the patient. This can be difficult. In some situations, the patient's wishes are not known. In other situations, the patient is too young or too cognitively compromised to have demonstrated their personal values. In yet other situations, these values and wishes are known and appear to conflict with accepted standards of clinical care. Consequently, in working with surrogate decision-makers, it is important for clinicians to clarify the goals and value system governing the choices made on behalf of the patient. It is also crucial that the patient's decisional capacity be assessed accurately and, in most circumstances, the patient's symptoms be treated aggressively so that they can contribute to the choices made to whatever extent possible.

3. Empirical Studies on Consent in Clinical Care

Nearly 400 studies performed since 1970 suggest that informed consent is an ideal that is often difficult to

achieve in clinical and research practice involving people with physical and mental illness (Sugarman et al. 1999). These data indicate, however, that issues faced in clinical care for mental illness are not radically different in degree or kind from those encountered in the treatment of physical illness (Stanley et al. 1988). Moreover, an extensive body of evidence on consent consistently demonstrates the poor readability, understandability, and ineffectiveness of formal consent documents in clinical care (Grossman et al. 1994). An important recent advance in the informed consent field has been the development of the MacArthur Competence Assessment Tool for Treatment (MacCAT-T) (Grisso et al. 1997). This is a structured interview designed to assess patients' clinical care decision-making abilities based on empirical work conducted over ten years by Appelbaum and Grisso. While this instrument is best utilized as a tool for approaching decisional capacity evaluation, not as a definitive 'test' of competence, it will introduce greater consistency of consent standards to both scientific and clinical arenas.

Evidence on consent in the care of mental illness reveals many key challenges for clinicians (Roth et al. 1982). Studies of psychotic, affective, and dementia disorders, for instance, have shown that symptoms of mental illness may adversely impact information-based or more strictly cognitive aspects of consent (Roberts 1998). In the 1995 MacArthur Treatment Competence Study (Appelbaum and Grisso 1995), the decisional abilities of patients with schizophrenia or major depression, patients with ischemic heart disease, and non-patient community volunteers (total $n = 498$) were assessed. When acutely ill, the psychiatric patients in this study had greater difficulties with decision-making on several formal cognitive measures than did the medically ill patients and community participants. The problems were more severe for individuals with psychotic symptoms than for depressed individuals. Clinical treatment led to significant improvement of these deficits, resulting in closer resemblance across the three groups.

Smaller studies echo these findings. An early study by Schachter et al. (1994) involving 38 schizophrenia patients revealed that the presence of conceptual disorganization, delusions, and hallucinations was associated with patients' poor performance on the information elements of consent. Appelbaum et al. (1981) assessed 50 people admitted to a mental health inpatient unit and found that most were unable to consent to hospitalization. Later this team evaluated 100 voluntarily hospitalized psychiatric patients, finding that a significant minority did not understand information presented to them on the purpose of hospitalization and discharge (Appelbaum et al. 1998). Beck (1988) evaluated the capacities of 56 psychiatric patients on antipsychotic medications, revealing that they had limited understanding and appreciation of their treatment. Stanley et al. (1988) compared the

reasoning ability of elderly individuals suffering from major depression ($n = 45$) and dementia ($n = 38$) with elderly control subjects ($n = 20$). The cognitively impaired dementia patients and, to a lesser extent, the depressed patients had significant deficits on a number of decision-making items. Similarly, Marson et al. (1995) used a clinical vignette-based instrument to measure the level of competence exhibited in medical treatment decisions by individuals with dementia and healthy comparison subjects. Not unexpectedly, this work revealed that the severity of cognitive impairment in dementia correlated with lack of ability to formulate logical reasons for treatment choice, i.e., addressing the 'rational reasons' dimension of decisional capacity. Interestingly, global measures of dementia severity such as the Mini-Mental Status Exam were not predictive of clinical decision-making capacity.

Studies of advance directives and surrogate decision-making also have important implications for mental health care. Several survey and interview studies have revealed that advance directives are underutilized (Bradley et al. 1998) and that patient preferences are rarely identified prospectively (Seckler et al. 1991). In terms of surrogate decision-making, several studies have shown that most elderly patients would prefer that a family member such as a spouse or adult child serve as an alternative decision-maker (High 1989). Studies of hypothetical decision-making by proxies suggest that they may not assess patients' wishes accurately, however Sachs et al. (1994) studied 42 decisionally capable dementia patients, 64 proxy decision-makers, and 60 well elderly individuals and found that the decision-makers' predictions of patient preferences were overall discordant with the patients' own stated wishes. Similarly, Warren et al. (1986) studied participation choices made by proxy decision-makers and 168 decisionally capable nursing home patients with respect to hypothetical, minimal risk research projects. The results demonstrated that the surrogate decision-makers used criteria other than the patient's predicted preferences in making these choices.

Clear evidence is also emerging that the barriers to consent may be amenable to intervention, offering significant hope (Roberts et al. 2000). Interventions in which information is presented verbally and in writing, using clear and understandable language, have been shown to improve patients' understanding of consent decisions (Carpenter et al. 2000). The timing of consent information and the provision of information in a relaxed manner or on repeated occasions is important to patients' abilities to integrate and apply this data (Roberts 1998). Treatment of symptoms, ranging from hallucinations to physical pain, has been observed to improve the quality of patients' decision-making. Finally, clinician behaviors that reflect how the consent process is valued, e.g., by giving time to the consent dialogue, have been found to be valued by

patients and surrogate decision-makers (Roberts and Roberts 1999). The observation that educational efforts, symptom treatment, and attitudinal obstacles can be addressed suggests a number of constructive approaches to consent in the context of clinical care for the mentally ill.

See also: Assessment and Application of Therapeutic Effectiveness, Ethical Implications of; Civil Rights; Ethical Dilemmas: Research and Treatment Priorities; Ethics and Psychiatry; Ethics and Values; Ethics for Biomedical Research Involving Humans: International Codes; Free Will and Action; Fundamental Rights and Constitutional Guarantees; Placebos, Use of; Ethical Aspects; Privacy of Individuals in Social Research: Confidentiality; Psychotherapy: Ethical Issues; Research Conduct: Ethical Codes; Research Subjects, Informed and Implied Consent of; Responsibility: Philosophical Aspects

Bibliography

- Appelbaum B C, Appelbaum P S, Grisso T 1998 Competence to consent to voluntary psychiatric hospitalization: A test of a standard proposed by APA. *American Psychiatric Association. Psychiatric Services* **49**: 1193-6
- Appelbaum P S, Grisso T 1988 Assessing patients' capacities to consent to treatment. *New England Journal of Medicine* **319**: 1635-8
- Appelbaum P, Grisso T 1995 The MacArthur Competence Study I, II, III. *Law and Human Behavior* **19**: 105-74
- Appelbaum P, Gutheil T 1991 *Clinical Handbook of Psychiatry and the Law*, 2nd edn. Williams & Wilkins, Baltimore, MD
- Appelbaum P, Mirkin S, Bateman A 1981 Empirical assessment of competency to consent to psychiatric hospitalization. *American Journal of Psychiatry* **138**: 1170-76
- Beck J 1988 Determining competency to consent to neuroleptic drug treatment. *Hospital Community Psychiatry* **39**: 1106-8
- Bradley E H, Peiris V, Wetle T 1998 Discussions about end-of-life care in nursing homes. *Journal of the American Geriatric Society* **46**: 1235-41
- Carpenter W T Jr, Gold J M, Lahti A C et al. 2000 Decisional capacity for informed consent in schizophrenia research [see comments]. *Archives of General Psychiatry* **57**: 533-8
- Grisso T, Appelbaum P S, Hill-Fotouhi C 1997 The MacCAT-T: A clinical tool to assess patients' capacities to make treatment decisions. *Psychiatric Services* **48**: 1415-19
- Grossman S A, Piantadosi S, Covahey C 1994 Are informed consent forms that describe clinical oncology research protocols readable by most patients and their families? *Journal of Clinical Oncology* **12**: 2211-15
- High D M 1989 Standards for surrogate decision making: What the elderly want. *Journal of Long Term Care Administration* **17**: 8-13
- Kaufmann C L, Roth L H, Lidz C W, Meisel A 1981 Informed consent and patient decisionmaking: The reasoning of law and psychiatry. *International Journal of Law and Psychiatry* **4**: 345-61
- Lambert P, Gibson J M, Nathanson P 1990 The values history: An innovation in surrogate medical decision-making. *Law, Medicine, and Health Care*. **18**: 202-12
- Lidz C 1984 *Informed Consent: A Study of Decisionmaking in Psychiatry*. Guilford Press, New York

- Marson D C, Cody H A, Ingram K K, Harrell L E 1995 Neuropsychologic predictors of competency in Alzheimer's disease using a rational reasons legal standard. *Archives of Neurology* **52**: 955–9
- Roberts L W 1998 The ethical basis of psychiatric research: Conceptual issues and empirical findings. *Comprehensive Psychiatry* **39**: 99–110
- Roberts L, Roberts B 1999 Psychiatric research ethics: An overview of evolving guidelines and current ethical dilemmas in the study of mental illness. *Biological Psychiatry* **46**: 1025–38
- Roberts L W, Fraser K, Quinn D, McCarty T, Severino S K 1999 Persons with mental illness. In: Bennahum D (ed.) *Managed Care: Financial, Legal, and Ethical Issues*. Pilgrim Press, Cleveland, OH
- Roberts L W, Warner T D, Brody J L 2000 Perspectives of patients with schizophrenia and psychiatrists regarding ethically important aspects of research participation. *American Journal of Psychiatry* **157**: 67–74
- Roth L H, Lidz C W, Meisel A, Soloff P H, Kaufman K, Spiker D G, Foster F G 1982 Competency to decide about treatment or research: An overview of some empirical data. *International Journal of Law and Psychiatry* **5**: 29–50
- Sachs G A, Stocking C B, Stern R, Cox D M, Hougham G, Sachs R S 1994 Ethical aspects of dementia research: Informed consent and proxy consent. *Clinical Research* **42**: 403–12
- Schachter D, Kleinman I, Prendergast P, Remington G, Schertzer S 1994 The effect of psychopathology on the ability of schizophrenic patients to give informed consent. *Journal of Nervous and Mental Disorders* **182**: 360–2
- Seckler A B, Meier D E, Mulvihill M, Paris B E 1991 Substituted judgment: How accurate are proxy predictions? [see comments] *Annals of Internal Medicine* **115**: 92–8
- Stanley B, Stanley M, Guido J, Garvin L 1988 The functional competency of elderly at risk. *Gerontologist* **28** (Suppl): 53–8
- Sugarman J, McCrory D C, Powell D, Krasny A, Adams B, Ball E, Cassell C 1999 Empirical research on informed consent. An annotated bibliography. *Hastings Center Report* **29**: S1–42
- Warren J W, Sobal J, Tenney J H, Hoopes J M, Damron D, Levenson S, Deforge B R, Muncie H L, Jr. 1986 Informed consent by proxy. An issue in research with elderly patients. *New England Journal of Medicine* **315**: 1124–8

L. W. Roberts

Psychoanalysis: Adolescence (Clinical–Developmental Approach)

Until relatively recently, the study of adolescence was synonymous with the theory and practice of psychoanalysis. The major contributors to adolescent development came from eminent clinicians and theorists, such as Anna Freud (1936), Peter Blos (1962), Erik Erikson (1950). The adolescent crisis, psychosexual maturation, identity formation, and the second separation–individuation phase are all psychoanalytic concepts that have remained alive ever since they were first introduced during the twentieth century. Since

then, the fields of adolescent development and adolescent psychopathology have differentiated greatly, and a myriad of theories and models, many developed in antagonism to psychoanalysis, have evolved. Grand theories and clinical observations have far less currency at the start of the twenty-first century than do systematic cross-sectional and longitudinal studies.

The discovery of adolescence by psychoanalysts became the entry point of the study of development across the life cycle. If an essential new set of tasks develops in adolescence, such as the creation of identity, then there was reason also to explore similarly significant transformations in adulthood. The important initial point was to overcome the singular psychoanalytic focus on early childhood, especially the oedipal years. In this vein, Erikson and others discovered stages of adulthood that reach well into old age. Thus adolescent development has to be viewed in the context of a shifting conception of psychoanalytic developmental psychology which views early childhood as only one important period of development.

But there is certainly no consensus on the issue of development in psychoanalytic circles: Even now, most psychoanalysts believe that the formative periods in life are the early attachment years, and the oedipal period between the ages of 5 and 7. If there was a shift, it was less to explore the later stages of life and more to understand the earliest, pre-oedipal stage of development. It is this fundamental notion of early childhood that has denied the life cycle theorists, with all their creativity, a very significant role in psychoanalytic practice.

However, adolescence, as the intermediary space between childhood and adulthood, has become a significant aspect of most psychoanalysts' thinking about development and psychopathology. This phase of life represents a theoretical compromise, a recognition that new transformations occur beyond the oedipal years, without needing to give up the primacy of early childhood for personality development and symptom formation.

With the majority of innovations about adolescence and the life cycle coming from outside psychoanalysis, it would be easy to dismiss the analytic approach. But many modern adolescent clinicians and researchers recognize that psychoanalysis has changed greatly, and that, when applied carefully, it continues to be an important interpretive method in the social and behavioral sciences. But such modern versions of psychoanalytic theory and research must include the many innovations of adjacent fields (e.g., Smelser 1999), especially the cognitive sciences, psychiatric epidemiology, anthropology, and academic developmental psychology.

The approach taken in this article is to integrate psychoanalytic concepts in an integrative framework called clinical-developmental psychology. This approach represents a synthesis of psychoanalytic concepts, especially Eriksonian identity theory, object

relations psychoanalysis as well as cognitive-developmental theory. Thus, the integrated theoretical, clinical, and research perspective does not represent psychoanalysis proper, nor does it represent traditional cognitive psychology in the tradition of Jean Piaget. It is, instead, a working out of the contradictions of two significant paradigms of human development that have formed the building blocks of a new synthesis. This new approach forms a dynamic and transformational theory of adolescent development and the lifespan, within a research paradigm and a differentiated treatment approach.

In general, psychoanalysis has become far more open to these kinds of syntheses than ever before. Bowlby's contribution (1969), for example, represents an integration of psychoanalytic and ethological thinking. The result is neither psychoanalysis nor ethology, but a theory of attachment that is one of the most significant advances of psychoanalytic thought. Modern psychoanalysis has become far more open, because of the rapid developments in academic developmental psychology and other fields and an increasing openness to integrate these approaches.

1. Psychoanalysis

The popular notion is that Freud's theory was about unconscious motivation, sexuality, aggressive drives, and the three structures, id, ego, and superego. Freud is, furthermore, credited with having created a comprehensive theory of human emotions. These general conceptions about Freud's contributions are not incorrect, but they are incomplete. Freud was equally dedicated to cognition and logic. He was convinced that psychiatric symptoms and psychological conflicts could be deciphered following a logical, hermeneutic strategy of interpretation. He also described in the 'Introductory Lectures' (Freud 1933), as well as in 'Mourning and Melancholia,' a theory of human relationship, one that was to become the dominant modern psychoanalytic perspective called 'object relations theory.'

These two Freudian approaches, cognitive and relational, have expanded over the decades and were joined by a major third psychoanalytic perspective, the life cycle approach. Interestingly, the cognitive branch of psychoanalysis came to overshadowed by the far more elegant and researchable cognitive contributions of academic developmental psychology, especially the work of cognitive and social cognitive psychologists. The life cycle approach, building primarily on Erikson's work (e.g., Erikson 1950) was acknowledged in psychoanalytic circles but never really shaped central psychoanalytic theory and practice. However, as mentioned above, it did help to extend the traditional focus on the early oedipal and pre-oedipal years at least into adolescence, as a second period of separation.

The life cycle approach in psychoanalysis changed into a research paradigm primarily through the work of George Vaillant (1977). Vaillant introduced a hierarchy of defenses in the tradition of Anna Freud, extended their development beyond adolescence into adulthood, and defined 'families of defenses,' defined not only in negative terms but also as adaptive styles. Vaillant showed, for example, how projection, with its pathological preoccupation with others, is related to a higher level adaptive defense which he calls 'empathy.' Vaillant applied this hierarchy of defenses to a fascinating set of longitudinal studies, especially the so-called Grant Study of Harvard men selected for their health. These men have now reached their eighties, and the research on them has demonstrated that a set of psychoanalytic constructs that were widely viewed as unresearchable, such as defense mechanisms, can be studied rigorously, have a developmental core, and are one way to trace the evolution of development in adulthood.

The psychoanalytic approach to adolescent development and the lifespan requires us to address C. G. Jung as well. His insights have been impressive and continue to garner a larger readership interested in development, meaning, and spirituality. Jung was really the first psychoanalytic thinker to dedicate significant effort to understand the transformations a person goes through during life. He was also fascinated by the collective symbols and unconscious processes of development and transformation in human development. While Jung had some interest in adolescents, his central developmental contribution was about mid-life, what he viewed as a typical shift from the external world to an inner and spiritual life. It is not difficult to create a bridge between these ideas to a theory of adolescent development. Many of Jung's ideas apply to adolescent themes of creating a sense of collective belonging to a generation, to a preoccupation with 'making it in the world,' and a preoccupation with being unique.

But it is the so called object relations strand of psychoanalysis of Fairbairn (1952) and Winnicott (1965) and others that has become the organizer of modern psychoanalysis, in the form of Kohut's self-psychology, as relational psychoanalysis, and as a research paradigm in the form of attachment theory. These relational approaches have reshaped the psychological and psychoanalytic understanding of adolescence. In the earlier version, the adolescent crisis was viewed as stemming from the need to separate, to overcome childhood identifications, and to create autonomy. For that reason, Blos called this phase the second separation-individuation phase (the first one being the one that Margaret Mahler described for the toddler). In the modern relational version of the adolescent, separation is far less important than are new relational capacities, and new forms of interconnections between the adolescent, the family, and society. The psychoanalytic adolescent of the twenty-

first century is far less of a rebel who pushes the parents away than a person yearning for support and closeness, and wishing to find a new balance between family relationships and new intimate relationships. It is impossible to disaggregate what represents a new reality for families, what is a new adolescent ethos, and what is a reframing of psychoanalytic theory. All forces are at play, and while there is no single psychoanalytic scholar who is developing a new theory of adolescence, many clinicians and researchers are contributing to this relational shift.

Erikson's contribution, while having had limited resonance in psychoanalysis, continues to exert considerable influence in developmental psychology, despite the fact that he did not use traditional empirical methods of sampling and proof. He showed that understanding adolescence means to go deeper, to understand the cognitive and emotional dimensions of meaning-making. At the time of writing, no one has yet been able to combine these kinds of keen observations and theoretical synthesis when studying adolescent life. Erikson's conceptualization of adolescent identity formation, and the processes it includes (crisis, diffusion, foreclosure, moratorium), remain the standard by which subsequent contributions are measured, even if those contributions might be far more precise and valid empirically.

Since Erikson introduced his work, much has changed in the world of adolescents. Given the social dimension of the construct, one would expect cultural and historic changes (what Erikson (1968) called 'the historic moment') to influence the very nature of the adolescent experience. But also, new theories and research evidence have reshaped our understanding of how meaning-making and identity formation develop in the second decade of life. It is these cultural changes of adolescence, and the new developments inside and outside psychoanalysis, that gave impetus to a clinical-developmental synthesis of adolescent theory, research, and practice (see *Developmental Sciences, History of*).

2. Clinical-developmental Assumptions of Adolescence and the Life Cycle

2.1 Development of Representations of Self and Other

Clinical-developmental psychology of adolescence has a foundation in a relational and interactionist perspective. Such a psychoanalytic and cognitive perspective views interpersonal relationships and their internalizations as essential dimension of human development, symptomatology, and more recently, psychological treatment. Some of the pioneers of cognitive-developmental psychology (e.g., Piaget 1926) took the interactionist idea and applied it to an

abstract, epistemic, logical subject of self in relation to an equally cognitive and abstract general 'other.'

Evidence has accumulated since then, that significant relationships with less abstract 'others,' such as parents, teachers, siblings, and friends are constructed actively and interpreted by the self, and that one result is an internal set of self and object representations. The concept that relationships and experiences with important others become represented internally is common, despite many conceptual differences, to a number of theories of psychology and psychoanalysis. These theories include classical psychoanalysis and branches of the modern psychoanalytic schools, such as self-psychology, object relations theory, and attachment theory. They also include the developmental theories built on the ideas of Vygotsky, Piaget, Mead, and Baldwin.

Although psychoanalysts and attachment theorists all share the view that the weaving of various significant relationships creates the inner fabric of the self, the clinical-developmental psychologist examines the *typical transformations* of relationships over the course of the lifespan (Noam and Fischer 1996). There is now evidence from a number of cross-sectional and longitudinal studies for the clinical-developmental view that internal representations of the self and others change over time in childhood, adolescence and even adulthood (e.g., Broughton 1978, Noam et al. 1995).

This view owes much to Piaget and Kohlberg, whose early work on the moral judgment of the child (Kohlberg 1969, Piaget 1932) revealed that children's constructions of social rules and relationships change systematically as they grow. Piaget and Kohlberg's ideas have been refined; new cognitive-developmental perspectives on the child's relationship to the social world in general, and to important others, now exist throughout the developmental literature.

While clinical-developmental psychology is built on an understanding of normative cognitive development, significant theoretical and clinical extensions of the traditional cognitive models have occurred. The psychoanalytic frame plays a very important role in these theoretical extensions. For example, there is greater awareness that representations of significant relationships can be quite resistant to change, despite their potential for transformation. Even an adult tends to represent the parents as relatively powerful figures, at times continuing to view them from the perspective of a child. But although the inner representational life is conservative in nature, most people are capable of changing their perspectives on significant relationships, and are able to do what attachment theorist have called 'updating the working model.'

Clinical-developmental psychology recognizes the fluctuations between progress and regress in development, the resistance to changing one's early representations, and the evolution and progression that often take place simultaneously. These fluctuations are especially strong during the adolescent years, when

close relationships with the family have to be brought into a new balance with peer influences and identity development.

2.2 Normative Development

In order to understand the developmental dimensions inherent in adolescent psychopathology, it is essential to understand what occurs typically in ‘normal development.’ How else can we know which challenges arise naturally during this phase, and which represent the beginnings of serious deviation and maladaptation? For example, anxieties in early childhood are common (e.g., fears of separation, monsters, darkness, etc.) and are not necessarily a cause for clinical concern. But severe separation anxiety in middle childhood and adolescence is not typical, and such anxiety often requires clinical attention. Another example is presented by Offer and colleague who have argued against Erikson’s notion of serious adolescent crisis as a part of normal development. They present evidence that ‘normal’ adolescence is far less conflictual and risky than thought previously.

While theory development and clinical observation, the traditional psychoanalytic tools, can help us to understand what is developmentally normal and what is not, it is only through empirical studies with community samples that these distinctions can properly be made. Such studies can also help us to determine what the prevalence of anxiety, aggression, depression, suicidality, etc., is in the child and adolescent population. How do they vary by socioeconomic class, geographical location, ethnic group, and gender? And how do these problems change over time? Answers to some of these questions have already come from large-scale longitudinal epidemiological studies (for a review, see Costello and Angold 1995).

These findings give us the tools to apply, in a meaningful way, the concepts of delay and aberrant pathways that have proved to be so useful to studies of mental retardation, delinquency, and psychosis. We can, for example, trace a delinquent’s moral development, allowing us to see that it is delayed in comparison to that of his nondelinquent peers (e.g., Gibbs et al. 1984). Findings such as these give us a developmental focus; they allow us to see that it is not only behavior or brain functioning that distinguishes delinquents from nondelinquents, but also level of adolescent socio-moral maturity. More importantly, once we understand the moral delay, we can, among other interventions, begin to focus our interventions on supporting the process of maturation of the cognitive system that guides moral development.

Clinical-developmental psychology, however, is not just an application of the study of normal development to the realm of psychological dysfunction. It is also a reformulation of a number of traditionally held assumptions of normative theory. Key among these

reformulations is the view that adolescent symptoms may be expressed differently depending on the individual’s level of development (e.g., Noam 1988). Typically, in early phases of development, impulsive, action-oriented symptoms are more frequent, while later in development, internalizing symptoms such as depression arise with greater frequency. Therefore, the idea that higher stages of development are necessarily more adaptive, as is suggested by many developmental theories, requires serious reconsideration. The clinical-developmental psychologist does not view development as progressing simply from egocentric, oral, unsocialized, or impulsive stages to those defined by the ability to take another person’s perspective and show tolerance, empathy, and ‘phallic’ strength. Developmental theories have assumed this view for too long.

But we now know that a similar developmental course that leads to increased self-knowledge and strengths in some people leads to alienation, self-deception, and self-destructive behavior in others (e.g., Noam et al. 1995). Reaching Piagetian formal operations, for example, can throw the adolescent into a dangerous identity crisis, providing the adolescent with so many potential ‘selves’ that no choice seems possible (e.g., Perry 1970). Chandler (1973) referred to this cognitive crisis of adolescence as the ‘epistemological loneliness of formal operations’ (p. 381).

2.3 Age Chronology

The clinical-developmental approach is far less focused on chronological age than are most research approaches to development and psychopathology. Unfortunately, it has become quite common to label a model or study as being an example of research in ‘developmental psychopathology’ if age is introduced as a significant variable. But chronological age is a very crude indicator of the underlying processes of biological maturation, cognitive capacity, peer and family relationships, and one’s understanding of one’s self. Age does, of course, play a role in development in most of these domains, but there is frequently a significant variation in maturity in each of these areas within a single age group.

There are, of course, also age trends in some forms of psychopathology. For example, suicide rates rise dramatically in adolescence, as do other disorders such as depression, and conduct problems. Schizophrenia is typically an illness that begins in late adolescence or young adulthood. Age, therefore, can serve as a simple way to organize an understanding of many underlying pathogenic processes. However, the simplicity of this approach is deceptive (see Rutter 1989).

In psychopathological development, as in normal development, we often find that chronological age is

by no means a guarantor that basic cognitive and social-cognitive processes have occurred. For example, despite textbook claims to the contrary, many adolescents and adults never achieve formal-operational thought (i.e., tables in a typical psychology textbook list adolescence next to formal operations, as if all adolescents function at that cognitive stage). While many adolescents have achieved a formal operational level of thinking, others continue to function at the level of concrete operations, and indeed may remain at this level for the rest of their lives.

Consequently, if we use chronological age as the principle developmental marker, then the important variations that occur in every group of normally-developing individuals will receive insufficient attention. For the clinical-developmental psychologist, this suggests that chronological age needs to be supplemented with a domain-specific understanding of a person's capacities in different areas of development.

But, although capacities are often measured in terms of chronological age, the more sophisticated measurements of development that we have as we enter the twenty-first century call on us to subsume the idea of chronological age into our understanding of progressive adaptation and skill acquisition. In so doing, our research finds significant links between development and psychopathology in adolescents, where chronological age plays only a minor role. Level of social cognitive functioning and the shape of the internal self and object representations, rather than age, are often associated with specific disorders.

2.4 Lifelong Development and Recovery

Significant developmental changes can, and regularly do, occur long after childhood and adolescence. The lifelong process of development has been described eloquently by Erikson (1950), Vaillant (1977), Baltes (1997), and others. But despite a great deal of progress in understanding the achievement of salient life tasks, meaning making, and cognition, little progress has been made in applying this understanding to psychological treatment and recovery. Indeed, little is known even now about typical problems and psychological dysfunctions that arise with development throughout life.

Critical to clinical-development is the understanding that at each point in life, fundamental shifts in cognitive, emotional, social, and interpersonal development are possible, and that these shifts follow a certain logic and organization. For example, adults who identify strongly with work, authority, and control, can become disappointed when others do not live up to their expectations. These inclinations to disappointment, perfectionism, and excessive control might be long-standing, stemming from compulsive character traits that were formed in childhood. But

they might also stem from the way people, as adults, define their basic categories of self, their world, and their relationships. Perhaps the categories most central to them are responsibility, predictability, and accountability. These categories and their salience can change over time into a greater acceptance of human frailty and of the limitations of control and predictability. As complex ways of understanding and experiencing change, past rigidities are also revisited. Each new chapter in a person's life typically rewrites earlier chapters. Understanding the ways in which people can transform their representational world allows the developmental therapist and developmentally-oriented psychoanalyst to help the patient revisit old relationships, overcome old vulnerabilities, work through past traumas, and help to prevent new ones from occurring. But the clinical-developmental perspective puts a great deal of emphasis on the self-righting and healing potential that is triggered under the right therapeutic conditions.

3. Extending the Psychoanalytic Paradigm: Differential Therapy with Adolescents

The next brief section will deal with using clinical-developmental principles to differentiate adolescent functioning and briefly introduce differential treatment implications. For a more detailed discussion see (Noam 1988), where four different levels of adolescent functioning and intervention strategies are described. Because of space considerations only two adolescent positions will be differentiated here. This section will show that there is no single treatment approach for adolescents, even youth who share the exact same age, but that the approach will differ fundamentally, not only by diagnosis, but also by developmental organization of self and relationship representation.

3.1 Reciprocal–Instrumental Self-complexity

Establishing treatments for adolescents who function at the reciprocal–instrumental self-complexity level is difficult. First, these adolescents are viewed by society as delayed developers, and often have many problems adjusting to the demands placed on them. They tend to be oppositional, even delinquent, and have a concrete and instrumental approach to relationships. Adding to these problems is the fact that play therapy is not appropriate at this age, because it seems infantilizing to them. Many intervention specialists take account of the verbal capacities of these adolescents in devising a treatment strategy. Because these adolescents are often quite verbal, therapists expect the youngsters to talk about their problems. But talking therapies generally do not work, nor do treatments that require a great

deal of self-reflection and insight. The therapist who tries such an approach is usually greeted with contempt and hostility, and can expect communication to break down.

And yet, it is these adolescent girls and boys who are at risk of developing problems at school, dropping out, using drugs and alcohol, behaving aggressively, and running away from home. They need a great deal of support and mentoring, in the forms of individual therapy and group interventions, to learn how to negotiate life, and how to develop goals that do not 'get them into trouble.' If adults are conducting the intervention, it is important they orient the work toward joint goals, giving the adolescent a great deal of room to decide what should happen. It is very important that therapists do not get caught up with questions such as 'what kind of therapy am I pursuing,' but instead that they remain extremely flexible, leaving the office when it is appropriate, to play, eat, or walk. It is not by chance that coaches and child-care workers are often successful with these adolescents; they are often perceived as nonthreatening and available when the adolescent needs them.

It is also essential to create contexts of peer learning and therapy for these adolescents. These environments can help the adolescents to learn how to overcome their difficulties in empathizing with others. A combination of developing group intimacy and confronting problem behavior by peers is powerful in helping the adolescent overcome an over-reliance on self-protection and move toward a more mature developmental understanding of the self and the world.

3.2 Mutual-Inclusive Self-complexity

At this point, talking therapy has a real potential for making the adolescent feel understood and supported. Sharing one's inner life with adults and peers, and framing it in psychological terms, becomes a powerful motivation toward intimacy and recovery. Group conformity is at its height, and so is the constant fear of losing the 'relationship base,' of abandonment by parents, teachers, and friends. It is this fear of losing those who define the self, and of not fitting in, that is a core feature of this adolescent's adolescence.

The ability of these youngsters to frame experiences in psychological terms, to describe feeling states and interactions with important others, encourages most therapists to explore patterns and motivations, autonomous self-observation and critical judgment about self and relationships. These explorations usually lead to surprisingly superficial descriptions, which frustrate both therapist and client. Surprised by this 'lack of self,' the therapist may interpret this state as an early, and primitive, manifestation of a separation-individuation problem. The clients, consequently, feel

incompetent, sensing that they are not living up to expectations. Feeling inadequate leads to feeling hopeless, and clients are thereby silenced. The silence, in turn, makes the therapist and client often feel quite uncomfortable, increasing the potential for low self-esteem and depression.

There is no therapeutic model that clearly fits this stage of adolescence, which is surprising in light of the fact that so many adolescents who enter therapy function at this level. One reason might be that we did not know about this developmental gestalt before the emergence of social cognition. Existing therapy methods that use supportive strategies run the risk of under-challenging the patient. The relationship can easily become stale as the adolescent waits for some challenge or some guidance in framing the problems in new ways.

Insight-oriented therapy, on the other hand, requires an observation of systematic patterns of the self which is not available to adolescents at this developmental stage. Following a set of goals that is experienced as self-chosen, questioning existing conventions, and creating new ones, requires a differentiation between the part of the self steeped in prescribed pathways and conventions, and the part that is reflecting, doubting, and questioning. Therapists can be sure that they have an ally in the patient, who can contain cognitively, at least under supportive conditions, both sides of the self. For that reason, more traditional psychodynamic therapy can be applied, which becomes even more viable as the person grows into full adulthood.

4. Conclusion

The psychoanalytic perspective continues to provide important impulses for the study and treatment of adolescents and their families. Erikson provided significant theoretical tools, and Bowlby introduced a research paradigm that has been amazingly productive. But further integrations of perspectives, especially pertaining to the cognitive dimensions of adolescent thought, have proven necessary. The clinical-developmental approach not only expands the theoretical tools for the study of adolescent development and adolescent risk: It also introduces a developmental typology of intervention and prevention strategies. Intervention advances with adolescents have been slow, despite the introduction of many strategies. It is possible that progress has been slow in part because the emphasis has been primarily on differentiating disorders, rather than on understanding these disorders in the context of adolescent development.

See also: Adolescent Development, Theories of; Adolescent Health and Health Behaviors; Adolescent

Vulnerability and Psychological Interventions; Bowlby, John (1907–90); Erikson, Erik Homburger (1902–94); Freud, Sigmund (1856–1939); Psychoanalysis, History of; Psychoanalysis in Clinical Psychology

Bibliography

- Baldwin J M 1902 *Social and Ethical Interpretations in Mental Development*. Macmillan, New York
- Baltes P B 1997 On the incomplete architecture of human ontogeny: Selection, optimization, and compensation as foundation of developmental theory. *American Psychologist* **52**: 366–80
- Blos P 1962 *On Adolescence: A Psychoanalytic Interpretation*. Free Press, New York
- Bowlby J 1969 *Attachment and Loss, Vol. I: Attachment*. Basic Books, New York
- Broughton J 1978 *The Development of Concepts of Self, Mind, Reality, and Knowledge. New Direction in Child Development: Social Cognition*. Jossey-Bass, San Francisco, CA
- Chandler M J 1973 Egocentrism and anti-social behavior: The assessment and training of social perspective taking skills. *Developmental Psychology* **9**: 326–32
- Costello E J, Angold A 1995 Developmental epidemiology. In: Cicchetti D, Cohen D J (eds.) *Developmental Psychopathology*. Wiley, New York, Vol. 1, pp. 23–56
- Erikson E 1950 Growth and crisis of the healthy personality. *Psychological Issues* **1**: 50–100
- Erikson E 1968 *Identity, Youth, and Crisis*. Norton, New York
- Fairbairn R 1952 *Psychoanalytic Studies of the Personality*. Routledge & Kegan Paul, London
- Freud A 1936 *The Ego and the Mechanisms of Defense*. International Universities Press, New York
- Freud S 1933 The anatomy of the mental personality, Lecture XXXI. In: *New Introductory Lectures on Psycho-analysis*, trans. W J H Sprott. W W Norton, New York, Inc. 3
- Gibbs J C, Arnold K D, Ahlborn H H, Chessman F L 1984 Facilitation of socio-moral reasoning in delinquents. *Journal of Consulting and Clinical Psychology* **52**: 37–45
- Kohlberg L 1969 Stage and sequence: The cognitive-developmental approach to socialization. In: Glosin D (ed.) *Handbook of Socialization, Theory and Research*. Rand McNally, New York, pp. 347–480
- Mead G H 1934 *Mind, Self, and Society*. University of Chicago Press, Chicago
- Noam G G 1988 A constructivist approach to developmental psychology. In: Nannis E, Cowan P (eds.) *Developmental Psychopathology and Its Treatment*. Jossey-Bass, San Francisco, CA, pp. 91–122
- Noam G G, Chandler M, LaLonde C 1995 Clinical-developmental psychology: Constructivism and social cognition in the study of psychological dysfunctions. In: Cicchetti D, Cohen D J (eds.) *Developmental Psychopathology, Vol. 1: Theory and Methods*. Wiley, New York, pp. 424–64
- Noam G G, Fischer K W 1996 (eds.) *Development and Vulnerabilities in Close Relationships*. Erlbaum, Hillsdale, NJ
- Offer D, Schonert-Reichl K 1992 Debunking the myths of adolescence: Findings from recent reviews. *Journal of the American Academy of Child and Adolescent Psychiatry* **31**: 1003–14
- Perry W S 1970 *Forms of Intellectual and Ethical Development in the College Years*. Holt Rinehart, New York

- Piaget J 1932 *The Moral Judgement of the Child*. Harcourt, New York
- Rutter M 1989 Age as an ambiguous variable in developmental research: Some epidemiological considerations from developmental psychopathology. *International Journal of Behavioral Development* **12**: 1–34
- Smelser N 1999 *The Social Edges of Psychoanalysis*. University of California Press, Berkeley, CA
- Vaillant G E 1977 *Adaptation to Life*. Little, Brown, Boston
- Winnicott D W 1965 *The Maturational Processes and the Facilitating Environment*. International Universities Press, New York

G. G. Noam

Psychoanalysis: Current Status

1. Freud's Theory of the Mental Apparatus

Freud's (1953a) starting point was his study of hysterical patients and the discovery that, when he found a way to help these patients piece together a coherent account of the antecedents of their conversion symptoms, dissociative phenomena, and pathological affective dispositions, all these psychopathological phenomena could be traced to traumatic experiences in their past that had become unconscious. In the course of a few years Freud abandoned his early efforts to recover repressed material by means of hypnosis, and replaced hypnosis with the technique of 'free association,' an essential aspect of psychoanalytic technique until the present time.

Practicing this method led Freud to several lines of discovery: to begin, he conceptualized unconscious mechanisms of defense that opposed the recovery of memories by free association. He described these mechanisms, namely repression, negation, isolation, projection, introjection, transformation into the opposite, rationalization, intellectualization, and most important, reaction formation. The latter involves overt, chronic patterns of thought and behavior that serve to disguise and disavow opposite tendencies linked to unconscious traumatic events and the intrapsychic conflicts derived from them. The discovery of reaction formations led Freud to the psychoanalytic study of character pathology and normal character formation, and still constitutes an important aspect of the contemporary psychoanalytic understanding and treatment of personality disorders (for practical purposes, character pathology and personality disorders are synonymous concepts).

A related line of development in Freud's theories was the discovery of the differential characteristics of conscious and unconscious thinking. Freud differentiated conscious thinking, the 'secondary process,' invested by 'attention cathexis' and dominated by

sensory perception and ordinary logic in relating to the psycho-social environment, from the 'primary process' of the 'dynamic unconscious.' That part of the unconscious mind he referred to as 'dynamic' exerted constant pressure or influence on conscious processes, against the active barrier constituted by the various defensive operations, particularly repression. The dynamic unconscious, Freud proposed, presented a general mobility of affective investments, and was ruled by the 'pleasure principle' in contrast to the 'reality principle' of consciousness. The 'primary process' thinking of the dynamic unconscious was characterized by the absence of the principle of contradiction and of ordinary logical thinking, the absence of negation and of the ordinary sense of time and space, the treatment of a part as if it were equivalent to the whole, and a general tendency toward condensation of thoughts and the displacement of affective investments from one to another mental content.

Finally, Freud proposed a 'preconscious,' an intermediate zone between the dynamic unconscious and consciousness. It represented the storehouse for retrievable memories and knowledge and for affective investments in general, and it was the seat of day-dreaming, in which the reality principle of consciousness was loosened, and derivatives of the dynamic unconscious might emerge. Free association, in fact, primarily tapped the preconscious as well as the layer of unconscious defensive operations opposing the emergence of material from the dynamic unconscious.

This model of the mind as a 'place' with unconscious, preconscious, and conscious 'regions' constituted Freud's 'topographic theory.' He eventually replaced it with the 'structural theory' namely, the concept of three interacting psychic structures, the ego, the superego and the id (Freud 1961).

1.1 The Id

The id is the mental structure that contains the mental representatives of the 'drives,' that is, the ultimate intrapsychic motivations that Freud (1955) described in his final, 'dual drive theory' of libido and aggression, or metaphorically, the sexual or life drive and the destruction or death drive. In exploring unconscious mental processes, what at first appeared to be specific traumatic life experiences turned out to reflect surprisingly consistent, repetitive, intrapsychic experiences of a sexual and aggressive nature.

Freud (1953a) was particularly impressed by the regularity with which his patients reported the emergence of childhood memories reflecting seductive and traumatic sexual experiences on the one hand, and intense sexual desires and related guilt feelings, on the other. He discovered a continuity between the earliest wishes for dependency and being taken care of (the psychology, as he saw it, of the baby at the mother's

breast) during what he described as the 'oral phase' of development; the pleasure in exercising control and struggles around autonomy in the subsequent 'anal phase' of development, (the psychology of toilet training); and, particularly, the sexual desire toward the parent of the opposite gender and the ambivalent rivalry for that parent's exclusive love with the parent of the same gender. He described this latter state as characteristic of the 'infantile genital stage' (from the third or fourth to the sixth year of life) and called its characteristic constellation of wishes and conflicts the positive Oedipus complex. He differentiated it from the negative Oedipus complex, that is, the love for the parent of the same gender, and the corresponding ambivalent rivalry with the parent of the other gender. Freud proposed that oedipal wishes came to dominate the infantile hierarchy of oral and anal wishes, becoming the fundamental unconscious realm of desire.

Powerful fears motivated the repression of awareness of infantile desire: the fear of loss of the object, and later of the loss of the object's love was the basic fear of the oral phase, directed against libidinal wishes to possess the breast; the fear of destructive control and annihilation of the self or the object was the dominant fear of the anal phase directed against libidinal wishes of anal expulsion and retentiveness, and the fear of castration, 'castration anxiety,' the dominant fear of the oedipal phase of development, directed against libidinal desire of the oedipal object. Unconscious guilt was a dominant later fear, originating in the superego and generally directed against drive gratification (see Sect. 1.3). Unconscious guilt over sexual impulses unconsciously equated with oedipal desires constitutes a major source of many types of pathology, such as sexual inhibition and related character pathology.

Freud described the oral phase as essentially coinciding with the infantile stage of breast feeding, the anal phase as coinciding with struggles around sphincter control, and the oedipal stage as developing gradually during the second and through the fourth years, and culminating in the fourth and the fifth years of life. This latter phase would then be followed by more general repressive processes under the dominance of the installation of the superego, leading to a 'latency phase' roughly corresponding to the school years, and finally, to a transitory reactivation of all unconscious childhood conflicts under the dominance of oedipal issues during puberty and early adolescence.

The drives represent for human behavior what the instincts constitute for the animal kingdom, that is, the ultimate biological motivational system. The drives are constant, highly individualized, developmentally shaped motivational systems. Under the dominance of the drives and guided by the primary process, the id exerts an on-going pressure toward gratification, operating in accordance with the pleasure principle. Freud initially equated the drives with primitive

affects. After discarding various other models of unconscious motivation, he ended up with the dual drive theory of libido and aggression.

He described the libido or the sexual drive as having an 'origin' in the erotogenic nature of the leading oral, anal, and genital bodily zones; an 'impulse' expressing the quantitative intensity of the drive by the intensity of the corresponding affects; an 'aim' reflected in the particular act of concrete gratification of the drive; and an 'object' consisting of displacements from the dominant parental objects of desire.

The introduction of the idea of an aggressive or 'death' drive, arrived at later in Freud's writing, stemmed from his observations of the profound self destructive urges particularly manifest in the psychopathology of major depression and suicide, and of the 'repetition compulsion' of impulse driven behavior that frequently seemed to run counter to the pleasure principle that supposedly governed unconscious drives. He never spelled out the details of the aggressive drive as to its origins. Freud described drives as intermediate between the body and the mind; the only thing we knew about them, Freud (1957a) suggested, were 'representations and affects.'

1.2 The Ego

While the id is the seat of the unconscious drives, and functions according to the 'primary process' of the dynamic unconscious, the ego, Freud (1961) proposed, is the seat of consciousness as well as of unconscious defense mechanisms that, in the psychoanalytic treatment, appear as 'resistance' to free association. The ego functions according to the logical and reality based principles of 'secondary process,' negotiating the relations between internal and external reality. Guided by the reality principle, it exerts control over perception and motility; it draws on preconscious material, controls 'attention cathexes,' and permits motor delay as well as selection of imagery and perception. The ego is also the seat of basic affects, particularly anxiety, as an alarm signal against the danger of emergence of unconscious, repressed impulses. This alarm signal may turn into a disorganized state of panic when the ego is flooded with external perceptions that activate unconscious desire and conflicts, or with overwhelming, traumatic experiences in reality that resonate with such repressed unconscious conflicts, and overwhelm the particularly sensitized ego in the process.

Freud originally equated the 'I,' that is, the categorical self of the philosophers, with consciousness. Later, once he established the theory of the ego as an organization of both conscious and unconscious functions, he at times treated the ego as if it were the subjective self, and at other times, as an impersonal

organization of functions. Out of this ambiguity evolved the contemporary concept of the self within modern ego psychology as well as in British and American object relations and cultural psychoanalytic contributions (Kernberg 1984). An alternative theory of the self was proposed by Kohut (1971), the originator of the self-psychology approach within contemporary psychoanalysis.

Nowadays, an integrated concept of the self as the seat of subjectivity is considered an essential structure of the ego, and the concept of 'ego identity' refers to the integration of the concept of the self. Because developmental processes in early infancy and childhood are now better understood, an integrated self concept usually goes hand in hand with the capacity for an integrated concept of significant others. An unconscious tendency toward primitive dissociation or 'splitting' of the self-concept and of the concepts of significant objects runs counter to such integration.

Character, from a psychoanalytic perspective, may be defined as constituting the behavioral aspects of ego identity (the self-concept) and the internal relations with significant others, (the internalized world of 'object relations'). The sense of personal identity and of an internal world of object relations in turn reflects the subjective side of character. It was particularly the ego psychological approach—one of the dominant contemporary psychoanalytic schools—that developed the analysis of defensive operations of the ego, and of pathological character formation as a stable defensive organization that needed to be explored and resolved in the psychoanalytic treatment. In the process, ego psychology importantly contributed to the psychoanalytic treatment of personality disorders.

Freud (1957c) conceptualized narcissism as the libidinal investment of the ego or self, in contrast to the libidinal investment of significant others ('objects'). In proposing the possibility of a withdrawal of libidinal investment from others with an excessive investment in the self as the basic feature of narcissistic pathology, he pointed to a broad spectrum of psychopathology, and thus first stimulated the contribution of Abraham (1919), and later those of Grunberger (1979), Jacobson (1964), Kernberg (1975), Klein (1957), Kohut (1971), and Rosenfeld (1964). Thus crystallized the description of the narcissistic personality as a disorder derived from a pathological integration of a grandiose self as a defense against unbearable aggressive conflicts, particularly around primitive envy.

1.3 The Superego

In his analysis of unconscious intrapsychic conflicts between drive and defense, Freud regularly encountered unconscious feelings of guilt in his patients, reflecting an extremely strict, unconscious infantile

morality which he called the superego. This unconscious morality could lead to severe self-blame and self-attacks, and particularly to abnormal depressive reactions, which he came to regard as expressing the superego's attacks on the ego. It was particularly in studying normal and pathological mourning, that Freud (1957c) arrived at the idea of excessive mourning and depression as reflecting the unconscious internalization of the representation of an ambivalently loved and hated lost object. In unconsciously identifying the self with that object introjected into the ego, the individual now attacked his own self in replacement of the previous unconscious hatred of the object; and the internalization of aspects of that object into the superego reinforced the strictness of the individual's preexisting unconscious infantile morality.

Freud traced the origins of the superego to the overcoming of the Oedipus complex via unconscious identification with the parent of the same gender: in internalizing the oedipal parent's prohibition against the rivalry with him or her and the unconscious death wishes regularly connected with such a rivalry, and against the incestuous desire for the parent of the other gender, this internalization crystallized an unconscious infantile morality. The superego, thus based upon prohibitions against incest and parricide, and a demand for submission to, and identification with the oedipal rival, became the guarantor of the capacity for identification with moral and ethical values in general.

Freud also described the internalization of the idealized representations of both parents into the superego in the form of the 'ego ideal.' He suggested that the earliest sources of self esteem, derived from mother's love, gradually fixated by the baby's and small child's internalizations of the representations of the loving mother into the ego ideal, led to the parental demands also becoming internalized. In other words, normally self-esteem is maintained both by living up to the expectations of the internalized idealized parental objects, and by submitting to their internalized prohibitions. This consideration of self-esteem regulation leads to the clinical concept of narcissism as normal or pathological self-esteem regulation, in contrast to the theoretical concept of narcissism as the libidinal investment of the self.

2. Psychoanalytic Treatment

2.1 The Psychoanalytic Theory of Psychopathology

This theory proposes that the clinical manifestations of the symptomatic neuroses, character pathology, perversions, sexual inhibitions, and selected types of psychosomatic and psychotic illness reflect unconscious intrapsychic conflicts between drive derivatives following the pleasure principle, defensive operations

reflecting the reality principle, and the unconscious motivations of the superego. Unconscious conflicts between impulse and defense are expressed in the form of structured conflicts between the agencies of the tripartite structure. There are ego defenses against impulses of the id; the superego motivates inhibitions and restrictions in the ego; at times the repetitive, dissociated expression of id impulses ('repetition compulsion') constitutes an effective id defense against superego pressures. The resolution of unconscious conflicts implies the analysis of all these intersystemic conflicts.

All these conflicts are clinically expressed by three types of phenomena: (a) inhibitions of normal ego functions regarding sexuality, intimacy, social relations, and affect activation; (b) compromise formations between repressed impulses and the defenses directed against them; and (c) dissociative expression of impulse and defense. This latter category implies a dominance of the splitting mechanisms referred to before. These have acquired central importance in the understanding of severe character pathology as reflected in contemporary psychoanalytic thinking.

2.2 The Structural Formulation of the Psychoanalytic Method

Psychoanalytic treatment consists, in essence, in facilitating the reactivation of the pathogenic unconscious conflicts in the treatment situation by means of a systematic analysis of the defensive operations directed against them. This leads to the gradual emergence of repressed impulses, with the possibility of elaborating them in relation to the analyst, and their eventual adaptive integration into the adult ego. Freud (1953c) had described the concept of 'sublimation' as an adaptive transformation of unconscious drives: drive derivatives, converted into a consciously tolerable form, are permitted gratification in a symbolic way while their origin remains unconscious. The result of this process is an adaptive, non-defensive compromise formation between impulse and defense. In analysis, the gradual integration into the patient's conscious ego of unconscious wishes and desires from the past and the understanding of the phantasied threats and dangers connected with them facilitates their gradual elaboration and sublimatory expression both in the consulting room and in everyday life.

2.3 The Object Relations Theory Formulation of Psychoanalytic Treatment

In the light of contemporary object relations theory, the formulation based upon the structural theory (resolution of unconscious conflicts between impulse and defense) has changed, in the sense that all

unconscious conflicts are considered to be imbedded in unconscious internalized object relations. Such internalized object relations determine both the nature of the defensive operations and of the impulses against which they are directed. These internalized object relations constitute, at the same time, the 'building blocks' of the tripartite structure of id, ego, and superego. Object relations theory proposes that the gradual analysis of intersystemic conflicts between impulse and defense (structured into conflicts between ego, superego, and id, decomposes the tripartite structure into the constituent conflicting internalized object relations. These object relations are reactivated in the treatment situation in the form of an unconscious relation between self and significant others replicated in the relation between patient and analyst, that is, the 'transference.'

The transference is the unconscious repetition in the 'here and now' of unconscious, conflicting pathogenic relationships from the past. The transference reflects the reactivation of the past conflict not in the form of a memory, but in the form of a repetition. This repetition provides essential information about the past, but constitutes at the same time, a defense in the sense that the patient repeats instead of remembering. Therefore, transference has important informative features that need to be facilitated in their development, and defensive features that need to be therapeutically resolved once their nature has been clarified. Transference analysis is the fundamental ingredient of the psychoanalytic treatment.

2.4 The Psychoanalytic Treatment Process

Gill (1954) in a classical definition that is still relevant today, proposed the definition of psychoanalysis as a treatment that facilitates the development of a 'regressive transference neurosis' and its resolution by means of interpretation alone, carried out by the analyst from a position of technical neutrality.

The concept of a regressive transference neurosis refers to the gradual gathering into the relationship with the analyst of the patient's most important past pathogenic experiences and unconscious conflicts. The concept of a regressive transference neurosis has been largely abandoned in practice because, particularly in patients with severe character pathology, transference regression occurs so early and consistently that the gradual development of a regressive transference neurosis is no longer a useful concept.

Gill's proposal that the resolution of the transference be achieved 'by interpretation alone,' refers to 'interpretation' as a set of the psychoanalyst's interventions that starts with 'clarification' of the patient's subjective experiences communicated by means of free association, expands with the tactful 'confrontation' of aspects of the patient's patterns of behavior that are expressed in a dissociated or split off manner from his

subjective awareness, and thus complements the total expression of his intrapsychic life in the treatment situation, and finally evolves into 'interpretation per se.' Interpretation per se implies the formulation of hypotheses regarding the unconscious meanings in the 'here and now' of the patient's material, and the relation of these unconscious meanings with the 'there and then' of the patient's unconscious, past pathogenic experiences.

The concept of 'technical neutrality' refers to the analyst's impartiality regarding both impulse and defense, with a concerned objectivity that provides a helpful collaboration with the patient's efforts to come to grips with his intrapsychic conflicts.

This definition of the nature of psychoanalytic treatment needs to be complemented with the contemporary concepts of 'transference,' 'countertransference,' 'acting out' and 'working through.'

2.5 An Object Relations Theory Model of the Transference and Countertransference

Modern object relations theory proposes that, in the case of any particular conflict around sexual or aggressive impulses, the conflict is imbedded in an internalized object relation, that is, in a repressed or dissociated representation of the self ('self representation') linked with a particular representation of another who is a significant object of desire or hatred ('object representation'). Such units of self-representation, object representation and the dominant sexual, dependent or aggressive affect linking them are the basic 'dyadic units,' whose consolidation will give rise to the tripartite structure. Internalized dyadic relations dominated by sexual and aggressive impulses will constitute the id; internalized dyadic relations of an idealized or prohibitive nature the superego, and those related to developing psychosocial functioning and the preconscious and conscious experience, together with their unconscious, defensive organization against unconscious impulses, the ego. These internalized object relations are activated in the transference with an alternating role distribution, that is, the patient enacts a self representation while projecting the corresponding object representation onto the analyst at times, while at other times projecting his self representation onto the analyst and identifying with the corresponding object representation. The impulse or drive derivative is reflected by a dominant, usually primitive affect disposition linking a particular dyadic object relation; the associated defensive operation is also represented unconsciously by a corresponding dyadic relation between a self representation and an object representation under the dominance of a certain affect state.

The concept of countertransference, originally coined by Freud as the unresolved, reactivated transference dispositions of the analyst is currently defined

as the total affective disposition of the analyst in response to the patient and his/her transference, shifting from moment to moment, and providing important data of information to the analyst. The countertransference, thus defined, may be partially derived from unresolved problems of the analyst, but stems as well from the impact of the dominant transference reactions of the patient, from reality aspects of the patient's life, and sometimes from aspects of the analyst's life situation, that are emotionally activated in the context of the transference developments. In general, the stronger the transference regression, the more the transference determines the countertransference; thus the countertransference becomes an important diagnostic tool. The countertransference includes both the analyst's empathic identification with a patient's central subjective experience ('concordant identification') and the analyst's identification with the reciprocal object or self representation ('complementary identification') unconsciously activated in the patient as part of a certain dyadic unit, and projected onto the analyst (Racker 1957). In other words, the analyst's countertransference implies identification with what the patient cannot tolerate in himself/herself, and must dissociate, project or repress.

At this point, it is important to refer to certain primitive defensive operations that were described by Klein (1952) and her school in the context of the analysis of severe character pathology. Primitive defensive operations are characteristic of patients with severe personality disorders, and emerge in other cases during periods of regression. They include splitting, projective identification, denial, omnipotence, omnipotent control, primitive idealization, and devaluation (contempt). All these primitive defenses center on splitting, i.e., an active dissociation of contradictory ego (or self) experiences as a defense against unconscious intrapsychic conflict. They represent regression to the phase of development (the first two to three years of life) before repression and its related mechanisms mentioned are established.

Primitive defensive operations present important behavioral components that tend to induce behaviors or emotional reactions in the analyst, which, if the analyst manages to 'contain' them, permit him to diagnose in himself projected aspects of the patient's experience. Particularly 'projective identification' is a process in which: (a) the patient unconsciously projects an intolerable aspect of self experience onto (or 'into') the analyst; (b) the analyst unconsciously enacts the corresponding experience ('complementary identification'); (c) the patient tries to control the analyst, who now is under the effect of this projected behavior; and (d) the patient meanwhile maintains empathy with what is projected. Such complementary identification in the countertransference permits the analyst to identify himself through his own experience with the aspects of the patient's experience communicated by

means of projective identification. This information complements what the analyst has discovered about the patient by means of clarification and confrontation, and permits the analyst to integrate all this information in the form of a 'selected fact' that constitutes the object of interpretation. Interpretation is thus a complex technique that is very much concerned with the systematic analysis of both transference and countertransference.

2.6 Contemporary Trends of the Psychoanalytic Method

Contemporary psychoanalytic technique can be seen as having evolved from a 'one person psychology' into a 'two person psychology' and then into a 'three person psychology.' The concept of 'one person psychology' refers to Freud's original analysis of the patient's unconscious intrapsychic conflicts by analyzing the intrapsychic defensive operations that oppose free association. The 'two persons psychology' refers to the central focus on the analysis of transference and countertransference. In the views of the contemporary intersubjective, interpersonal and self psychology psychoanalytic schools, the relationship between transference and countertransference is mutual, in the sense that the transference is at least in part a reaction to reality aspects of the analyst, who therefore must be acutely mindful of his contribution to the activation of the transference. The so-called 'constructivist' position regarding transference analysis assumes that it is impossible for the analyst to achieve a totally objective position outside the transference/countertransference bind.

In contrast, the contemporary 'objectivist' position, represented by the 'three person psychology' approaches of the Kleinian school, the French psychoanalytic mainstream, and significant segments of contemporary ego psychology proposes that the analyst has to divide himself between one part influenced by transference and countertransference developments, and another part that, by means of self reflection, maintains himself/herself outside this process, as an 'excluded third party,' who, symbolically, provides an early triangulation to the dyadic regression that dominates transference developments. This triangulation in the treatment situation becomes particularly important in the treatment of severe personality disorders.

The 'repetition compulsion' as a resistance of the id is very likely a form of acting out as a defense against emotional containment of an extremely painful or traumatic set of experiences. 'Working through' refers to the repeated elaboration of an unconscious conflict in the psychoanalytic situation. It is a major task for the analyst, who has to be alert to the subtle variation in meanings and implications of what on the surface

may be appear to be an endless repetition of the same conflict in the transference. The patient elaboration of the conflict that presents itself with these repetitive characteristics also implies the function of 'holding' originally described by Winnicott (1965). It consists of the analyst's capacity to withstand the onslaught of primitive transferences without retaliation, abandonment of the patient, or a self devaluing giving up, and the maintenance of a working relationship (or 'therapeutic alliance') that addresses itself consistently to the healthy part of the patient, even when the latter is under the control of his most conflicting behaviors. Bion's (1967) concept of 'containing' is complementary to 'holding,' in the sense that holding deals mostly with the affective disposition of the analyst, and containing with his cognitive capacity to maintain a concerned objectivity and focus on the 'selected fact,' permitting the integration in the analyst's mind what the patient can only express in violently dispersed or split off behavior patterns.

'Dream analysis' developed in the context of the method of free association, and constituted, in Freud's (1953b) view, a 'royal road to the unconscious.' Freud's discovery of primary process thinking derived from his method of dream analysis. By now, psychoanalytic thinking has evolved into the view that there are many 'royal roads' to the unconscious. The analysis of character defenses, for example, or of particular transference complications, may be equally important avenues of entry into the patient's unconscious mind.

The technique of dream analysis consists, in essence, in asking the patient to free associate to elements of the 'manifest content' of the dream, in order to arrive at its 'latent' content, the unconscious wish defended against and distorted by the unconscious defensive mechanisms that constitute the 'dream work,' and have transformed the latent content into the manifest dream. The latent content is revealed with the help of the simultaneous analysis of the way in which the dream is being communicated to the analyst, the 'day residuals' that may have triggered the dream, the unconscious conflicts revealed in it, and the dominant transference dispositions in the context of which the dream evolved. Dreams also provide some residual, universal symbolic meanings that may facilitate the total understanding of the latent content.

'The analysis of character' may be the single most important element of the psychoanalytic method in bringing about fundamental characterological change. Character analysis is facilitated by the patient's use of reaction formations, that is, his defensively motivated character traits, as transference resistances. Thus, the activation of defensive behaviors in the transference, reflecting the patient's characterological patterns in all interpersonal interactions, facilitates both the analysis of the underlying unconscious conflicts, and in the process, the resolution of pathological character patterns. The result is an increase in the

patient's autonomy, flexibility, and capacity for adaptation.

The overall objective of psychoanalytic treatment is not only the resolution of symptoms and pathological behavior patterns or characteristics, but fundamental, structural change, that is, the expansion and enrichment of ego functions as the consequence of resolution of unconscious conflict and the integration of previously repressed and dynamically active id and superego pressures into ego potentialities. Such change is reflected in the increasing capacity for both adaptation to, and autonomy from, psychosocial demands and expectations, and an increased capacity for gratifying and successful functioning in love and work.

3. Treatment Results: Research on Outcome

The Menninger Psychotherapy Research Project, a naturalistic study comparing psychoanalysis, psychoanalytic psychotherapy, and supportive psychotherapy, showed psychoanalysis to be the most effective of these approaches with patients presenting relatively good ego strength, while patients with severe ego weakness—what nowadays would be described as presenting severe personality disorders or borderline personality organization—improved most with psychoanalytic psychotherapy (Kernberg et al. 1972). That research also showed how important supportive elements were throughout all modalities of treatment (Wallerstein 1986). A comprehensive review of outcome studies on psychoanalytic psychotherapy and psychoanalysis by Bachrach and collaborators (Bachrach et al. 1985) concluded that the improvement rates are in the 60–90 percent range, but it also pointed to limitations and problems in the methodology utilized.

Recently, studies regarding the treatment process and outcome of psychoanalysis and psychoanalytic psychotherapy have become more precise in defining the specific treatment variables of psychotherapeutic and psychoanalytic treatments, and several systematic studies on psychoanalytic psychotherapies and psychoanalysis are in progress (Fonagy 1998). A recent study by the Stockholm Outcome of Psychoanalysis and Psychotherapy Project (STOPPP) has found, on the basis of a relatively large patient population, that psychoanalytic treatment, in comparison with psychoanalytic psychotherapy, obtained a significantly higher degree long range symptomatic improvement (Sandell et al. 1997).

See also: Conventions and Norms: Philosophical Aspects; Ethical Issues in the 'New' Genetics; Normative Aspects of Social and Behavioral Science; Norms; Research Conduct: Ethical Codes; Research Ethics; Research; Research Funding: Ethical Aspects; Value Pluralism

Bibliography

- Abraham K 1919 A particular form of neurotic resistance against the psychoanalytic method. In: *Selected Papers on Psycho-Analysis*. Brunner/Mazel, New York, 1979, pp. 303–11
- Bachrach H M, Weber J J, Murray S 1985 Factors associated with the outcome of psychoanalysis. Report of the Columbia Psychoanalytic Research Center (IV). *International Review of Psychoanalysis* **12**: 379–89
- Bion W R 1967 *Second Thoughts. Selected Papers on Psychoanalysis*. Basic Books, New York
- Fonagy P 1998 *An Open Door Review of Outcome Studies in Psychoanalysis*
- Freud S 1953a *Three Essays on the Theory of Sexuality*, standard edn. Hogarth Press, London, **7**: 125–245
- Freud S 1953b *Fragment of an Analysis of a Case of Hysteria*, standard edn. Hogarth Press, London, **7**: 3–122
- Freud S 1953c *The Interpretation of Dreams*, standard edn. Hogarth Press, London, **4**: 1–338, **5**: 339–625
- Freud S 1955 *Beyond the Pleasure Principles*, standard edn. Hogarth Press, London,
- Freud S 1957a *Instincts and their Vicissitudes*, standard edn. Hogarth Press, London, **14**: 109–40
- Freud S 1957b *Mourning and Melancholia*, standard edn. Hogarth Press, London, Vol. 14, pp. 237–58
- Freud S 1957c *On Narcissism*, standard edn. Hogarth Press, London, Vol. 14, pp. 69–102
- Freud S 1961 *The Ego and the Id*. Standard edn. Hogarth Press, London, Vol. 19, pp. 3–66
- Gill M 1954 Psychoanalysis and exploratory psychotherapy. *Journal of the American Psychoanalytic Association* **2**: 771–97
- Grunberger B 1979 *Narcissism: Psychoanalytic Essays*. International Universities Press, New York
- Jacobson E 1964 *The Self and the Object World*. International Universities Press, New York
- Kernberg O F 1975 *Borderline Conditions and Pathological Narcissism*. Jason Aronson, Inc., New York
- Kernberg O F 1984 *Severe Personality Disorders. Psychotherapeutic Strategies*. Yale University Press, New Haven, CT
- Kernberg O F, Burnstein E, Coyne L, Appelbaum A, Horwitz L, Voth H 1972 Psychotherapy and Psychoanalysis: Final Report of the Menninger Foundation's Psychotherapy Research Project. *Bulletin Menninger Clinic* **36**: 1–275
- Klein M 1952 Notes on some schizoid mechanisms. In: Riviere J (ed.) *Developments in Psycho-Analysis*. Hogarth Press, London, pp. 292–320
- Klein M 1957 *Envy and Gratitude*. Basic Books, New York
- Kohut H 1971 *The Analysis of the Self*. International Universities Press, New York
- Racker H 1957 The meaning and uses of countertransference. *Psychoanalytic Quarterly* **26**: 303–57
- Rosenfeld H 1964 On the psychopathology of narcissism: A clinical approach. *International Journal of Psycho-Analysis* **45**: 332–7
- Rosenfeld H 1987 *Impasse and Interpretation*. Tavistock, New York
- Sandell R, Blomberg J, Lazar A 1997 When reality doesn't fit the blueprint: doing research on psychoanalysis and long-term psychotherapy in a public health service program. *Psychotherapy Research* **7**: 333–44
- Sandell R, Schubert J, Blomberg J, Carlsson J, Lazar A, Bromberg J 1997 The influence of therapist factors on outcomes of psychotherapy and psychoanalysis in the Stockholm Outcome of Psychotherapy and Psychoanalysis Project

- (STOPP). *Annual Meeting of the Society for Psychotherapy Research*, Geilo, Norway
- Wallerstein R 1986 *Forty-two Lives in Treatment: A Study of Psychoanalysis and Psychotherapy*. Guilford Press, New York
- Winnicott D 1965 *The Maturation Processes and the Facilitating Environment*. International Universities Press, New York

O. F. Kernberg

Psychoanalysis, History of

With the centenary of the birth of psychoanalysis, historians have begun to perceive that the biography of Sigmund Freud (1856–1939) and the interpretation of psychoanalysis constitutes an important chapter in twentieth-century cultural and intellectual history both as a body of theory and as cultural practice. An ongoing attempt to determine the meaning of Freud and psychoanalysis has led to the publication of hundreds of volumes, and the historiography of psychoanalysis has found its own place within the history of science. Historical literature about psychoanalysis ranges from the self-representations and autobiographical works of Freud to the first official biography, from moderate revisionism and contextualized approaches to the harsh, and at times hostile, clinical, feminist, epistemological, hermeneutical, and psychobiographical critiques of today. Each culture has absorbed Freud and psychoanalysis differently; but, a general history of psychoanalysis as a single science can be catalogued in three ways: (a) as a history of the psychoanalytic theories from Freud to the present; (b) as a history of the psychoanalytic movement, its institutions and practices; (c) as a self-reflection on the historiography of psychoanalysis itself.

1. A History of the Psychoanalytic Theories from Freud to the Present

The development of Freud's theory of the unconscious and his psychotherapeutic method can be understood on the basis of his professional development in the context of a process of differentiation and specialization of the sciences, in particular medicine and philosophy, in the late nineteenth century. Historical works of the kind, for example studies of the scientific and social context and the institutional structure of the discipline, have created a methodologically sophisticated and broad view of the psychoanalytic discipline, including empirical research and epistemological interpretation.

Various types of hermeneutical studies of psychoanalysis itself, that is, visions of it as a system of

symbols and as a method for culture-decoding rather than as a type of natural science, have broadened the view of the history of psychoanalysis. Jean Laplanche and Jean-Baptiste Pontalis's *Vocabulaire de la Psychanalyse* (1967), the first conceptual dictionary of psychoanalysis, is based on the techniques of the history of philosophy and the French tradition of the epistemology of the sciences. As the basic concepts of the psychoanalytic theory have undergone changes that reflect the development of the sociopolitical climate, the authors have incorporated the later concepts of Freud and his followers as well as critics who created new schools and theories of psychoanalysis (e.g., Melanie Klein, Jacques Lacan). Their attempts have stimulated a whole universe of separate hermeneutical and epistemological studies.

Freud's psychoanalysis with its keywords *subjectivity* and the *scientific empirical method* has been interpreted as a characteristic product of Viennese modernity, and has been seen as a parallel to phenomena like Zionism and Austromarxism. Bourgeois society and European modernity experienced significant crises in Vienna of the *fin de siècle*, where psychoanalysis was founded in the last decades of the Austro-Hungarian Monarchy. On a sociopolitical level, the multinational state tended to search for a renewal that included different forms of living. During the period between 1890 and 1910, extraordinary achievements were made in the areas of music, literature, art, and science. These were fundamentally influential for new developments in observation and thought in the twentieth century, and the psychoanalytic theory enjoyed its first heyday in the years before World War I. The following factors were generative forces behind the *fin de siècle* creativity: the rise of new cultural founding figures of the liberal bourgeoisie, the influence of the Jewish intelligentsia, and the upward mobility of immigrants from the Austro-Hungarian Empire; the resistance against the political dirigism, bureaucracy and the repression of the Church and censorship; the cultural networks established with European centers and circles abroad; the import of foreign ideas; and the intellectual exchange between the submetropolises of the countries of the monarchy (Nautz and Vahrenkamp 1993).

From 1859 until 1938, Sigmund Freud lived and worked in Vienna, and the significance of Vienna as the site of psychoanalysis' foundation and development is evident. Socio-politico-cultural studies have dealt with Freud's *fin de siècle* (see Schorske 1981, McGrath 1986), viewing psychoanalysis within Viennese culture, for example at the juncture of a specific time and place. Special interest has focused on Freud's Jewish identity and the relationship between Judaism and psychoanalytic theory, ranging from minimalistic to maximalistic explanations.

Psychoanalytic theory and Freud's biography were and in many ways still are intertwined in a unique way. Therefore, the historian is in the position of having to

treat psychoanalytic concepts as both strictly personal documents and as data governed by the architecture of the theoretical system. In Freud's early texts, the evolution of psychoanalytic theory is presented linearly, as though it were fundamentally the same as any other scientific field, but there are serious historiographic difficulties entailed in approaching psychoanalysis in this fashion. Freud's self-analysis, his autobiographical writings, and self-created image of psychoanalysis as a science and various theoretical concepts of psychoanalysis (e.g., seduction theory, dream interpretation, Oedipal conflict, resistance, repression, transference, and countertransference) have become prominent in the field of psychoanalytic historiography, including Freud's own changes of theory throughout his life, as well as post-Freudian developments (e.g., drive-psychology, ego-psychology, the development from defense to adaptation, from instinctual to motivational diversity).

2. A History of Freud's Biography, the Psychoanalytic Movement, and its Institutions and Practices

In the historiography of psychoanalysis, the genre of Freud studies has been differentiated from the historical examination of the psychoanalytic movement in its local national centers and within specific cultures. Biographical studies dominate over sociohistorical works, and the focus has been on the struggle over legacy and legitimacy rather than on a social history of the profession.

2.1 Freud Biographies

Three groups of historical literature about Freud have been described: (a) publications of primary sources, for example Freud's correspondences; (b) publications of historical materials concerning case studies and theoretical works of Freud; and (c) critical-academic studies about psychoanalysis. The history of Freud biography and the psychoanalytic movement began with Freud's autobiographical works *On the History of the Psychoanalytic Movement* (1914) and *An Autobiographical Study* (1925), and in his own historical papers, Freud—in his one-sided but coherent story—suggests that the history of psychoanalysis is basically the history of refusals, resistances, and the necessary regroupings of the orthodoxies (ideas, institutions and people/practitioners themselves). His story of psychoanalysis appears to be synonymous with the history of its splits, schisms, and dissidents. But, just as the proliferation of schools and the consequent battles among and splittings within psychoanalytic institutes stimulated abundant polemics, these upheavals also stimulated reworkings of almost every important item in the history of Freud biography.

Dreams from *The Interpretation of Dreams*, patients from *The Studies on Hysteria* and later case studies have been analyzed and reanalyzed from new perspectives; all the schisms and quarrels in the history of psychoanalysis were rediagnosed by the latter-day schismatics and schism-haters. Based on newly accessed archival material and historical information, early speculations and widely differing interpretations have been reconsidered. Freud biographies have addressed almost every aspect of Freud's life (profession, social status, political views, culture, religion, affection-emotion), sometimes in a general fashion, sometimes in a very specific manner, and reflecting the time and the standpoint of their authors. To forestall future falsifications in popular biographies and early attempts at studying Freud's biography by his pupils and followers, Freud's daughter Anna Freud felt prompted to authorize Ernest Jones to write a first full-scale biography (Young-Bruehl 1998). Freud studies proper began with the pioneering research of Siegfried Bernfeld; his scholarship was invaluable to Ernest Jones when putting together his trilogy *The Life and Work of Sigmund Freud* (1953–57), yet, at least some of Bernfeld's findings were suppressed in Jones'. As a member of the first generation of loyal followers, Jones' monumental work became the precipitant for a third type of historical accounts: critiques of Jones's biography, or what might be called second-order biographies.

For a long time, Freud's prominence nurtured the illusion of an isolated, ahistoric development of psychoanalysis. A cultural-historical approach to Freud became more intense as the 1960s unfolded, as biography was in general too role-model-oriented. With Henri Ellenberger's *The Discovery of the Unconscious* (1970), a mode of Freud biography contextualizing Freud began to flourish, taking the precursors of Freud's ideas more into account than Jones did.

2.2 The Psychoanalytic Movement

Interest in supporters, students, and contributors of Freud has long been marginal or limited to only the most well-known representatives. With the publication of *The Minutes of the Vienna Psychoanalytic Society* (Nunberg and Federn 1962–75), a new focus was put on the institutional framework, the form in which these individuals cooperated scientifically, and how they organized their exchanges with other social structures. Works on Freud's friendships and professional relationships include the dynamical and sociological aspects of the formation of groups, the production of scientific knowledge, the structure of mastery and slavery (often basic to the formation of psychoanalytic institutions), as well as more contextually embedded accounts of specific psychoanalytic cultures. Further, *Freud and his Followers* (Roazen 1975) stimulated studies of the family networks and

subnetworks; as a consequence, biographies of early pupils and later analysts have emerged. Their adherent or schismatic relations to Freud figured largely in these accounts. A special body of historical literature has focused on the role and influence of female patients and analysts. The founding of new psychoanalytic schools has created a new intellectual-historical approach, including the professionalization and institutionalization of the discipline in different cultures and at different times. From the early days of psychoanalysis, the internationalization of psychoanalysis had its own motor and effect on theory and practice (e.g., Kutter 1992, 1995). Taking geographical and temporal particularities into account, a Western (European and North American) perspective has been criticized.

Economic and political circumstances, as well as cultural and sociohistorical connections, have all influenced the theoretical, practical, and organizational developments of psychoanalysis as a discipline and are part of the production of psychoanalysis as a corpus of knowledge. Fenichel (1998) has observed the impacts of outside reality on the production of theory and practice during the years when Central European Jewish psychoanalysts, candidates in training and their families were forced to flee National Socialist Germany and Austria and the occupied countries in Europe. The production of scientific findings and the transformation and transfer of knowledge as social processes in which internal and external factors are in a constant exchange does not minimize the role of the founder of the science of the unconscious in its beginning, but it destroys the illusion of a single isolated discovery. The shifts in the history of psychoanalysis can be seen and interpreted as a complex, dialogical, multi-author enterprise and depend on the studies of the special scientific tradition and the study of psychoanalytic microcultures (see Hale 1995, Lockot 1985, Roudinesco 1982).

3. A Self-reflection on the Historiography of Psychoanalysis Itself, and the Problems of Methodology

The history of psychoanalysis over the past generation has been a rapidly growing, controversy-ridden, and attention-attracting area. Over the past 30 years, each of the historical visions has inspired and/or provoked a revision, and each revision a counter-revision. The history of psychoanalysis has to take into account the peculiar, complex, and powerful interplay among psychoanalysis, historiography, and social, cultural, and political ideology; because this has not been recognized, many widely disparate accounts of psychoanalytic history have been proffered. Neighboring fields, such as the history of medicine, philosophy, psychology and science, women's studies, the history of law, crime, and deviance, the history of

professions, the history of sexuality, the history of the body, and cultural studies, literature, queer/gay/gender studies, have had their impacts on the historiography of psychoanalysis as well.

Studies in the area of Freud biography and the pre- and early history of psychoanalysis have received more attention. Because connections with later periods seem less powerful and convincing, the fixation on Freud himself and, even more particularly, on the originary decade of Freud's discoveries have meant that later developments during and especially since Freud's lifetime are receding in prominence.

The various schools of psychoanalysis have fashioned for themselves idealized genealogies, lines of intellectual filiation between the great minds of the past and contemporary practitioners, tracing a suitable line of evolution, and have been selective of their usable past. They may also assist in sharpening contested interdisciplinary or intradisciplinary boundaries in which 'positive' episodes are memorialized, taken as characteristic of the essence, while 'negative' ones are omitted, serving to legitimate particular theories and methodologies and delegitimizing others. Historical accounts of psychoanalytic theory and practice have to include thoughts and speculations about the special acquisition and transmission of psychoanalytic knowledge.

A 'coherent' picture of Freud and the psychoanalytic movement has undergone a profound change. In the historiography of psychoanalysis, at least two different schools of thought have been described: the 'official' history of Freud, which has been largely though not exclusively produced and consumed by practicing psychoanalysts, and a 'revisionist' approach that has emerged in the last three decades and has been mainly produced by people working outside of institutional psychoanalysis (e.g. historians, psychologists, philosophers, sociologists, political scientists). For a long time, the two groups have debated separately—with a trend to specialization and fragmentation thereby repeating the fragmentation of the object they are analyzing—and have tried to avoid other findings. But lately, a dialogue between historians and clinicians—both interested in history—concerning the history of psychoanalysis has arisen, and there have been attempts to overcome prejudices. One outcome of this dialogue has been the formation of local and international organizations and discussion groups and the establishment of special journals and periodicals. Still, the issues of language, including the problems of translation and the questions concerning authorship and audience remain central, as in other scientific disciplines. Language questions have been correlated with speculative histories on Freud and the psychoanalytic movement.

After the historical optimism of the mid-century period, the height of the professional prestige of psychoanalysis, when there was remarkably little criticism from historians of psychiatry, a revisionist

historiography of psychoanalysis was stimulated; traditional biography and the general longitudinal narrative were followed by the study of institutions introducing the ideal of objective, systematic, and quantitative analysis, having in mind methodologically sophisticated, empirically substantial, and sociologically-oriented scholarship. On the other hand, as scholars have argued, the demystifications of hagiographical idealizations have often only generated 'heroic' neo-marxist and Foucauldian remystifications. Revisionist and hagiographical literature have been diagnosed by contemporary scholars as both reductionistic and lacking self-reflection.

Psychoanalysis as a theory and a profession emphasizes the value and virtue of self-awareness and in its therapeutical practices often employs a developmental mode of inquiry; nevertheless, with a few exceptions, it has remained resolutely unintrospective about the methodological and ideological conditions under which it writes the narrative of its own past. That is all the more peculiar in light of the past decade's extensive and sophisticated discussion concerning different conceptual, methodological, and epistemological strategies for writing the history of science generally. Two approaches have been observed, the internalist approach, stressing cognitive development, and the externalist approach, focusing on the social, economic, political, and professional determinants of scientific theory and practice. All history writing has been diagnosed perspectively as the expression of a particular historical environment, culture, background, and self-consciousness among cultural-critics. LaCapra (1988) emphasizes the Freudian point that any historian's relationship with the past, and through that with the present, is a transference relationship, and that, in this transference, repetition and denial come more easily than working through and mourning (Smith and Morris 1992).

Jones' idealization of Freud has been rejected, and, particularly in the medium of psychobiography, a much more contextualized, conflicted, and complicated founder of psychoanalysis has emerged in the historical literature. Jones did not interpret blanks in his documentation as secrets or resistances, but nowadays psychologically informed biographers consider the fact that many of the most important things in people's lives are never recorded.

Contrary to the sophisticated histories of Freud's unfolding discoveries, historians have lamented the difficulties of psychobiography, and some of the key methodological problems concerning childhood experiences, the workings of the mind, and the character of thought, have been widely recognized—although such recognition has not yet informed a major historiographical work on psychoanalytic biography and history writing.

Despite the progressive de-Freudianization of psychiatry and psychology in general, Freud's place as one of the supreme minds of the twentieth century

remains secure for most historians. Psychoanalysts have altered psychoanalytic theory over the course of time, their theoretical and methodological findings depending on their analysands, their social context, and the historico-political changes in general. Historiography has altered accordingly, considering and tolerating the complexity instead of simplistic and popular versions of psychoanalysis.

See also: Cognitive Psychology: History; Defense Mechanisms; Ego Psychology and Psychoanalysis; Feminist Theory: Psychoanalytic; Freud, Sigmund (1856–1939); Gender and Feminist Studies; Gender and Feminist Studies in Economics; Gender and Feminist Studies in History; Gender and Feminist Studies in Political Science; Gender and Feminist Studies in Psychology; Gender and Feminist Studies in Sociology; Jung, Carl Gustav (1875–1961); Klein, Melanie (1882–1960); Personality Disorders; Personality Theories; Personality Theory and Psychopathology; Psychiatry, History of; Psychoanalysis: Adolescence (Clinical–Developmental Approach); Psychoanalysis in Clinical Psychology; Psychoanalysis in Sociology; Psychoanalysis: Overview; Psychology, History of; Psychology: Overview; Psychotherapy, History of; Psychiatric Aspects; Social Psychology; Transference in Psychoanalysis; Unconscious: History of the Concept

Bibliography

- Ellenberger H F 1970 *The Discovery of the Unconscious. The History and Evolution of Dynamic Psychiatry*. Allen Lane, London
- Fenichel O 1998 *119 Rundbriefe Europa (1934–1938), Amerika (1938–1945)* [Mühlleitner E, Reichmayr J (eds.)]. Stroemfeld Verlag, Basel, Frankfurt am Main
- Forrester J 1994 A whole climate of opinion: rewriting the history of psychoanalysis. In: Micale M S, Porter R (eds.) *Discovering the History of Psychiatry*. Oxford University Press, New York, Oxford
- Freud S 1914 Zur Geschichte der psychoanalytischen Bewegung. In: *Gesammelte Werke X*, S Fischer, Frankfurt am Main [1957 On the history of the psychoanalytic movement. In: *The Standard Edition of the Complete Psychological Works of Sigmund Freud XIV*. Hogarth Press, London]
- Freud S 1925 Selbstdarstellung. In: *Gesammelte Werke XIV*. S Fischer, Frankfurt am Main [1959 An autobiographical study. In: *The Standard Edition of the Complete Psychological Works of Sigmund Freud XX*. Hogarth Press, London]
- Grubrich-Simitis I 1993 Zurück zu Freuds Texten. Stumme Dokumente sprechen machen. S Fischer, Frankfurt am Main [In: Grubrich-Simitis I 1996 *Back To Freud's Texts: Making Silent Documents Speak*. Yale University Press, New Haven, CT]
- Hale N G Jr 1995 *The Rise and Crisis of Psychoanalysis in the United States: Freud and the Americans, 1917–1985*. Oxford University Press, New York
- Haynal A, Falzeder E (eds.) 1994 *100 Years of Psychoanalysis: Contributions to the History of Psychoanalysis*. H. Karnac, London

- Jones E 1953–57 *The Life and Work of Sigmund Freud* 1st edn. Basic Books, New York, 3 Vols.
- Kearns K 1997 *Psychoanalysis, Historiography, and Feminist Theory. The Search for a Critical Method*. Cambridge University Press, Cambridge, UK
- Kurzweil E 1989 *Freudians: a Comparative Perspective*. Yale University Press, New Haven, CT
- Kutter P (ed.) 1992, 1995 *Psychoanalysis International: a Guide to Psychoanalysis Throughout the World*. Frommann-Holzboog, Stuttgart, Bad Cannstatt
- LaCapra D 1988 History and psychoanalysis. In: Meltzer F (ed.) *The Trail(s) of Psychoanalysis*. University of Chicago Press, Chicago
- Laplanche J, Pontalis J-B 1967 *Vocabulaire de la Psychanalyse*. Presses Universitaires de France, Paris [1973 *The Language of Psychoanalysis*. Hogarth Press and the Institute of Psychoanalysis, London]
- Lockot R 1985 *Erinnern und Durcharbeiten: Zur Geschichte der Psychoanalyse und Psychotherapie im Nationalsozialismus*. Fischer Taschenbuch Verlag, Frankfurt am Main
- Lorenzer A 1984 *Intimität und Soziales Leid*. Archäologie der Psychoanalyse. S Fischer, Frankfurt am Main
- Loewenberg P 1995 *Fantasy and Reality in History*. Oxford University Press, New York
- McGrath W J 1986 *Freud's Discovery of Psychoanalysis*. Cornell University Press, Ithaca, NY
- Mühlleitner E 1992 *Biographisches Lexikon der Psychoanalyse. Die Mitglieder der Psychologischen Mittwoch-Gesellschaft und der Wiener Psychoanalytischen Vereinigung 1902–38*. Edition Diskord, Tübingen
- Nautz J, Vahrenkamp R (eds.) 1993 *Die Wiener Jahrhundertwende. Einflüsse, Umwelt, Wirkungen*. Böhlau, Wien
- Nunberg H, Federn E 1962–75 *The Minutes of the Vienna Psychoanalytic Society*. International Universities Press, New York, 4 Vols
- Reichmayr J 1994 *Spurensuche in der Geschichte der Psychoanalyse*. Fischer Taschenbuch, Frankfurt am Main
- Roazen P 1975 *Freud and His Followers*. Knopf, New York
- Roudinesco E 1982–86 *La bataille de cent ans. Histoire de la psychanalyse en France*. Editions Ramsay, Paris, 2 Vols. [1990 *Jacques Lacan & Co: A History of Psychoanalysis in France 1925–1985*. University of Chicago Press, Chicago]
- Schorske C E 1981 *Fin-de-Siècle Vienna*. Vintage Books, New York
- Smith J, Morris H (eds.) 1992 *Telling Facts. History and Narration in Psychoanalysis*. Johns Hopkins University Press, Baltimore, MD
- Young-Bruehl E 1998 *Subject to Biography: Psychoanalysis, Feminism, and Writing Women's Lives*. Harvard University Press, Cambridge, MA

E. Mühlleitner

Psychoanalysis in Clinical Psychology

1. Historical Background to Psychoanalysis

As is well known, Freud's discovery of the talking cure was really that of an intelligent patient (Anna O) and her physician (Breuer). The patient found out for herself that certain symptoms disappeared when she

succeeded in linking up fragments of what she said and did in an altered state of consciousness (which we would now call dissociative) with forgotten impressions from her waking life. Breuer's remarkable contribution was that he had faith in the reality of the memories which emerged and did not dismiss the patient's associations as products of a deranged mind. Breuer, of course, left the therapeutic battlefield as soon as his patient's speech and behaviors started to go beyond the boundaries marked out by turn-of-the-century Viennese society.

Psychotherapy had to wait for a more powerful intellect who would not take fright at the instinctual elements in the psychological world of so-called civilized man. At first Freud sought to unearth the traumas which he believed lay at the origins of the neuroses. Later, when confronted by evidently incorrect statements, he modified his theory assuming consistency between recollection and childhood psychic reality rather than physical reality. The issue of accuracy of memories of childhood sexual trauma remains controversial, although its relevance to psychoanalytic technique is at best tangential. Freud's technique, however, was dramatically modified by his discoveries. The intense emotional relationship between patient and physician, which had its roots in catharsis following hypnotic suggestion, had gradually subsided into what was principally an intellectual exercise to reconstruct the repressed causes of psychiatric disturbance from the fragments of material derived from the patient's associations. It was a highly mechanistic approach reminiscent of a complex crossword puzzle. In the light of therapeutic failures, however, Freud once more restored the emotional charge into the patient-physician relationship. However, in place of hypnosis and suggestion, he used the patient's emotional reactions to the therapist. He thought that these reactions were explicable as transferences of the patients' feelings about their past relationships on to the current analytic relationship. Instead of seeing the patient's intense emotional reaction to the therapist as an interference, Freud came to recognize the importance of transference as a representation of earlier relationship experiences which could make the reconstruction of those experiences in analysis highly meaningful to that individual (see *Therapist-Patient Relationship*). Freud's early clinical work lacked some of the rigor which came to characterize classical psychoanalysis. His occasional encouragement to his patients to join him on holiday might now be considered boundary violations. What is perhaps less well known is that Freud remained somewhat skeptical about the effectiveness of psychoanalysis as a method of treatment. Indeed, autobiographies of some of his patients testify to his great flexibility as a clinician and use of nonpsychoanalytic techniques, including behavioral methods. Nor was Freud the only clinician to use psychoanalytic ideas flexibly. The Hungarian analyst Sandor Ferenczi was

one of the first to discover the treatment of phobic disorders by relaxation and exposure, although his discovery had little impact on subsequent clinical practice.

The technique of psychoanalysis after Freud's death came to be codified with those (such as Alexander and French, and Frieda Fromm-Reichmann) who attempted to revive or retain Freud's original clinical flexibility being subjected to powerful intellectual rebuttals. In reality, psychoanalysts probably continued to vary in the extent to which they observed the ideals of therapeutic neutrality, abstinence, and a primarily interpretive stance, but these deviations could no longer be exposed to public scrutiny for fear of colleagues' forceful condemnation. Personal accounts of analyses with leading figures yield fascinating insights into variations in technique, principally in terms of the extent to which the analyst made use of personal relationship. There has been an ongoing dialectic throughout the history of psychodynamic approaches between those who emphasize interpretation and insight and those who stress the unique emotional relationship between patient and therapist as the primary vehicle of change. The controversy dates back to disputes concerning the work of Ferenczi and Rank but re-emerged with the first papers of Balint and Winnicott in London opposing a Freudian and Kleinian tradition, and somewhat later in the United States with Kohut opposing classical ego psychology.

Since the early 1990s, the pluralistic approach of modern psychoanalysis has brought out into the open many important dimensions along which psychoanalysts' techniques may vary. In particular, the recent trend to consider analyst and patient as equal partners engaged in a mutual exploration of meaning has directly challenged many of the classical constructs. This controversy is too fresh to permit a conclusive assessment. It is evident, however, that fairly fundamental reconsideration of the aims and techniques of psychoanalytic therapy is currently underway.

2. The Basic Assumptions of Psychoanalysis

Notwithstanding the theoretical heterogeneity of psychoanalytically based therapeutic approaches, there is probably a core set of assumptions to which all psychodynamic therapists would, to a greater or lesser extent, subscribe. These can be basically summarized as follows: (a) intrapsychic conflict is an all-pervasive aspect of human experience; (b) the mind is organized to avoid displeasure arising out of conflict and maximize the subjective sense of safety; (c) defensive strategies are adopted to manipulate ideas and experience in order to minimize displeasure; (d) psychological disturbance arises developmentally, so an adaptation that was rational and reasonable at an earlier developmental phase may leave residues that

cause maladjustment in adulthood; (e) social experience is organized into relationship representations; self and other are depicted in specific interactions charged by powerful emotional content; (f) these relationship representations inevitably re-emerge in the course of psychoanalytic treatment.

2.1 Technique—Principal Features

2.1.1 Neutrality and abstinence. Based on the classical framework of libidinal theory, Freud explicitly enjoined against the analyst giving in to the temptation of gratifying the patient's sexual desire. Obviously, this is primarily an ethical issue. However, within the psychoanalytic context it also justifies the analyst's stance of resisting the patient's curiosity or using the therapeutic relationship in any way that consciously or unconsciously could be seen as motivated by the need to gratify their own hidden desires. Within this classical frame of reference, the patient must also agree to forgo significant life changes where these could be seen as relevant to current psychotherapeutic work. In practice, such abstinence on the part of the patient is rare. Yet long-term psychodynamic treatment may founder if the emotional experiences of the therapy are obscured by the upheavals of significant life events.

Abstinence primarily ensures the neutrality of the therapist. The therapist must resist the temptation to direct the patient's associations and must remain neutral irrespective of the subject matter of the patient's experiences or fantasies. It is perhaps this aspect of the psychoanalyst's therapeutic stance that makes psychoanalysts most vulnerable to ridicule, but it is probably genuinely critical for the therapist to retain emotional distance from the patient to a degree which enables the latter to bring fantasies and fears of which they feel uncertain. Nevertheless, neutrality at its worst denies the possibility of sensitivity; recent literature on the process and outcome of psychotherapy makes it clear that the therapist's genuine concern for the patient must be manifest if significant therapeutic change is to be achieved.

2.1.2 Mechanisms of defense. Classical psychoanalytical theory and its modern equivalent (ego psychology) see conflict at the core of mental functioning. They see defenses as adaptations to intrapsychic conflict. Within object relations theories, defenses are seen as helpful to the individual to maintain an authentic or 'true' self representation or a nuclear self (see *Defense Mechanisms*). Attachment theory understands defenses as assisting in the maintenance of desirable relationships. The Klein-Bion model makes limited use of the notion of defense mechanisms but uses the term in the context of more complex hypo-

thetical structures called defensive organizations. The term underscores the relative inflexibility of some defensive structures which are thus best conceived of as personality types. For example, narcissistic personality disorder combines idealization and destructiveness; genuine love and truth are devalued. Such a personality type may have been protective to the individual at an earlier developmental stage, and has now acquired a stability or autonomy which must be rooted in the emotional gratification which such a self-limiting form of adaptation provides.

Irrespective of the theoretical frame of reference, from a therapeutic viewpoint clinicians tend to differentiate between so-called primitive and mature defenses based on the cognitive complexity entailed in their functioning. In clinical work, primitive defenses are often noted together in the same individual. For example, individuals loosely considered 'borderline' tend to idealize and then derogate the therapist. They maintain their self-esteem by using splitting (clear separation of good from bad self-perception) and then projection. Projective identification is an elaboration of the process of projection. An individual may ascribe an undesirable mental state to the other through projection but when the other can be unconsciously forced to accept the projection and experience its impact, the defense becomes far more powerful and stable. The analyst's experience of a fragment of the patient's self state has in recent years been considered an essential part of therapeutic understanding.

Through projective identification the patient can experience a primitive mode of control over the therapist, whether in fantasy or in actualized form. Bion argued that when the self is experienced as being within another person (the therapist), the patient frequently attempts to exert total control over the recipient of the projection as part of an attempt to control split-off aspects of the self. Bion also argued that not all such externalizations were of 'bad' parts of the self. Desirable aspects of the self may also be projected, and thus projective identification can be seen as a primitive mode of communication in infancy. There are other aspects of projective identification which we commonly encounter clinically. These include the acquisition of the object's attributes in fantasy, the protection of a valued aspect of the self from internal persecution through its evacuation into the object, and the avoidance or denial of separateness. It is thus a fundamental aspect of interpersonal relationships focused on unconscious fantasy and its appreciation is critical for the adequate practice of long-term psychotherapy.

Some mechanisms of defense are thought to be more characteristic of the less severe psychological disorders (e.g., depression, anxiety, obsessive-compulsive disorders, etc.). It is beyond the scope of this article to consider the various defense mechanisms in detail. Since Anna Freud's classical work, these defense mechanisms have become fairly generally

accepted, if only as terms of mild rebuke between mental health professionals. Thus we generally accept that motivated repression may be associated with momentary forgetting of conflictual contents, that denial may be invoked by individuals wishing to disavow the emotional significance of an experience, that reaction formation is helpful in turning terror into aggression, and that identification with the aggressor may be the only solution available to a maltreated child who becomes an abuser in adolescence.

The diagnostic significance of defense mechanisms is controversial. Some researchers have claimed that an individual's habitual mode of defense has predictive value beyond that of psychiatric diagnosis. Given the theoretical ambiguity that surrounds the concept, it is unlikely that its use as part of a diagnostic formulation is justified. It does, however, assist the psychoanalytic clinician in conceptualizing the patient's reactions. In particular, these defenses are likely to be encountered in the course of psychoanalytic treatment at moments of intense emotional resistance. For example, it is common to observe patients experiencing considerable difficulty in recalling the contents of treatment sessions, yet their memory for other less central aspects of their lives appears to be exceptional. Being alert to the presence of defenses is important, not because it provides an opportunity for confrontation with ideas which the patient seems reluctant to acknowledge, but rather because it alerts the analyst to the presence of underlying anxieties which need to be tackled if resistance to the therapy is to be overcome.

2.1.3 Modes of therapeutic action. The primary mode of the therapeutic action of psychoanalytic psychotherapy is considered to be *insight*. Insight may be defined as the conscious recognition of the role of unconscious factors in current experience and behavior. Unconscious factors encompass unconscious feelings, experiences, and fantasies. Insight is more than mere intellectual knowledge. Thomä and Kächele (1987) consider insight to be equidistant from emotional experience and intellect. Etchegoyen (1991) distinguished descriptive insights from demonstrated (ostensive) insights which represent a more direct form of knowing, implying emotional contact with an event one has experienced previously.

Although specific formulations of the effect of insight depend on the theoretical framework in which explanations are couched, there is general agreement that insight's therapeutic effect is to integrate mental structures in some way (Thomä and Kächele 1987). Kleinian theorists see the healing of defensively created splits in the patient's representation of self and others as crucial. One may be more specific by specifying split or part-objects as isolated representations of intentional beings whose motivation is insufficiently well understood for these to be seen as coherent beings. In this case insight could be seen as a

development of the capacity to understand internal and external objects in mental state terms, thus lending them coherence and consistency. The same phenomenon may be described as an increasing willingness on the part of the patient to see the interpersonal world from a third-person perspective.

A simple demonstration to the patient of such an integrated picture of self or others is not thought to be sufficient. The patient needs to work through a newly arrived at integration. Working through is a process of both unlearning and learning: actively discarding prior misconceptions and assimilating learning to work with new constructions. The technique of working through is not well described in the literature, yet it represents the critical advantage of long-term over short-term therapy. Working through should be systematic and much of the advantage of long-term treatment may be lost if the therapist does not follow through insights in a relatively consistent and coherent manner.

In contrast to the emphasis on insight and working through are those clinicians who, as we have seen, emphasize the relationship aspect of psychoanalytic therapy (Balint, Winnicott, Loewald, Mitchell, and many others). This aspect of psychoanalytic therapy was perhaps most eloquently described by Loewald when he wrote about the process of change as: 'set in motion, not simply by the technical skill of the analyst but by the fact that the analyst makes himself available for the development of a new "object-relationship" between the patient and the analyst ...' (Loewald 1960 pp. 224–5). Sandler and Dreher (1996) have recently observed 'while insight is aimed for, it is no longer regarded as an absolutely necessary requirement without which the analysis cannot proceed.' There is general agreement that the past polarization of interpretation and insight on the one hand and bringing about change by presenting the patient with a new relationship on the other was unhelpful. It seems that patients require both and both may be required for either to be effective.

It has been suggested that change in analysis will always be individualized according to the characteristics of the patient or the analyst. For example, Blatt (1992) suggested that patients who were 'introjective' (preoccupied with establishing and maintaining a viable self-concept rather than establishing intimacy) were more responsive to interpretation and insight. By contrast, anaclitic patients (more concerned with issues of relatedness than of self-development) were more likely to benefit from the quality of the therapeutic relationship than from interpretation.

2.2 Psychoanalysis in Clinical Psychology

While intensive long-term psychoanalysis is relatively rarely offered nowadays as a therapy except to those engaged in psychotherapeutic training, long-term psychoanalytic psychotherapy, using the same principles but with session frequency at once or twice a week is

perhaps the most common long-term treatment offered for personality and emotional problems by clinical psychologists. Clinical psychology is increasingly concerned with the evidence base of the treatments that psychological therapists engage in. What hope is there in this era of empirically validated treatments, which prizes brief structured interventions, for a therapeutic approach which defines itself by freedom from constraint and preconception, and counts treatment length not in terms of number of sessions but in terms of years? Can psychoanalysis ever demonstrate its effectiveness, let alone cost-effectiveness?

There can be no excuse for the thin evidence base of psychoanalytic treatment. In the same breath that psychoanalysts often claim to be at the intellectual origin of other talking cures (e.g., systemic therapy, cognitive behavior therapy), they also seek shelter behind the relative immaturity of the discipline to explain the absence of evidence for its efficacy. Yet the evidence base of these 'derivatives' of psychoanalytic therapy has been far more firmly established than evidence for psychoanalysis itself. Of course, there are reasons given for this—reasons such as the long-term nature of the therapy, the complexity of its procedures, the elusiveness of its self-declared outcome goals, and the incompatibility of direct observation with the need for absolute confidentiality. None of these reasons stands up to careful scrutiny, however.

What is surprising, given this unpropitious backdrop is that there is in fact some suggestive evidence for the effectiveness of psychoanalysis as a treatment for psychological disorder.

2.2.1 The evidence base of psychoanalytic treatment. Psychoanalysts have been encouraged by the body of research which supports brief dynamic psychotherapy (see *Psychological Treatment, Effectiveness of*). A meta-analysis of 26 such studies has yielded effect sizes comparable to other approaches (Anderson and Lambert 1995). It may even be slightly superior to some other therapies if long-term follow-up is included in the design. One of the best-designed Randomized Controlled Trials (RCTs), the Sheffield Psychotherapy Project (Shapiro et al. 1995), found evidence for the effectiveness of a 16-session psychodynamic treatment based on Hobson's (1985) model in the treatment of major depression. There is evidence for the effectiveness of psychodynamic therapy as an adjunct to drug dependence programs (Woody et al. 1995). There is ongoing work on a brief psychodynamic treatment for panic disorder (Milrod et al. 1997). There is evidence for the use of brief psychodynamic approaches in work with older people (Thompson et al. 1987). There are some excellent studies confirming the cost-effectiveness of psychoanalytic brief therapy in a naturalistic clinical setting.

There are psychotherapy process studies which offer qualified support for the psychoanalytic case. For example, psychoanalytic interpretations given to clients which are judged to be accurate are reportedly associated with relatively good outcome (Crits-Christoph et al. 1988). There is even tentative evidence from the reanalysis of therapy tapes from the NIMH (National Institutes of Mental Health) Treatment of Depression Collaborative Research Program that the more the process of a brief therapy (CBT, cognitive behavior therapy, IPT, Interpersonal therapy) resembles that of a psychodynamic approach, the more likely it is to be effective.

Evidence is available to support therapeutic interventions which are clear derivatives of psychoanalysis. However, most analysts would consider that the aims and methods of short-term, once-a-week psychotherapy are not comparable to 'full analysis.' What do we know about the value of intensive and long-term psychodynamic treatment? Here the evidence base becomes somewhat patchy.

The Boston Psychotherapy Study (Stanton et al. 1984) compared long-term psychoanalytic therapy (two or more times a week) with supportive therapy for clients with schizophrenia in a randomized controlled design. There were some treatment-specific outcomes, but on the whole clients who received psychoanalytic therapy fared no better than those who received supportive treatment. In a more recent randomized controlled study (Bateman and Fonagy 1999), individuals with a diagnosis of borderline personality disorder (BPD) were assigned to a psychoanalytically oriented day hospital treatment or treatment as usual. The psychoanalytic arm of the treatment included therapy groups three times a week as well as individual therapy once or twice a week over an 18-month period. There were considerable gains in this group relative to the controls and these differences were not only maintained in the 18 months following discharge but increased, even though the day hospital group received less treatment than the control group. A further controlled trial of intensive psychoanalytic treatment of children with chronically poorly controlled diabetes reported significant gains in diabetic control in the treated group which was maintained at one-year follow-up (Moran et al. 1991). Experimental single case studies carried out with the same population supported the causal relationship between interpretive work and improvement in diabetic control and physical growth (Fonagy and Moran 1991). The work of Chris Heinicke also suggests that four- or five-times-weekly sessions may generate more marked improvements in children with specific learning difficulties than a less intensive psychoanalytic intervention (Heinicke and Ramsey-Klee 1986).

One of the most interesting studies to emerge recently was the Stockholm Outcome of Psychotherapy and Psychoanalysis Project (Sandell 1999). The study followed 756 persons who received national

insurance funded treatment for up to three years in psychoanalysis or in psychoanalytic psychotherapy. The groups were matched on many clinical variables. Four- or five-times-weekly analysis had similar outcomes at termination when compared with one to two sessions per week psychotherapy. However, in measurements of symptomatic outcome using the SCL-90, improvement on three-year follow-up was substantially greater for individuals who received psychoanalysis than those in psychoanalytic psychotherapy. In fact, during the follow-up period, psychotherapy patients did not change but those who had had psychoanalysis continued to improve, almost to a point where their scores were indistinguishable from those obtained from a nonclinical Swedish sample.

Another large pre-post study of psychoanalytic treatments has examined the clinical records of 763 children who were evaluated and treated at the Anna Freud Centre, under the close supervision of Freud's daughter (Fonagy and Target 1996). Children with certain disorders (e.g., depression, autism, conduct disorder) appeared to benefit only marginally from psychoanalysis or psychoanalytic psychotherapy. Interestingly, children with severe emotional disorders (three or more axis I diagnoses) did surprisingly well in psychoanalysis, although they did poorly in once or twice a week psychoanalytic psychotherapy. Younger children derived greatest benefit from intensive treatment. Adolescents appeared not to benefit from the increased frequency of sessions. The importance of the study is perhaps less in demonstrating that psychoanalysis is effective, although some of the effects on very severely disturbed children were quite remarkable, but more in identifying groups for whom the additional effort involved in intensive treatment appeared not to be warranted.

The Research Committee of the International Psychoanalytic Association has recently prepared a comprehensive review of North American and European outcome studies of psychoanalytic treatment (Fonagy et al. 1999; also see <http://www.ipa.org.uk>). The committee concluded that existing studies failed to unequivocally demonstrate that psychoanalysis is efficacious relative to either an alternative treatment or an active placebo, and identified a range of methodological and design problems in the 50 or so studies described in the report. Nevertheless, the report is encouraging to psychoanalysts. A number of studies testing psychoanalysis with 'state of the art' methodology are ongoing and are likely to produce more compelling evidence over the next years. Despite the limitations of the completed studies, evidence across a significant number of pre-post investigations suggests that psychoanalysis appears to be consistently helpful to patients with milder (neurotic) disorders and somewhat less consistently so for other, more severe groups. Across a range of uncontrolled or poorly controlled cohort studies, mostly carried out in Europe, longer

intensive treatments tended to have better outcomes than shorter, nonintensive treatments. The impact of psychoanalysis was apparent beyond symptomatology, in measures of work functioning and reductions in health care costs.

3. Conclusions

Psychoanalytic notions are no longer considered 'outlandish' from the point of view of empirical psychology. New discoveries from the experimental human sciences have yielded much data consistent with psychoanalytic ideas. None of these studies can 'prove' psychodynamic ideas as the two pertain to independent realms of human functioning: the laboratory and the consulting room. However, taken as a whole, they point to the existence of underlying psychological processes posited by psychodynamic theoreticians when offering their model of human development and the pathological outcomes that may arise from it. Psychodynamic theory is unlikely ever to be comprehensively validated in empirical studies. All that we can hope for is that the general model of a person, painstakingly assembled on the basis of neuropsychological, social, and cognitive-psychological investigations, will not be too dissimilar from the vision of human functioning which has emerged from the clinical endeavors of psychodynamically oriented practitioners.

See also: Defense Mechanisms; Ego Psychology and Psychoanalysis; Feminist Theory: Psychoanalytic; Freud, Sigmund (1856–1939); Psychoanalysis: Adolescence (Clinical–Developmental Approach); Psychoanalysis: Current Status; Psychoanalysis, History of; Psychoanalysis in Sociology; Psychoanalysis: Overview; Psychological Treatment, Effectiveness of; Psychology of the Real Self: Psychoanalytic Perspectives; Therapist–Patient Relationship; Transference in Psychoanalysis

Bibliography

- Anderson E M, Lambert M J 1995 Short-term dynamically oriented psychotherapy: A review and meta-analysis. *Clinical Psychology Review* **15**: 503–14
- Bateman A, Fonagy P 1999 The effectiveness of partial hospitalization in the treatment of borderline personality disorder—a randomised controlled trial. *American Journal of Psychiatry* **156**: 1563–9
- Blatt S J 1992 The differential affect of psychotherapy and psychoanalysis with anaclitic and introjective patients: The Menninger Psychotherapy Research Project revisited. *Journal of the American Psychoanalytic Association* **40**: 691–724
- Crits-Christoph P, Cooper A, Luborsky L 1988 The accuracy of therapists' interpretations and the outcome of dynamic psychotherapy. *Journal of Consulting and Clinical Psychology* **56**: 490–5

- Etchegoyen H 1991 *The Fundamentals of Psychoanalytic Technique*. Karnac, London
- Fonagy P, Kachele H, Krause R, Jones E, Perron R, Lopez L 1999 *An Open Door Review of Outcome Studies in Psychoanalysis*. International Psychoanalytical Association, London
- Fonagy P, Moran G S 1991 Studies of the efficacy of child psychoanalysis. *Journal of Consulting and Clinical Psychology* **58**: 684–95
- Fonagy P, Target M 1996 Predictors of outcome in child psychoanalysis: A retrospective study of 763 cases at the Anna Freud Centre. *Journal of the American Psychoanalytic Association* **44**: 27–77
- Heinicke C M, Ramsey-Klee D M 1986 Outcome of child psychotherapy as a function of frequency of sessions. *Journal of the American Academy of Child Psychiatry* **25**: 247–53
- Hobson R F 1985 *Forms of Feeling: The Heart of Psychotherapy*. Basic Books, New York
- Loewald H W 1960 On the therapeutic action of psycho-analysis. *International Journal of Psycho-analysis* **41**: 16–33
- Milrod B, Busch F, Cooper A, Shapiro T 1997 *Manual for Panic-focused Psychodynamic Psychotherapy*. American Psychiatric Press, Washington, DC
- Moran G, Fonagy P, Kurtz A, Bolton A, Brook C 1991 A controlled study of the psychoanalytic treatment of brittle diabetes. *Journal of the American Academy of Child and Adolescent Psychiatry* **30**: 926–35
- Sandell R 1999 *Long-term Findings of the Stockholm Outcome of Psychotherapy and Psychoanalysis Project (STOPPP)*. Paper presented at the Psychoanalytic Long-term Treatments: A Challenge for Clinical and Empirical Research in Psychoanalysis, Hamburg, Germany
- Sandler J, Dreher A U 1996 *What do Psychoanalysts Want? The Problem of Aims in Psychoanalysis*. Routledge, London and New York, Vol. 24
- Shapiro D A, Rees A, Barkham M, Hardy G, Reynolds S, Startup M 1995 Effects of treatment duration and severity of depression on the maintenance of gains after cognitive-behavioral and psychodynamic-interpersonal psychotherapy. *Journal of Consulting and Clinical Psychology* **63**: 378–87
- Stanton A H, Gunderson J G, Knapp P H, Vancelli M L, Schnitzer R, Rosenthal R 1984 Effects of psychotherapy in schizophrenia: I. Design an implementation of a controlled study. *Schizophrenia Bulletin* **10**: 520–63
- Thomä H, Kächele H 1987 *Psychoanalytic Practice. I: Principles*. Springer-Verlag, New York
- Thompson L W, Gallagher D, Breckenridge J S 1987 Comparative effectiveness of psychotherapies for depressed elders. *Journal of Consulting and Clinical Psychology* **55**: 385–90
- Woody G E, McLellan A T, Luborsky L, O'Brien C P 1995 Psychotherapy in community methadone programs: A validation study. *American Journal of Psychiatry* **152**: 1302–8

P. Fonagy

Psychoanalysis in Sociology

Psychoanalysis, as developed by Sigmund Freud and his followers, has had a major impact upon sociological theory and modern sociology. Freud's central

discoveries—the unconscious, sexual repression, the Oedipus complex, and the like—have been deployed by sociologists to interpret and discuss the self and human subjectivity, gender and sexuality, the family and socialization, language and ideology, as well as the formation of cultural identities and forms of political domination. Notwithstanding this impact, sociology as a discipline has long had a difficult, and indeed fraught, relationship with psychoanalysis. Sociologists have criticized psychoanalytical theory on the grounds of its methodology, epistemology, and ontology. Despite these criticisms, many contemporary sociologists remain engaged with, and some strongly committed to, the psychoanalytic tradition in order to conceptualize the relation between the individual and society, especially the complex, contradictory ways that human subjects acquire and reshape the ideas, values, symbols, beliefs, and emotional dispositions of the wider society. This has been particularly evident over recent decades, in which Freudian themes and psychoanalytic motifs have been used to analyze sexual politics, issues of identity and lifestyle, as well as the debate over modernity and postmodernism.

1. Freud and the Sociological Imagination

Freudian psychoanalysis shares with sociology a primary preoccupation with the fate of the individual or self in the context of social relationships and the wider cultural process. While Freud's own writings were primarily derived from his clinical work with patients, and to that degree are at variance with the core methodologies of mainstream social science, his characterization of human personality and self-identity has clear parallels with, say, Thomas Hobbes' theory of human nature or Karl Marx's account of the self-seeking individual in organized capitalist society. Yet whereas both Hobbes and Marx stress the impact of social forms in the constitution of the self, Freud's methodological starting point is the individual psyche, principally the instinctual impulses and libidinal longings that shape the human imagination. The self for Freud is radically fractured or divided, split between consciousness of identity and the repressed unconscious. The biographical trajectory of the self, according to Freud, is carried on against the backdrop of this radical otherness of the unconscious, a domain of the psyche that infuses three agencies of the psyche: id, ego, and superego. The id, lying at the root of unconscious desire, is that which cannot be symbolized yet constantly strives for expression in our daily lives, manifesting itself in dreams, day-dreams, slips of the tongue, and the like. It is a hidden area of the self which knows no reality, logic, or contradiction; the unconscious is at the root, says Freud, of how people for example can simultaneously love and hate their

parents. The superego, as internal anchoring of cultural prohibition, is founded in this id. Desire, according to Freud, infiltrates all human intentions, ideals, and imperatives. Similarly, Freud sees parts of the ego as interwoven with the force of the id, the self arising as a product of the unconscious.

While Freud's writings on human personality and the constitution of the self have proven of interest to sociologists, it is his account of the relations between self and society—primarily his late writings on civilization—that has most influenced the development of modern sociology. In his late writings, Freud comes to see human beings living under the destructive force of a terrifying death drive, as well as strict cultural prohibitions on sexual desire and enjoyment. These themes are set out in his magisterial books *Beyond the Pleasure Principle* (1920) and *Civilization and its Discontents* (1930). Civilization, writes Freud, is repressive. Society imposes severe psychic demands upon individuals to achieve cultural conformity, demands that produce intense personal misery and neurotic suffering. Some sociologists have interpreted this aspect of Freud's sociology as at one with the conservative social thought of Hobbes, likening it to the theorem that cultural reproduction demands the overcoming of chaotic passion. But it is clear that this involves a radical misunderstanding of Freud's position. The dualisms that Freud uses to analyze modern civilization (consciousness/unconscious, desire/repression, the pleasure principle/the reality principle, Eros/Thanatos) suggests a disjunction between self and world. According to Freud, ambivalence is at the core of the individual's relation to the self, to others, and to society.

Freud's writings on the fate of the self in contemporary culture have strongly influenced sociological debates—from Herbert Marcuse to Michel Foucault. Too much repression, Freud says, leads to intense unconscious anguish, hostility, and rage. At such a point, the intensification of unconscious desire can release the 'mental dams' of sexual repression in a far-reaching way. The issue of the subjective seeds of social and political transformation are thus at the heart of Freud's theoretical contribution to sociology and social theory.

2. The Integration of Psychoanalysis and Sociology

With regard to social analysis and sociological theory, Freudian ideas have loomed large in the sociological conceptualization of human subjectivity and interpersonal relationships, sexuality and modern culture, as well as the mix of reason and irrationality in politics and history. There have been three key sociological developments in this connection: first there has been several early and influential articulations of psycho-

analysis and sociology developed in Europe (particularly as represented in the work of the Frankfurt school of sociology) and North America (especially in the writings of Talcott Parsons); second, structuralist and poststructuralist appropriations of psychoanalysis as refracted in sociological debate; and finally, feminist and postmodern social theory has been significantly shaped by psychoanalysis.

2.1 Early Disciplinary Articulations

In Europe, Herbert Marcuse and Theodor Adorno—leading members of the Frankfurt school of critical theory—turned to Freud in order to reconceptualize the relation between self and society. The political motivation prompting this turn to Freud had its roots in Marcuse's and Adorno's attempts to conceptualize the rise of fascism, Nazism, and also the spread of bureaucratic capitalism. From Freud's theory of the Oedipus complex, and particularly his theorem concerning the repression of infantile sexuality, Marcuse and Adorno developed the notion of 'the authoritarian personality.' Driven by a desire for conformity and clear rules, the authoritarian personality was viewed by Marcuse and Adorno as a character type strongly prevalent in the German middle classes, a character type who hungered for strong leadership, social order, and regulation. Not only in Nazi Germany, however, was this personality type to be found; in the advanced liberal societies of the West, tendencies towards authoritarianism and conformism are increasingly evident. Marcuse thus spoke of the emergence of 'one-dimensional man.'

Marcuse's radical Freudianism, in particular, won a wide audience in the 1960s—not only in social science circles, but among student activists and sexual liberationists. Arguing that the so-called 'sexual revolution of the 1960s' did not seriously threaten power structures of the established social order, Marcuse sought to show how demands for freedom were routinely rechanneled for commercial interests. The core of his analysis rested upon the distinction he drew between basic and surplus repression. Basic repression he defined as the minimum level of psychological renunciation demanded by the social structure and cultural order. Repression that is surplus, by contrast, refers to the intensification of self-restraint demanded by asymmetrical relations of power. Marcuse describes the 'monogamic-patriarchal' family, for example, as a site of surplus repression. Interestingly, while Marcuse saw signs of surplus repression increasingly everywhere in late capitalist society, he remained remarkably optimistic about the possibilities for social and cultural change.

On the other side of the Atlantic, a different approach to the integration of psychoanalysis and sociology was fashioned. The core thematic of this

approach concerned social order, socialization, and the reproduction of the social system. The grand theorist of American sociology, Talcott Parsons, employed Freudian ideas to understand how basic symbols and values are internalized by human subjects throughout the socialization process. According to Parsons's appropriation of Freud for social theory, the structure of human personality is an outcome of an internalization of desired objects, role relationships, and ethico-cultural values that make up the broader social network. In this approach, it is the linkage of personality structure, the social system, and the cultural system that is stressed. Unlike Marcuse's and Adorno's emphasis on the social manipulation of the unconscious, Parsons finds a kind of pre-established harmony between the individual and society.

2.2 *Structuralist and Poststructuralist Psychoanalytic Sociology*

For many years the integration of psychoanalysis and social analysis developed by the Frankfurt school was commonly regarded as the most sophisticated and important work in this subfield of modern sociology. From the late 1960s onwards, however, the impact of French theory, particularly structuralist and post-structuralist philosophy, became increasingly influential in terms of theorizing the social dimensions of psychoanalysis. The key figure in this connection was Freud's French interpreter, Jacques Lacan. Seeking to rework the core concepts of psychoanalysis in the light of modern linguistics, Lacan argued that the unconscious exemplifies key features of language; as Lacan famously argues, 'the unconscious is structured like a language.' The subject or 'I,' according to Lacan, is not self-transparent, but is rather located in a system of signification from which identity is fashioned. For Lacan, intersubjectivity is at the center of psychological functioning and its disturbances; distortions or pathologies at the level of the self are, says Lacan, located in 'the discourse of the other.'

It is perhaps Lacan's essay 'The mirror stage as formative of the function of the I' (1949) that has come to exert most influence on contemporary sociological theory. In the essay, Lacan conceptualizes the small infant's initial recognition of itself in a mirror or reflecting surface, and of how this generates a sense of identity. Through the mirror, says Lacan, the infant makes an imaginary identification with its reflected image, an identification which the infant reacts to with a sense of jubilation and exhilaration. But the mirror image of the self for Lacan is, in fact, a distortion; the mirror lies. The mirror stage is radically 'imaginary,' in Lacan's theorization, since the consolingly unified image of selfhood which it generates is diametrically opposed to the bodily fragmentation and lack of coordination of the child. These imaginary traps and distortions are a universal and timeless feature of self-organization, and Lacan sees such illusions as directly

feeding into and shaping pathologies of the self in contemporary culture.

Lacan was not especially interested in the social applications of psychoanalysis; it was one of his followers, the French Marxist political philosopher Louis Althusser, who brought Lacanian theory into the center of key debates in sociology. In his important essay '*Ideology and Ideological State Apparatuses*' (1971), Althusser analyzed ideology in terms of the process by which individuals come to understand and relate to themselves in a manner which supports dominant class relations. According to Althusser, ideology provides an imaginary identity, an imagined map for locating oneself in the wider social network. Echoing Lacan, Althusser uses the notion of the mirror stage to deconstruct ideology. There is a duplicate mirror-structure at the heart of ideology says Althusser, a structure that grants to the self an ideological mirror in which it can recognize itself and other people. Althusser calls this process 'interpellation'—the capturing of the individual within the net of received social meanings.

The Lacanian/Althusserian account of the de-centering of the subject has been highly influential in recent sociological theory, and has impacted upon debates concerning agency, structure, class, social fragmentation, and cultural order. The social-theoretical work of Paul Hirst, Barry Hindess, Stuart Hall, Etienne Bailbar, Pierre Machery, Fredric Jameson, and Slavoj Žižek is, in differing ways, indebted to the Lacanian/Althusserian theory of the subject and its ideological subjection.

2.3 *Feminism, Postmodernism, and Psychoanalytic Sociology*

It is not only in studying politics and social change that psychoanalysis has become an important theoretical tool for sociology, but also in debates concerning gender and sexual politics. Freudian ideas have been incorporated into social theory. In the sociology of sexuality, the sociology of the family, and especially the social theory of gender transformation, Freudian psychoanalysis has played a vital role in expanding sociologists' understanding of the subjective and affective components of human social relationships.

The psychoanalytic perspective made a forceful entry into contemporary feminist sociological theory in Juliet Mitchell's pioneering book, *Psychoanalysis and Feminism* (1974). Mitchell deployed Freudian and Lacanian psychoanalytic ideas as a means of connecting a discussion of gender power with an Althusserian/Marxist theory of late capitalist society. Against this theoretical backdrop, she asserted that definitions of masculinity and femininity are framed through linguistic and historical structures—with man as a self-determining, autonomous agent, and woman as a lacking other. Such gender dualism, according to Mitchell, is highly conducive to capitalist social

regulation—the split between private and public, the pathologies of the familial life, and the like. Mitchell's ideas, while criticized in some feminist sociological circles, had a lasting impact on psychoanalytically-oriented feminist sociology. Especially in terms of Lacanian feminist approaches, the writings of authors such as Julia Kristeva, Luce Irigaray, Jacequeline Rose, and Judith Butler have significantly influenced sociological debate in recent years.

In the United States, the feminist theories of Nancy Chodorow, Jessica Benjamin, and Jane Flax have been influential in contemporary sociology. These feminist authors draw from the psychoanalytic perspective, but rather than turn to Lacan and French psychoanalysis their work selectively incorporates the insights of Freudian and post-Freudian (especially object-relational) theory. In Chodorow's work, it is part of an attempt to understand the psychic components of female and male socialization, especially in terms of the unconscious forces that shape gender roles. In Benjamin's work, psychoanalysis is deployed to rethink the dynamics of domination and submission within the wider frame of gender, society, and history. In Flax's discussion of psychoanalysis, feminism, and postmodernism, it is primarily a set of philosophical observations about the development of gender relations and the sociology of sexuality and intimacy.

Related to the intertwining of psychoanalysis and feminism, postmodernist appropriations of psychoanalysis have also been influential in recent sociological interventions. This has been especially true for the contested modernity/postmodernity debate, in which sociologists as diverse as Anthony Giddens, Alain Touraine, and Zygmunt Bauman have drawn from Freud to analyze anew the self and self-identity. In the writings of Michel Foucault, Jean-François Lyotard, and Jacques Derrida, which have also had a strong influence on sociological theory, there has been much debate about the fate of the individual or 'death of the subject' in postmodernist culture. In all of these discussions, psychoanalysis has provided sociology with conceptual tools for questioning and deconstructing the enlightenment and rationality of progress.

See also: Critical Theory: Contemporary; Critical Theory: Frankfurt School; Ego Psychology and Psychoanalysis; Feminist Theory; Feminist Theory: Psychoanalytic; Freud, Sigmund (1856–1939); Gellner, Ernest (1925–95); Parsons, Talcott (1902–79); Postmodernism in Sociology; Psychoanalysis, History of; Structuralism, History of; Structuralism, Theories of; Unconscious: History of the Concept

Bibliography

Althusser L 1984 Ideology and ideological state apparatuses. *Essays on Ideology*. Verso, London

Chodorow N 1978 *The Reproduction of Mothering: Psychoanalysis and the Sociology of Gender*. University of California Press, Berkeley, CA

Elliott A 1999 *Social Theory and Psychoanalysis in Transition: Self and Society from Freud to Kristeva*. Free Association Books, London and New York

Freud S 1935–74 *The Standard Edition of the Complete Psychological Works of Sigmund Freud* [trans. Stachey J]. Hogarth Press, London

Lacan J 1977 *Écrits: A Selection*. Tavistock Press, London

Marcuse H 1956 *Eros and Civilization*. Ark, London

Mitchell J 1974 *Psychoanalysis and Feminism*. Allen Lane, London

Smelser N J 1998 *The Social Edges of Psychoanalysis*. University of California Press, Berkeley, CA

A. Elliott

Psychoanalysis: Overview

By psychoanalysis is generally understood, in the first instance, the psychological theory and corresponding clinical practice developed by Sigmund Freud in the first four decades of the twentieth century, and in most usage the term refers also to at least some of the developments that proceed directly from Freud, such as Jungian, Kleinian, and Lacanian theory, though the proper boundaries of psychoanalysis are hard to mark and subject to dispute. This article will, after a brief sketch of Freud's development, survey the principal theoretical issues raised by psychoanalytic theory, the main developments in psychoanalysis after Freud, and the relation of psychoanalysis to other fields of human enquiry.

1. Freud's Life and Writings

Freud was born of a Jewish family in Freiberg (now in the Czech republic) in 1856. He qualified as a physician at the University of Vienna, becoming early involved in neurological research under the tutelage of the physiologist Ernst Brücke, a positivist and Darwinian. After three years of medical practice at the General Hospital in Vienna, Freud received an appointment as lecturer in neuropathology. A period spent at the Salpêtrière in Paris exposed Freud to the French neurologist Jean Charcot's use of hypnotherapy to treat mental disorder. In 1886 Freud set up in private practice in Vienna, applying Charcot's methods, and in 1893 he published with the Viennese physician Josef Breuer *Studien über Hysterie* [*Studies on Hysteria*], an account of the use of hypnosis to effect the cathartic recall of forgotten traumatic experience, underpinned by a schematic theory of mental dissociation. This led Freud to attempt to formulate a general theory of mental functioning in strict neurological terms, his

'Project for a scientific psychology' of 1895, which he abandoned without completing. Thereafter, Freud concentrated on the analysis of clinical material provided by his patients and undertook a self-analysis. The explorations of these years led to the publication in 1900 of *Die Traumdeutung* [*The Interpretation of Dreams*], Freud's first full statement of psychoanalytic ideas. Here Freud defended the substitution of free association for hypnosis as a clinical method. Over the following decade Freud published *Zur Psychopathologie des Alltagslebens* [*The Psychopathology of Everyday Life*] (1904) and *Drei Abhandlungen zur Sexualtheorie* [*Three Contributions to the Theory of Sexuality*] (1905), and began to acquire a following, leading to the founding in 1910 of the International Psychoanalytic Association. The spread of psychoanalysis in Europe and the United States was accompanied by internal disagreement and the formation of alternative schools, most notably those of Alfred Adler and Carl Jung. Freud's later writings focused on theoretical questions and the application of psychoanalysis to cultural issues. Forced by the Anschluss to flee Austria, Freud died in London in 1939. (Jones 1953–7 and Gay 1988 are classic intellectual biographies of Freud. Decker 1977 analyses Freud's intellectual historical context.)

2. Fundamental Conceptual Issues in Psychoanalytic Theory

Though subject to fierce criticism from many quarters, Freud's theories achieved over the course of the twentieth century a massive degree of influence in all of the humanities and human sciences, and contemporary publication suggests that their power to stimulate is by no means exhausted. The theoretical dissemination of psychoanalytic ideas far exceeds the prevalence of psychoanalysis as a clinical practice: though institutes of psychoanalysis are established and thrive in most Western countries, psychoanalytic therapy has been consistently repudiated by the medical establishment and accorded no recognized place in psychiatric treatment. The greater clinical influence of psychoanalysis has been achieved through dilution, its transmutation into less demanding and theoretically committed forms of treatment: psychotherapy in its myriad contemporary forms, though it most often represents itself as wholly distinct from psychoanalysis, is barely conceivable without its historical relation to Freud.

Both the influence exerted by psychoanalysis and the chronic controversy surrounding it are closely connected to the highly ambiguous, indeterminate status of psychoanalytic theory, reflected in the host of epistemological, methodological, and conceptual problems which it sets. Freud presented his theories, which he revised constantly, in terms that display very

clearly the importance for his development of diverse strands in nineteenth-century thought, including philosophy and a number of the natural sciences (Darwinian biology, sexology, evolutionary anthropology, brain anatomy, associationist psychology; see Sulloway 1979 and Kitcher 1992). Furthermore, the range of phenomena that Freud regarded psychoanalysis as subsuming grew continuously, from specific forms of individual pathology (hysteria, neurosis) and abnormal psychological explananda (dreams, parapraxes) to mental functioning in general, and thence to morality, religion, art, and the social order. These factors—the continual development and interdisciplinary scope of Freud's theories, and the heterogeneity of the theoretical sources on which Freud drew—help to account for the paradoxical appearance that psychoanalysis gives of being at once an attempt to extend naturalistic, mechanistic, and reductionist modes of explanation to the human sphere, and a hermeneutical enterprise fundamentally more akin to literary criticism and philosophical reflection than to biology.

The conceptual indeterminacy of psychoanalytic theory may be analyzed under three headings.

2.1 The Structure of Psychoanalytic Theory

Though Freud appealed increasingly to the application of psychoanalytic ideas to social and cultural phenomena as a source of corroboration, and some post-Freudian developments have sought to shift the emphasis in this direction, psychoanalytic theory is primarily a theory of individual mental functioning. There is, therefore, a basic distinction to be drawn between the trans-individual applications of psychoanalysis and the psychoanalytic theory of the individual mind, which has logical priority in all contexts of psychoanalytic explanation.

The psychoanalytic theory of the individual mind may be divided for analytical purposes into three levels: the clinical level at which interpretations of individuals are formulated and case histories composed; low-level hypotheses concerning particular kinds of psychological configuration, such as Freud's theories of psychosomatic hysterical disorders, obsessional neurosis, psychosis, mourning and melancholia, disturbances of memory, sexual orientation, formation of male and female identity, and character types; and the highest, unifying level of psychoanalytic theory, called by Freud 'metapsychology,' consisting of abstract, semiphilosophical assumptions about the nature and general form of the mind.

At the center of the metapsychology is Freud's concept of the dynamic unconscious. The notion of the unconscious had an important place in nineteenth-century thought (Ellenberger 1970, Whyte 1979), versions of the concept being discoverable in philos-

ophers such as Arthur Schopenhauer, Friedrich Nietzsche, and Eduard von Hartmann, and in the nascent empirical psychology of Friedrich Herbart, Gustav Fechner, Hermann von Helmholtz, and Wilhelm Wundt. Occurrences of the concept of the unconscious in nineteenth-century writings tended, however, to be either theoretically undeveloped (as in Schopenhauer), or to play a merely ancillary role in metaphysical speculation (as in Hartmann), or to be qualified by uncertainty regarding the reference and explanatory value of the concept (as in the empirical psychologists). Freud may be said to have given the concept a fully scientific character, insofar as he both brought the philosophical conception of the unconscious into close relation with problems of psychological explanation and gave the concept a theoretical and empirical determinacy which it lacked previously.

Freud conceived the unconscious as an instinctually grounded source of motivation which is directly responsible for the ubiquity of conflict in normal mental life and substantially independent from cognition in its mode of operation. The nature of unconscious mental activity is explicated through Freud's theories of wish-fulfillment, of primary (as opposed to secondary) process thinking, and of the pleasure (as opposed to the reality) principle, these theories being closely associated with a developmental story. Though Freud asserted the autonomy of psychological characterizations of the unconscious from neurological speculation in *The Interpretation of Dreams*, and always claimed that the unconscious is composed of 'ideas' rather than somatic states, i.e., that it has an intentional character, at all periods of his development Freud regarded 'economic' descriptions of mental life—descriptions in terms of quantities of neural energy—as also essential to the metapsychology. Psychoanalytic theory thus incorporates suggestions of both dualism and physicalist reductionism.

Within this framework, the metapsychology underwent continual modification, the most important development being Freud's move from an account that ties the existence of the dynamic unconscious directly to repression—the first topographical model of *Ucs.*, *Pcs.*, and *Cs.*, articulated in Freud's metapsychological papers of 1915—to the structural model of ego, id, and superego presented in *Das Ich und das Es* [*The Ego and the Id*] (1923), according to which the unconscious is aboriginal, and distributed across the major portion of the ego as well as the id. Whereas the first model implies that unconscious mentality is created only through aberrations in conscious mental life, which remains the core of subjectivity, the second model implies that consciousness requires a special 'making conscious' of mental states and that human subjectivity by its nature extends beyond consciousness. The structural model may thus be argued to impact in a quasi-philosophical manner on the traditional and commonsensical conception of the self or

person. (For scholarly expositions of Freud's metapsychological concepts, see Laplanche and Pontalis 1967.)

Though Freud's own writings stress the interdependence of the three levels, there is some vacillation in Freud between regarding the metapsychology as the crowning and defining achievement of psychoanalysis, and reducing it to a mere superstructure that might be radically overhauled and is replaceable, in principle if not in practice, by neurological science. Questions may thus be raised concerning the relations of the three levels, and in this context various possibilities present themselves. Psychoanalysis may be identified with either a theoretical edifice offering a general, philosophically significant model of the mind; or a cluster of lower-level theories aiming at the formulation of psychological laws in specific subdomains of psychology; or an accumulated body of case studies of individuals serving primarily to specify a therapeutic technique. These options—in numerous variations and combinations—have been defended and explored by theorists in close connection with different views of the epistemological foundations of psychoanalysis.

2.2 Epistemological and Methodological Issues

Because of the physicalist–reductionist strain in psychoanalysis, and the positivistic character of the conception of knowledge appealed to by Freud in his claims for the scientificity of psychoanalysis, there has been a natural (and, in the English-speaking world, dominant) tendency to consider the question of the empirical warrant for psychoanalytic theory in the context of scientific methodology. This mode of approach to psychoanalysis is associated with a negative estimate of the objectivity and cognitive value of psychoanalysis.

Two major writings in this sphere are Karl Popper (1963, Chap. 1) and Adolf Grünbaum (1984). Popper conducted an enormously influential attack on psychoanalysis in the course of defending falsifiability as the key to scientific method. According to Popper, psychoanalysis does not open itself to refutation, and so fails to satisfy the condition of falsifiability which, in his account, provides the necessary and sufficient condition for determining the scientific standing and candidacy for rational acceptability of a theory. In view of its immunity to counterevidence, Popper holds, psychoanalysis must be classified (alongside Marxism) as a 'pseudo-science.' Grünbaum, operating with a different conception of scientific method, maintains in opposition to Popper that psychoanalysis does meet the conditions for evaluation as a scientific theory, but proceeds to offer a detailed critique of what he takes to be Freud's inductive reasoning. Freudian theory reposes, Grünbaum argues, on the claims that only psychoanalysis can give correct

insight into the cause of neurosis, and that such insight is causally necessary for a durable cure. Grünbaum then emphasizes the empirical weakness of psychoanalysis' claim to causal efficacy, and presses the frequently voiced objection that the therapeutic effects of psychoanalysis may be due—for all that Freud is able to demonstrate to the contrary—to suggestion. Grünbaum argues further that, even if the clinical data could be taken at face value, the inferences that Freud draws are unwarranted.

Different conceptions of science carry different implications for the epistemological status of psychoanalysis, and a sufficiently anti-rationalistic conception of science, such as Paul Feyerabend's, may succeed in eliminating the epistemological discrepancy between psychoanalysis and the natural sciences; but this strategy carries obvious costs. It is generally agreed that, without repudiating altogether the concept of a distinctive scientific method in the manner of Feyerabend, the prospect of defending the epistemological credentials of psychoanalysis in the terms supplied by the philosophy of natural science is poor. Attempts to test psychoanalytic hypotheses experimentally in controlled, extraclinical contexts have been inconclusive (see Eysenck and Wilson 1973). The comparison with cognitive science, which has successfully claimed for itself empirical support, may appear to aggravate the epistemological situation of psychoanalysis. (For discussion of psychoanalysis' scientificity, see Hook 1964 and Edelson 1984.)

An alternative approach to the epistemological issue is to abandon the assumption that psychoanalysis must be shown to conform to the model of explanation of the natural sciences in order for it to be held to have cognitive value. This involves breaking with Freud's own understanding of psychoanalysis, but it may draw support from the fact that Freud also on numerous occasions emphasizes the alignment of psychoanalysis with, and its dependence on, commonsense principles of reasoning: the everyday assumptions about the mind that are embedded in commonsense ('folk') psychology.

A diversity of approaches falls under this heading, but they share agreement that the right way to understand psychoanalysis is to regard it as reworking the routes of psychological understanding laid down in our ordinary, everyday grasp of ourselves, and validated independently from scientific methodology.

Some philosophers, encouraged by Ludwig Wittgenstein's writings on the philosophy of psychology and remarks on Freud (Wittgenstein 1978), explored the notion that psychoanalytic explanation trades off the concept of a *reason* rather than that of a cause, and tried to suggest that psychoanalytic interpretation is directly continuous with the practice of explaining agents' actions by citing their reasons. Though psychoanalysis is thereby freed from the burden of having to appeal to strict causal laws, the epistemological grounds of which prove so problematic, an evident

difficulty faces this approach, arising from the tension between the assumption of the agent's rationality which is presupposed by the very application of the concept of a reason, and the nonrational or irrational character of the mental connections postulated in psychoanalytic explanation. The resulting concept of 'neurotic rationality' employed by some philosophers reflects this discomfort.

However, more satisfactory solutions along these lines can be developed by making the relation of psychoanalysis to commonsense psychology less direct. One such approach regards psychoanalysis as a *theoretical extension* of commonsense psychology. It is argued that the familiar schema of practical reason explanation, formalized in the practical syllogism, is in psychoanalytic explanation fundamentally modified by the substitution of the concepts of wish and fantasy for those of belief and desire, and by the replacement of a direct, intrapsychic relation between desire and mental representation for the syllogistic relation between practical reasoning and action. The basis for assigning content and explanatory role to wishes and fantasies—thematic linkages, relations of meaning—remains, however, the same. Melanie Klein's development of Freud's theories, which attributes enormous importance to the role of fantasy in mental life, is standardly appealed to by proponents of this approach. In this view, the overarching epistemological ground for psychoanalysis lies in its capacity to offer a unified explanation for phenomena—dream, psychopathology, mental conflict, sexuality, and so on—that commonsense psychology is unable or poorly equipped to explain (see *Irrationality: Philosophical Aspects*), while exploiting the same interpretative methodology as commonsense psychology and complementing commonsense psychological explanations. (See Wollheim 1974, 1991, Wollheim and Hopkins 1982, Hopkins 1992, Neu 1991, Cavell 1993, Levine 1999.)

An alternative epistemological route from commonsense psychology to psychoanalysis is to proceed by way of *philosophical theory*: given an independently elaborated philosophical theory of psychoanalytically relevant aspects of human subjectivity, intersubjectivity, and representation, it becomes possible to interpret psychoanalysis in the terms provided by this theory and thereby extend to psychoanalysis whatever cognitive authority the theory possesses. Proponents of this approach typically draw on different philosophical traditions from the commonsense-extension theorists, whose philosophical allegiances characteristically reflect a mixture of empiricism and Wittgenstein. (The two approaches are, however, not necessarily incompatible.)

Jürgen Habermas and Jacques Lacan provide two clear examples of this mode of approach. Habermas' (1968, Chaps. 10–11) hermeneutic account asserts a complete separation of psychoanalysis from the natural sciences, this association being attributed to a

naturalistic and scientific misconception of psychoanalysis on Freud's part, and seeks instead to integrate psychoanalysis with communication theory and to set it in the context of a nontraditional conception of rationality and cognitive interest. The unconscious and its symbolism are understood by Habermas in terms of specific forms of distortion of communication and interruptions to self-communication; psychoanalytic interpretation and treatment are understood correspondingly as directed to retrieving elements that have suffered exclusion from the field of public communication, and to engendering a form of self-reflection that will undo self-alienation and thereby emancipate. (Ricoeur 1965 offers an alternative hermeneutic approach.)

Lacan's reading of Freud (see Evans 1996) evinces a similar methodological structure: a general theory of subjectivity and representation, in Lacan's case deriving from multiple sources including Ferdinand de Saussure and G. W. F. Hegel, is employed to elucidate psychoanalytic concepts. The concept of the unconscious is explicated by Lacan in terms of paradoxes arising necessarily from the attempt to represent the self and its objects, and constraints put on interpersonal desire by its symbolic mediation, while Freud's theories of the nature of unconscious mental processing are recast in terms borrowed from structural linguistics.

2.3 Further Theoretical Issues

An issue that has plagued psychoanalysis since its inception concerns the conceptual possibility of mental phenomena being unconscious. Nineteenth-century philosophy of psychology raises frequently the question whether the posit of unconscious mental phenomena is warranted or even intelligible, Franz Brentano for example devoting to it a chapter of his *Psychologie vom empirischen Standpunkte* [*Psychology from an Empirical Standpoint*] (1874). Freud's own response was to emphasize the dogmatism of those who reject the notion of unconscious mental phenomena in advance of considering what explanatory gains may be made by employing it, these being of course, in Freud's view, decisively favorable. It is fair to say that beyond this point little progress has been made with the problem, which to that extent remains without a satisfactory solution. This does not, however, constitute a serious embarrassment for psychoanalysis, since the identity of mental phenomena with consciousness maintained by opponents of psychoanalysis is as hard to supply with a philosophical proof as the concept of unconscious mental phenomena is hard to make intuitively acceptable. The issue has in any case been largely overtaken, or at any rate moved down the contemporary philosophical agenda, by the confident

positing of unconscious mental processing witnessed by Noam Chomsky and cognitive psychology.

In its place a new question has emerged concerning what it may mean to speak of mental content which is prelinguistic, unconceptualized, or nonpropositional. Psychoanalytic theory, most markedly in its Kleinian development but also in Freud, appears to require the existence of such content, in view of what it claims about the mental life of infants and primary process thinking. The opposition here is between those who wish to take such talk at face value, as referring to mental states with no less reality than the beliefs and desires attributed in commonsense psychology, and those who regard it as violating *a priori* conditions on the possession of mental states, and are consequently obliged to either reject or offer nonrealistic readings of psychoanalytic accounts of the infant's mind and primary process (Cavell 1993). The prominence of this issue in contemporary discussion of psychoanalysis reflects the continuing shift in twentieth-century philosophy, also evidenced in Habermas's and Lacan's reconstructions of psychoanalysis, away from Cartesianism towards a consideration of mental states in terms of their conditions of attribution and thus in relation to language.

A further set of questions regarding the nature of psychoanalysis is raised by the recent ascent of cognitive science (see *Cognitive Science: Philosophical Aspects*). Freud's unpublished *Project* allows itself to be recast in contemporary, computational terms, and it has been argued that the same sort of reconstruction should be applied to Freud's properly psychoanalytic theories. This would open up the possibility of integrating psychoanalytic theory with the models and hypotheses of present-day cognitive psychology, and further theoretical integration with evolutionary psychology may also be envisaged.

This raises the conceptual question whether psychoanalysis is essentially a subpersonal psychological theory, or whether it is as much a theory of the fully personal level as commonsense psychology. The former is necessary if there is to be direct theoretical integration of psychoanalysis with cognitive psychology; if psychoanalysis is a personal-level psychology, then its relation to cognitive psychology will prove as problematic as that of commonsense psychology. Examination of Freud's writings shows that the question is hard to answer. On the one hand, a move in the direction of subpersonality would appear to be implied in the explanatory shift away from consciousness, and to be required if psychoanalysis is to succeed in offering explanations where commonsense psychology gives out. On the other hand, to regard unconscious motivation and content as on a par with involuntary bodily processes, and as no more intrinsically tied to the self or subjectivity than a neural event, would appear to reflect a deep misunderstanding of the nature and purpose of psychoanalytic explanation. Plausibly, psychoanalysis needs

to be read as containing both personal and sub-personal elements, but to say this is, of course, still to leave the matter unclear.

Questions of possible theoretical integration aside, cognitive psychology poses an epistemological challenge to psychoanalysis. If cognitive psychology should prove capable of explaining the same set of explananda as psychoanalysis, and if its explanations are not isomorphic with those of psychoanalysis, then psychoanalytic theory would lose its claim to explanatory value: Freudian theory would simply have been supplanted, and the commonsense-extension theorist's defense of psychoanalysis would collapse. In such circumstances the only prospect of legitimating psychoanalytic theory would lie in allying the theory with some anti-naturalistic philosophical outlook. Critics of psychoanalysis who share Popper's misgivings about psychoanalysis' scientificity, and those who believe that Freud remained too close to commonsense psychology and failed to carry through a sufficiently strict program of scientific naturalism in psychology, regard this *dénouement* as the most likely.

As this section has shown, discussion of psychoanalysis is inseparable from a multitude of broad philosophical issues: it cannot be maintained that psychoanalysis is unproblematically grounded on experience, nor that there is any one self-evidently correct reading of the conceptual structure of psychoanalytic theory. No other major psychological theory finds itself in the same position. Consequently, even if it is agreed that psychoanalysis should be removed from the context of the philosophy of science, no determinate account of the nature or epistemology of psychoanalytic theory can be arrived at without importing general philosophical commitments. The inevitable result is that the estimation of psychoanalytic theory is embroiled with a number of fundamental philosophical antinomies, above all the conflict between scientific naturalism and anti-naturalism in philosophy and the human sciences.

3. Post-Freudian Developments in Psychoanalytic Theory

The vast range of developments of Freud's ideas may be roughly distinguished into those that attempt an explicit synthesis of psychoanalysis with some other body of theory in the human sciences and those that seek to remain within the terms of Freud's specifically psychological program, aiming either to carry forward Freud's own theories, or to formulate rival, non-Freudian forms of psychoanalysis.

The former will be surveyed in the next section. Post-Freudian developments of the latter kind (see Eagle 1984, Frosh 1987) have in some cases been driven by purely theoretical considerations, reflecting particular views of the relation of the clinical to the theoretical components in psychoanalysis, of the

relation of psychoanalytic theory to the natural sciences, and of the epistemology of psychoanalysis. Instances include the attempts of American theorists such as Merton Gill, Karl Pribram, Emmanuel Peterfreund, David Rapaport, and B. B. Rubinstein to reconstruct Freud's metapsychology in the terms of the natural sciences or to discard it altogether; and, at the opposite end of the spectrum, the endeavors of Jean-Paul Sartre and Ludwig Binswanger to re-express what they take to be Freud's central insights in a philosophical idiom wholly purified of naturalistic elements, 'existential psychoanalysis.'

The more fruitful and influential post-Freudian developments in psychoanalytic theory, however, have taken place in close association with clinical work, and thus engage with Freud's empirically significant, low-level hypotheses, as much as with his metapsychology. The earliest instances are provided by Freud's dissenting contemporaries (see Brown 1961). Jung's 'analytical psychology' disputes Freud's view of the importance of sexuality, and more fundamentally asserts the direct and determining participation in psychic life of unconscious mental contents with a wholly impersonal content and origin, psychological 'archetypes' composing a 'collective unconscious' (see Nagy 1991). Again, Adler's monistic interpretation of human behavior in terms of responses to feelings of inferiority, and Otto Rank's theory of the trauma of birth as the source of mental disorder compete directly with Freud's view of the determinants of mental life.

Only Jung succeeded in founding a school of psychoanalysis that is fully independent of Freud; the ideas of Adler, Rank, and other early figures in the psychoanalytic movement, such as Sandor Ferenczi, were instead either discarded or incorporated into later developments in Freudian theory. The most presently active traditions in psychoanalytic thought are those of Lacan, Klein, and the school of ego psychology which descends from Anna Freud. These schools are geographically anchored in, respectively, France, Britain, and the United States.

As indicated above, the tendency in Lacan's reading of Freud is to raise psychoanalysis to the level of a philosophical discourse, an approach which—encouraged also by the work of Jacques Derrida—has been endorsed by many French philosophers (see, e.g., Henry 1985) and literary theorists who have made use of Lacanian ideas. While some elements in Lacan, such as his claim that the unconscious is structured like a language, may be regarded as reconstructive, other parts of Lacanian theory, such as his account of the infant as acquiring its ego through imaginary identification with a mirror image, offer novel extensions to Freudian theory. Other theses of Lacan's, such as his conception of psychoanalytic treatment as directed towards the acceptance of a fluid self-identity rather than firmly integrated ego, and his view of human discontent as deriving from the metaphysics of

subjectivity rather than the hedonic costs of renouncing instinctual gratification, seem indisputably revisionary.

The Kleinian development of Freud, referred to also as the British school of object-relations, rests on an innovation in clinical practice. The psychoanalysis of children undertaken by Klein greatly expanded the body of data available to psychoanalytic theory. On the basis of the new evidence concerning the mental life of very young children afforded by her employment of the play technique, Klein hypothesized that relations to external objects are standardly—in childhood and throughout adult life—facilitated and mediated by representations of objects in fantasy. On that basis Klein developed a synchronic and diachronic model of the mind as incorporating a fantasized ‘inner world’ composed of ‘internal figures’ and exhibiting certain key structures (called by Klein the ‘depressive’ and ‘paranoid-schizoid positions’) and patterns of activity (splitting, projective identification) determining crucial aspects of the subject’s relation to objects in reality (see Segal 1989, Hinshelwood 1990, and *Object Recognition: Theories*). Some aspects of Kleinian theory remain highly controversial—particularly Klein’s affirmation of Freud’s concept of the death instinct, and the richness of the mental life that her account attributes to the child in its first few months of life—but it has been vigorously extended in the post-Kleinian work of Wilfred Bion, Donald Winnicott, Hannah Segal, and others, especially with reference to psychosis (an area largely neglected by Freud) and symbolism (see Anderson 1992).

Ego psychology developed out of work by Anna Freud, whose *Das Ich und die Abwehrmechanismen* [*The Ego and the Mechanisms of Defence*] (1936) sought to shift the emphasis of psychoanalytic interpretation away from the contents of the id toward the modes of operation employed by the ego in its task of mediating the relations of drives to external reality on the grounds that only in this way can the specificity of an individual’s psychology be grasped and addressed therapeutically. (Anna Freud and her followers clashed with Kleinian members in the course of schismatic ‘Controversial Discussions’ held at the London Institute for Psycho-Analysis in the 1940s.) This theoretical reorientation became more pronounced in the work of Heinz Hartmann in the United States, who produced a systematic account of the ego as comprising a sphere of autonomy and as serving the function of adaptation. Later developments stemming from Hartmann include Margaret Mahler’s work on the infant–mother relation and Heinz Kohut’s ‘self-psychology.’

The diversity of post-Freudian schools is frequently cited as an argument against the cognitive value of psychoanalysis. Whether this implies a lack of convergence of the kind that may properly be taken to impugn the objectivity of a theory is, however, a moot point. Plausibly, the diversity of psychoanalytic theory

derives in large part from their adjustment to different therapeutic techniques—entailing different explanatory foci—and at least some of the disagreement that appears to remain may be attributed to differences of formulation rather than of substance (implicit versions of Kleinian ideas, for example, can be detected readily in Lacan and ego psychology). The degree to which different psychoanalytic theories are intertranslatable, and to which consilience may be discovered among the clinical interpretations of differently schooled analysts, has, however, yet to be determined.

4. *The Relation of Psychoanalytic Theory to other Human Sciences*

Wherever the human sciences conceive their object in intentional or semantic terms—that is, in all but the most austere reductionist contexts—the possibility of establishing a continuity with psychoanalytic theory exists and has been explored to some degree; in most of the social sciences, a variety of interfaces with psychoanalysis have been formulated. Freud’s own writings on culture, religion, and social theory—*Totem und Tabu* [*Totem and Taboo*] (1913), *Massenpsychologie und Ich-Analyse* [*Group Psychology and the Analysis of the Ego*] (1921), *Die Zukunft einer Illusion* [*The Future of an Illusion*] (1927), *Das Unbehagen in der Kultur* [*Civilization and its Discontents*] (1930), *Der Mann Moses und die monoththeistische Religion* [*Moses and Monotheism*] (1939)—provide models for the trans-individual application of psychoanalysis but have not been adopted in mainstream social and cultural thinking (similarly Jung, whose exploration of cultural anthropology was more intensive than Freud’s); probably the element in Freud to have most influenced these spheres has been, closer to the psychology of the individual, his theory of moral development (Deigh 1996).

The most extensive applications and developments of psychoanalytic ideas beyond the psychology of the individual have taken place in contexts where psychoanalysis has promised to fill an explanatory lacuna in some other body of theory and vice versa; here the possibility of a genuine synthesis is created.

Marxism found itself in just such a situation, and elicited highly original developments in the critical theory of the Frankfurt school. A basic consonance of Marxism (regarding its theory of ideology, alienation, and the social superstructure) with psychoanalysis is established by their shared concern with the analysis of forces conditioning and limiting rationality and freedom. The theoretical need of Marxism for psychoanalysis derives from the absence from classical Marxist theory of any account of the subjective process whereby ideology is instantiated and rendered effective at the level of the individual. The corresponding need of psychoanalysis consists in its failure to take account of or make conceptual provision for the culturally

variable mediation of human desire and the permeation of sociohistorical conditions to unconscious mental functioning (Freud's insensitivity to the role of cultural variables was asserted early on by psychoanalytic theorists such as Karen Horney).

In this light, the social holism of Marx and the methodological individualism of Freud could be regarded as complementary rather than opposed, and capable of yielding in conjunction a form of social psychology which would unify the instinctual and cultural, and individual and social, dimensions of human existence. Theories that sought such a unification—specifically, a comprehensive theory of capitalism and of features of modernity such as mass culture and fascism—were advanced by Erich Fromm, Wilhelm Reich, Herbert Marcuse, Max Horkheimer, T. W. Adorno, and others (see Held 1980, Chap. 4). Habermas's reading of Freud also belongs to this tradition.

The project of critical theory has been partially recapitulated in the intense preoccupation of feminist theory with psychoanalysis over recent decades (Mitchell 1975 and Chodorow 1978 are early, seminal works). Here the theoretical alliance has been more ambivalent, however. On the one hand psychoanalysis has seemed to offer feminist theorists a unique and necessary theoretical route to a deep understanding of women's subjectivity; no other psychological theory appears to carry a similar promise of allowing a unified understanding of sexuality, gender, and power relations. On the other hand, psychoanalysis has appeared to be antithetical to the purposes of feminism, on account of its reflections of Freud's own patriarchy and conservatism, and the putative biological determinism and essentialism of Freudian theory, whereby existing social relations and identities are universalized. Resolutions to this dilemma have typically taken the form of a partial historicization of Freudian theory (sometimes appealing to the work of Michel Foucault) and an exploration of post-Freudian (especially Lacanian) theory, aimed at constructing a form of psychoanalysis adequately oriented towards the cognitive-emancipatory interests of women.

See also: Ego Psychology and Psychoanalysis; Feminist Theory: Psychoanalytic; Freud, Sigmund (1856–1939); Klein, Melanie (1882–1960); Psychoanalysis: Adolescence (Clinical–Developmental Approach); Psychoanalysis: Current Status; Psychoanalysis, History of; Psychoanalysis in Clinical Psychology; Psychoanalysis in Sociology; Psychology of the Real Self: Psychoanalytic Perspectives; Transference in Psychoanalysis

Bibliography

Anderson R (ed.) 1992 *Clinical Lectures on Klein and Bion*. Routledge, London

- Assoun P-L 1976 *Freud: la philosophie et les philosophes*. Presses Universitaires de France, Paris
- Brown J 1961 *Freud and the Post-Freudians*. Penguin, Harmondsworth, UK
- Cavell M 1993 *The Psychoanalytic Mind: From Freud to Philosophy*. Harvard University Press, Cambridge, MA
- Chodorow N 1978 *The Reproduction of Mothering: Psychoanalysis and the Sociology of Gender*. University of California Press, Berkeley, CA
- Clark P, Wright C (eds.) 1988 *Mind, Psychoanalysis and Science*. Blackwell, Oxford, UK
- Decker H 1977 *Freud in Germany: Revolution and Reaction in Science, 1893–1907*. International Universities Press, New York
- Deigh J 1996 *The Sources of Moral Agency: Essays in Moral Psychology and Freudian Theory*. Cambridge University Press, Cambridge, UK
- Eagle M 1984 *Recent Developments in Psychoanalysis: A Critical Evaluation*. Harvard University Press, Cambridge, MA
- Edelson M 1984 *Hypothesis and Evidence in Psychoanalysis*. University of Chicago Press, Chicago
- Ellenberger H 1970 *The Discovery of the Unconscious: The History and Evolution of Dynamic Psychiatry*. Basic Books, New York
- Evans D 1996 *An Introductory Dictionary of Lacanian Psychoanalysis*. Routledge, London
- Eysenck H, Wilson G (eds.) 1973 *The Experimental Study of Freudian Theories*. Methuen, London
- Frosh S 1987 *The Politics of Psychoanalysis: An Introduction to Freudian and Post-Freudian Theory*. Macmillan, London
- Gay P 1988 *Freud: A Life for Our Time*. Dent, London
- Grünbaum A 1984 *The Foundations of Psychoanalysis*. University of California Press, Berkeley, CA
- Habermas J 1968 *Erkenntnis und Interesse*. Suhrkamp Verlag, Frankfurt am Main [1971 *Knowledge and Human Interests*. Beacon, Boston]
- Held D 1980 *Introduction to Critical Theory: Horkheimer to Habermas*. University of California Press, Berkeley, CA
- Henry M 1985 *Généalogie de la psychanalyse: le commencement perdu*. Presses Universitaires de France, Paris [1993, *The Genealogy of Psychoanalysis*. Stanford University Press, Stanford, CA]
- Hinshelwood R 1990 *A Dictionary of Kleinian Thought*. Free Association Press, London
- Hook S (ed.) 1964 *Psychoanalysis, Scientific Method and Philosophy*. New York University Press, New York
- Hopkins J 1992 *Psychoanalysis, interpretation, and science*. In: Hopkins J, Savile A (eds.) *Psychoanalysis, Mind and Art: Perspectives on Richard Wollheim*. Blackwell, Oxford, UK
- Jones E 1953–57 *Sigmund Freud: Life and Work*. Hogarth Press, London, 3 Vols.
- Kitcher P 1992 *Freud's Dream: A Complete Interdisciplinary Science of Mind*. MIT Press, Cambridge, MA
- Laplanche J, Pontalis J B 1967 *Vocabulaire de la psychanalyse*. Presses Universitaires de France, Paris [1973 *The Language of Psychoanalysis*. Methuen, London]
- Levine M (ed.) 1999 *Analytic Freud: Philosophy and Psychoanalysis*. Routledge, London
- Mitchell J 1975 *Psychoanalysis and Feminism*. Penguin, Harmondsworth, UK
- Nagy M 1991 *Philosophical Issues in the Psychology of C. G. Jung*. State University of New York, Albany, NY
- Neu J (ed.) 1991 *The Cambridge Companion to Freud*. Cambridge University Press, Cambridge, UK

- Popper K 1963 *Conjectures and Refutations: The Growth of Scientific Knowledge*. Routledge & Kegan Paul, London
- Ricœur P 1965 *De l'interprétation: essai sur Freud*. Éditions du Seuil, Paris [1970 *Freud and Philosophy: An Essay in Interpretation*. Yale University Press, New Haven, CT]
- Segal H 1989 *Klein*. Fontana, London
- Sulloway F 1979 *Freud: Biologist of the Mind*. Basic Books, New York
- Whyte L 1979 *The Unconscious Before Freud*. Friedman, London
- Wittgenstein L 1978 *Lectures and Conversations on Aesthetics, Psychology and Religious Belief*. Blackwell, Oxford, UK
- Wollheim R (ed.) 1974 *Freud: A Collection of Critical Essays*. Anchor Doubleday, New York [Reprinted 1977 as *Philosophers on Freud: New Evaluations*. Aronson, New York]
- Wollheim R 1991 *Freud*, 2nd edn. Fontana Collins, London
- Wollheim R, Hopkins J (eds.) 1982 *Philosophical Essays on Freud*. Cambridge University Press, Cambridge, UK

S. Gardner

Psychobiology of Stress and Early Child Development

Stressors that threaten well-being activate physiological systems that, while fostering immediate survival, can interfere with repair of the body, defense against illness, and development. Two systems play major roles in stress: the sympathetic adrenomedullary (SAM) system that produces adrenaline and other catecholamines, and the limbic-hypothalamic-pituitary-adrenocortical (LHPA) system that produces the steroid hormone cortisol in humans. Of these two systems, the LHPA system exerts slower, more persistent actions, and is often the focus of research on stress physiology. A growing literature indicates that conditions associated with stress, maltreatment and neglect during fetal and early development alter regulation of the LHPA system in ways that could heighten vulnerability later to stress-related disorders. During early life, individual differences in temperament interact with qualities of the caregiving environment to predict cortisol responses to psychosocial stressors. Combined, these growing literatures on normal and atypical development promise to explicate mechanisms through which early experiences affect physical and psychological health.

1. Psychobiology of the LHPA System

Animal studies constitute the bulk of research on this system. Determining their implications for human development is a major challenge. Current perspectives drawn from the animal research will be presented in this section.

Cortisol is the end product of a cascade of events that begins with stimulation of cells in an area of the brain called the hypothalamus. These cells secrete

peptide hormones (CRH and AVP) that, in turn, cause cells in the pituitary gland (once thought of as the master endocrine gland) to produce yet another biochemical, ACTH. ACTH is secreted into the bloodstream where it is picked up by receptors on the cortex of the adrenal glands, stimulating the adrenal to secrete cortisol (Johnson et al. 1992). Humans produce cortisol even when they are not stressed. Nonstressed or basal cortisol production follows a daily rhythm. Like a thermostat, basal levels are maintained through negative feedback loops; the primary feedback sites are located in two areas of the brain, called the hypothalamus and hippocampus. Through multiple pathways in the brain, internal and external threats to well-being cause increased activity in this LHPA system that raise cortisol levels over baseline (de Kloet 1991). Responses to psychological stressors operate through pathways involving an area of the brain called the amygdala that also plays important roles in organizing behavioral reactions to threat. CRH, the peptide that begins the cascade of events that stimulates cortisol production, is produced in many brain areas, including sites such as the amygdala, where it is believed to play a major role in regulating threat reactive behavior and in stimulating both the LHPA and the SAM systems to react to psychological stressors. One major theory relating stress to depression focuses on repeated stressors and the orchestrating role of extra-hypothalamic CRH (Nemeroff 1996). The cumulative impact of stressors on health has been described as allostatic load, or the health-impairing costs of survival-facilitating activity of the LHPA system (McEwen 1998).

Understanding how the LHPA system promotes both health and disease is a core problem in stress research. Both hypo- and hyper-responsivity of the LHPA system are associated with disorder (de Kloet et al. 1998). Disorders of hypo-responsivity include those related to failure to contain immune inflammatory responses and to maintain the responsivity of areas of the brain involved in emotions and information processing; hyper-responsivity disorders include those related to immune suppression, cardiovascular disease, and affective psychopathology. One line of reasoning traces these contrasting impacts through the two types of receptors for cortisol that mediate two modes of the hormone's effects (de Kloet et al. 1998). The first, a 'proactive' or 'permissive' mode, maintains basal activity of the system, regulates the threshold of the system's response to stressors, promotes coordination of the daily rhythm in the hormone, and helps maintain the viability of brain systems involved in selective attention, sensory integration, and response selection. Low levels of circulating cortisol affect this mode through stimulating MR or Type I receptors in the central nervous system. The second, a 'reactive' or 'suppressive' mode, operates through GR or Type II receptors that mediate highly catabolic events promoting increases in energy availability, modulation

of the immune system, and threat-related learning. GRs are occupied increasingly as cortisol rises above basal levels into stress ranges. Both proactive and reactive modes support health and survival. However, the reactive mode, if sustained, threatens health. In the hypothalamus and hippocampus, occupation of GR or Type II receptors also terminates the stress response, serving as a brake on the duration of the reactive/suppressive stress mode. In animals, early experiences appear to alter the balance of MRs/GRs, lowering the level of hormone at which the system shifts from proactive to reactive functioning and allowing the reactive, suppressive mode to continue for longer periods before cortisol returns to basal levels.

A recent line of developmentally relevant research focuses on mechanisms relating the LHPA system to systems involved in growth and development of the brain and body. Of particular note are studies examining the interaction of glucocorticoids, MR/GR balance and serotonin (5-HT) receptor activity in the hippocampus (López et al. 1998), and studies examining the interaction of glucocorticoids, growth hormone, and insulin growth factors (e.g., IGF-1) (Johnson et al. 1992). In animals, conditions that elevate glucocorticoids chronically also increase basal glucocorticoid levels, affect MR/GR ratios, and lower serotonin activity in the hippocampus. These conditions include stress during pre- and postnatal development. These changes are associated with heightened vulnerability to stressors later in life, as well as problems with behavior regulation, response inhibition, and sensitivity to self-administration of drugs of abuse (de Kloet et al. 1998). Critically, again in animals, conditions associated with responsive caregiving during postnatal development reduce the impact of prenatal stress and buffer the impact of postnatal stressors (Caldji et al. 1998, Levine 1994).

2. Studies in Infants and Young Children

2.1 Measurement Issues

Research in children has been stimulated by the development of cortisol assays that use samples of free-flowing saliva (Kirschbaum and Hellhammer 1994). Unfortunately, salivary sampling does not allow the assessment of activity higher in the system (i.e., at the pituitary and brain levels). Blood sampling is needed to assess ACTH concentrations, and a spinal tap is typically needed to measure CRH. Particular drugs are usually given to probe the physiology of the LHPA system in ways that bypass individual differences in how the person perceives and interprets events. Because of their invasiveness, none of these techniques is used typically with normally developing children.

In addition, the researcher who studies the LHPA system must attend to many confounding variables that constrain how studies can be conducted. Because of the daily rhythm in the hormones, all samples must be conducted at the same time of day. Careful attention needs to be given to the child's health, what he or she has eaten and drunk recently, and the familiarity of the context (Kirschbaum and Hellhammer 1994). These constraints have limited the study of this system in children. Despite this, a literature is developing which promises to inform our understanding of early risks for later physical and emotional pathology.

2.2 Developmental Changes in Responsivity

Prenatally, the LHPA system produces cortisol by the beginning of the second trimester. In animals, the mother's cortisol affects the fetus, resulting in the development of fewer MRs and larger LHPA responses to stressors (Barbazanges et al. 1996). In humans, fetal vulnerability (low birthweight, prematurity) increases with increased maternal psychosocial stress and LHPA activity (Sandman et al. 1997).

The full-term neonate exhibits robust, graded elevations in cortisol that sensitize to repeated pain-eliciting events (e.g., heel lances) and habituates to repeated handling stressors (e.g., physical examinations) (Gunnar 1992). Responsiveness to handling stressors appears to decrease at around 3 months of age, after which stressors such as physical examinations fail to provoke elevations in cortisol for most infants. Another marked decrease in cortisol responsivity is observed between 6 and 12 months of age, after which during infancy it becomes difficult to provoke elevations in cortisol, on average, to many mild stressors (Gunnar 2000).

2.3 Relationships as Regulators of the HPA system

By 6–12 months, social interactions with caregivers mediate cortisol responses. Thus, brief maternal separations fail to elevate cortisol if substitute caregivers are sensitive and responsive, while elevations are noted when substitute caregivers are not (Gunnar 2000). Similarly, the availability of the parent in secure attachment relationships buffers elevations in cortisol to events that elicit wariness and behavioral distress, while elevations are noted for toddlers exhibiting similar wariness/distress if the attachment relationship is insecure (Spangler and Schieche 1998). Elevations in cortisol have also been noted for toddlers with the type of disorganized/disoriented attachment more typically noted among infants of depressed and/or abusive parents. Higher home baseline levels of cortisol have also been reported among 3-year-olds whose

mothers were clinically depressed during the child's first year of life. These data are consistent with the literature in animals cited above relating maternal care to stress reactivity and regulation.

Social relations with peers are potent stimuli of the LHPA system in human adults and animals (Kirschbaum and Hellhammer 1994). For children, this appears true at least as early as the preschool years. For toddlers and preschoolers, childcare involving long hours of interaction with like-aged playmates produces elevations in cortisol for many children (Dettling et al. 1999). Even short periods of interaction in half-day preschool programs may elevate cortisol for some children. Children who show activation of the LHPA system in peer settings are those with problematic peer relations; peer rejection, low dominance status, and social withdrawal tend to characterize these children. Problematic styles of interacting (e.g., anger, aggression, and poor self-control) that foster peer rejection also correlate with higher and more responsive patterns of cortisol production in preschool peer groups (Gunnar et al. 1997).

3. Temperament

Both extreme shyness and problems in effortful control have been associated with LHPA system activity in young children. Extreme shyness has been hypothesized to reflect altered thresholds for activation of stress physiology involving central fear circuits in the brain (i.e., amygdala-bed nucleus pathways). Thus, increased responsiveness of the LHPA system has long been hypothesized for extremely shy, socially anxious children (Kagan et al. 1987). Strong evidence for this hypothesis is largely lacking when children are assessed under conditions of social challenge (Gunnar et al. 1997); however, shyness, whether measured by teacher report or direct observation, has been found to correlate positively with cortisol sampled at home (de Haan et al. 1998). It is not clear why cortisol correlates more readily with shyness when home hormone concentrations are examined. It may be that under conditions of challenge, moderating factors such as coping strategies and social support need to be taken into account, while home levels are more reflective of tonic influences.

Rothbart has identified effortful control as a higher-order dimension of temperament that emerges from factor analyses of temperament questionnaires during the preschool years. Attention focusing and the capacity to inhibit prepotent responses form the core of this dimension, which Posner and Rothbart (1994) argue reflects the development of an attentional network involving the anterior cingulate gyrus which in animals also plays a role in containing the cortisol stress response (Diorio et al. 1993). In children, effortful control modifies relations between negative emotionality and behavior (Eisenberg et al. 1994). In

preschool-aged children, parent and teacher reports of effortful control tend to be negatively correlated with cortisol levels and measures of cortisol responsivity in classroom settings (Gunnar et al. 1997). Whether these correlations reflect mechanisms linking the LHPA system to the neurobiology of effortful control or the impact of effortful control on the behaviors that create stressful social encounters has not been determined.

4. Neglect, Maltreatment, and Trauma

Based on the animal research, early neglect and abuse should alter the activity of the developing LHPA system. So far, there is only a small literature on this in humans. Because of the multiple genetic and experiential risk factors typically noted in children with maltreating backgrounds, isolating the impact of adverse early experiences on the human LHPA system promises to be difficult. A full discussion of this emerging literature is beyond the scope of this article. Two findings related to physical and sexual abuse seem consistent enough to mention. First, during periods of active abuse young children do exhibit elevated cortisol levels and heightened responsivity to CRH challenge suggestive of a hyper-responsive LHPA system. Once the child's life circumstances are stabilized, however, for most children this hyper-responsiveness diminishes. The exceptions may be (a) depressed maltreated children who show a flattening of the daily rhythm in cortisol production (Hart et al. 1996), and (b) maltreated children with chronic post-traumatic stress disorder (PTSD) and who have higher baseline cortisol and catecholamine levels that are correlated with the duration of abuse (De Bellis et al. 1999).

In addition, the relations between growth failure, maltreatment, and LHPA system bear comment. Physical abuse and neglect sometimes produce significant growth delays due to reduced pituitary production of growth hormone (GH), cell insensitivity to GH, and blunted GH responses to provocative challenges, alterations that all revert to normal when the child is removed from the maltreating circumstances (Alanese et al. 1994). Theoretically, the LHPA system may be involved in the etiology of this syndrome, called variously psychosocial short stature or psychosocial dwarfism (Johnson et al. 1992). Children adopted from deprived, orphanage conditions exhibit growth failure, and six years after adoption have elevated baseline cortisol levels that were correlated with length of orphanage experience (Gunnar 2000).

5. Future of Research

Psychobiological studies of stress in young children have burgeoned since the development of techniques to measure cortisol noninvasively in saliva. Com-

munity studies of children negotiating the normative stressors of childhood are beginning to make connection with studies of children from highly stressful, adverse life circumstances and those with clinical disorders. As in studies of adults and animals, individual differences in response to stressors are common and related to the child's social circumstances and temperament. The challenge for the future is to determine the relation between LHPA function in early development and child outcomes and to explicate the relevant molecular processes.

See also: Attachment Theory: Psychological; Biopsychology and Health; Shyness and Behavioral Inhibition; Stress, Neural Basis of; Stress: Psychological Perspectives; Temperament and Human Development

Bibliography

- Albanese A, Hamill G, Jones J, Skuse D, Matthews D R, Stanhope R 1994 Reversibility of physiological growth hormone secretion in children with psychosocial dwarfism. *Clinical Endocrinology* **40**: 687–92
- Barbazanges A, Piazza P V, Le Moal M, Maccari S 1996 Maternal glucocorticoid secretion mediates long-term effects of prenatal stress. *Journal of Neuroscience* **16**: 3943–9
- Caldji C, Tannenbaum B, Sharma S, Francis D, Plotsky P M, Meaney M J 1998 Maternal care during infancy regulates the development of neural systems mediating the expression of fearfulness in the rat. *Proceedings of the National Academy of Sciences of the United States of America* **95**: 5335–40
- De Bellis M D, Baum A S, Birmaher B, Keshavan M S, Eccard C H, Boring A M, Jenkins F J, Ryan N D 1999 A. E. Bennett Research Award. Developmental traumatology. Part 1. Biological stress systems. *Biological Psychiatry* **45**: 1259–70
- de Haan M, Gunnar M R, Tout K, Hart J, Stansbury K 1998 Familiar and novel contexts yield different associations between cortisol and behavior among 2-year-olds. *Developmental Psychobiology* **33**: 93–101
- de Kloet E R 1991 Brain corticosteroid receptor balance and homeostatic control. *Frontiers in Neuroendocrinology* **12**: 95–164
- de Kloet R, Vreugdenhil E, Oitzl M S, Joels A 1998 Brain corticosteroid receptor balance in health and disease. *Endocrine Reviews* **19**: 269–301
- Dettling A C, Gunnar M R, Donzella B 1999 Cortisol levels of young children in full day childcare centers: Relations with age and temperament. *Psychoneuroendocrinology* **24**: 519–36
- Diorio D, Viau V, Meaney M J 1993 The role of the medial prefrontal cortex (cingulate gyrus) in the regulation of hypothalamic–pituitary–adrenal responses to stress. *Journal of Neuroscience* **13**: 3839–47
- Eisenberg N, Fabes R A, Nyman M, Bernzweig J, Pinuelas A 1994 The relations of emotionality and regulation to children's anger-related reactions. *Child Development* **65**: 109–28
- Gunnar M R 1992 Reactivity of the hypothalamic–pituitary–adrenocortical system to stressors in normal infants and children. *Pediatrics (suppl.)* **90**: 491–7
- Gunnar M R 2000 Early adversity and the development of stress reactivity and regulation. In: Nelson C A (ed.) *The Effects of Adversity on Neurobehavioral Development. The Minnesota Symposia on Child Psychology*. Lawrence Erlbaum Associates, Mahwah, NJ, Vol. 31, pp. 163–200
- Gunnar M R, Tout K, de Haan M, Pierce S, Stansbury K 1997 Temperament, social competence, and adrenocortical activity in preschoolers. *Developmental Psychobiology* **31**: 65–85
- Hart J, Gunnar M R, Cicchetti D 1996 Altered neuroendocrine activity in maltreated children related to symptoms of depression. *Developmental Psychobiology* **8**: 201–14
- Johnson E O, Kamilaris T C, Chrousos G P, Gold P W 1992 Mechanisms of stress: A dynamic overview of hormonal and behavioral homeostasis. *Neuroscience and Biobehavioral Reviews* **16**: 115–30
- Kagan J, Reznick J S, Snidman N 1987 The physiology and psychology of behavioral inhibition in children. *Child Development* **58**: 1459–73
- Kirschbaum C, Hellhammer D H 1994 Salivary cortisol in psychoneuroendocrine research: Recent developments and applications. *Psychoneuroendocrinology* **19**: 313–33
- Levine S 1994 The ontogeny of the hypothalamic–pituitary–adrenal axis: The influence of maternal factors. *Annals of the New York Academy of Sciences* **746**: 275–88
- López J F, Chalmers D T, Little K Y, Watson S J 1998 A. E. Bennett Research Award. Regulation of serotonin 1A, glucocorticoid, and mineralocorticoid receptor in rat and human hippocampus: Implications for the neurobiology of depression. *Biological Psychiatry* **43**: 547–73
- McEwen B S 1998 Stress, adaptation, and disease: Allostasis and allostatic load. *Annals of the New York Academy of Sciences* **840**: 33–44
- Nemeroff C B 1996 The corticotropin-releasing factor (CRF) hypothesis of depression: New findings and new directions. *Molecular Psychiatry* **1**: 336–42
- Posner M I, Rothbart M K 1994 Attentional regulation: From mechanism to culture. In: Bertelson P, Elen P, d'Ydewalle G (eds.) *International Perspectives on Psychological Science*. Lawrence Erlbaum Associates, Mahwah, NJ, Vol. 1, pp. 41–54
- Sandman C A, Wadwa P D, Chicx-DeMet A, Dunkel-Schetter C, Porto M 1997 Maternal stress, HPA activity, and fetal-infant-outcome. *Annals of the New York Academy of Sciences* **814**: 266–75
- Spangler G, Schieche M 1998 Emotional and adrenocortical responses of infants to the strange situation: The differential function of emotion expression. *International Journal of Behavioral Development* **22**(4): 681–706

M. R. Gunnar

Psychohistory

1. Its Meaning

The term *psychohistory* (originally written *psychohistory*) emerged in America in the 1950s to designate historical writings utilizing psychology, for the most part depth psychology, as their primary frame of

reference or mode of explanation. Two of its tenets in particular distinguish psychohistory sharply from other systematic approaches to the human past. The first of the two, which is rarely spelled out even in methodological writings, is that the facts of human history consist entirely of what people did whether spontaneously or under pressure of circumstances. To the psychohistorian, an intellectual revolution thus signifies thinkers changing their minds; a decline in fertility translates as women bearing fewer children; urbanization conveys people living increasingly in cities. Other historians may agree that even such seemingly impersonal historic developments all come down to human doings, yet whenever other historians narrate or explain developments of the sort they tend in the first instance to see belief systems or birth rates or urban:rural ratios changing rather than people thinking or acting differently. The other highly distinctive tenet of psychohistory is, by contrast, fully and even emphatically explicit in the psychohistorical literature: that the determinants of human conduct, whether individual or collective, are largely unconscious. Historians of other schools will often freely acknowledge deeply subjective causation of historic human behavior only to shy away from it in their own research as too dauntingly difficult to reconstruct, let alone demonstrate. To the psychohistorian, however, whatever the difficulties involved, there is simply no other way for history to be understood (see *History: Theories and Methods*).

2. Its Development

Psychohistory originated as applied psychoanalysis, with Freud and his disciples drawing on their clinical findings now and again to illuminate historical facts (see *Psychoanalysis: Overview*). At first, those clinical findings bore heavily on the patterns and pitfalls of individual emotional development, with a strong emphasis on pregenital sexuality as it emerges in childhood bit by bit and then undergoes repression even while continuing to inform emotional life. Accordingly, the earliest applications of psychoanalysis to history—including those by Freud himself, who long set the standard for his followers—were biographic and focused on the infantile concerns latent in the words and deeds of eminent or paradigmatic historic figures (see *Biography: Historical*). A landmark work in this category was psychoanalyst Erik H. Erikson's *Young Man Luther* of 1958, which argued that Luther's bold quest for spiritual autonomy was that of his whole age writ large (see *Erikson, Erik Homburger (1902–94)*). Unlike biography in relation to historiography as a whole, psychobiography remains to this day the basic as well as the leading form of psychohistory. At the same time, it was as quick as psychoanalysis itself to move beyond its initial Freudian orthodoxy (see *Freud, Sigmund (1856–1939)*).

Post-Freudian concepts utilized in psychobiography when this is practiced as applied psychoanalysis have notably included C. G. Jung's 'individuation,' Alfred Adler's 'inferiority complex,' Erik H. Erikson's 'identity crisis,' Heinz Kohut's 'narcissistic rage,' and Jacques Lacan's 'language of the id.'

Psychohistory long continued to show its derivation from psychoanalysis even where it broadened its scope beyond psychobiography. As late as the 1970s psychohistorians who explored social groups, whole cultures, or long-range collective developments, did so as a rule only in terms of the individual psychodynamics thought to have determined them, tracing their salient characteristics to formative experiences widely shared by the individuals involved, and most often to pathogenic childrearing practices on the order of abrupt weaning or intrusive parenting. Correspondingly, the preferred subjects under the head of group behavior were instances of social pathology such as witchcraft, or of political pathology such as Nazism. The most systematic psychohistorical approach to group behavior to emerge on these individualistic premises is prosopography duly slanted: the profiling of the likes of saints or witches or Nazis in terms of variables of psychoanalytical significance in their individual family histories—ages of parents and siblings, for instance, or again early parent or sibling losses (see *Prosopography (Collective Biography)*).

A prominent early alternative to the individual personality as the basic unit even of large-scale, broad-gauged psychohistorical accounts or analyses was the study of a mass leader whose following interacted with him like a single composite being. Freud himself pioneered this departure from the classic individualistic model that underlay his own science of psychoanalysis. In his *Group Psychology and the Analysis of the Ego (Massenpsychologie und Ichanalyse)* of 1922 he treated mass phenomena for the most part as simultaneous, parallel regression by individuals, but in a couple of highly suggestive passages he speculated on human consciousness as originally shared by a Darwinian primal horde from which a leader eventually emerged with a mind and will of his own; such shared group consciousness, he considered, might revive at critical historic junctures under the spell of a charismatic leader who threw back to that aboriginal tribal chief. Freud (1939) worked this speculative construction into a highly imaginative reading of the Moses story, with Moses reincarnating the tribal father supposedly slain and later exalted by his sons acting in concert. Given the very nature of this concept of group consciousness as a reversion to a tribal leader and following, subsequent psychohistorical treatments of leader and led have mostly dealt with political collectivities such as Lincoln's or Nixon's America, or again William II's and especially Hitler's Germany.

On the Freudian model of group process, then, group identity constitutes a regressive throwback to an original shared consciousness in the face of a strong

leader. But a rival conception of shared mental life, one that makes it out as permanent and leaderless, came into psychohistory independently. Its *locus classicus* lies outside of psychohistory: in the 'collective consciousness' (*conscience collective*) that Emile Durkheim in his *Suicide* of 1897 inferred from the patterned regularity of suicide rates over and beyond the particular subjective causation of each individual suicide. An early specimen of such a leaderless group model in psychohistory was Caroline E. Playne's *The Neuroses of the Nations* of 1925, which owed nothing to psychoanalysis and indeed wholly ignored it. When William L. Langer, in a resounding presidential address to the American Historical Association in 1957, called the mastery and use of depth psychology the historian's 'next assignment,' he expressly meant its application to leaderless group phenomena, which he exemplified by the epidemic of religious and cultural pathology that followed in the wake of the Black Death of 1348. (Oddly, Langer made no reference to a pioneering work in this mold just then published by his Harvard University colleague Walter Abell (1957) about how the waxing and waning insecurities of medieval European life were projected into Europe's public art of the time.) By the 1970s the dynamics of 'group process,' as it came to be called, topped the psychohistorical agenda, with collective fantasy as its privileged medium of analysis. An exemplary early product of this approach to group process was Paul Monaco's (1976) work, which teased the deep-lying national anxieties of France and Germany in the 1920s out of the recurrent motifs and techniques of each country's most popular movies of those years. Another choice area, still under exploited, for ferreting out the unconscious underpinnings of group activity is historical demography (see *Historical Demography*), especially the adjustments of aggregate birth rates to the felt needs of a population with regard to its overall numbers.

As the term 'group process' itself implies, explanation in psychohistory gradually has shifted from unconscious motives and meanings to unconscious mechanisms governing thought and conduct, while conscious purposes have been even more marginalized in the process. This shift reflects the continuing self-distancing of psychohistory from its parent discipline of psychoanalysis, focused as the latter is on latent impulses originating in childhood. Nowhere is this shift more evident than in the concept of traumatic reliving, which matured within psychohistory well before it was taken up in psychoanalysis. Psychoanalysis had first confronted adult psychic trauma during World War One in the form of battle shock for which, as for other emotional ailments, it sought a childhood referent that in this case was not forthcoming. Comparative psychobiography established two patterned, alternative reactions to emotional trauma: either a continual recall accompanied by the full, tormenting original affect, or a repression of the

traumatic affect, and sometimes even of the traumatic memory itself, followed by a contrived, controlled, disguised reexperience or reliving of the trauma to the apparent end, never attained, of mastering the attendant affect that could not be mastered at the time of the experience itself. These same mechanisms have subsequently been shown to operate as well on the collective level in instances of group trauma, and to operate independently of the widely varying reactions of the individuals composing the traumatized group. Precursively, William L. Langer in handing out his 'next assignment' conceived of the after effects of the Black Death, notably its morbid reenactments such as the Dance of Death, as just such a case of group traumatic reliving.

3. *Its Problematic*

Psychohistorical research is uniquely demanding of its practitioners because of its concern with subjective phenomena that are rarely if ever directly recorded. To understand a piece of the human past psychohistorically, researchers must first master the relevant facts at least as thoroughly as the practitioners of any other mode of inquiry. But much more, they must become immersed in those facts, make them vicariously their own, and feel out their inner determinants by dint of this imaginative identification with their original experience. The method is strikingly like the one promoted by Henri Bergson under the name of 'intuition': to study a subject intellectually from the outside until one enters into it subjectively through a core insight around which everything known about it will then fall into place. Such intimate, personal self-projection into a historic subject is treble problematic. For one thing, it presupposes a basic equivalence of mental and emotional constitution as between researchers and their human subjects, however remote those subjects may be from them historically, and even if they are groups rather than individuals like themselves. For another, researchers' identification with their subject must be thoroughgoing, yet unless they can also sustain some critical awareness outside of that identification the insights gained through it will be lost. And for a third, pieces of the researchers' own mental underworld may slip unnoticed into their reading of a subject. This danger, often called 'countertransference' by loose analogy with psychoanalysts projecting into their patients, is the risk most cautioned against by psychohistorians themselves. Actually, an affinity in researchers for their subject's inner experience can, on the positive side, sensitize or alert them to connections and meanings they might otherwise miss. Besides, a purely projective finding will not stand up in any case inasmuch as it will fail to fit all the evidence, let alone satisfy anyone apart from the researchers themselves.

Researchers will feel the rightness of their inner

grasp of a subject when it not only pulls together all the relevant facts at their disposal, but points the way to unsuspected additional corroborative evidence. Yet the discovery of such new evidence, however suggestive, is in itself no proof that their inner grasp of their subject is universally valid. Because of the very nature of psychohistorical explanation, to validate a psychohistorical thesis takes more than just a wealth of corroborative material, even if the thesis itself is what turned up key pieces of that material. To pass muster, a psychohistorical thesis must also carry subjective conviction, beyond the researchers themselves, with others willing and able to put themselves in the subject's place in their turn and see whether the proposed unconscious source of the thought or action at issue checks out subjectively with them too.

This second, subjective criterion of verification means in brief that a psychohistorical thesis must be emotionally as well as intellectually compelling. Consequently it can rely on no pre-given theory for support. Yet however direct its appeal, its need for subjective on top of objective verification demands too much effort, not to add goodwill, of critics accustomed instead to merely material and logical criteria of validation. As a substitute test of validity I therefore propose that, in psychohistory as in any science, a thesis is confirmed if, first of all, the known evidence all runs its way, if moreover it could potentially be refuted by new evidence, and if finally any piece of what it purports to explain cannot very well be explained otherwise. To take a simple example of this last stipulation from my own thesis that Hitler was traumatized by his mother's iodoform poisoning during her terminal cancer treatment: why else did he put it to his generals in late 1944 that his Ardennes offensive had caught the Allies off guard because they believed he was dead already 'or at all odds am down with cancer somewhere and can't live any more or drink any more' except that deadly iodoform poisoning induces a burning thirst together with an inability to drink? (Binion 1976, p. 143).

4. *Its Crisis Years*

As a historical school in its own right, psychohistory flourished above all in the US during the 1960s and 1970s. Even then, like all new departures, it was highly controversial, especially within the historical professions, because of its constitutive tendency to undercut all other modes of historical explanation. It has since suffered considerably from the decline of psychoanalysis, with which it is rightly associated in that the two still share at least the premise of unconscious motives and meanings. Experimental graduate programs in the field have all folded, while course offerings on college campuses have dwindled to the vanishing point. The ranks of researchers in the field are fast

thinning even as the stringent standards of verification invoked for it by critics and adepts alike are discouraging prospective practitioners. Fewer psychohistorical books and articles are appearing year by year. Of the two major specialized journals, *The Psychohistory Review* ceased publication in 1998 while *The Journal of Psychohistory* has lost most of its initial circulation. Perhaps worst of all for the discipline, the consensus of outsiders is that its scholarly output is falling off qualitatively as well.

On the other hand, psychohistory has left its impress on most other kinds of historical inquiry, especially those that emerged or developed later than psychohistory itself. Biographies and even general histories now routinely incorporate psychohistorical materials and interpretations. The study of mentalities often verges on psychohistory. So does that of historic memory, especially in the concern with memory gaps or distortions that has dominated the field. Gender history and gay history, as well as the history of emotions, lean heavily on psychohistorical concepts and findings. Many a term launched in psychohistory, such as Erikson's 'identity crisis,' has become common coinage elsewhere, while the mechanism of traumatic reliving by individuals and groups alike has found its way into a broad variety of scholarly contexts, including even a quantitative history of testation in late medieval Tuscany (Cohn 1992) (see *Identity in Childhood and Adolescence*).

The fact that psychohistory has lost much of its vitality of the 1960s and 1970s, or again that it is being put to subordinate uses by historians of other orientations, does not affect its constitutive claim to provide a privileged access to causality in history. Given this inherently persuasive claim, neither the recent loss of momentum by psychohistory as a method in its own right nor its present auxiliary status in historical research can very well be regarded as permanent.

See also: Emotions, History of; Historiography and Historical Thought: Current Trends; Psychoanalysis, History of

Bibliography

- Abell W 1957 *The Collective Dream in Art: A Psycho-historical Theory of Culture Based on Relations between the Arts, Psychology, and the Social Sciences*. Harvard University Press, Cambridge, MA
- Aldrich R 1993 *The Seduction of the Mediterranean: Writing, Art, and Homosexual Fantasy*. Routledge, London
- Binion R 1976 *Hitler among the Germans*. Elsevier, New York
- Binion R 1994 Fiction as social fantasy: Europe's domestic crisis of 1879–1914. *Journal of Social History* 27: 679–99
- Brink A 1989 *Bertrand Russell, A Psychobiography of a Moralist*. Humanities Press International, Atlantic Highlands, NJ

- Cohn S K Jr. 1992 *The Cult of Remembrance and the Black Death. Six Renaissance Cities in Central Italy*. Johns Hopkins University Press, Baltimore, MD, pp. 116–17, 288
- Demos J P 1970 *A Little Commonwealth: Family Life in Plymouth Colony*. Oxford University Press, New York
- Erikson E H 1959 *Identity and the Life Cycle*. International University Press, New York
- Freud S 1939 *Der Mann Moses und die Monotheistische Religion*. Verlag Albert de Lange, Amsterdam
- Graziano F 1992 *Divine Violence: Spectacle, Psychosexuality, & Radical Christianity in the Argentine 'Dirty War.'* Westview Press, Boulder, CO
- Kohut T 1985 Mirror image of the nation: An investigation of Kaiser Wilhelm II's leadership of the Germans. In: Strozier C B, Ofer D (eds.) *The Leader: Psychohistorical Essays*. Plenum, New York
- Langer W L 1958 The next assignment. *American Historical Review* 63: 283–304
- Mack J E 1977 *A Prince of our Disorder: The Life of T. E. Lawrence*. Weidenfeld and Nicolson, London
- Markowitz J 1969 *The Psychodynamic Evolution of Groups*. 1st edn. Vantage Press, New York
- Monaco P 1976 *Cinema and Society: France and Germany during the Twenties*. Elsevier, New York
- Roper L M 1994 *Oedipus and the Devil: Witchcraft, Sexuality, and Religion in Early Modern Europe*. Routledge, New York
- Saslow J M 1986 *Ganymede in the Renaissance: Homosexuality in Art and Society*. Yale University Press, New Haven, CT

R. Binion

Psycholinguistics: Overview

1. Introduction

Psycholinguistics is a field that combines methods and theories from psychology and linguistics to derive a fuller understanding of human language. From psychology, it inherits experimental methodology and a body of knowledge about processes in perception, memory, attention, learning, and problem-solving. From linguistics, it derives detailed descriptions of specific languages, rigorous accounts of the shape of grammar, and ideas about the nature of human language (Osgood and Sebeok 1967).

The basic issue that motivated the establishment of psycholinguistics as a separate field of study was the problem of the 'psychological reality' of linguistic concepts. For example, speakers of English can form the plurals of nouns by adding the suffix *-s*. This process allows us to form the plural *chandeliers* from *chandelier*. But do we actually use a productive rule to produce this plural or do we simply retrieve the plural *chandeliers* from our long-term memory as a unit? Psycholinguistic research shows that, in fact, both rote and rule are operative at various times in language production (Pinker 1999).

This issue of psychological reality applies to all levels of language structure and usage, including articulatory phonetics, auditory phonetics, phonology, morphology, lexicon, syntax, semantics, and pragmatics. It applies to all of the constructs of linguistics, including rules, grammars, paradigms, trees, segments, words, and morphemes. When we hear words, do we break them up into their component phonemes and distinctive features (see *Phonology*), or do we recognize them as wholes? When we listen to sentences, do we actually construct tree-like representations of the types proposed by linguists or do we somehow extract meaning without building up formal structures? If we do use formal grammars to listen and speak, what is the exact shape of the grammars that we use?

Exploration of the psychological reality of linguistic structures immediately leads us to two related fields of study. The first is developmental psycholinguistics, or the study of child language acquisition (Fletcher and MacWhinney 1995). If we believe that adults form the plurals of nouns by adding *-s*, we need to consider how young children can learn to apply this rule to produce 'cats' and not 'foots' or 'tooths.' If we argue that they simply learn each form by rote, how can we account for the fact that they make errors like 'feets' and the fact that they can produce the plural for a new word like 'wug' even without having been given that form explicitly?

An even more difficult issue involves how rules of grammar might be processed in the brain. Addressing this question has led psycholinguists to explore issues in neurolinguistics (Stemmer and Whitaker 1998) and cognitive neuroscience (Gazzaniga 1997). When the field of psycholinguistics first developed in the 1950s, psychologists knew little about the detailed functioning of the human brain and were forced to treat it as a 'black box.' However, as our understanding of the functioning of the human brain grew during the 1980s and 1990s, it became clear that a precise understanding of the functioning of human language would have to make reference to neural mechanisms. The interactive approach to cognition (McClelland and Rumelhart 1986) used artificial neural networks to model the processing of human language. Directly opposed to connectionism was Fodor's (1983) modular approach to cognition that emphasized the independence of separate cognitive modules for each level of linguistic structure. Researchers have attempted to test the contrasting predictions of the interactive and modular approaches using standard experimental methodology. However, this work has indicated that neither of the strong positions can be maintained (Simpson 1994). To better understand the mechanisms involved, psycholinguists are now trying to link experimental methodology to methods for the imaging of the human brain during language processing. On this level, it appears that processing works in terms of interactive modules.

This article will examine research in six core areas of psycholinguistics: spoken word recognition, sentence comprehension, sentence production, message construction, memory limitations, and cross-linguistic comparisons. In addition to these core areas, psycholinguists are involved in the study of reading, conversational interaction, figurative language, text comprehension, aphasia, child language disorders, gesture, prosody, neurolinguistic imaging, animal communication, and language evolution. However, our analysis here will focus on these six core areas.

2. Word Recognition

When we listen to speech, we perceive words as following each other in clear temporal succession, like the beads on a string or the boxcars on a railway train. However, these perceptions underestimate the extent to which words are actually being blurred together by coarticulation and assimilation. Extreme examples of this blurring occur in forms such as *'supchu'* for 'What's up with you?' However, even nonphrasal sequences such as 'I owe you a yo-yo' show similar blurring. Even without blurring, the job of segmenting sentences into words would be a tough one. Consider a phrase such as 'my catalog value for Mark's piece.' Within this phrase, there are fragments that match a variety of other possible words, such as 'mike,' 'eye,' 'I,' 'cat,' 'at,' 'cattle,' 'log,' 'you,' 'Val,' 'ark,' 'arks,' 'are,' and 'form.' The reason why we tend not to hear these alternative forms is fairly simple. If we decided to commit ourselves to having heard the word 'mike,' we would end up with the nonsensical segmentation of 'mike at a log value for Mark's piece.' At each point during sentence perception, many of these alternative words are partially active in a short list of competitors. Competitors that recognize larger segments such as 'catalog' are preferred over those that recognize smaller strings or which break up larger strings into pieces, such as 'cat a log.' If a competitor leaves an unrecognized fragment, then its own recognition is weakened. These constraints work together to guarantee a maximally satisfactory segmentation. Although segmentation is primarily driven by the competition of words for matches, it is also facilitated by stress patterns, pauses, and other prosodic patterns (Norris 1994).

Segmentation relies on word recognition, but this process is also highly dynamic. Words do not have clear and invariant forms. Consonants leave no clear and invariant imprint on the auditory stream, since they are heavily blended into vowels. Because of differences in their vocal tracts, men and women produce vowels in radically different ways. Dialect differences and variations in speech level and further variability to word forms. Again, the solution to this problem involves lexical competition. For example, in the NAM model (Goldinger et al. 1989), the sound

corresponding to a word like 'deep' will activate a neighborhood of similar words, such as 'deal,' 'dear,' 'peep,' or 'keep.' Within this domain of rhyming forms, competition is particularly keen between words that share the same initial segments, such as 'deep,' 'deal,' 'dear,' and 'deed.' The ability of a word to dominate in this competition is a function of the strength of its match to the features of the input.

These competitive models can be given a specific neuronal instantiation in terms of a neural network model based on self-organizing feature maps (Mäkkiläinen 1993). In these models, an unorganized, but interconnected, flat sheet of simulated neural tissue is trained to recognize a set of input words. As learning progresses, words that share features move to adjacent areas in the feature map. In our example, the words 'deep,' 'deal,' 'dear,' and 'deed' would occupy a small neighborhood on the larger map. The network operates in a winner-take-all fashion so that, when 'deep' starts to receive the highest level of activation, it will inhibit the activation of the competitors through a series of lateral connections. Maps of this type mimic actual cortical tissue in their use of position to represent featural structure and their reliance on lateral inhibition to sharpen the outcome of a competition.

The study of the precise mechanics of word recognition has relied on studies of word reading rather than listening. The advantage of studying visual word recognition is that experimenters can use the laboratory computer to control tightly the display of words on the computer screen. The tasks most commonly used in this area are naming and lexical decision. Both tasks are sensitive to frequency, neighborhood effects, grammatical relations, and priming effects. For example, the word 'doctor' will prime a subject's naming of the word 'nurse' and the word 'govern' will prime naming of the words 'government' and 'governing,' but not 'misgovern.' This means that the time to read the word 'nurse' out loud will be less when it follows 'doctor,' than when it follows a control word like 'house.' The details of these various priming effects provide psycholinguists with a powerful method for mapping out the shape of the mental lexicon or dictionary (Balota 1994).

3. Sentence Comprehension

At one time, psycholinguists thought that the process of word recognition preceded the process of sentence comprehension. The idea was that we must first identify all the words in a sentence before we can feed these words to a comprehension mechanism that decides what it all means. In the late 1970s, Marslen-Wilson (1975) and others showed that sentence processing is not blocked out in this way. Instead, processing keeps up with word recognition in a fully

incremental, or online fashion. Both word recognition and sentence comprehension attempt to keep up with speech exactly as it comes in, although sometimes there is a slight lag, particularly for sentence comprehension. You might notice this lag if you find that you are trying to read a newspaper or watch a television program while someone is talking to you. You hear their words, but you may not be processing exactly what they say. They may then stop and ask you whether you have been listening. Somewhat dishonestly, you reply that you have and to prove this you repeat the last seven or eight words they have said. In fact, you were not really understanding the message in any very deep way. Rather, you were squirreling away the words in a superficial form in your short-term memory, hoping not to have to actually do the work of fully comprehending the message.

3.1 Incrementalism

The previous example is interesting, because it illustrates the exception to the general rule of incrementalism. Generally speaking, we process incoming material incrementally, both lexically and conceptually. Take as an example the sentence, 'The boy chased the baboon into the bedroom.' As soon as we hear 'the' and 'boy,' we immediately begin to relate them to each other. We then relate this unit to the following verb 'chased.' Milliseconds after hearing 'chased,' we begin to construct an interpretation of the activity in which there is a boy doing some chasing. By the time we hear 'baboon,' we can begin to sketch out the figure that the boy is chasing. We do not need to wait until we have heard all the words in the sentence to begin to extract these meanings. In this sense, sentence processing is both interactive and incremental—we tend to make decisions about material as soon as it comes in, hoping that the decisions that we make will not be reversed. Although processing is incremental, it only builds as much structure as it needs in order to keep words related. The real job of comprehension is delayed until more of the message is heard.

3.2 Garden-pathing

There are times when the initial decisions that we have made take us down the garden path. A classic example of garden-path processing occurs with sentences such as 'The communist farmers hated died.' It often takes the listener awhile to realize that it was the 'communist' that died and that it was the 'farmers' who hated the 'communist.' Inclusion of a relativizer to produce the form, 'The communist that farmers hated died' might have helped the listener sort this out. A somewhat different example is the sentence, 'The horse raced past

the barn fell.' Here, we need to understand 'raced past the barn' as a reduced relative clause with the meaning 'The horse who was raced past the barn.' If we do this, the appearance of the final verb 'fell' after 'barn' no longer comes as a surprise.

Garden paths arise when a word or suffix has two meanings, one of which is very common and one of which is comparatively rare (MacDonald et al. 1994). In a sentence like 'The horse raced past the barn fell' the use of the verb 'raced' as a standard transitive verb is much more common than its use as the past participle in a reduced passive. In such cases, the strong meaning quickly dominates over the weak meaning. By the time we realize our mistake, the weak meaning is fully suppressed by the strong meaning and we have to try to comprehend the sentence from scratch. A classic garden-path example from Lashley (1951) is the sentence 'Rapid righting with his uninjured left hand saved from destruction the contents of the capsized canoe.' When this sentence is read aloud, listeners find it extremely difficult to understand the second word as 'righting' rather than 'writing.'

3.3 Lexical Effects

Current models of sentence processing emphasize the extent to which lexical and syntactic ambiguities of individual words trigger competing syntactic structures (Trueswell and Tanenhaus 1994). For example, there are two readings of the sentence 'Flying planes can be dangerous.' The planes can be either dangerous to their pilots and passengers or dangerous to on-lookers down on the tarmac. Both interpretations of the participle 'flying' are fairly strong. Because the two readings are of similar strength, they can compete with each other and no garden-pathing arises. Another example of this type is 'He bought her pancakes' in which 'her' can be either the possessor of the pancakes or the recipient of the pancakes. Both meanings are strong and can compete with each other during sentence processing, yielding a clear ambiguity.

Sometimes lexically based expectations can be fairly complex. Consider these sentences in which the verbs 'criticize' and 'apologize' set up contrasting expectations:

John criticized Mary, because she hadn't delivered the paper on time.

John apologized to Mary, because he hadn't delivered the paper on time.

John criticized Mary, because he hadn't delivered the paper on time.

John apologized to Mary, because she hadn't delivered the paper on time.

Processing of the first two sentences is quick and easy, because the gender of the pronoun matches that of the expected agent. However, the processing of the second pair is more problematic, because the gender of

the pronoun forces the selection of an unexpected cause or of the criticism or the apology (McDonald and MacWhinney 1995).

3.4 Modularity

Although processing on individual linguistic levels is highly incremental, the interaction between levels is not immediate. During the first 300 milliseconds after hearing a word, we attend primarily to its auditory shape, rather than to the degree to which it fits into some grammatical context. Take as an example the sentence 'The sailors took the port at night.' Here the word 'port' could refer to either the wine or the harbor. We can ask subjects to listen to sentences like this while watching a computer screen. Directly after subjects hear the word 'port' we can present one of these three words on the computer screen: 'wine,' 'harbor,' and some control word such as 'shirt.' If we do this, we will find that the recognition of both 'wine' and 'harbor' is facilitated in comparison to the control word 'shirt.' If we change the sentence to something like 'The sailors drank the port at night,' we might expect that the context would bias the subject to respond more quickly to 'wine' than to 'harbor,' because one is not likely to drink a harbor. However, there is evidence that both 'wine' and 'harbor' are facilitated in comparison to the control word 'shirt,' even when the context tends to bias the 'wine' reading of 'port.' This facilitation is fairly short-lived and the contextually appropriate reading soon becomes dominant.

This type of result indicates that, in the first fraction of a second after hearing a word, we rely most strongly on auditory cues to guide our processing. This is not to say that context is not present or not being used as a cue. However, during the first fraction of a second, we need to focus on the actual auditory form in order to avoid any 'hallucinatory' effects of paying too much attention to context too soon in processing.

4. Sentence Production

There are many similarities between sentence comprehension and sentence production. In both activities, we rely heavily on the words in our lexicon to control syntactic structures. Both activities make use of the same patterns for determining grammatical structures. The most important difference is that, during sentence production, we are in full control of the meanings we wish to express. In comprehension, on the other hand, we are not in control and have to follow the ideas of others.

The production of sentences involves at least four processes (Levelt 1989). The first process is message construction. This process takes our goals and intentions and builds up a thread of ideas to be

articulated. The process of lexical access then converts these ideas in word form. The third process uses positional patterning to order words into phrases and clauses. The fourth process activates a series of verbal gestures through articulatory planning. As in the case of comprehension, these four stages are conducted not in serial order, but in parallel. Even before we have finished the complete construction of the message underlying a sentence, we begin the process of articulating the utterance. Sometimes we find out in the middle of an utterance that we have either forgotten what we want to say or do not know how to say it. It is this interleaved, incremental, online quality of speech production that gives rise to the various speech errors, pauses, and disfluencies that we often detect in our own speech and that of others.

Speech errors come in many different forms. Some involve simple slurring of a sound or retracing of a group of words. Others provide more dramatic evidence of the nature of the language planning process. Some of the most entertaining speech errors are spoonerisms, which owe their name to an English clergyman by the name of William Spooner. Instead of 'dear old queen,' Spooner produced 'queer old dean.' Instead of 'ye noble sons of toil,' he produced 'ye noble tons of soil.' Instead of 'I saw you light a fire in the back quad, in fact you wasted the whole term,' he said 'I saw you fight a liar in the back quad, in fact you tasted the whole worm.' These errors typically involve the transpositions of letters between words. Crucially, the resulting sound forms are themselves real words, such as 'liar,' 'queer,' and 'worm.' The tendency of these errors to produce real words is known as the 'lexical bias' in speech errors and indicates the extent to which the lexicon itself acts as a filter or checker on the articulatory process.

Another illuminating group of errors is named after a character named Mrs. Malaprop in the play *The Rivals* by Sheridan. Some examples of malapropisms include 'Judas Asparagus' for 'Judas Iscariot,' 'epitaphs' for 'epithets,' or 'croutons' for 'coupons.' Some malapropisms arise when uneducated speakers attempt to produce unfamiliar words, but many of these errors are true slips of the tongue. In a malapropism, the two words have a similar sound but a very different meaning. The fact that speakers end up producing words with quite the wrong meaning suggests that, at a certain point during speech planning, the output processor handles words more in terms of their phonological form than the meaning they express. It is at this point that malapropisms can occur.

When two words end up competing directly for a single slot in the output articulatory plan, the result is a lexical blend. For example, a speaker may want to describe both the flavor and the taste of some food and end up talking about its 'flaste.' Or we might want to talk about someone's 'spatial performance' and end up talking about their 'perfacial performance.' These

errors show the extent to which words are competing for placement into particular slots. When two words are targeting the same slot, one will usually win, but if there is no clear winner there can be a blend.

Another remarkable property of speech errors is the way in which grammatical markers seem to operate independently of the nouns and verbs to which they attach. Consider an exchange error such as 'the floods were roaded' for 'the roads were flooded.' In this error, the plural marker *-s* and the past tense marker *-ed* go to their correct positions in the output, but it is the noun and the verb that are transposed. This independent behavior of stems and their suffixes indicates that words are being composed from their grammatical pieces during sentence production and that grammatical markers contain clear specifications for their syntactic positions.

Some of the earliest thinking on the subject of speech errors was done by Sigmund Freud at the beginning of the century. Freud believed that slips of the tongue could provide evidence for underlying anxieties, hostilities, and worries. From this theory, arose the notion of a 'Freudian slip.' We now know that the majority of speech errors are not of this type, but it still appears that at least some could be viewed in this way. For example, if a speaker is discussing some activities surrounding a local barber, he might say 'He made hairlines' instead of 'He made headlines.' These errors indicate contextual influences from competing plans. However, even rare examples of this type of contextual influence seldom reveal underlying hostilities or neuroses.

5. Message Construction and Perspective

The basic goal of both sentence comprehension and sentence production is the linking of spoken forms to underlying conceptual meanings. In the 1970s, psycholinguists thought of these conceptual structures in terms of links between simple sentences or propositions. These propositions represented relations that had no clear relation to the actual physical perspective of the speaker. More recently, work in philosophy, robotics, psychology, and cognitive linguistics has recast this view of meaning structures in terms of dynamic activities that are linked to the embodied perspective of the speaker.

Consider what occurs when a speaker relates a fable, such as the story of 'the Tortoise and the Hare.' In this story, there are basically three perspectives: one for the tortoise, one for the hare, and one for the narrator. The narrator says 'the hare decided to rest.' At this point, the narrator is taking the perspective of the hare and describing the hare's actions from that perspective. Next, the narrator shifts to the perspective of the tortoise and describes how he passes the sleeping hare. These various shifts and blends of perspective

are important aspects of all narrative structures. However, we can see their impact everywhere in language. For example, if we want to point out that there is a dent in the side of the refrigerator, we may use the impersonal construction, 'someone knocked a dent in the refrigerator' or we may use the pseudo-passive, 'the refrigerator got a dent knocked into it.' In the impersonal we are forced to point a finger at an unspecified perpetrator of the deed. In the pseudo-passive, we background or de-emphasize the doer of the activity, focusing instead on the result. These choices all reflect variations in perspective.

Another choice we make is between alternative perspectives on a single action. For example, we could describe a picture by saying, 'the girl gave a flower to the boy.' In this case, we are taking the viewpoint or perspective of the girl and describing the activity of giving from her perspective. Alternatively, we could say, 'the boy got a flower from the girl.' In this alternative form, we view the action from the perspective of the recipient. Depending on whether we choose to view the action from the viewpoint of the giver or the receiver, we will use either the verb 'give' or the verb 'get' Language is full of choices of this type, and sentence production can be viewed as involving a continual competition between choices for perspectives. The selection of pronouns is determined by the need to link up referents to earlier perspectives. Shifts from the present tense to the past also reflect movements in perspective to various points in time.

6. Cognitive Resources

Our ability to process language is shaped by our basic cognitive capacities. Even people with severe mental retardation acquire a basic control over spoken language. On the other hand, normal speakers vary in the degree to which they can remember long strings of words. The notion of a verbal working memory (Gathercole and Baddeley 1993) has figured heavily in attempts to derive the psychological reality of linguistic rules. For example, a sentence with multiple center embeddings, such as 'the man the dog the cat scratched bit smiled' are impossible for most people to understand. Psycholinguists believe that the problem here lies not in the violation of a linguistic rule, but in the fact that this structure places unrealistic demands on memory. Variations in memory span or capacity have also been related to differences in ability to learn words in a new language (Gupta and MacWhinney 1997), process complex sentences (Just and Carpenter 1992), and acquire reading skills in school.

Capacity limitations also shape the ways in which we form long-term memories from language. In general, the exact surface form of sentences often fades within seconds after we hear them (Potter and Lombardi 1990). This is particularly true of the

commonplace, predictable material that is used in textbooks or in experiments on sentence processing. However, if one examines memory for more lively, charged, interpersonal communications, a very different picture emerges. For this type of discourse, memory for the specific wording of jokes, compliments, insults, and threats can be extremely accurate, extending even over several days (Keenan et al. 1977). Often we are careful to note the exact phrasing of language spoken by our close associates, since we assume that this phrasing contains important clues regarding aspects of our interpersonal relations. For material that contains no such interpersonal content, we focus only on underlying meaning, quickly discarding surface phrasings.

7. Cross-linguistic Comparisons

All languages have sound segments, syllables, words, phrases, and sentences. In all languages, processing is online, incremental, and interactive. However, languages differ markedly in the particular grammatical contrasts they choose to mark and the devices that they use to mark them. In languages such as Warlpiri or Hungarian, word order is remarkably flexible. In languages such as Navajo or Eskimo, a single complex word may express what English does in a seven-word sentence.

Psycholinguistic comparisons of processing in different languages (MacWhinney and Bates 1989) use a simple common task to examine the effects of radically different structures. In one type of experiment, subjects listen to a sentence such as 'The eraser is pushed by the cat' and have to point to a picture that shows who is the actor. Consider a comparison of these sentences from English and Spanish:

The lion kisses the cow.

El león besa la vaca. (The lion kisses the cow).

A major difference between these two languages revolves around the use of variable word orders. In Spanish, it is possible to invert the word order and produce *la vaca besa el león*, while still meaning that the lion is kissing the cow. This inversion is even clearer if the particle *a* is added to mark the direct object, as in these variant orderings in which the two nouns are moved into different places around the verb *besa*.

El león besa a la vaca.

A la vaca besa el león.

Besa el león a la vaca.

Besa a la vaca el león.

These differences between English and Spanish can be traced to the relative cue validities of word order and object marking in the two languages. In English, it is virtually always the case that the noun that appears before the verb is the subject of the sentence. If the verb is an active verb, the preverbal noun is almost

always the actor. This means that the cue of preverbal positioning is an extremely reliable guide in English to assignment of the actor role. In Spanish, there is no such simple rule. Instead, the best cue to assignment of the actor role is the presence of the object marker particle *a*. If one of the nouns in a two-noun sentence is marked by *a*, then we know that the other noun is the agent or subject.

Other languages have still other configurations of the basic cues to sentence interpretation. For example, Hungarian makes reliable use of a single case marking suffix on the direct object. German uses the definite article to mark case and number. The Australian language Warlpiri marks the subject with a variety of complex case markings. Navajo places nouns in front of the verb in terms of their relative animacy in the Navajo 'Great Chain of Being' and then uses verbal prefixes to pick out the subject. These languages and others also often rely on number agreement between the verb and the noun as a cue to the subject. English also requires subject-verb agreement, but this cue is often missing or neutralized. In languages such as Italian or Arabic, subject-verb agreement marking is extremely clear and reliable.

When a monolingual speaker comes to learning a second language, they need to fundamentally retune their sentence-processing mechanism. First, they need to acquire a new set of grammatical devices and markings. Second, they need to associate these new devices with the correct cue validities. Third, they need to reorganize their expectations for particular sequences of cues and forms. Initially, the learner simply transfers the cues, cue validities, and habits from the first language to processing of the second language. Over time, the cue validities change smoothly, until they eventually match that of the native monolingual (de Groot and Kroll 1997).

8. Summary

In this article we have examined six core issues being addressed by psycholinguists. These areas are central to the study of psycholinguistics, because they allow us to evaluate the psychological reality of linguistic formalisms. The interaction between psycholinguistics and linguistic theory has been intense, reciprocal, and dynamic. We have learned that language processing is interactive and incremental in regard to word recognition, sentence comprehension, sentence production, and message construction. We know that cross-linguistic differences in processing are closely related to differing levels of reliability for the cues involved and the ways in which structures impose demands on verbal memory capacity. Some versions of linguistic theory have moved to adapt to some of these findings from psycholinguistics (Pollard and Sag 1994). However, a satisfactory resolution of the core

issue of the psychological reality of linguistic structures has not yet been obtained.

See also: Computational Psycholinguistics; Connectionist Models of Language Processing; First Language Acquisition: Cross-linguistic; Language Acquisition; Language and Thought: The Modern Whorfian Hypothesis; Second Language Acquisition

Bibliography

- Balota D A 1994 Visual word recognition: The journey from features to meaning. In: Gernsbacher M (ed.) *Handbook of Psycholinguistics*. Academic Press, San Diego, CA
- de Groot A M B, Kroll J F 1997 *Tutorials in Bilingualism*. Erlbaum, Mahwah, NJ
- Fletcher P, MacWhinney B J (eds.) 1995 *The Handbook of Child Language*. Blackwell, Oxford, UK
- Fodor J 1983 *The Modularity of Mind: An Essay on Faculty Psychology*. MIT Press, Cambridge, MA
- Gathercole V, Baddeley A 1993 *Working Memory and Language*. Erlbaum, Hillsdale, NJ
- Gazzaniga M (ed.) 1997 *The Cognitive Neurosciences*. MIT Press, Cambridge, MA
- Goldinger S D, Luce P A, Pisoni D B 1989 Priming lexical neighbors of spoken words: Effects of competition and inhibition. *Journal of Memory and Language* 28: 501–18
- Gupta P, MacWhinney B J 1997 Vocabulary acquisition and verbal short-term memory: Computational and neural bases. *Brain and Language* 59: 267–333
- Just M, Carpenter P 1992 A capacity theory of comprehension: Individual differences in working memory. *Psychological Review* 99: 122–49
- Keenan J, MacWhinney B J, Mayhew D 1977 Pragmatics in memory: A study in natural conversation. *Journal of Verbal Learning and Verbal Behavior* 16: 549–60
- Lashley K 1951 The problem of serial order in behavior. In: Jeffress L A (ed.) *Cerebral Mechanisms in Behavior*. Wiley, New York
- Levelt W J M 1989 *Speaking: From Intention to Articulation*. MIT Press, Cambridge, MA
- MacDonald M C, Pearlmuter N J, Seidenberg M S 1994 Lexical nature of syntactic ambiguity resolution. *Psychological Review* 101(4): 676–703
- MacWhinney B J, Bates E (eds.) 1989 *The Crosslinguistic Study of Sentence Processing*. Cambridge University Press, New York
- Marslen-Wilson W 1975 Sentence perception as an interactive parallel process. *Science* 189: 226–7
- McClelland J L, Rumelhart D E 1986 *Parallel Distributed Processing*. MIT Press, Cambridge, MA
- McDonald J L, MacWhinney B J 1995 The time course of anaphor resolution: Effects of implicit verb causality and gender. *Journal of Memory and Language* 34: 543–66
- Miikkulainen R 1993 *Subsymbolic Natural Language Processing*. MIT Press, Cambridge, MA
- Norris D 1994 Shortlist: A connectionist model of continuous speech recognition. *Cognition* 52: 189–234
- Osgood C, Sebeok T A (eds.) 1967 *Psycholinguistics: A Survey of Theory and Research Problems With a Survey of Psycholinguistic Research, 1954–1964*. Indiana University Press, Bloomington, Indiana

- Pinker S 1999 *Words and Rules: The Ingredients of Language*. Basic Books, New York
- Pollard C, Sag I 1994 *Head-driven Phrase Structure Grammar*. Chicago University Press, Chicago
- Potter M, Lombardi L 1990 Regeneration in the short-term recall of sentences. *Journal of Memory and Language* 29: 633–54
- Simpson G B 1994 Context and the processing of ambiguous words. In: Gernsbacher M (ed.) *Handbook of Psycholinguistics*. Academic Press, San Diego, CA
- Stemmer B, Whitaker H 1998 *Handbook of Neurolinguistics*. Academic Press, New York
- Trueswell J C, Tanenhaus M K 1994 Toward a lexicalist framework for constraint-based syntactic-ambiguity resolution. In: Clifton C, Rayner K, Frazier L (eds.) *Perspectives on Sentence Processing*. Erlbaum, Hillsdale, NJ, pp. 155–79

B. J. MacWhinney

Psychological Anthropology

Psychological anthropology considers connections between the individual and sociocultural milieu, including cultural influences on personality and psychological foundations of society and culture. The field comprises studies of child rearing, self-representation, emotion, motives, moral reasoning, cognition, dreaming, mental disorder, gender relations, violence, racism, and cultural symbolism, among other things. Theoretical and methodological perspectives include cultural psychology, ethnopsychology, psychoanalytic anthropology, cross-cultural studies of child development, evolutionary psychological anthropology, and cognitive anthropology (see *Cognitive Anthropology*).

Although it shares concerns with several related fields (i.e., social psychology, cultural psychology (in psychology), cross-cultural psychology, and ‘applied’ psychoanalysis), psychological anthropology differs from these enterprises in being strongly committed to ethnographic fieldwork in diverse cultures. This engagement with culture (and cultures) leads psychological anthropologists to put two demanding questions to theories of mind and personality: they ask if such theories adequately consider cultural influences on the individual, and whether they can in any way illuminate cultures, particularly the symbolic content of expressive culture and the logic of local knowledge. Whereas the first question poses doubts about the universal validity of Western psychology, the second, pointing in the opposite direction, encourages interest in psychoanalysis and cognitive science, psychoanalysis because it addresses content and cognitive science because it addresses logic.

Some cultural determinists or culturalists define their work as ‘cultural psychology.’ In their view,

'psychological anthropology' implies universalism and reductionism and, correlatively, too little concern with local cultures. Concurring with this appraisal, still others narrow their concerns to 'ethnopsychology,' to local cultural ideas about the mind or person. Yet for many others 'psychological anthropology' remains a comprehensive enterprise that allows for both cultural influences on subjectivity and for psychological interpretations of society and culture. It may also include evolutionary perspectives on human nature and attention to precultural social environment in child development.

1. Culture and Personality

Franz Boas, the founder of modern American anthropology, was interested in general psychology or human nature as well as 'folk psychology,' the way in which the 'genius of peoples' is shaped by local history and culture (Stocking 1992). Many of his students, however, were more influenced by his interest in this second, cultural psychology than by his interest in general psychology. Margaret Mead argued for the plasticity of sexuality and gender relations, and she rebutted Piaget and others effectively who were asserting similarity between the thinking of 'primitives' and children. Ruth Benedict contended that cultures inculcate radically different worldviews and, therefore, that subjectivity varies substantially from one culture to another. In yet another version of this cultural determinism, the neo-Freudian psychoanalyst Abram Kardiner, working with several anthropologists, suggested that the 'primary institutions' of a culture (i.e., family organization, child-rearing practices) produce a distinctive 'basic personality structure.' This shared personality core, in turn, informs 'secondary institutions' (e.g., art, myth, ritual, politics) through projection. The model guided attempts to measure basic personality structure with projective personality tests and to document the effects of child rearing with cross-cultural studies, such as those carried out by John W. M. Whiting and his many students.

The principal tenets of this 'culture and personality' theorizing and research were congruity between personality and culture, and reproduction of shared personality and culture through child-rearing practices. By the late 1960s, these tenets were in doubt. Life histories and projective personality testing had demonstrated shared psychological traits in communities, but they had also revealed much individual variation. The assumption that child-rearing practices have robust consequences for personality and culture also seemed less certain. Edward Sapir, Boas's most brilliant student, expressed reservations about culture and personality well before these empirical results. In his view, personality and culture are interrelated, partly because they have a common nexus in symbolic communication and social interaction, but he also emphasized that individual personality is not simply a

passive effect of social environment. Observing that cultural knowledge and practices are unevenly distributed in groups and that the same cultural practices can have diverse meanings and emotional entailments for individuals, Sapir inferred that the individual is the creative locus and foundation of culture. Although he had misgivings about it, Sapir concluded that psychoanalysis was the best available language for describing individual psychology.

More in keeping with Sapir's thinking, some psychoanalytically oriented anthropologists focused on expressions of human nature and individual personality in cultural beliefs and practices (e.g., Géza Róheim, George Devereux, Weston La Barre, and Melford E. Spiro). Róheim, the first of these psychoanalytic anthropologists, was especially critical of culture and personality for ignoring human nature and for underestimating cultural and psychological diversity within societies and communities.

2. Social Anthropology and Psychoanalysis

In the early twentieth century British anthropologists turned their attention from social evolution to actual social life, including questions about social actors. Accordingly, they were interested in the newly emerging fields of social psychology and psychoanalysis. Both C. G. Seligman and W. H. R. Rivers read Freud. Bronislaw Malinowski especially was intrigued by the conflict between sexual impulses and social norms in psychoanalytic theorizing. A pragmatic empiricist, Malinowski nonetheless had difficulty grasping hidden structures and processes, whether social or psychological. Assuming that intrapsychic complexes were evident in observable social relations, he argued that the Oedipus complex takes a quite different form in the matrilineal Trobriands. When the response among psychoanalysts to this intervention was less than positive, Malinowski's already ambivalent interest in psychoanalysis waned considerably (Stocking 1986).

Various expressions and degrees of ambivalence about psychoanalysis were in fact common in British social anthropology. While social anthropologists sought to understand culture in terms of social relations, it was apparent that the drama of social life involves motivated social actors with personal histories. They could not help but notice, moreover, that expressive culture often includes oedipal themes and bodily imagery (e.g., Meyer Fortes, Edmund Leach, and Victor Turner but especially John Layard, Kathleen Gough, T. O. Beidelman, and L. R. Hiatt).

3. Psychological Anthropology

As the basic tenets of culture and personality became more questionable, the field began in the 1960s to reinvent itself under the heading of 'psychological anthropology.' There was less interest in typifying

entire cultures and more concerted efforts to explore specific issues. Tensions between culturalist and psychoanalytic or otherwise psychological approaches have persisted, however. Recently, the theoretical scene has been further complicated by the emergence of evolutionary psychological anthropology, a perspective that competes with psychoanalytic anthropology in some respects and complements it in others.

3.1 Human Nature

Culturalists contend that human nature lacks content, except perhaps for a capacity for learning language and culture. Many psychoanalytically oriented anthropologists are convinced, however, that human beings are desiring, emotional, and moral creatures as well as creatures capable of learning language and culture (Ingham 1996). For their part, evolutionary psychological anthropologists argue that human beings come equipped with inclinations and capabilities oriented toward ensuring inclusive fitness. The evolutionary and psychoanalytic perspectives converge in some respects, particularly in their attention to deception and sexuality. Proposals by evolutionary psychological anthropologists and evolutionary psychologists that childhood experience may shape reproductive strategies also converge in interesting ways with thinking in psychoanalytic anthropology. The ethnographic evidence for these varying views of human nature is ambiguous. While cultural diversity argues against an essential human nature, recurrent themes in the ethnographic data appear to have just the opposite implication. In support of a common human nature, it is noticed, for example, that oedipal themes and culture heroes and mischievous tricksters are nearly everywhere present in myth and folklore (Johnson and Price-Williams 1996).

3.2 Incest Aversion

Although there are attempts to explain incest aversion in cultural terms, the most lively discussions involve psychoanalytic and evolutionary, neo-Westermarckian perspectives. Wolf's work in particular evokes debate, partly because it purports to have devastating consequences for psychoanalysis. Wolf adduced data on Chinese *simpua* marriage (a son marries an adopted daughter with whom he has a 'sisterlike' relationship) to argue that familiarity in childhood inhibits sexual desire (the *simpua* marriages result in higher rates of infertility and divorce). Spain (1987) responded that both Freud and Westermarck predict no incestuous behavior as a result of normal child development. Subsequently, Wolf (1995), again dismissing Freud, drew on attachment theory to argue that the familiarity effect begins as early as the second year of life,

presumably before the oedipal period. Nonetheless, a further response to Wolf seems likely. Wolf did not investigate mother-child relationships or childhood sexuality, and he does not consider that attachment theory is a branch of psychoanalysis and that there is clinical literature that implicates early attachment in the development and resolution of childhood oedipality.

3.3 Embodiment and Desire

It has become increasingly common to assert that the body and desire are socially or culturally constructed. 'Embodiment' in this thinking refers not to the bodily grounding of the self or mind but to the way in which culture or discourse shapes ideas about the body and bodily experience. The experience of embodiment is then expressed in 'narratives' about the body.

A different approach to embodiment assumes a more active, motivated subject in which motivation and body image reflect unconscious object relations and internalized moral values as well as the more immediate effects of social discourse. In a study of a recovered anorexic woman, for example, Banks (1998) shows that the woman's changing personal narratives about eating had more to do with her emotional experiences, personal relationships, and religious convictions than dominant cultural discourses about beauty and thinness. A fantasy story about a cookie jar, written in a diary at 11 years of age, implies that the woman's anorexia during adolescence was a way of organizing memories of maternal neglect in early life and religiously inflected anxieties about the dangers of sexuality. The temptation of the cookies and the decision to forego them in the story anticipated the way in which anorexia would become a defense against sexuality during adolescence. Far from being a passive self, the woman was an active subject who responded to discourses about the body in terms of her personal experience and deeply internalized religious convictions.

Culturalists and discourse theorists say that sexual desire and romantic love are 'socially constructed' or 'invented' along with the body, and, to be sure, they provide ample ethnographic material in support of this contention. It is not clear, however, how discourse fashions motivation or affect in desire and love. Cognitive and culturalist anthropologists theorize that motives are represented in the mind as cultural schemas (D'Andrade and Strauss 1992). A case has been made for this approach but, again, it is uncertain how schemas 'stick' in the mind or how they acquire affective force. As Banks's study suggests, cultural schemas may take hold in intrapsychic object relations, that is, in unconscious representations of emotionally significant relationships, especially with parents. If this is so, it does not suffice to say that the

body or desire is socially constructed. Developmental studies of individuals challenge simplified versions of social constructionism but so, too, does the evidence for cultural universals. One of these challenges is the finding that more or less durable passionate attachment is virtually universal in human societies (Jankowiak 1995).

3.4 Attachment, Oedipus Complex, Socialization

Observations in various cultures suggest that warm, nurturing parenting and paternal presence in early life lay an emotional basis for sociability and positive self-regard. They also suggest that secure early attachment postpones and reduces sexual activity in adolescence and lays an emotional foundation for durable love relationships and warm, empathic parenting. An evolutionary argument relates these findings to reproductive strategy, that is, it is theorized that during the course of human evolution secure attachment in childhood promoted durable passionate attachment and high-quality parenting in adulthood, whereas insecure attachment promoted increased sexual activity and multiple relationships. The first pattern supposedly favored reproductive success where supportive environments resulted in lower death rates, whereas the second pattern was more likely to lead to reproductive success where adverse material conditions resulted in higher mortality (Chisholm 1992). Another formulation finds the Oedipus complex in these scenarios. In this argument, secure attachment results in early onset and resolution of oedipal attachments and, eventually, in an emotional capacity for enduring passionate attachment, whereas insecure attachment promotes more persistent and sexualized forms of oedipal attachment and, consequently, promiscuity and other forms of sexualizing and objectifying sexual partners. Accordingly, the secure form of the Oedipus complex supported the transition to pair-bonding and increased paternal investment in offspring during early human evolution (Ingham 1999).

Attachment is probably everywhere part of early child development, but this need not imply that secure attachment is everywhere produced in the same way. On the contrary, evidence suggests that diverse parenting styles and family arrangements can promote secure attachment. Whether it can be added that attachment and the Oedipus complex vary greatly from society to society continues to be a subject for debate, however. Some writers think they do (Spain 1992), while others find the evidence less than persuasive (Spiro 1992). Attachment and oedipality aside, there is no doubt that children are socialized into their cultures. Many studies show that children learn their cultures from an early age, and that this learning informs perception, emotional expression, and moral reasoning as well as social attitudes and worldview. Many such studies contest assumptions about what

constitutes 'normal' development. Shweder (1991), for example, argues that theories of moral development may say more about Western concerns with individual rights than anything about the universal course of moral development.

3.5 Self and Emotion

Culturalists and postmodernists argue that the self and emotion, like embodiment and desire, are culturally determined or socially constructed, and, consequently, that they too vary greatly from one culture to another. One often-repeated idea is that the self is more individualistic or 'egocentric' in modern societies and more communal or 'sociocentric' in traditional societies. Implicit in these arguments is an assumption that what people think about themselves and how and what they feel are heavily constrained by words and idioms. It is argued, for example, that Russians are more expressive emotionally than Americans. In support of this contention, it is noted that Russians have a richer store of words and phrases for describing emotion and more linguistically marked emotional bodily expression than Americans (Wierzbicka 1998).

It can be objected, however, that linguistic constructionism risks typifying whole peoples and underestimating psychological complexities. Although self-descriptions may differ somewhat from culture to culture in how much they refer to social ties, studies in various cultures show that individuals are often better described as having both individualistic and communal inclinations, and, moreover, that this is so even in cultures that would seem to be highly communal (see, e.g., McHugh 1989). Research in the US suggests that the common assertion that Americans are unusually individualistic may be doubly simplistic, not only because Americans are communal as well as individualistic but also because their individualism assumes different forms depending on social class and neighborhood.

What is proving more interesting for many researchers than local lexicons is the way in which self-representation and emotional expression are organized through the stories people tell about themselves and to themselves. Research is being done, for instance, on how individuals use self-narratives to reconstruct and negotiate social identities in diasporic communities. Interest in self-narratives is part of a broader trend toward methodologically rigorous discourse-centered ethnography in psychological anthropology. Speech productions are tape-recorded, transcribed with technical notation, and examined for narrative style and speech style. They are then further examined for what they express about the self and culture. Since discourse is inherently dialogical—interlocutors are both speakers and listeners—attention to discourse translates the abstract relation

between self and society, psyche and culture, into a more concrete mutual relation between interlocutors. In studies of personal narratives about religious conversion, Stromberg (1991) shows how conversion makes culture personally meaningful by transforming referential symbols into constitutive symbols. In the conversion process, symbols that refer to the external world become part of self-identity. At the same time, they may become a personal language for managing intrapsychic conflict, say between desire and moral duty. In one of his studies a man feels powerless and insufficiently masculine. Speech style, body posturing, and cultural symbolism imply that this self-organization involves wishes for masculine power, memories of a brutal father, and feelings of having been overpowered by women, especially his mother and motherlike wife. His dreams represent a masculine power as phallic, flying horses, and the mother figure as a vaginallike tornado. Recommitment to Christian faith occurs after the man identifies in his dreams with a potent, athletic Christ (Stromberg 1991).

Another set of issues concerns how universal emotions or common emotional predicaments are organized in particular cultures. One example is grieving, arguably a universal experience. In some cultures mortuary rituals involve multiple burials and reburials or cremations and re-cremations in which the deceased is figuratively and literally deconstructed and reconstructed. Beliefs associated with such practices sometimes link the mother with the deceased, and the rituals themselves may involve eating the flesh of the deceased and giving the flesh of the deceased or its symbolic equivalent to relatives. In a comparative study of these practices, Stephen (1998) draws on Melanie Klein's object-relations psychoanalysis to theorize that many emotional losses resonate with separation from the mother at some level and may, therefore, activate unconscious traces of child rage and guilt and related needs to make reparation.

3.6 *Unconscious Meaning and Motivation in Myth and Ritual*

Because it seems to speak to symbolic content in this way, efforts to understand the unconscious meaning and motivation of expressive culture draw more on psychoanalysis than any other psychology. Psychoanalytically oriented anthropologists have often argued that relations with supernatural figures recreate or repair imaginary relations with parents (see Obeyesekere 1981). Various studies suggest that puberty rites and adolescent initiation rites may be oriented toward resolving oedipal ties to parents and orienting young people toward marriage. Ancient Greek myths can be read as implying that adolescent transition rites function in this way. Perhaps the exception that proves the rule, the story of Oedipus may be a deviant form of more common myths in which the hero marries and becomes king after passing

through an initiatory trial with a monster (see Goux 1993). Other elements in Western culture may have similar socializing functions. In a rigorously close reading of the text and its midrashic commentaries, Paul (1996) shows that mythic patricide and incestuous desire are recurrent though disguised themes in the five books of Moses. The reading suggests the Torah activates unconscious oedipal fantasies and related guilt feelings but that it also represents both paternal prohibitions and a substitute for the maternal object. In effect, the Torah may serve the growing boy or girl in the Hebrew community as a model for normative displacement and transformation of childhood attachments.

Culturalists frequently complain that psychoanalytic interpretations of cultural symbolism are simplistically universalistic and reductionistic. This may be a fair criticism in some cases—there are examples of overly mechanical and insensitive uses of psychoanalytic interpretation in the anthropological literature—but it seems unfair in many others, including those cited here. Psychoanalytic anthropologists are usually careful to consider ethnographic details and conscious personal and cultural meanings. Indeed, they contend that the evidence for unconscious meanings and motives lies precisely in such ethnographic details and conscious meanings.

4. *Basic Dilemma, Contemporary Perspectives*

As in the past, psychological anthropologists are struggling with an empirical and theoretical bind involving cultural particularism and psychological universalism, the influence of the sociocultural milieu on the individual, on the one hand, and the motivated agency of psychological subjects, on the other. Advocates of cultural psychology and ethnopsychology try to resolve the dilemma on the side of culture. They sometimes say their viewpoint allows for an active self—this is the point of their replacing the language of cultural determinism with that of social and cultural constructionism—but it is often difficult to see how their self is more than a passive product of culture. While their culturalist approach opens productive questions about how individuals are shaped or constrained by their sociocultural milieus, usually it leaves motivation and the perplexing, enigmatic features of those milieus in question. Universalistic psychoanalytic and evolutionary approaches have a reverse set of strengths and weaknesses. They offer theories and hypotheses about motivation but they are less well equipped to deal with social context and cultural influence.

Cognizant of both horns of the dilemma, many psychological anthropologists are exploring ways of conceptualizing the interplay between self and other, personality and culture. On the culturalist/constructionist side of the discussion, the growing interest in discourse-centered ethnography is arguably the most

noteworthy development (Besnier 1994). While it may sidestep motive and volition, it encourages attention to symbolic expression; it takes the culturalist approach beyond reified notions like individualism and sociocentrism to more detailed understanding of the complexities of subjectivity and culture; and, finally, because speech is inherently dialogical, it conduces toward a view in which the subject is not only the locus of cultural influence but also an active, speaking agent in the production and reproduction of culture. On the psychoanalytic side, there is effort to renew the encounter between psychoanalysis and social anthropology (Augé 1999, Heald and Deluz 1994). In this approach, culture mediates social relations and is not reducible to individual psychology but may tap into unconscious fantasy and emotion. In a dialogical version of this approach, individuals and groups influence each other through pragmatic discourse. More particularly, their rhetorical practices exert social and political effects by suppressing, mobilizing, and redirecting common sense, moral reasoning, and unconscious motivation, especially around issues of attachment, sexuality, aggression, and love. Finally, attention is given to the ways in which individuals resist these tactics or accede to them according to their social positions and psychological dispositions (Ingham 1996). In an attempt to integrate discourse theory with culturalist, cognitive, and psychoanalytic approaches, Nuckolls (1996) argues that cultural knowledge everywhere emerges in conversation and that it is always a paradoxical, dialectical mixture of diverse desires, goals, and rationalizations. At the same time, the dissonance in cultural knowledge engenders desires for coherence and synthesis. Unity and agreement remain elusive, however, and when achieved soon give way to further dialogue and disagreement. What is true of cultural knowledge in these respects is evidently true of psychological anthropology, itself an ongoing conversation. However appealing synthesis may be, it seems that psychological anthropology is entrapped in paradox and dialectic no less than its subject matter.

See also: Benedict, Ruth (1887–1948); Boas, Franz (1858–1942); Cultural Psychology; Emotion: History of the Concept; Incest Prohibition, Origin and Evolution of; Malinowski, Bronislaw (1884–1942); Myth in Religion; Psychoanalysis: Overview; Ritual; Sapir, Edward (1884–1939)

Bibliography

- Augé M 1999 *The War of Dreams: Exercises in Ethno-fiction* (trans. Heron L). Pluto Press, London
- Banks C G 1998 Illness narratives and human development: Changing metaphors about the body in a case of anorexia nervosa. In: Denzin N K (ed.) *Cultural Studies: A Research Annual*, Vol. 3. JAI Press, Stanford, CT
- Besnier N 1994 The evidence from discourse. In: Bock P K (ed.) *Handbook of Psychological Anthropology*. Greenwood Press, Westport, CT
- Chisholm J S 1992 Putting people in biology: Toward a synthesis of biological and psychological anthropology. In: Schwartz T, White G M, Lutz C A (eds.) *New Directions in Psychological Anthropology*. Cambridge University Press, Cambridge and New York
- D'Andrade R G, Strauss C A (eds.) 1992 *Human Motives and Cultural Models*. Cambridge University Press, Cambridge, UK
- Goux J-J 1993 *Oedipus, Philosopher* (trans. Porter C). Stanford University Press, Stanford, CA
- Heald S, Deluz A (eds.) 1994 *Anthropology and Psychoanalysis: An Encounter Through Culture*. Routledge, London
- Ingham J M 1996 *Psychological Anthropology Reconsidered*. Cambridge University Press, Cambridge, UK
- Ingham J M 1999 The Oedipus complex and the signifying chain: A view from anthropology. *Journal for the Psychoanalysis of Culture & Society* 4: 50–62
- Jankowiak W (ed.) 1995 *Romantic Passion: A Universal Experience?* Columbia University Press, New York
- Johnson A, Price-Williams D 1996 *Oedipus Ubiquitous: The Family Complex in World Folk Literature*. Stanford University Press, Stanford, CA
- Obeyesekere G 1981 *Medusa's Hair: An Essay on Personal Symbols and Religious Experience*. University of Chicago Press, Chicago
- McHugh E L 1989 Concepts of the person among the Gurungs of Nepal. *American Ethnologist* 16: 75–86
- Nuckolls C W 1996 *The Cultural Dialectics of Knowledge and Desire*. University of Wisconsin Press, Madison, WI
- Paul R A 1996 *Moses and Civilization: The Meaning Behind Freud's Myth*. Yale University Press, New Haven, CT
- Shweder R A 1991 *Thinking Through Cultures: Expeditions in Cultural Psychology*. Harvard University Press, Cambridge, MA
- Spain D H 1987 The Westermarck-Freud incest-theory debate: An evaluation and reformulation. *Current Anthropology* 28: 623–45
- Spain D H 1992 Oedipus Rex or edifice wrecked? Some comments on the universality of oedipality and on the cultural limitations of Freud's thought. In: Spain D H (ed.) *Psychoanalytic Anthropology After Freud: Essays Marking the Fiftieth Anniversary of Freud's Death*. Psyche Press, New York
- Spiro M E 1992 Oedipus redux. *Ethos* 20: 358–76
- Stephen M 1998 Devouring the mother: A Kleinian perspective on necrophagia and corpse abuse in mortuary ritual. *Ethos* 26: 387–409
- Stocking Jr G W 1986 Anthropology and the science of the irrational: Malinowski's encounter with Freudian psychoanalysis. In: Stocking Jr G W (ed.) *Malinowski, Rivers, Benedict and Others: Essays on Culture and Personality*. University of Wisconsin, Madison, WI
- Stocking Jr G W 1992 Polarity and plurality: Franz Boas as psychological anthropologist. In: Schwartz T, White G M, Lutz C A (eds.) *New Directions in Psychological Anthropology*. Cambridge University Press, Cambridge, UK
- Stromberg P G 1991 Symbols into experience: A case study in the generation of commitment. *Ethos* 19: 102–26
- Wierzbicka A 1998 Russian emotional expression. *Ethos* 26: 456–83

Wolf A P 1995 *Sexual Attraction and Childhood Association: A Chinese Brief for Edward Westermarck*. Stanford University Press, Stanford, CA

J. M. Ingham

Psychological Climate in the Work Setting

Psychological climate is defined as 'employees' perceptions of their work environment' and has been the focus of considerable research in organizational psychology. Measures of psychological climate are intended to assess work environments as they are perceived and interpreted by employees (James et al. 1990). Examples of psychological climate dimensions include, but are not limited to, role clarity, job importance, leader support, and workgroup cooperation.

Research has linked psychological climate variables to work-related outcomes such as job satisfaction, job involvement, and job performance. Furthermore, research has attempted to determine whether there is a hierarchical model to psychological climate. That is, do the individual climate variables all represent assessments of the degree to which organizational events can contribute to desired ends (as described by one's personal values) and reflect an underlying higher order construct?

The next sections will examine the development of psychological climate research from its conception to its current state of affairs and probable future research directions.

1. The Origins of Psychological Climate Research

Beginning with Lewin et al.'s (1939) discussion of 'social climates' in the workplace, assessing employee climate perceptions became increasingly interesting to organizational researchers. However, the idea of psychological climate, as it is studied now, was not first comprehensively discussed until the late 1960s. Litwin and Stringer (1968) presented a model of the determinants of motivated behavior in organizations that included the concept of organizational climate. It should be pointed out that in the early stages of psychological climate research it was referred to as organizational climate, though that term is now used to describe shared perceptions of organizational policies and practices, rather than individual employees' perceptions of characteristics of the workplace. According to Litwin and Stringer (1968), the organization is seen as generating an organizational climate, which in turn either positively or negatively affects particular motivational patterns of employees. They theorized that climate could provide managers with the link between an organization's procedures/prac-

tices and the concerns and needs of individual employees. This link would help managers determine what practices would stimulate employee motivation. Litwin and Stringer (1968) defined several important dimensions of organizational climate, including structure (perception of formality and constraint in the organization), challenge (perception of challenge, demand for work, and opportunity for sense of achievement), reward and support (emphasis on positive reinforcement rather than punishment), and social inclusion (sociability, belonging, and group membership). Finally, Litwin and Stringer (1968) tested their hypotheses in an experiment that simulated eight days of actual organizational life in three companies with very different organizational climates. The organizations differed in terms of the amount of structure (strict procedures vs. more flexible procedures), responsibility given to employees (e.g., low vs. high), reward and punishment (e.g., only punishment vs. a combination of both punishment and rewards), warmth and support (isolated employees vs. friendly workgroups), cooperation and conflict (e.g., cooperation is praised or not), and risk and involvement (e.g., organization stressing conservatism vs. stressing challenge). Litwin and Stringer's (1968) results indicated that type of organizational climate was related to employee satisfaction and employee performance. Overall, they found that what they called an achieving climate (e.g., high employee responsibility, cooperation among employees is stressed, tolerate conflict and support risk taking) led to the best outcomes for the employees (i.e., high satisfaction) and the organization (i.e., high performance).

Litwin and Stringer's (1968) work provided the impetus for increased research on psychological climate in work settings. This subsequent research identified several problems that confounded much of the original psychological climate research. The most important of these was a 'level of analysis' problem.

1.1 Distinction between Psychological Climate and Organizational Climate

James and Jones (1974) conducted a review of the climate literature and commented that, while organizational climate research had received a considerable amount of attention, the 'conceptual and operational definition, measurement techniques, and ensuing results' (James and Jones 1974 p. 1096) were highly diverse and contradictory. These problems were causing organizational climate to be viewed as a somewhat nebulous construct. James and Jones (1974) argued that previous climate research could be organized into three different, but not mutually exclusive, approaches to defining and measuring organizational climate. They designated these approaches as: (a) the multiple-measurement-organizational attribute approach,

which views climate as a set of organizational attributes measured by a variety of methods; (b) the perceptual measurement-organizational attribute approach, which uses only perceptual tools to measure climate but still sees climate as an organizational level variable; and (c) the perceptual measurement-individual attribute approach, which views climate as perceptual and an individual level phenomenon. James and Jones (1974) believed that researchers needed to differentiate between climate measured at the organizational level and at the individual level. They proposed that psychological climate was an individual attribute, whereas organizational climate was a situational attribute of the organization. Furthermore, by establishing these new definitions, researchers would be able to focus on either psychological climate or organizational climate which would, in turn, lead to research that would better be able to define both constructs and how they both related to other variables.

Later studies helped shape the boundaries between psychological climate and organizational climate (Schneider 1975). Psychological climate was defined as individual descriptions of organizational practices and procedures, which are useful in understanding the influence of the organizational environment on individual performance and satisfaction. Organizational climate, on the other hand, refers to a collective description of the organizational environment, and is thought to relate to organizational level outcomes. This distinction allowed psychological climate researchers to focus squarely on individual level issues, such as the relationship between psychological climate and various outcome variables (e.g., individual job performance). However, as the definition of psychological climate became somewhat solidified, researchers began to debate whether there was a distinction between psychological climate and job satisfaction. This debate was settled when researchers reached the consensus, through theoretical arguments (James and Jones 1974) as well as empirical research (Schneider and Snyder 1975), that psychological climate reflected descriptions of the work environment whereas job satisfaction reflected an affective evaluation of the work environment.

1.2 Hierarchical Model of Psychological Climate

As the definition of psychological climate became generally accepted, researchers began to expand the list of perceptual variables they used to measure individuals' descriptions of their work environment. However, research on the measurement properties of psychological climate also began to indicate that these psychological climate variables often clustered within four higher order factors: role, job, leadership, and workgroup (James and James 1992). While the exact

number and specific names of psychological climate variables clustered under these higher order factors changes depending on the scale used, the following examples should provide a general illustration of these higher order factors. The 'role' higher order factor is often made up of role ambiguity, role conflict, and role overload. The 'job' higher order factor often comprises job challenge, job autonomy, and job importance. The 'leader' higher order factor is most often made up of leader trust and support, leader goal facilitation, and leader hierarchical influence. Finally, the workgroup higher order factor often comprises work group cooperation, work group friendliness, and workgroup pride. Furthermore, this early research viewed these four psychological climate factors as 'distinct cognitive organizing principles for perceptual variables' (James and James 1989 p. 93).

However, James and James (1989) argued that there is a further hierarchical relationship among psychological climate dimensions. Specifically, they suggested that the judgment process of appraising organizational events in relation to one's personal values is common to each climate dimension. These value-based appraisals represent assessments of the degree to which organizational events can contribute to desired personal ends. Moreover, these judgments were hypothesized to reflect an underlying general factor (PCg), which represents an overall evaluation of whether the work environment is personally beneficial or detrimental to one's well-being. Research by James and James (1989) as well as others (e.g., Brown and Leigh 1996) has shown support for the idea of PCg. The PCg model, because it presents climate as one overarching construct, subsequently made examining the relationship between psychological climate and other variables much simpler.

2. Current Research

Current research falls into two categories: research that focuses on the relationship between psychological climate variables and other constructs (e.g., job performance); and research that focuses on the hierarchical structure of psychological climate.

2.1 Psychological Climate and Work-related Outcome Variables

Research has demonstrated that psychological climate perceptions provide an important mediating link between organizational characteristics and consequential work outcomes such as employee attitudes, motivation, psychological well-being, and performance. A recent meta-analysis (Parker et al. 1998) examined the relationships between individual-level higher order climate perceptions (e.g., role, job,

workgroup, leader) and work outcomes such as employee work attitudes, psychological well-being, motivation, and performance. It was found that individuals' climate perceptions had significant relationships with their job satisfaction, work attitudes, psychological well-being, motivation, and performance. Furthermore, each of the climate categories exhibited a somewhat different pattern of relationships with the various outcomes. For example, employees' job (e.g., importance, autonomy) and role (e.g., overload, clarity) perceptions appeared to have the weakest relationships with all of the outcome variables. Interestingly, the leader (e.g., goal emphasis, support) and work group (e.g., cooperation, warmth) constructs were most predictive of employees' work attitudes. A similar pattern was evident for the relationships of the various psychological climate dimensions with employee motivation and performance. For psychological well-being, however, job and leader perceptions provided the strongest relationships. Overall, it appears that employees' climate perceptions have stronger relationships with their work attitudes than with their motivation and performance. This pattern of relationships indicates that the effects of climate on motivation and performance may be mediated by employees' work attitudes. This hypothesis was tested in the study and supported. Thus, the relationship between employees' motivational and behavioral reactions to perceptions of their work environment and their motivation and performance does seem to be mediated by their overall evaluations of these perceptions.

Most studies that have considered the relationship between psychological climate variables and work-related outcome variables have assumed a linear relationship. Current research is investigating whether this is the case.

2.2 Nonlinear Relationships Between Psychological Climate and Outcome Variables

Researchers have debated over whether the relationships between certain climate dimensions (e.g., job autonomy) and outcome variables (e.g., job satisfaction) are linear or nonlinear in nature. For example, job autonomy is assumed by some researchers to follow an inverted U shape. That is, while medium levels of job autonomy are thought to be more desirable than low levels, very high levels of job autonomy are potentially harmful since it implies difficulty in decision making and high responsibility on the job. Other researchers, however, have assumed a linear relationship between job autonomy and job satisfaction: the more autonomy workers experience, the more satisfied they are with the job.

Previous studies examining psychological climate variables have shown support for both the linear and

nonlinear points of view. However, the nonlinear relationships found in prior research have been somewhat weak, often explaining only a small amount of additional variance (1–2 percent) over the linear model. In summary, the evidence is inconclusive as to which type of relationship exists.

However, research examining methodological problems associated with using a common response format may provide an explanation for these weak curvilinear findings. The use of a common response format has been shown to introduce shared item variance (i.e., common methods variance, or CMV) and elevate relationships among job perceptions (Glick et al. 1986). However, one can measure constructs using a multimethod approach. By measuring constructs with more than a single method one should be able to address the common response format problem mentioned above. A recent study (Baltes et al. 1999b) used both multitrait-multimethod (MTMM) data and monomethod data to test the relationships between climate dimensions that have been hypothesized to have nonlinear relationships with job satisfaction. Strikingly different results were found between the monomethod and MTMM data. Whereas the monomethod results mirrored earlier inconclusive findings, the MTMM data indicated that nonlinear equations explained significantly more of the relationship between all three climate dimensions and job satisfaction. These results suggest the following. First, the use of MTMM data allows one to test for nonlinear effects more effectively. Second, nonlinear quadratic effects better model the relationship between role clarity, role overload, job autonomy, and the outcome variable job satisfaction. These results, of course, have implications for both measurement and interpretation of psychological climate variables and work-related outcomes.

2.3 Hierarchical Structure of Psychological Climate

As mentioned earlier, recent climate research has used the James and James PCg model to justify collapsing multiple indicators of psychological climate into one overarching appraisal of the work environment (Brown and Leigh 1996, James and James 1992). However, it may be premature for psychological climate researchers to accept the construct validity of PCg because the James and James (1989) study relied on a single response format (i.e., traditional Likert scale). As mentioned earlier, the use of a common response format (i.e., monomethod data) has been shown to introduce shared item variance, which can elevate relationships among constructs. As a result, the intercorrelations among climate dimensions that generate support for PCg may be the result of nonrandom measurement errors caused by common methods variance. The hypothesis that support for

PCg is the result of common methods variance has received some support (Parker 1999). However, research efforts that use multiple measures of psychological climate are required to more appropriately test whether PCg is the result of common methods variance (James and James 1989). A study conducted by Baltes et al. (1999b) used multitrait-multimethod data collection to examine the hierarchical model of psychological climate (PCg). Two different self-report response formats were used to measure 12 psychological climate constructs. The PCg model, when tested with multitrait-multimethod data, did not fit the data. The lack of support for the PCg model appears to be caused by a strong common methods bias, which caused a 35 percent bias (i.e., inflation) in the observed correlations among the constructs. These results do not support the practice of collapsing psychological climate dimensions into an overarching construct. Furthermore, support was also not found for collapsing to the aforementioned four higher order constructs: role, job, leader, and workgroup. This result is perhaps the most troublesome given the extent of prior research that has considered the relationship between these higher order constructs and work-related outcome variables. These results are the product, of course, of one study and need to be replicated before previous research linking higher order psychological climate factors to outcome variables is considered obsolete.

3. Future Research

The results of current and past research will probably lead to three streams of future research. First, with respect to the relationship between psychological climate and various work outcomes, future research needs to test more complicated mediation and/or moderation models that include other work-related variables, such as employee work attitudes and/or other individual difference variables. By testing more complicated models researchers will be better able to discern how psychological climate fits into the workplace environment.

Second, if the relationships between these climate variables and job satisfaction are nonlinear then perhaps the important substantive question to ask employees is not the traditional one of how much they have of, for example, role clarity, but how what they have differs from what they would like. That is, the discrepancy between what they would like and what they have becomes the important question to ask since this discrepancy would indicate where they fall along the nonlinear curve. Knowing where an employee falls on the regression curve would allow one to predict what effect more or less of a climate dimension would have. Just knowing how much an employee has of, for example, job autonomy, would not allow a researcher

to effectively hypothesize what effect more or less job autonomy would have on the job satisfaction of that employee. Future research should not just ask the traditional question of how much an employee has of certain climate dimensions, but focus on the discrepancy between what they have and what they desire to better understand and predict outcome variables such as job satisfaction.

Finally, future research should assess psychological climate dimensions with other methods, such as essays or interviews, and test the PCg model. Whereas some current research, as mentioned earlier, has called into question the existence of a higher order psychological climate factor, more research with multiple measurement methods needs to be completed to address this issue more adequately.

See also: Organizational Climate

Bibliography

- Baltes B B, Bauer C, Altmann R 1999a Psychological climate and job satisfaction: Multitrait-multimethod evidence for nonlinear relationships. Paper presented at the 107th annual meeting of the American Psychological Association, Boston, MA, August 20–7
- Baltes B B, Parker C P, LaCost H, Altmann R, Huff J, Young S 1999b A Multitrait-Multimethod (MTMM) examination of hierarchical models of psychological climate. Poster presented at annual meeting of the Society for Industrial and Organizational Psychology, Atlanta, GA, 14th Annual Meeting, April 24–May 2
- Brown S P, Leigh T W 1996 A new look at psychological climate and its relationship to job involvement, effort, and performance. *Journal of Applied Psychology* **81**: 358–68
- Glick W H, Jenkins G D, Gupta N 1986 Method versus substance: How strong are underlying relationships between job characteristics and attitudinal outcomes? *Academy of Management Journal* **29**: 441–64
- James L R 1982 Aggregation bias estimates of perceptual agreement. *Journal of Applied Psychology* **67**: 219–29
- James L R, James L A 1989 Integrating work environment perceptions: Explorations into the measurement of meaning. *Journal of Applied Psychology* **74**: 739–51
- James L R, James L A 1992 Psychological climate and affect: Test of a hierarchical dynamic model. In: Cranny C J, Smith C, Stone E F (eds.) *Job Satisfaction: How People Feel About Their Jobs and How it Affects Their Performance*. Lexington Books, New York, pp. 89–117
- James L R, James L A, Ashe D K 1990 The meaning of organizations: The role of cognition and values. In: Schneider B (ed.) *Organizational Climate and Culture*. Jossey-Bass, San Francisco, CA, pp. 40–89
- James L R, Jones A P 1974 Organizational climate: A review of theory and research. *Psychological Bulletin* **81**: 1096–112
- Lewin K, Lippitt R, White R K 1939 Patterns of aggressive behavior in experimentally created 'social climates.' *Journal of Social Psychology* **10**: 271–99
- Litwin G H, Stringer R A 1968 *Motivation and Organizational Climate*. Harvard University Press, Cambridge, MA
- Parker C F 1999 A test of alternative hierarchical models of

- psychological climate: PCg, satisfaction, or common method variance? *Organizational Research Methods* 2: 257–74
- Parker C F, Young S, Baltes B B, Altmann R, Huff J, Lacost H 1998 A meta-analysis of relationships between climate perceptions and work outcomes. Paper presented at the annual meeting of the Society for Industrial and Organizational Psychology, Dallas, TX, 13th Annual Conference, April 23–26
- Schneider B 1975 Organizational climates: An essay. *Personnel Psychology* 28: 447–79
- Schneider B, Snyder R A 1975 Some relationships between job satisfaction and organizational climate. *Journal of Applied Psychology* 60: 318–28

B. B. Baltes

Psychological Development: Ethological and Evolutionary Approaches

1. Introduction: The Science of Animal Motives

Classical ethology, the ‘biology of animal behavior’ (Eibl-Eibesfeldt 1970, Hinde 1982), compares descriptions of how different species act in their natural environment, and in families and societies. Lorenz proposed that genetically determined motor programs or Innate Release Mechanisms (IRMs), triggered by environmental Sign Stimuli, confer ‘dispositions to learn’ that benefit the species. By comparing spontaneous behavior rituals and sequences of display or response between individuals, ethologists seek evidence of innate impulses in different species, and determine the genetic relationships and evolutionary lineage (see *Comparative Method, in Evolutionary Studies*). Ethology developed from zoology and field studies of wild animals. It accepts that animals have motives and awareness that can be understood by accurate description of their behaviors in nature, and how these develop. Darwin’s *Expression of the Emotions in Man and Animals* (1872) inaugurated ethological science (see *Darwin, Charles Robert (1809–82)*). He related human feelings to those one can infer in animals from their expressions.

Tinbergen (1951) identified innate signaling behaviors of communication—in mating, nest building, care of young, territorial defence, hunting or escape from predators, sharing environmental resources, etc. (see *Evolutionary Social Psychology*). He defined four time-scales of behavioral determination: over evolutionary time; in development of the individual; in control of behavioral tendencies over hours, days, or months; and in adaptive functional response to immediate environmental events. He used field experiments to identify that stimuli or behavior signals are ‘releasers’ of innate activities.

Scientists research animal senses, often very different from ours, and the ways animals move to

investigate their environment. Communicative signals are instructive, because they have evolved to be perceived. Rhythmic light flashes, visible movements of body parts, vocal or other sounds, or vibrations, convey effective messages, even in primitive species. How animals react to signal patterns gives further evidence about internal processes that guide attention, movements, and emotions. In humans, innate hearing and vision has been explored by studying the vocal sounds, face expressions, and body movements that infants prefer, and the emotions infants express (Fernald 1992, Papousek and Papousek 1987, Stern 2000, Trehub 1990, Trevarthen 1998, 1999).

As in Darwin’s day, evidence for species-specific motives and feelings in humans comes from comparing spontaneous expressions of people in different cultures, and by observing infants before they have had the chance to learn (Eibl-Eibesfeldt 1989). Newborn infants communicate and imitate, and children show systematic changes in behaviors and interests (Stern 2000, Trevarthen 1998). As children’s motives develop, they change the behavior of caregivers, stimulating ‘intuitive parenting’ and the sharing of meanings (Papousek and Papousek 1987).

2. Learning, Cognition, and Communication

Efforts have been made to relate ethological theory to concepts of experimental psychology and ‘conditioning’ in simplified laboratory situations (Hinde 1982). After decades of rejection by psychologists of ‘mentalist’ interpretations, the idea that animals may share motives, feelings, and conscious experiences comparable to those of human beings has again become scientifically acceptable (see *Human Cognition, Evolution of; Evolution of Cognition: An Adaptationist Perspective*). Psychologists postulate internal ‘representations’ that process perceptual information and control actions of subjects in complex circumstances. More recently, attention has turned to the evolved functions of emotional regulatory systems in the brain, and their role in self-awareness and intelligence (Damasio 1994) (see *Emotions, Evolution of*).

Cognitive psychology acknowledges that human behavior is regulated by conscious states and intentions, and that learning integrates emotions and mental representations. Nevertheless, cognitive science remains individualistic and rational—focused on reality. Complex intelligence is analyzed in terms of perceptual categorizations and hierarchical representational systems. Different modes of intelligence are thought to be formulated anatomically in identifiable brain regions. In particular, social intelligence is considered to pass, in evolution and development, through grades of cognitive representation, and levels of awareness of cognition in other individuals—from immediate, ‘thoughtless’ sensorimotor reactions, to verbalizable, rationally formulated ‘theories of mind’

(see *Evolution of Cognition: An Adaptationist Perspective*). Dynamic Systems Theory, opposed to linear causal explanation, attributes behavior patterns and mental representations to 'emergent properties' thrown up in complex systems comprising large numbers of elements—neural connections, perceptual feature detectors, sensorimotor loops, motoric forces, representations in memory, social transactions, etc. In neither of the dominant contemporary theories of complex mental life and behavioral coordination is evolved anticipatory motivation given a determinant role.

No adaptive behavior of an animal subject is without innate 'constraints' due to prefunctional configuration (morphogenesis) of sensory and motor structures, brain tissue, and nerve connections, which map both the body and its future engagement with the outside world in a coherent, subjective space and time. What is innate is an active neurobiological initiative that investigates the environment, applying specific anticipatory, probing processes of behavior. 'Critical periods,' or ages in which active investigations of experience change, correspond to brain growth periods that are generally more open to learning (see *Lifespan Development: Evolutionary Perspectives*).

3. Behavior from Dyads to Large Social Groups

Subjective motives communicated to other individuals become intersubjective (Bråten 1998). In the more intelligent, behaviorally complex species there is a rich capacity for display of mental impulses—of attention, intention, and emotion. Intersubjective relations assume central importance in reproduction, feeding, establishment of cooperative groups in families or much larger societies. Interspecific collaborations, including predator–prey interactions, control behavior in dyads and larger groups, regulating ecological relations between animals psychologically.

Sociobiology has formulated theories to explain the evolutionary advantages of various forms of conflict and cooperation within and between communities, and special complementary adaptations of individuals in the service of the group, or to propagate a genetic type. Reductive cost-benefit analyses, of energy balance sheets for motor and metabolic investment, or of transmission and survival of genes in populations (reproductive success), have prevailed, which can be explained by an unwillingness of quantitative biologists to attribute self-regulatory motives and emotions to animals as individuals, or members of families and cooperative groups (see *Evolution, Natural and Social: Philosophical Aspects*). They see behaving animals as driven by forces of competition for future success that are indifferent to any aspirations for, or pleasures in, life in the present. The genetic history of humans indicates that there has been rapid adaptation to an

extremely large range of environments, and new adaptability by cooperative learning of behavioral traditions (Cavalli-Sforza and Cavalli-Sforza 1995) (see *Genes and Culture, Coevolution of; Cultural Evolution: Overview*).

The benefits of competition have been exaggerated. Research on the psychopathology of infanticide or mass killing by individuals indicates that direct competition by assault results from a loss of pleasure in the kinds of altruistic and affectionate motives of play that lead to creative invention of new and beneficial motives, ideas, or activities in a group (Bekoff and Byers 1998). Communication between members of a species and between differently adapted species without concern for gain leads to intricate dynamic ecological systems and mutual dependency. Some behaviors of predator and prey in natural ecological equilibrium are cooperative, or mutually supportive. Skilled collective understanding of the environment by intelligent social species depends on positive motives for sharing enjoyment of experience and skilled action, such as are abundantly shown by the young of many species when they play together.

4. Attachment, Companionship, and Parental Support of Early Development

Emotions linked to innate behaviors regulate parent–offspring attachments and caregiving (see *Attachment Theory: Psychological*). In many species, mutual adaptations of motives in parents and young facilitate cooperative encounters that nurture the young, and parenting strategies limit the offspring's dependency. The food, care, and support that nestlings or nurslings receive sustain the developing body and brain. Human infants show attachment to the physiological or bodily benefits from mothering (Schore 1994), the biological importance of which was first demonstrated experimentally in monkeys. Infants also respond to psychological expressions of pleasure and interest carried by the voice, facial expressions, and gestures of the parent, and these lead to 'proto-conversations' and companionship (Trevvarthen 1998, 2001b; Trevvarthen and Aitken 2000). Infants and parents seek contact and affective communication, and deprivation of sympathetic parenting has detrimental effects on both emotional and cognitive development of a child.

Pleasure and pain may be the main emotions or neurochemical systems seeking parental nurturance in mammals, including humans (Panksepp 1998), but more complex emotions, including self–other conscious states and relational emotions such as pride and shame, regulate intersubjective motives and assist the learning of a wide range of socially and culturally important skills in infants (Stern 2000, Draghi-Lorenz

et al. 2001) (see *Self-conscious Emotions, Psychology of*). Infants seek emotional confirmation to develop knowledge and skills (Trevvarthen and Aitken 2000).

5. *Motives and Benefits of Play*

Animal motives are revealed not only in behaviors that immediately benefit the individual, or the propagation and flourishing of the group or species. Their essential creative aspect is vivid in play (Bekoff and Byers 1998). Ethological and psychological theories of play appear confused by a desire to find a function outside motivation itself. However, by its nature, and by any definition, play behavior largely is determined by the anticipatory processes of behavioral control, and not any immediate adaptive consequence. Play behaviors are more apparent in the more social species that have elaborate intersubjective capacities or social awareness. By giving strong and continuous messages of exuberant motives, play invites sympathetic attention and, indeed, participation by others in changing outcomes and risks. Human children, like young of other species, can play alone, practising and shaping perceptuo-motor skills. But behaviors exhibited in self-satisfying play have a potential to be social, and they are responsive to provocative or 'teasing' interactions with other individuals. Play fighting between young in a cooperative family group is affiliative, consolidating relations and training awareness of mutual intentions. Many play behaviors anticipate functions of mature 'innate' behavior patterns. Thus, play expresses motives for behaviors that require exercise of motor and cognitive processes for their maturation. It would seem pointless to attempt a cognitive rationalization of play if play is the prior behavior, adapted to generate intelligence and regulate cognition.

Research with infants confirms that play develops relationships and skills, as well as creating rituals and 'formats' of cooperation. The imaginative or 'symbolic' play of young children foreshadows the creative efforts made by young adult athletes, and the striving of outstanding craftsmen, artists and scientists throughout their lives (Bekoff and Byers 1998). Art, appreciation of beauty in artefacts and ceremony in human communities may be related to universal rhythms and modes of intimate communication between infants and their partners (Dissanayake 2000).

6. *'Communicative Musicality' and the Pulse of Human Thought and Imagination*

Animal communication employs precisely timed rhythms, modulations of expressive activity, and accurate interaction of movements in communicating individuals (Tinbergen 1951). Infants' expressions,

and their refined discrimination of temporal, prosodic, harmonic, and melodic features of vocalization or musical sounds, beginning before birth, indicate that human language communication draws on rich pre-verbal talents for expressive gesture and perception (Trehub 1990). Human sensitivity for the emotive rhythms, qualities, and narrative sequences of expressive song and dance, may be evolved with bipedal standing and walking, which free the erect body, and especially the hands and voice, for flexible poly-rhythmic gestures that can richly express dynamic emotions and their blends and changes in interpersonal engagements (Trevvarthen 1999). Emotions are conveyed in regulated 'contours' or 'melodies' of muscle activity. They are held together by a coherent time-generating process in the brain that creates a hierarchy of acts and succession of perceptual events in one 'mind time' (Pöppel 1994). Infants discriminate the pulse, expressive quality, narrative succession, and contingency of response in adult expressions.

7. *The Biology of Human Imitation, Cultural Learning, and Language*

Monkeys and apes learn in social play, and imitate, but parental support for learning is limited in sub-humans, and the most elaborate monkey play is between peers. Apes play over a wider range of ages, and their play assists development of family and group bonds, and they imitate how adults forage and use transformed objects as 'tools' for getting food. No apes show the intricate intersubjective exchanges of human vocal and gestural play with infants soon after birth, or the constructive and narrative-making fantasy games of toddlers. Donald (1991) argues from comparative cognitive psychology and the fossil record that it is likely that the first bipedal hominids were communicating nonverbally, by gestural mimetic narratives, about nonpresent experiences (see *Cultural Evolution: Theory and Models*). He concludes that grammar of language evolved out of the rhythmic syntax of gestural mimesis (see *Evolution and Language: Overview*). Study of developments towards language of infants and toddlers supports this theory (Locke 1993).

Human beings have unique motives for cultural learning—the acquisition of artificial knowledge and skills. These govern progress from preverbal forms of communication and play (Trevvarthen 1998). Human ethology is necessarily a science concerned with these motives. All cultural cooperation among humans, however artificial, depends on a uniquely productive capacity for 'mirroring' feelings, interest, and purposes of other persons, and for solving cognitive, practical, moral, and artistic problems collaboratively (Bruner 1996). Studies of imitation reveal that a neonate human has a remarkably complete image of move-

ments in the expressive parts of another person's body, an image that is both receptive through the constellation of all senses, and capable of forming a matching movement in communicative interactions (Bråten 1998, Nadel and Butterworth 1999).

Human language learning depends upon joint attention and sympathetic understanding of intentions (Tomasello and Farrar 1986). Toddlers acquire use of words by following convergent interest to grasp meanings in other persons' speech (Locke 1993). Long before verbal communication, an infant is proficient in 'protoconversational' games, and in participation in rhythmic rituals of action games and songs, which parents are highly motivated to share, intuitively matching and complementing the emotions expressed by the infant (Trevvarthen 1998).

8. The Evolutionary Neuroembryology of Motives and Emotions

Discoveries in neurobiology enhance our knowledge of adaptive motive systems and emotions and their role in self awareness and communication (Schoe 1994). Brain mechanisms of emotion in brainstem regions influence the higher functions of imagination, thought, and language (Panksepp 1998) and exercise an influence over the maturation of cortical cognitive systems in both prenatal and postnatal development of mammals, including human beings (Trevvarthen 2001a). Cortical mechanisms for perceiving postural, facial, and vocal expressions, as well as manual gestures, for remembering experiences associated with different expressed emotions, and for generating such expressions as elaborate narratives of internal impulses and feelings, are enlarged greatly in humans as compared to homologous brain areas of other species (Damasio 1994, Trevvarthen 2001a). Play is associated with manifestations of neurochemical well-being or 'happiness' (Panksepp 1998). Often young animals play vigorously with body movements at a particular age, and it has been noted that in this 'sensitive period' the cerebellum of the hind brain is undergoing accelerated growth and acquiring refinements of circuitry necessary for accurate and rapid motor control.

9. Future Directions

Only recently has over two centuries of research on the comparative anatomy of brains and their development been comprehensively integrated with psychological theories of cognition and learning. A focus on the forebrain cortex, and the effects of cortical lesions on human awareness, memory, and motor skills had left the core systems that incorporate motivational and emotional controls out of the picture. Recognition

that acquired skills and knowledge are directed, focused, selected, and integrated by subcortical systems leads to questions about how adaptations of behavior and awareness that anticipate experience and shape learning are generated. Ethological theory directly supports new enterprises in human genetics, neuroethology, and developmental neuroscience, and applied fields, such as education and psychiatry. We anticipate a more balanced natural science of mental life integrating the processing of empirical experience and reason with inherited emotive impulses for both individual experience and cooperative social life. Important questions will arise concerning what determines human affairs and our relation to the natural world.

See also: Animal Cognition; Behavioral Genetics; Psychological Perspectives; Comparative Psychology; Primate Behavior: Significance for Understanding Humans; Psychological Development: Ethological and Evolutionary Approaches; Sensitive Periods in Development, Neural Basis of

Bibliography

- Bekoff M, Byers J A 1998 *Animal Play: Evolutionary, Comparative and Ecological Approaches*. Cambridge University Press, New York
- Bråten S (ed.) 1998 *Intersubjective Communication and Emotion in Early Ontogeny*. Cambridge University Press, Cambridge, UK
- Bruner J S 1996 *The Culture of Education*. Harvard University Press, Cambridge, MA
- Cavalli-Sforza L L, Cavalli-Sforza F 1995 *The Great Human Diasporas: The History of Diversity and Evolution*. Addison-Wesley, Reading, NY
- Damasio A R 1994 *Descartes Error: Emotion, Reason and the Human Brain*. Putnam, New York
- Dissanayake E 2000 *Art and Intimacy: How the Arts Began*. University of Washington Press, Seattle and London
- Donald M 1991 *Origins of the Modern Mind*. Harvard University Press, Cambridge, MA
- Draghi-Lorenz R, Reddy V, Costall A 2001 Rethinking the development of 'non-basic' emotions: A critical review of existing theories. *Developmental Review* (in press)
- Eibl-Eibesfeldt I 1970 *Ethology: The Biology of Behavior*. Holt, Rinehart and Winston, New York
- Eibl-Eibesfeldt I 1989 *Human Ethology*. Aldine de Gruyter, New York
- Fernald A 1992 Human maternal vocalizations to infants as biologically relevant signals: An evolutionary perspective. In: Barkow J H, Cosmides L, Tooby J (eds.) *The Adapted Mind: Evolutionary Psychology and the Generation of Culture*. Oxford University Press, Oxford, UK
- Hinde R A 1982 *Ethology: Its Nature and Relation with Other Sciences*. Fontana, London
- Locke J L 1993 *The Child's Path to Spoken Language*. Harvard University Press, Cambridge, MA and London
- Nadel J, Butterworth G (eds.) 1999 *Imitation in Infancy*. Cambridge University Press, Cambridge, UK

- Panksepp J 1998 *Affective Neuroscience: The Foundations of Human and Animal Emotions*. Oxford University Press, New York
- Papousek H, Papousek M 1987 Intuitive parenting: A dialectic counterpart to the infant's integrative competence. In: Osofsky J D (ed.) *Handbook of Infant Development*, 2nd edn. Wiley, New York
- Pöppel E 1994 Temporal mechanisms in perception. *International Review of Neurobiology* 37: 185–202
- Schore A N 1994 *Affect Regulation and the Origin of the Self: The Neurobiology of Emotional Development*. Erlbaum, Hillsdale, NJ
- Stern D N 2000 *The Interpersonal World of the Infant: A View from Psychoanalysis and Development Psychology*, 2nd edn. with new Intro. Basic Books, New York
- Tinbergen N 1951 *The Study of Instinct*. Clarendon Press, Oxford, UK
- Tomasello M, Farrar M J 1986 Joint attention and early language. *Child Development* 57: 1454–63
- Trehub S E 1990 The perception of musical patterns by human infants: The provision of similar patterns by their parents. In: Berkley M A, Stebbins W C (eds.) *Comparative Perception: Vol. 1, Mechanisms*. Wiley, New York
- Trevarthen C 1998 The concept and foundations of infant intersubjectivity. In: Bråten S (ed.) *Intersubjective Communication and Emotion in Early Ontogeny*. Cambridge University Press, Cambridge, UK
- Trevarthen C 1999 Musicality and the intrinsic motive pulse: Evidence from human psychobiology and infant communication. *Musicae Scientiae*, Special Issue (1999–2000): 157–213
- Trevarthen C 2001a The neurobiology of early communication: Intersubjective regulations in human brain development. In: Kalverboer A F, Gramsbergen A (eds.) *Handbook on Brain and Behavior in Human Development*. Kluwer Academic Publishers, Dordrecht, The Netherlands
- Trevarthen C 2001b Intrinsic motives for companionship in understanding: Their origin, development, and significance for infant mental health. *Infant Mental Health Journal* 22(1–2): 95–131
- Trevarthen C, Aitken K 2000 Intersubjective foundations in human psychological development. *Journal of Child Psychology and Psychiatry* 42(1): 3–48

C. Trevarthen

Psychological Therapies: Emotional Processing

It is now well established that systematic exposure to phobic stimuli markedly reduces associated fear and several theories about the psychological mechanisms underlying the efficacy of exposure therapy have been advanced (Craske 1999). Emotional processing theory (e.g., Foa and Kozak 1986) initially was advanced as one explanation of the effectiveness of exposure therapy in the treatment of anxiety disorders. This article traces the origin of the concept of emotional processing, describes emotional processing theory and how it accounts for the effectiveness of psychological

therapies in treating anxiety disorders, and discusses recent developments in the concept of emotional processing.

1. History of the Emotional Processing Concept

The concept of emotional processing has its origin in Lang's (1977) analyses of fear-relevant imagery in the context of behavior therapy and fear reduction. In studying the procedure of systematic desensitization, Lang et al. (1970) found three predictors of successful treatment: greater initial heart-rate reactivity during fear-relevant imagery, greater concordance between self-reported distress and heart rate elevation during fear-relevant imagery, and systematic decline in heart-rate reactivity with repetition of the imagery. On the basis of these findings, Lang (1977, p. 863) suggested that 'the psychophysiological structure of imagined scenes may be a key to the emotional processing which the therapy is designed to accomplish'. While Lang did not expand on the concept of emotional processing, in subsequent papers he further elaborated on the theory of emotional imagery in behavior therapy (e.g., Lang 1984).

Lang's (1977, 1984) theory holds that a fear image is a cognitive structure containing stimulus, response, and meaning information that serves as a program to avoid or escape from danger. For example (see Fig. 1), the fear structure of a dog phobic includes representations of dogs, representations of physiological and behavioral responses in the presence of dogs (e.g., rapid heart-rate, sweating, running away), and representations of meanings associated with dogs (e.g., 'Dogs are dangerous') and with responses to dogs (e.g., 'Rapid heart rate and sweating mean that I am afraid'). The fear structure is activated by environmental input that matches some of the information stored in the structure. According to Lang (1977, p. 867), 'the aim of therapy could be described as the reorganization of the image unit in a way that modifies the affective character of its response elements.' He further proposed that for image modification to occur, the fear structure must be at least partially activated. Accordingly, it is this aspect of the theory that accounts for the relationship between physiological measures (e.g., elevated heart-rate reactivity) and good outcome in exposure therapy for fear. However, Lang did not specify how activation of the fear structure brings about alterations in the structure.

The concept of emotional processing was subsequently taken up by Rachman (1980, p. 51) who proposed a working definition of emotional processing as 'a process whereby emotional disturbances are absorbed, and decline to the extent that other experiences and behaviors can proceed without disruption.' Accordingly, successful processing would be evidenced by a person's ability to confront a previously distressing event without evidencing signs of emotional distress. By contrast, several clinical phenomena

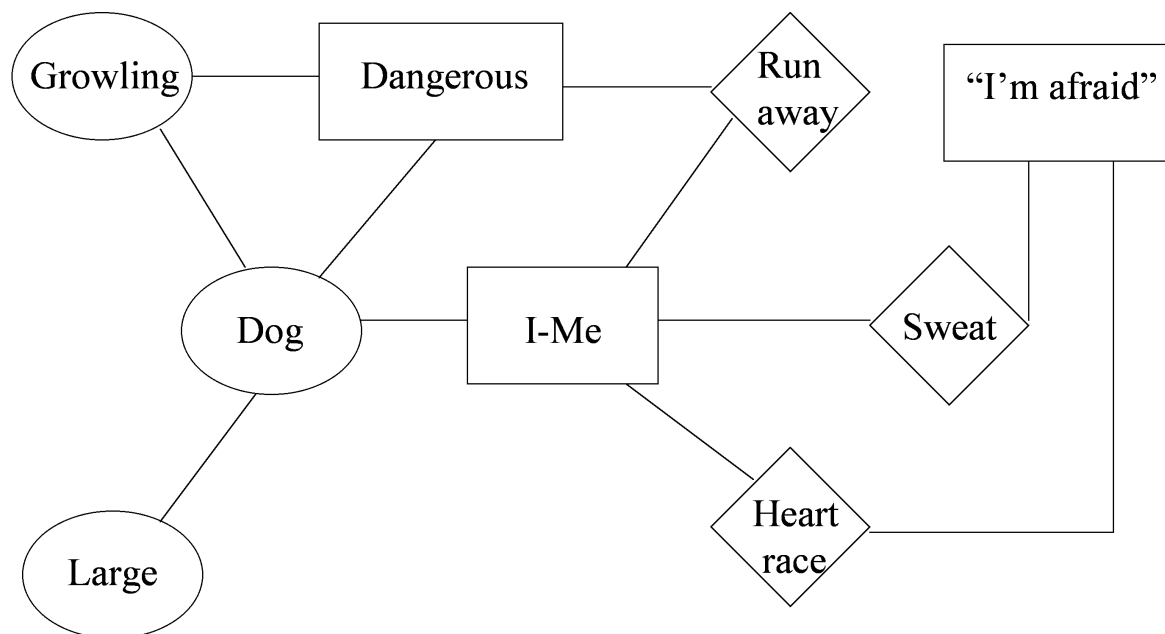


Figure 1

A schematic diagram of the fear structure of an individual with a dog phobia. The ovals indicate stimulus representations, diamonds indicate response representations, and rectangles indicate meaning representations. Lines indicate associations among representations

such as persistent fears and obsessions, failures to benefit from treatment designed to reduce fears, and the return of fear after apparently ‘successful’ treatment would be evidence of failed or incomplete emotional processing.

Rachman (1980) noted that the majority of people are able successfully to process the upsetting events that occur in their lives and proposed a number of situational, personality, and contextual variables that are likely to produce difficulties in processing an emotionally disturbing event. For example, difficulties in emotional processing are likely to occur in generally anxious and introverted individuals who confront intense, dangerous, unpredictable, and uncontrollable events. Rachman further discussed factors that can facilitate emotional processing, such as prolonged exposures to fear relevant stimuli or images without distraction. Rachman’s analysis does not constitute a theory of emotional processing. Rather, it is an attempt to describe the sort of phenomena for which an adequate theory of emotional processing may provide an explanation. Indeed, as noted by Foa and Kozak (1986), Rachman’s approach to defining and analyzing emotional processing suffers from circular reasoning: the fear reduction that is attributed to successful emotional processing is the same evidence used to infer that successful emotional processing has occurred.

2. *Emotional Processing Theory: Basic Concepts and Mechanisms Underlying Treatment*

2.1 *The Cognitive Structure of Fear*

Building on Lang’s (1977, 1984) theory, Foa and Kozak (1986) conceptualized fear as a memory structure that serves as a blue print for escaping or avoiding danger that contains information about the feared stimuli, fear responses, and the meaning of these stimuli and responses. When a person is faced with a realistically threatening situation (e.g., a car accelerating at you) the fear structure supports adaptive behavior (e.g., swerving away). However, a fear structure becomes pathological when the associations among stimulus, response, and meaning representations do not accurately reflect reality in that harmless stimuli or responses are erroneously interpreted as being dangerous. Furthermore, characteristic differences between different anxiety disorders are thought to reflect the operation of specific pathological fear structures (cf. Foa and Jaycox 1999). For example, the fear structure of individuals with panic disorder with agoraphobia is characterized by fear of physical sensations associated with their panic symptoms (e.g., rapid breathing, heart palpitations) and interpretation of these sensations as meaning that they are having a heart attack or are going crazy. As a result of this

misinterpretation, panic disordered individuals avoid locations they anticipate will give rise to panic attacks or similar bodily sensations, such as physical exertion.

By contrast, the fear structure of individuals with posttraumatic stress disorder (PTSD) involves associations among trauma-related harmful stimuli and similar but non-harmful stimuli, erroneous interpretation of safe stimuli as dangerous (e.g., 'All men are rapists'), and erroneous interpretations of responses to mean self-incompetence (e.g., 'My symptoms mean I can't cope with this'). Accordingly, when encountering a trauma reminder, the PTSD sufferer experiences strong emotional reactions similar to those that occurred during the trauma. Support for these hypotheses comes from a study comparing traumatized individuals with and without PTSD to one another and to nontraumatized individuals using the Posttraumatic Cognitions Inventory (PTCI), a measure of negative cognitions about the self, negative cognitions about the world, and self blame. Individuals with PTSD showed higher scores on all three scales than did the other two groups who did not differ from one another (Foa et al. 1999).

2.2 Emotional Processing as Modification of the Fear Structure

If fear behavior reflects the activation of an underlying cognitive fear structure, then changes in the fear structure will result in corresponding changes in behavior. Foa and Kozak (1986) proposed that psychological interventions known to reduce fear, such as exposure therapy, achieve their effects through modifying the fear structure. They further suggested that, 'The hypothesized changes in the fear structure may be conceptualized as the mechanism for what Rachman (1980) has defined as emotional processing' (Foa and Kozak 1986, p. 22).

2.3 Necessary Conditions for Emotional Processing

According to emotional processing theory (Foa and Kozak 1986), two conditions are necessary for therapeutic fear-reduction to occur. Similar to Lang (1977, 1984), Foa and Kozak suggested the fear structure must be activated in order for it to be available for modification. Activation occurs when a person is exposed to information that matches some of the information stored in the fear structure. But activation by itself is not sufficient. New information that is *incompatible* with the pathological aspects of the fear structure must also be available and incorporated into the existing structure for emotional processing to occur. Indeed, exposure to information that is *consistent* with the existing fear structure would be expected to strengthen existing associations. Although experiences may have the effect of either strengthening

or weakening existing fear structures, the term emotional processing is reserved for modifications that result in fear-reduction.

2.4 Indicators of Emotional Processing and Relation to Treatment Outcome

Fear and emotional processing are theoretical constructs that cannot be observed directly. Rather, they must be inferred through observable variables. Because emotional processing is being offered as an *explanation* of fear reduction, it is important that the indicators of emotional processing be different from the indicators of fear reduction. To avoid circularity, Foa and Kozak (1986) distinguished between measures of emotional processing during the course of exposure therapy and treatment outcome. Specifically, they proposed that emotional processing can be inferred from: (a) self-reports (e.g., 'I'm afraid') and physiological reactivity (e.g., elevated heart rate) that are indicative of fear activation; (b) gradual decreases of these measures during continued exposure to fear-relevant information; and (c) diminished initial fear reactions across successive exposures. In contrast to these indicators, Foa and Kozak viewed treatment outcome as a much broader notion that encompasses variables not specifically targeted in treatment, such as the quality of a person's job performance, personal relationships, and their degree of life satisfaction.

Several studies are consistent with the hypothesized relationship between indicators of emotional processing and treatment outcome. For example, individuals with PTSD experience distressing intrusive thoughts about their traumatic experience; carefully avoid people, places, and activities that remind them of the traumatic event; and experience disruptions in memory, concentration, and sleep. An effective treatment for PTSD is imaginal exposure to memories of the traumatic experience (Shalev et al. 1996). It has been found that individuals who exhibited fear activation during imaginal reliving, followed by decline across sessions achieved better outcome (e.g., fewer intrusive thoughts, less avoidance, and improvements in memory, concentration, and sleep) than individuals who did not exhibit fear activation or who exhibited fear activation with minimal habituation across sessions (cf. Foa 1997). The validity of these predictors of outcome has been demonstrated in other anxiety disorders as well, such as obsessive-compulsive disorder and specific phobias (Foa and Kozak 1986).

2.5 Sources of Corrective Information

It is easy to see how exposure therapy, in which fearful individuals are helped to confront feared images or situations, results in activation of a fear structure. But what kind of information does exposure therapy

provide to correct erroneous associations among the stimulus, response, and meaning representations in the fear structure?

One important source of corrective information is *habituation*, by which repeated exposure to a stimulus that reliably elicits a specific response results in a decline in that response. Foa and Kozak (1986) proposed that habituation of physiological responses within an exposure therapy session, one of the indicators of emotional processing, constitutes an important source of information for modifying aspects of the fear structure. For example, habituation generates new response information about the *absence* of physiological responding in the presence of the phobic stimulus that is incompatible with prior response information. Physiological habituation also counters erroneous meaning representations underlying the cognition that fear will persist until escape occurs. Repeated and prolonged exposure may also provide corrective information as to the realistic likelihood of feared consequences. For example, the fear structure of an individual with obsessive compulsive disorder may include meaning representations with unrealistically high estimations of the probability of contracting a severe illness from touching doorknobs. As a result, doorknobs are carefully avoided thereby preventing the person from learning that touching doorknobs does not pose a serious health risk. Treatment through exposure (e.g., touching dirty doorknobs) and response prevention (e.g., refraining from washing) provides the corrective information that the probability of becoming sick after touching a doorknob is actually extremely low.

3. Recent Advances in Emotional Processing Theory

3.1 Mechanisms Underlying Natural Recovery from Trauma

Studies investigating patterns of reactions to a trauma over time indicate that individuals differ in their ability to successfully process the traumatic event and thus recover naturally with the passage of time. While most trauma victims process the trauma successfully, a substantial minority fails to do so and consequently develops chronic PTSD (Foa 1997). Why do some traumatized individuals recover and others go on to develop chronic disturbances? Foa (1997) suggested that three common mechanisms underlie both natural recovery from trauma and reduction of PTSD severity via treatment. The first mechanism is *emotional engagement* with the trauma memory, which in the context of exposure therapy has been termed fear activation. In relation to natural recovery, emotional engagement refers to fear activation that occurs upon encountering a trauma reminder in the natural en-

vironment. Just as fear activation has been found positively associated with treatment outcome, the absence of emotional engagement has been found to be associated with poor natural recovery following a traumatic event. For example, individuals who *dissociate* during or shortly after experiencing a traumatic event experience greater PTSD symptom severity later on, compared to those who do not dissociate. The dissociation phenomenon connotes the absence of emotional engagement with the trauma memory and hence, according to emotional processing theory, will retard emotional processing. A delay in the peak reaction to a traumatic event may also be conceptualized as the absence of emotional engagement and thus is expected to hinder emotional processing and subsequent recovery. Consistent with this hypothesis, individuals whose peak symptom severity occurred within two weeks of the traumatic event were less symptomatic 14 weeks later than were individuals whose symptoms peaked between 2–6 weeks after the event. A second mechanism hypothesized to underlie both natural recovery and improvement with treatment is the *degree of organization or articulation* of the trauma narratives (Foa 1997). In support of this hypothesis, a higher degree of narrative articulation shortly after the trauma predicted greater recovery (lower PTSD symptom severity) three months later. In the same vein, exposure therapy for PTSD was associated with increased organization and decreased disorganization of the trauma narrative over the course of treatment, and those changes were correlated with symptom improvement.

The third hypothesized mechanism for recovery is a *change in core cognitions* about the world and the self. As discussed earlier, two basic dysfunctional cognitions are thought to underlie PTSD: the world is completely dangerous and the self is totally incompetent. It follows that recovery would be associated with less negative cognitions (Foa and Jaycox 1999). Indeed, in a longitudinal study of crime victims, individuals who retained their PTSD diagnosis three months after their assault exhibited more negative cognitions compared to those who no longer had PTSD. Similarly, cognitions about the world and self were less negative after treatment for PTSD than they had been before treatment (see Foa 1997, for discussion of the preceding two studies).

The ways in which exposure therapy is thought to promote fear activation and modification of erroneous cognitions have been explicated earlier. But how does emotional processing theory account for natural recovery? It is proposed that natural emotional processing and the consequent recovery occurs via repeated activation of the trauma memory in daily life through engagement with trauma-related thoughts and feelings, sharing them with others, and confronting trauma-reminder stimuli. In the absence of repeated traumatization or 'emotional breakdown,' these natural exposures contain information that disconfirms the

common post-trauma cognition that the world is a dangerous place. For example, a rape victim may repeatedly encounter men that are physically similar to the perpetrator, thus activating the trauma memory structure. As these encounters do not lead to additional assaults, the natural initial postrape inclination to view the world as entirely dangerous is not confirmed and thus gradually subsides. In addition, repeated thinking or talking about the traumatic event promotes habitation of negative emotions and the generation of an organized trauma narrative. By contrast, individuals who avoid engaging with the traumatic memories by either suppressing related thoughts and feelings or avoiding trauma-related situations, do not have the opportunity to form an organized narrative or disconfirm the initial post-trauma negative cognitions and thus develop chronic disturbances. Indeed, mothers who lost their babies to Sudden Infant Death Syndrome and were able to share their loss with a supportive social network resolved their grief more satisfactory 18 months after the trauma than did mothers who were not able to share their grief with supportive others (Lepore et al. 1996).

3.2 Emotional Numbing and Emotional Processing in PTSD

In an apparent paradox, individuals with PTSD frequently report experiencing intense negative emotions when reminded about the traumatic (re-experiencing) while also having difficulty experiencing and expressing emotions in situations that call for an emotional response (emotional numbing). To explain this paradox, Litz et al. (2000) hypothesized that the strong emotional reactions cued by trauma reminders among individuals with PTSD temporarily depletes their emotional resources, thereby reducing emotional responsiveness to subsequent events. Consistent with this hypothesis, veterans with PTSD who were first shown a video tape of combat images and then asked to rate a series of slides depicting emotionally positive images (e.g., a baby) were emotionally more reactive to the combat video, but less reactive to the positive slides than veterans without PTSD. In contrast, no differences between groups emerged when the veterans viewed neutral images (e.g., furniture) before rating the positive slides. Thus, veterans with PTSD displayed emotional numbing for the positive slides, but only after first experiencing negative emotional arousal.

The research by Foa and co-workers (see Foa 1997) on the consequences of emotional nonengagement and by Litz et al. (2000) on the relationship between arousal and emotional numbing suggests two ways in which extreme emotional reactivity following a traumatic event may interfere with emotional processing and recovery. According to Foa acute emotional reactivity can lead to extensive avoidance, thereby

preventing exposure to corrective information. Litz and co-workers proposed that chronic emotional reactivity can lead to emotional numbing, thereby reducing emotional engagement with the trauma memory.

3.3 The Effects of Corrective Experiences: Alteration of Existing Cognitive Structures or Development of New Cognitive Structures?

Foa and Kozak (1986) defined emotional processing as a change in the pathological elements of an existing fear structure following exposure to corrective information. A plausible alternative is that exposure to corrective information results in the formation of a *new* memory structure that coexists with the old fear memory rather than replacing it (Bouton 1988). Accordingly, a person's behavior in a given situation will be determined by which of the two structures is activated. From Foa and Kozak's perspective, the return of fear following successful treatment represents a further change in the fear structure. Bouton's theory would view relapse as a reduction in the accessibility of the new nonfearful memory. Support for Bouton's hypothesis comes from research conducted by Craske and co-workers (cf. Craske 1999), demonstrating that certain procedural variations in treatment by exposure therapy produced less between-session fear reduction, but prevented the long-term return of fear at follow-up.

3.4 Worry and Depressive Rumination

Emotional processing concepts have been utilized in the analysis of worry (Borkovec and Lyonfields 1993) and depressive rumination (Teasdale 1999). Although somewhat different analyses have been offered for these two phenomena, both Borkovec and Teasdale suggested that cognitive processing in one channel or at one level of meaning (linguistic/conceptual or propositional) may interfere with processing in another channel or level of meaning (imagery or implicational). Specifically, on the basis of studies showing that verbal articulation of fearful material results in less physiological arousal than imagining the same material (e.g., Vrana et al. 1986), Borkovec and co-workers have proposed that worry serves to inhibit emotional imagery. The short-term consequence of worry is to minimize emotional distress by limiting activation of the fear structure. However, the long-term consequence of avoiding full activation of the fear structure, as discussed above in the context of emotional numbing and emotional nonengagement, is to inhibit emotional processing and thereby maintain the worry.

Although the concept of emotional processing has generally been limited to the field of anxiety, Teasdale and co-workers (cf. Teasdale 1999) have proposed an

emotional processing model of depression. The model distinguishes between two different, but interacting levels of cognitive meaning, and proposes three different configurations of the interactions between these levels of meaning, also referred to as the 'three modes of mind.' One level of meaning is propositional, corresponding to the kind of information that can be conveyed in a single sentence. The other, implicational, represents higher order meaning derived from recurring patterns or themes. In the case of depression, the implicational meaning of experiences such as failing a class or the break-up of an important relationship may be a failure in all aspects of one's life. Importantly, the model posits that emotions are directly generated by the implicational level, while the propositional level of meaning only effects emotional states indirectly through the continual cycling of information between the two levels of meaning. Depressive rumination involves the cycling of depressogenic information between these two levels of meaning, thereby generating and maintaining dysphoria.

Teasdale (1999) further suggested that each meaning level can operate in one of two modes. In the *direct* mode, processing occurs in real time, whereas in the *buffered* mode information is allowed to accumulate before it is processed further. Although both levels of meaning can be operating in direct mode at the same time, only one level can be buffered at a given time. This yields three possible configurations, each one having different consequences for the person's emotional experience. When neither level of meaning is buffered, 'mindless emoting' occurs, in which people identify themselves with their emotional reactions and engage in little self-reflection. When the propositional level of meaning is buffered, individuals are said to be in the 'conceptualization/doing' configuration in which they think about themselves and their emotions as objects and develop goal-oriented strategies to cope with their problems and emotions. For example, depressive rumination frequently involves the person thinking about the nature, causes, and consequences of their depression. The third configuration, called 'mindful experiencing/being,' occurs when the implicational level is buffered and is characterized by 'integrated cognitive-affective inner exploration, use of present feelings and "felt senses" as a guide to problem solution and resolution, and a non-evaluative awareness of present subjective-self-experience' (p. S77).

Within this model of depression, Teasdale (1999) has defined effective emotional processing as 'processing that leads to modification of the affect-related schematic models [i.e. implicational level of meaning] that support processing configurations maintaining dysfunctional emotion' (p. S65). He further hypothesizes that emotional processing will occur only when the implicational system is buffered, thereby allowing the integration of new and old information. Accordingly, two general strategies can be used to promote

changes in implicational meaning. One strategy is to help depressed individuals create and encode in memory new implicational meanings, and many current cognitive therapy techniques (e.g., cognitive restructuring) help in doing so, although such techniques may run the risk of maintaining the conceptualizing/doing configuration. A second strategy, unique to Teasdale's theorizing, is to help people learn to shift from either of the two modes of mind that do not promote emotional processing (mindless emoting and conceptualizing/doing) to the one that does (mindful experiencing/being). This involves exercises in mindfulness, similar to some forms of meditation, in which individuals learn to become aware of, and acknowledge depressogenic thoughts without exclusively focusing on them. Teasdale found that a treatment program combining traditional cognitive therapy interventions with mindfulness exercises was more effective in preventing relapses among individuals with recurrent major depression than was cognitive therapy alone.

4. Concluding Comments

Emotional processing theory was advanced by Foa and Kozak (1986) primarily to account for the effectiveness of exposure therapy in the reduction of anxiety. However, consideration of its historical antecedents (Lang 1977, Rachman 1980) indicates the concept of emotional processing originally was conceived more broadly. Rachman (1980) in particular proposed that the concept should encompass a range of phenomena that includes, but is not limited to, anxiety and the efficacy of exposure therapy. In keeping with this, recent advances in such topic areas as natural recovery following trauma (Foa 1997), PTSD symptom structure (Litz et al. 2000), worry (Borkovec 1993) and relapse following recovery from depression (Teasdale 1999) have extended the domain of emotional processing theory and demonstrate its continued heuristic value.

At the same time, questions remain for future research as to the nature and limits of emotional processing, two of which are briefly noted. First, does exposure to corrective information alter existing cognitive structures, as proposed by Foa and Kozak (1986), or does it result in the acquisition of a new memory (Bouton 1988, Craske 1999)? Resolution of this issue has implications for understanding the process of relapse and development of relapse-prevention strategies. Second, to what extent do procedures that promote emotional processing of fear and anxiety promote the emotional processing of other problematic emotional reactions, such as sadness/depression, anger, or guilt? In exposure therapy, an important source of corrective information is habituation. However, it may be that these other emotions do not habituate (Pitman et al. 1991) and that alternative procedures may be needed to promote their

emotional processing, such as traditional cognitive restructuring techniques and mindfulness strategies (Teasdale 1999). Thus, resolution of this issue has implications for the design of interventions for other problematic (nonfear) emotional reactions.

See also: Anxiety and Anxiety Disorders; Anxiety and Fear, Neural Basis of; Behavior Psychotherapy: Rational and Emotive; Behavior Therapy: Psychiatric Aspects; Behavior Therapy: Psychological Perspectives; Emotion, Neural Basis of; Emotions and Health; Emotions, Psychological Structure of; Fear Conditioning; Habituation; Panic Disorder

Bibliography

- Borkovec T D, Lyonfields J D 1993 Worry: Thought suppression of emotional processing. In: Krohne H W (ed.) *Attention and Avoidance*. Hogrefe & Huber, Seattle, WA, pp. 101–18
- Bouton M E 1988 Context and ambiguity in extinction of emotional learning: Implications for exposure therapy. *Behaviour and Research Therapy* **26**: 13–149
- Craske M G 1999 *Anxiety Disorders: Psychological Approaches to Theory and Treatment*. Westview Press, Boulder, CO
- Foa E B 1997 Psychological processes related to recovery from a trauma and an effective treatment for PTSD. In: Yehuda R, McFarlane A (eds.) *Psychobiology of PTSD Annals of the New York Academy of Sciences*, New York, pp. 410–24
- Foa E B, Ehlers A, Clark D M, Tolin D F, Orsillo S M 1999 The Posttraumatic Cognitions Inventory (PTCI): Development and validation. *Psychological Assessment* **11**: 303–14
- Foa E B, Jaycox L H 1999 Cognitive-behavioral theory and treatment of posttraumatic stress disorder. In: Spiegel D (ed.) *Efficacy and Cost-effectiveness of Psychotherapy*. American Psychiatric Association, Washington, DC, pp. 23–61
- Foa E B, Kozak M J 1986 Emotional processing of fear: Exposure to corrective information. *Psychological Bulletin* **99**: 20–35
- Lang P J 1977 Imagery in therapy: An information processing analysis of fear. *Behavior Therapy* **8**: 862–86
- Lang P J 1984 Cognition in emotion: Concept and action. In: Izard C, Kagan J, Zajonc R (eds.) *Emotion, Cognition and Behavior*. Cambridge University Press, New York, pp. 193–206
- Lang P J, Melamed B G, Hart J 1970 A psychophysiological analysis of fear modification using an automated desensitization procedure. *Journal of Abnormal Psychology* **76**: 220–34
- Lepore S J, Silver R C, Wortman C B, Wayment H A 1996 Social constraints, intrusive thoughts, and depressive symptoms among bereaved mothers. *Journal of Personality and Social Psychology* **70**: 271–82
- Litz B T, Orsillo S M, Kaloupek D, Weathers F 2000 Emotional processing in posttraumatic stress disorder. *Journal of Abnormal Psychology* **109**: 26–39
- Pitman R K, Altman B, Greenwald E, Longpre R E, Macklin M L, Poire R E, Steketee G S 1991 Psychiatric complications during flooding therapy for posttraumatic stress disorder. *Journal of Clinical Psychiatry* **52**: 17–20
- Rachman S 1980 Emotional processing. *Behaviour and Research Therapy* **18**: 51–60
- Shalev A, Bonne O, Eth S 1996 Treatment of posttraumatic stress disorder: A review. *Psychosomatic Medicine* **58**: 165–82

Teasdale J D 1999 Emotional processing, three modes of mind and the prevention of relapse in depression. *Behaviour and Research Therapy* **37**: S53–S77

Vrana S R, Cuthbert B N, Lang P J 1986 Fear imagery and text processing. *Psychophysiology* **23**: 247–53

F. B. Foa and S. P. Cahill

Psychological Treatment Delivered Nontraditionally

The term ‘psychotherapy’ generally connotes an interactive, face-to-face process in which a trained mental health professional attempts to help a client or group of clients effect personal change. Authority and responsibility for facilitating such change reside with the professional, whose presence and expertise are deemed necessary to guide and control the meetings in accordance with sound clinical and ethical practices.

Since the 1970s an array of interventions for delivering psychological assistance have come into being that deviate significantly from the traditional psychotherapy format. In general these newer formats enjoy widespread popularity among the general public and their use is less directly controlled by a mental health professional. Such alternative psychological interventions include self-help and mutual support groups, radio and TV shows dispensing psychological advice, interactive computer software packages for addressing a wide variety of personal problems, self-help books, self-help audio and video tapes, and most recently a spate of cyberspace offerings. The latter category includes self-help guides, single session e-mail and bulletin board advice, ongoing e-mail counseling, real-time chat groups, and most recently, web telephony and videoconferencing.

A small body of research evaluating effectiveness focuses primarily on paraprofessionals and self-help books, and shows these alternatives to be sometimes as effective as more classic psychotherapy. However, much more research is needed. Added benefits of alternative formats are lower cost, convenience of access and use, and the availability of psychological aid for people who otherwise either would not, or could not, utilize traditional psychotherapy. Examples include the homebound, those in remote areas, people who consider psychotherapy stigmatizing, and individuals with a strong need for privacy and independence when dealing with their psychological problems.

This article first examines some of the social and cultural factors contributing to the development and use of these newer forms of psychological treatment. The treatments themselves are then reviewed along two dimensions. The first dimension is the source of the help: is it professional, paraprofessional, nonpro-

fessional, or some mix? The second dimension is the nature and degree of contact between help giver and help seeker. Finally, the status of research, along with issues about quality, safety, and protection of the public are briefly discussed.

1. Social Climate and the Rise of New Forms of Psychological Treatment

Viewed as a group, the newer models of psychological assistance, as compared with traditional forms of psychotherapy, tend to involve less direct professional contact and control, even when the treatment is designed or delivered by a mental health professional. This shift away from strictly professional delivery of mental health services reflects an intersection of several social developments. Observations about social trends are based on developments in the USA and may differ in applicability to other parts of the world.

At a broad societal level, considerable value has come to be placed on empowering the individual, as opposed to almost total reliance on institutions and professionals for information, support, and assistance. This stems in part from a post-Watergate, post-Vietnam weakening of US public confidence in its major institutions, including the corporate world, government, and the medical establishment. Various activist movements (e.g., those concerning civil rights, women, consumer rights, gays and lesbians), beginning in the 1960s and continuing to the present, also reinforce values about personal entitlement and empowerment. A result of this trend is that people now demand a greater measure of involvement and control over their physical and mental health care.

Technological advances in communications are a further factor in stimulating individual desire for greater control of over personal well-being. Before the 1980s many personal issues such as alcohol and drug abuse, child molestation, family violence, sexual orientation, bulimia and anorexia nervosa, cancer, sexually transmitted diseases, Alzheimer's disease, fetal alcohol syndrome, mental illness, to name but a few, were topics rarely spoken of in public. Early television talk shows played an important role in shining a spotlight on such previously closeted issues. As people became increasingly sophisticated about the relationship between stress and health, and the potential damage that various stressful life events can cause, the public also demonstrated its eagerness for help.

However, scientific surveys of the prevalence and incidence of mental health problems, at least insofar as the US public is concerned, make it clear that we do not now have, and probably never will have, a sufficient number of trained mental health professionals to meet the needs of all those reporting diagnosable psychological problems. Meanwhile, the need for psychological services continues to grow, fed in part by dramatic advances in medical science. More

people with serious chronic medical conditions are surviving, many into old age, putting numerous stresses on both patients and their families.

At the same time, the growth of managed health care, with its focus on cost containment, has resulted in limiting access to care, especially for mental health concerns (see *Managed Care and Psychotherapy in the United States*). This is so despite a growing body of research from the field of psychoneuroimmunology (Hafen et al. 1996), demonstrating that psychological stressors have a negative impact on the immune system, general health, and overall functioning (see *Psychoneuroimmunology*). The same body of research also shows the social support and the opportunity to discuss problems constructively can positively affect health, enhance the quality of life, and even promote longevity. Western medicine, though desirous of promoting health and preventing the occurrence or the worsening of disease, fails fully to capitalize on what is known about the mind-body connection and its impact on health and well-being.

Further, under managed care, doctors and nurses, traditionally an important source of informal psychological assistance, now have less time to spend with their patients than in the past. The ironic result is a public increasingly aware of the connection between stress and physical and psychological illness, but finding affordable psychological services growing more scarce.

To summarize, four factors stand out as particularly influential in promoting nontraditional psychological treatments: (a) the growing public awareness and desire for help at the very time that for many the availability of affordable professional psychotherapy is decreasing; (b) the value placed on personal entitlement and empowerment in today's society; (c) an ongoing revolution in electronic forms of communication, including the very rapid growth and easy accessibility of information and services via the Internet; (d) the rapid dissemination of information by today's highly competitive mass media businesses. Given these forces, it is not surprising that enterprising people have, and no doubt will continue to find, alternative, accessible, less costly (although sometimes financially quite profitable) ways of delivering psychological treatments.

One such example of a break from traditional psychotherapy began in the late 1970s, as the number and type of self-help/mutual support groups multiplied rapidly (Goodman and Jacobs 1994). Such groups had existed for decades, the best known being Alcoholics Anonymous and Recovery, Inc. (see *Alcoholics Anonymous*). But suddenly groups sprang up for almost every conceivable personal problem from substance abuse, to parenting concerns, to being a victim of violence, to life transitions such as divorce, retirement, or bereavement, to lifestyle issues such as sexual orientation, transvestitism, open marriages, as well as groups for almost every disease listed by the

World Health Organization. Millions of people joined them and, although some groups utilize professional leaders or professionally developed ideas, many have little or no professional input. Sociologist Frank Reissman (Gartner and Reissman 1984) terms the heart of the self-help group movement the 'helper-therapy' principle—the idea that when people suffering from a common problem come together for mutual support, they not only receive help from others, particularly from those more experienced in coping with the problem, but also give help, and in so doing are themselves strengthened and empowered. In the 1980s many health professionals held negative views about self-help groups, asserting that lay people left on their own were likely to harm each other by dispensing bad advice. Less openly discussed were their concerns over loss of professional control and income. Over time professionals have become more supportive of self-help groups as both an adjunct to traditional treatment and as a useful, member-governed intervention in its own right. Today many professionals are directly involved with groups in a variety of roles such as consultants, researchers, leaders, co-leaders with group members, and even as participants who share the group's common concern.

More recent alternatives to traditional psychotherapy have been generated as a result of the ongoing technological revolution in electronic communication, which continues to offer new means of interacting to large numbers of people. On the Internet, some formats take advantage of real-time communication to deliver psychological aid, while others necessitate time delays between the posting of messages requesting psychological help and the responses to those messages. One recent review of psychological assistance on the Internet (Barak 1999), listed 10 types of interventions: (a) information resources on psychological concepts and issues; (b) self-help guides; (c) psychological testing and assessment; (d) help in deciding to undergo therapy; (e) information about specific psychological services; (f) single-session psychological advice through e-mail or e-bulletin boards; (g) ongoing personal counseling and therapy through e-mail; (h) real-time counseling through chat, web telephony, and videoconferencing; (i) real-time and delayed-time support groups, discussion groups, and group counseling; (j) psychological and social research. The next section offers a framework for organizing the many and varied nontraditional intervention formats offered to the public.

2. A Framework For Viewing Nontraditional Forms Of Psychological Assistance

Two basic elements in most traditional psychotherapy are (a) live, real-time, face-to-face contact between a help seeker and a help giver and (b) the help giver

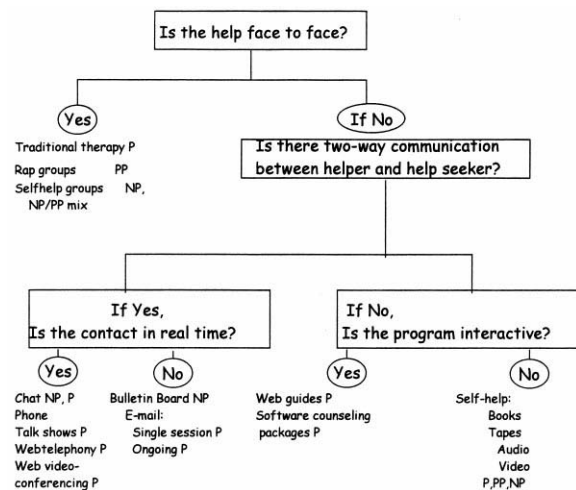


Figure 1

Nature and type of psychological treatments

being a trained mental health professional. This section outlines alternative psychological treatments in terms of their departure from these two elements, that is, alternative formats for delivering psychological aid are shown according to two major categorizations.

The first categorization is the nature of the contact between help seeker and the source of the help: is it live and face-to-face? If not, is there some type of two-way communication between the helper and help seeker? If so, is it in real time or delayed time? (Note that in either case, the true identity of either party may not be known.) If not, does the help seeker interact and receive feedback from a packaged program such as those that can be bought and used on personal computers or accessed via the Internet? If not, what is the format for the delivery of the noninteractive aid?

The second categorization refers to the training of the person offering the mental health aid: is the person a trained mental health professional, a paraprofessional, or a nonprofessional? The intention is to provide a framework for categorizing any type of psychological treatment. For the examples listed in Fig. 1, P = professional, PP = paraprofessional, and NP = nonprofessional. The listing is not intended to be exhaustive.

The relatively unregulated nature of the various forms of nontraditional psychological treatment being offered to the lay public has raised concerns about their safety, quality, and effectiveness. Research into the use of paraprofessionals vs. professionals, self-help books vs. therapist-delivered services, and computer-assisted vs. therapist-delivered treatment generally shows these alternatives to be similar in effectiveness to professionally delivered psychotherapy (Christensen and Jacobson 1994). But not only is that body of scientific studies relatively small, some newer

computer-based formats have never been studied. Thus, many important questions remain unanswered. For example, are some nontraditional formats more appropriate for certain types of people, or certain types or levels of severity of problem?

Further, the relatively unregulated nature of the Internet leaves open the possibility of poor quality offerings, untested claims, and even outright charlatanism among those designing and offering psychological assistance. In the interest of assuring quality psychological services for the public, priority needs to be given to research, as well as education for both Internet users and service providers, about appropriate standards.

See also: Psychotherapy: Ethical Issues; Psychotherapy Integration; Psychotherapy Process Research; Social Support and Health; Social Support and Recovery from Disease and Medical Procedures; Social Support and Stress; Social Support, Psychology of; Support and Self-help Groups and Health

Bibliography

- Barak A 1999 Psychological applications on the Internet: A discipline on the threshold of a new millennium. *Applied and Preventative Psychology* 8: 231–45
- Christensen A, Jacobson N S 1994 Who (or what) can do psychotherapy: The status and challenge of nonprofessional therapies. *Psychological Science* 5: 8–14
- Gartner A, Reissman F 1984 *Self-help Revolution*. Human Services Press, New York
- Goodman G, Jacobs M K 1994 The self-help, mutual support group. In: Fuhriman A, Burlingame G M (eds.) *Handbook of Group Psychotherapy*. Wiley, New York, pp. 489–526
- Hafen B Q, Karren K J, Frandsen K J, Smith N L 1996 *Mind/Body Health The Effects of Attitudes, Emotions and Relationships*. Allyn and Bacon, Boston, MA

M. K. Jacobs

Psychological Treatment, Effectiveness of

1. Introduction

Despite its relative youth, the field of psychotherapy has faced many challenges, including one that disputed the very purpose it intended to accomplish. In a review of the effectiveness of psychotherapy, Eysenck (1952) concluded that 74 percent of patients, drawn from 24 studies, and diagnosed as neurotic, were found to make progress considered equivalent to a similarly diagnosed group of individuals who received no treatment.

This finding stimulated years of controversy regarding the effectiveness of psychotherapy, as well as a determined and focused effort on the part of researchers to establish empirically the beneficial effects, if any, of psychotherapy. In addition to reporting the findings supporting the efficacy of psychotherapy treatment, the remainder of this discussion will present related findings from this effort: (a) the durable effects of psychological treatment, (b) the factors leading to patient improvement, and (c) the disturbing evidence that suggests that not all patients benefit from treatment. Of related importance to this discussion is the paradoxical outcome that often accompanies many scientific investigations; namely, that as each question is sufficiently answered, more questions take its place.

2. The Definition of Psychotherapy

In general, psychological treatment is a method designed to change a person's thoughts, feelings, or behaviors with the intention of facilitating increased functioning and life adjustment. Approaches to psychotherapy are numerous and diverse including such orientations as cognitive, behavioral, client centered, gestalt, and existential therapies. All of the various approaches involve a confiding relationship with a trained therapist focused on a patient in need of assistance.

3. The Effectiveness of Psychotherapy

As mentioned above, the conclusion made by Eysenck stimulated both the frequency and quality of studies designed to evaluate the outcome of therapy. Researchers pursuing evidence to refute Eysenck's claim have investigated a number of domains that are considered relevant to distinguishing the effects of psychotherapy. Assessments of the benefits of psychotherapy have focused on such domains as an increase in problem-solving skills, reduction in symptoms, improvements in interpersonal and social functioning, and health. The changes noted in patients treated with psychotherapy have been compared to groups of individuals who have not been exposed to therapy. Reviews of this literature have confirmed the beneficial effects of therapy, concluding that treated patients consistently fare better than untreated patients (Lambert and Bergin 1992, 1994).

An often used statistical technique, meta-analysis, has enabled researchers to summarize individual collections of empirical data, comparing the effect sizes of treatment, placebo, and control groups. For example, using meta-analysis, Smith et al. (1980) reported that the average person in treatment is better off at the end of treatment than 80 percent of the untreated

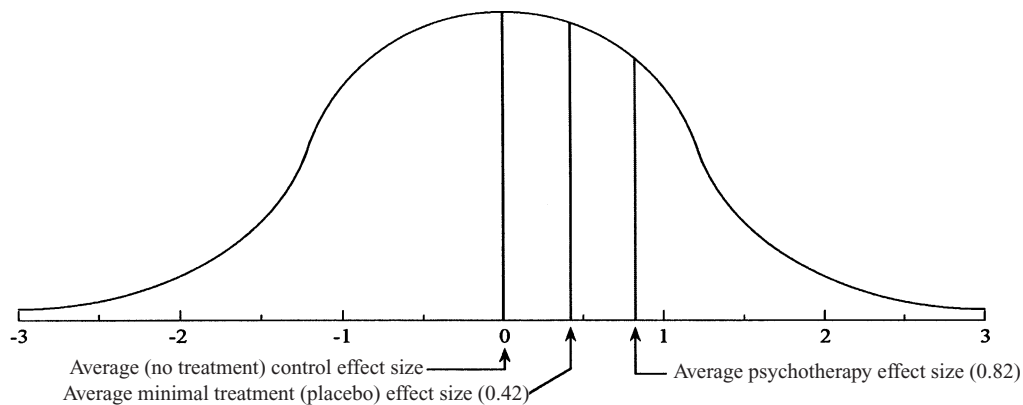


Figure 1

Comparison of placebo and psychotherapy effects in relation to no-treatment control

persons. A graphic representation of the effect sizes (treatment, placebo, and control) found in a number of meta-analytic reviews is presented in Fig. 1.

4. Treatment Duration and Lasting Effects

In contrast to the length of treatment associated with early theories of psychoanalytic psychotherapy, the extant literature suggests that substantial benefits from treatment occur in relatively short durations, and equally important, these gains are maintained. Dose-effect designs, derived from the dosage model introduced by Howard et al. (1986), have confirmed a positive rate of effect as the dose (i.e., number of sessions) increases, with diminishing returns at higher dosage levels. Studies from the dose-effect literature suggest that, for many clinical disorders, a range of 10 to 20 sessions is sufficient for 50 percent of patients to return to normal functioning. Moreover, promising results occur as the dosage level is increased to the extent that 75 percent of patients can be expected to recover substantially after 26 to 50 sessions (Lambert et al. 2001).

Although a majority of patients experience significant benefits from psychotherapy, there is also evidence that different clusters of psychological symptoms respond to treatment at distinct rates of improvement. It appears that symptoms of poor morale respond most quickly, followed by chronic symptoms of distress, with symptoms intrinsic to the character of the patient recovering the slowest. Observed trends in patient improvement have led some researchers to propose a phase model of recovery in which personal welfare, symptom reduction, and interpersonal and social functioning improve in sequence.

In addition to results supporting relatively short recovery periods for a majority of patients, there is evidence that beneficial gains from psychotherapy are

enduring. Although adjustment from psychotherapy treatment does not prevent recurrences of psychological disturbance, it appears that many patients maintain healthy functioning for prolonged periods of time subsequent to treatment. However, exceptions to the notion of lasting treatment gains are noted in the conclusions of a review by Nicholson and Berman (1983), which included patients suffering from symptoms of substance abuse, eating, recurrent depression, and personality disorders.

5. Factors Important to Psychotherapy Outcome

In addition to the research focused on supporting the effectiveness of psychotherapy, considerable emphasis has been placed on the factors attributable to positive patient outcome. Findings from this research fall into three general categories: patient variables, common variables, and specific intervention variables.

5.1 Patient Variables

Although a number of studies have investigated the impact therapists have on psychotherapy outcome, there is surprisingly little evidence that suggests therapists or their techniques are central and essential factors affecting patient outcome. On the contrary, it is in fact the patient that is most likely to determine successful outcome. The literature focusing on the study of patient variables is too voluminous to repeat, and thus only the most important findings are reported here.

Each patient entering treatment brings a diverse array of factors when presenting to the clinician. These include a variety of psychological disorders, historical backgrounds, stresses, and the quality of their social support networks to name just a few. Patient variables found to have relationships with outcome are severity

of disturbance, motivation and expectancy, capacity to relate, strength of ego, psychological mindedness, and ability to recognize relevant problems (Garfield 1994). Therefore, patients suffering from challenging symptoms as well as relative deficits in numerous patient variable categories are likely to experience negligible improvement during treatment. For example, the borderline, alcoholic patient with suicidal tendencies who is forced into treatment, believing that most marital problems are a result of an insensitive spouse, is likely to benefit less from treatment than the depressed patient who voluntarily begins treatment and is determined to make personal changes that will lead to marital harmony.

Although patient variables have been conveniently categorized, the impact each has on the outcome of therapy is not equal. In general, the severity of disturbance is the patient variable that is most related to the outcome of the treatment. Patients with serious mental disorders, such as the schizophrenic, schizoaffective, and bipolar disorders are typically treated primarily with psychopharmacological management, suggesting the challenge these disturbances provide for psychotherapy alone. Personality disturbances also appear immutable to change in the short term use of psychotherapy.

5.2 Common Factors

Because outcome researchers have consistently failed to find sufficient evidence indicating superior effect sizes for particular theoretical modalities, some have concluded that psychotherapy equivalence is a result of factors common to all theories of psychotherapy. Whereas some view these beneficial ingredients as both necessary and sufficient for change, others perceive that these ingredients act in coalition with variables that are unique to particular interventions. It is important to note that regardless of the position held, factors common to all treatments are likely to account for the greatest amount of the improvement in patients undergoing psychotherapy.

In a review of the empirical research, Lambert and Bergin (1994) grouped many of the shared active ingredients of a variety of psychotherapies into three categories: support, learning, and action factors. The factors included in these categories seem to operate most actively during the in-session process of therapy, resulting in a collaborative working relationship that allows clients to confront and cope with reality more effectively. Factors that engender a collaborative relationship include a client's increased sense of trust, security, and safety, along with a reduction in tension, threat, and anxiety that in turn promotes a client's awareness of problems and ultimately new and more healthy perspectives on fears, appropriate risk taking, and motivations for resolving problems in interpersonal relationships.

Factors receiving the most attention in the literature are the common factors considered to be the core of therapy (empathy, warmth, and positive regard). A number of related terms such as acceptance, tolerance, therapeutic alliance, working alliance, and support are also used to describe these ingredients (see *Therapist–Patient Relationship*). In a study designed to explore which of several therapist variables distinguished most consistently between more effective and less effective therapist trainees, Lafferty et al. (1991) concluded that a client's perception of empathy is important to outcome, because less effective therapists, those with patients experiencing more distress as a result of treatment, were found to have lower levels of empathy.

Findings from studies investigating the therapeutic relationship have suggested that a positive working alliance between the patient and therapist is a necessary condition for patient improvement. When asked for factors that were most helpful in their treatment, patients tended to endorse the personal qualities of their therapists, using adjectives such as 'sensitive,' 'honest,' and 'gentle.' The results of similar studies confirm the import that patients place on the characteristics of therapists (Horvath and Greenberg 1994).

Additional evidence supporting the necessity for a strong working relationship is the positive association repeatedly found between scores on measures of early treatment alliance and psychotherapy outcome. In clinical trial studies comparing multiple treatments for particular psychological disorders, patients' positive assessment of the therapeutic alliance is one of the best predictors of psychotherapy outcome (e.g., Krupnick et al. 1996). Although a significant amount of evidence points to common factors as mediators of patient change, the notion of beneficial ingredients unique to specific therapies cannot be ruled out, as occasionally such interventions can be shown to contribute to successful outcomes in patients independent of common factors (see *Psychological Treatments: Randomized Controlled Clinical Trials*).

5.3 Specific Interventions

The drive to empirically validate the efficacy of specific interventions is due, in part, to the field of psychotherapy's commitment to theoretically based approaches, as well as the influence of political and economic forces. Collectively, these factors have led to the use of comparative outcome studies, which generally avoid the ethical and methodological problems associated with no-treatment and placebo controls, while enabling researchers to study the effectiveness of particular techniques or theories by comparison.

5.3.1 Anxiety disorders. Specific behavioral treatments have been found to be effective in treating

patients with anxiety disorders. To date, two treatments, systematic exposure and applied relaxation techniques, have been shown to be efficacious in treating a variety of anxiety disorders. Exposure is a broad term that encompasses both interoceptive and situational exposure interventions. Interoceptive procedures attempt to reliably produce panic-type sensations, with the intent of severing or weakening the patient's associations between certain bodily cues and their respective reactions to panic. The technique of situational exposure, also known as *in vivo* exposure, refers to the patient's repeated intentional encounters with a feared stimulus. Given the focus placed on the catalyst of fear, it is not surprising that clearly identifying the stimulus that induces the patient's reactions is crucial for the success of these procedures.

Evidence suggests that the most effective treatment strategy includes identification of feared stimuli, confrontative periods of exposure sufficient in length to reduce the patient's level of anxiety, and a focus on thoughts providing self-efficacy or performance accomplishment. A general strategy employing the use of exposure techniques appears to be the only efficacious intervention for agoraphobia (DeRubeis and Crits-Christoph 1998). Similar techniques have been shown to be successful for social phobic, panic, and post-traumatic stress disorder patients as well.

The results of a form of relaxation, referred to as applied relaxation, appear to show promise in treating specific anxiety disorders. Applied relaxation entails instructing patients in the skill of progressive muscle relaxation. Patients are generally taught how to progressively relax specific muscle groups. Whereas the evidence supporting the use of applied relaxation techniques alone is mixed in the panic disorder literature, it appears there is support for treating generalized anxiety disorder with relaxation training.

Additional evidence suggests that the treatment of specific disorders is most successful when particular interventions are combined with specific theory-based therapies. For instance, panic disorder is treated most effectively when a cognitive-behavioral theory of therapy is used in tandem with behavioral techniques. Barlow et al. (1989) compared relaxation training (RT), imaginal exposure plus cognitive restructuring (E+C), and a combination of these two interventions (RT+E+C). Outcome was determined by the percentage of patients experiencing panic attacks. Results indicated that 36 percent of the WL (wait-list control) patients, 60 percent of the RT patients, 85 percent of the E+C patients, and 87 percent of the RT+E+C patients were free of panic attacks at the end of treatment.

5.3.2 Depression. Like anxiety-based disorders, there is some evidence supporting the use of specific behavioral techniques in the treatment of depression;

however, in the case of depression, these procedures are no more effective than cognitive, interpersonal, and psychodynamic-interpersonal therapies. For example, behavioral activation alone, which refers to a patient scheduling activities that promote mastery as well as pleasure experiences, has been found to be as effective as cognitive therapy. Whereas behavioral activation is a procedure that makes concerted efforts to involve patients in constructive and pleasurable activities, cognitive therapy emphasizes the modification of maladaptive thoughts, beliefs, attitudes, and behaviors of individuals. Although some researchers predicted the superiority of cognitive therapy's ability to successfully treat depressive patients (Hollon and Beck 1986), the results of large well-controlled studies suggest that interpersonal and psychodynamic-interpersonal therapies are efficacious as well. In contrast to the focus that cognitive therapy places on maladaptive thoughts, beliefs, and behaviors, the interpersonal theory perceives depression as the consequence of problems in the interpersonal history of the patient.

5.3.3 Eating disorders. As eating disorders are generally characterized by harmful cognitions surrounding weight and physical image, cognitive and behavioral interventions have been the most extensively researched treatments. In particular, cognitive behavior therapy has been the focus of much research in the treatment of bulimia nervosa, showing promising results. However, it appears that interpersonal therapy provides generally equivalent results when long term follow-up findings are considered (Fairburn et al. 1993).

5.3.4 Substance abuse disorders. Reviews of the treatment of substance abuse disorders suggest that various therapy modalities have been found effective in treating patients. Therapies that provide generally equivalent results are cognitive, cognitive behavior, and psychodynamic derived therapies. Whereas patients abusing alcohol have been treated, with similar success, using psychodynamic, cognitive behavior, and psychoeducational treatment modalities, opiate dependent patients have been treated with cognitive and supportive-expressive psychodynamic therapies with generally equivalent results.

Some caution should be applied when interpreting the above findings, however, as the results mentioned are typically based on randomized clinical trials. To date, there is little evidence suggesting that these findings generalize to the clinical setting. The question of whether one technique is comparatively unique and more effective than another is a consistent topic in the literature of psychotherapy. There is currently little

evidence supporting superiority of particular interventions, although future research may note unique contributions. For instance, there is a growing body of evidence suggesting that the use of therapy manuals allow for the detection of discernable differences in the behavior of therapists, while at the same time enhancing the effects of specific therapy procedures. Perhaps these contributions will allow for more accurate comparisons in outcome studies. The search for specific techniques and common factors that distinguish successful treatment has proved to be difficult because of the myriad factors that likely exist in the interactions between therapists and clients.

6. The Negative Effects of Psychotherapy

Findings from the study of the progress and outcome of patients undergoing psychotherapy have resulted in both expected and unexpected conclusions. Most patients receiving treatment improve, a minority of patients remain unchanged, and contrary to the intent of psychotherapy, some patients actually worsen as a result of treatment. Independent reviews of the literature suggest that there are patients who fail to achieve beneficial gains from treatment. Although accurate estimates of patient worsening are difficult to make because the results of such incidents are not often the focus of psychotherapy outcome studies, estimates of patient worsening vary between 0 and 15 percent (Lambert and Bergin 1992). Similar to patient improvement, it appears that the phenomenon of deterioration is equivalent across theoretical orientations. In addition, treatment failures are reported in all client populations, treatment interventions, and group and family therapies.

Despite what appears to be an equitable distribution of treatment failures, specific factors have been associated with deterioration. Consistent with the influence that patient variables often have in treatment success, patient characteristics appear closely linked to treatment failure. Severity of mental illness alone is predictive of unsuccessful outcome, and severe levels of mental illness in combination with specific interventions designed to breach, challenge, and undermine entrenched coping strategies and defense mechanisms are more likely to result in treatment failure. Consequently, psychotic, borderline, schizophrenic, and bipolar patients are those patients most frequently experiencing deterioration during the course of treatment. Similar findings have been reported in the psychoanalytic literature, as patients lacking quality interpersonal relationships, low anxiety tolerance at the inception of treatment, and low motivation were more likely to worsen when treated with psychoanalysis and purely supportive therapies, whereas patients treated with supportive-expressive psychotherapy enjoyed better outcomes. Deterioration is not unique to individual therapy, however. Reviews of

group psychotherapy outcome have reported a positive correlation between treatment failure and patient variables such as low participation, poor self-esteem, poor self-concept, and more significant needs for fulfillment. Additional personal characteristics associated with patient deterioration include hostility, interpersonal dysfunction, and negative expectations of treatment.

The above findings suggest that it is important to consider the type of intervention most appropriate for a particular client; however, harm is not the fault of patient and intervention alone. The results of studies investigating therapists' contributions to the phenomenon of deterioration have suggested that specific therapists' behaviors can lead to treatment failure. For instance, the results of a study investigating the relationship between therapist's emotional adjustment, empathy, directiveness, support, and credibility, and patient outcome found that a therapist's empathic abilities were the most predictive of patient outcome (Lafferty et al. 1991). Additional findings of the study suggested that effective therapists were also likely to provide more direction and support than ineffective therapists. These researchers were also interested in the similarity in values held by the therapists and patients, and found that less-effective therapists emphasized values, such as the importance placed on a life of comfort and excitement, whereas effective therapists considered intellectual values, such as reflection, as more important.

Another study suggested that patient deterioration is related to therapist countertransference reflecting disappointment, hostility, or irritation (Mohr 1995). In addition, there was some indication that client deterioration is more prevalent when clinicians underestimate their client's illness or overestimate their client's progress in treatment. Research exploring the importance of leadership styles in group therapists suggests that confrontational and aggressive group therapists tended to have more treatment failures in their groups. These therapists were observed to be more likely to insist upon immediate client disclosure, emotional expression, and change in attitude.

7. Conclusion

Years of clinical research have produced compelling evidence for the value of psychotherapy. Contrary to Eysenck's controversial publication regarding the efficacy of psychotherapy, outcome studies have confirmed that patients benefit in a variety of domains from treatment. According to the dose-effect literature, benefits derived from therapy transpire in relatively brief periods of time. For many clinical disorders, generally one-half of all patients return to normal functioning in 10 to 20 sessions. An additional 25 percent of patients experience substantial improvement when dosage levels are increased to 30 to 50 sessions. The benefits from treatment appear to be

durable, as the results of follow-up studies suggest that the effects of treatment endure for at least one to two years subsequent to treatment.

Not all patients experience the beneficial effects of treatment, and this finding has inspired the investigation of factors facilitating both positive and negative change. The results of such studies have indicated that a patient's level of disturbance is most predictive of outcome; however, desire to change, interpersonal skills, ego strength, psychological mindedness, and awareness of relevant problems have all been found to be related to outcome. In addition to patient variables predictive of successful outcome, there are characteristics of a therapeutic relationship that facilitate a patient's sense of trust, safety, and security. According to patients' reports, therapist empathy and a strong therapeutic alliance are most predictive of treatment outcomes.

Although the use of specific interventions has showed promising results in the case of particular disorders, in general the search for superior treatment strategies has been disappointing, and according to some researchers unnecessary, given the complexity of potential interactions between therapists, clients, and modes of therapy. Regretfully, not all patients improve from treatment, indeed, some actually worsen. Knowledge of this phenomenon is somewhat limited; however, there is evidence that suggests that treatment failure is best predicted by severity of mental illness. Furthermore, factors inhibiting the development of a strong therapeutic alliance also predict negative outcome.

Although researchers have contributed empirical evidence supporting the effectiveness of psychotherapy, they are just learning to understand why patients tend to improve. In time, it is likely that focused and determined research will eventually lead to greater insights into this important question.

See also: Assessment and Application of Therapeutic Effectiveness; Ethical Implications of; Program Evaluation; Psychological Treatments, Empirically Supported; Psychological Treatments: Randomized Controlled Clinical Trials; Psychotherapy and Pharmacotherapy, Combined; Psychotherapy: Clinical Practice Guidelines; Psychotherapy Process Research; Therapist–Patient Relationship

Bibliography

- Barlow D H, Raske M G, Cerny J A, Klosko J S 1989 Behavioral treatment of panic disorder. *Behavior Therapy* **23**: 507–28
- DeRubeis R J, Crits-Christoph P 1998 Empirically supported individual and group psychological treatments for adult mental disorders. *Journal of Consulting and Clinical Psychology* **66**: 37–52
- Eysenck H J 1952 The effects of psychotherapy. *Journal of Consulting and Clinical Psychology* **16**: 319–24

- Fairburn C G, Jones R, Peveler R C, Hope R A, O'Conner M 1993 Psychotherapy and bulimia nervosa: Longer-term effects of interpersonal psychotherapy, behavior therapy, and cognitive behavior therapy. *Archives of General Psychiatry* **50**: 419–28
- Garfield S L 1994 Research on client variables in psychotherapy. In: Bergin A E, Garfield S L (eds.) *Handbook of Psychotherapy and Behavior Change*, 4th edn. Wiley, New York, pp. 190–228
- Hollon S D, Beck A T 1986 Cognitive and cognitive-behavioral therapies. In: Garfield S L, Bergin A E (eds.) *Handbook of Psychotherapy and Behavior Change*, 3rd edn. Wiley, New York, pp. 443–82
- Horvath A O, Greenberg L S (eds.) 1994 *The Working Alliance: Theory, Research, and Practice*. Wiley, New York
- Howard K I, Kopta S M, Krause M S, Orlinsky D E 1986 The dose-effect relationship in psychotherapy. *American Psychologist* **41**: 159–64
- Krupnick J L, Sotsky S M, Simmens S, Moyer J, Elkin I, Watkins J, Pilkonis P A 1996 The role of the therapeutic alliance in psychotherapy and pharmacotherapy outcome: findings in the National Institute of Mental Health Treatment of Depression Collaborative Research Program. *Journal of Consulting and Clinical Psychology* **64**: 532–9
- Lafferty P, Beutler L E, Crago M 1991 Therapeutic relationship and improvement as perceived by clients and therapists. *Journal of Consulting and Clinical Psychology* **57**: 76–80
- Lambert M J, Bergin A E 1992 The achievements and limitations of psychotherapy research. In: Friedman D K (eds.) *The History of Psychotherapy: A Century of Change*. American Psychological Association, Washington, DC, pp. 360–90
- Lambert M J, Bergin A E 1994 The effectiveness of psychotherapy. In: Bergin A E, Garfield S L (eds.) *Handbook of Psychotherapy and Behavior Change*, 4th edn. Wiley, New York, pp. 143–89
- Lambert M J, Hansen N B, Finch A E 2001 Patient-focused research: using outcome data to enhance treatment effects. *Journal of Consulting and Clinical Psychology* **69**
- Mohr D C 1995 Negative outcome in psychotherapy: A critical review. *Clinical Psychology Science Practice* **2**: 1–27
- Nicholson R A, Berman J S 1983 Is follow-up necessary in evaluating psychotherapy? *Psychological Bulletin* **93**: 261–78
- Smith M L, Glass G V, Miller T I 1980 *The Benefits of Psychotherapy*. Johns Hopkins University Press, Baltimore
- Wampold B E 1997 Methodological problems in identifying efficacious psychotherapies. *Psychotherapy Research* **7**: 21–43

M. J. Lambert and E. J. Hawkins

Psychological Treatments, Empirically Supported

Changes in the US healthcare system, set in full motion during the past decade and continuing to evolve, are likely to have a profound impact on the way psychotherapy is practiced. Two primary forces are driving the changes: (a) the increasing penetration of managed care and other types of cost-conscious medical plans, and (b) the development and proliferation of clinical practice guidelines (see *Psychotherapy*:

Clinical Practice Guidelines). The most notable impact of these trends has been the requirement for greater accountability on behalf of service providers. Essentially, managed care and practice guidelines ask the question: does a treatment work and, if so, is it cost effective? Although these forces are influencing every area of healthcare, the current paper will discuss their specific impact on the practice of psychotherapy, and explore why the identification, promotion, and dissemination of empirically supported treatments (ESTs; i.e., those with controlled research data supporting their efficacy, see below for the specific criteria used to make this determination) is essential to the survival of psychological interventions as a first-line treatment for emotional disorders.

1. Managed Care

More than any previous event, managed care organizations (or any other system monitoring the utilization and cost of service such as health maintenance organizations, capitated contracts with providers, etc.) are reshaping the practice of psychotherapy (see *Managed Care and Psychotherapy in the United States*). In the traditional fee-for-service model, decisions about the cost and length of treatment were primarily a function of choices made by the doctor and patient. The allocation of resources (i.e., cost of psychotherapy) was of little concern to the clinician. In fact, the fee-for-service model encourages the provision of service, as more service creates more income. However, in response to the increased costs of psychotherapy, and in particular to the perceived 'endless' nature of psychotherapy, managed care organizations are pressuring clinicians to allocate decreasing (less than usual) amounts of service.

While to date, the focus of managed care organizations cost cutting has been almost entirely on limiting the number of sessions a patient receives, ultimately, in order to compete, managed care organizations will have to focus on the quality and efficacy of the psychotherapeutic interventions as well. Managed care organizations are clearly motivated to cut costs. However, they must also satisfy the consumer (i.e., patient), and payor (e.g., employer providing health benefits) and therefore have to balance cost cutting with patient outcome. Simply reducing the length of care will not accomplish this goal, because many, and perhaps the majority, of psychotherapists are trained in 'long-term' therapeutic interventions which by nature require more sessions than managed care organizations are willing to allocate. Simply reducing the length of therapies devised to take place over a longer period of time is likely to result in ineffective outcomes, leading both to consumer dissatisfaction and increased costs down the road as the severity of the disorder may increase, becoming less responsive to treatment. Thus, concern for the effectiveness of an intervention will eventually temper the managed care

organizations focus on economics. Ultimately, managed care organizations will be interested in clinicians providing the 'optimal intervention...the least extensive, intensive, intrusive and costly intervention capable of successfully addressing the presenting problem' (Bennett 1992, p. 203).

2. Practice Guidelines

A second force that will have a significant impact on the practice of psychotherapy is the establishment of clinical practice guidelines and treatment consensus statements. As noted by Smith and Hamilton (1994), 'Guidelines are now being developed because there is a perception that inappropriate medical care is sometimes provided and that such inappropriate care has both health and economic consequences' (p. 42). It is well known that in all medical fields there is significant variability in treatments delivered to patients with various illnesses. Patients are not necessarily receiving the most effective treatment. Consequently, in order to ensure that patients uniformly receive the optimal intervention—whether it be a type of medication, surgical procedure, or psychotherapeutic intervention—clinical practice guidelines are being developed. For the most part, the development of practice guidelines relies on systematically evaluating scientific and clinical input.

One organization developing practice guidelines is the federal Agency for Healthcare Policy and Research (AHCPR). In order to create impeccable guidelines, only treatments with documented efficacy from randomized controlled trials are considered. Therefore, for an intervention to be included as a first-line treatment, research evidence attesting to its efficacy must be available. Without this evidence, one cannot conclude if a treatment is effective or ineffective, or make judgments about its relative efficacy compared with other available treatments.

The recommendations of the clinical practice guidelines are quite clear. Consider the wording from the AHCPR guideline for Depression (Depression Guideline Panel 1993). When psychotherapy is going to be selected as the sole treatment, the guideline states 'Psychotherapy should generally be time-limited, focused on current problems, and aimed at symptom resolution rather than personality change as the initial target. Since it has not been established that all forms of psychotherapy are equally effective in major depressive disorder, if one is chosen as the sole treatment, it should have been studied in randomized controlled trials' (Depression Guideline Panel 1993, p. 84). In addition to endorsing specific treatments for depression that had sufficient empirical evidence (e.g., cognitive, behavioral, interpersonal), the report goes on to state 'Long-term therapies are not currently indicated as first-line acute phase treatments' (p. 84).

Along the same line as treatment guidelines, consensus statements are likely to carry a significant

amount of weight in determining what treatments have been shown to be effective as well. For example, in 1991 a Consensus Development Conference on the Treatment of Panic Disorder was held, sponsored by the National Institute of Mental Health. As was true in the process of developing the AHCPR treatment guidelines, the merit of various treatments was determined by the available scientific evidence. The format of the conference was to let 'each camp put its best data forward' and the panelists would evaluate and weigh the evidence to formulate a consensus statement (Wolfe and Maser 1994). As in the case of treatment guidelines, specific treatments were judged to be effective for panic disorder: Several pharmacological compounds and cognitive behavioral psychotherapy (cf. Panic Consensus Statement published in Wolfe and Maser 1994). Once again, as is true of treatment guidelines, the wording of the Panic Consensus Statement as to the use of treatments not supported by empirical evidence, whether they be psychological or pharmacological, is quite clear, 'One risk of maintaining individuals in nonvalidated treatments of panic disorder is that misplaced confidence in the therapy's potential effectiveness may preclude application of more effective treatment.' The statement spells out a specific concern about the use of psychotherapies without demonstrated effectiveness for panic disorder, 'The nature of the therapeutic relationship makes it difficult for the patient to seek additional or alternate treatment.'

As one can imagine, as clinical guidelines and treatment consensus statements continue to emerge for a wide array of emotional disorders, they will have a significant impact on the way clinicians practice psychotherapy. In effect, these documents set standards of care, which if ignored, leave the clinician both ethically and legally vulnerable. Health insurance companies and managed care plans are now providing their practitioners with copies of guidelines and asking that they be followed. The implications are clear: failure to provide an empirically supported treatment from these guidelines, when one exists, may constitute malpractice in the eyes of the payor. For example, a number of states have passed, or are in the process of passing, legislation that give guidelines the 'force of law' by protecting clinicians who follow the guidelines from malpractice litigation (Barlow and Barlow 1995).

3. Identifying, Promoting, and Disseminating Empirically Supported Treatments

Recognizing the lack of utilization of ESTs and considering the trends noted above, the Society of Clinical Psychology (Division 12 of the American Psychological Association) established a task force to identify, promote, and assist in the dissemination of ESTs. The original Task Force on Promotion and

Dissemination of Psychological Procedures (now named the Committee on Science and Practice; CSP) was created by David H. Barlow, then president of the Society. The CSP is now a standing committee in the Society of Clinical Psychology, and the work will continue into the future (the most up-to-date reports and a list of committee members are available on the Internet at www.apa.org/divisions/div12).

As noted in a recent paper summarizing the accomplishments and future directions of the CSP (Weisz et al. 2000), there are two broad goals guiding the CSP: to influence both practice and science. Ideally, the influence on practice would be to improve mental health care for the public by encouraging and supporting empirically supported practice and by fostering clinical training that emphasizes reliance on ESTs. As for science, the CSP hopes to stimulate improvement of the quality and clinical utility of treatment research so that it is most relevant to the concerns of clinicians (see Weisz et al. 2000 for specific details).

3.1 Identifying Empirically Supported Treatments

The first mission of the original Task Force was to define criteria to determine whether treatments qualified as ESTs. The following criteria were established (Chambless et al. 1998). For treatments to be considered by the Task Force, research must have been conducted with treatment manuals. This requirement was set because treatment manuals make for better research design, and the use of a treatment manual allows readers to know what specific procedures have been supported. For example, although many interventions fall under the rubric of behavior therapy (e.g., exposure, relaxation training, cognitive restructuring), only exposure and response prevention has been demonstrated as efficacious for obsessive-compulsive disorder. Stating that behavior therapy is effective for obsessive-compulsive disorder does not allow the reader to know what specific strategies were tested and how they were applied. Use of a treatment manual solves this dilemma.

As for whether or not the treatment qualifies as an EST, the Task Force distinguishes between two levels of EST 'well-established' and 'probably-efficacious.' EST status is determined by the results from group design and single-case design studies. With regard to group design studies, to qualify as a well-established treatment, at least two studies conducted by different investigators must qualify as supporting studies. The following criteria must be satisfied in order to qualify as a supporting study: (a1) treatment is superior to pill placebo or psychological placebo, or (a2) treatment is equivalent to an already established treatment; (b) a treatment manual was used in the study; (c) characteristics of the client sample must be clearly specified. A treatment may not meet criteria for well-established

Table 1

Examples of empirically supported treatments

Well-established treatments

Anxiety and stress:

Cognitive behavior therapy for panic disorder with and without agoraphobia

Cognitive behavior therapy for generalized anxiety disorder

Exposure treatment for agoraphobia

Exposure/guided mastery for specific phobia

Exposure and response prevention for obsessive-compulsive disorder

Stress inoculation training for coping with stressors

Depression:

Behavior therapy for depression

Cognitive therapy for depression

Interpersonal therapy for depression

Health problems:

Behavior therapy for headache

Cognitive-behavior therapy for bulimia

Multicomponent cognitive-behavior therapy for pain associated with rheumatic disease

Multicomponent cognitive-behavior therapy with relapse prevention for smoking cessation

Problems of childhood:

Behavior modification for enuresis

Parent training programs for children with oppositional behavior

Marital discord:

Behavioral marital therapy

Probably efficacious treatments

Anxiety:

Applied relaxation for panic disorder

Applied relaxation for generalized anxiety disorder

Cognitive behavior therapy for social phobia

Cognitive therapy for obsessive-compulsive disorder

Couples communication training adjunctive to exposure for agoraphobia

EMDR for civilian post-traumatic stress disorder (PTSD) Rothbaum

Exposure treatment for PTSD

Exposure treatment for social phobia

Stress inoculation training for PTSD

Relapse prevention program for obsessive-compulsive disorder

Systematic desensitization for animal phobia

Systematic desensitization for public speaking anxiety

Systematic desensitization for social anxiety

Chemical abuse and dependence:

Behavior therapy for cocaine abuse

Brief dynamic therapy for opiate dependence

Cognitive-behavioral relapse prevention therapy for cocaine dependence

Cognitive therapy for opiate dependence

Cognitive-behavior therapy for benzodiazepine withdrawal in panic disorder patients

Community reinforcement approach for alcohol dependence

Cue exposure adjunctive to inpatient treatment for alcohol dependence

Project CALM for mixed alcohol abuse and dependence

Social skills training adjunctive to inpatient treatment for alcohol dependence

Depression:

Brief dynamic therapy

Cognitive therapy for geriatric patients

Reminiscence therapy for geriatric patients

Self-control therapy

Social problem-solving therapy

Table 1
Continued

Probably efficacious treatments (continued)

Health problems:

Behavior therapy for childhood obesity
Cognitive-behavior therapy for binge eating disorder
Cognitive-behavior therapy adjunctive to physical therapy for chronic pain
Cognitive-behavior therapy for chronic low back pain
EMG biofeedback for chronic pain
Hypnosis as an adjunct to cognitive-behavior therapy for obesity
Interpersonal therapy for binge-eating disorder
Interpersonal therapy for bulimia
Multicomponent cognitive therapy for irritable bowel syndrome
Multicomponent cognitive-behavior therapy for pain of sickle cell disease
Multicomponent operant-behavioral therapy for chronic pain
Scheduled, reduced smoking adjunctive to multicomponent behavior therapy for smoking cessation
Thermal biofeedback for Raynaud's syndrome
Thermal biofeedback plus autogenic relaxation training for migraine

Marital discord:

Emotionally focused couples therapy for moderately distressed couples
Insight-oriented marital therapy

Problems of childhood:

Behavior modification of encopresis
Cognitive-behavior therapy for anxious children
Exposure for simple phobia
Family anxiety management training for anxiety disorders

Sexual dysfunction:

Hurlbert's combined treatment approach for female hypoactive sexual desire
Masters and Johnson's sex therapy for female orgasmic dysfunction
Zimmer's combined sex and marital therapy for female hypoactive sexual desire

Other:

Behavior modification for sex offenders; Marshall et al. (1991)
Dialectical behavior therapy for borderline personality disorder
Family intervention for schizophrenia
Habit reversal and control techniques
Social skills training for improving social adjustment of schizophrenic patients
Supported employment for severely mentally ill clients

Source: Chambless et al. (1998)

Citations for evidence supporting each treatment can be found in Chambless et al. (1998). Updates of the EST list will be published on the Internet at www.apa.org/divisions/div12

but still satisfy the less conservative criteria for a probably efficacious treatment: two studies show that the treatment is more effective than a waiting list control group, or only one supporting study exists, or if more than one exists, all were conducted by the same investigator. (For details concerning the definition of EST, including use of single subject designs, and resolution of conflicting data, the reader is directed to the Task Force report itself.)

Based upon these criteria, a list of ESTs was published in 1995 (Task Force on Promotion and Dissemination of Psychological Procedures 1995) and two subsequent updates have appeared (Chambless et

al. 1996, 1998). As can be seen in Table 1, the Task Force has identified more than 70 ESTs (Chambless et al. 1998). It is important to note that it is not assumed that treatments without empirical evidence supporting them are ineffective, but instead, in the absence of scientific data, no conclusion can be drawn about its efficacy or lack thereof.

4. Promotion and Dissemination

Deciding upon criteria and identifying ESTs is only the first step in the process towards increasing their

utilization so that the general public benefits. The second step is to call attention to ESTs so that practitioners, health insurers, and the general public is aware of their efficacy (promotion), and to be sure that practitioners are able to use these treatments (dissemination).

4.1 Promotion

As noted by Barlow (1994), despite the considerable data that exist demonstrating the efficacy of psychological interventions, they seem to be taking a backseat to pharmacological approaches in clinical practice guidelines. Why? Perhaps one reason is that, unlike pharmaceutical companies that spend a significant amount of money promoting the use of their treatments to consumers and providers, no such profit motivated organization exists for psychotherapeutic interventions. As a result, many patients are not aware that effective psychological treatments exist, and thus, do not seek them out. Clearly, more effort needs to be focused educating the public about ESTs. As a first step towards this goal, the CSP is about to launch a website listing a description of ESTs for a range of disorders so that 'consumers' of mental health services can find out what types of treatments have been shown to be effective for the problem/disorder for which they are seeking treatment.

4.2 Dissemination

Identifying and promoting ESTs will not solve all the problems noted above. The next step is to ensure that the treatments are available. Unfortunately, as suspected by the CSP, a recent survey of clinical psychology doctoral programs and internship programs revealed that there is insufficient training in ESTs (Crits-Christoph et al. 1995). For example, as noted above, the two of the first-line treatments for depression listed by the AHCPR Depression Treatment Guideline were cognitive therapy and interpersonal therapy. The survey by Crits-Christoph et al. (1995) revealed that among clinical psychology internship programs—the place where clinical psychologists receive the bulk of supervised clinical experience—only 59 percent of programs provided supervision in cognitive therapy for depression, and a mere 8 percent of programs provided supervision in interpersonal psychotherapy. These are two psychotherapeutic treatments listed as first-line interventions for depression in the AHCPR Depression Treatment Guideline, thus, the lack of training in these areas is of significant concern. University doctoral programs were slightly better; 80 percent providing supervision in cognitive therapy and 16 percent in interpersonal psychotherapy. It is

important to note, however, that having supervision available does not require students to receive it. As a result, these numbers do not indicate that 80 percent of students are in fact receiving training in cognitive therapy, and thus, the percentages are an overestimation of the number of students receiving supervision in that approach. For example, only 14 percent of internship programs required training in cognitive therapy for depression and only 3 percent in interpersonal therapy. Unfortunately, training in ESTs for depression is the highest. As noted in the Crits-Christoph et al. (1995) survey, only a handful of programs require training in a range of ESTs. For example, the survey revealed that only 7 percent of internship sites provide supervision in interpersonal therapy (IPT) for bulimia, 5 percent in CBT for social phobia. The numbers for doctoral programs, though higher, are still not all that promising: 21 percent offer supervision in IPT for bulimia, 24 percent in CBT for social phobia. Each of these treatments represents a first-line intervention for the respective disorder (with some of the treatments being the only empirically supported psychotherapeutic intervention), so it is not as though training in other EST approaches is occurring. Unfortunately, training in other ESTs is similar to the low numbers for depression, bulimia, and social phobia.

If training at the graduate and internship level is lacking, one can imagine how difficult it will be to disseminate these treatments to practitioners already practicing in the field. To date, no formal process exists to disseminate and train ESTs to practitioners. A perfect example of the consequence of this is that despite the strong data supporting cognitive behavioral treatment for panic disorder, a recent survey found that the majority of patients with panic disorder receiving psychological treatment receive dynamic psychotherapy (which has not been empirically supported as a first-line treatment) (Goisman et al. 1999). Indeed, as noted by Calhoun et al. (1998) in a recent paper discussing training in ESTs, reaching professionals out in the field is the most problematic area. With a rapidly changing data base, clinicians trained 10 years ago are unlikely to be up-to-date with the newer, empirically supported psychotherapies. Indeed, the majority of references supporting ESTs have appeared in the past 10 years (Calhoun et al. 1998). How can clinicians become motivated to learn ESTs? Even if they are motivated, how can treatments be disseminated in a way to provide high integrity supervision to insure they are properly trained? The CSP plans to devote a considerable amount of effort to develop a model that would facilitate training professionals in ESTs, because as discussed above, this is an essential activity for the survival of psychotherapy as a viable first-line treatment for emotional disorders. For those interested, a list of manuals outlining ESTs is available (cf. Sanderson and Woody 1995; available at www.apa.org/divisions/div12).

5. Criticism of the EST Movement

The movement towards ESTs are not welcomed by everyone within our profession, and indeed, several authors have raised important concerns and criticisms (e.g., Garfield 1998, Kovacs 1995, Nathan 1998) regarding the identification of 'validated' therapies. Although space does not permit a discussion of the specific issues here, in fact, the CSP agrees that there are difficulties and risks in establishing so-called psychotherapy guidelines—or any treatment guidelines for that matter (cf. Pincus 1994 concerning the development of psychiatry guidelines)—yet the CSP and other professional organizations believe the benefits clearly outweigh the risks (Pincus 1994, Sanderson 1995), and thus, will continue the work outline above (cf. Weisz et al. 2000 for a discussion of the future directions of the CSP).

See also: Adolescence, Psychiatry of; Adolescent Development, Theories of; Adolescent Vulnerability and Psychological Interventions; Antisocial Behavior in Childhood and Adolescence; Clinical Treatment Outcome Research: Control and Comparison Groups; Conduct Disorder; Managed Care and Psychotherapy in the United States; Psychological Treatment, Effectiveness of; Psychological Treatments: Randomized Controlled Clinical Trials; Psychotherapy: Case Study; Psychotherapy: Clinical Practice Guidelines; Social Competence: Childhood and Adolescence

Bibliography

- American Psychiatric Association 1993 Practice guidelines for major depressive disorder in adults. *American Journal of Psychiatry* **150**(4): 1–26 (Suppl. 5.)
- Barlow D H 1994 Psychological interventions in the era of managed care. *Clinical Psychology: Science and Practice* **1**: 109–22
- Barlow D H, Barlow D G 1995 Guidelines and empirically validated treatments: Ships passing in the night? *Behavioral Healthcare Tomorrow* **4**: 25–9
- Bennett M J 1992 The managed care setting as a framework for clinical practice. In: Feldman J L, Fitzpatrick R J (eds.) *Managed Mental Healthcare: Administrative and Clinical Issues*. American Psychiatric Press, Washington, DC, pp. 203–18
- Calhoun K S, Moras K, Pilkonis P A, Rehm L P 1998 Empirically supported treatments: Implications for training. *Journal of Consulting and Clinical Psychology* **66**: 151–62
- Chambless D L, Sanderson W C, Shoham V, Johnson S B, Pope K, Crits-Christoph P, Baker M, Johnson B, Woody S R, Sue S, Beutler L, Williams D, McCurry S 1996 Empirically validated therapies: A project of the division of clinical psychology, American psychological association, task force on psychological interventions. *The Clinical Psychologist* **49**: 5–15
- Chambless D L, Baker M J, Baucom D, Beutler L E, Calhoun K S, Crits-Christoph P, Daiuto A, DeRubeis R, Detweiler J, Haaga D A F, Johnson S B, McCurry S, Mueser K T, Pope K S, Sanderson W C, Shoham V, Stickle T, Williams D A,

- Woody S R 1998 Update on empirically validated therapies: II. *The Clinical Psychologist* **51**: 3–16
- Crits-Christoph P, Frank E, Chambless D L, Brody C, Karp J F 1995 Training in empirically validated therapies: What are clinical psychology students learning? *Professional Psychology: Research and Practice* **26**: 514–22
- Depression Guideline Panel 1993 *Depression in Primary Care: Volume 2. Treatment of Major Depression. Clinical Practice Guideline, Number 5*. Rockville, MD (Department of Health and Human Services, Public Health Service, Agency for Healthcare Policy and Research. AHCPR Publication No. 93-0551). US Government Printing Office, Washington, DC
- Garfield S 1998 Some comments on empirically supported treatments. *Journal of Consulting and Clinical Psychology* **66**: 121–5
- Goisman R, Warshaw M, Keller M B 1999 Psychosocial treatment prescriptions for generalized anxiety disorder, panic disorder, and social phobia. *American Journal of Psychiatry* **156**: 1819–21
- Kovacs A 1995 We have met the enemy and he is us! *The Independent Practitioner* **15**(3): 135–7
- Nathan P E 1998 Practice guidelines: not yet ideal. *American Psychologist* **53**: 290–9
- Pincus H A 1994 Treatment guidelines: Risks are outweighed by the benefits. *Behavioral Healthcare Tomorrow* **4**: 40–45
- Sanderson W C 1995 Can psychological interventions meet the new demands of health care? *American Journal of Managed Care* **1**: 93–8
- Sanderson W C, Woody S 1995 Manuals for empirically validated treatments. A project of the division of clinical psychology, American psychological association, task force on psychological interventions. *The Clinical Psychologist* **48**: 7–11
- Smith G R, Hamilton G E 1994 Treatment guidelines: Provider involvement is critical. *Behavioral Healthcare Tomorrow* **4**: 40–45
- Task force on promotion and dissemination of psychological procedures 1995 Training in and dissemination of empirically-validated psychological treatments. *The Clinical Psychologist* **48**: 3–23
- Weisz J R, Hawley K M, Pilkonis P A, Woody S R, Follette W C 2000 Stressing the other three Rs in search of empirically-supported treatments: Review procedures, research quality, relevance to practice and public interest. *Clinical Psychology: Science and Practice* **7**: 243–58
- Wolfe B E, Maser J D (eds.) 1994 *Treatment of Panic Disorder: A Consensus Statement*. American Psychiatric Association Press, Washington, DC

W. C. Sanderson

Psychological Treatments: Randomized Controlled Clinical Trials

The randomized controlled trial is one method of ascertaining the effectiveness of a treatment. In its simplest form, a controlled trial consists of the recruitment of a group of participants with particular characteristics and their allocation by chance to one of two treatment conditions. These conditions should

differ only in the treatment procedure A whose effects are being studied. Thus one condition will include A, and the other condition will not. Randomization is used to ensure that events other than treatment that might affect outcome will be equally balanced across the two treatment conditions. If measures of key features of the disorder, for example, the principal symptoms, are taken before and after treatment, then it is possible to test whether the null hypothesis of no effect of treatment A is true or not. While more complex experimental designs are often used, and the choice of the control condition is often difficult, all controlled clinical trials adhere to this basic conceptual form.

1. Development of Controlled Clinical Trials in Psychology

Parallel developments in two clinical disciplines, medicine and psychology, beginning in the 1930s, edged both fields toward the use of controlled trials. In an early study spanning both medicine and psychology (Chappell and Stephenson 1936), the authors demonstrated that an educational group program resulted in greater improvement and less relapse in a group of peptic ulcer patients than the control group, both groups having received the same medical care. Another study, the Cambridge-Somerville Youth Project, begun in the late 1930s and continuing with a long-term follow-up into the 1950s, was one of the earliest trials of a psychological treatment (Teuber and Powers 1953). This large-scale, well-designed study of 650 boys thought to be at high risk of developing delinquency was aimed at evaluating a prevention program. The outcome, as these individuals were followed over time, demonstrated that the treatment group fared less well than the control group in terms of a range of delinquent behaviors. Such negative results, while disappointing, are important because they allow ineffective approaches to treatment to be abandoned.

These early controlled trials were the exception rather than the rule. Most studies of the effectiveness of psychological therapies consisted of treating a series of patients without a control group. Hence, it was impossible to be sure whether the results were due to spontaneous recovery or to the treatment, let alone specific aspects of the treatment rather than a non-specific (placebo) effect. Two developments catalyzed the use of the controlled trial. Hans Eysenck, Professor of Psychology at London University, reviewed the psychotherapy treatment literature in 1952 and again in 1966 (see Eysenck 1952, 1966). On both occasions Eysenck concluded that there was little evidence that psychotherapy was effective, that is, that the results did not exceed the expected rate of spontaneous recovery. While this opinion resulted in a vigorous debate and defense of psychotherapy, including the

addition of a few treatment trials that demonstrated some success for various forms of psychotherapy that Eysenck had not considered, Eysenck's conclusions were widely accepted, reaffirming the need for well-conducted controlled trials. Some, however, argued that the effects of psychological therapies, because they were aimed at 'inner' essentially unmeasurable changes, could never be adequately assessed in a controlled trial. This problem was largely overcome by the next development.

The second, and most important, catalyst for the use of controlled trials of psychological treatments, was the emergence of behavior therapy (see *Behavior Analysis, Applied; Behavior Therapy: Psychological Perspectives* for further details). Eysenck 1966, p. 39) had concluded that 'With the single exception of the psychotherapeutic methods based on learning theory, results of published research ... suggest that the therapeutic effects of psychotherapy are small or non-existent ...' Because behavior therapy was based on learning theory, the focus of treatment was directly on the problem behaviors comprising the clinical syndrome. This allowed for the development of objective measures of treatment outcome (see *Behavioral Assessment* for a fuller discussion of objective evaluation methodology). In addition, because treatment was scientifically based, its practitioners were more likely to use scientific methods, including the randomized controlled trial, to assess the effectiveness of these new treatments.

2. The Flow of Research

Although the controlled trial is the definitive method of ascertaining the effectiveness of a treatment, it is not the only method used in the research process aimed at the development of a new therapy. The first step in the development of a new therapy, the initial formulation of its conceptual basis and procedures, may stem from basic psychological research or theory, and/or from clinical observation. The first step in the development of a new therapy is usually an uncontrolled observational study, namely treatment of a series of patients. If a sufficient number of patients respond favorably to the new treatment, then the next step is to document the treatment procedures in a manual so that it can be replicated by others. The first controlled study may compare treatment to no treatment, a step that is relatively easy to carry out with a small number of patients, providing information as to whether the new treatment is better than doing nothing, and allowing the results to be roughly compared with those of standard therapy for the same condition, if such a therapy exists. One objective method of comparing the effects of different therapies is to calculate the effect size (the change from pretreatment to post-treatment in a group of patients divided by the standard

deviation of the change). An effect size of 0.3 would be regarded as small, one of 0.5 as moderate, and 1.0 as large.

The next step is to ascertain whether the promising new treatment has specific effects over and above what is usually termed the placebo or nonspecific effect (see Sect. 3). Hence, it is important to compare the new treatment to a second treatment containing nonspecific features common to all therapies, but not containing the specific procedures aimed at ameliorating the behaviors being studied. Single case controlled studies may also be useful at this stage of development to more precisely delineate the active components of treatment (see *Single-case Experimental Designs in Clinical Settings*).

Once the new treatment has been demonstrated to have specific effects on a particular disorder, the next step will probably consist of a comparison with the standard therapy for the condition. At this stage related studies are likely to be carried out at different centers. The results of such studies can be compared using a procedure known as meta-analysis in which effect sizes are compared across studies. It may be clear from the results of such an analysis that the new treatment is better or worse than standard treatment, or that it is useful for a particular subgroup of individuals, or that it has advantages such as being less costly but similarly effective. On the other hand the results of such comparisons may be unclear. Some studies may demonstrate positive effects while others may show no effect. Such conflicting results are often due to the initial studies having too few patients enrolled to determine accurately the effectiveness of the new treatment as compared with other treatments. In this case a study involving several clinical sites may be needed to recruit a sufficient number of patients in a short enough time, to obtain more definitive results.

2.1 Multisite Studies

Multisite studies are necessarily more complex to carry out than a study at a single site. They are also more expensive than single-site studies and for that reason should only be entered upon when the preliminary work described above has been completed, and the need for such a study has been precisely formulated. The investigators involved must agree to a common experimental protocol. The treatment(s) being studied must be well documented in a treatment manual and therapists at each of the sites trained in the delivery of the treatments. Moreover, the therapists should be carefully supervised throughout the study to ensure that they are delivering the treatment(s) in accord with the protocol. Tape recordings of treatment sessions may also be audited to ensure fidelity to the treatment. In addition, the assessments used must be carried out in a standard manner and done com-

parably between sites. This may mean that assessors will also be trained and supervised across sites in the particular assessment methods. One advantage of a multisite over a single-site study is that the effects of treatment are examined in different sites that may have unique characteristics in terms of patient population, therapist characteristics, and so on. Hence, generalizability of the results of treatment to the clinical world is greater than in the single-site study.

Another reason to conduct a multisite study is when treatment for a relatively rare disorder is being studied. If the study were to be done at only one site it might take many years before sufficient participants entered the study. A multisite study would provide more rapid results. Another reason for considering a multisite study is when the question of interest shifts to what treatment or treatments might benefit those who do not respond to an initial effective treatment. Because only treatment failures will be entered into such a study, a large number of individuals must be treated in order to generate a sufficiently large sample for the study.

2.2 Effectiveness Research

As noted above, a distinction has arisen between efficacy and effectiveness controlled trials. Trials to establish efficacy are considered to be carried out in refined settings with carefully chosen patients and expert therapists. Effectiveness studies on the other hand are considered to be carried out in clinical practice settings with 'real life' patients and possibly less expert therapists. Some would consider efficacy studies to be of limited value in assessing the outcome of psychological therapies in contrast to effectiveness studies. However, if the flow of research described above is taken into consideration it is evident that efficacy and effectiveness studies lie on a continuum rather than being different types of studies, and that research along the entire continuum is needed. Moreover, the scientific questions being studied differ along the continuum that constitutes the flow of research.

Once a psychological treatment has been established as useful for a particular disorder in a number of single-site studies (efficacy studies) and, if needed, a multisite study, research as to how to transfer the treatment to other settings may be needed (effectiveness studies). This step in the development of a treatment may involve controlled studies of modifications of the treatment. Such modifications may include simplification of the treatment for wider dissemination to less skilled therapists and investigation of the necessary clinic environment to support the therapy. For example, in most controlled trials therapists are carefully supervised, and the question arises whether such supervision is essential to the effectiveness of the therapy. Additionally, the type of

patient encountered in different clinical settings may differ, again requiring modification of the therapeutic procedures. Once these questions have been resolved, often using controlled trials, an effectiveness study can be designed in one or more clinical settings using the staff of such clinics as the therapists, and enrolling all the patients with the particular disorder into the study as they appear in the clinic. A control condition in such cases might be the usual care that participants would receive in such a clinic.

3. Essential Elements of a Controlled Trial

While the basic design of two groups differing only in A (the components of a new treatment) suffices for early stages of research, other questions will need to be addressed. Among the most problematic is how to control for nonspecific elements of psychological treatments. Treatments can be regarded as a package containing both specific procedures aimed at the particular problem, together with elements common to all therapies, such as therapist attention, and therapeutic instructions that include the rationale for therapy together with engendering an expectancy that treatment is likely to be successful.

Research has demonstrated that nonspecific elements lead to significant behavior change. Hence, it is important to know to what extent procedures specific to the therapy contribute to a successful outcome over and above nonspecific elements. This raises problems for psychological therapies because there is no placebo pill. This problem has been solved in several ways. The first solution is to use a nonspecific therapy, that is, one without the specific therapeutic procedures being examined but which also has face validity as a therapy with which to compare the active treatment. Some controlled trials may compare a psychotherapeutic procedure with medication and with a pill placebo as a control condition, thus solving the nonspecific effect question in a different manner. Finally, it is possible to compare a new therapy with a standard therapy for the condition which has already been demonstrated more successful than a nonspecific therapy.

3.1 The Use of Treatment Manuals

The reports of early studies of psychological therapies usually described the treatment in a paragraph or two. Such descriptions are obviously insufficient to allow other investigators or therapists to provide the same treatment to their patients. The treatment manual was developed to solve this problem. Such manuals, now regarded as an essential element of a controlled trial of psychological therapies, describe the treatment pro-

cedures in sufficient detail for others to replicate them. Therapists are carefully trained in the procedures, which the manual facilitates.

There is some controversy regarding the clinical use of treatment manuals, which have now become a common method of disseminating new therapies to clinicians. Some argue that manuals lock the therapist into a rigid method of providing treatment, thus eliminating the opportunity to individualize the treatment. Others suggest that manuals provide essential guidelines for therapists and that the experienced therapist will be able to individualize treatment to the extent necessary. There is relatively little experimental data on the effectiveness of the use of manuals in the clinic; however existing data suggests that manuals are likely to enhance therapy probably by ensuring that the therapist stays on track during the treatment process. See *Clinical Psychology: Manual-based Treatment* for a fuller discussion of this important issue.

3.2 Controlling the Fidelity of Treatment

A further essential element of a clinical trial is a method to ensure that the therapy provided is faithful to the treatment manual. This is usually ensured in two ways. First, therapists are trained in the use of the manual and practice the treatment on several patients before the controlled trial begins. During this period the therapists are supervised by an expert. Second, once the trial begins, therapy sessions are audiotaped, and a random sample of such tapes is audited. The auditor will use a checklist to ensure that all the elements of treatment are being provided, and that no foreign elements have crept in. Regular feedback is provided to the therapists, and the degree of therapist compliance with the treatment protocol should be reported in the paper describing the study.

4. Outcomes of Controlled Trials

There are several potential outcomes from a randomized clinical trial. First and foremost is whether the treatment is effective as compared to the control group; second, whether there are specific predictors as to which patients do best with the treatment under study; third, whether variables that mediate the outcome of treatment can be identified. Such mediators may throw light on how a given therapy works.

4.1 Assessment of the Primary Outcome

The primary outcome for a controlled trial is the main indicator or indicators that characterize the syndrome.

For example, for panic disorder, the number and severity of panic attacks would comprise the main outcome; for depression a measure of the severity of depression; for bulimia nervosa, a measure of binge eating and purging. Because these measures test the primary hypothesis addressed, that is, the null effect of the treatment in question as compared with a control group, the study should stand or fall depending on whether the primary outcome is positive or not. If the primary outcome is positive then secondary measures—usually associated psychopathology—are also examined for the effect of treatment. Because some symptoms such as mild anxiety or depression are normative in a community, the extent to which the treated population now approximates community norms may also be reported.

4.2 Examination of Predictors

It is usual in a controlled trial to assess a series of variables that are hypothesized to affect the outcome of treatment. Examples of potential predictors include: the duration and severity of the problem, the age and social class of the patient, and associated psychopathology. If predictors of sufficient strength are found they may be used in future trials to select patients most likely to succeed with the treatment, thus refining the indications for treatment. In general, the search for such predictors has not been very successful. Usually such predictors are sufficient for a general description of the patient most likely to improve, but not specific enough to identify such patients before treatment begins. An alternative method of prediction is to examine the progress of patients during treatment. In many psychological treatments patients obtain significant benefit early in therapy. Patients who do not demonstrate this early response tend to do poorly. Hence, it may be possible to identify those most likely to do badly with a particular psychotherapy early in treatment and to use a different treatment approach with such patients.

4.3 Identifying Treatment Mediators

A mediator is a variable that is responsible for an effective treatment outcome. Hence, it is a result of a therapeutic procedure that occurs prior to the benefits of treatment and that is significantly related to the outcome. Rather than measuring such variables before treatment begins, as is the case with predictors, potential mediators are measured early in the course of treatment. Potential mediators are usually selected based on precise hypotheses about how the treatment works. The identification of mediators can throw important light on the mechanisms underlying the

effectiveness of a given psychological treatment and may provide insight as to how to improve the treatment. See also *Psychotherapy Process Research* in which interactions between therapist and patient, and their effect on outcome, form the question of interest.

5. Are Controlled Psychological Treatment Trials Useful?

While controversy over the use of controlled trials to evaluate psychological treatments continues, the argument has shifted as to whether such trials can reliably inform clinical work and social policy. Some argue that uncontrolled trials conducted in 'real' clinical settings with real patients would be more informative. This argument ignores the necessity for a sequenced flow of research in the development of a new psychological therapy. Moreover, without a control group, it is impossible to determine whether it is the treatment or some other variable that is responsible for the outcome. Despite this ongoing controversy, the controlled trial has become the accepted method of evaluating the effectiveness of psychotherapy. Most individuals would prefer that their medical treatment was informed by controlled therapeutic trials; similarly, consumers of psychotherapy should demand that these procedures are well tested.

6. Future Developments

It is likely that the use of randomized controlled trials to evaluate the effectiveness of psychological treatments will accelerate. At present a relatively large number of treatment procedures have been identified as effective for specific psychological problems (see *Psychological Treatment, Effectiveness of; Psychological Treatments, Empirically Supported* for a detailed review of such treatments). Many, but not all, such therapies fall under the rubric of behavior therapy. These research efforts have now generated a small, but growing number of multisite trials in which the treatment is put to a more definitive test, compared with another effective psychotherapy, or with medication either separately or in combination with psychotherapy. As is the case in medical research, it would be expected that this trend will continue, and that multisite trials will be increasingly used in psychotherapy research.

Because no one psychotherapy is effective with all those suffering from a particular psychological disorder, research is beginning to examine the efficacy of a second treatment approach for those who fail initial therapy. Moreover, there is much interest in the effectiveness of therapist-supported bibliotherapy.

Such developments may ultimately lead to algorithms guiding the manner in which psychological treatments should be sequenced for the treatment of a given condition. For example, treatment might begin with the least costly treatment followed by increasingly expensive (and probably more effective) treatments for those who fail one or more initial treatment.

In the somewhat further future the biological mechanisms through which psychotherapies work will become a focus of research, ultimately bringing together psychological and biological mechanisms underlying behavior change.

See also: Clinical Treatment Outcome Research: Control and Comparison Groups; Psychological Treatment, Effectiveness of; Psychological Treatments, Empirically Supported; Single-case Experimental Designs in Clinical Settings

Bibliography

- Beutler L E, Crago M 1991 *An International Review of Programmatic Studies*. American Psychological Association, Washington, DC
- Campbell D T, Stanley J C 1963 *Experimental and Quasi-Experimental Designs for Research*. McNally, Chicago, IL
- Chappell M N, Stephenson I T 1936 Group psychological training in some organic conditions. *Mental Hygiene*, **20**: 588–97
- Eysenck H J 1952 The effects of psychotherapy: An evaluation. *Journal of Consulting and Clinical Psychology* **16**: 319–24
- Eysenck H J 1966 *The Effects of Psychotherapy*. International Science Press, New York
- Goldfried M R, Wolfe B E 1998 Toward a more clinically valid approach to therapy research. *Journal of Consulting and Clinical Psychology* **66**: 143–50
- Kazdin A E 1998 *Research Design in Clinical Psychology*, 3rd edn. Allyn and Bacon, Boston, MA
- Kraemer H C, Prunyn J P 1990 The evaluation of different approaches to randomized clinical trials. *Archives of General Psychiatry* **47**: 1163–9
- Persons J B, Silberschatz G 1998 Are results of randomized clinical trials useful to psychotherapists? *Journal of Consulting and Clinical Psychology* **66**: 126–35
- Russell R L, Orlinsky D E 1996 Psychotherapy research in historical perspective: Implications for mental health care policy. *Archives of General Psychiatry* **53**: 708–15
- Teuber N L, Powers E 1953 Evaluating therapy in a delinquency prevention program. *Proceedings of the Association for Research in Nervous & Mental Disease* **31**: 138–47
- Ullman L P, Krasner L 1965 *Case Studies in Behavior Modification*. Holt, Rinehart and Winston, New York
- Wilson G T 1998 The clinical utility of randomized controlled trials. *International Journal of Eating Disorders* **24**: 13–29

W. S. Agras

Psychologism, History of

The history of psychologism is the history of an accusation and of a threat. To label an author an advocate of psychologism is to claim that he or she uses psychological concepts and ideas in the wrong place. The accusation first became popular in the German-speaking world with the rise of experimental psychology in the second half of the nineteenth century. The most famous episode in the history of the accusation is the German ‘psychologism controversy’ at the beginning of the twentieth century. It will be the central topic of this article. It is widely accepted that this controversy played an important role in shaping the character of Western philosophy in general and analytic philosophy in particular. It also influenced subsequent debates over ‘sociologism.’

1. The Psychologism Controversy in Early Twentieth-century Germany

From the 1880s onwards, a number of German logicians and epistemologists started to attack some of their philosophical colleagues for presenting and endorsing the following kinds of ‘psychologistic’ arguments:

Logic studies specific ‘laws of thought.’ Psychology is the study of all kinds of laws of thought. *Therefore*, logic is a part of psychology.

Normative disciplines must be based upon descriptive-explanatory disciplines. Logic is a normative discipline with respect to human thinking. Psychology is *the* descriptive-explanatory discipline with respect to human thinking. *Therefore*, logic must be based upon psychology.

Logic is concerned with judgments, concepts, and inferences. Judgments, concepts, and inferences are human mental entities. *Therefore*, logic is a theory about human mental entities.

The touchstone of logical truth is self-evidence. Self-evidence is a human mental experience. *Therefore*, logic is about a human mental experience (and thus a part of psychology).

We cannot conceive of a radically different logic. The limits of conceivability are our mental limits. *Therefore*, logic is relative to the human species, and thus a subfield of human psychology.

In some cases there was considerable dispute concerning the question who actually held these views. The first argument can be found in Theodor Lipps and Gerardus Heymans, the second in Wilhelm Wundt, the third in Wilhelm Jerusalem and Christoph Sigwart, the fourth in Theodor Elsenhans, and the fifth in Benno Erdmann. All the accused had published widely on psychological issues, and some of them openly advocated treating psychology as the foundation of philosophy.

Arguments against these and other allegedly psychologistic authors were provided by many writers, amongst them the logician and mathematician Gottlob Frege, the phenomenologist Edmund Husserl, and the Neokantians Hermann Cohen, Paul Natorp, Heinrich Rickert, and Wilhelm Windelband. Of these and the many other critics, Husserl was undoubtedly the most successful. Husserl's arguments appeared in the *Prolegomena* to his *Logische Untersuchungen*. The following arguments figured centrally in the debate:

If logical rules were based upon psychological laws, then all logical rules would have to be as vague as the underlying psychological laws. *But*, not all logical rules are vague. *Therefore*, not all logical rules are based upon psychological laws.

If laws of logic were psychological laws, then they could not be known *a priori*. They would be more or less probable rather than certain, and justified only by reference to experience. *But*, laws of logic are known *a priori*; they are justified by apodictic self-evidence, and certain rather than probable. And *therefore*, laws of logic are not psychological.

If logical laws were psychological laws, they would refer to psychological entities. *But*, logical laws do not refer to psychological entities. And *therefore*, logical laws are not psychological laws.

All forms of psychologism imply relativism. Psychologistic thinkers take logic to be true only for the human species. But this contradicts the meaning of truth. Truths are unchanging, and independent of the human mind.

A good deal of the psychologism controversy concerned the validity and originality of Husserl's arguments. Several commentators maintained that Husserl's first three arguments all begged the question. Others challenged Husserl's claim that the laws of logic did not imply matters of fact. Husserl's critics also denied that truths were eternal or independent of human thought.

The debate over psychologism was vigorous. This can be seen, for instance, from the fact that more than 130 philosophers were accused of advocating psychologism and that there were as many symptoms and definitions of psychologism as there were philosophical schools (Kusch 1995). It was not merely the relationship between logic and psychology that was at issue, but the relationship between psychology and philosophy as a whole. This was a pressing issue not least because the practitioners of the then new (experimental) psychology worked in philosophy departments, and because between 1873 and 1913 the number of full professorships held by these 'psychologists' increased from one to ten. Of the 39 full professorships in philosophy in 1892, practitioners of experimental psychology held three; of the 42 full professorships in 1900 they occupied six; while of the 44 full professorships in 1913 they had already gained 10. Practitioners and advocates of the new psychology sometimes

claimed that psychology provided the key to real progress in philosophy or that it was the heart of philosophy (Ash 1995, Kusch 1995, 1999).

This last-mentioned contention was opposed by a quickly growing group of 'pure' philosophers—led by the Neokantians, and by Eucken, Dilthey, and Husserl. They deemed it harmful to philosophy and its progress that experimental psychology should be regarded as part and parcel of philosophy. To put it another way, these philosophers believed that the role of the philosopher, as suggested by the academic success of experimental psychology, needed purification.

The prolonged debate between the advocates and opponents of psychology as a philosophical discipline was eventually brought to an end by World War I. The war brought about an atmosphere in which attacks on one's colleagues were regarded as utterly inappropriate. Moreover, it also led to a clear division of labor between pure philosophers and psychologists: while pure philosophy concentrated on the ideological task of celebrating the German 'genius of war,' experimental psychology focused on the training and testing of soldiers. After the war, both academic pure philosophy and experimental psychology had to cope with, and accommodate to, an intellectual environment that was hostile to science, rationality, and systematic knowledge. The project of a scientifically minded philosophy with experimental psychology as its central pillar quickly lost support. Advocates and practitioners of experimental psychology therefore had to find new ways of justifying their work. Many of them continued their wartime involvement with applied psychology (Kusch 1995).

2. The Influence of the Psychologism Controversy

2.1 Psychologism and Sociologism

The way the psychologism controversy was fought by the defenders of philosophical purity became an important model for subsequent disputes between advocates and opponents of 'naturalism' (the view that philosophical questions can be answered using the results of the natural and social sciences). Many later arguments against 'sociologism' or 'biologism' closely mirror Frege's and Husserl's ideas.

For instance, in the German-speaking world the psychologism controversy of the first two decades of the twentieth century was followed by the dispute over the nature and philosophical aspirations of the sociology of knowledge. Accusations of 'sociologism' surfaced quickly and 'sociologism' was seen by many as the 'successor' to psychologism. Again many philosophers felt that a special science—this time sociology—was trying to occupy philosophical terrain.

Critics of the sociology of knowledge accused it of both psychologism and sociologism, and in so doing, they drew on Husserl's work in particular. Defenders of the sociology of knowledge tried a similar feat in presenting their work as a bulwark against psychologism (Meja and Stehr 1990). This idea has become a general feature of the field to this day (Bloor 1983). The close link between the two 'isms' has been transmitted to subsequent Anglophone thinking mainly through the work of Sir Karl Popper.

2.2 Psychologism and Philosophy

German-speaking philosophy of the turn of the century exerted a powerful and lasting influence on philosophy in many other countries. Frege's and Husserl's (1970) antipsychologism thus became the dominant view in Western philosophy for much of the twentieth century. This is not to deny that other countries also had their own debates over the relation between psychology and philosophy. In the UK, for instance, George Frederick Stout and James Ward emphasized the philosophical significance of psychology; British Idealists like Francis Herbert Bradley and Brand Blanshard were sympathetic to psychology even though they denied its philosophical credentials; and realists like John Cook Wilson, Bertrand Russell, Herbert Joseph, and Herbert Prichard insisted on a separation. For Prichard, psychology was not even a 'proper science' (Passmore 1994).

The institutional separation of psychology from philosophy eventually removed one important source of conflict. Once institutional separation had been achieved, 'pure' philosophers and 'impure' philosophers (i.e., psychologists) were no longer in direct competition over academic chairs. On the plane of ideas, however, the psychological challenge has remained. Throughout the twentieth century, psychology exerted a powerful pull on philosophers who wished to put philosophy on the sure path of a natural science. Willard Van Orman Quine explicitly called for a return to psychologism in the 1960s. Such calls have kept the psychologism accusation alive. Amongst the accused one can find the names of Rudolf Carnap, Michael Dummett, Peter Geach, Nelson Goodman, Thomas Kuhn, John McDowell, Karl Popper, Wifrid Sellars, and Ludwig Wittgenstein.

It is now common amongst historians of philosophy and psychology to lament the outcome of the psychologism controversy. The separation of philosophy from psychology delayed the development of research programs like 'cognitive science' that combine psychological experimentation with philosophical aspiration.

See also: Determinism: Social and Economic; Experimentation in Psychology, History of; Husserl,

Edmund (1859–1938); Knowledge, Sociology of; Logic and Cognition, Psychology of; Psychology: Historical and Cultural Perspectives; Psychology: Overview

Bibliography

- Ash M G 1995 *Gestalt Psychology in German Culture 1890–1967: Holism and the Quest for Objectivity*. Cambridge University Press, Cambridge, UK
- Bloor D 1983 *Wittgenstein: A Social Theory of Knowledge*. Columbia University Press, New York
- Frisby D 1983 *The Alienated Mind: The Sociology of Knowledge in Germany 1918–33*. Heinemann, London
- Haaparanta L (ed.) 1994 *Mind, Meaning and Mathematics: Essays on the Philosophical Views of Husserl and Frege*. Kluwer, Dordrecht, The Netherlands
- Husserl E 1970 *Logical Investigations*, Vol. I, trans. Findlay J N. Routledge, London
- Kusch M 1995 *Psychologism: A Case Study in the Sociology of Philosophical Knowledge*. Routledge, London
- Kusch M 1999 *Psychological Knowledge: A Social History and Philosophy*. Routledge, London
- Meja V, Stehr N 1990 *Knowledge and Politics: The Sociology of Knowledge Dispute*. Routledge, London
- Notturmo M A 1985 *Objectivity, Rationality and the Third Realm: Justification and the Grounds of Psychologism: A Study of Frege and Popper*. Nijhoff, Dordrecht, The Netherlands
- Passmore J 1994 *A Hundred Years of Philosophy*, 2nd edn. Penguin, London
- Popper K 1945 *The Open Society and Its Enemies, Vol. II, The High Tide of Prophecy: Hegel, Marx, and the Aftermath*. London: Routledge

M. Kusch

Psychology and Economics

Economics without psychology has not succeeded in explaining important economic processes and 'psychology without economics' has no chance of explaining some of the most common aspects of human behavior (Katona 1951).

Psychology as the science of human behavior naturally also encompasses the study of economic behavior. From this perspective, there is a considerable overlap between the two disciplines of psychology and economics, as manifested in such subdisciplines as economic psychology, behavioral or psychological economics, economic cognition, and other hybrids. Despite this multiple overlap, both disciplines use rather different approaches, rest on entirely different

assumptions, and focus on different phenomena of interest. This article first portrays the points of contact and difference between psychology and economics. A generalized exchange assumption is then identified as a theoretical interface connecting both disciplines. The implications of this basic exchange assumption afford an organizing framework for a review of empirical research findings.

1. Comparison of the Two Disciplines

1.1 Points of Contact

As disciplines psychology and economics are multiply interrelated. Both serve each other as explanans and explanandum. First and foremost, psychological theories and knowledge have been amply applied to explain problems genuinely studied in economics. Vice versa, economic theory has also been applied to some fields located traditionally more within the realm of psychologists rather than economists. In most general terms, one might say that both disciplines meet in the area of decision making.

Psychological theories and knowledge help to explain how scarce resources are handled so as to satisfy needs. This approach focuses on psychological mechanisms and processes underlying consumption, investment, saving, and other economic behavior. Although early classics of economic theory explicitly or implicitly acknowledged psychological variables, there was no attempt to incorporate the psychological research systematically. Only Katona (1951) recognized that purely economic variables such as income insufficiently explain economic behavior such as consumption. He assumed that psychological variables such as affect, attitudes, expectations, and others intervene between economic stimuli and consumer reactions. Until today, the study of consumer behavior including marketing and advertising (see *Consumer Psychology; Psychology and Marketing; Advertising, Psychology of*) has remained by far the most successful and most pervasive application of psychology to economic behavior. This is particularly so in the area of consumer information processing (CIP), which exclusively relies on psychological theory. CIP deals with how consumers attend to, encode, store, and retrieve product relevant information and how they integrate this information to arrive at a judgment or choice. Examples of other interdisciplinary applications are fiscal psychology, that deals with attitudes towards taxes and the behavior of tax evasion, environmental behavior such as energy conservation or contingent valuation, pricing in business, household saving, and investing.

Whereas psychologists study stock market behavior, Gary Becker (1976) applied a purely economic approach to explain marriage, crime, and having

children—all behaviors that have traditionally been of more interest to psychologists than to economists. By assuming a maximizing utility behavior, stable preferences, and a market equilibrium—the key assumptions in economics—a wide range of human behaviors can be integrated and explained in terms of costs and benefits (Levinger 1965).

Another point of contact concerns the impact of economic variables on states and behaviors that are significant for the individual's well-being such as poverty, debt, or unemployment as related to happiness, physical and psychological health, and suicide. Conversely, more psychological variables such as achievement motivation or religious conviction can have an impact on economic success.

1.2 Points of Difference

As described above, psychology and economics differ not so much in their content domains but in the research methods and theoretical premises and frameworks. First of all, whereas psychologists study *individual* behavior, the research object of economists are *markets*, not individuals or even households. This distinction is most important as other major disagreements and often seemingly irreconcilable differences stem from it. Economists presuppose that markets behave rationally, following rules that maximize utilities. Although individuals may not adhere to these rules—as psychology has shown over and over—the assumption is that irrationalities on the individual level will cancel each other on the market level, resulting in rational markets. This assumption is unrealistic from psychology's view. A whole research tradition has accumulated evidence for how individuals deviate from normative rules not at random but in a systematic fashion. Given such systematic biases in judgment and decision-making, cancellation at the aggregate level becomes an illusion.

Psychology is inductive whereas economics is deductive. Psychologists observe individual behavior in order to test hypotheses concerning the underlying mediating processes. Although these hypotheses are often derived from higher-order theories, these theories are typically domain-specific and controversial, rather than universal and commonly shared. Economists on the other hand are more interested in predictors rather than causes, are more interested in outcomes rather than processes, and share widely accepted key assumptions such as utility maximizing behavior. Based on these key assumptions, economists deduce formalized models to predict particular outcomes. Whether the predictors are causally relevant at the individual level is only of interest to the extent that such causal mediators result in accurate models.

Clearly, this juxtaposition describes the extremes. Economists and psychologists have developed some common ground, combined interests, and a fruitful

dialog, mainly incorporating psychology into economic decision making. Some of this work will be reviewed in the following. However, given that psychology is mainly concerned with individual behavior and economics with aggregate behavior, it is interesting that the mediating link—how individual behavior translates into aggregate behavior—has largely been neglected by both sides and their respective interdisciplinary approaches. Aggregate behavior, however, is clearly not simply aggregated individual behavior. For example, although consumer sentiment predicts consumer spending at the aggregate level—and is indeed a better predictor than income or other economic variables—it does not do so at the individual level.

Research methods constitute an additional source of divergence. Psychological research relies mainly on highly controlled experiments mostly conducted in the laboratory. The experimental situation is quite often rather different from the setting where the corresponding behavior naturally occurs. In particular, economists have questioned whether the use of deception in experiments leads to valid research inferences. Nevertheless, a considerable part of economic research has adopted the experimental method from psychology, with and without deception.

2. Exchange Rules as an Interface for Research in Psychology and Economics

The common denominator of most empirical research in the interface of psychology and economics lies in the assumption of a basic *exchange paradigm* that governs all kinds of research topics such as pricing, marketing, consumer choice, fairness of negotiation, coalition formation, investment, etc. The ultimate goal of economical exchange is to minimize the total input to actions and transactions and to maximize the total output combined over several dimensions or resources (money, effort, time, security, etc.). Within this paradigm, success means to obtain high output from low input, failure means that the output is low relative to the input. Fairness means that different parties' (actors, companies') output–input ratio is balanced. Note that in order to calculate an output–input ratio, it is essential that resources of all kinds be translatable to a joint scale so that effort can be paid in money, security can be traded against reduced profit, information costs are compensated by higher quality, and innovative ideas increase product quality, market price, and organizational efficiency (i.e., reduce workload and coordination loss). Only when all these different kinds of resources can be exchanged and can compensate each other, different action alternatives or decision options can be represented on a monotonic preference scale. Monotonicity or equivalence is indeed a precondition for all rational behavior within this basic exchange paradigm.

The rationality principle that underlies models of economic behavior as well as the irrational tendencies emphasized in psychological approaches revolve around diverse aspects of monotonicity. Indeed, a review of research on psychology and economics can be organized in terms of rules and violations of monotonicity on the one hand and problematic aspects of the underlying normative monotonicity assumptions on the other.

3. Rules and Violations of Monotonicity: Research Examples

3.1 Transitivity of Preference

Any quantitative scaling of economic goods (e.g., products, shares, investment options, job applicants) a, b, c must fulfill the transitivity constraint: If $a < b$ and $b < c$, then $a < c$. For instance, if the price that can be obtained for object b is higher than the price for a , and the price for c is even higher than for b , then c must also obtain a higher price than a . Otherwise, rules of exchange would either break down completely, or the joint scaling of prices, preferences, and attitudes would be severely complicated. However, in reality, this seemingly obvious normative rule can be violated systematically. Psychological models predict intransitivities under specified conditions. In a multi-attribute utility framework, each object a, b, c is characterized by multiple aspects, describing advantages or disadvantages. When comparing a pair or subset of objects, some aspects are more salient and receive higher weights than other objects. Intransitivities can arise when the preference of b over a and of c over b is due to different aspects than the comparison of a and c . For instance, when choosing among three meals in a restaurant, somebody may prefer b over a because b seems more delicious, and c over b because c seems even more delicious. But when comparing c to a , the price difference may become more salient, and a may now be preferred to c because a is less expensive.

3.2 Independence of Context

The independence principle implies that preference orders be invariant over context conditions; if $a < b$ in context x , then $a < b$ should also hold in context y . Again, empirical evidence and multi-attribute theories are in conflict with this premise of consistent exchange. Let a and b represent two cars the value of which is mainly determined by two attributes, performance and economy. Let a be superior in performance and b superior in economy. The relative advantage of either car in one attribute exactly compensates for its disadvantage in the other attribute so that a and b are represented on the same iso-

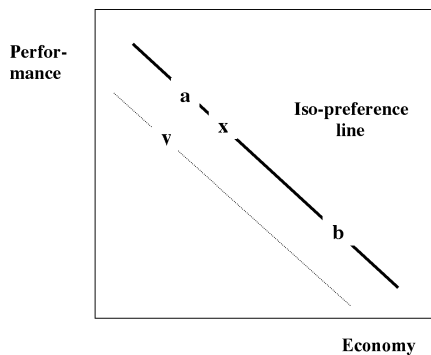


Figure 1
Graphical illustration of similarity and attraction effects

preference line (Fig. 1). Given such a choice situation, the likelihood of consumers to choose *b* rather than *a* depends crucially on the presence of further alternatives. If the comparison is made in the context of a third car *x* that lies on the same iso-preference line but much closer to *a* than to *b*, then the context stimulus *x* becomes a rival for cars like *a*, inducing a tendency to prefer *b* over *a*. In contrast to this so-called ‘similarity effect,’ a context stimulus *y* that also lies close to *a*, but on a slightly inferior iso-preference line, functions like a decoy that serves to highlight the dominance of *a* over context stimuli and thereby increases the likelihood to choose *a* rather than *b*. This opposite influence of context has been termed ‘attraction effect’ (Huber and Puto 1983). Context effects like these have strong implications for marketing, consumer choice, and investment decisions involving similarly attractive multi-attribute options. For example, by including a new car in the range of an automobile manufacturer, the attraction of an already existing car may increase (attraction effect) or decrease (similarity effect).

3.3 Framing of Decision Options

Verbal labeling or circumscribing of economic targets must not affect their monotonic ordering for rational exchange to work properly. However, countless studies demonstrate preference reversals as a function of mere verbal framing. Decisions under uncertainty often involve the choice between two largely equivalent options, a P-option with a relatively high probability of success but a relatively low outcome, and an O-option with a higher outcome in case of success but a relatively low success probability. The tendency to choose the more risky O-option depends strongly on whether the verbal framing focuses on positive or negative outcomes. If the decision problem is framed positively in terms of possible gains, a

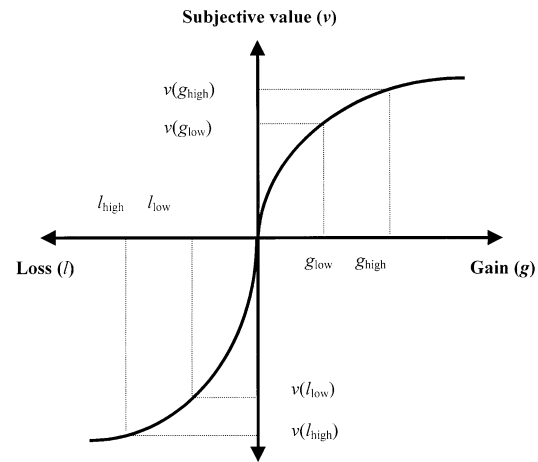


Figure 2
S-shaped subjective value function, according to prospect theory

majority of people would show risk aversion and choose the more secure P-option. If, however, the same decision problem is framed negatively in terms of the possible losses, preference for the O-option becomes more likely.

To account for such framing effects, psychologists and economists refer to prospect theory (Kahneman and Tversky 1984)—a very prominent model of economic choice. In particular, outcome framing effects are commonly explained by the assumption of an S-shaped subjective value function in which both the positive value of gains and the negative value of losses increase in a negatively accelerated fashion, as portrayed in Fig. 2. Twice as extreme an outcome is not worth twice as much, psychologically. The value of winning or losing \$200 is less than twice the value of winning or losing \$100, respectively. Such an S-shaped value function is consistent with the finding that positive outcome framing increases the tendency to choose the P-option (because the higher value of the O-option is worth less, due to the S-shaped function, than the higher probability of the P-option). Negative framing favors the choice of the O-option (because the higher negative outcome of the O-option is less extreme subjectively than the higher probability of the P-option).

3.4 Exchangeability and Compensability of Resources

An essential premise of exchange is that qualitatively different resources can be exchanged and aggregated in a compound value function. Thus, when calculating the lower limit of the price for a product, the costs for

development, production, advertising, and taxing have to be translated onto the same joint scale as the aesthetic appeal, the sonorous brand name, and the scarcity of the product. Monotonicity should not only hold within the singular components, but must also be conserved after exchanges involving different resources.

3.5 Pricing and Choice

One apparent implication of the exchangeability of resources is that differences in attractiveness should correspond to differences in the price that can be obtained for products or goods. A preferred or chosen object should justify a higher price than a less preferred or not-chosen object. However, many experiments show a preference reversal such that whereas the P-option is consistently preferred to an O-option (of the same expected value) on a *choice task*, the O-option is preferred over the P-option on a *pricing task* (Slovic 1995). Although most people prefer the secure P-option when making a choice, they are willing to pay a higher price for the O-option—an often-replicated finding that cannot be derived from prospect theory. One interpretation for the disparity of pricing and choice, though not the only one, is that the pricing task makes the outcome dimension (typically indicated in monetary terms) more salient, whereas the choice task makes the probability of winning more salient.

3.6 Effort and Accuracy

Another implication of considering qualitatively different resources as exchangeable is that amount of work and effort can be paid in money and, likewise, that work load is commonly considered a positive monotonic predictor of quality or returns. Accordingly, investing additional work should increase, and must not decrease, the quality of a product or the accuracy of an organizational decision. Otherwise, the economical value of labor and effort expenditure would be negative, creating a severe exchange paradox. Nevertheless, at least under specified conditions, working or thinking too much can impair rather than improve rational decisions (Wilson and Schooler 1991).

3.7 Information Costs and Expertise

Extending the same principle to information costs, a related implication is that the value of a product, or the future gains to be expected from rational decisions, should increase monotonically with the amount of information gathered, or with the expertise of consultants and decision makers. Information costs, much like production costs, must be subtractable from the gross income on the same monetary scale.

Of course, there is ample evidence to support this plausible assumption as, for instance, evidence showing that the reliability of decisions (e.g., choices among job applicants, prediction of growth, market analyses) increases with the number of diagnostic indicators. However, there is also a notable body of opposite evidence pointing out the less-is-more principle. As Gigerenzer and Goldstein (1996) have shown in pertinent computer simulations, increasing the number of predictors and increasing the complexity and sophistication of algorithms may actually reduce the validity of predictions and decisions. This seemingly paradoxical result can be expected in highly uncertain environments when indicators are error-prone and fallible so that reliance on an increasing number of indicators often means increasing reliance on error variance. To understand the less-is-more effect, the logical principle of rationality has to be replaced by the notion of ecological rationality. Moreover, bounded rationality highlights the need to pit accuracy against information costs and effort load and to be content with ‘satisficing’ rather than optimal actions and decisions.

One particularly intriguing variant of the less-is-more effect pertains to personalized knowledge or expertise. The additional knowledge acquired by experts as compared with lay-people, and the increased costs of expert consultation, may not be justified by a corresponding improvement of products or decisions. For example, at the stock market, experts who are highly informed about shares and market rules may be outperformed by novices simply because experts try to utilize too many fallible indicators (Borges et al. 1999). At the same time, the degree of overconfidence is likely to increase with increasing expertise. Thus, as experience and qualification increases, subjective confidence may increase at a faster rate than accuracy and actual achievement (Yates et al. 1998).

3.8 Monotonicity of Indices

Economical decisions must often rely on indices, such as the daily value of a share on the stock market, as an index of the corresponding company’s (potential) performance. It is commonly assumed that indices correlate highly with the corresponding true properties on the same monotonic joint scale. For instance, it is likely that those companies whose share value index recently outperformed the market are expected to be more successful in the near future than those companies that lately failed to match the market. Indices are thus often regarded as monotonic functions of underlying true properties. However, this seemingly plausible assumption fails to take statistical regression into account, which is the source of many illusions and erroneous predictions. Regression is an inevitable result of error variance. To the extent that the reliability of an index is less than perfect, high index

		Other party		
		Defect	Cooperate	
Own party	Defect	-20 / -20	+30 / -10	Defect
	Cooperate	-30 / +30	-10 / -10	Cooperate

Figure 3
Pay-off matrix exemplifying a social dilemma situation

values are more likely to overestimate and low values are more likely to underestimate the true attribute to be assessed. Therefore, high values (of shares at the stock market) can be expected to be less successful after a while, whereas low values can be expected to be more successful. Ironically, then, investing in funds that recently outperformed the market may lead to less profit than investing in recently less successful funds (Hubert 1999). Thus, the assumption of a positive monotonic relation between an index and the true value of the underlying attribute is counteracted by an inverse relation between the error component of the index and the underlying attribute.

3.9 Absolute and Relative Gains and Losses

When the context, or state of the world, changes, monotonic orderings must be conserved, as already explained with reference to the independence principle. If decision option *a* yields a higher relative payoff than option *b* under condition *x*, and the relative payoff of *a* also dominates *b* under the complementary condition *y*, *a* is assumed to dominate *b* in general. Game theory is concerned with dramatic violations of this rule within social and economical dilemmas. The payoff structure of many negotiation and competition dilemmas is such that the payoff that one party gains from action *a* is higher than the payoff gained from action *b*, regardless of whether another party chooses strategy *x* or *y*. But although $a > b$ given either *x* or *y*, it is wiser and more profitable to choose action *b* rather than *a*. Such a payoff matrix is shown in Fig. 3, with reference to an embargo situation involving two parties each having two options: to cooperate (not delivering any goods) or to defect (violating the embargo to increase own profit). The situation can be exploited if one defects and the other party cooperates (yielding a high +30 payoff). However, if the other party also defects, violations of the embargo are sanctioned (−20). Nevertheless, defecting leads to higher payoff (i.e., lower embargo costs) than cooperating regardless of whether the other party defects (−20 vs. −30) or cooperates (+30 vs. −10). In spite

of the *relative* dominance of defecting over cooperating, given both opponent strategies, cooperation can have a strong *absolute* advantage in the long run, just because defecting is likely to be reciprocated by the other party so that the absolute costs will be clearly higher (−20 or −30) than when the other party cooperates (+30 or −10). Therefore mutual fairness and solidarity is in the long run a more profitable strategy (moderate costs of −10 for both parties) than reciprocal defection (stable pattern of −20 for both parties).

4. Conclusion

The above examples represent only a small subset of possible monotonicity breaks. They illustrate that at the crossroads of psychology and economics interesting research problems emerge wherever the interchangeability principle is violated. This offers a rather fruitful perspective for the role of psychology in economic research. Economists often refer to psychology when the real data do not meet the theoretical predictions. The connotation is that psychology represents irrationality or some source of error which is responsible for the incongruency of economic models and real market behavior. However, the above examples demonstrate that violations of economic assumptions are not necessarily irrational. There may be good reasons for focusing on different features when comparing different products; the regressiveness of indices is an unavoidable problem; and the very fact that buying more information may sometimes lower the quality of decisions reflects the intrinsic uncertainty of the probabilistic world. Rather than attributing violations of rational norms to human deficits, these violations can be as well interpreted as reflecting unrealistic simplifications of the normative models used to investigate economic behavior.

See also: Entrepreneurship, Psychology of; Money and Finances, Psychology of

Bibliography

- Becker G S 1976 *The Economic Approach to Human Behavior*. University of Chicago Press, Chicago
- Borges B Goldstein D G Ortmann A LGigerenzer G 1999 Can ignorance beat the stockmarket? In: Gigerenzer G, Todd P M, ABC Research Group (eds.) *Simple Heuristics that make us Smart*. Oxford University Press, Oxford, UK, pp. 59–72
- Gigerenzer G, Goldstein D G 1996 Reasoning the fast and frugal way: Models of bounded rationality. *Psychological Review* **103**: 650–69
- Huber J, Puto C 1983 Market boundaries and product choice: Illustrating attraction and substitution effects. *Journal of Consumer Research* **10**: 31–44
- Hubert M 1999 The misuse of past-performance data. In: Lifson L E, Geist R A (eds.) *The Psychology of Investing*. Wiley, New York, pp. 147–57

- Kahneman D, Tversky A 1984 Choices, values, and frames. *American Psychologist* **39**: 341–50
- Katona G 1951 *Psychological Analysis of Economic Behavior*. McGraw-Hill, New York
- Levinger G 1965 Marital cohesiveness and dissolution: An integrative review. *Journal of Marriage and the Family* **27**: 19–28
- Slovic P 1995 The construction of preference. *American Psychologist* **50**: 364–71
- Wilson T D, Schooler J W 1991 Thinking too much: Introspection can reduce the quality of preferences and decisions. *Journal of Personality and Social Psychology* **60**: 181–92
- Yates J F, Lee J-W, Shinotsuka H, Patalano A L, Sieck W R 1998 Cross-cultural variations in probability judgment accuracy: Beyond general knowledge overconfidence? *Organizational Behavior and Human Decision Processes* **74**: 89–117

K. Fiedler and M. Wänke

Psychology and Marketing

Marketing is an applied field. It is synonymous in everyday parlance with the economic activity of selling, and especially with advertising. The field, however, is construed differently by most professional practitioners. The American Marketing Association defines marketing as, 'the process of planning and executing the conception, pricing, promotion, and distribution of ideas, goods, and services to create exchanges that satisfy individual and organizational goals' (emphasis added). While marketing draws upon a wide range of social sciences, including anthropology, economics, and sociology, the emphasis on satisfying goals implies that insight into the psychology of the customer is particularly important for the marketer. Thus a company might offer a new microwave oven that is the product of work by electrical and industrial engineers, home economists, financial analysts, etc. A marketer would bring to this effort a focus on how consumers will think about the product and how it will fit with their behavior.

Marketing is hence an applied field in that interventions are designed with regard to the psychological responses of consumers. Marketing research is commonly used to learn more about these responses. In practice this can be as simple, and as particularistic, as surveys asking about product preferences or usage and qualitative research such as focus groups eliciting more in-depth reactions from consumers. Academic researchers typically conduct more complex studies. An example is the work of Simonson (1990) on the variety of brands consumers choose for consumption over multiple occasions. He finds that the set of brands chosen reflects greater variety when the choices for multiple consumption occasions are made at a single point in time (i.e., simultaneously) vs. being made just

prior to each consumption occasion (i.e., sequentially). This observation can be applied to marketing situations such as the design of retail stores. If a store, for instance, is intended as a convenience store to be used for items consumed immediately, it will not be necessary for it to carry as large a selection of products as with, say, a traditional grocery store that is used to build the household pantry inventory.

Psychology and marketing are thus intimately related. Marketing is based on an assessment of psychological response. The goal of marketing research is to enhance the applicability of these assessments to marketing situations of interest. There are two approaches to achieving this goal and each leads to a different view of exactly how psychology and marketing are connected.

1. Effects Application

One approach is that marketing research can focus on *effects application* (EA) (Calder et al. 1981, Calder and Tybout 1999). Research starts with psychological notions and proceeds by creating a research setting that allows the mapping of observed effects directly into situations of marketing interest. If the research setting is similar enough to the 'real' situation, the research is thought to generalize to these situations. From the perspective of this approach, it is important to conduct research with people who match the demographics and other characteristics of target consumers. Likewise the research setting, including independent variable treatments, should match the types of experiences people would have in the marketplace. Effects application studies are applied by extrapolating observed effects to a situation of interest.

The Simonson study described above can be viewed as an example of EA research. The research was no doubt motivated by psychological considerations, such as notions of consumers' desire for variety and their motivation to simplify complex tasks. The important thing is that although this research is motivated by psychological considerations it is the *effect*, greater variety with simultaneous versus sequential purchases, that is of primary interest. And the issue of application hinges on whether the effect can be directly generalized to marketing situations of interest. Because the original study was conducted with student subjects in a laboratory setting, the applicability to purchases in the marketplace might be viewed as suspect. For this reason, Simonson and Winer (1992) sought to enhance applicability by demonstrating the same effect with actual consumer purchase data from in-store scanner records. They looked at whether consumers bought a greater variety of yogurt flavors when they were buying larger quantities of yogurt. The argument from an EA perspective is that scanner data, in which actual

consumers spent their own money in real stores, makes the effect more directly applicable to marketing contexts than the laboratory study.

Other studies in the marketing literature use a related but somewhat different effects strategy. These studies seek to demonstrate the robustness of effects by demonstrating them across a variety of settings. Thus Strahilevitz and Myers (1998) draw upon notions of mental accounting to predict that the promise of a donation to charity will be a more effective incentive for consumer purchase when the product for which the incentive is offered is frivolous rather than practical. They find support for their prediction across three experiments that employ different stimuli. The robustness of the phenomenon, rather than the correspondence between the demonstrations and naturally occurring settings, provides the rationale for applying the effect to marketing contexts.

The EA approach views the relationship between psychology and marketing as follows: psychology provides important examples of psychological phenomena. It is the task of marketing research to investigate these effects in special research settings designed to be applicable to marketing situations of interest. Marketers thus draw from the psychological literature possible effects that must be observed in research settings that, by virtue of their similarity, are applicable to marketing situations. Or, alternatively, enough different settings are used to suggest robustness.

2. Theory Application

A second major approach leads to a very different view of the relationship between psychology and marketing, as well as what indeed it means for a discipline like marketing to be an applied field. This approach is called *theory application* (TA) (Calder et al. 1981, Calder and Tybout 1999). TA studies are designed to be applicable in a different way than are EA studies. The focus is on theory as an explanation of observed effects. It is not the effects that are expected to generalize; it is the explanation. Effects are used to test the theory. It is the theory that is expected to generalize to marketing situations.

This focus on theory testing leads in practice to marketing research studies that look quite different from EA studies. There is no requirement that research participants match consumers on demographic or other characteristics (unless this is relevant to the test of theory), nor is it important that things such as real money or natural marketing contexts be used in TA studies. Moreover, replicating the same effect over many situations is not regarded as particularly informative. This is because TA studies seek only to establish a strong explanation, not to generalize by

virtue of similarity of the research setting and marketing situation or on the basis of robustness. Instead of focusing on whether the greater variety with simultaneous choices effect can be obtained in a setting that sufficiently resembles marketing situations, a TA approach would concentrate on developing the theoretical explanation.

It might, for example, be hypothesized in connection with the variety-seeking research discussed above that the simultaneous choice setting is more cognitively demanding than the sequential choice setting and, thus, leads consumers to employ cognitive strategies that simplify the task. Simplification might be achieved by expanding the set of selected items, thereby obviating the need to make finer discriminations among the items. Here we have an explanation of the variety-seeking effect as opposed to the effect itself. Evidence for this explanation, and against rival explanations (i.e., that simultaneous choice makes potential satiation more salient and thereby encourages variety seeking), is the goal of TA research. Application to marketing situations is based on the rigor with which the theory has been tested and how well the theory describes the marketing situation.

Research by Meyers-Levy and Peracchio (1995) provides a good example of TA research. These authors examined the processes that underlie consumers' responses to full color, black and white, and color-highlighted advertisement pictures. Petty and Cacioppo's (1983) elaboration likelihood model (ELM) served as a starting point for Meyers-Levy and Peracchio's theorizing. ELM describes two routes to persuasion: a central route where people are motivated by relevance to process detailed arguments and a peripheral route where people are not motivated to process detailed arguments and instead respond to peripheral cues. Meyers-Levy and Peracchio reasoned that the correspondence between the cognitive resources made available and those required to process the advertisement claim might moderate processing via the central route.

Specifically, when processing motivation is high and the advertisement claim requires few resources to substantiate it, advertisements that include color will be more persuasive than black and white ones. This is expected because individuals will have sufficient cognitive resources to process the color and the color will substantiate and reinforce the advertisement claim. However, when more resources are needed to assess the advertisement claim, either a black and white picture or one in which color is used to highlight only those aspects of the picture that are relevant to the claim, is expected to be more persuasive than a full color advertisement. This is because individuals will lack sufficient resources to substantiate the advertisement claim and process full color information.

Meyers-Levy and Peracchio's hypotheses received support in two laboratory experiments conducted with student subjects. As a result, their integration of ELM

with the notions of cognitive resource matching is preferred to an explanation based on ELM alone. Further, they demonstrate that cues are not *per se* either central or peripheral, but rather depend on the match between the resources required and available. It is Meyers-Levy and Peracchio's theoretical explanation of their effects, rather than the effects themselves, that is applied to marketing contexts beyond the particular one they examined.

As the above discussion makes clear, the research setting and subjects are not of primary interest in TA research, it is the explanation being tested. Meyers-Levy and Peracchio's studies would not generalize to marketing situations based on similarity between the research setting and subjects. However their theorizing might be applied to any number of marketing contexts. This application would use the theory to design an intervention for the situation of interest that according to the theory would work in a desired way. The explanation provided by the theory would predict an outcome but it would do so based on the design of the intervention suggested by the theory.

An example of this kind of theory application may be helpful. Suppose there is an interest in communicating with consumers who are in the market for a new car. Presumably such individuals are motivated to engage in detailed processing of message claims. Meyers-Levy and Peracchio's theorizing would suggest that full color advertisements be used to substantiate and reinforce claims related to styling (which may be apprehended with relatively few resources) but black and white advertisements be employed when presenting detailed information about the car's features and warranty. The design of the advertisements thus constitutes an application of the theory and leads to different expectations for the ads.

The important point about the above application is, again, that it does not depend on matching the settings of studies to the particular concern with advertising. It only matters that the theory has received support and that it describes the marketing situation of interest. Note also that there is always the possibility that there are other, unanticipated factors involved in the situation of interest that may not be taken into account by the theory. In our advertising example, for instance, the color resolution or the particular colors chosen might be unappealing, in which case the use of color might undermine rather than reinforce the style claim. Such possibilities have two implications. One is that any application via TA research has the potential to fail (just as with EA research where it is impossible to match on all differences between the research setting and situations of interest). The other is that it may be desirable to conduct *intervention testing* (Calder et al. 1981, Calder and Tybout 1999), in which more factors are allowed to vary freely than in the typical theory test-oriented TA study. This can increase confidence in the application, but because this will ordinarily be a weak test of the theory, it does not increase support of

the theory *per se*, and the application still depends most critically on the prior TA research.

3. Conclusion

To return to the relationship between psychology and marketing, as we have seen, one account of this relationship is provided by the EA approach. Psychological phenomena motivate EA studies that are applied by virtue of the similarity of the research setting to marketing situations of interest. The alternative TA approach, however, provides a very different perspective. On this view the relationship of psychology and marketing is much more reciprocal and synergistic, as demonstrated in the discussion of Meyers-Levy and Peracchio's (1995) work.

With the TA approach marketing research is primarily engaged in testing psychological theories. This work is no different from the theory-testing work that goes on in psychology. It is only that marketing researchers have selected certain theories for further development and testing; these theories are ones that are considered especially promising for application. This applicability is based on the relevance of the theory to marketing situations. The success of the application depends most critically on the rigor of the TA studies supporting the theory, both within psychology and marketing. As experience with such applications increases, assisted in some cases by intervention testing, there will be an additional source of feedback about theories. It is the cooperative aspect of testing theories, through both TA research and the process of theory application, which constitutes the relationship between psychology and marketing. At the beginning of the twenty-first century, marketing continues to explore both paths to application.

See also: Advertising: General; Advertising, Psychology of; Decision and Choice: Economic Psychology; Marketing Strategies; Services Marketing

Bibliography

- Calder B J, Phillips L W, Tybout A M 1981 Designing research for application. *Journal of Consumer Research* **8**: 197–207
- Calder B J, Tybout A M 1999 A vision of theory, research, and the future of business schools. *Journal of the Academy of Marketing Science* **27**: 359–66
- Meyers-Levy J, Peracchio L A 1995 Understanding the effects of color: How the correspondence between available and required resources affects attitudes. *Journal of Consumer Research* **22**: 121–38
- Petty R E, Cacioppo J T 1983 *Attitude Change: Central and Peripheral Routes to Persuasion*. Springer/Verlag, New York
- Simonson I 1990 The effect of purchase quantity and timing on variety seeking behavior. *Journal of Marketing Research* **27**: 150–62

- Simonson I, Winer R S 1992 The influence of purchase quantity and display format on consumer preference for variety. *Journal of Consumer Research* 19: 133–8
- Strahilevitz M, Myers J G 1998 Donations to charity as purchase incentives: How well they work may depend on what you are trying to sell. *Journal of Consumer Research* 24: 434–46

B. J. Calder and A. M. Tybout

Psychology: Historical and Cultural Perspectives

Psychology occupies a middle position among the sciences, between methodological orientations derived in part from the physical and biological sciences and a subject matter extending into the social and human sciences. The struggle to create a science of both subjectivity and behavior and the effort to develop professional practices utilizing that science's results provide interesting examples for the reach, and also the limits, of scientific ideals in modern life. This article presents a brief account of the social and cultural relations of psychological thought and practice from the eighteenth through the twentieth centuries.

1. Routes to Institutionalization, 1700–1914

From the earliest times philosophical arguments about the nature of humans and of human knowledge have included claims about mental states and psychological processes. Protestant scholastics first used the term 'psychology' in discussions of the relation of soul and body in the sixteenth century. Christian Wolff's systematic accounts of 'rational' and 'empirical' psychology published in 1732 and 1734, respectively, marked out a distinct subject of scholarship, with journals and practitioners, a subject matter and intended methods of study. But that subject generally remained a teaching and research field within philosophy and pedagogy. The introduction of so-called 'physiological' psychology in the middle third of the nineteenth century and the institutionalization of laboratory instruction on the model of the natural sciences soon thereafter had a major impact on the development of the discipline. However, experimental psychology never succeeded in dominating the field entirely. The institutionalization of psychology took different forms in different parts of Europe and the USA.

1.1 Britain and France

Britain was the home of the statistical research practices pioneered by Francis Galton. These practices

targeted not psychological processes assumed to be essentially similar in all individuals, but rather distributions of performances *among* individuals. In *Essay on Human Faculty* (1883) and other works, Galton attempted to show that both physical and mental capabilities are quantitatively distributed in the same way, and that both are therefore inherited to the same—large—extent. In a 1904 paper, Charles Spearman took the approach a step further by distinguishing between 'general intelligence,' or 'g,' a factor underlying all performances in a test series and presumed to be hereditary, and so-called 's' factors accounting for differential performance on specific tests, presumed to be teachable. By the second third of the twentieth century, this emphasis on individual differences expressed in statistical terms had become a predominant research mode in both academic and applied psychology in the English-speaking world. Even so, as late as the 1920s there were only six university chairs for psychology in England. Psychological practitioners of various kinds far outnumbered academics in the membership of the British Psychological Association at the time of its founding in 1901.

In France, philosophers Hippolyte Taine and Théodule Ribot shared a coherent vision of the field as a synthesis of medical and philosophical approaches. However, Ribot's university course in psychology, first offered at the Sorbonne in 1885, was located in the Faculty of Letters rather than the Faculty of Sciences, while the laboratory demonstrations took place in the Faculty of Medicine. Despite this fragmented situation, he encouraged younger scholars, such as the physician Pierre Janet and the biologist Alfred Binet, to adopt a natural scientific approach.

When he became director of the first psychological laboratory in France, located in the Sorbonne's Faculty of Sciences, in 1890, Binet attempted to establish an explicitly biological science of higher mental processes, which he called 'individual psychology.' In 1894 he founded France's first scientific psychological journal, *L'Année Psychologique*. A commission from the Ministry of Education led to his publication, with Theodore Simon, of the first intelligence tests in 1905. The tests' purpose was not to measure intelligence directly—Binet doubted that this was possible—but rather to establish practical criteria for separating 'subnormal' from normal children, in order to provide the former with special education. But schoolteachers opposed the use of the tests in France, and here, as in England, extensive academic institutionalization of psychology did not follow.

1.2 Germany and the USA

Germany is generally regarded as the homeland of scientific psychology. In response to Immanuel Kant's claim that mental events, lacking the attribute of space, could not be measured, Johann Heinrich Herbart presented a program for the measurement of

sensations in 1816. In 1860, physicist and philosopher Gustav Theodor Fechner realized the program, with modifications, in his psychophysics—the measurement of relations between external stimuli and just-noticeable differences in sensation. Working in part with Fechner's techniques and in part with methods derived from laboratory physiology, Wilhelm Wundt established the first continuously operating laboratory for 'physiological psychology,' as he called it, in Leipzig in 1879. Other approaches widely received at the time included the 'medical psychology' of Rudolph Hermann Lotze, Franz Brentano's descriptive *Psychology from an Empirical Standpoint* (1874) and the *Völkerpsychologie* launched in 1860 by Moritz Lazarus and Haim Steinthal, which took an ethnological, linguistic, and historical approach. Other experimenters, such as Georg Elias Müller and Carl Stumpf, worked parallel to Wundt, but pursued quite different research programs.

In Wundt's and other German laboratories, in contrast to the situation in the UK or France, experimenter and subject were generally equal in status and often changed roles. German experimenters employed mechanical apparatus to control stimulus presentation, but they supplemented their data charts with extended records of their subjects' self-observations. In this way they engaged in an instrument-aided version of the self-discovery traditional to members of the German educated middle classes. Thanks to journals such as the *Zeitschrift für Psychologie und Physiologie der Sinnesorgane*, founded in 1890, as well as the Society for Experimental Psychology, founded in 1904 with Müller as first chairman, the 'new psychology' was on firm ground in Germany by 1905. But there was no agreement on the subject matter and methods of the discipline.

One reason for this was what Kurt Danziger (1990) has called the 'positivist repudiation' of Wundt by experimenters, including Müller, Hermann Ebbinghaus and Oswald Külpe. Though Wundt denied that experimental methods were sufficient to study the higher mental processes, these scientists extended apparatus-driven experimental techniques from sensation and perception to memory and thinking. Wundt opposed premature efforts to apply laboratory techniques in real-world situations, but William Stern, Karl Marbe, and others assessed the veracity of witnesses' testimony in court and tested the performances of schoolchildren at different times of day or the skills of industrial workers. In addition, an explicitly humanistic philosophical tradition persisted, with competing conceptions of the subject matter, method, and practical uses of psychology.

At the turn of the century, neo-Kantians and other philosophers attacked the intrusion of 'psychologism' into their field; in 1912, over 110 German teachers of philosophy signed a public statement opposing the appointment of psychologists to professorships in philosophy. But this protest failed, because state

officials were not convinced that the discipline had any obvious link to professional or civil service training. Until the Nazi era, experimenting psychologists in Germany maintained their own laboratories, journals, and association, but generally continued to compete for chairs in philosophy.

Wundt's US students transferred the new experimental psychology from Germany to the USA in the 1880s and 1890s, but the positivistic concepts they employed to justify using such tools were quite different from Wundt's. The sheer size of the country as well as the decentralized structure of the emerging US university supported rapid institutionalization. By 1910 there were more psychological laboratories in the USA than universities in Germany. The founding of the American Psychological Association in 1892 predated that of corresponding European societies. Despite this rapid growth, the new discipline exhibited some continuity with the past. Instruction in psychology had long been part of the required philosophy courses taught by college presidents such as James McCosh at Princeton. These courses and their teachers encouraged an orientation toward moral issues and concentration on useful knowledge rather than the emphasis on empirical foundations for philosophy of mind prevalent in Germany.

Also formative of the discipline in the USA was the work of Darwin and Spencer. Evolutionary thinking reinforced the emphasis on biological functions vs. mental faculties, and also encouraged comparison of adult humans with children and animals. Such views supported evolutionary theories of cognition like those of James Mark Baldwin, and also granted psychologists so inclined the authority to present themselves as agents of human betterment. Education and child study thus came to be of central concern to US psychology; here John Dewey, G. Stanley Hall, and Edward Thorndike were the opinion leaders, though they advanced different research and reform programs.

William James combined science and reform in his own way. Himself an evolutionist in certain respects, he created, with Dewey and James Rowland Angell, a distinctly American functional psychology. Though committed to psychology as a natural science, he criticized the tendency to substitute psychologists' conceptions of reality for their subjects' reported experiences in his classic text, *The Principles of Psychology* (1890). He favored a more expansive conception of the subject matter of psychology, but his later efforts to study the experiences of psychics and mystics with the same objectivity as those of 'normal' adults was not widely accepted. Instead, an emphasis on social usefulness that presupposed an engineering model of science became central to the establishment of psychology in the USA.

Disagreements on the scope and methods of psychology in the USA paralleled those in Germany. A group of experimentalists led by Edward Bradford

Titchener, which began to meet separately from the APA in 1904, did not oppose applied work *per se*, but insisted on rigorous method both within and outside the lab. In contrast, activists like Hall, who pioneered the use of questionnaires in the USA, were less concerned with laboratory-style rigor than with translating moral issues into scientific ones.

In this atmosphere, intelligence testing became more popular in the USA than in France. Henry H. Goddard, director of a training school for so-called 'feeble-minded' children, propagated the tests as instruments of human betterment. Lewis M. Terman revised the Binet–Simon scale for use in US schools in 1915, and thus linked 'mental age' to school class years. This move proved well suited to US schools as sorters of a socially and ethnically diverse population.

1.3 Common Features of the 'New' Psychology

Despite these multiple routes to institutionalization and differences in approach, certain common features of the 'new' psychology can be identified. One of these common features was a reliance on measuring instruments to establish objectivity and acquire scientific standing. With their apparatus for the controlled presentation of stimuli and measuring reaction times, experimental psychologists reconstituted the object to which their efforts were addressed. What had been mental and moral capacities became psychical functions; and the sensing, perceiving, conscious mind became an instrument that functioned, or failed to function, in a measurably 'normal' way.

A second common feature of the 'new' psychology was the use of physiological analogies based in turn on mechanical physics and technology. The term 'inhibition,' for example, blended organic and machine metaphors and applied them to both human action and to society. Soon after scientists and engineers applied the idea of energy conservation to human labor in order to create a science of work intended to make the 'human motor' run more efficiently, Emil Kraepelin and others extended the effort to 'mental work'; Hugo Münsterberg gave the result the name 'psychotechnics.'

A third common feature of the 'new' psychology was a studied vagueness about the mind–body relationship. Many psychologists asserted some version of psychophysical parallelism or claimed a functional relationship of mind and brain, but few were very precise about the nature of that relationship. A fourth common feature was the contested use of the term 'experimental' itself. Until late in the century, the term *psychologie expérimentale* referred in both France and Germany to seances. The experimentalists actively opposed spiritualism and attempted to expose quack practitioners, but James, Pierre Janet, and others studied altered mental states in psychics and mystics.

This broader view was not widely accepted, due to a fifth common feature of the 'new' psychology—a

tendency to restrict its subject matter to topics that could be addressed by the natural scientific methods and apparatus then available, such as psychophysics, sensory psychology, attention span, and retention. One result was an uneasy tension between efforts to preserve the notion of a volitional, active mind and the actual stuff of experimental research—measurable reactions to external stimuli. Another result was the exclusion of social or 'crowd' psychology; measuring instrument experimental methodology was plainly not applicable to group behavior.

A sixth common feature of the new psychological science was its gendered dimension. The head–heart dichotomy and the worship of the (female) 'beautiful soul' persisted through the nineteenth century; but its role in the 'new' psychology was ambivalent. The generalized, 'normal' adult mind that the experimentalists usually claimed to be their subject matter was at least implicitly the common property of both sexes, but the vocabulary and practices of objective science carried unmistakably masculine symbolism.

2. The Era of Competing Schools, 1910–45

The struggle for intellectual dominance in early twentieth-century psychology has been depicted since the 1930s as a battle of competing 'schools.' This view has its uses, but it conveys the false impression that all schools competed on an equal basis everywhere. Behaviorism captured both expert and popular attention in the USA in the 1920s, but the new approach was hardly taken seriously in other countries until after 1945. The 'reflexology' of Russian physiologists Ivan Pavlov and V. M. Bekhterev did not become a dominant approach in psychology even in the USSR until the 1940s. Gestalt psychology and other initiatives from Germany were received with interest but also with skepticism in other countries. Psychoanalysis had established itself as an international movement by the 1920s, but acquired few academic adherents at the time.

In German-speaking Europe, the 'crisis of psychology' announced by Vienna professor Karl Bühler in a 1927 book reflected continuing disagreement about subject matter and methods. The most widely received view internationally was that of Gestalt psychology, developed by Max Wertheimer, Wolfgang Köhler, and Kurt Koffka. They claimed, among other things, that immediately perceived structures (*Gestalten*) and relationships rather than isolated sensations are the primary constituents of consciousness, and employed holistic vocabulary to ground a rigorously natural-scientific world view. Felix Krueger, head of the Leipzig school of 'holistic psychology' (*Ganzheitspsychologie*), emphasized the role of feeling in perception and espoused Neo-Romantic cultural conservatism. William Stern's personalism focused on the individual as a 'psychophysical whole' in a way consistent with political liberalism.

The situation of psychology as a profession in Germany was equally contested. The challenge of philosopher Eduard Spranger's 'humanistic' psychology, as well as alternative professional practices such as handwriting analysis advocated by Ludwig Klages and typological personality diagnostics such as that of Ernst Kretschmer, raised the pressure to develop modern research instruments congruent with German cultural tradition. In Austria, the work of the Vienna Psychological Institute formed a bridge between old and new, theory and practice, Europe and America. Karl Bühler's department of general psychology conducted basic cognition research in the tradition of Brentano. At the same time, in rooms located at the city's adoption center, the institute's department of child and youth psychology, led by Charlotte Bühler, created performance measures for assessing the behavioral development of infants modeled on those of Arnold Gesell at Yale. In addition, Rockefeller Foundation funding supported the sociographic and survey research of the institute's Research Center for Economic Psychology under Paul Lazarsfeld in the late 1920s and early 1930s. All this put the Vienna institute, along with those in Jena and Hamburg, in the forefront of the transition to practice-oriented basic research in German-speaking psychology.

In the USA, multiple versions of behaviorism competed for attention in the 1920s. As proclaimed by John B. Watson in 1913, radical behaviorism excluded consciousness altogether from psychological science in favor of 'prediction and control' of behavior; in his later writings Watson advocated Pavlovian conditioning as a form of social engineering. Far more significant within the discipline at the time were the social science and child development programs generously funded by the Laura Spellman Rockefeller Foundation. The workers in these programs were not doctrinaire behaviorists, but they believed that measuring children's growth and IQ test scores over time would produce scientific norms of human development, and hoped to utilize this knowledge to rationalize society. Critics of behaviorism called on Gestalt psychology for support, while Harvard professor Gordon Allport and other prominent psychologists advocated a person-centered conception of psychology.

The 1930s were dominated by competing versions of neobehaviorism. Edward Tolman tried to integrate purposive motivation and cognitive processes into behavior theory, going so far as to claim that white rats framed 'hypotheses' as to which maze route would yield an expected food reward. Clark Hull developed an elaborate model for learning theory based on what he took to be Newton's physics, and tried to expand the model from the habit hierarchies of classical conditioning to personality theory. Finally, B. F. Skinner developed operant conditioning in the 1930s, producing careful measurements of the relative likelihood of simple behaviors such as rats or pigeons

pressing a bar to obtain pellet, of food under rigorously controlled conditions and suspending all efforts to explain such behavior.

A prominent nonbehaviorist effort to bring systematic theorizing into psychology was that of Berlin émigré Kurt Lewin. Lewin advocated experimental study of ideal-typical behavioral situations, exemplified in his Iowa studies of 'democratic' and 'authoritarian' leadership in children's groups. Lewin and his US competitors shared an admiration for classical physics and a willingness to draw upon philosophy of science, especially operationism and logical positivism, to legitimate their positions. They differed in their basic conceptual foundations and also in the physics they chose to emulate. The competition was resolved only by the rapid fragmentation of the discipline in the 1950s.

In Britain and France psychology remained relatively weakly institutionalized in this period in comparison with Germany and the USA. Yet precisely this situation enabled a wide range of alternatives to US behaviorism to flourish. In the UK, Cyril Burt, originally a London school official and later professor at University College, London, adapted and expanded Charles Spearman's concept of general and specific intelligence in studies of educational performance, delinquency, and so-called 'backward' children. He then developed a mathematical basis for factorial approaches to intelligence and personality testing. Controversy over accusations that he manipulated or even invented some of his twin research and other data did not emerge until after his death. During the same period, Cambridge professor Frederick Bartlett published the pioneering study *Remembering* (1932), in which he established the role of learned schemata in retention and laid the foundations for considering memory as a process of active reconstruction rather than rote recall.

In France, psychology remained divided between medicine and philosophy; there was no separate degree until 1947. One result was that alongside the strictly experimental work of Henri Piéron, Binet's successor as director of the Psychological Laboratory at the Sorbonne, philosophers and sociologists considered psychological issues in broader and less positivistic ways. One example was the debate over Lucien Lévy-Bruhl's concept of 'primitive' mentality, which contributed to the emergence of the 'mentalities' concept of the *Annales* school in history. In French-speaking Switzerland, biologist and philosopher Jean Piaget, building in part upon Eduard Claparède's functional psychology, began his pioneering studies of cognitive development in children.

2.1 Dynamics of Professionalization to 1945

The turning point for the public visibility of professional psychology in the USA came with the mass

use of intelligence tests in the US Army during World War II. The route of application ran here not from the 'normal' to the 'pathological,' but rather from socially marginal populations—the so-called 'feeble-minded' and schoolchildren—to 'normal' adults. The interaction of applied psychology and the professional officer corps reshaped both the aims of intelligence testing, the test instrument itself, and ultimately conceptions of the objects being assessed. Intelligence became not intellectual or problem-solving capacity alone, but a sum of skills and (presumably hereditary) aptitudes for certain kinds of learning.

'Binet testing,' as it was then called, fueled the professionalization of psychology in both the USA and UK during the 1920s. Quantitative assessment and classification instruments spread rapidly in both basic research and professional practice, primarily because the products thus created supported the functions required by administrators, initially in schools and later also in industry and social service agencies. During this period the field became more open to women; but a gender hierarchy emerged, with industrial psychology remaining male-dominated, while female 'Binet testers' and social workers took on more people-oriented functions. During World War II, fields of application included the employment of social psychology in morale research and applied human relations, incorporation of psychophysics and experimental psychology into studies of human-machine interactions, and diagnostic testing in clinical psychology, as well as the use of intelligence and personality testing in personnel management. All this led in turn to significant basic research programs in the postwar period.

The professionalization of psychology in Germany took a rather different course. During World War I, efforts focused on the adaptation of techniques from psychophysics to develop sound-ranging devices, and to test the visual discrimination ability of drivers and pilots. More important after the Nazi takeover of power were the rapid growth of military psychology as a result of German rearmament, and the resulting shift from psychotechnical skills testing to 'intuitive' character diagnosis. The primary purpose here was elite officer selection, rather than sorting large numbers of average recruits. Paper-and-pencil and skills tests were secondary to the extended observation of officer candidates in simulated command situations. The personality characteristics sought had considerable affinities to the traditional virtues of the Prussian officer—the will to command and the ability to inspire troop loyalty. In contrast, diagnostic efforts based on Nazi 'race psychology' could not be translated into professional practice.

In the USA, too, personality diagnostics ultimately became a road to professionalization. However, in contrast to Germany, quantitative methods based on techniques of factor analysis developed by L. L. Thurstone and others predominated, despite com-

petition from 'projective' tests such as the Rorschach in the 1930s and 1940s. The history acquired a gendered dimension in the construction of 'female' and 'male' traits in early personality research.

3. *The Postwar Era: 'Americanization' and the Alternatives*

In the USA, the postwar years saw explosive expansion and differentiation in both the scientific and professional realms. The establishment of a divisional structure within the APA in 1947—already negotiated during the war—reflected this process. Despite the optimism of the time, it proved difficult to subsume all aspects of psychology's protean identity within single university departments or graduate programs. Fragmentation was most obvious in the different research practices institutionalized in experimental, social, and personality psychology.

Cognition returned to the laboratory in this period, supported in part by reliance on methodological conventions patterned on those of the behaviorists, such as rigorous separation of independent and dependent variables, in part by computer-centered models of mental processes. In contrast to the statistical inference techniques and the corresponding computational models of mind that came to be preferred in cognition research, the preferred research tools in educational psychology were the correlational methods pioneered by Galton. A comparable methodological split distinguished neobehavioristic learning theory from experimental social psychology and personality theory. Nonetheless, experimental studies of social influence on perception by Solomon Asch and of prejudice by Gordon Allport, as well as T. W. Adorno and colleagues' *The Authoritarian Personality* study (1950) captured the imagination of many in the field. The popularity of such studies reflected a widespread tendency of the period to psychologize, and thus individualize, social problems. Meanwhile, developmental psychology took the work of Piaget as a touchstone for numerous studies closely related to the practical needs of schools for age-related developmental norms.

By the 1970s, the sheer numbers of psychologists (over 70,000, over 100,000 by the end of the century) had reached levels that could not have been imagined 50 years earlier. The growth was worldwide, but more than two-thirds of the total were Americans. The openness of both discipline and profession to women continues; today more than half the doctorates in the field go to women. Gender divisions also continue, with women being most numerous in developmental and educational psychology and men in experimental, industrial, and personnel psychology. Nonetheless, the broad institutional anchorage of psychology in the USA was more than sufficient to assure that the

research and professional practices instituted there would spread throughout the world, and that alternative viewpoints coming from Asia, Africa, or Latin America would generally be marginalized.

The most important exceptions to the overall trend were the impact of Piaget in developmental psychology, and the reception of work by British psychologists Hans Eysenck and Raymond Catell in personality testing and diagnostics. In cognition research, the work of Bartlett and the achievements of Soviet researchers such as Alexander Luria were mobilized to lend respectability and theoretical sophistication to the resurgent field in the USA. Nonetheless, the all-pervasive influence of computer metaphors and the associated information-processing models were plainly of Anglo-American origin.

The history of the psychological profession after 1945 continued to be affected by contingent local circumstances. The rise of clinical psychology in the USA was originally driven by the need to deal with large numbers of mentally ill veterans after World War II. The new field ultimately brought forth its own basic research in both clinical and academic settings, which led to the emergence of scientific communities based on methodological norms quite different from those of experimental or developmental psychologists. In addition, an eclectic, so-called 'humanistic' psychology movement arose in opposition to both behaviorism and psychoanalysis, and became widely popular in psychotherapy, social work, and the emerging field of counseling psychology.

In Europe, the rise of clinical psychology came approximately 10 years later than in the USA. There, in contrast to the USA, the supremacy of personality diagnostics and its quantitative tools had already been established in basic research before the professionalization of the clinical field. Another important difference was that clinical training in academic settings in Europe was based far more on cognitive and behavioral techniques than on psychoanalysis. Barriers to the academic institutionalization of psychoanalytic research and training were surmounted only in exceptional cases.

What had been a predominantly European field at the beginning of the twentieth century has become deeply dependent on US research styles and professional practices. US predominance has been contested by dissident local-language movements, most notably in France and Germany. Most significant, however, is the contrast between US predominance worldwide and the insecure standing of trained psychologists in the USA itself. Vagueness and confusion in the use of the term 'psychologist' in public discussion have been remarkably consistent over time; the term itself lacks legal protection in any case. The fact that the popularity of self-help books does not depend on whether their authors are psychologists or not indicates that even in the USA, where most of the world's psychologists live and work, trained academics

and professionals can hardly claim hegemony over psychological discourse in the public sphere to the degree that physical scientists can in their fields.

4. Conclusion

Given this situation, it might well be asked why such a shakily legitimated field has acquired such an important role in the twentieth-century. Roger Smith (1998) suggests that the discipline drew its authority from and simultaneously gave voice to a culture and society in which everyone 'became her or his own psychologist' (p. 577). Nikolas Rose (1996) argues that psychological practices make possible particular kinds of social authority, assembled at first *ad hoc*, then grafted onto all activities aimed at simplifying the administration of modern life by producing calculable individuals and manageable social relations, from law and penal administration to education and parenting. No single science or profession has monopolized the codification and certification of these activities.

The predominance of behaviorism in the USA in the middle third of the twentieth century was an episode in a much larger story. However, it is a characteristic episode, for both the discourse of prediction and control and its associated practices have persisted, even as the so-called cognitive revolution reintroduced mentalistic vocabularies. One reason for this appears to be that not only the members of the discipline and profession called psychology, but the modern culture and society in which they function, require and may even desire both technocratic discourse and the instruments that embody and enact it.

See also: Allport, Gordon W (1897–1967); Authoritarian Personality: History of the Concept; Behaviorism; Behaviorism, History of; Binet, Alfred (1857–1911); Broadbent, Donald Eric (1926–93); Campbell, Donald Thomas (1916–96); Cattell, Raymond Bernard (1905–98); Cognitive Psychology: History; Cognitive Science: History; Disciplines, History of, in the Social Sciences; Erikson, Erik Homburger (1902–94); Freud, Sigmund (1856–1939); Galton, Sir Francis (1822–1911); Hall, Granville Stanley (1844–1924); Heider, Fritz (1896–1988); Helmholtz, Hermann Ludwig Ferdinand von (1821–94); James, William (1842–1910); Janet, Pierre (1859–1947); Jung, Carl Gustav (1875–1961); Kant, Immanuel (1724–1804); Kohlberg, Lawrence (1927–87); Khler, Wolfgang (1887–1967); Lewin, Kurt (1890–1947); Lorenz, Konrad (1903–89); Luria, Aleksander Romanovich (1902–77); Pavlov, Ivan Petrovich (1849–1936); Piaget, Jean (1896–1980); Psychology: Overview; Rogers, Carl Ransom (1902–87); Savage, Leonard J (1917–71); Skinner, Burrhus Frederick (1904–90); Stevens, Stanley Smith (1906–73); Thorndike, Edward Lee (1874–1949); Tversky, Amos

(1937–96); Vygotskij, Lev Semenovic (1896–1934); Watson, John Broadus (1878–1958); Wundt, Wilhelm Maximilian (1832–1920)

Bibliography

- Ash M G 1995 *Gestalt Psychology in German Culture 1890–1967: Holism and the Quest for Objectivity*. Cambridge University Press, Cambridge, UK
- Benetka G 1995 *Psychologie in Wien. Sozial- und Theoriegeschichte des Wiener Psychologischen Instituts 1922–1938*. Wiener Universitätsverlag, Vienna
- Benjamin L (ed.) 1992 The history of American psychology. *American Psychologist* **74**: 109–335
- Capsheuw J H 1999 *Psychologists on the March: Science, Practice and Professional Identity in America, 1929–1969*. Cambridge University Press, Cambridge, UK
- Carson J 1999 *Talents, Intelligence and the Construction of Human Difference in France and America, 1750–1920*. Princeton University Press, Princeton, NJ
- Danziger K 1990 *Constructing the Subject: Historical Origins of Psychological Research*. Cambridge University Press, Cambridge, UK
- Dehue T 1995 *Changing the Rules: Psychology in the Netherlands, 1900–1945*. Cambridge University Press, Cambridge, UK
- Gardner H 1985 *The Mind's New Science: A History of the Cognitive Revolution*. Basic Books, New York
- Geuter U 1992 *The Professionalization of Psychology in Nazi Germany* (trans. Holmes, R) Cambridge University Press, Cambridge, UK
- Gould S J 1981 *The Mismeasure of Man*. W. W. Norton, New York
- Herman E 1996 *The Romance of American Psychology*. University of California Press, Berkeley, CA
- Joravsky D 1989 *Russian Psychology: A Critical History*. Blackwell, Oxford, UK
- Kusch M 1995 *Psychologism*. Routledge, London
- Lück H, Grünwald H, Geuter U, Miller R, Rechten W 1987 *Sozialgeschichte der Psychologie*. Leske und Budrich, Opladen, Germany
- Morawski J G (ed.) 1988 *The Rise of Experimentation in American Psychology*. Yale University Press, New Haven, CT
- O'Donnell J M 1985 *The Origins of Behaviorism: American Psychology, 1870–1920*. New York University Press, New York
- Oppenheim J 1985 *The Other World: Spiritualism and Psychological Research in England, 1850–1914*. Cambridge University Press, Cambridge, UK
- Richards R J 1987 *Darwin and the Emergence of Evolutionary Concepts of Mind and Behavior*. University of Chicago Press, Chicago, IL
- Rose N 1985 *The Psychological Complex: Psychology, Politics and Society in England, 1839–1939*. Routledge, London
- Rose N 1996 *Inventing Our Selves: Psychology, Power and Personhood*. Cambridge University Press, Cambridge, UK
- Ross D 1972 *G. Stanley Hall: The Psychologist as Prophet*. University of Chicago Press, Chicago, IL
- Scarborough E, Furumoto L 1989 *Untold Lives: The First Generation of American Women Psychologists*. Columbia University Press, New York
- Smith L 1986 *Behaviorism and Logical Positivism: A Re-assessment of the Alliance*. Stanford University Press, Stanford, CA

- Smith R 1992 *Inhibition: History and Meaning in the Sciences of Mind and Brain*. University of California Press, Berkeley, CA
- Smith R 1998 *The Norton History of the Human Sciences*. W. W. Norton, New York
- Sokal M M (ed.) 1987 *Psychological Testing and American Society*. Rutgers University Press, New Brunswick, NJ
- Van Ginneken J 1992 *Crowds, Psychology and Politics 1871–1899*. Cambridge University Press, Cambridge, UK
- Vidal F 1994 *Piaget Before Piaget*. Harvard University Press, Cambridge, MA
- Zenderland L 1998 *Measuring Minds: Henry Herbert Goddard and the Origins of American Mental Testing*. Cambridge University Press, Cambridge, UK

M. G. Ash

Psychology of the Real Self: Psychoanalytic Perspectives

Theories of the self have led to more confusion than understanding in the development of psychoanalytic thinking. Freud's ambiguous use of the term *Ich*, referring to the self both as a whole person—subjective experiential—and to the ego (Strachey 1966) theoretical, more objective and mechanical—led to considerable confusion about the relationship between the self and the ego which persists to the beginning of the twenty-first century.

The matter was further confused (Bettelheim 1982) by the abstract mistranslations of Freud's use of the word soul. An additional factor contributing to the confusion was that Freud and the pioneer analysts mostly worked with neurosis and, therefore, with oedipal levels of conflict. This is a developmental level at which the self had already become autonomous and assumed its functions and capacities. As a result, it did not cause enough of the kind of clinical problems that might have drawn attention.

However, when analysts turned their attention to patients with psychoses and character disorders, they were confronted with disorders of the self. These patients, development was arrested not at the oedipal but at the preoedipal level and, therefore, their deep difficulties were predominantly with the functioning of the self. The lack of a fully developed conception of the self, which had not previously been an obstacle, now became one.

From this basic conflict there emerged two main-streams of thought with regard to the concept of the self, each of which emphasized one of the two senses in which Freud used the term—either the whole person or as a special part of the mind with its own unique capacities and functions. Those who emphasized *Ich* as ego tended to be working mostly with neurotics and oedipal conflicts and placed heavy emphasis on the structural theory (ego, id, superego).

Those who emphasized the other meaning of the term *Ich*—the self as the whole person—tended to be working with psychoses and character disorders, and preoedipal levels of conflict and difficulties with the self.

This latter group, who emphasized the self as a whole person, further fractionated into a number of subgroups based on varying definitions of the self, its nature, and its functions as follows:

(a) Jung (1953) saw the self as a primordial image or archetype expressing the individual's need for unity, wholeness, and its highest aspirations.

(b) Rank (1929) focused on the psychological aspects of the emergence of the physical self at birth and emphasized separation anxiety as springing from birth trauma.

(c) Adler (1940) saw the life-style (repetitive patterns derived from early childhood) as a defensive overcompensation for inferiority feelings about the self, therefore blocking the emergence of the real, creative self.

(d) Horney (1950) postulated an idealized self which produced a pseudo unity or wholeness that blocked the emergence of what she called the real self.

(e) Sullivan (Mullahy 1970) felt that the essence of the self was determined by its functions: (i) the fulfillment of needs; (ii) the maintenance of security.

A reconciliation of both senses of the use of the word *Ich* or self—whole person and ego—has more recently been undertaken by theoreticians in ego psychology and object relations theory. Hartmann (1958) distinguished between the self as a whole person, the ego as a system (part of the tripartite structure of id, ego, superego), and the self representation as the intrapsychic representation of the self. He also defined narcissism as the libidinal investment of the self representation. Then Jacobson (1964), in order to work with psychotics, carried the concept further by defining the self representation as the unconscious, preconscious, and conscious intrapsychic representations of the bodily and mental self in the system ego. Kernberg (1982) used the term character to refer to the self as a whole person and preferred to use the term self to refer to the sum total of intrapsychic self representations in intimate contact with the sum total of object representations in the ego. In contrast to Hartmann, he felt that the self and object representations are both libidinally and aggressively invested.

Kohut (1971), from his work with narcissistic personality disorders, viewed the self as consisting of fused self objects which were invested with narcissistic libido that had its own line of development separate from the development of object representations and object libido. In other words, there was no separation of self and object, but rather maturation of self objects from primitive to mature forms.

This focus on the self as an intrapsychic entity, when kept in concert with the other perspectives of de-

velopmental object relations theory, can lead to a broader, more inclusive and comprehensive concept of the real self and its disorders. This view reconciles both definitions of *Ich*—the self and the ego—bringing together the personal, subjective experiential with the objective, mechanistic theoretical.

1. The Real Self: Structure, Development, and Function

1.1 Clinical

A few definitions will facilitate discussion of the structure of the real self.

1.1.1 Self-image. The image an individual has of himself at a particular time and in a specific situation. It consists of his body image and the mental representation of his state at the time. This image may be conscious or unconscious, realistic or distorted (Moore and Fine 1968).

1.1.2 Self-representation. A more enduring schema than self-image, constructed by the ego out of the multitude of realistic and distorted self images which an individual has had at different times. It represents the person as he consciously and unconsciously perceives himself and may be dormant or active (Moore and Fine 1968).

1.1.3 Supraordinate self-organizations. Because subjective experiences may be organized by multiple self representations, the 'I' of one state of mind may not necessarily be the same as the 'I' of another state of mind. This term is used for the organization and patterning of the various subordinate self-images and representations (Horowitz and Zilben 1983). It connects them and provides a continuity between them and a sense of unity and wholeness.

2. The Real Self

In keeping with ego psychology and object relations theory, the term 'real self' is used in the intrapsychic sense of the sum of self and object representations with their related affects.

The term 'real' is synonymous with healthy or normal. It is used to emphasize that the representations of the real self have an important conscious reality component even though unconscious and fantasy elements occur. It also indicates that the real self has an important reality function: to provide a

vehicle for self-activation and the maintenance of self-esteem through the mastery of reality.

The term real self also helps to differentiate it from the false selves of the borderline and narcissistic patients that are based primarily on defensive fantasy, not on reality, and are directed toward defense, not toward self-expression in reality. When using the term real self, both the collective subordinate self-representations and their supraordinate organization are referred to.

The real self exists as a parallel partner of the ego and has its own development, its own capacities, and its own psychopathology. The self and the ego develop and function together in tandem, like two horses in the same harness. If the ego is arrested in development, so is the self. One aspect of the self could be viewed as the representational arm of the ego, although it is obviously more than that. Similarly, one aspect of the ego, since it deals with volition and will and with the activation and gratification of individuating wishes, could be viewed as the executive arm of the self. However, it is also obviously more than that as its primary function is maintaining intrapsychic equilibrium.

Erickson (1968) referred to the dual and inseparable nature of the self-ego as follows: Identity formation can be said to have a self aspect and an ego aspect. What could be called the self identity emerges from experiences from which temporarily confused (subordinate) selves are successfully reintegrated in an ensemble of roles which also secure social recognition. One can speak of ego identity when one discusses the ego's synthesizing power in the light of its central psychosocial function and a self identity when the integration of the individual's self role images are under discussion.

To differentiate between personal identity [*or in my terms the real self*—JFM] and ego identity: personal identity is a conscious feeling based on two simultaneous observations: the perception of the self sameness and continuity of one's existence in time and space, and the perception of the fact that others recognize one's sameness and continuity. Ego identity, however, is the quality of that existence in its subjective aspect; it is the awareness of the fact there is a self sameness and continuity to the ego synthesizing methods.

The real self, then, provides an internal repertoire that, although finite and fixed, is varied and flexible enough to blend the need for real self-expression with the external roles required by adaptation.

3. Development of the Real Self

The ingredients or building blocks of the real self consist of the biologic and genetic endowment, the child's experience of proprioceptive and sensory sensations from his or her own body, as well as pleasure in increasing mastery in coping with the environment

plus the interactions with a caring object as described by Mahler and McDevitt (1982).

I will assume the infant has two basic points of reference from which he builds up his self schema: one, his own inner feelings or states forming the primitive core of the self on the one hand and, two, his sense of the care given by the libidinal object on the other hand. Insofar as the infant's development of the sense of self takes place in the context of the dependency on the mother, the sense of self that results will bear the imprint of her caregiving.

Psychoanalytic child observation researchers have traced the emergence of the self in normal children in the first three years of life. The issues of this child observation research are quite complex and beyond the scope of this paper. The contributions of four researchers: Mahler (1975), Stern (1985), Bowlby (1969), and Schore (1994) are briefly summarized.

Mahler (1975) derived intrapsychic hypotheses from observations of the child's behavior. Her work with child psychotics led her to theorize that the child's self emerges from a symbiotic relationship with the mother. She's described this occurring in four stages: autistic, symbiotic, separation-individuation and on the way to object constancy. Each stage brought further separation of the self from the maternal object representation. She emphasized the vital importance of the mother's support for the child's emerging self.

Stern (1985) similarly based his theories of intrapsychic development from observations of the children's behavior. He found that there was no symbiotic stage that the child was prewired to see the mother perceptually as separate from birth and that the child was an active co-creator of its own development in relationship with the primary caretaker. He also described four stages: emerging self, core self, inter-subjective self, and verbal self.

Bowlby (1969), unlike Mahler and Stern, made no intrapsychic hypotheses but limited his findings to what can be observed from the child's behavior, thus providing a firmer scientific base that could be replicated. He theorized that at birth the child had no attachment to the mother, and that the primary task was for the child to find a secure attachment. He divided that attachment into two types—the secure and the insecure. This insecure was then subdivided into avoidant, resistant, disorganized, and disoriented.

Schore (1994), based on neurobiological rather than psychoanalytic research, reported that the development of the brain is experience-dependent on the interaction with the primary caretaker or the mother. The genes are responsible for the physical neurons but it is the interaction with the mother during the crucial early stages of development between 10 to 12 months and 14 to 16 months that is responsible for the wiring of those neurons. It is the wiring that occurs during these crucial stages that helps the self to emerge and to take on its functions in the regulation of affect and socio-emotional relationships. Thus, the mother is

found to be as important neurologically as psychologically.

Although the self becomes autonomous and able to perform its function at three years, further increments are added during the phallic-oedipal phase, latency, and adolescence. As the child passes through each later psychosocial stage, the resolution of it provides certain capacities for the self such as industry, initiative, and generativity as described by Erickson (1968). The whole process of development of the self culminates in the normal identity crisis of adolescence, where the adolescent must test, select, and integrate the self images derived from the crises of childhood in the light of the climate of youth and sexual maturity.

3.1 Capacities of the Self

Simultaneously and in parallel with the maturation of ego functions, and with self and object representations becoming whole, the now whole, separate, real self becomes autonomous and takes on its capacities. These capacities were clinically identified mostly by their impairment and repair in borderline and narcissistic patients. This list is probably far from complete but forms an adequate clinical working scheme:

(a) Spontaneity and aliveness of affect. The capacity to experience affect deeply with liveliness, joy, vigor, excitement, and spontaneity.

(b) Self-entitlement. From early experiences of mastery coupled with parental acknowledgment and support of the emerging self, the sense builds up that the self is entitled to appropriate experiences of mastery and pleasure, as well as to the environmental input necessary to achieve these objectives. This sense, of course, is sorely deficient in the borderline disorder, and pathologically inflated in the narcissistic disorder.

(c) Self-activation, assertion and support. The capacity to identify one's unique individuating wishes and to use autonomous initiative and assertion to express them in reality and to support and defend them when under attack.

(d) Acknowledgment of self-activation and maintenance of self-esteem. To identify and acknowledge to oneself that one's self (in both senses of the term) has coped with an affective state and/or an environmental issue or interaction in a positive, adaptive manner. This acknowledgment is the vehicle for autonomously fueling adequate self-esteem.

(e) Soothing of painful affects. The capacity to autonomously devise means to limit, minimize, and soothe painful affects.

(f) Continuity of self. The recognition and acknowledgment through an effective supraordinate organization that the 'I' of one experience is continuous over time and related to the 'I' of another experience.

(g) Commitment. To commit the self to an objective or a relationship and to persevere, despite obstacles, to attain the goal.

(h) Creativity. To use the self to change old, familiar patterns into new, unique, and different patterns.

(i) Intimacy. The capacity to express the self fully in a close relationship with minimal anxiety about abandonment or engulfment.

(j) Autonomy. The capacity to autonomously regulate self-esteem and socio-emotional relationships.

4. The Healthy and Impaired Real Self

For a clinical example of the operation of the real self, consider a young psychologist who submits his first version of his PhD thesis to his advisory committee and is told that in some areas it lacks information and in others the writing is poor.

Since the cornerstone of his real self is his sense of self-entitlement, he feels entitled to mastery and whatever input that this requires, plus his capacities for self-activation, assertion, and commitment. He realizes at the outset that the purpose of submitting his thesis at that time was to learn its deficiencies so that he could correct them. He then applies his creativity, commitment, and self-assertion to find ways to learn the necessary knowledge and improve his writing. By mastering his reality problem, he has provided an autonomous feedback to support his self-esteem or self-image, a support which is also soundly based on reality.

To extend the metaphor of the psychology student to contrast the operation of the borderline, false defensive self with the real self: Because of his difficulty with self-activation, assertion and commitment, and creativity, the student might procrastinate and delay submitting his thesis way beyond the deadline. The critical comments would be felt as a reinforcement or proof of his inadequate self and would precipitate an abandonment depression to which he would respond by helpless clinging to a teacher or other authority, or by avoiding the project altogether. This, of course, would further reinforce his sense of an inadequate self.

See also: Culture and the Self (Implications for Psychological Theory); Cultural Concerns; Ego Psychology and Psychoanalysis; Person and Self: Philosophical Aspects; Personality and Conceptions of the Self; Psychoanalysis: Overview; Self: History of the Concept; Self: Philosophical Aspects

Bibliography

Adler A 1940 *The Theory and Practice of Individual Psychology*.
Harcourt Brace, New York

- Bettelheim B 1982 *Freud and Man's Soul*. Alfred A. Knopf, New York
- Bowlby J 1969 *Attachment and Loss: Attachment*. Basic Books, New York, Vol. 1
- Erickson E 1968 *Identity, Youth, and Crisis*. W W Norton, New York, pp. 208–31
- Freud A 1937 *The Ego and the Mechanisms of Defense*. Hogarth Press, London
- Freud S 1964 New introductory lectures on psychoanalysis. Lecture XXXI: The dissection of the psychical personality. In: Stracheys (ed.) *Standard Edition*. Hogarth Press, London, Vol. 22, pp. 57–80
- Hartmann H 1958 *Ego Psychology and the Problem of Adaptation*. International Universities Press, New York
- Horney K 1950 *Neurosis and Human Growth*. W W Norton, New York
- Horowitz M, Zilben N 1983 Regressive alterations in the self concept. *American Journal of Psychiatry* **140**(3), March
- Jacobson E 1964 *The Self and the Object World*. International Universities Press, New York
- Jung C G 1953 Two essays on analytic psychology. In: *Collected Works* 7, Pantheon, New York
- Kernberg O 1982 Self, ego, affects, drives. *Journal of the American Psychoanalytic Association* **30**: 893–917
- Kohut H 1971 *The Analysis of the Self*. International Universities Press, New York
- Mahler M S 1975 *The Psychological Birth of the Human Infant*. Hutchinson and Co., London
- Mahler M, McDevitt J 1982 Thoughts on the emergence of the self with particular emphasis on the body self. *Journal of the American Psychoanalytic Association* **30**: 827–47
- Masterson J F 1983 *The Real Self: A Developmental, Self, and Object Relations Approach*. Brunner/Mazel, Inc., New York
- Moore B, Fine B (eds.) 1968 *Glossary of Psychoanalytic Terms and Concepts*. American Psychoanalytic Association, New York
- Mullahy P 1970 *Psychoanalysis and Interpersonal Psychiatry: The Concepts of H. S. Sullivan*. Science House, New York
- Rank O 1929 *The Trauma of Birth*. Harcourt Brace, New York
- Schore A N 1994 *Affect Regulation and the Origin of the Self*. Lawrence Erlbaum Associates, NJ
- Stern D 1985 *The Interpersonal World of the Infant*. Basic Books, New York
- Strachey J (ed.) 1966 In collaboration with Freud A. *The Complete Psychological Works of Sigmund Freud*. Hogarth Press, London

J. F. Masterson

Psychology: Overview

Psychology is both an academic discipline and a profession. The concept of psyche originates from the Greek language and is frequently translated as 'soul.' The term has been used in religious and philosophical treatises since the antiquities, before an academic discipline explicitly designated as psychology was established in the sixteenth century. Until the nineteenth century, psychology was a special field within philosophy. During the first half of the twentieth century, it was institutionalized in universities as an

independent science. In the second half of the century, an increasing number of positions for university graduates with a degree in psychology were provided in public service, business organizations, and private practice. Today, there are approximately 300,000 scientifically qualified psychologists worldwide, most of whom are employed outside academic institutions.

Although psychology often presents itself as a unitary field, it comprises a considerable diversity of domains, theories, and methods. Despite its independent status, psychology shares historical roots and current concerns with other disciplines and professions. Both the ramifications of psychology into diverse branches and its position among related disciplines and professions will be outlined in turn.

1. Conceptions and Fields ('Psychologies') within Psychology

The soul was conceived as the principle which distinguishes the living body from the dead and manifests itself in consciousness and movement, or mind and behavior. The soul and its mental and behavioral manifestations have been studied from various perspectives. The paradigmatic perspective emphasizes theoretical and methodological principles and affiliations to 'schools.' The universal perspective seeks for knowledge that is valid for the human species, or even all vertebrates endowed with intelligence and mobility. In contrast, the differential perspective explores the variations occurring within and between species. In the analytical perspective, researchers refine the existing knowledge by discovering ever more details, while adherents of the holistic perspective endeavor to comprehend higher-order structures. The theoretical perspective calls for the description, classification, and causal explanation of mental and behavioral phenomena. The pragmatic perspective views the soul as the source of happiness, sociability, and efficiency, and raises questions of utility and morale.

Specialized research programs and practical applications have emerged from these perspectives. They have been institutionalized as curricular units, classifications for books and journals, and as divisions within and between scientific and professional associations. Some programs have competed for dominance and presented themselves as complementary 'psychologies' within psychology.

1.1 Paradigmatic: The Three Major Schools

Scientific paradigms are defined by their explanatory principles, the concomitant concepts and terms, and their preferred methods and values. They are established and maintained by a community of scientists that endorse one and only one paradigm. Modern psychology has spawned three major theoretical paradigms: cognitivism, behaviorism, and depth

psychology, all of which continue longstanding philosophical traditions.

In the tradition of rational philosophy, cognitivism regards human beings as reasonable, conscious, reflexive, and self-organizing. In this view, each person is endowed with insight into natural, logical, and social laws, and may utilize their free will and foresight to plan and execute actions directed towards valued goals. Moreover, humans are seen as sociable and communicative. Language is a tool for representing imaginations and intentions both in a person's own mind and in his or her interaction with others. Authors with a cognitive orientation (e.g., Broadbent, Köhler, Piaget) rely on introspection and self-reports (e.g., think-aloud protocols in studies of problem solving), but also collect behavioral data (e.g., operations in the course of problem solving). Typically, cognitivistic studies employ methods from mathematics, physics (e.g., field theory), or engineering (e.g., information theory) to elucidate the intricacies of mind and behavior.

In the tradition of materialism and utilitarianism, behaviorism (in Russia: reflexology) focuses on overt behavior (e.g., locomotion, instrumental acts) elicited by external stimuli (e.g., color signals) and internal stimuli (e.g., thirst), or emitted to take advantage of opportunities in the natural and social environment (e.g., food, praise). According to this conception, a human being is an automaton capable of forming stimulus-response associations (habits) and of adapting to contingencies between acts and their rewards or punishments (reinforcements). Behaviorism assumes innate, species-specific habits, or instincts (e.g., mating). Authors such as Hull, Pavlov, and Skinner have studied the acquisition of new habits and the shaping of behavior by reinforcements. Orthodox behaviorism opposes introspection, and relies exclusively on objective behavioral data.

In the tradition of mysticism, depth psychology posits a pervading influence of unconscious ideas and affects on conscious thought, language, and action. Freud's psychoanalysis, the most popular paradigm within depth psychology, contends that the unconscious originates from personal conflicts in early childhood when anxiety-provoking tendencies (primarily hostility towards the parent of opposite sex) are repressed and rejected from the conscious mind. In his analytical psychology, Jung even postulates an innate, species-specific unconscious consisting of archetypes. These archetypes are explained as fundamental experiences (e.g., motherhood, fire) which have been acquired and transmitted over generations. Depth psychology believes in the power of the enlightened human mind to reveal unconscious contents. However, the revelation of unconscious contents requires skillful interpretations. Beyond straightforward observations, interpretations should offer explanations for the meaning of symbols from the unconscious (e.g., the moon as a symbol of femininity) which are

encountered in dreams and other fantasies, as well as in achievements (e.g., in the arts) and failures (e.g., slips of the tongue).

The three major theoretical paradigms gained wide recognition during the first half of the twentieth century; they were also engaged in mutual antagonisms. However, no final verdict has been reached in the battles for dominance. There were some partially successful attempts to reconcile the conflicting paradigms. At present, conflicts between 'schools' have attenuated and orthodoxy has been superseded by eclecticism. For instance, numerous behaviorists now regard cognitions (e.g., attitudes and beliefs) as determinants of behavior, despite their subjective nature, while cognitive theorists accept the notion of unconscious ideas (e.g., subliminal perceptions).

1.2 Universal, Analytic: General Psychology, Biological Psychology

General psychology investigates universal and basic features of mind and behavior. Originally, it was launched in the spirit of philosophical idealism that acknowledged as scientific only universal and basic statements probing into the essence of matters. Ever since, general psychology has continued the quest for invariances and fundamental notions. It has predominantly pursued four objectives:

(a) Taxonomies. Cognitive functions have been categorized as sensation and perception, imagination and memory, concept formation, reasoning, problem solving, and language. Similarly, behavioral functions were classified as motives, drives, intentions, emotions, motor patterns, and actions.

(b) Structural analysis. Both cognitive and behavioral structures have been researched extensively. An example is semantic networks, that is, organizations of concepts representing domains of knowledge (e.g., automobiles), and action sequences (e.g., mountain climbing).

(c) Functional analysis; that is, the elaboration of the conditions which enable, prevent, and modify cognitive and behavioral processes and their outcomes (e.g., task complexity as a condition of learning).

(d) Formal (e.g., mathematical) models. General psychology designs models of internal mechanisms underlying psychic phenomena (e.g., the model of working memory). These models also serve to explain the effects of salient conditions.

In pursuit of these four objectives, general psychology is striving for progressive refinement. For structural and conditional analyses as well as for tests of formal models, experimental methods have become indispensable. Therefore, general psychology is sometimes equated with experimental psychology.

General psychology can be extended to encompass the exploration of the biological bases of cognitive and behavioral phenomena—the domain of biological

psychology. Using advanced measurement techniques, biological psychology locates and monitors mechanisms and processes of the brain, the peripheral nervous system, the endocrine glands, and other physiological systems which constitute the biological substrate of psychological functions such as language, memory, perception, and emotion, and which mediate cognitive and motivational states such as wakefulness and anxiousness.

1.3 Differential, Analytic: Animal, Developmental, Social Psychology, Individual Differences

Since the eighteenth century, the monopoly of general psychology has been challenged by romanticism, which emphasized the principle of organic growth in the natural world and in cultural settings, and by evolution theory (Darwinism). Both movements viewed variation as a fundamental natural principle. Romanticism and evolution theory provided the intellectual climate for the emergence of new psychological approaches that made variations their object of study:

(a) Animal psychology, also designated as comparative psychology, compared the behavior and the intelligence of species, including humankind. Much of this work was based on the assumption of a mental evolution (i.e., a gradual increase in intelligence from lower animals to humans), and a behavioral evolution (primary instincts like love or hate, transmitted from prehistoric man to modern-day humans, and secondary instincts acquired individually).

(b) Child psychology commenced with the study of mental and behavioral changes in humans during the first years of life, and was later extended to developmental psychology covering the whole life-span. In the beginning, child psychology frequently had an evolutionary background, since child development was interpreted as a replication of the evolution of the human species. Subsequently, evolution theory lost its major impact, and the study of human development continued with a generic interest in the changes that occur during people's lives.

(c) Social psychology documented and reflected the variations in cognitions, habits, and rules that are attributable to culture and social groups within cultures (e.g., the family). It drew cross-cultural comparisons of social structures, above all language, art, customs, power, education, kinship. Favorite topics were the evolution of social structures (social Darwinism) as well as the impact of external (e.g., geographical) factors on social life. Another influential branch was mass psychology that attributed to crowds the loss of rationality. Crowds were described as emotional, expressive, irresponsible, and addicted to charismatic leaders. Finally, social psychology restricted itself to the analysis of three basic issues: social relations (e.g., centralized vs. decentralized groups),

social cognitions (e.g., attitudes, stereotypes), and social processes (e.g., communication, group problem solving).

(d) The psychology of individual differences emerged from theories of human faculties, both intellectual (e.g., reasoning) and moral (e.g., altruism). The study of individual differences benefited from sociometrics which assessed frequency distributions of crime, income, etc. in populations. Measures of intelligence, motivation, and performance were found to be distributed as the Gaussian bell-shaped curve with the mean representing the most frequent case. From the genetic point of view, the positions of parents and their offspring in distributions were compared. Obvious correlations between parents and their children could be attributed to hereditary factors. However, sociocultural factors (e.g., educational style) were also examined as causes for transmissions of mental and behavioral dispositions over generations.

The studies on variations have modified the taxonomies and models from general psychology to account for inter- and intra-individual differences. New models were constructed which consider the ontogenetic, phylogenetic, social, and individual factors governing mental and behavioral variations. Thus, developmental and animal psychology have designed stage models of ontogenetic and phylogenetic changes, and social psychology has developed models of group formation and group cohesion.

Methodologically, differential approaches have conducted experiments, but have also collected a large body of data by field observations and surveys (e.g., longitudinal studies in developmental psychology).

1.4 Holistic: Personality Psychology

While human life manifests itself in a diversity of phenomena that vary with social context and over the life-span, the experience of a unitary mind and of consistency in behavior prevails in human consciousness. Upon this experience the concepts of self and ego are based. These two concepts replace the concept of soul in modern psychology. Self and ego represent an internal system which integrates diverse and even conflicting functions (e.g., inhibition of recollections), developmental stages (e.g., transitional stages from being protected in childhood to assuming responsibility in adulthood), and social roles (e.g., private vs. professional) to establish a sense of personal identity in consciousness and activity.

The unitary structure of mind and the consistency of behavior are the central issues of personality psychology. The concept of personality originally designated the coherent and integrated organization of persons, reflecting their life-long identity and individuality. Thus far, personality psychology has followed the philosophy of holism, that truth and valuable action can only be attributed to complete,

integrated structures. Presently, postmodern approaches question the assumptions of continuous integrality and self-identity. They prefer the assumption of a patchwork personality consisting of discontinuous and even contradictory features (e.g., role conflicts, different attitudes for different settings).

The method of understanding has been advocated as the most useful approach to personality. Understanding has been conceived as a hermeneutic technique that enables direct communication between an observer and his or her object. Adherents of the hermeneutic position have contributed in-depth character studies and biographies. Influential and popular hermeneutic approaches to personality development were provided by depth psychology (see Sect. 1.1). Numerous authors have acknowledged the artistic but have criticized the scientific relevance of hermeneutic contributions. As alternatives, they have proposed more rational models for personality theory. Factor theories of personality and theories of self-regulation have met with great success. Factor theories are based on measurements of individual differences. Employing mathematical techniques, they reduce the variety of measures to a minimum of personality dimensions. Extraversion–introversion and neuroticism are among the most frequently identified dimensions. Theories of self-regulation attempt to conceptualize the processes by which individuals maintain their identity and consistency, as well as those mechanisms through which persons induce or cope with changes in their personality. Central to these theories are self-concept, self-awareness, and self-management.

Due to its diversity of methods and theoretical principles, personality psychology is the most heterogeneous among the fields treated in this article. Its holistic claims, however, have merited its frequent appraisal as the most substantial field in psychology.

1.5 Pragmatic: Applied Psychology

Psychology has served three practical purposes: assessment, intervention, and evaluation. A large number of testing methods are available to assess the intelligence and the personality, as well as the specific aptitudes and attitudes of individuals. Based on these assessments, persons can be selected (e.g., for vocational careers), and classified for specific treatments (e.g., therapy plans). Besides persons, objects (e.g., sales products) and situations (e.g., work places) are also subjected to psychological assessments. Interventions include training, therapy, rehabilitation, and work organization programs. Evaluation refers primarily to public programs (e.g., educational projects, health campaigns). Evaluation methods are designed to monitor outcomes and costs of these programs, and to offer advice on how to improve their efficacy.

Some have claimed that a generic applied psychology can be adapted to all domains of civilization. Indeed, psychological services have been offered for a

plethora of specialized fields such as architectural psychology, sports psychology, and traffic psychology. Firmly established are the following:

(a) Clinical psychology for the diagnosis and therapy of mental and behavioral disorders in individuals, couples, and small private groups. Clinical psychologists typically treat neurotic and psychosomatic disorders (e.g., anxiety, compulsion, high blood pressure), as well as behavioral and social deviations (e.g., aggressiveness, substance abuse).

(b) Educational psychology is predominantly applied in educational institutions (preschool and school programs, special schools, vocational training, etc.), where psychologists advise educators and offer guidance to students. Educational psychologists also deal with educational processes resulting from peer interactions, media exposure, etc.

(c) Work and organizational psychology emphasizes personnel development, training, and work organization in various industries and on the administrative level.

(d) Law psychology is divided into several branches. Forensic psychologists assess witnesses and defendants in court cases; in criminal psychology, the objective is the prevention of delinquency and psychological support for inmates of penal institutions. Moreover, psychologists act as consultants to administrative courts (e.g., by rendering opinions in custody disputes).

(e) Economic psychology encompasses market research, consumer communication, and macroeconomics (i.e., the relationship of income, consumption, and savings).

The traditional view of applied psychology is based on the scientist–practitioner model which assumes that practical expertise is an extension of basic theoretical and methodological expertise. Alternatively, applied psychology may be conceptualized as representing an independent field of practical expertise that originates from the direct interaction with the problems encountered in the various branches of psychology. In this latter view, the scientific value of applied psychology rests on the documentation and analysis of practical problems, and on the validation of psychological procedures and theories. For instance, intelligence tests are not directly derived from basic theories of reasoning. They are carefully selected compilations of various tasks that have long been familiar in educational institutions and workplaces. Psychologists have standardized and validated these tasks to meet specific diagnostic requirements (e.g., school and job entry examinations).

1.6 Conclusions

The programs and paradigms of psychology form a fairly coherent and stable structure. Nevertheless, they exhibit inclinations towards isolating themselves. Typically, fields such as biological and social psy-

chology establish their own research programs, and each field holds its own conferences. Deficits in interdisciplinary exchange have also been claimed to exist between basic psychological research and applied psychology. On the other hand, there is a considerable overlap between various fields of specialization. For instance, research in social psychology may be duplicated in developmental psychology, as development occurs in social contexts and is interpreted to a large extent as a socialization process.

A pervasive feature of psychology is the fragmentation due to differences in perspective and historical background. Traditionally, boundaries between different 'psychologies' have been lamented and have given rise to efforts at unification and integration. Recently, however, disintegration and deregulation have also been interpreted as the results of diversity and subjectivism, and the fragmented state of psychology has been defended as a desirable adaptation to postmodern culture.

2. The Position of Psychology among Academic Disciplines and Professional Fields

Psychology has often been hailed as the heart of the life sciences. Psychological problems, methods, and theories relate to the humanities as well as to the social and the natural sciences. From the humanistic point of view, psychology is the basis of all scientific inquiry, since it studies the human mind as the source of knowledge. Furthermore, it provides the foundation for moral conduct as it elucidates the human condition.

Until the nineteenth century, scientific research of mind and behavior, and of individuals and cultures was conducted primarily under the label of philosophy. Moreover, psychological doctrines have been cultivated in law schools, medicine, and theology. In the late nineteenth century, psychology was institutionalized as a specialized academic discipline, along with other methodologically and theoretically advanced fields from the philosophical faculty such as linguistics, history, ethnology, sociology, and biology. A similar development took place in the practical domain. When psychology became a profession, some related branches of practice had already been firmly established (e.g., teachers in public schools). With an academic and professional system steadily growing and splitting into branches, psychology followed two courses. As an independent scientific discipline and profession, psychology develops specific theories, methods, and practical techniques to cultivate an exclusive expertise. In terms of its interdisciplinary position, various related disciplines and professions continue to contribute psychologically relevant findings and methods.

In the second part of this article, the position of psychology among academic disciplines and social

service professions will be examined. Both disciplinary and interdisciplinary projects will be regarded within four contexts.

2.1 The Metaphysical Context: Psychology, Theology, Occultism, Epistemology

Soul is a central concept in psychology and in theology. Theology defines the soul as an immaterial substance inherent in individuals. As far as the theological conception of the soul is holistic, it corresponds to the notion of a unitary self in personality theory (see Sect. 1.4). Unlike psychology, however, theology assumes a metaphysical world of external wisdom and ultimate values as the origin and destination of the soul. Moreover, theology posits interactions between the soul and metaphysical agents (e.g., gods, angels). This belief extends to miracles beyond rational explanation (e.g., resurrection from the dead), and to particular vicissitudes of the soul (e.g., creation, condemnation, and redemption).

While psychology as a modern doctrine refrains from metaphysical speculations, it does consider religious experiences and practices from a theoretical and a practical point of view. With regard to the former, a field designated as psychology of religion analyzes religious experiences and practices as social cognitions and customs. Regarding the latter, ministers are trained in pastoral psychology that is aimed at improving their ability to cope with the demands of their profession. Psychological programs, especially psychotherapy, may include religious notions and exercises to serve clients who are believers. Most churches agree with psychology in respect to the desirability of mundane peace and harmony. Therefore, clerical counseling services include or cooperate with psychological agencies. During and after natural and technical disasters (e.g., earthquakes, airplane accidents), members of the clergy and psychologists trained in emergency psychology join to provide immediate help to victims and their families.

While religious beliefs were discounted as superstitions in the era of enlightenment, some authors continued to believe in the existence of immaterial beings (e.g., spirits, demons) and rationally inexplicable phenomena (e.g., accurate prophecies, communication with spirits, movement of physical objects by spiritual forces). The movement of occultism suggested that these phenomena may be based on natural conditions that have thus far eluded all scientific scrutiny (e.g., 'animal magnetism'). Parapsychology has resumed the investigations of occult phenomena. Numerous studies have revealed deception as the origin of allegedly supernatural events. Nevertheless, there have remained some unusual demonstrations and reports that could not be explained scientifically.

In accord with epistemology, the field of metaphysics has not been fully abandoned in modern

psychology. Beyond religious beliefs, metaphysics has been interpreted as cognitive contents that transcend personal experience. In this sense, the metaphysical approach is maintained in nativistic theories of psychology such as the assumption of archetypes in analytical psychology (see Sect. 1.1) that allegedly are transmitted to the unconscious from preceding generations. Moreover, fundamentals of perception and reasoning (e.g., the dimensions of space and time, the concept of causality) have been postulated which serve as prerequisites for the acquisition of new experiences, but are not the result of individual experience.

2.2 The Cultural Context: Psychology, the Humanities, and the Social Sciences

Psychology has shared topics with all the humanities and social sciences. Linguistics is an example. The relationship between language, mind, and society has long been deemed worthy of scrutiny. Consequently, the interest in language as the object of research is common to a variety of disciplines (e.g., sociology, history). While their priorities differ, their efforts converge in providing a comprehensive, interdisciplinary understanding of language. Psychological principles are posited by many authors as providing the basis for the structure of languages (e.g., memory capacity, interactive minds). In psychology itself, language is also an important topic. Cognitive, developmental, social, and other fields of psychology investigate verbal representations and verbal behavior (e.g., words and concepts, language development in children). Within some psychology departments, the study of speech comprehension and production (both oral and written) has been institutionalized as a special field called psycholinguistics. However, the intricacies of specific languages (e.g., Dutch, Chinese) usually remain unexplored in psychological research. Expanding expertise pertaining to the grammar, vocabulary, morphology, and literature of a plethora of modern-day, as well as ancient languages has spawned language-specific research and training. This has been institutionalized in language departments that conduct extensive studies on specific languages. Moreover, many language departments pursue a program of linguistics that elaborates general theories of language.

A similar example from the social sciences is ethnology. Culture-specific lifestyles and beliefs have been salient topics in the discussion of human nature and human interactions on all levels of complexity (e.g., parent–child relationships, government of states). Much emphasis has been placed on comparing ‘primitive’ with ‘civilized’ lifestyles, in the hope that such studies would reveal the evolution of humankind. Frequently, cultural studies have been assigned to psychology, and psychological principles were invoked to explain diverse cultural phenomena. Jung (see Sect. 1.1), for instance, has found evidence for archetypes in folk mythology. Widespread interest in

foreign and primitive cultures has inspired research expeditions to distant countries. Such projects have provided invaluable information on human cultures (including their geography, language, technical skills, etc.). Moreover, they have led to the development of advanced research methods. A result of these efforts has been the establishment of fields such as ethnology and anthropology. Related disciplines, including psychology, have contributed to and benefited from ethnological research. Within psychology, there still exists the subdivision of cross-cultural psychology.

2.3 The Natural Context: Psychology, Biology, Physiology

Some psychological research has been initiated during the nineteenth century by biologists and physiologists in order to study brain functions which eluded the research methods of the time. Physiological psychology was promoted as a new scientific discipline that employed experiments to investigate higher mental functions (e.g., sensation, attention, learning). By the end of the twentieth century, brain research, neurophysiology, and biochemistry have come closer to the explanation of how cognition and behavior are represented in and controlled by the nervous system. Above all, brain research has determined the location of psychologically relevant neural functions (e.g., centers for speech production and comprehension), and has elucidated the organization of neural networks that interact during psychological processes (e.g., links between cortical correlates of perceptions and subcortical correlates of emotions in the appraisal process). Equally vital is the understanding of the hormonal control of arousal in the nervous system (e.g., due to adrenaline discharge), including the regulation of the sleep–wake cycle, alternations between activity and relaxation, and mood changes. In order to continue the research in this area, a new field called neuropsychology has been established.

The study of individual differences was instigated by efforts to ascertain the inheritance of character traits and skills (including deviant character and ingenuity). Studies on phenotypes have demonstrated the substantial influence of genetic factors on personality and behavior. Biological research programs are close to unraveling the mystery of the genetic code that is the substrate of inherited cognitive and behavioral dispositions.

Ethology deserves to be mentioned as a traditional interface between psychology and biology. It focuses on species-specific behavior with an emphasis on adaptation to the environment. Central issues are instincts as long-term, species-specific adaptations of drives and related behavior patterns (e.g., sex drive and mating behavior), and key stimuli eliciting instinctive behavior (e.g., the scheme of childlike characteristics which triggers affectionate behavior). Currently, ethological studies in biology outnumber

those in psychology, as the latter discipline has turned to questions of individual learning and socialization. However, psychology, in particular animal psychology and biological psychology (see Sect. 1.2) has continually incorporated ethological findings into its theories. The existence of neural and hormonal correlates of instinctive behavior has long been postulated, and neurophysiological research is in the process of providing the evidence for these assumptions.

2.4 The Context of Healthcare: Psychology and Medicine

Health is a physical and a mental state. Physical ailments have a psychological impact, and psychological disorders may cause physical diseases. The role of psychosocial factors has long been recognized in medicine, and special categories of neurotic and psychosomatic diseases have been defined. Neuroscience has demonstrated the effects of psychological stress on the immune system, and the new field psychoimmunology has been established.

The treatment of neurotic and psychosomatic disorders has become the domain of certified therapists trained in clinical psychology (see Sect. 1.5) rather than medicine. These therapists implement psychological diagnostic procedures and therapies such as cognitive and behavior therapy. A rapidly growing field in the health services is health psychology that designs, conducts, and validates programs of disease prevention (e.g., dental care, breaking nicotine addiction) and rehabilitation (e.g., after myocardial infarctions).

While all these psychological fields center on patients, another approach focuses on physicians: medical psychology. Obviously, diagnostic and therapeutic procedures may engender unconventional and emotional interactions between physicians and patients, and the outcome of treatments may depend on the social skills of physicians. Therefore, medical psychology studies and attempts to enhance doctor–patient relationships. For instance, doctors are advised on how to cope with stressful experiences (e.g., death of patients), and with problems of intimacy (e.g., during urological examinations).

2.5 Conclusions

There exists a tripartition of psychologically relevant research and practice: (a) generic involvement in topics which various fields of psychology (e.g., developmental, social psychology) share with other disciplines (e.g., sociology, pedagogy, biology), (b) specializations in object-centered disciplines (e.g., linguistics, ethnology, neurophysiology), and (c) specializations within psychology (e.g., psycholinguistics, cross-cultural psychology, neuropsychology). A complex network of

academic disciplines and social service professions has flourished in recent years, and within this network psychology occupies a respectable position with links to a variety of related fields.

However, the existing interdisciplinary network has also been criticized on the grounds of insufficient cooperation and of causing paralyzing conflicts. In particular, overlapping research in psychology and other independent disciplines has been indicted as a waste of resources. Why must there be linguistics and psycholinguistics, ethnology and cross-cultural psychology, neurophysiology and neuropsychology, organizational psychology and sociology, pedagogy and educational psychology, etc.? There are voices that advocate a re-integration of the presently diverging branches of the humanities, the social and the natural sciences. Some view psychology as the designated leader in this movement. Presently, however, the vision of a department of psychology pervading the entire university is highly unrealistic.

In regard to interdisciplinary integration, the attitudes diverge also within psychology itself. Some psychologists pride themselves in the academic and professional identity which psychology has achieved by its own methodological and theoretical approaches, and therefore stand opposed to a comprehensive integration, while others emphasize the benefits to be gained by merging resources with related disciplines. Two new interdisciplinary fields are exemplary for the movement towards integration: cognitive science and neuroscience. The highly sophisticated methodology of these disciplines (e.g., computer technology, high-resolution measurements of brain processes) may provide answers to fundamental psychological problems and may facilitate the development of revolutionary practical applications (e.g., decoding the memory traces in the brain could advance more effective learning procedures). However, major objections against a full integration of psychology into the new fields of cognitive science and neuroscience remain: mind may be more than a complex web of biological and technical structures and processes.

See also: Behaviorism; Broadbent, Donald Eric (1926–93); Cognitive Psychology: Overview; Developmental Psychology; Freud, Sigmund (1856–1939); Jung, Carl Gustav (1875–1961); Köhler, Wolfgang (1887–1967); Motivation and Actions, Psychology of; Pavlov, Ivan Petrovich (1849–1936); Personality Psychology; Piaget, Jean (1896–1980); Skinner, Burrhus Frederick (1904–90); Social Psychology; Watson, John Broadus (1878–1958)

Bibliography

- Hilgard E R 1987 *Psychology in America*. Harcourt Brace Jovanovich, New York
 Koch S (ed.) 1959/1963 *Psychology: A Study of Science*. McGraw-Hill, New York, Vols. 1–6

Schönpflug W 2000 *Geschichte und Systematik der Psychologie (History and Structure of Psychology)*. Beltz/Psychologie Verlags Union, Weinheim, Germany

W. Schönpflug

Psychometrics

Modern test theory makes free use of the idea of a latent variable to explain answer choices in each test item. A latent trait can be transformed by any smooth order-preserving function, and thus does not have the measurement structure that is often mistakenly attributed to it. The notion of a manifold is used to explain this one-dimensional structure. This manifold is not a latent variable, however, and arc length or distance along this manifold from a fixed point is a measure of ability or trait level that is manifest. This arc length measure is suggested as an alternative to latent trait values in order to avoid the confusion that the idea of a latent trait often causes.

1. Introduction

Terms like 'psychometrics' are constantly-moving targets for one seeking a definition, but some frequently encountered features of psychometric data are indicated in the next section. But many readers will think of test scores as prototypical data. What characteristics of the people taking tests do test scores and the data that are used to compute them reflect? Test scores are numbers, but not all the structure of numbers are meaningful representations of something about examinees. Section 3 considers how test scores define nearness or proximity of examinees, and the status of the algebraic operations of summing and differencing. Section 4 relates the psychological concept of a trait to the mathematical notion of a curve in high-dimensional space or a manifold, and considers arc length along this curve as a type of measurement.

2. Definitions and Common Themes

According to one popular dictionary, psychometrics means 'mental measurement.' But we need a better definition, since the status of both mind and measurement as concepts are no longer as unequivocal as they were at the beginning of the twentieth century. This article considers as psychometric most quantitative models for the behavior of humans and other animals that is controlled by the central nervous system.

Defined as broadly as this, it may seem surprising that there are any themes that seem to be common to psychometric research. Nonetheless, there are a few, and they may be motivated by the following questions:

What variables should be observed? In the natural sciences, where seven constants are said to define the universe, which variables to observe is seldom an issue. But in behavioral research, variables are made to measure by researchers, often in the form of tests or scales developed to reflect vaguely defined concepts such as 'depression' or 'ability.' For any popular concept, there are likely to be dozens and even hundreds of such measures available.

What does a variable value mean as a measure? That is, what information can be considered useful in a set of test scores, response times, or other observations, and what can be considered irrelevant? Stevens (1946) proposed a classification into nominal, ordinal, interval, and ratio scales according to the amount of algebraic structure usable in the data. Viewing test scores as nominal scales only treats numbers as labels for examinees. Ordinal scales represent their order structure or ranking. Interval scales also reflect a stable property of the examinees in the ratio of differences, and ratios of scores are meaningful for ratio scales. But so many writers have criticized this system as inadequate that this issue needs revisiting here, at least briefly.

How many variables should be observed? Judging by research reported in widely read journals, a great number. Because it is difficult to decide which measure of, say, intelligence to use, researchers tend to use several, hoping to cover the ground and perhaps to head off complaints from supporters of particular measures. Moreover, access to replicates, usually people, is often expensive and may involve ideological and political sensitivities. Consequently, psychometric data are often high dimensional, especially relative to sample sizes typically used. Researchers may hope that they can compensate for limited sample sizes by observing more variables; but this belief is both naive and dangerous.

Is there some simple low-dimensional structure in the high-dimensional data? Often a one-dimensional structure is assumed, as when choices on tens or even hundreds of test items are reduced to a single test score, but in other cases special psychometric technology is applied to data to reveal both the low dimensionality implicit in the data, and how the data are approximated within this low-dimensional structure. Principal components or factor analysis is, therefore, a standard tool in behavioral research.

3. What does a Psychometric Measurement Mean?

It is fair to say that most researchers don't concern themselves with this question, but it is a hard issue to avoid, since sooner or later numbers are interpreted, and usually by people whose limited exposure to these issues makes them vulnerable to misinterpretation of

the results. Is a change in score from 10 to 5 on a depression scale as important as one from 30 to 25? Probably not, but is there at least some useful information in score changes such as these?

A psychometric measurement system is a mapping from a set of people into some system of numbers. It has a purpose, namely to use some of the structural characteristics of numbers to reflect or represent analogous characteristics of people. For example, assigning a social security number just distinguishes one person from another, but assigning a test score tends to mean much more, and at least is intended to reflect an order relation presumed to hold among people. But numbers, and especially real numbers, are structured in many different ways, and the question is; which of these ways should be considered as reflective of characteristics of people, and which should be ignored?

Mathematics, at least as an applied science, is the study of structured sets. By sets, of course, we mean collections of people, drawn from some population X , including the set of nobody, which we indicate by $\emptyset$. Sets A and B can be combined in two ways: unions $A \cup B$, meaning the people who belong to either A , B , or both; and intersections $A \cap B$, meaning only the people who belong to both.

There are three main types of structure that pre-occupy mathematicians:

(a) topological: structure that defines nearness or neighborhoods,

(b) measure or probability: structure that defines size, and

(c) algebraic: rules for connecting two or more elements, that is, identifying certain pairs, triples, etc.

These structures are increasing in level of organization, with topological being in a sense minimal, and algebraic being the most sophisticated.

We shall have a look at topological structure because it seems important in explaining why test scores are so widely accepted as psychometric variables, and is an aspect of measurement that Stevens and other writes have neglected. But measure and algebraic structure are left for others to others to deal with. (See *Memory Models: Quantitative*.) Test scores are by and large accepted by both ordinary citizens and researchers as useful summaries of human behavior. Why this is so, in spite of their murky status among philosophers, mathematicians, and statisticians, probably derives from their topological structure and its relation to analogous properties of behavior that seem to us reasonable.

Topology has a bad reputation for being abstruse, but it deals with a type of structure corresponding to intuitions rooted in experience. Topology is about nearness or neighborhoods defined within a population X of people. One may also think of the topology of X as its *texture*, and in these remarks we will consider whether test scores have gained their accepted

place in most cultures because they reflect assumptions of a textural nature about how people vary.

A topology is a family T of subsets drawn from a set X such that:

(a) both nobody $\emptyset$ and everybody X belong to T ;

(b) the union of no matter how many sets in T is also in T ; and

(c) the intersection of a finite number of sets in T is also in T .

Sets in T are called *open*, and any open set containing a specified person is called a *neighborhood* for that person. For example, a person's neighborhood can be defined as those other persons sharing the same postal code, or having gone to the same university, or inhabiting the same dwelling, or of the same age, or any combination of these characteristics. Or, in this context, as those who have been assigned the same score on a test. The second part of the definition of a topology ensures that neighborhoods can be expanded as far as needed, even to include everybody, and the third says that neighborhoods can be shrunk, but only a finite number of times. What happens when we expand and shrink neighborhoods is the essential practical issue that the topological structure seeks to address.

Since the topology of test scores may reflect something of the topology of the people assigned the scores, let's switch to considering the topology of test scores, and ask whether this reasonably captures something about the population to which they are assigned. Because test scores are real numbers, an open set is defined as the set of test scores between two limits, and this topology is of a specialized variety called *metric*. Of course, test scores also define a topology of the examinees that have taken them, since we can define an open set of examinees as one containing all examinees whose test scores are within specified limits.

Figure 1 illustrates examinee neighborhoods for a score of 2 on a 50-item test, with the neighborhoods being defined by three different test-scoring strategies. The test is simulated, but this does not matter. Nor does it matter that we have chosen the relatively extreme score of 2; neighborhoods for more average scores will show the same features as shown in these three plots, but with less compelling graphics.

The points labeled E display what can be called an *even* topology on examinees. They indicate the number of examinees among 1000 whose scores differ from 2 by varying amounts. Each score separation value therefore defines a neighborhood. Everybody is, for example, no more than 48 points from 2, and the size of these neighborhoods shrinks smoothly as we reduce the score separation, until we see that only a few people are less than 5 points away from 2. We call this an *even* topology because neighborhoods shrink smoothly.

Suppose, now, we arbitrarily add 10 to all scores of 30 or above. This doesn't change the order of the

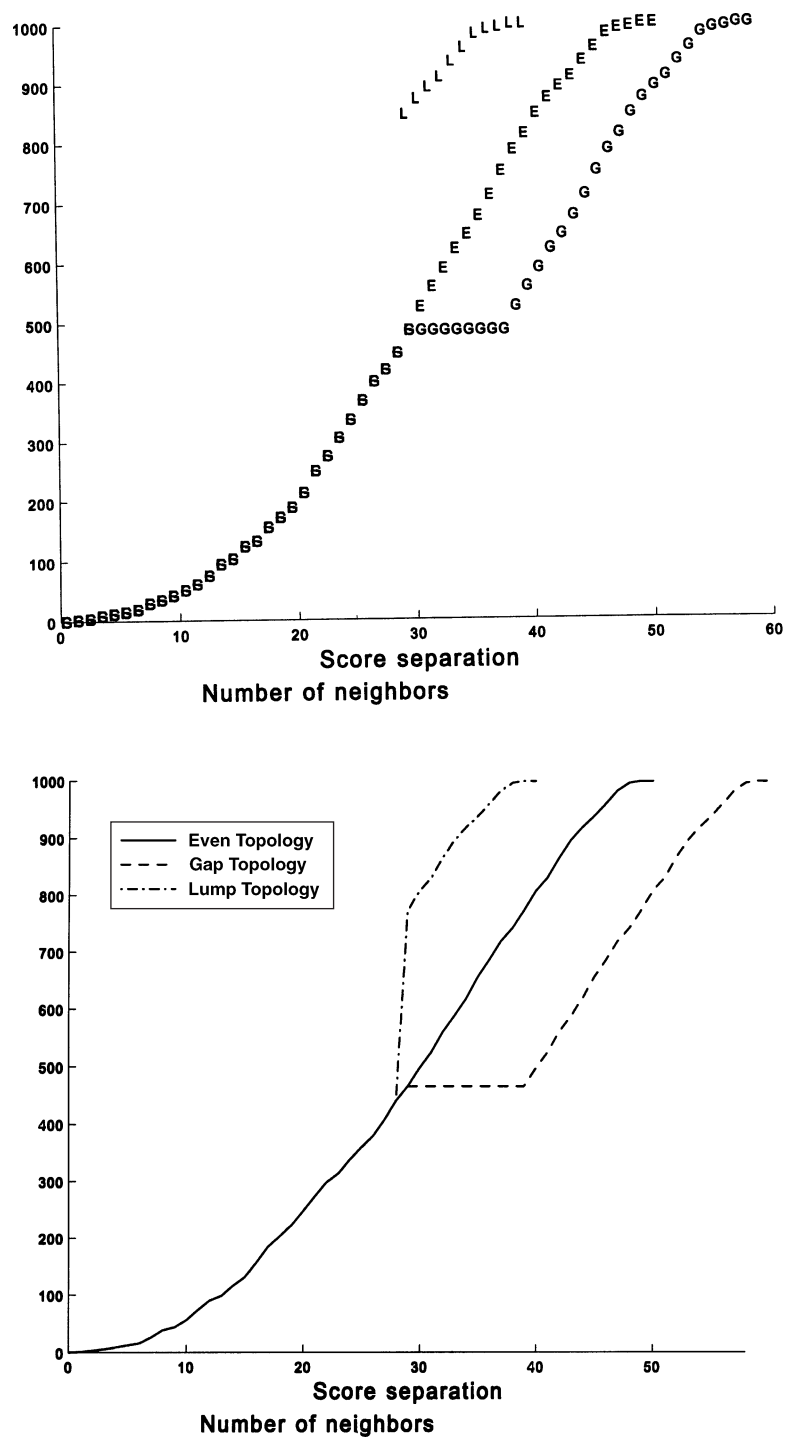


Figure 1

Number of examinees with scores separated by various amounts from a score of 2 according to three different topologies for a 50-item test given to 1000 examinees. Points labeled E have the even topology, corresponding to scores without gaps or lumps, those labeled G have the gap topology, and those labeled L have the lump topology

scores, but it certainly changes the neighborhoods. The points labeled G show that neighborhood size fails to shrink for score separations between 28 and 38 because of the gap in the scores. This is a *gap* topology. If the population being scored had such a gap, it would mean that there would be nobody capable of producing a set of choices on items corresponding to a test score between 30 and 40.

Or, alternatively, let us combine all scores between 30 and 40 into 30, and subtract 10 from all scores above 40. The neighborhood plot is shown as points labeled L, and we see that neighborhood size shrinks abruptly for score separations above 28 with respect to a score of 2. This is a *lump* topology because, put in people terms, there is a large group of examinees who will necessarily produce identical scores. Do people lump up in this way with respect to their behavior? Perhaps sometimes, but mostly not.

Both the gap and lump topologies that are shown in Fig. 1 are defined by test-scoring methods that preserve the order of test scores, but most of us would not be happy attributing either of these textures to subsets of people. We like to imagine that people are evenly distributed in the sense that (a) if we observe enough behavior, only a very small number of people will display identical behavior, and (b) nevertheless all behavior patterns are potentially exhibited by someone. This is the sense in which the even topology determined by most test scoring methods defines a texture for the population that has neither gaps nor lumps, and tends to correspond to intuitions that we have about how performance on a test is related to the ability that it is supposed to reflect.

A transformation that leaves topology unchanged is called *continuous*, a term that we may roughly translate into *smooth*. If there are no gaps or lumps, then, for example, substituting $y = \exp(x)$ or $z = \exp(x)/[1 + \exp(x)]$ for score x will not introduce lumps or gaps, whereas the two transformations indicated above will, and are clearly not smooth.

Stevens' ordinal scale is, then, too simple. It says too little about the texture of either the data or the population generating the data, and something like the concept of an even topology is needed to explain why those who work with test scores are clearly so comfortable with them as descriptions of the behavior of people.

4. Latent Traits and Manifolds

Few who have worked with colleagues in departments of psychology or faculties of education, or taught statistics to students in these places, will quarrel with the statement that finding low-dimensional structure in high-dimensional data is perhaps the central statistical challenge in these fields. Consequently,

statistical procedures such as stepwise regression, principal components analysis, and structural equation modeling are essential in most contemporary research, and new methods for this problem appear regularly in our methodology journals.

The latent trait is one way to motivate the concept of a low dimensional structure. Test and scale designers often appeal to constructs such as 'proficiency,' 'intelligence,' and 'creativity' in describing what scores on their instruments mean, and freely appeal to a conceptual link between usually unidimensional test scores and these constructs. Of course, it is only test scores, and the choices that generate them, that we can actually observe, and so these constructs are said to be latent or unobservable variables whose structure is reflected in these test scores. Similarly, early factor analysts liked to refer to collections of such latent traits that generate the observed tendency for rather large number of test scores to exhibit low dimensional correlational structure. The five variables in the NEO Personality Inventory, introversion, agreeableness, conscientiousness, neuroticism, and openness, (Costa and McCrae 1985) used in personality research are one of the more successful outcomes of this type of thinking.

Figure 2 displays three *item characteristic curves* for an actual test, these being defined as the relation between the probability $P_i(\theta)$ of getting test item i correct to one's position along a one-dimensional continuum θ . Continuum θ is not observed, but rather inferred from the data. These three curves were estimated by a nonparametric technique described in Ramsay (1991). Probabilities of success on these three items are displayed as functions of four different types of continuum: the first extending over the whole real line; the second containing only nonnegative numbers; the third within the unit interval $[0,1]$; and the fourth to be explained below. Now, since we are not actually observing θ , there is no deep reason for preferring one of these continua over the others, so that the transformations taking us from one to any other are all legitimate modifications of this model. What are not arbitrary, however, are the probability values $P_i(\theta)$ themselves. These are anchored in the data, and must conform to actual proportions of success for groups of people who have neighboring values of θ or, since smooth transformations preserve topology, neighboring values of either a transformation of θ , or of test scores themselves.

Consequently, the upper panel of Fig. 3 shows what is invariant with respect to transformations of θ . By plotting probabilities against each other, we see that probabilities of success migrate smoothly within the unit cube. The sinuous curve that this figure displays also has the even topology in the sense that it has no gaps or lumps. In this plot, the role of θ or any of its transformations has disappeared, except of course, for its topology, since the curve has inherited the even topology of real numbers.

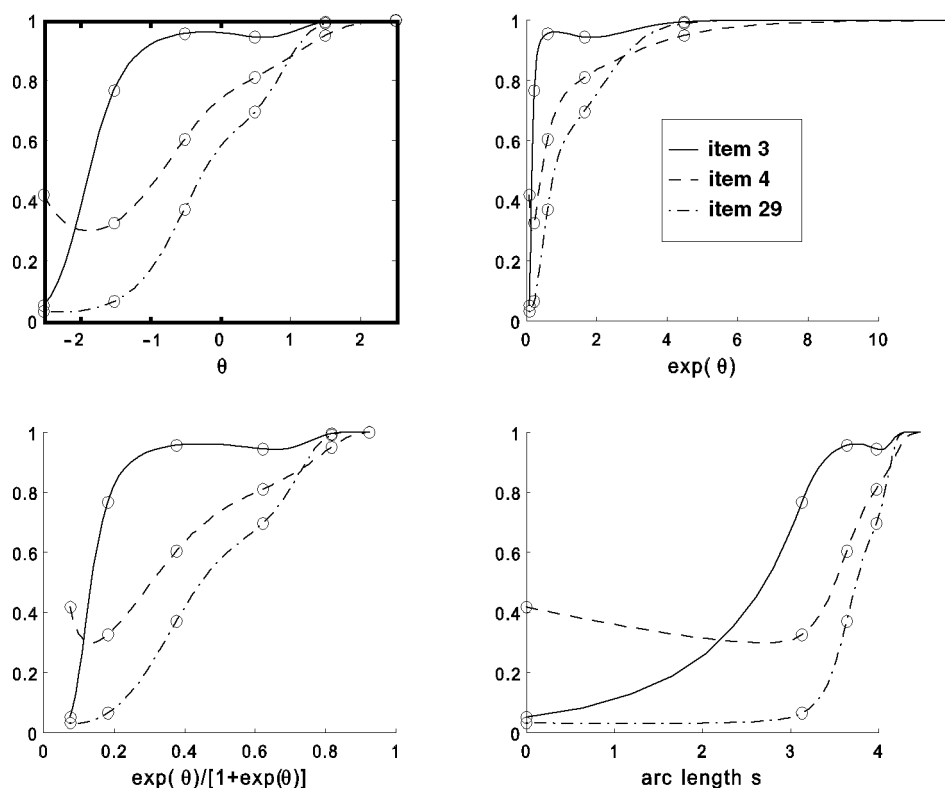


Figure 2

The probabilities of success on three test items as a function of positions on four types of traits. Traits in the upper right and lower left are transformations of that in the upper left, and the trait in the lower right, arc length, is observable rather than latent

The curve in Fig. 3 is a *manifold* in mathematical jargon. Indeed, a smooth manifold, and, if we can compute derivatives of item characteristic curves, a *differentiable manifold*. The subdiscipline dealing with such structures is *differential geometry*. The variable, or any smooth transformation of it, is called a *chart* for the manifold. And with good historical reason, too, since a familiar example of a two-dimensional manifold is the surface of the earth, and latitude and longitude are used to chart this structure, but this is only one example among a large number of alternative charting systems used by cartographers.

What is essential here is that the notion of a manifold captures both the even topological texture discussed in Section B, and something else that is much stronger. Any latent variable charting the manifold can be freely transformed by a smooth and order-preserving function. But there is one charting system that is intrinsic to the manifold, and therefore not at all latent. This is *arc length* s , the distance along the manifold from its beginning, or from any designated point within the manifold. It measures how far

someone at the lowest level of performance has to go within probability space to get to a specified performance criterion. Sums and differences of arc lengths can be calculated, and this makes s a ratio scale variable in Stevens' sense. Indeed, it is more, since s cannot be multiplied by an arbitrary positive constant, and it therefore is an *absolute* scale. Each examinee has an s -value, and, since values of s also possess a measure structure, we can approximate the distribution of these values. The probability density function for s is shown at the bottom of Figure 3. It may be argued that arc length s is the *natural* or *intrinsic* measure of test performance. For further discussion of the manifold perspective on modern test theory, see Ramsay (1996). (See also *Fechnerian Psychophysics* for another application of this concept.)

Of course, this is a rather elementary, albeit enormously important, example of a manifold. What comes out of a principal components analysis of p variables is a flat k -dimensional manifold, although there are certainly researchers who have considered how to construct curved manifolds. Multidimensional

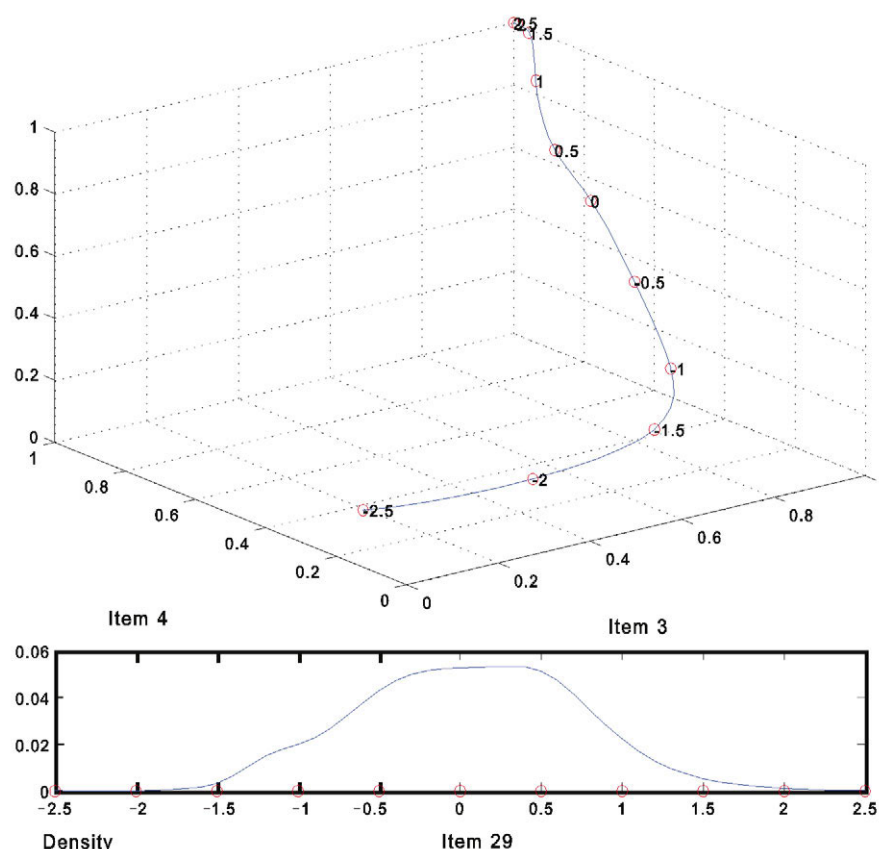


Figure 3

The curve in the upper plot results from plotting probabilities of success for the three items in Fig. 2 against one another. This curve is called a probability manifold. Numbers indicate various trait values on the scale of θ indicated in Fig. 2. The lower panel displays the distribution of trait values

scaling uses distance information, which is essentially topological, to identify similar flat structures, but within which interesting curved substructures are often easily visible.

The advantages of manifolds, as opposed to latent variables, are:

(a) Manifolds are manifest. One can understand how they relate to the data.

(b) Manifolds have topological, measure-theoretic, metric, and algebraic structures that justify using operations such as computing derivatives, for example, to measure the amount of information conveyed by an item about an examinee.

(c) Manifolds do not confuse people. Nobody is misled in using sums, differences, and other operations within these manifolds, whereas the same operations applied to latent variables do not survive the legitimate (i.e., continuous, differentiable, etc.) transformations that may be applied to them.

(d) There is an established, and even rather old, body of mathematical theory, namely differential

geometry, that can be used to develop new methods for data analysis.

It may be that we need to replace models expressed in terms of latent traits by explicit references to the manifolds that they implicitly define.

See also: Classical (Psychometric) Test Theory; Combinatorial Data Analysis; Conjoint Analysis Applications; Dimensionality of Tests: Methodology; Measurement Theory: History and Philosophy; Measurement Theory: Probabilistic; Memory Models: Quantitative; Multidimensional Scaling in Psychology; Scaling and Classification in Social Measurement; Test Theory: Applied Probabilistic Measurement Structures

Bibliography

Costa P T Jr., McCrae R R 1985 *The NEO Personality Inventory Manual*. Psychological Assessment Resources, Odessa, FL

- Hambleton R K, Swaminathan H 1991 *Fundamentals of Item Response Theory*. Sage, Newbury Park CA
- Lord F M 1980 *Applications of Item Response Theory to Practical Testing Problems*. Erlbaum, Hillsdale, NJ
- Ramsay J O 1991 Kernel smoothing approaches to nonparametric item characteristic curve estimation. *Psychometrika* **56**: 611–30
- Ramsay J O 1996 A geometrical approach to item response theory. *Behaviormetrika* **23**: 3–17
- Stevens S S 1946 On the theory of scales of measurement. *Science* **103**: 677–80
- van der Linden W J 1997 *Handbook of Modern Item Response Theory*. Springer, New York

J. O. Ramsay

Psychoneuroimmunology

In defending the organism against the invasion of foreign material, the immune system is critical for the maintenance of health. In times past, the immune system was considered a self-regulating, autonomous agency of defense. Research conducted since the mid-1970s, however, has revealed that immunoregulatory processes are influenced by the brain and that neural and endocrine functions and behavior are influenced by the immune system. Psychoneuroimmunology, then, is the study of the interactions among behavior, neural and endocrine function, and immune system processes (Ader 1981a). The basic premise underlying psychoneuroimmunology is that adaptive responses reflect the operation of a single, integrated network of defenses. While each component of this defensive network evolved to serve specialized functions at special times, each also monitors and responds to information derived from the others. Thus, it is not possible to achieve a complete understanding of immunoregulatory processes without considering the organism and the internal and external environment in which immune responses take place. The evidence for brain-immune system interactions, with particular but selected emphasis on the role of behavioral and psychosocial factors in modulating immune function, is the subject of this article.

First, however, a brief reminder that the immune system's defense of the organism against foreign, 'nonself' material (antigens) is executed by white blood cells, primarily T and B lymphocytes, capable of responding clonally to antigens and retaining the 'memory' of that encounter. Antibody-mediated immunity refers to the exposure of antigens to bone-marrow derived B cells which produce antibodies that protect the organism against extracellular infectious agents. Cell-mediated immunity is provided by thymus-derived T cells that defend against intracellular parasitic and viral infections. An integrated immune response actually involves interactions among

specialized subpopulations of T cells (e.g., helper, cytotoxic), B cells, other white blood cells such as macrophages, and substances (cytokines) secreted by activated immune cells. Lymphocyte proliferation in response to *in vitro* stimulation by mitogens (plant lectins) that preferentially influence T and/or B cells is a measure of the capacity of these lymphocytes to divide in response to nonspecific foreign stimuli. Natural killer (NK) cells are large, granular lymphocytes that are implicated in tumor metastases and in the recognition and defense against viral agents without having had prior experience with that antigen.

1. Historical Perspective

The first sustained program of research on brain-immune system interactions was initiated by Russian investigators in the 1920s. Derived directly from a Pavlovian perspective, these investigators studied the conditioning of immune responses. There was reliable evidence for conditioned alterations in nonspecific immunologic reactions; the evidence for a conditioned change in antibody production was less convincing (Ader 1981b). Early studies on the effects of brain lesions on anaphylactic responses began in the 1950s and were pursued by investigators in the former Soviet Union and in the United States who studied the effects of hypothalamic lesions and electrical stimulation on several parameters of immunologic reactivity.

In the 1950s and 1960s, Fred Rasmussen, Jr., a virologist, and Norman Brill, a psychiatrist (the first such interdisciplinary collaboration), began a series of studies on the effects of stress (avoidance conditioning, physical restraint, electric shock, social 'crowding') on susceptibility to experimentally-induced infectious diseases (e.g., Rasmussen et al. 1957). Depending on the nature of the stressor, these stimuli increased susceptibility to some viral infections and decreased susceptibility to others. These experiments stimulated only a few other investigators to study the effects of early life experiences and stressful psychosocial circumstances on susceptibility to infectious and neoplastic diseases in animals.

In the 1970s, things began to change. Several dramatic observations led to sustained research on the relationship between the brain and the immune system. Hadden et al. (1970) documented the existence of receptors on lymphocytes, providing the first observations linking the immune system to the sympathetic nervous system; Ader and Cohen (1975) demonstrated behaviorally conditioned suppression of the immune system; Bartrop et al. (1977) described immunologic effects associated with the bereavement that followed the death of a spouse; Besedovsky and Sorkin (1977) began their studies of a neuroendocrine-immune system network and the effects of immune responses on neural and endocrine function; the sympathetic innervation of lymphoid tissues was documented

(Felten and Felten 1991); and Blalock and Smith (1980) showed that lymphocytes, themselves, were capable of producing neuropeptides. These seminal studies provided the impetus for new research on receptors for hormones and neurotransmitters on lymphocytes, the effects of sympathetic nervous system (SNS) activity on immunity, conditioned immune responses, depression and immunity, the effects of stressful life experiences on immune function and resistance to disease, psychosocial factors in the progression of AIDS, and immunologic effects on behavior.

2. Brain-Immune System Communication

Two pathways link the brain and the immune system: autonomic nervous system activity and neuroendocrine outflow via the pituitary. Primary (bone marrow, thymus) and secondary (e.g., spleen, lymph nodes) lymphoid organs are innervated with sympathetic nerve fibers that form close neuroeffector junctions with lymphocytes and macrophages and release a variety of neuropeptides. The stimulation or interruption of these connections can influence immune responses. Chemical sympathectomy, for example, suppresses antibody- and cell-mediated immunity; hypothalamic lesions generally inhibit and stimulation potentiates immunologic reactivity.

Conversely, activation of the immune system is accompanied by changes in hypothalamic, autonomic, and endocrine processes. There is, for example, an increase in the firing rate of neurons in the hypothalamus that corresponds to the time of peak antibody production; sympathetic activity (turnover of norepinephrine) is increased in the spleen and hypothalamus; and immune responses initiated by viral infections are associated with increases in circulating levels of ACTH and corticosterone.

Because lymphocytes carry receptors for a variety of hormones and neuropeptides, the cellular interactions that mediate humoral and cellular immunity can be modulated by the neuroendocrine environment in which these responses occur. Lymphocytes bear receptors for corticotropin-releasing factor (CRF), ACTH, and endogenous opioids. There are direct immunomodulatory effects of CRF and ACTH, but their major *in vivo* effects are exerted through interactions with other hormones and immune system products, the most conspicuous of which is the release of adrenocortical steroids. There is an extensive literature documenting the anti-inflammatory and immunosuppressive effects of corticosteroids—at least in pharmacologic doses. At physiologic doses, adrenocortical steroids are essential for normal immune function and, depending on steroid concentration, antigen concentration, and the temporal relationship between hormonal and antigenic stimulation, corticosteroids can be immunoenhancing.

Recent data have also shown that lymphocytes activated by antigenic stimuli are capable of producing neuropeptides and hormones. Thus, the brain is capable of detecting and responding to signals released by an activated immune system. Cytokines, nonantibody messenger molecules released by activated immune cells, besides regulating cellular interactions, are another means by which the immune system communicates with the CNS and, thereby, influences behavior. The precise site(s) at which cytokines act within the brain has not been elaborated. Nonetheless, it is known that cytokines are electrophysiologically, endocrinologically and behaviorally active. Symptoms of sickness, for example, decreased activity, loss of interest in the environment, reduced food intake, fatigue, and cognitive disorders and affective changes can be induced in healthy subjects by the central or peripheral injection of proinflammatory cytokines. Indeed, the recent therapeutic use of interferons in human disease has been associated with neurologic and psychiatric side effects (Meyers 1999).

At the neural and endocrine levels, then, there is abundant evidence of interactions between the brain and the immune system.

3. Behavioral Influences on Immune Function

3.1 Effects of 'Stress' on Immune Function

The link between behavior and immune function goes back to the earliest observations of a relationship between psychosocial factors and susceptibility to immunologically mediated disease processes. There is compelling evidence that psychosocial factors influence the onset or severity of rheumatoid arthritis and other autoimmune diseases, diseases effected by immune reactions generated against an individual's own tissues. Similarly, stressful life experiences are associated with increased susceptibility to infectious diseases and the reactivation of latent viral infections. The death of a spouse, for example, has been associated with depression and an increased morbidity and mortality in response to diseases presumed to involve immune defenses. Bereavement and/or depression is also accompanied by a depression of immune function (reduced lymphoproliferative responses, impaired NK cell activity). In none of these instances, however, has it been demonstrated that these health effects are the result of stressor-induced changes in immune function (see *Stress and Health Research*).

Other 'losses,' such as marital separation and divorce and other chronic stressors such as caregiving for disabled patients, have also been associated with decreases in lymphocyte function and NK cell activity and a depressed antiviral response to an influenza immunization. In individuals seropositive for Epstein-Barr virus (EBV), there is also an increase in anti-EBV

antibody titers reflecting compromised cell-mediated immune system control over the latent virus. Even more seemingly innocuous circumstances such as taking examinations results in a transient increase in levels of distress and a corresponding decrease in lymphoproliferative responsiveness, NK cell activity, percentage of helper T-lymphocytes, and interferon production by stimulated lymphocytes. Exam periods are also associated with elevated EBV titers and an increase in the incidence of self-reported symptoms of infectious illness.

Clinical studies have described an association between some immunologically mediated diseases and clinical depression. Multiple sclerosis (MS), a T cell mediated autoimmune disease, is associated with depressive states. Systemic lupus erythematosus (SLE) is a B cell mediated autoimmune disease in which depression is frequently reported. Depressive features have also been described in HIV-infected individuals. Thus, considerable attention has been given to the effects of clinical depression on immune function. Based on a meta-analysis of the literature (Herbert and Cohen 1993), depression is reliably associated with a decline in both enumerative and functional measures of immunity such as lymphoproliferative responses and NK cell activity. Depressed patients also show elevated antibody titers to herpes simplex virus (HSV) and cytomegalovirus and a decreased ability to mount a specific cell-mediated response to varicella zoster virus. The decline in cellular immunity is comparable to the decline observed in elderly patients who are more susceptible to herpes zoster infection than younger adults (Irwin et al. 1998). Other measures of immune function in depressed patients are characterized by a high variability within and between studies.

Typically, acute laboratory stressors such as unsolvable puzzles or public speaking have been reported to increase numbers of cytotoxic lymphocytes and the number and activity of NK cells along with a decrease in lymphoproliferative responses. These data, too, are characterized by high variability that is, at least in part, related to individual differences in autonomic nervous system reactivity. For example, subjects who are characterized as high SNS reactors in response to an acute laboratory stressor show a greater decrease in lymphoproliferative responses, a greater increase in the number of suppressor/cytotoxic T cells and, perhaps, a greater increase in NK cell activity than low reactors.

It is clear that, while stressful life experiences influence immune function in humans, there are large individual differences due, presumably, to personality and dispositional factors that influence the individual's perception and appraisal of the environmental circumstances, the individual and social resources available to cope effectively with the demands of the environment, and the biological substrate upon which the stressful stimuli are superimposed.

Most of the evidence for stress-induced alterations in immunity comes from basic research on animals. Stressors such as restraint, electric shock, avoidance conditioning, and noise can disrupt immune function and increase susceptibility and/or mortality to a variety of infectious agents (e.g., HSV, influenza, Coxsackie virus, Sindbis virus) or allow an otherwise inconsequential exposure to a pathogen to develop into clinical disease (Moynihan and Ader 1996, Sheridan et al. 1994).

In contrast, though, some of these same stressors *decreased* susceptibility to poliomyelitis virus infection, experimental allergic encephalomyocarditis and *Plasmodium berghei*, a rodent malaria. Stressful life experiences also exert a protective effect on the development of experimental allergic encephalomyelitis, a murine model of MS. In the case of experimentally induced arthritis, increased or decreased susceptibility depends on the experimental model of arthritis and the nature of the stressor.

Prenatal and early life experiences such as interruptions of mother-young interactions, the social environment in which animals live, exposure to predators or pheromones emitted by stressed conspecifics, and subjecting animals to environmental conditions over which they have no control (e.g., physical restraint, electric footshocks, forced exercise) are among the psychological and social manipulations that alter neuroendocrine states and modulate immune responses. Antibody production is generally suppressed by such stressful stimulation.

Although differential housing (frequently mislabeled as 'crowding' or 'isolation') may not qualify as a stressor, the manner in which animals are housed or a change in housing conditions is sufficient to influence the antibody response to different antigens. Primary and/or secondary antibody responses in rodents are suppressed by restraint or electric shock stimulation, depending on the nature of the antigen and the timing of the stressor in relation to the antigenic stimulation. Other stressors, however, are associated with enhanced antibody production. Similarly, cell mediated processes (e.g., graft rejection, delayed type hypersensitivity reactions) and nonimmunologically specific responses (e.g., lymphocyte proliferation in response to mitogens, NK cell activity, cytokine production) generally are suppressed in animals subjected to stressors but, again, the nature and timing of the stressor, the particular response being measured as well as the species, strain and sex of the animal will determine the direction and magnitude of the effects.

A few attempts have been made to discriminate between stressors over which the individual does or does not have control, for example, avoidable vs. unavoidable, escapable vs. inescapable, signaled vs. unsignaled electric shock in animals. Although differences have been observed in some studies, the results have been contradictory. In humans, too, studies involving control—or perceived control—over stress-

ful stimulation have yielded inconsistent results. Given the number of variables identified as having an effect on the assessment of immunologic reactivity, however, there are insufficient data to warrant any conclusions at this time.

While many hypotheses have been proffered, the neuroendocrine mechanisms presumed to mediate the effects of stressful life experiences on immune function remain elusive. For one thing, our understanding of neuroendocrine-immune system interactions under normal and stressful conditions is incomplete. Glucocorticoids, for example, are generally immunosuppressive, and it is generally assumed that corticosteroid elevations, the most common manifestation of the stress response are responsible for the immunosuppression associated with stressful life experiences. There are many examples of stress-induced, corticosteroid-mediated alterations of immune responses. But, there are very many other observations of stress-induced alterations of immunity that are independent of adrenocortical activity. Intermittent and continuous schedules of inescapable electric footshocks, both of which elevate corticosterone levels, effect an analgesic response to subsequent footshock. Only the intermittent shock, however, is an opioid-mediated analgesia and only the intermittent footshock suppresses NK cell activity, illustrating the immunomodulating potential of endogenous opioids.

Even without being able to detail the extensive literature, it is clear that the response to stressful life experiences involves complex neural, endocrine, and immune response interactions (Glaser and Kiecolt-Glaser 1994, Rabin 1999). Different stimuli commonly referred to as stressors—and commonly thought to elicit a representative ‘stress’ response—can have different effects on the same immune response. Conversely, the same stressor can have different effects on different immune responses. Also, the intensity and chronicity of the stressor are among the several parameters of stimulation which influence immune responses. In general, then, we have an incomplete picture of the direction, magnitude, and duration of the effects of stressful life experiences on immunity. We do, however, know something about the factors that contribute to the observed variability. They include:

- (a) the quality and quantity of stressful stimulation;
- (b) the capacity of the organism to cope effectively with stressful circumstances;
- (c) the quality and quantity of immunogenic (or pathogenic) stimulation;
- (d) the temporal relationship between stressful stimulation and immunogenic stimulation;
- (e) the parameters of immune function and the sample times chosen for measurement;
- (f) the experiential history of the organism and the prevailing social and environmental conditions upon which stressful stimulation and immunogenic stimulation are superimposed;

- (g) a variety of host factors such as species, strain, age, sex, nutritional state and current disease; and
- (h) interactions among these several variables.

3.2 Conditioning Effects

Behaviorally conditioned alterations of immune function provide one of the more dramatic lines of evidence linking the brain and the immune system. In a prototypical study (Ader and Cohen 1975), a novel, distinctively-flavored drinking solution (e.g., saccharin), the conditioned stimulus (CS) is paired with the injection of an immunomodulating drug, the unconditioned stimulus (UCS). When animals are subsequently injected with antigen, some of them are re-exposed to the CS, alone. When the immunosuppressive drug, cyclophosphamide (CY), was used as the UCS, conditioned animals re-exposed to the CS showed an aversion to the saccharin solution and an attenuated antibody response relative to conditioned animals that were not re-exposed to the CS or nonconditioned animals that were exposed to saccharin. Several independent investigators have verified and extended these findings.

The effects of conditioning have been consistent under a variety of experimental conditions (Ader and Cohen 2001). Studies have documented the acquisition and/or extinction of conditioned nonspecific host defense responses such as NK cell activity and different cell-mediated as well as antibody-mediated immune responses using different unconditioned and conditioned stimuli. Also, conditioning is not limited to changes associated with taste aversion learning; several studies have described the acquisition and extinction of conditioned ‘stress’ effects. Moreover, there is no consistent evidence of a relationship between conditioned behavioral and conditioned immune changes. Taste aversions can be expressed without concomitant changes in immune function, and conditioned changes in immunologic reactivity can be obtained without observable avoidance conditioning. Also, the available data indicate that conditioned immunosuppressive responses cannot be attributed to either stress-induced or conditioned elevations of adrenocortical steroids.

The acquisition and extinction of a conditioned *enhancement* of immunologic reactivity have been observed using antigens rather than pharmacologic agents as UCSs. In the first such study (Gorczynski et al. 1982), CBA mice were anesthetized, shaved, grafted with skin from C57BL/6J mice, and left bandaged for 9 days. In response to the grafting of allogeneic tissue, there was, as expected, an increase in the number of precursors of cytotoxic T lymphocytes (CTLp) that could react against the foreign (nonself) tissue. The conditioning procedures were repeated when CTLp numbers returned to baseline levels. On the fourth such trial, the procedures were repeated again, but the tissue graft was omitted. Approximately half the

animals in each of several experiments showed an increase in CTLp in response to the grafting procedures alone. Half of the 'responder' mice were given additional conditioning trials and half received extinction trials (sham grafting). When tested again, all the animals given additional conditioning trials showed the conditioned increase in CTLp whereas none of the 'responder' mice that experienced unreinforced trials showed a conditioned response.

In another study on conditioned immunoenhancement, mice were immunized repeatedly with low doses of a common laboratory antigen following exposure to a gustatory CS. Three weeks after the last trial, half the animals in each of the several groups were re-exposed to the CS, alone, or were re-exposed to the CS in the context of a low-dose booster injection of the same antigen. The latter groups were added to determine if conditioning could be more sensitively indexed by asking if a CS could potentiate the effects of a minimally effective dose of antigen rather than initiate the production of antibody by itself. In this experiment, neither the CS nor the booster dose of antigen, alone, were sufficient to elicit a robust antibody response. However, antibody production was enhanced when conditioned mice were re-exposed to the CS in the context of re-exposure to a minimally immunogenic dose of that same antigen. Recently, others have observed a conditioned enhancement of antibody production after a single conditioning trial and without the need for a booster injection of antigen.

Data on conditioning in humans is limited. The anticipatory (conditioned) nausea that frequently precedes immunosuppressive cancer chemotherapy was also associated with anticipatory suppression of proliferative responses to T cell mitogens. In healthy subjects, the pairing of a distinctive flavor with injections of adrenaline resulted in an enhancement of NK cell activity when subjects were subsequently re-exposed to the CS alone. Preliminary data on cell-mediated immunity derived from a study in which subjects were repeatedly injected with tuberculin obtained from a green vial and with saline drawn from a red vial. On the test trial, the contents of the vials were switched. Saline administered to the arm previously treated with tuberculin did not elicit a skin reaction, but there was an attenuated response in the arm previously injected with saline. These studies lend credence to earlier anecdotal reports of what would now appear to be examples of conditioned alterations of immunologic reactivity in human subjects or patients.

4. Biologic Impact of Behaviorally-induced Alterations of Immune Function

Many studies have related stressful life experiences to changes in immunity and/or changes in the onset or progression of disease (Biondi and Zinnino 1997), and

these correlational studies are, indeed, provocative. The association between stressful life events and increased morbidity and the association between stressful life events and alterations in immune function do not, however, establish a causal chain linking stressful life events, immune function, and health or disease. There are, at present, few, if any, human studies in which altered resistance to disease has been demonstrated to be a direct result of biologically relevant changes in immune function induced by psychosocial or stressful life experiences. Such data are being collected from animal models of spontaneously occurring disease or models in which it is possible to precipitate adaptive responses to immunogenic stimuli experimentally.

In mice infected with influenza virus, for example, there is a delay in the production of virus-specific antibody in restrained relative to unrestrained animals. In mice inoculated with HSV, prolonged periods of restraint or repeated periods of electric shock suppressed NK cell activity and the development of HSV-specific cytotoxic T lymphocytes. Also, higher titers of infectious virus were recovered from the restrained mice. Other recent studies have shown that psychosocial factors can reactivate latent HSV. While restraint was ineffective, disruption of the social hierarchy within a colony of mice by repeatedly introducing an intruder into a group cage increased aggressive behavior, activated the hypothalamic-pituitary-adrenal axis (HPA), and resulted in reactivation of latent HSV in more than 40 percent of the latently infected animals.

In an NK-sensitive tumor model in which lung metastasis is related to NK cell function, it has been found that, depending upon when the stressor was imposed in relation to the effect of NK cells on the metastatic process, forced swimming, acute alcohol intoxication and social confrontations decreased NK cell activity and increased lung metastases. This finding is consistent with strain, age, and sex differences in NK cell activity and the development of lung metastases.

In defending the organism against infection, inflammatory processes are an essential component in the healing of wounds. Both the SNS and the HPA axis are capable of modulating inflammatory and immune processes. Thus, stressful life experiences would be expected to influence wound healing as, in fact, it does. Chronically stressed caretakers of Alzheimer's patients took 24 percent longer than control subjects to heal a standard, experimentally produced wound in the skin, and there was a 40 percent delay in wound healing in dentistry students tested just before an examination period (when perceived stress was high) compared to healing during a summer vacation (when perceived stress was low). Paralleling this human research, recent animal studies have found that mice restrained for 3 days before and 5 days after wounding showed a reduced inflammatory response, an elevated

corticosterone level and a dramatic delay in the rate at which the wounds healed.

These animal models of wound healing, NK-sensitive tumors and reactivation of latent virus infection will enable research designed to examine the mechanisms underlying the effects of stressful life experiences and determine whether they are, as suggested, due to behaviorally-induced alterations in immune function.

The spontaneous development of a systemic lupus erythematosus-like disease in New Zealand mice was used to assess the biologic impact of conditioned alterations in immune responses since a suppression of immunologic reactivity would be in the biological interests of these animals. Half the weekly treatments received by the lupus-prone mice consisted of saccharin plus an ip injection of saline (the CS) rather than the CS plus an injection of cyclophosphamide. Under this pharmacotherapeutic regimen, the onset of disease was delayed using a cumulative amount of immunosuppressive drug that was not, by itself, sufficient to alter progression of the autoimmune disorder.

Consensual validity for these observations was provided by a study in which re-exposure to a CS previously associated with CY accelerated mortality in conditioned mice in response to a transplanted plasmacytoma. It has also been shown that protection against adjuvant-induced arthritis could be achieved by exposing animals to a CS previously paired with an immunosuppressive drug or with a stressful (immunosuppressive) experience. Among mice previously conditioned by the pairing of a CS with an immunosuppressive drug, re-exposure to the CS following transplantation of allogeneic tissues prolonged survival of the allografts. Clearly, conditioned immunologic effects can have a profound biologic impact on the development of autoimmune disease, neoplastic processes, and graft survival. Thus far, there is only one clinical case study in which a conditioning protocol was successfully used to reduce the total amount of cytoxan used in the treatment of a child with SLE (Olness and Ader 1992).

5. Conclusions

'Psychoneuroimmunology' is an interdisciplinary field of study that has developed and now prospers by exploring and tilling fertile territories secreted by the arbitrary if not illusory boundaries of the biomedical sciences. Disciplinary boundaries and the bureaucracies they created are biological fictions that can restrict imaginative research and the transfer and application of technologies. They lend credence to Heisenberg's assertion (1958) that 'What we observe is not nature itself, but nature exposed to our method of questioning' (Ader 1995).

A paradigm shift is occurring, however. Based on research conducted since the mid-1970s, it has become evident that immunoregulatory processes can no longer be studied as the independent activity of an autonomous immune system. We now know of at least two pathways through which the brain communicates with the immune system: autonomic nervous system activity and neuroendocrine outflow via the pituitary. Both pathways generate chemical signals that can be perceived by the immune system via receptors on the surface of lymphocytes and other immune cells. Conversely, an activated immune system generates chemical signals that can be perceived by the nervous system. Thus, there is a bidirectional exchange of information between the brain and the immune system and all immunoregulatory processes take place within a neuroendocrine milieu that is responsive to the individual's perception of and adaptive responses to physical and psychosocial events occurring in the external world.

It should not be surprising, then, that stressful life experiences or Pavlovian conditioning could influence immune function. The conditioned suppression or enhancement of immune responses has been demonstrated in different species and under a variety of experimental conditions. Also, stressful life experiences influence immune function. Both conditioning and stressful stimulation exert biologically meaningful effects in that they appear to be implicated in altering the development and/or progression of immunologically mediated disease processes. While these are reproducible phenomenon, the extent to which one can generalize from one stressor or from one parameter of immune function to another is limited. Also, the direction and/or magnitude of the effects of behavioral factors in modulating immune responses depend on the nature of the behavioral circumstances; the nature of the antigenic stimulation and the temporal relationship between them; the immune response and time at which it is measured; a variety of host factors; and the interactions among these variables.

The neuroendocrine pathways underlying the effects of conditioning or stressful life experiences on immune responses are not yet known. It would appear, though, that the neural and endocrine changes associated with changes in behavioral states and the network of connections between the brain and the immune system provide multiple pathways through which behavioral processes could influence immune defenses. The existence of these bidirectional pathways reinforces the hypothesis that immune changes constitute an important mechanism through which psychosocial factors could influence health and disease.

See also: Endocrinology and Psychiatry; Stress and Cardiac Response; Stress, Neural Basis of; Stress: Psychological Perspectives

Bibliography

- Ader R (ed.) 1981a *Psychoneuroimmunology*. Academic Press, New York
- Ader R 1981b A historical account of conditioned immunobiologic responses. In: Ader R (ed.) *Psychoneuroimmunology*. Academic Press, New York
- Ader R 1995 Historical perspectives on psychoneuroimmunology. In: Friedman H, Klein T W, Friedman A L (eds.) *Psychoneuroimmunology, Stress and Infection*. CRC Press, Boca Raton, FL
- Ader R, Cohen N 1975 Behaviorally conditioned immunosuppression. *Psychosomatic Medicine* **37**: 333–40
- Ader R, Cohen N 2001 Conditioning and immunity. In: Ader R, Felten D L, Cohen N (eds.) *Psychoneuroimmunology*, 3rd edn., Academic Press, New York
- Bartrop R W, Lazarus L, Luckhurst E, Kiloh L G, Penny R 1977 Depressed lymphocyte function after bereavement. *Lancet* **i**: 834–37
- Besedovsky H, Sorkin E 1977 Network of immune-neuroendocrine interactions. *Clinical and Experimental Immunology* **27**: 1–12
- Biondi M, Zannino L-G 1997 Psychological stress, neuroimmunomodulation, and susceptibility to infectious diseases in animals and man: A review. *Psychotherapy and Psychosomatics* **66**: 3–26
- Blalock J E, Smith E M 1980 Human leukocyte interferon: Structural and biological relatedness to adrenocorticotrophic hormone and endorphins. *Proceedings of the National Academy of Science, USA* **77**: 5972–74
- Felten S Y, Felten D L 1991 Innervation of lymphoid tissue. In: Ader R, Felten D L, Cohen N (eds.) *Psychoneuroimmunology*, 2nd edn. Academic Press, New York
- Glaser R, Kiecolt-Glaser J K (eds.) 1994 *Handbook of Human Stress and Immunity*. Academic Press, San Diego, CA
- Gorczynski R M, Macrae S, Kennedy M 1982 Conditioned immune response associated with allogeneic skin grafts in mice. *Journal of Immunology* **129**: 704–9
- Hadden J W, Hadden E M, Middleton E 1970 Lymphocyte blast transformation. I. Demonstration of adrenergic receptors in human peripheral lymphocytes. *Journal of Cellular Immunology* **1**: 583–95
- Heisenberg W 1958 *Physics and Philosophy: The Revolution in Modern Science*, 1st edn. Harper, New York
- Herbert T B, Cohen S 1993 Depression and immunity: A meta-analytic review. *Psychology Bulletin* **113**: 472–86
- Irwin M, Costlow C, Williams H, Artin K H, Levin M J, Hayward A R, Oxman M N 1998 Cellular immunity to varicella-zoster virus in patients with major depression. *Journal of Infectious Diseases* **178S**: 104–8
- Meyers C A 1999 Mood and cognitive disorders in cancer patients receiving cytokine therapy. *Advances in Experimental Medicine and Biology* **461**: 75–81
- Moynihan J A, Ader R 1996 Psychoneuroimmunology: Animal models of disease. *Psychosomatic Medicine* **58**: 546–58
- Olness K, Ader R 1992 Conditioning as an adjunct in the pharmacotherapy of lupus erythematosus. *Journal of Developmental and Behavioral Pediatrics* **13**: 124–5
- Rabin B S 1999 *Stress, Immune Function, and Health*. Wiley-Liss, New York
- Rasmussen A F, Marsh J T, Brill N Q 1957 Increased susceptibility to herpes simplex in mice subjected to avoidance-learning stress or restraint. *Proceedings of the Society for Experimental Biology and Medicine* **96**: 183–9
- Sheridan J F, Dobbs C, Brown D, Zwilling B 1994 Psycho-

neuroimmunology: Stress effects on pathogenesis and immunity during infection. *Clinical Microbiological Review* **7**: 200–12

R. Ader

Psychopharmacotherapy, Costs of

Psychopharmacotherapy is one of the mainstays of the treatment of psychiatric disorders besides social therapy and psychotherapy. Psychiatric disorders essentially comprise the affective disorders (depression, bipolar, i.e., manic-depressive disorder), the schizophrenias, the anxiety disorders (panic disorder, generalized anxiety, post-traumatic stress disorder, obsessive-compulsive disorder, social phobia), the eating disorders (bulimia and anorexia nervosa), the substance abuse and dependence disorders including alcoholism, the personality disorders, and the dementias (with Alzheimer-type dementia being the cause in about 70 percent of cases). They are highly prevalent (Table 1) and frequently take a chronic or chronic remittent course. Therefore, it is easily understandable that the total costs of psychiatric disorders compare very well with the costs of the leading somatic illnesses (Table 2). The prevalences of depression and of the eating disorders are increasing. The WHO burden of illness study (Murray and Lopez 1997) suggests that depression will be the leading cause of disability in the developed countries in 2020. Psychiatric disorders (except the dementias) manifest in early adulthood thus limiting the abilities of patients for social achievement.

Psychopharmacotherapy essentially comprises seven classes of drugs:

- (a) neuroleptics for the treatment of psychotic states, especially the schizophrenias;
- (b) antidepressants for the treatment of depression of whatever origin as well as for the long-term treatment of anxiety disorders;
- (c) mood stabilizers (lithium, some anticonvulsants) for the short-term treatment of mania and for the long-term relapse and recurrence prevention of affective disorders;
- (d) minor tranquilizers (essentially benzodiazepines) for the short-term treatment of anxiety and agitation of whatever origin and of alcohol withdrawal syndromes;
- (e) hypnotics (essentially benzodiazepines and benzodiazepine-like drugs) for short-term treatment of insomnia;
- (f) anticraving compounds (bupropion for smoking cessation, acamprosate for alcohol abstinence, opiate antagonists for alcoholism and opiate dependence) which, however, are not yet approved in all countries; and

Table 1

Annual rate of behavioral disorders in the USA
(adapted from NIMH)

Disorder	Percent afflicted
Anxiety disorders (all)	13
Depressive disorders (all)	10
Alcohol abuse/dependence	7
Drug abuse/dependence	3
Severe cognitive impairment	3
Antisocial personality	2
Schizophrenia	1
Somatization	0.2

(g) cognition enhancers (nootropics such as piracetam, ginkgo biloba) for the nonspecific treatment of dementias and acetylcholinesterase inhibitors specifically for Alzheimer-type dementia.

Concerning the eating disorders, only one drug, the antidepressant fluoxetine, is approved for supportive treatment of bulimia.

Psychopharmacotherapy competes with psychotherapy in the treatment of mild to moderate depression and the anxiety disorders. The relative merits of both alternatives are subject to debate which is fueled by a lack of scientific data, ideology, and economic interests of the health care providers involved. As a rule of thumb, pharmacotherapy and psychotherapy may be thought not to be alternatives but supplements whenever a psychiatric disorder is moderate to severe, although a synergistic action remains to be definitely proved.

Life expectancy increased dramatically in the twentieth century, and this was accompanied by an increase in the prevalence of age-associated diseases. In the field of psychiatry these are the dementias. Thus, although the number of healthy life years is generally increasing, the prevalence of chronic diseases is increasing in parallel. Consequently, the demand for health care services is increasing. Costs for drug therapy are increasing overproportionately because drug development has become increasingly complicated and costly. The mean investment necessary for developing a single new chemical entity to approval amounts to about US\$500 million.

Worldwide, health care is one of the most rapidly growing markets with positive impacts on employment. Nevertheless, governments internationally tend to restrict the growth because at least part of it is publicly funded. The fundamental but unanswered question is what proportion of the gross domestic product society is willing to invest in health care. Obviously, however, most people value health and thus health care very highly. Thus, disorders, i.e., the patients suffering from them, compete for limited resources. Economic criteria have become a third dimension besides efficacy and effectiveness as well as tolerability and safety when selecting between treatment options. It is easily conceivable that psychiatric disorders, faced with persistent stigma, have a hard standing in this competition.

Pharmacoeconomics, a young scientific discipline, has developed a sound scientific basis for the evaluation of the economic dimension (Gold et al. 1996). This is indispensable in the interests of an ethically justifiable allocation of resources. The question is

Table 2

Estimated annual total costs of some psychiatric and somatic diseases (American Diabetes Association 1996, American Heart Association 1998, Davidson et al. 1999, Greenberg et al. 1993, 1999, Whetten-Goldstein et al. 1998)

Disease	Estimated total cost (\$US billion)	Year considered
Schizophrenia	33	1991
Depression	43.7	1990
Bipolar (manic-depressive) disorder	25.6	1991
Alcoholism	98.6	1991
Anxiety disorders	42.3	1990
Alzheimer's disease	87.6	1991
Coronary heart disease	43	1990
Coronary heart disease	99.8	1999
Cancer	104	1991
Acquired immunodeficiency syndrome (AIDS)	66	1991
Multiple sclerosis	100.8	1994
Diabetes mellitus	91.8	1992
Hypertensive disease	33.3	1999
Congestive heart failure	21	1999
Stroke	45.3	1999

whether a new, more costly drug allows one to achieve better outcomes so that it is worth its higher costs. In principle, a new drug might even spare money in that the better outcome might help avoid costly, higher-level health care such as inpatient treatment and reduce indirect costs, e.g., by less absenteeism.

1. Basic Terms of Pharmacoeconomics

Pharmacoeconomics is the science of measuring *costs* and *outcomes* associated with the use of pharmaceuticals in health-care delivery. Its objective is to improve public health through rational decision making when selecting among alternative therapies, e.g., for formularies and their impact on costs and outcomes. Several methodological guidelines have been published (e.g., CCOHTA 1994).

Outcomes are the possible end result of a particular treatment. Outcomes depend on the *effectiveness* of the treatment under the influence of all the confounding variables in real life. Therefore, costs have to be measured under real-life conditions. In contrast, the experimental *efficacy* trials are designed to eliminate as completely as possible all confounding influences. The outcome parameter chosen should have high everyday relevance, e.g., the rate of full remissions of depression.

Costs are the resources consumed by the illness and by its treatment. *Direct costs* are those directly related to treatment such as prescribed medication, self-medication, consulting the physician, blood tests, transportation, psychotherapy, occupational therapy, informal care by relatives, etc., as well as costs for hospital and nursing-home treatment. *Direct costs* include those for diagnosis and treatment of side effects and complications of the initial therapy. There is methodological controversy as to whether the costs of social support (welfare, unemployment benefit, disability pension, housing benefit, etc.) should be included in the direct costs. A complete evaluation of direct costs should consider the opportunity costs, i.e., the lost potential of the resources invested to provide more efficient benefits in other fields.

Indirect or future costs are those related to lost productivity due to illness, i.e., due to absenteeism, premature pensioning because of disability, and premature death (e.g., by suicide) as well as those related to health care consumption during the years of life gained owing to the intervention. Concerning psychiatric disorders, costs of violence and crime also have to be considered. *Intangible costs* comprise those attributable to impairments of the quality of life, e.g., of the patients themselves and their relatives and acquaintances involved in the care.

The economic impact of treatment extends into the future. If costs and benefits accrue during different periods, costs and savings have to be discounted for the remaining life-span of the patient sample in question. Discount rates are supposed to reflect the

rates of return on private-sector investments and are generally set at 3–5 percent.

If the costs are low and the outcomes excellent the decision in favor of the treatment is obvious as is its rejection if the costs for poor outcomes are high. Therefore, pharmacoeconomic analyses make sense only when both the costs and the quality of treatment are high or low. The economic *efficiency* then is the ratio of costs divided by the outcomes. When allocating resources the drug yielding the most favorable outcomes at the lowest cost will be favored.

2. Types of Pharmacoeconomic Analysis

The basic requirement of a pharmacoeconomic analysis is the definition of the perspective, e.g., that of the patient, the health care provider, the health insurance as the payer, or the societal perspective of government. The categories of costs to be considered depend on this perspective. For example, from a provider's perspective, the value of an antidepressant drug under capitation may be expressed in terms of cost per reduction in some depression scale score. From the patient's perspective, the value may be in cost per full treatment response with least side effects. From the societal perspective, the value may be in cost per patient returning to continuous employment at or above the level of previous education. Simply to compare the daily costs of medication does not fulfill the criteria of pharmacoeconomic analysis.

The principle types of analysis include cost-minimization analysis, cost-effectiveness analysis, cost-utility analysis and cost-benefit analysis. *Cost-minimization* analysis (CMA) assumes that the interventions in question yield identical outcomes; this is rarely valid. *Cost-effectiveness* analysis (CEA) compares the costs of alternative treatments in terms of a specific health outcome such as life-years gained, e.g., by reducing mortality from suicide, or responder rates. *Cost-utility* analysis (CUA) adjusts the life-years gained by some measure for quality of life. Thus, the results of CUA are typically expressed as costs per quality adjusted life years (QALYs). Quality of life is assessed either by some rating scale for health-related quality of life or, more typically, by determining the preferences of patients suffering from specific diseases for one or the other disease state as opposed to perfect health where perfect health is set as '1' and death as '0.' In *cost-benefit* analysis (CBA), the QALYs gained are transformed into monetary terms using the human capital approach or willingness-to-pay approach.

As there is no one-to-one relationship between therapeutic intervention and outcome and its financial sequelae, the impact of these uncertainties has to be estimated. This is done by *sensitivity analysis*, where key variables are systematically modified and the economic consequences calculated yielding best- versus worst-case scenarios.

3. Pharmacoeconomic Study Designs

There are two principle approaches to pharmacoeconomic analysis: the clinical trial approach (experimental or observational), and the modeling or systems approach. The former means gathering cost data (retrospectively or, preferably, prospectively) in clinical trials. The latter '*in vitro*' approach applies decision-tree analysis to data available from published literature and databases.

Current guidelines favor the clinical trial approach as the 'gold standard.' Modeling, however, can overcome limitations implicit in clinical trials. These include the selection bias concerning the patients recruited in randomized, strictly controlled, and therefore artificial efficacy trials, thus limiting generalizability. Furthermore, their observation period rarely exceeds a few months whereas economic consequences may extend far into the future. Efficacy trials necessarily are underpowered for economic analysis because sample size estimation is based on the critical difference and variability of the efficacy parameter. Obviously, the critical difference and variability of economic data will fundamentally differ. Clinical trials using the mirror image design, i.e., comparing costs and benefits in a defined period before and after administering a specific drug, are of limited value because at least in psychiatric disorders there is a spontaneous decrease of the costs per outcome during the course of illness irrespective of the mode of treatment.

The putatively best experimental design is a prospective, randomized, 'naturalistic,' i.e., non-blind, adequately powered clinical trial. Both principle approaches, the clinical trial approach and modeling, have their weaknesses and strengths so that they are not really exclusive alternatives but complementary.

Friedberg et al. (1999) found that studies sponsored by drug companies were less likely to report unfavorable results than studies funded by nonprofit organizations. This finding could be due to various sources of bias which, however, if present, still have to be identified.

The organization of health-care supply, the health-care utilization patterns, and the prices differ considerably between and even within cultures. Therefore, the results of pharmacoeconomic analyses cannot simply be transferred from one health care market to another.

4. Pharmacoeconomic Findings with Neuroleptics

Since the 1990s new neuroleptics have been introduced (amisulpride, clozapine, olanzapine, quetiapine, risperidone, sertindole, zotepine), although not yet approved in all countries. These compounds are called 'atypical' because they have a lower incidence of extrapyramidal side effects (e.g., Parkinsonism) than conventional neuroleptics (e.g., haloperidol, fluphenazine).

Moreover, they possibly have superior efficacy concerning cognitive dysfunction and so-called negative symptoms as well as quality of life in schizophrenia. These advantages might improve compliance, which is most important in long-term relapse and recurrence prevention. The costs of daily treatment with atypical neuroleptics exceed those of conventional neuroleptics by a factor of 2–8 depending on the compound and the country considered. On the other hand, neuroleptic medication costs generally make only about 4–8 percent of direct costs.

The pharmacoeconomic analyses actually available for neuroleptics are limited to amisulpride, clozapine, olanzapine, risperidone, and sertindole in schizophrenia. The majority of studies address clozapine, olanzapine, and risperidone. The costs are mostly considered from the payer's perspective, i.e., addressing direct costs. Most are cost-effectiveness analyses, some cost-utility analyses. The majority follow the modeling approach. The economic analyses based on clinical efficacy trials have the limitations discussed above, especially those restricting the analysis to study completers. This is far from the naturalistic, real-world approach necessary. There is only one randomized, naturalistic, adequately powered trial comparing clozapine with conventional neuroleptics, where clozapine significantly reduced the frequency and duration of inpatient treatment; the cost data are to be reported. All in all, the current evidence suggests that clozapine is cost-effective for neuroleptic-refractory schizophrenia and that olanzapine and risperidone therapy may be cost-neutral despite higher medication costs (Revicki 1999). Nevertheless, the access of patients to the new therapeutic chances is restricted on economic grounds, although this varies internationally.

5. Pharmacoeconomic Findings with Antidepressants

Since the 1980s, a number of new antidepressants have been introduced (selective serotonin reuptake inhibitors (SSRI): citalopram, fluoxetine, fluvoxamine, paroxetine, sertraline; the selective norepinephrine reuptake inhibitor reboxetine; the serotonin–norepinephrine reuptake inhibitor venlafaxine; the norepinephrine auto- and heteroreceptor blocker mirtazapine; the reversible and selective inhibitor of monoamine oxidase moclobemide), although not yet approved in all countries. Their efficacy is comparable to that of traditional tricyclic antidepressants (TCAs such as amitriptyline, imipramine, desipramine, nortriptyline), but owing to selectivity their tolerability and safety are superior. Safety is of specific concern in view of the suicide risk, although antidepressants are involved only in the minority (< 10 percent) of suicides. These advantages help reduce drop-out rates although to a limited degree (about 25 percent versus

29 percent). Nevertheless, this might be particularly relevant in long-term recurrence prevention which is necessary in about 60 percent of patients in whom antidepressants are indicated. There is a dramatic underrecognition and undertreatment worldwide (Davidson et al. 1999, Tylee et al. 1999). Undertreatment causes a doubling of direct costs due to increased utilization of other medical services (Manning and Wells 1995) and increases indirect costs due to absenteeism. The advantages of innovative antidepressants might increase acceptability and thus utilization of antidepressants. The costs of daily treatment with selective antidepressants, however, exceed those of TCAs by a factor of 2–20 depending on the compound and the country considered. On the other hand, antidepressant medication costs generally constitute only about 3–6 percent of direct costs.

Pharmacoeconomic analyses published cover all innovative antidepressants except reboxetine. All but one (Gould et al. 1995) are restricted to depression. The costs are mostly considered from the payer's perspective. Most are cost-effectiveness analyses, only a few cost-utility and cost-benefit analyses. There are no randomized, open-label, naturalistic, adequately powered trials but some studies analyzed large, complete samples of HMO participants. By far the majority follow the modeling approach. There are some outstanding analyses in which classes of compounds are compared (e.g., CCOHTA 1997, Casciano et al. 2000). Overall, the current evidence suggests that extended-release venlafaxine might be more cost-effective than SSRIs and that the cost-effectiveness of SSRIs might equal or even exceed that of TCAs, despite higher medication costs (Conner et al. 1999, Casciano et al. 2000). However, more research is needed before this can be taken for granted. Accordingly, there is debate as to whether the access of patients to innovative antidepressants should be restricted or whether they should be recommended as first-choice drugs (CCOHTA 1997).

The limited evidence available suggests that psychotherapy is more likely to be cost-effective the more severely afflicted the patients are (Gabbard et al. 1997). The cost of psychotherapy is about 3–6 times higher than that of antidepressant drug therapy. One modeling study (Antonuccio et al. 1997) suggests, however, that cognitive behavioral therapy (CBT) might be more cost-efficient than fluoxetine treatment of depression owing to lower recurrence rates after psychotherapy. In a CUA based on a randomized controlled trial (Lave et al. 1998), nortriptyline was slightly superior to interpersonal psychotherapy (IPT). One meta-analysis (Gould et al. 1995) suggests that CBT is superior to imipramine and cost-effective in panic disorder. Economic analyses addressing the relative merits of pharmacotherapy and psychotherapy are completely lacking for prominent diseases such as obsessive-compulsive disorder, generalized anxiety, and social phobia.

6. Pharmacoeconomic Findings with Mood Stabilizers

Mood stabilizers by definition reduce the risk of relapses and recurrences and thus putatively hospitalization. Therefore, they may be expected to reduce direct costs (about 60 percent for hospitalization). Lithium is the only drug reducing mortality essentially by reducing the suicide rate (Ahrens 1997). Only a few pharmacoeconomic analyses have been published. Lithium was more cost-effective than carbamazepine from the payer's perspective in a modeling approach based on retrospectively gathered data from a randomized controlled clinical trial (Dardennes et al. 1999). Valproate was more cost-effective than lithium over 1 year from the payer's perspective in a modeling approach (Keck et al. 1996); the prophylactic efficacy of valproate, however, remains to be established.

7. Pharmacoeconomic Findings with Tranquilizers and Hypnotics

There are almost no economic analyses published on tranquilizers or hypnotics. Yuen et al. (1997) found that hospital stay was longer and costs were higher in elderly people if sedative-hypnotic medications were prescribed, even if prescription adhered to established guidelines. Research is needed especially to clarify whether the higher costs of the newer hypnotics (zaleplon, zolpidem, zopiclone) compared with benzodiazepines are economically justified.

8. Pharmacoeconomic Findings with Anticraving Drugs

Long-term acamprosate adjunctive to conventional therapy was found to be cost-effective in the treatment of alcoholism in the German setting in a single, retrospective modeling study (Schädlich and Brecht 1998). Opiate antagonists (Howard et al. 1996), opiate agonist (methadone) substitution, and bupropion have not been evaluated economically.

9. Pharmacoeconomic Findings with Cognitive Enhancers

There are no economic analyses published on nootropics. The acetylcholinesterase inhibitors donepezil (Foster and Plosker 1999) and rivastigmine (Fenn and Gray 1999) seem to be at least cost-neutral, especially by delaying the need for institutionalization. The current evidence is restricted, however, to a few observational studies and modeling analyses using the cost-effectiveness approach from the payer's perspective.

tive. Economic analyses on galanthamine, an acetylcholinesterase inhibitor and nicotinic agonist expected to be approved, remain to be published.

10. Conclusions and Some Precautions

The innovative neuroleptics and antidepressants not only have the potential for a balance of costs despite higher costs of medication, but might even save direct costs essentially by reducing the risk of institutionalization owing to lower drop-out rates in comparison with traditional compounds. This conclusion, however, is based essentially on decision analyses, i.e., modeling. More research, particularly large-scale, randomized trials under naturalistic conditions, is needed. From a systemic perspective, the potential cost savings can only be realized if the institutions concerned, e.g., hospitals, reduce their capacities accordingly. About 80 percent of the costs of hospitals are common costs, mostly wages and salaries. Whereas in a free health-care market the savings might be realized by market forces, in systems bound in economic welfare a political will is needed to enforce the inevitable dismissal of staff.

See also: Alzheimer's Disease: Antidementia Drugs; Antidepressant Drugs; Assessment and Application of Therapeutic Effectiveness, Ethical Implications of; Health Economics; Placebo Studies (Double-blind Studies); Prophylaxis in Psychiatry; Psychiatric Care, Economics of; Psychological Treatment, Effectiveness of; Psychological Treatments, Empirically Supported; Psychopharmacotherapy: Side Effects; Psychotherapy and Pharmacotherapy, Combined

Bibliography

- Ahrens B 1997 Mortality studies and the effectiveness of drugs in long-term treatment. *Pharmacopsychiatry* **30** (Suppl. 1): 57–61
- Antonuccio D O, Thomas M, Danton W G 1997 A cost-effectiveness analysis of cognitive behavior therapy and fluoxetine (Prozac) in the treatment of depression. *Behavior Therapy* **28**: 187–210
- Bacaltchuk J, Trefiglio R P, Oliveira I R, Hay P, Lima M S, Mari J J 2000 Combination of antidepressants and psychological treatments for bulimia nervosa: a systematic review. *Acta Psychiatrica Scandinavica* **101**: 256–64
- CCOHTA 1994 *Canadian Coordinating Office for Health Technology Assessment 1994. Guidelines for the Economic Evaluation of Pharmaceuticals: Canada*. Canadian Coordinating Office for Health Technology Assessment (CCOHTA), Ottawa, ON
- CCOHTA 1997. *Canadian Coordinating Office for Health Technology Assessment 1997. Selective Serotonin Reuptake Inhibitors (SSRIs) for Major Depression. Part II. The Cost-effectiveness of SSRIs in Treatment of Depression: Canada*. Canadian Coordinating Office for Health Technology Assessment (CCOHTA), Ottawa, ON
- Casciano R, Arikian S R, Tarride J-E, Casciano J, Doyle J J 2000 Antidepressant selection for major depressive disorder: the budgetary impact on managed care. *Drug Benefit Trends* **12**: 6–16
- Conner T M, Crismon M L, Still D J 1999 A critical review of selected pharmacoeconomic analyses of antidepressant therapy. *Annals of Pharmacotherapy* **33**: 364–72
- Dardennes R, Lafuma A, Watkins S 1999 Prophylactic treatment of mood disorders: a cost-effectiveness analysis comparing lithium and carbamazepine. *Encephale* **25**: 391–400
- Davidson J R T, Meltzer-Brody S E, Goldberg D, Wittchen H-U, Magruder K M, Lecrubier Y, Ballenger J C, Tylee A, Schulberg H C, Nutt D J 1999 The underrecognition and undertreatment of depression: what is the breadth and depth of the problem? Discussion. *Journal of Clinical Psychiatry* **60** (Suppl. 7): 10–11
- Fenn P, Gray A 1999 Estimating long term cost savings from treatment of Alzheimer's disease: a modelling approach. *Pharmacoeconomics* **16**: 165–74
- Foster R H, Plosker G L 1999 Donepezil: pharmacoeconomic implications of therapy. *Pharmacoeconomics* **16**: 99–114
- Friedberg M, Saffran B, Stinson T J, Nelson W, Bennett C L 1999 Evaluation of conflict of interest in economic analyses of new drugs used in oncology. *Journal of the American Medical Association* **282**: 1453–7
- Gabbard G O, Lazar S G, Hornberger J, Spiegel D 1997 The economic impact of psychotherapy: a review. *American Journal of Psychiatry* **154**: 147–55
- Gold M R, Russell L B, Siegel J E, Weinstein M C (eds.) 1996 *Cost-effectiveness in Health and Medicine*. Oxford University Press, Oxford, UK
- Gould R A, Otto M W, Pollack M H 1995 A meta-analysis of treatment outcome for panic disorder. *Clinical Psychological Review* **15**: 819–44
- Greenberg P E, Sisitsky T, Kessler R C, Finkelstein S N, Berndt E R, Davidson J R T, Ballenger J C, Fyer A J 1999 The economic burden of anxiety disorders in the 1990s. *Journal of Clinical Psychiatry* **60**: 427–35
- Greenberg P E, Stiglin L E, Finkelstein S N, Berndt E R 1993 The economic burden of depression in 1990. *Journal of Clinical Psychiatry* **54**: 405–18
- Howard M O, McGuffin R W, Saxon A J, Sloan K L, Walker R D 1996 Clinical issues related to the costs of alcoholism. *Pharmacoeconomics* **9**: 134–45
- Keck P E Jr, Nabulsi A A, Taylor J L, Henke C J, Chmiel J J, Stanton S P, Bennett J A 1996 A pharmacoeconomic model of divalproex vs. lithium in the acute and prophylactic treatment of bipolar I disorder. *Journal of Clinical Psychiatry* **57**: 213–22
- Lave J R, Frank R G, Schulberg H C, Kamlet M S 1998 Cost-effectiveness of treatments for major depression in primary care practice. *Archives of General Psychiatry* **55**: 645–51
- Manning W, Wells K B 1995 How can care for depression become more cost-effective? *Journal of the American Medical Association* **274**: 1931–4
- Murray C J L, Lopez A D 1997 Alternative projections of mortality and disability by cause 1990–2020: global burden of disease study. *Lancet* **349**: 1498–504
- Revicki D A 1999 Pharmacoeconomic studies of atypical antipsychotic drugs for the treatment of schizophrenia. *Schizophrenia Research* **35** (Suppl.): S101–9
- Schädlich P K, Brecht J G 1998 The cost effectiveness of acamprosate in the treatment of alcoholism in Germany:

- economic evaluation of the prevention of relapse with acamprosate in the management of alcoholism (PRAMA) study. *Pharmacoeconomics* **13**: 719–30
- Tylee A, Gastpar M, Lepine J-P, Mendlewicz J 1999 DEPRES II (Depression Research in European Society II): a patient survey of the symptoms, disability and current management of depression in the community. *International Clinical Psychopharmacology* **14**: 139–51
- Whetten-Goldstein K, Sloan F A, Goldstein L B, Kulas E D 1998 A comprehensive assessment of the cost of multiple sclerosis in the United States. *Multiple Sclerosis* **4**: 419–25
- Yuen E J, Zisselman M H, Louis D Z, Rovner B W 1997 Sedative-hypnotic use by the elderly: effects on hospital length of stay and costs. *Journal of Mental Health Administration* **24**: 90–7

J. Fritze

Psychopharmacotherapy: Side Effects

Psychopharmacotherapy refers to the treatment implications of the science of psychopharmacology. Psychopharmacology is the study of drug-induced alterations in the cognitive function and behavior of humans that aims to develop treatment strategies for neurological and psychiatric disorders. Rapid technological progress has led to an abundance of scientific techniques in neurochemical and animal studies on the interaction between drug action and behavior (Feldman et al. 1997).

Most psychoactive drugs exert their effects first by interfering with the communication between nerve cells (neurons). At the molecular level this occurs by their interaction with receptors of neurotransmitters located on the surface or inside the neuron. This step initiates a cascade of events that eventually lead to alterations in the neurotransmitter and/or hormonal systems. Nonspecific drug action at any level in this process results in side effects.

The term side effect implies an undesirable (adverse) drug effect; a particular side effect of a drug for a given condition may be a desirable one for another condition. For instance, benzodiazepines may be prescribed for anxiety disorders, but as a side effect they cause sedation. On the other hand, they may also be prescribed for the treatment of insomnia as a primary indication. Indeed, the psychoactive potential of several drugs was discovered as a result of unexpected but beneficial side effects on psychiatric patients who were taking these drugs for other therapies (see below). The spectrum of drug effects originate largely from their effects on multiple neurotransmitter systems, and the interactions of these systems with each other and other sites of action in the central nervous system.

At times it is a challenge in the practice of psychiatry to be able to determine if a given behavior is a

symptom of an illness or a side effect of drug treatment, as the latter may mimic the former. For instance, untreated patients with schizophrenia frequently display decreased spontaneity, lack of motivation, and social isolation. These symptoms resemble some of the extrapyramidal system side effects induced by conventional antipsychotics. A practical management of this potential difficulty involves careful assessments of the patient's functions e.g., sleep patterns, daily living routines, and sexual function prior to treatment.

1. Historical Perspective

Treatment with substances that alter mood, thinking, and cognition date back to prehistoric times (Spiegel 1996). Various psychoactive substances have been used in religious ceremonies, rituals, and for healing of pain.

Alcohol is the oldest drug used for its calming and sedative effects. The disinhibition caused by alcohol is described in the first book of the Old Testament. Apart from consumption in social settings, alcohol was introduced for medical use in the nineteenth century as a general anesthetic. The narrow therapeutic index (the ratio of toxic over desired dose) of alcohol limited its usefulness for this purpose.

Opium (*papaver somniferum*), hellebore, and *rauwolfia* roots are among the oldest compounds that were extracted from plants in prehistoric and ancient times. Morphine was isolated from opium in the beginning of the nineteenth century and has been available as an analgesic with considerable addiction potential.

In ancient times, different forms of hellebore root were used as emetics, laxatives, and for mental conditions such as mania and melancholy. Risk of seizures limited its use.

Rauwolfia serpentina roots were used by Indians to treat snake bites and also in some mental conditions. Reserpine, a *rauwolfia* alkaloid was noted to have a calming effect in patients with schizophrenia. Because reserpine induces severe depression as a side effect, its use as an antipsychotic has been abandoned.

Cocaine is an alkaloid derived from the shrub *Erythroxylon coca*; it originated in South America where it was used for its stimulating effects. It was isolated in the nineteenth century and is still used as a local anesthetic. Paranoid delusions, hallucinations, seizures, and dependency are some of the adverse effects of cocaine.

The discovery of psychoactive medications has resulted from the side effect profiles of drugs used in medicine. For instance, the antipsychotic chlorpromazine was initially used to help patients with presurgical preparation and as an adjuvant to anesthetics. Its sedative and calming effects easily found its way to treatment in psychiatric disorders as a primary

indication soon after its antipsychotic properties were discovered. Similarly, iproniazid (an antidepressant of the monoamine oxidase inhibitor (MAOI) type) was originally developed as a drug for the treatment of tuberculosis. In clinical drug trials, patients treated with this drug were observed to become stimulated and cheerful. This observation led to the investigation of the antidepressant effect of iproniazid.

2. Genetic Factors and Side Effect Vulnerability

Side effects are not uniformly observed in everyone. While some patients never develop a certain side effect, others are affected from day one of treatment. The strongest determinant of this interindividual variation is believed to be in the genetic complexion. Variations in the sequence of the DNA called polymorphisms are believed to play a role in expression of a wide spectrum of reactions to drugs: from full response to non-response; minimal to severe side effects. The completion of the Human Genome Project will aid in the identification of patient subgroups according to their drug response patterns. For instance, certain variations in the gene that codes for the enzyme CYP2D6 lead to slow metabolism of some drugs such as tricyclic antidepressants, therefore these patients run higher levels of drugs and are more vulnerable to side effects (Schatzberg and Nemeroff 1998). The application of pharmacogenetics will be the identification of drugs that will be both efficacious and well tolerated for each individual according to his or her genetic makeup.

3. Assessment and Management of Side Effects

Compared to the rating scales used to assess the symptoms of diseases, there are relatively few rating scales for the assessment of side effects. The most comprehensive of these is the Systematic Assessment for Treatment Emergent Events which has been designed for use in clinical drug trials. Udvalg for Kliniske Undersøgelser Side Effect Rating Scale is also comprehensive and was developed for use in clinical trials and in routine practice. The Rating Scale for Extrapyramidal Side Effects (Simpson–Angus Extrapyramidal System Scale), Barnes Akathisia Rating Scale, and Abnormal Involuntary Movement Scale were developed to measure the extrapyramidal system side effects, akathisia, and dyskinetic movements, respectively, that are associated with antipsychotic drug use (American Psychiatric Association 2000). Even though each side effect ultimately requires individual management, a comprehensive approach is vital for general management. Such an approach involves a baseline assessment, careful choice of medication, patient and family education, minimization of the number of drugs (to decrease potential side effects and to avoid drug interactions), single

bedtime dosing, the availability of clinicians to discuss the side effects, adjustment in dosage, counteractive drug treatment, and continuous assessment.

4. Noncompliance

Among factors that play a role in treatment non-compliance, such as denial of the illness and lack of supervision, one of the strongest predictors is early onset and persistence of side effects. Noncompliance by itself is a predictor of relapse in patients with chronic mental disorders who are recommended to receive lifelong maintenance treatment. One of the studies in this area found that half of the schizophrenic patients took less medication than recommended, and the reluctance was associated with the side effects of antipsychotic drugs (Van Putten 1974). On the positive side, many studies indicate improvement in treatment compliance with patient education. Particular approaches in patient education have been proposed that maximize the information conveyed while avoiding suggestibility (Balon 1999).

5. Classification of Side Effects

One of the common characteristics of psychoactive drugs is their capability to cross the blood-brain barrier and reach the central nervous system. Consequently, side effects of psychoactive drugs may be generated both centrally and peripherally. Here, the distinction between drug action and drug effect is also relevant: drug action refers to alterations produced by the drug at the molecular level, whereas drug effect (as well as side effect) refers to the observed reactions of the organism. Below, we will examine the behavioral side effects of psychoactive medications; a review of physiological systems is beyond the scope of this chapter and can be found elsewhere (Schatzberg and Nemeroff 1998).

5.1 Autonomic

The autonomic nervous system controls homeostasis in the different functional systems in the body mostly through involuntarily mechanisms. Major neurotransmitter systems involve those of acetylcholine and norepinephrine (β or α receptors).

Autonomic side effects of psychoactive drugs are most commonly due to *anticholinergic effects*. Because acetylcholine is widely distributed in the central and peripheral nervous systems, adverse anticholinergic effects can be observed both behaviorally and systemically.

Most antiparkinsonian (benztropine, trihexiphenidyl) and the majority of the antidepressant and antipsychotic drugs produce anticholinergic effects.

Among the tricyclic antidepressants, tertiary amine tricyclics (amitriptyline, clomipramine, doxepin, imipramine, trimipramine) produce more anticholinergic side effects compared to secondary amine tricyclics (desipramine, nortriptyline, protriptyline) and tetracyclic antidepressants (amoxapine, maprotiline). Compared to tricyclic antidepressants, anticholinergic side effects with selective serotonin reuptake inhibitors (SSRIs) are minimal. Conventional low-potency antipsychotics (drugs in this category, such as chlorpromazine, thioridazine, and mezoridazine require higher doses to produce equal efficacy than the high-potency antipsychotics) have significant anticholinergic properties compared to high-potency antipsychotics (fluphenazine, haloperidol, perphenazine, pimozide, thiotixene, trifluoperazine). Among the new generation antipsychotics, clozapine has the most potent anticholinergic effects, as well as having properties of acetylcholine release. Olanzapine also possesses anticholinergic effects, but less than clozapine. Risperidone, quetiapine, and ziprasidone have minimal anticholinergic activity. Peripheral anticholinergic effects include warm skin, flushing, dryness, dilated pupils (and blurred vision), increased intraocular pressure, dry mouth (if unrecognized this can cause dental problems, such as caries), tachycardia (increased heart rate), reduced acid secretion in the stomach, reduced pulmonary secretions, diminished bowel motility (constipation), urinary retention, and delayed or retrograde ejaculation.

Central anticholinergic effects of psychoactive drugs involve features such as memory impairment, confusion, disorientation, delirium, visual hallucinations, and worsening of preexisting psychotic symptoms. Combination of psychoactive drugs with anticholinergic effects, such as a tricyclic antidepressant, and a low-potency antipsychotic will highly increase the likelihood of an adverse reaction. Geriatric patients are particularly sensitive to anticholinergic side effects and deserve special consideration in the choice of psychoactive drugs and evaluation of a deteriorating course of their clinical situation (see *Antidepressant Drugs*).

5.2 Dependence

Dependence is an important component of the side effect spectrum of some of the psychoactive drugs (barbiturates, benzodiazepines). The *Diagnostic and Statistical Manual of Mental Disorders, Fourth Edition* defines substance dependence as a maladaptive pattern of substance use, leading to clinically significant impairment or distress, as manifested by three or more of the following criteria: tolerance, withdrawal, consumption of large amounts or over a longer than intended time period, persistent desire to cut down, significant effort to obtain the substance, social,

occupational sacrifice, and continued usage despite known hazards. It is of note that in this definition a biopsychosocial approach is undertaken. The treatment models for dependency are conducted in a multidisciplinary fashion that address physical, psychological, and environmental factors. Chronic drug use involves tolerance, physical dependence, and psychological dependence. Even though related, tolerance and physical dependence are different concepts:

Tolerance refers to the need of increased amount of a drug with repeated usage in order to achieve the same magnitude of effect. Interestingly, not all of the behavioral effects of drugs show similar tolerance patterns; in the case of benzodiazepines, for instance, tolerance develops for the sedative effects, but not for the anxiolytic effects. Tolerance is a reversible condition; abstinence from the drug over a period of time will reduce the amount of drug needed to exert the particular effects. Cross-tolerance is a phenomenon where the tolerance for a particular drug effect operates for another drug effect as well. It is demonstrated by the diminished anticonvulsant effects of the barbiturates in people with chronic alcohol use. Both changes in cellular physiology and behavioral conditioning systems have been postulated as underlying mechanisms for the development of tolerance.

Physical dependence refers to the condition where withdrawal symptoms emerge upon discontinuation of the drug. Withdrawal symptoms are specific to different drugs and quite unpleasant, so that continued use is sought to avoid them. Most withdrawal symptoms are separate from the ones in the intoxication state and can be identified without difficulty.

5.3 Endocrinological and Metabolic

Most antipsychotics elevate serum prolactin secondary to the blockade of dopamine. Increased prolactin may cause menstrual irregularities in females, gynecomastia (abnormal enlargement of the breast) in males, and galactorrhea (spontaneous milk flow) in both females and males. Amenorrhea (complete cessation of menses) occurs frequently in women taking antipsychotics. These changes persist with treatment and resolve with discontinuation. Galactorrhea and amenorrhea may also be observed with tricyclic antidepressants.

Hypothyroidism (decreased function of the thyroid gland) may be associated with long-term lithium therapy. The elderly and females are at greater risk for lithium-induced hypothyroidism. Weight gain is a common side effect of antipsychotics, particularly of new generations (clozapine, olanzapine, risperidone, sertindole, quetiapine), cyclic antidepressants, MAOIs, and SSRIs. Impact of weight gain on general health in the long run remains to be determined. In contrast, stimulants (amphetamine, dextroamphet-

amine, methylphenidate), and certain SSRIs are associated with appetite suppression and weight reduction. Stimulants are also associated with adverse growth effects (height and weight suppression); however, these effects are compensated with drug holidays.

5.4 Insomnia

Insomnia is the most common transient side effect of stimulants and MAOIs.

5.5 Neurological/behavioral

5.5.1 Cognitive effects. Even though the motor and cognitive functions of the brain do not operate independently from each other, specific neuropsychological tests that capture primarily motor or cognitive functions have been applied to measure the effects of psychoactive drugs in these domains. While some studies of conventional antipsychotics showed worsening performance on psychomotor speed, attention, and working memory, recent publications do not support a significant worsening in working memory, but even indicate improvements in some attentional measures (Goldberg and Weinberger 1996). The only double-blind study comparing the new generation antipsychotic clozapine and the conventional antipsychotic haloperidol revealed a significant interaction on measures of executive/attention and visuospatial functions. Further analysis revealed that these effects resulted from worsening performance of the haloperidol treated patients. Patients treated with clozapine showed an insignificant improvement (Buchanan et al. 1994). Anticholinergic doses have been associated with poor verbal learning, verbal fluency, and motor speed. Lithium may also interfere with memory and motor speed functions. Benzodiazepines may cause slowed cognitive function, impaired memory, and behavioral disinhibition.

5.5.2 Motor effects. Conventional antipsychotic drugs (haloperidol, fluphenazine), and to a lesser extent, new generation antipsychotic drugs (risperidone, olanzapine), some of the selective serotonin reuptake inhibitors (sertraline, fluoxetine), and cyclic antidepressants (amoxapine) affect motor functions of the extrapyramidal system. Acute conventional antipsychotic treatment interferes with simple repetitive motor tasks such as finger tapping. These effects are related to the drugs' dopamine blocking properties in the basal ganglia. Lithium and valproic acid may cause hand tremors. Stimulant use may be associated with development or worsening of tics.

Acute motor effects of the conventional antipsychotics include parkinsonism, dystonic reactions, and akathisia. The major chronic effect of antipsychotic use is tardive dyskinesia. *Parkinsonism* is a word derived from Parkinson's Disease, which is characterized by tremor, rigidity, and bradykinesia (slowed movement). All three features are observed in susceptible patients in the initial phases of treatment. Tremor with a regular rate may be observed in the hands and feet, and is present at rest. Tremor around the mouth also occurs. Akinesia (absence of movement) of the facial muscles sometimes lead to lack of expression which is known as masked face. Patients walk slowly, sometimes with shuffling, short steps, and the normal arm swing is diminished. Rigidity of muscles can be felt on examination; cogwheeling (increased tone in all directions of movement and superimposed ratchet-like resistance) may be elicited. Handwriting is characteristically small and tremulous.

One of the most disturbing side effects of antipsychotic medications (mostly of the high-potency conventional ones and some of the new generation antipsychotics) is akathisia, a subjective experience of motor restlessness. Patients complain of not being able to sit still, they pace in hallways, constantly move their legs to help the discomfort. They describe it as '... I want to run a hundred miles ...' Significant akathisia may lead to insomnia because of having trouble lying in bed. It may be accompanied by a feeling of anxiety and agitation. Restlessness and agitation are frequently observed as part of psychosis in the absence of any drug treatment. Drug-induced akathisia therefore is very important to diagnose differentially for appropriate treatment. If it is mistaken as psychosis-related agitation, increased antipsychotic drug dose will only make it worse. Akathisia has been identified as a predictor of antipsychotic noncompliance. Several treatment options are available.

Acute dystonic reaction is another side effect of the conventional antipsychotics that requires immediate treatment. Dystonic reactions consist of abruptly developing muscle contractions which frequently involve the muscles of the face, tongue (protrusions, twisting), neck (spasmodic torticollis), jaw (trismus), eyes (oculogyric crisis, blephorospasm), and limbs. It is most common in young men in the initial phases of treatment. It is treatable with anticholinergic and antihistamine drugs. Prophylactic treatment with these drugs may increase the chances of compliance.

Tardive dyskinesia is a long-term side effect of conventional antipsychotics characterized by involuntary choreiform (random, brisk) or athetoid (slow, regular continuing twisting) movements of the head, face, tongue, limbs, and the trunk. These movements may range from mild and unrecognized to severe and incapacitating. Movements around the mouth are the most common. Dyskinesia may be exacerbated with stress and disappears with sleep. About 10–20 percent of patients treated with conventional antipsychotics

for more than a year develop tardive dyskinesia. Children, women, the elderly, patients with brain damage or mood disorders constitute high-risk groups. While tardive dyskinesia is associated with antipsychotic use, 1–5 percent of patients with schizophrenia who lived before the antipsychotic era have been reported to exhibit similar dyskinetic movements. Thus, tardive dyskinesia may be related to the pathophysiology of schizophrenia. The new generation antipsychotics (clozapine, olanzapine, risperidone, quetiapine) have significantly lower incidence of extrapyramidal system side effects and tardive dyskinesia compared to the conventional antipsychotics. This is a feature of the new generation antipsychotics that distinguish them from the conventional antipsychotics. Numerous animal studies have demonstrated that clozapine fails to induce extrapyramidal system related movements, and it is the only antipsychotic that might have a potential to prevent tardive dyskinesia from deteriorating (see *Schizophrenia, Treatment of*).

5.5.3 Sedation. Sedation is a common side effect with most antipsychotics, mood stabilizers, benzodiazepines, barbiturates, and antidepressants (tertiary amine cyclic antidepressants, MAOIs, trazodone, sometimes with other SSRIs, venlafaxine, and nefazodone). In particular clinical situations, such as severe agitation, sedation may be desirable, whereas in others such as a negative symptoms of schizophrenia (affective flattening, lack of spontaneous speech, avolition), or depression with hypersomnia, it would not be. The choice of drug therefore largely depends on the clinical necessity. In undesirable conditions (night time dosing, reduction in dose) switching to a less sedating medication is considered.

5.5.4 Seizure. Antipsychotic drugs, particularly the ones with low potency reduce the seizure threshold. In a patient with no history of seizure disorder, incidence is low and no prophylactic treatment is warranted. Patients with a history of seizure disorder deserve special consideration. Clozapine increases risk of seizures in a dose-related fashion. Anticonvulsants such as valproic acid can benefit the patient to continue treatment with clozapine. Among the antidepressants, clomipramine and amoxapine may lower the seizure threshold while incidence of seizures with SSRIs is low, but present. Tricyclic antidepressants, MAOIs, anticonvulsants, lithium, and antipsychotics may also induce myoclonic (irregular rapid muscle contraction) movements.

5.5.5 Hypomania/mania. Even though it remains controversial, one of the important side effects of antidepressant treatment is the induction of hypomania/

mania. In a patient with no history of mania, undergoing antidepressant treatment, emergence of manic features merits consideration of the possibility of drug-induced mania. In patients with bipolar disorder, on the other hand, the risk of mania induction with sole antidepressant (in the absence of a mood stabilizer) treatment is established.

5.5.6 Delirium. Infrequently, worsening of psychosis, confusion, and agitation may occur with antipsychotic treatment. Most cases are, however, associated with concomitant anticholinergic use.

5.5.7 Neuroleptic malignant syndrome. This is a potentially life-threatening complication of the antipsychotic treatment characterized by muscle rigidity, hyperthermia, mental state changes (stupor, agitation, confusion), and autonomic dysfunction (blood pressure instability, increased pulse, respiration, sweating). The incidence is about 0.1 percent of patients exposed to conventional antipsychotics. High-potency antipsychotics may have a higher risk than low-potency antipsychotics, even though rarely it has also been reported with clozapine. Rhabdomyolysis (muscle necrosis) may occur that leads to myoglobinuria (passing a muscle-related protein—myoglobin—in urine) and subsequent renal failure. Mortality may approach 12 percent. Immediate intervention is necessary.

5.6 Placebo Effects

A placebo is a pharmacologically inactive agent administered in a group of patients in a double-blind clinical trial to improve the true estimate of the active agents efficacy and side effects. Placebo effects may be relevant to the concept of side effects from the perspective of experiences of drug effects in the absence of drug action. Substantial evidence suggests emergence of physiological changes associated with placebo administration. While placebo drug efficacy approaches significant percentage of patients in some clinical trials, incidences of side effects are also variably present (see *Placebo Studies (Double-blind Studies)*).

5.7 Sexual

In males impotence, decreased libido, changes in quality of orgasm, priapism (painful erection), and erectile and ejaculatory disturbances have been reported with antipsychotic use. In women orgasmic dysfunction and decreased libido are frequently observed with antipsychotic treatment. Sexual dysfunction associated with SSRI use impairs quality of life, and is a reason for noncompliance. Bupropion and nefazodone may have a more favorable profile com-

pared to TCAs, MAOIs, and venlafaxine. In contrast, some antidepressants are prescribed for the treatment of premature ejaculation.

6. Quality of Life and Side Effects

The World Health Organization has defined Quality of Life as:

'An individual's perception of their position in life in the context of the culture and value systems in which they live and in relation to their goals, expectations, standards and concerns. It is a broad ranging concept incorporating in a complex way the person's physical health, psychological state, level of independence, social relationships, personal beliefs and their relationship to salient features of the environment' (WHOQOL 1995).

This definition emphasizes a comprehensive approach in the assessment of quality of life. Disciplines of medicine, particularly cancer research have concentrated on global measures of quality of life, as symptom improvement by itself does not account for it. In psychiatry, the slow pace of research in this area may be due to difficulties of valid and reliable quality of life assessments from psychiatric patients. Pioneered by Lehman (Lehman et al. 1982), renewed interest has led to research in debilitating psychiatric disorders such as schizophrenia and demonstrated the significance of a multidisciplinary treatment model. While the strongest predictor of symptom improvement is identified as compliance with recommended drugs, its sole impact on quality of life in the long term is yet to be elucidated. The issue gets complicated due to the side effect profile of psychoactive drugs, which may mimic the symptoms of diseases, and also due to the direct impact of side effects on quality of life. For patients with schizophrenia, symptoms, their severity, and subjective distress caused by side effects accounted for half the variance in quality of life (Awad et al. 1997).

In one study where 53 schizophrenic outpatients treated with depot neuroleptics were studied, 70 percent complained about side effects when 94 percent were found to experience them. Among the patients who denied having side effects, 87 percent had some form of it (Larsen and Gerlach 1996).

The recognition of the negative impact of tardive dyskinesia on, for instance, patients' self esteem, incapacity, and perception by society has been essential towards targeting more favorable treatments. The introduction of new generation psychoactive drugs, particularly antidepressants and antipsychotics is therefore encouraging. Such drugs have not only led to superior efficacy, but also have more favorable side effect profiles. However, there remains significant room for further improvement in side effect profiles of psychoactive drugs.

See also: Alzheimer's Disease: Antidementia Drugs; Antidepressant Drugs; Drugs and Behavior, Psy-

chiatry of; Medicalization: Cultural Concerns; Patient Adherence to Health Care Regimens; Prophylaxis in Psychiatry; Psychopharmacotherapy, Costs of; Psychotherapy and Pharmacotherapy, Combined

Bibliography

- American Psychiatric Association 2000 *Handbook of Psychiatric Measures*. APA, Washington, DC
- Awad A G, Voruganti L N, Heslegrave R J 1997 A conceptual model of quality of life in schizophrenia: Description and preliminary clinical validation. *Quality of Life Research* **6**(1): 21-6
- Balon R 1999 *Practical Management of the Side Effects of Psychotropic Drugs*. Dekker, New York
- Bernstein J G 1995 *Handbook of Drug Therapy in Psychiatry*. Mosby-Year Book, St Louis, MO
- Bloom F E, Kupfer D J (eds.) 1994 *Psychopharmacology the Fourth Generation of Progress*. Raven, New York
- Buchanan R W, Holstein C, Breier A 1994 The comparative efficacy and long-term effect of clozapine treatment on neuropsychological test performance. *Biological Psychiatry* **36**(11): 717-25
- Feldman R S, Meyer J S, Quenzer L F 1997 *Principles of Neuropsychopharmacology*. Sinauer Associates, Sunderland, MA
- Goldberg T E, Weinberger D R 1996 Effects of neuroleptic medications on the cognition of patients with schizophrenia: A review of recent studies. *Journal of Clinical Psychiatry* **57**(Suppl. 9): 62-5
- Kaplan H I, Sadock B J 1999 *Comprehensive Textbook of Psychiatry*. Williams and Wilkins, Baltimore, MD
- Larsen E B, Gerlach J 1996 Subjective experience of treatment, side-effects, mental state and quality of life in chronic schizophrenic out-patients treated with depot neuroleptics. *Acta Psychiatrica Scandinavica* **93**: 381-8
- Lehman A F, Ward N C, Linn L S 1982 Chronic mental patients: The quality of life issue. *American Journal of Psychiatry* **139**(10): 1271-6
- Schatzberg A F, Nemeroff C B 1998 *The American Psychiatric Press Textbook of Psychopharmacology*. American Psychiatric Press, Washington, DC
- Spiegel R 1996 *Psychopharmacology: An Introduction*, 3rd edn. Wiley, Chichester, UK
- Van Putten T 1974 Why do schizophrenic patients refuse to take their drugs? *Archives of General Psychiatry* **31**(1): 67-72
- World Health Organization 1999 *WHOQOL Annotated Bibliography*. WHO, Geneva, Switzerland

N. R. Schooler and H. Gunduz

Psychophysical Theory and Laws, History of

1. Introduction

Psychophysics is a scientific approach to the measurement of mental processes. The empirical laws of psychophysics are based on such measures as the magnitudes of stimuli just noticeably different from

each other, the proportion of errors made when comparing two stimuli, the time required to complete such judgments, and self-reports of the confidence in comparative judgments or the strength of a stimulus. Some laws define relations between physical events and their mental representations such as sensations and feelings. The theories of psychophysics provide explanations of these empirical relations and suggest neural mechanisms that give rise to experimental results.

2. Nineteenth-century Psychophysics

The first law defined the 'just noticeable difference' (jnd) in sensation. A jnd occurred, phenomenally, when an increment in the size of a physical stimulus evoked a sensation of a difference in magnitude. The German physiologist Ernst Heinrich Weber (1834, 1846/1978) discovered, for many sense modalities, the increment in a physical stimulus needed to evoke a jnd. Leipzig physicist Gustav Fechner (1860) quantified Weber's observations and created the First Law of Psychophysics, Weber's Law,

$$\Delta s/S = c \quad (1)$$

where S is the magnitude of the physical stimulus, Δs is the increment, and c is called Weber's Constant. For discriminating lifted weights $c = 0.043$. Thus, an initial weight of 1000 g will require an increment of 43 g (to $S + \Delta s = 1043$ g,) before the new weight will feel just noticeably different from 1000 grams.

The subjective nature of these self-reports led Fechner to wonder whether the person experiencing a just noticeable difference could identify which of the two stimuli, S or $S + \Delta s$, was larger. Fechner discovered that two stimuli may feel just noticeably different, but may be identified correctly in only about 75 percent of the comparisons made. Fechner described these experiments and officially established the field of psychophysics in the two-volume *Elemente der Psychophysik* (1860/1966).

In *Elemente* Vol. 1, Fechner created a theory of how humans comparing two sensory stimuli, such as a pair of lights for brightness, two weights for heaviness, or two lines for length, make errors in their judgments. The theory assumed that the sensory system provided the mind with measurements of the external world. However, the process by which the sensory/nervous system created an internal sensory representation of an external physical stimulus introduced error into the measurement. A fixed weight might at one time feel heavier, or lighter, depending upon the amount of error added to, or subtracted from, the true sensory value for the external stimulus. Extending the theory of error created by Gauss in 1809 for physical measurement devices, Fechner invented a theory of judgment about invisible mental events that allowed

him to measure the size of the error in the nervous system.

Figure 1 illustrates the essential features of Fechner's theory as it applies to the comparison of two just noticeably different weights, W_a and W_b . Along the abscissa are true sensory values, μ_a and μ_b , for the two weights undergoing comparison by lifting each once. Two Gaussian (Normal) probability distributions describe the error of sensory measurement and surround the true sensory values of μ_a and μ_b . Fechner postulated that a (decision) threshold, T , for deciding which weight was the larger existed exactly midway between μ_a and μ_b , that is at

$$T = \mu_a + (\mu_b - \mu_a)/2 \quad (2)$$

A lifted weight sensed to be greater than this threshold was thought to be the heavier weight.

The shaded area of Fig. 1 to the right of T represents the probability of making an error of judgment in deciding that the smaller of two stimuli was in fact the larger. The symmetric nature of Fechner's theory predicts that the probability of judging the larger of two stimuli to be the smaller equals the same probability, and is shown by the shaded area to the left of the threshold, T . The two error probabilities increase as the variability in the sensory system increases. Such an increase in variability causes a greater spread in each Gaussian probability distribution.

Fechner's goal was to measure the invisible variability of the sensory system in units of the physical stimuli. As a measure of variability Fechner chose the value $\sqrt{2}\sigma$, Gauss's measure of the spread of the Gaussian distribution. The value of σ , a parameter of the Gaussian probability distribution, creates a scale of measurement units along the abscissa of Fig. 1. The unknown value for T is located an unknown number of σ units above the true sensory value μ_a . But, once the probability of an error of judgment is known, one can determine from tables (many created by Fechner) the corresponding distance in σ units of the (decision) threshold T above the true sensory value μ_a . For example, if the probability of an error in judging which weight is larger equals 0.25, then the (decision) threshold, T , is 0.674σ units greater than the true sensory value μ_a .

To determine the value of σ , Fechner replaced the unknown values of μ_a and μ_b by the physical values that gave rise to them, W_a and W_b . The distance from μ_a to T is defined as $t = 0.675\sigma$. That is,

$$t = 0.675\sigma = (W_b - W_a)/2, \quad \text{or} \quad \sigma = (W_b - W_a)/(2 \cdot 0.675) = (W_b - W_a)/1.35 \quad (3)$$

Using the example above, the smaller of two just noticeably different weights equaled 1000 g and the larger 1043 g (and the probability of an error equaled 0.25). Thus the measure of variability is

$$\sigma = (1043 \text{ g} - 1000 \text{ g})/1.35 \approx 32 \text{ g} \quad (4)$$

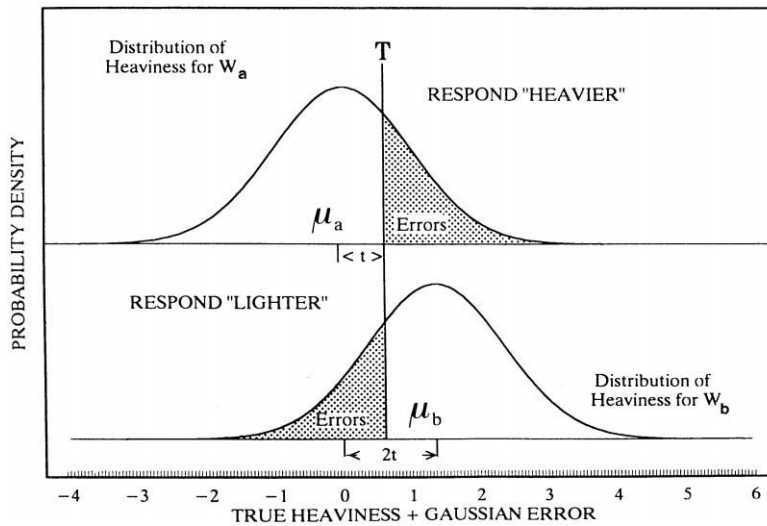


Figure 1

Fechner's theory of mental judgment proposed that Gaussian variability perturbed nervous system transmission of peripheral sensory stimuli. Weight W_a was incorrectly judged heavier than weight W_b whenever the sensory value for W_a exceeded the value T . Similarly, W_b was incorrectly judged lighter than W_a whenever the sensory value for W_a was less than T . The probabilities of an error are represented by the shaded areas to the right and left of T .

Fechner summarized these ideas in a famous formula that is the Second Law of Psychophysics:

$$t = hD/2 \quad (5)$$

Here $h = 1/\sqrt{2}\sigma$ and D equals the difference between two physical stimuli, $W_b - W_a$.

The term h measures the sensitivity of the system. When σ is large, indicating great sensory variability and a small value of h , many errors occur and the system is said to be insensitive. A small value for σ leads to few errors, a larger value for h , and the system is said to be more sensitive. This first method of measuring the invisible, internal, variability of the sensory system established psychophysics as a truly scientific discipline. Measuring the unknown variability of the sensory/nervous system in units of physical stimuli marks a revolution in the scientific investigation of the human mind. This powerful theory became the basis for the fields of experimental psychology and statistical hypothesis testing.

Fechner also theorized, in *Elemente* Vol. 2, about a measure of how much sensation a stimulus generated. Placing each jnd in a sequence is a way of adding together phenomenally equal units of sensation. Using Weber's Law as a base, Fechner argued that on a continuous scale of sensation Weber's Constant,

$$c = \Delta s/S \quad (6)$$

is a differential equation. Integrating the differential equation is equivalent to adding together the jnds of

sensation increments to obtain a total amount of sensation. The result of the integration is called Fechner's Law of Sensation: the total amount of sensation generated by a stimulus is

$$\xi = k \cdot \log_e(S) \quad (7)$$

where S is the magnitude of the stimulus and ξ is the amount of sensation.

Fechner's Law is perhaps the most widely known of any psychophysical law. Yet criticisms of Fechner's idea never abated (Krueger 1989, Lockhead 1992, Murray 1993). The notion of adding together jnds to create a yardstick of sensory measurement is appealing, but the 'law' provides for no independent experimental means for testing its prediction. How is internal sensation to be measured? Or as Laming (1997) wonders, 'Can sensation be measured?'

The father of American Philosophical Pragmatism, C. S. Peirce, put Fechner's ideas to another test. Peirce argued that the sense of confidence in a judgment was a sensation directly related to the probability of making a correct judgment. Peirce's Law states that if C is the average confidence in a judgment and p is the probability of making a correct judgment, then

$$C = k \cdot \log_e[p/(1-p)] \quad (8)$$

Like Fechner's Law there is a logarithmic relation between two variables. But, in this case each variable can be measured through direct experimentation. Peirce and Jastrow (1884, see Link 1992, Vickers 1979), and later experiments, confirmed this prediction.

The profound impact of Fechner's Psychophysics led to Wilhelm Wundt's 1879 creation of the first psychological laboratory at Leipzig University. A new student, James McKeen Cattell, proposed a new measure of the amount of sensation generated by perfectly discriminable stimuli, as opposed to jnd stimuli. Reaction time, RT, the time taken to respond to the presence of a stimulus, could measure stimulus strength. Cattell's experiments showed remarkable decreases in response time as stimulus magnitude increased. Later Piéron (1914, see Welford 1980, Luce 1986) proposed the general principle: Piéron's Law:

$$RT = a + kS^{-\beta} \quad (9)$$

where a is a constant due to non-sensory effects, k and β are constants, and S is physical stimulus intensity.

3. Twentieth-century Psychophysics

Yet, measuring the magnitudes of perfectly discriminable stimuli remained elusive. Thurstone (1955, Bock and Jones 1968) proposed that every stimulus projects a value on an internal psychological continuum. For example, stimuli such as crimes may vary on a continuum of the seriousness of that crime. When asked to judge which of two crimes is more serious, a subject will compare the two crime's seriousness values and choose the crime with the greater seriousness value. This is the famous Law of Comparative Judgment.

Crimes vary in their degree of seriousness for each subject, but for a fixed subject each crime's seriousness is fixed. For a fixed crime, individual differences across subjects created variability in the seriousness of crime. Thurstone represented this variability by a Gaussian distribution of seriousness values. For a group of subjects the theory is formally identical to Fechner's, but the emphasis on measuring stimulus values, rather than variability, led to a powerful new direction in psychophysical research.

Thurstone proved that by comparing all stimuli against each other, each stimulus could be measured on the psychological continuum. Now, mental phenomena as elusive as attitudes could be measured. In an application to classical psychophysics, Guilford (1954) used the method to determine heaviness values for difficult-to-discriminate lifted weights.

Perplexing irregularities in the ordering of probabilities in the full paired comparison matrix were the basis for Clyde H. Coombs to propose an important extension of the Law of Comparative Judgment, the Theory of Ideal Comparisons. Rather than comparing two stimuli against each other, Coombs (1964) proposed that each stimulus was compared against an ideal defined on a psychological continuum. Choosing between two stimuli equally distant from the ideal led to indifference with respect to them and the peculiar response proportions in the paired comparison matrix. For example, a continuum of shades of gray may

contain an ideal gray. But, two shades of gray equally distant on either sides of the ideal would be equally preferred. The subject faced with choosing either of these stimuli would be indifferent to them and chose either with probability 0.50.

In spite of its immense value the Law of Comparative Judgment said nothing about how response time or confidence should vary when stimuli are compared against each other. D. M. Johnson (1939, see Link 1992) combined choice, confidence, and response times (RT) in an experiment on judgments of differences between line lengths and discovered Johnson's Law:

$$RT = a(10)^{b(1-C)} \quad (10)$$

where $0 \leq C \leq 1$ is a measure of confidence and a and b are constants. Johnson's Law, together with Peirce's Law, suggests a relationship between confidence and response probability.

S. S. Stevens (1975) dismissed the theoretical reasoning leading to Fechner's Law and the Law of Comparative Judgment, and proposed an alternative formulation for determining the amount of sensation generated by a single stimulus. For Stevens 'equal stimulus ratios produce equal sensation ratios.' This idea is the basis for the famous Power Law of Sensation,

$$\xi = \lambda S^\gamma \quad (11)$$

where ξ is the strength of response to a stimulus of intensity S , λ is a non-sensory scaling constant and γ measures the effects of peripheral sensory transducers that are unique to a sensory modality. Stevens (1975) determined Power Law exponents for many sensory quantities and gave convincing evidence of the satisfactory representation of stimulus strength provided by the Power Law.

Although the Power Law contradicts Fechner's Law, the Power Law must be consistent with Weber's Law. Link (1992) derived from the Wave Theory of Difference and Similarity the prediction that Stevens' exponents and Weber's Constants are reciprocals of each other. Using values reported by Teghtsoonian (1971, see Link 1992) the prediction held for sensory quantities investigated by Stevens such as shock, heaviness, loudness, finger span, vibration, saturation, taste, and brightness. Ekman's Law provides yet another relation between jnds and the Power Law by asserting that increments in sensation on a psychological continuum are a linear function of sensation magnitude, as if Weber's Law applied to both the psychological continuum and the physical stimulus continuum. Gescheider (1997) shows how Ekman's Law leads to Stevens' Power Law.

Cognitive factors also affect the magnitude of response. Baird (1996) suggests that beyond the sensation effects caused by the peripheral nervous system are cognitive factors that demand a reinterpretation of the Power Law. Proposing a Comple-

mentarity Theory of Psychophysics, Baird accounts for a wide variety of experimental results unaccounted for by the Power Law. In a different approach, Norwich (1993) developed a general theory of the sensory language of the brain. Using the fundamental equation,

$$F = kH \quad (12)$$

where F is a perceptual variable, k is a positive constant and H is a measure of entropy, Norwich derives both the Power Law and Fechner's Law from the same basis.

Fechner's proposal that the (decision) threshold, T , was halfway between the means of the two sensory distributions shown in Fig. 1 proved too restrictive. A major advance in traditional psychophysical modeling occurred when Tanner, Swets, and Birdsall (1954, see Swets 1964) allowed the decision threshold to vary. Calling the decision threshold a response criterion, c , the authors ushered in a new era of psychophysical studies. Experiments showed how a nonsensory component of performance, response bias, affected the position of c . This Theory of Signal Detection and Recognition greatly expanded the applications of psychophysical methods.

For choices between many stimuli Luce (1959, see Falmagne 1985) created the Choice Axiom. Let $S_1, \dots, S_n$ be a set of commensurate stimuli. Luce's Choice Axiom states that the probability of choosing stimulus S_i is

$$W(S_i)/(\sum_n W(S_n)) \quad (13)$$

where W_i is the psychophysical value of stimulus S_i . For example, given five breakfast cereals, $B_1, \dots, B_5$, the probability of choosing cereal B_i is

$$p_i = W(B_i)/(\sum_n W(B_n)) \quad (14)$$

Luce's Choice Axiom receives support from numerous experiments in psychology and economics.

The majority of psychophysical theories and experiments focus on the ability to discriminate between stimuli rather than on the similarity between them. Shepard (1957, see Borg and Lingoes 1987) proposed that similarity between two stimuli, (S_i, S_k) , is a continuously changing negative exponential function of the distance, d_{ik} , between them. The probability of correctly identifying a particular stimulus from a set of n stimuli is often assumed to be

$$p_{ik} = \exp(-d_{ik})/\sum_n \exp(-d_{in}) \quad (15)$$

a form similar to Luce's Choice Axiom.

Following Thurstone's suggestion that there are many underlying psychological continua, Anderson (1981) proposed an Information Integration Theory of how sensory values on different dimensions are combined to create perceptual experience. Anderson's cognitive algebra creates perceptual magnitudes from

a linear model of sensory values defined on the dimensions on which a stimulus projects its values. Subjectively, a fine dinner is a sum of the psychological values of its many components.

4. Conclusion

During nearly 185 years, the field of psychophysics introduced new measurements of, and theories about, mental processes. Following Weber's discovery of the jnd, Fechner created Weber's Law, the field of psychophysics, and a theory of error in sensory comparisons. Thurstone introduced the idea of psychological continua on to which stimuli project psychological values. Comparisons between two psychological values led to the Law of Comparative Judgment and measurements of stimuli not defined on a physical scale. Anderson proposed a cognitive algebra for adding psychological values combined across several such psychological continua.

Fechner's Law, that the amount of sensation equaled a logarithmic transformation of the physical stimulus intensity, provoked continual controversy. Piéron showed that the time required to respond to the presence of a stimulus obeyed a power law of stimulus intensity. Stevens proposed that the amount of sensation obeyed a Power Law with an exponent unique to a sensory modality.

But cognitive factors were also shown to affect judgments of stimuli. Coombs proposed that psychological values were compared against ideals rather than against each other, and suggested a method of locating an ideal. Signal Detection Theory provided for a decision criterion that could vary in response to cognitive factors. Theoretical ideas suggested new experiments. New empirical discoveries challenged old theories. Now, at the beginning of the twenty-first century, psychophysics is poised for major theoretical advances toward more refined and more extensive measurements of mental processes.

See also: Cognitive Psychology: History; Cognitive Psychology: Overview; Fechnerian Psychophysics; Memory Psychophysics; Sensation and Perception: Direct Scaling; Signal Detection Theory, History of; Signal Detection Theory: Multidimensional

Bibliography

- Anderson N H 1981 *Foundations of Information Integration Theory*. Academic Press, New York
- Baird J C 1996 *Sensation And Judgment: Complementarity Theory of Psychophysics*. Mahwah, Hillsdale, NJ
- Bock R D, Jones L V 1968 *The Measurement and Prediction of Judgement and Choice*. Holden-Day, San Francisco
- Borg I, Lingoes J 1987 *Multidimensional Similarity Structure Analysis*. Springer-Verlag, New York
- Coombs C H 1964 *A Theory of Data*. Wiley, New York
- Falmagne J C 1985 *Elements of Psychophysical Theory*. Oxford University Press, New York

- Fechner G T 1966 [1860] In: Boring E G, Howes D H (eds.) *Elements of Psychophysics*. Holt, Rinehart & Winston, New York
- Gescheider G A 1997 *Psychophysics: The Fundamentals*. Erlbaum, Mahwah, NJ
- Guilford J P 1954 *Psychometric Methods*, 2nd edn. McGraw-Hill, New York
- Krueger L E 1989 Reconciling Fechner and Stevens: Toward a unified psychophysical law. *Behavioral and Brain Sciences* **12**: 251–67
- Laming D R J 1997 *The Measurement of Sensation*. Oxford University Press, New York
- Link S W 1992 *The Wave Theory of Difference and Similarity*. Erlbaum, Hillsdale, NJ
- Lockhead G R 1992 Psychophysical scaling: Judgments of attributes or objects? *Behavioral and Brain Sciences* **15**: 543–58
- Luce R D 1986 *Response Times: Their Role in Inferring Elementary Mental Organization*. Oxford University Press, New York
- Murray D J 1993 A perspective for viewing the history of psychophysics. *Behavioral and Brain Sciences* **16**: 115–37
- Norwich K H 1993 *Information, Sensation and Perception*. Academic Press, San Diego, CA
- Stevens S S 1975 In: Stevens G (ed.) *Psychophysics: Introduction to Its Perceptual, Neural, and Social Prospects*. Wiley, New York
- Swets J A (ed.) 1964 *Signal Detection and Recognition by Human Observers: Contemporary Readings*. Wiley, New York
- Thurstone L L 1955 *The Measurement of Values*. University of Chicago Press, Chicago
- Vickers D 1979 *Decision Processes in Visual Perception*. Academic Press, New York
- Weber E H 1978 [1846] Der Tastsinn. In: Ross H E, Murray D J (eds.) *E. H. Weber: The Sense of Touch*. Academic Press, London
- Welford A T (ed.) 1980 *Reaction Times*. Academic Press, New York

S. W. Link

Psychophysics

How might the human mind be studied, scientifically?

It is a commonplace notion that the brightness of a light, the loudness of a sound, or the heaviness of a weight is one thing, and the luminance, the sound pressure level, or the physical weight in gm quite another. If a way could be found to measure brightness, loudness, and heaviness—not the physical attributes of luminance, sound pressure level, and weight, but the subjective experiences they engender—then such measurements might provide the foundation for a scientific study of the mind, analogous to physics.

The name of this enterprise is *psychophysics*. It has led to two quite distinct traditions of experimental research that I call *sensory discrimination* and *sensory judgment*.

1. Sensory Discrimination

Fechner ([1860] 1966) conceived the idea of using the ‘just noticeable difference’ (JND) between two stimulus magnitudes as the unit of subjective measurement. Earlier, Weber ([1834, 1846] 1978) had enunciated the law that now bears his name. If X_1 and X_2 are two sensory magnitudes such that X_2 is just noticeably greater than X_1 , then

$$(X_2 - X_1)/X_1 = \text{constant.} \quad (1)$$

Writing the constant, which is known as the Weber Fraction, as Θ ,

$$X_2 = X_1(1 + \Theta) \quad (2)$$

or

$$\ln X_2 - \ln X_1 = \ln(1 + \Theta), \quad (3)$$

another constant, which Fechner identified as the subjective value of a JND. It follows that the subjective experience of stimulus magnitude increases as $\ln X$, a relation known as Fechner’s Law.

1.1 Psychophysical Methods

The project of measuring subjective sensation by means of Eqn. (3) depends on accurate measurement of JNDs. Fechner directed systematic research towards the development of three methods.

1.1.1 The method of limits. Envisage that X_2 , greater than X_1 , is reduced progressively by small amounts until it is judged ‘equal’ and then by further small amounts until it is ‘less than’, or the process might be reversed with X_2 initially less than X_1 . This procedure identifies ‘thresholds’ above and below X_1 at which the judgment changes.

1.1.2 The method of average error. Envisage that X_2 is adjusted until it is judged ‘equal’ to X_1 , which is fixed. The probable error of adjustment can then be used as an alternative to the upper and lower thresholds identified by the method of limits.

1.1.3 The method of constant stimuli. Envisage that various X_2 s are presented together with a fixed X_1 and the subject asked to say which is greater. It turns out that the probability of a correct judgment increases smoothly with the difference between X_1 and X_2 . That relationship is known as the *psychometric function*.

1.2 Signal-detection Theory

The hundred years following Fechner's seminal work were spent arguing, first, how properly to analyze the judgments from these three psychophysical methods and, second, what the relationship between them was. That argument was resolved only with the advent of signal detection theory, which showed that these were the wrong questions to be asking. To put the matter succinctly, the presentation of a stimulus, be its magnitude X_1 or X_2 , provides the subject with information whether the stimulus magnitude is, in fact, X_1 or X_2 . This information is a random variable with a different distribution depending on which stimulus is presented. The optimum strategy for the subject is to report ' X_2 ' when the information in favour of X_2 exceeds a certain criterion and ' X_1 ' otherwise.

Signal-detection theory brought the idea of hypothesis testing by bisection of the likelihood-ratio (Neyman and Pearson 1933) into psychological theory. It identified two fundamental factors in sensory discrimination; there is (a) the information afforded to the subject (a property of the stimuli to be discriminated), and (b) the decision criterion (which is at the subject's disposal). These two factors are confounded in the traditional psychophysical methods.

It is common to model signal detection in terms of two normal distributions (unit variance, difference of mean d'), but it should be borne in mind that this is a *model* of the information and its function is simply to facilitate calculations. Nevertheless, in that spirit, let the mean of the distribution representing magnitude X be $\ln X$ and let the variance be $[\ln(1 + \Theta)]^2$. If a JND be defined as that stimulus difference which gives $d' = 1$, Eqn. (3) is obtained.

We cannot say, however, that signal-detection theory validates Fechner's Law (Eqn. (3)) because the signal-detection model is no more than a model and any other model delivering the same numerical predictions, more or less, will do as well. Such another model is composed of χ^2 distributions multiplied by scaling factors proportional to the magnitudes of the stimuli. It arises in this manner.

1.3 Differential Coupling

Weber's Law (Eqn. (1)) holds rather accurately for discrimination between two separate stimuli (two flashes of light in darkness, two bursts of noise in silence), and for many stimulus attributes, down to, but not below, absolute threshold. The amplitude of a pure tone (but not Gaussian noise) and the contrast of a sinusoidal grating are known exceptions. But if the task is to detect a flash of light added to a pre-existing background, the relationship is more complicated. For the detection of a sinusoidal grating van Nes and Bouman (1967) demonstrated a square-root relationship up to a transition illuminance, above which

detection conformed to Weber's Law, and the transition illuminance increased as the square of the wave-number of the grating. In addition, the psychometric function for detection is steeper than that for discrimination between two separate stimuli; and the Weber fraction for discrimination increases when one of the stimulus magnitudes lies below the absolute threshold. In short, Weber's Law, on which Fechner's Law is based, is just one component of a large body of related phenomena, a body sufficiently large and complex as to admit substantially only one explanation.

(a) The sensory pathway is differentially coupled to the physical stimulus so that only changes in sensory input are available as a basis for discrimination. When the two stimuli to be distinguished are presented separately, that means intrinsic random variability only. The energy of that intrinsic variability (a sample of Gaussian noise) has a χ^2 distribution multiplied by the noise power density, which, in this case, is proportional to the stimulus magnitude X . The logarithm of a χ^2 variable is, to a very good approximation, normal with variance independent of X . Weber's Law falls directly out of differential coupling.

(b) The steeper gradient of the psychometric function for detection is equivalent to the interposition of a square law transform, which, in turn, is directly analogous to the phenomenon of low-signal suppression in radio reception (Davenport and Root 1958). The stimulus to be detected adds a small perturbation to the Gaussian noise giving a noncentral χ^2 distribution of energy. The resultant central/noncentral χ^2 model accommodates additional phenomena within the general theory and is to be preferred to the logarithmic model underlying Fechner's Law. The distributions in this model are everywhere scaled in proportion to physical stimulus magnitude, so that there is no empirical justification (arising from the study of sensory discrimination) for any measure of sensation distinct from the physical (Laming 1997, Chaps. 3, 4).

(c) The differential coupling is realized in the balanced excitatory/inhibitory receptive fields of sensory neurons. The low-signal suppression is a consequence of transmission by all-or-none action potentials of one polarity only. The theory of sensory discrimination describes the flow of information through the early stages of sensory pathways and annexes that preconscious component of perception to the physical domain. There is no possibility of basing a scientific study of the mind on resolving power.

2. Sensory Judgment

A very wide range of sound energy levels are encountered in nature and sound pressure level is measured on a logarithmic scale; 10 decibels equates to a 10-fold

increase in power. If Fechner's Law were valid, people should say that 100 dB (a subway train entering a station) should sound twice as loud as 50 dB (quiet conversation in a library). Most people would say that the subway train is much more than twice as loud. Around 1930 this posed a problem for acoustic engineers explaining sound levels to their clients and led to the development of the *sones* scale (Stevens 1936). Loudness in sones increases twofold for each 10 dB increase in sound power.

2.1 Stevens' Power Law

Stevens (1957) returned to the problem of subjective measurement in the 1950s with three new experimental methods.

(a) *Magnitude estimation*, in which subjects assigned numbers to stimulus magnitudes;

(b) *Magnitude production*, in which subjects adjusted a variable stimulus magnitude to match a given number; and

(c) *Cross-modality matching*, in which subjects adjusted a stimulus magnitude on one continuum to match a given magnitude on another.

Taking the numbers assigned to the stimuli to be direct measures of subjective sensation, Stevens found that

$$N = aX^\beta, \quad (4)$$

where N is the mean assignment, a a constant, and β an exponent typical of the stimulus attribute. Equation (4) is known as *Stevens' Power Law* and its validity has been demonstrated for more than 30 different attributes.

2.2 Explanation of the Power Law

It is tempting to suppose, as Stevens did, that the sheer volume of experimental demonstration cracks the problem of subjective measurement. But those experiments invariably employed a geometric ladder of stimulus values. If subjects could be induced to assign numbers in like manner (to 'judge ratios'), then a power law relationship is inevitable, simply by regression of log number onto log stimulus magnitude. There are several reasons for supposing that this is all Eqn. (4) tells us.

(a) The power law relation can be perturbed by spacing the stimuli other than geometrically.

(b) The subjects in Stevens' laboratory consistently used numbers extending over a range of about 1 to 30. The estimate of the exponent β can therefore be calculated from the log range of stimulus values employed.

(c) Instructions to subjects are invariably illustrated with numerical examples. A wider spread of numbers in those examples, but examples which conform to the same nominal scale, produces a wider spread of

estimates in the experiment and therefore an increased value of β . Subjects accept guidance about how wide a range of numbers they should be using.

(d) The variance of the numerical assignments in experiments from Stevens's laboratory is about 100-fold greater than the equivalent variability implicit in sensory discrimination.

(e) But that variability is much reduced when two similar stimulus values (± 5 dB) are presented in succession. The correlation between successive log numerical assignments (then about +0.8) implies that each stimulus is used as a point of reference for the succeeding judgment.

So magnitude estimates are very imprecisely determined and are easily influenced by extraneous factors. Good power law data depends on balancing these extraneous influences by appropriate randomization of the schedule of stimulus presentations.

2.3 Relative Judgment

The correlation (e) above supports a theory of relative judgment in which each stimulus is compared to its predecessor. More to the point, that comparison is little better than ordinal; we can say that this sound is louder/softer/about the same as the previous one, but that is all. This idea (Laming 1984) accounts for several otherwise quite unrelated phenomena, including the quantitative variation of the autocorrelation (e) with stimulus differences. By focusing on the precision with which such judgments are made, this idea also relates magnitude estimation to category judgment and absolute identification—the stimuli and the subject are the same, only the response categories are different—and especially to the often repeated finding that the amount of information in a category judgment is limited to 2.3 bits, which is equivalent to the use of no more than five categories without error (Garner 1962, Chap. 3). But if human judgment is ultimately no better than ordinal, it affords no possibility of establishing a metric internal to the mind that is distinct from the physical.

3. How Might the Human Mind be Studied?

There are three ways in which these traditions of research might be continued.

3.1 Information Theory

(Note that I refer here to the mathematical theory of statistical information, not to the information theory (properly *communication theory*) of Shannon (1949), nor to the colloquial use of 'information' which has subsequently become fashionable in psychology.)

Discriminations between two separate stimuli conform accurately to the normal, equal variance, signal detection model (Tanner and Swets 1954). In that

model the parameter d'^2 measures the mean information afforded by a single presentation of either stimulus. A discrimination between two indistinguishable (identical) stimuli provides a natural zero; two looks at the same stimulus (cf. Swets 1964), or two simultaneous, correlated, signals, provide a natural concatenation of separate means; and a comparison between the informativeness of two different discriminations can easily be fashioned from statistical analysis. These are the criteria needed to establish ratio-scale measurement, and it is ratio-scale measurement of fundamental quantities such as mass, length, and time that has enabled the physical sciences to develop their present-day precision.

With that lesson in mind, psychologists have long striven to establish their own scales of measurement, but hitherto have always lacked empirical foundations sufficient to establish a ratio scale (Michell 1999). The parameter d'^2 bids fair to provide the first ratio-scale measure of specifically psychological origin; other ratio scales in psychology have been borrowed from extraneous disciplines. By this means experimental performance may be related to physical parameters of the stimuli (see Laming 1986, 1988).

3.2 The Precision of Sensory Judgment

Magnitude estimation and category judgment may be meaningfully analyzed using a multistimulus elaboration of the normal signal detection model, with a separate normal distribution for each stimulus value (the Law of Categorical Judgment, Case IC; Torgerson 1958). The inverse variance in units of log stimulus magnitude measures the precision of the judgments. That precision varies inversely with the square of the log stimulus range and also depends on the randomness of the stimulus presentation schedule. An 'analysis of variance' of sensory judgment provides a second vehicle of enquiry.

Inverse variance is Fisher's (1922) measure of information and is closely related to d'^2 . So this second suggestion employs the same mathematical theory as the first, but now directed toward the relationships between different experimental paradigms, including the classical psychophysical methods listed above. The lesson to be absorbed is that a magnitude estimate provides some limited information about the subject's assessment of the stimulus, but is never to be taken at its face value.

3.3 Sequential Interactions

It has been known since 1920 that successive psychophysical judgments are statistically related, but there is, as yet, no theoretical understanding what those interactions signify. Aggregate performance in

an experiment results from a pattern of sequential interactions driven by a (usually) random sequence of stimulus values. Trial-by-trial analysis provides a third way to study the mechanics of the human mind.

3.4 Internal Sensation

But none of these approaches measures subjective sensation. Internal experience is private to the subject; it cannot be shared with the experimenter except to the extent that it can be expressed in terms of some feature of the external world perceptible to both. Experimental paradigms are simply schemes for expressing that otherwise private experience. The strictly limited precision of any and every subject's judgments suggests that there is no more precise internal metric to be expressed.

See also: Fechnerian Psychophysics; Memory Psychophysics; Psychophysical Theory and Laws, History of; Signal Detection Theory; Signal Detection Theory, History of; Signal Detection Theory: Multidimensional

Bibliography

- Davenport W B Jr., Root W L 1958 *An Introduction to the Theory of Random Signals and Noise*. McGraw-Hill, New York
- Fechner G T [1860] 1966 *Elements of Psychophysics*. Holt, Rinehart and Winston, New York, Vol. 1
- Fisher R A 1922 On the mathematical foundations of theoretical statistics. *Philosophical Transactions of the Royal Society of London* **222A**: 309–68
- Garner W R 1962 *Uncertainty and Structure as Psychological Concepts*. Wiley, New York
- Laming D 1984 The relativity of 'absolute' judgements. *The British Journal of Mathematical and Statistical Psychology* **37**: 152–83
- Laming D 1986 *Sensory Analysis*. Academic Press, London
- Laming D 1988 Précis of *Sensory Analysis* and A re-examination of *Sensory Analysis*. *Behavioral and Brain Sciences* **11**: 275–96, 316–39
- Laming D 1997 *The Measurement of Sensation*. Oxford University Press, Oxford, UK
- Michell J 1999 *Measurement in Psychology: Critical History of a Methodological Concept*. Cambridge University Press, Cambridge, UK
- Neyman J, Pearson E S 1933 On the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society of London* **231A**: 289–337
- Shannon C E 1949 The mathematical theory of communication. In: Shannon C E, Weaver W (eds.) *The Mathematical Theory of Communication*. University of Illinois Press, Urbana, IL, pp. 1–91
- Stevens S S 1936 A scale for the measurement of a psychological magnitude: Loudness. *Psychological Review* **43**: 405–16
- Stevens S S 1957 On the psychophysical law. *Psychological Review* **64**: 153–81

- Stevens S S 1975 *Psychophysics: Introduction to its Perceptual, Neural, and Social Prospects*. Wiley, New York
- Swets J A (ed.) 1964 *Signal Detection and Recognition by Human Observers: Contemporary Readings*. Wiley, New York
- Tanner W P Jr., Swets J A 1954 A decision-making theory of visual detection. *Psychological Review* **61**: 401–9
- Torgerson W S 1958 *Theory and Methods of Scaling*. Wiley, New York
- van Nes F L, Bouman M A 1967 Spatial modulation transfer in the human eye. *Journal of the Optical Society of America* **57**: 401–6
- Weber E H [1834, 1846] 1978 *The Sense of Touch*. Academic Press, London

D. Laming

Psychophysiology

1. Introduction

1.1 What is Psychophysiology?

Psychophysiology, as the term implies, is the study of the interrelationships between mind and body. Although clearly related to behavioral and cognitive neuroscience, psychophysiology differs from traditional neuroscience both methodologically and conceptually. Methodologically, psychophysiolgists study primarily human subjects rather than lower animals, they usually measure physiological responses non-invasively, and the physiological responses are more molar as opposed to molecular. Conceptually, most psychophysiolgists focus primarily on understanding psychological events and secondarily on understanding the underlying physiological events.

1.2 Goals and History of Psychophysiology

The notion that one can understand psychological states and processes by means of observing or measuring physiological changes dates from the origins of the human race (e.g., blushing indicates embarrassment, sweating is associated with fear, etc.). However, it was not until approximately 1960 that psychophysiology was defined as a specific scientific discipline. The Society for Psychophysiological Research was founded in 1960 and the journal *Psychophysiology* was first published in 1963 under the leadership and vision of Albert Ax. The stated purpose of the society was and continues to be ‘to foster research on the interrelationships between the physiological and psychological aspects of behavior.’ Because only some of the highlights of psychophysiology can be identified here, the reader is referred to an introductory text by

Hugdahl (1995), a more advanced handbook edited by Cacioppo et al. (2000), and the journals *Psychophysiology* and *International Journal of Psychophysiology*.

1.3 Typical Psychophysiological Measures

A great many physiological responses are used as psychophysiological measures. These fall broadly into four categories. The first, and historically oldest, category includes responses of the autonomic (or involuntary) nervous system (ANS), that branch of the nervous system, which controls the actions of internal organs such as the heart, blood vessels, sweat glands, and digestive organs. It is divided into the sympathetic branch, which mobilizes body resources to prepare the organism to respond to the environment and deal with stressors and threat (e.g., by the increasing heart rate, respiration rate, and palmar sweating), and the parasympathetic branch, which acts to store and conserve energy by reducing heart rate and increasing digestive activity. As can be seen, the ANS controls many aspects of physiological function that serve as indicators of emotion or activation. Thus, a psychophysiolgist might use heart rate or palmar sweating as a measure of how anxious a person felt in a particular social situation, such as having to give a speech to an audience.

The second category of psychophysiological measures reflects the activity of the skeletal nervous system, as indicated by motor activity. Of particular interest have been general measures of overall muscle tension, such as tension in the forearm muscles, which can reflect broad states of arousal and activation, and measures of the activity of facial muscle activity, which can indicate states of emotion.

The third category of psychophysiological measures consists of electrophysiological indices of central nervous system activity, primarily electroencephalogram (EEG) and evoked response potential (ERP) measures. For instance, one might use the amount of alpha (8–12 Hz) rhythm in the EEG as an index of relaxation, or the magnitude of a particular ERP component elicited by a stimulus as a measure of attention to that stimulus. The fourth category of measures is derived from brain scanning techniques, such as positron emission tomography (PET), magnetic resonance imaging, MRI, or functional MRI scans. For example, one could use a PET-derived measure of metabolic activity in the frontal lobes to examine differences in executive function between patients with schizophrenia and normal subjects.

The third and fourth categories of measures are the subjects of individual articles in this encyclopedia (see *Event-related/ Evoked Potentials*; *Electroencephalography: Basic Principles and Applications*; *MRI (Magnetic Resonance Imaging) in Psychiatry*). There-

fore, the present discussion is confined to ANS activity and skeletal muscle activity. These measures have been used to investigate long lasting tonic states, transient phasic responses, and individual differences among people.

2. Long-lasting States

2.1 Alertness/Arousal/Activation

The terms alertness, arousal, and activation have a long and useful, albeit a bit checkered, history in psychophysiology. These terms generally denote a continuum of energy mobilization, from the low end of coma or deep sleep to the high end of excitement and extreme agitation. Primary psychophysiological indices of the state of arousal are levels of activity of the ANS, skeletal electromyographic (EMG) tension, and EEG. In keeping with the focus of this chapter, we will limit our discussion to the ANS and EMG measures.

Arousal has been manipulated in psychophysiological laboratories in a variety of ways, including having subjects perform a difficult arithmetic task; be threatened with impending electric shock; blow a balloon to bursting; and hold their hands in extremely cold water. In each of these tasks there is typically a significant increase in activity of the sympathetic division of the ANS (e.g., increased palmar skin conductance and heart rate) and EMG. In more 'real world' situations, arousal has been observed in parachutists just before jumping from an aircraft, and from students just before taking critical exams. Again, there is evidence of increased activity in the sympathetic division of the ANS and EMG.

Some problems with the simple global arousal concept were emphasized by Lacey (1967). Most important for the present discussion, he and his colleagues showed that the various psychophysiological indices of arousal (e.g., skin conductance and heart rate) were often not significantly correlated and, even more embarrassing for a simple arousal concept, these measures would go in opposite directions in many situations. For example, when subjects watched colored lights or listened to tones, skin conductance would increase but heart rate would decrease. Thus, different stimulus situations would reliably produce different patterns of physiological responding ('situational stereotypy' or 'stimulus specificity').

To this day arousal is a useful descriptive concept, but psychophysiolgists remain careful not to over-generalize from one response system to another, or from one stimulus situation to another. There is a far greater degree of specificity of response systems and situational reactions than was recognized by the early global arousal/activation concepts.

2.2 Emotion

The study of emotion has also had a long history in psychology and psychophysiology. Advances with both EMG and ANS measures since the 1960s have shed new light on our understanding of emotions.

Movement of facial muscles is a common way of expressing and communicating emotion and therefore facial EMG offers one measure of emotions. Even small covert changes in facial expression can be detected as small electrical changes with surface EMG recordings from various muscles of the face. Several researchers have found that the muscle that pulls up the corner of the lip as in a smile (*zygomaticus major*) is more active when subjects report positive emotion, whereas the muscle that pushes the eyebrows together as in a frown (*corrugator supercilii*) is more active when subjects report negative emotion. Whether facial EMG can distinguish various specific emotions (e.g., anger, fear, jealousy, disgust, etc.) is more controversial. Even more controversial is whether ANS responses can distinguish between different specific emotions. It is generally agreed that ANS activity occurs during emotional states, but the degree of specificity of that activity to different emotions remains an interesting and unsettled issue.

Another specific EMG measure that has proved useful in the study of emotions is the acoustic startle reflex. The acoustic startle response includes a series of involuntary muscle movements following a loud abrupt noise. One of the fastest and more reliable of these movements in humans is the eyeblink reflex. The eyeblink reflex can be measured easily by placing electrodes on the surface of the skin over the *orbicularis oculi* muscle just below the lower eyelid. For an overview of the methods and many implications associated with this measure, see Dawson et al. (1999). An important characteristic of the eyeblink startle reaction is that its magnitude is enhanced in the presence of unpleasant negative emotions and is diminished in the presence of pleasant positive emotions. This has been demonstrated many times by startling human subjects while they view unpleasant pictures (e.g., mutilated face) or pleasant pictures (attractive nude). It has been shown that changes in the acoustic startle eyeblink reflex can provide valid psychophysiological indices of the emotional state of the individual at the time that the reflex is elicited. A review of this literature and a proposed theoretical account can be found in Bradley et al. (1999).

3. Transient Stimulus-elicited Responses

3.1 Orienting, Defensive, and Startle Responses

Psychophysiolgists who focus on ANS and skeletal motor activity distinguish among three broad types of

transient responses seen in reaction to a wide variety of stimuli: the orienting response, the defensive response, and the startle response (Graham 1992). The orienting response is the response to any novel or significant, but non-aversive, stimulus. It is thought to signal a shift of attention to that stimulus and to facilitate stimulus processing. Its components include a transient increase in palmar skin conductance, heart rate deceleration, constriction of the blood vessels in the skin (usually measured in the fingertips), and dilation of the blood vessels of the head. The defensive response is the response to aversive stimuli, including painful or threatening stimuli. Its components include an increase in palmar skin conductance, an increase in heart rate, and vasoconstriction in both the blood vessels of the skin and the head. The startle response is the reaction to unexpected stimuli that are intense and have a rapid onset. While there are ANS components of the startle response, the component that has been most investigated, because of its reliable occurrence and slow habituation, is the startle eyeblink, as discussed above.

3.2 Classical Conditioning

A paradigm in which stimulus-elicited responses are studied that has been of continuing interest to psychophysiologicals is classical conditioning. In paradigms in which the conditioned stimulus (CS) is several seconds long, with the unconditioned stimulus (UCS) occurring at CS offset, the form of the conditioned response (CR) is generally observed to be that of a heightened orienting response to the CS (Hugdahl 1995). A general question frequently investigated has concerned the relationship between classical conditioning, a very simple form of learning, and the more complex forms of learning of which humans are capable, including verbally mediated processes. For instance, Dawson and his colleagues (Dawson and Schell 1985) studied the relationship between acquisition of the conditioned skin conductance response and the human subject's conscious awareness of the relationship between the CS and the UCS, asking whether a person who is unaware of the CS–UCS contingency will show conditioning. In a series of studies in which awareness of the CS–UCS relation was prevented or delayed by a distracting secondary task, they found that a CR was seen only in subjects who became aware of the CS–UCS relation. Moreover, CRs developed only at or after the point in time in a series of CS–UCS trials when subjects indicated awareness. Thus, these results indicate that conscious relational learning is necessary for human classical conditioning of ANS responses with neutral CSs. On the other hand, the relationship between awareness and conditioning may be different with certain paradigms used to establish the skeletal eyeblink conditioned response (Clark and Squire 1999).

In most psychophysiological studies of classical conditioning, the CSs have been fairly neutral stimuli such as simple tones or colored lights. In contrast, Öhman and his colleagues (see Hugdahl 1995) conducted a series of interesting studies of a different class of CSs, those variously characterized as potentially phobic, fear-relevant, or biologically prepared, such as pictures of spiders, snakes, or angry faces. Öhman and his colleagues have found that CRs (usually the conditioned skin conductance response) conditioned to potentially phobic stimuli (a picture of a snake) are harder to extinguish than are CRs conditioned to neutral stimuli (a picture of a flower). They require more non-reinforced presentations of the CS, and are resistant to instructional manipulation. That is, while responses conditioned to neutral stimuli can usually be abolished by instructing the subject that the UCS will no longer be delivered, so that the subject no longer has any cognitive expectancy that the UCS will follow the CS, responses conditioned to potentially phobic stimuli remain after such instructions, even in the absence of conscious expectancy of the UCS. Thus, psychophysiological techniques can be used in the laboratory to study the dynamics of 'irrational' conditioned responses, such as one may encounter in clinical phobias where the patient fears an object while realizing rationally that the object is harmless.

4. Psychopathology

Psychophysiology is particularly useful in the study of psychopathology because it offers nonverbal measures of psychological states and processes. It thus opens a window onto events that psychiatric patients are unlikely to be able to report verbally. Many psychophysiological measures have been studied with psychopathological patients but space limitation requires that we mention briefly only a few measures in schizophrenia and psychopathy. For a general review of a greater variety of measures and a greater variety of psychopathological conditions, see Keller et al. (2000).

4.1 Schizophrenia

Schizophrenia is a psychotic disorder characterized by hallucinations, delusions, and disorganized thinking and behavior. With ANS measures, especially skin conductance activity, there is usually tonic sympathetic hyper-arousal. On the other hand, with phasic orienting response ANS measures, there is usually hypo-reactivity, at least in sizable sub-groups of patients (for a review see Öhman 1981). Among the many unsettled issues is whether these ANS abnormalities are state-related and/or trait related. That is, if ANS abnormalities are present even before the schizophrenia patients show symptoms, and persist

after the symptoms disappear, they are considered trait measures. However, if the abnormalities are present only when the patients are exhibiting psychotic symptoms then they are considered state related. Although this issue is still unsettled, preliminary evidence suggests that tonic autonomic hyper-arousal is state related, whereas phasic autonomic hypo-reactivity may be trait related, especially when it is expressed relative to their general arousal state (Dawson et al. 1994). These findings suggest that hypo-reactivity may be related to a continuing vulnerability to schizophrenia, whereas the hyper-arousal may be a separate phase associated with the development of symptoms.

4.2 Psychopathy

Psychopathy is a serious personality disorder characterized by antisocial behavior, untruthfulness, irresponsibility, and lack of remorse or empathy. It has been hypothesized that the psychopath is primarily deficient in fear reactivity, and consistent with this hypothesis is the reliable finding that psychopaths show less fear conditioning as indicated by poor skin conductance classical conditioning (Lykken 1957).

Another psychophysiological measure of fear and negative affect mentioned earlier involves the acoustic startle eyeblink reflex. Recall that normal subjects give enhanced eyeblink startle reactions while observing unpleasant pictures, compared to startle while observing neutral or pleasant pictures. This is not true of psychopaths, or at least of significant subgroups of psychopaths (Patrick et al. 1993). This finding also is consistent with the hypothesized deficit in fear in some types of psychopathy.

5. Applied Psychophysiology

In addition to their use in investigating many questions of broad theoretical interest in behavioral and social sciences, psychophysiological techniques have also been used in a number of ways to deal with specific applied problems. The two most often encountered areas of applied psychophysiology are biofeedback and the detection of deception (lie detection).

5.1 Biofeedback

In the early development of learning theory, it was generally believed that while responses of the skeletal motor system could be instrumentally conditioned, responses of the ANS could only be classically conditioned. Because they were not under conscious, voluntary control, ANS responses would not respond to operant reinforcement contingencies. However, beginning in the early 1960s, investigators such as H. Kimmel, B. Engel, N. Miller, and D. Shapiro

began to demonstrate that rewarding organisms, including humans, for specific ANS activity increased the amount of such autonomic activity, and that this could occur whether or not a human subject was aware of the reinforcement contingency. Thus, it was demonstrated that people could learn to increase the size and number of skin conductance responses, or to increase or decrease heart rate or blood pressure, if they were rewarded for doing so by the arrival of pleasant stimuli or the avoidance of unpleasant ones. It was soon discovered that people could learn these things if their only reward was being told that they were acquiring a desired response or performing a task set for them by, for instance, lowering the pitch of a tone that was controlled by their heart rate, or moving the needle on a dial that indicated blood pressure.

These findings were soon applied to treating the symptoms of many psychosomatic disorders. Such disorders typically involve ANS dysfunction (migraine headache, hypertension, Raynaud's syndrome) or loss of control of skeletal muscle tension (tension headache). In clinical biofeedback applications, the patient is given an external stimulus, such as a tone or meter reading, that indicates the level of muscle tension, cerebral vasoconstriction, blood pressure, etc., and the patient's task is to lower the pitch of the tone, move the meter indicator to the left, etc. This approach has been used with varying degrees of success as a therapeutic method in a wide variety of psychosomatic disorders (Carlson et al. 1994), and the training parameters that maximize its effectiveness have been widely studied.

5.2 Detection of Deception

Use of psychophysiological techniques in the detection of deception began in the early part of the twentieth century. There are principally two different techniques in use, each with its strong advocates and critics. Both rely upon the fact that stimuli or information that are of significance or importance to a person will elicit greater ANS activity than will less significant, important stimuli. The Guilty Knowledge Technique, developed by Lykken (see Lykken 1981), relies on there being information known to the guilty person in a crime and to the polygraph examiner, but not to the innocent person. Thus, if a victim were wearing a blue shirt, a guilty person will orient more to the word 'blue' than to other colors when asked the color of the shirt and presented with alternatives, whereas the innocent person will not show differential orienting. The Control Question Technique, first well described by Reid and Inbau (1977) and extensively investigated by Raskin and his colleagues, compares responses to 'critical' questions about a particular crime ('Did you steal that red Porsche last Friday?') with 'control' questions that are presumed to elicit greater arousal among those innocent of that crime, and in fact that are presumed to elicit lying by the innocent, than do the critical questions ('Have you

ever stolen anything of value'?). Laboratory investigations with 'mock crimes' have indicated 90 percent or better accuracy rates with these techniques, with the most common error by far being the false guilty finding. However, a number of investigators have raised questions about the accuracy rate in real-world situations, particularly with respect to whether or not adequate control questions could be found when a person is accused of a serious crime.

6. Conclusion

Psychophysiology offers potentially powerful non-verbal, noninvasive, methodologies with which to study psychological processes in real time. Cognitive processes and emotional processes are studied in both normal and abnormal populations using psychophysiological measures and concepts. Thus psychophysiology lies at the interface of cognitive science, clinical science, and neuroscience and offers the capability to bridge these disciplines.

See also: Anxiety and Anxiety Disorders; Anxiety and Fear, Neural Basis of; Autonomic Classical and Operant Conditioning; Classical Conditioning, Neural Basis of; Cognitive Neuropsychology, Methodology of; Electroencephalography: Basic Principles and Applications; Emotion, Neural Basis of; Event-related/Evoked Potentials; Fear: Potentiation of Startle; Magnetoencephalography

Bibliography

- Bradley M M, Cuthbert B N, Lang P J 1999 Affect and the startle reflex. In: Dawson M E, Schell A M, Böhmelt A H (eds.) *Startle Modification: Implications for Neuroscience, Cognitive Science, and Clinical Science*. Cambridge University Press, Cambridge, UK
- Cacioppo J T, Tassinari L G, Bernston G G 2000 *Handbook of Psychophysiology*. Cambridge University Press, Cambridge, UK
- Carlson J C, Seifert A R, Birbaumer N 1994 *Clinical Applied Psychophysiology*. Plenum, New York
- Clark R E, Squire L R 1999 Human eyeblink classical conditioning: Effects of manipulating awareness of the stimulus contingencies. *Psychological Science* **10**: 14–18
- Dawson M E, Nuechterlein K H, Schell A M, Gitlin M, Ventura J 1994 Autonomic abnormalities in schizophrenia: State or trait indicators? *Archives of General Psychiatry* **51**: 813–24
- Dawson M E, Schell A M 1985 Information processing and human autonomic classical conditioning. In: Ackles P K, Jennings J R, Coles M G H (eds.) *Advances in Psychophysiology*, JAI Press, Greenwich, CT
- Dawson M E, Schell A M, Böhmelt A H 1999 *Startle Modification: Implications for Neuroscience, Cognitive Science, and Clinical Science*. Cambridge University Press, Cambridge, UK
- Graham F K 1992 Attention: The heartbeat, the blink, and the brain. In: Campbell B A, Hayne H, Richardson R (eds.) *Attention and Information Processing in Infants and Adults: Perspectives from Human and Animal Research*, Lawrence Erlbaum, Hillsdale, NJ

- Hugdahl K 1995 *Psychophysiology: The Mind-Body Perspective*. Harvard University Press, Cambridge, MA
- Keller J, Hicks B D, Miller G A (2000) Psychophysiology in the study of psychopathology. In: Cacioppo J T, Tassinari L G, Bernston G G (eds.) *Handbook of Psychophysiology*, Cambridge University Press, New York
- Lacey J I 1967 Somatic response patterning and stress: Some revisions of activation theory. In: Appley M H, Trumbull R (eds.) *Psychological Stress: Issues in Research*, Appleton-Century-Crofts, New York
- Lykken D T 1957 A study of anxiety in the sociopathic personality. *Journal of Abnormal and Social Psychology* **55**: 6–10
- Lykken D T 1981 *A Tremor in the Blood*. McGraw-Hill, New York
- Öhman A 1981 Electrodermal activity and vulnerability to schizophrenia: A review. *Biological Psychology* **12**: 87–145
- Patrick C J, Bradley M M, Lang P J 1993 Emotion in the criminal psychopath: Startle reflex modulation. *Journal of Abnormal Psychology* **102**: 82–92
- Reid J E, Inbau F E 1977 *Truth and Deception: The Polygraph 'Lie-detector' Technique*, 2nd edn. Williams and Wilkins, Baltimore, MD

M. E. Dawson and A. Schell

Psychosomatic Medicine

In its simplest sense, 'psychosomatic medicine' merely indicates that there is a (psyche) mind-body (soma) relationship relevant to human health. Drawing from this, many nonprofessionals believe that it is a medical specialty that concentrates on the psychological influences on disease, particularly certain 'psychosomatic diseases.' Actually, it is a health science that is concerned with the nature of the mind-body relationship, including its clinical applications.

1. Definition

More specifically, psychosomatic medicine is a field of health science based on the concept that the human organism is a *psychosomatic unity*, that biological and psychological processes are inextricably inter-related aspects of its function. This stands in contrast to the still predominant, traditional biomedical model (see Engel 1977) from the nineteenth century which holds that diseases represent cellular or organ pathology arising from intrinsic defects or external bio-physicochemical agents. Psychosomatic medicine conceptualizes diseases as occurring in persons, not their components, as a consequence of their transactions with an environment that is informational, social, and cultural, as well as bio-physicochemical. It notes that people are not simply biological organisms, but sentient ones with motives, thoughts, feelings, and relationships.

Mind-body unity should not be misunderstood to imply that mental and somatic phenomena cannot be

separately studied. Such reductionism is a legitimate tactic that permits the study of scientific questions limited to that system. Each organ or part function merits study by techniques appropriate to it. To understand *organismic* function, however, requires a perspective that includes data from all body systems, including the psychological. Psychosomatic medicine provides no theories about the nature of the underlying processes. What it provides is its holistic perspective for interpreting them. It draws on the data, theory, and techniques of all other relevant fields, integrating these in unique ways informed by its conceptual base. Thus, contrary to common lay belief, psychosomatic medicine is not a medical specialty. Nor is it a scientific discipline. Its practitioners come from a range of disciplines and clinical specialties.

Some have accused psychosomatic medicine of overweighting the importance of psychosocial factors in disease, trivializing biology. This had legitimacy during the 1940s and 1950s when a prominent segment of the field fell into this trap (see Sect 2.3). If this seems still true, that is because its main effort is directed to clarifying the effects of psychosocial factors on the body. But that emphasis is necessitated by the neglect of this work by the remainder of biomedicine. Psychosomatic medicine is concerned with *all* mind-body relationships, giving equal importance to biology.

Doubts have been raised about the aptness of the term. To some, the bipartite nature of 'psychosomatic' undermines the intended unity. But it is one, unhyphenated word. Others question the absence of reference to the social, noting that many relevant factors are such. Engel's (1977) alternative, 'biopsychosocial model' was meant to resolve that, as well as to circumvent the residual misinterpretation that biology is devalued. But if social is included, what about cultural? Indeed, the list could include economic and other factors. Moreover, these factors can impinge on the organism only via the psychological apparatus. 'Psychosomatic medicine' also has the virtue of historical continuity that includes a mass of research documentation.

2. History

2.1 Origins

From earliest times, people have been struck by instances in which physical symptoms followed noteworthy life events, and reports of these appear in ancient writings. With the beginnings of medical science, physicians added more such reports as well as describing illness outbreaks during wars and major upheavals. The idea that stressful life experiences caused illness gained wider currency. During the nineteenth century, philosophical conjectures about the holistic quality of human nature developed. This

was the period when 'psychosomatic' came into use by physicians, although in a sense partly different from our own.

2.2 Early Twentieth Century

With the turn of the century, Cannon's (1915) studies on the physiology of the emotions added key data revealing how these physiological responses were *adaptive* to the external threats that had triggered the emotions. Building on Claude Bernard's earlier work, he indicated (1932) how his findings fit into the process by which organisms maintain their integrity by adapting to environmental change via internal responses to maintain internal equilibrium: 'homeostasis.'

Clinical and epidemiological studies of better quality increased. By 1935, (Helen) Flanders Dunbar was able to cite some 2,251 references in her groundbreaking work, *Emotions and Bodily Disease*, which ushered in the modern, scientific era of psychosomatic medicine.

Further stimulus came from World War II. With astonishing foresight, the (US) National Research Council, anticipating casualties from war stress, convened a series of conferences among psychosomatic experts during the mid-1930s. This kindled additional research which culminated in 1939 with publication of the journal *Psychosomatic Medicine*, still the pre-eminent periodical in the field. Three years later, the American Psychosomatic Society was founded to foster scientific interchange and education. US military physicians observed how war trauma led to physical as well as mental disorders, and many returned with an interest in psychosomatic medicine. Psychosomatic societies arose in other countries and, later, international meetings began.

Psychoanalysis provided further impetus. Freud and his followers were greatly interested in physical symptoms. Much of their theorizing involved a simplistic transfer of the psychoanalytic psychology of hysterical symptoms, without regard to physiology. Nevertheless, as psychoanalytic activity flourished, it stirred interest among intellectuals, adding to the popularity of psychosomatic medicine. With this came the naive belief that many medical disorders were primarily psychogenic.

2.3 Specificity Theories

Paralleling major biomedical advances that elucidated the etiology and pathogenesis of disease, psychosomatic medicine undertook a search for specific psychological causative factors. Dunbar was a pioneer here too, observing that patients with several medical disorders had distinctive personality traits which she concluded were etiological. Others developed similar formulations. The identified diseases were termed

‘psychosomatic.’ This effort peaked with the contributions of Franz Alexander (1950). As a psychoanalyst, he focused on unconscious processes, but his work was far more elaborate and sophisticated than others. Most importantly and enduringly, he clarified that no purely psychological mechanism could explain the physiological processes of internal organs. Instead, Alexander reported that ‘specific dynamic constellations’ of motivations, defenses, and emotions characterized patients with each of a series of ‘psychosomatic’ diseases, and that life experiences which matched the vulnerabilities inherent in the particular constellation precipitated that illness. He also linked each constellation to specific physiological processes mediated through autonomic neural pathways that plausibly explained the pathogenesis of that disease. The brilliance of this work captured the imagination of many, and it became a major force in psychosomatic medicine. Belief developed that psychotherapy might cure ‘psychosomatic diseases.’

2.4 The Last Third of the Twentieth Century

More careful studies called into question much specificity research, which had been uncontrolled for observer bias, and utilized selected subjects. Skepticism grew and spread from specificity to involve all of psychosomatic medicine. Moreover, remarkable advances in biomedicine made other factors seem irrelevant: ‘With penicillin, who needs psychology?’ The influence of the field on the rest of medical science and practice withered.

Experiencing self-doubt and loss of external support, the field contracted and changed. Grand theories were eschewed. Research focused on better-defined questions and became methodologically tighter. Clinical studies became less common, especially those of therapy. Laboratory projects concentrated on elucidating basic psychophysiological relationships, often in normal subjects or even animals. If less exciting, this shift proved salutary, strengthening the scientific base and the quality of data.

Over time, the field recovered. Increasing studies of clinical problems embodied a new methodological rigor gained from models provided by the continuing basic research. A variety of subgroups within the psychosomatic society grew to a critical size and formed associations of their own, expanding work further. Acceptance by clinical biomedicine grew more slowly, however.

3. Theoretical Models of Mind–Body Relationships

The evidence of correlations between psychological and somatic processes is extensive. But how can this be understood? Five explanatory models can be adduced from scientific reports.

3.1 Concomitant Effects

This model rejects causality in the relationship. It postulates that both the psychological and physical findings are products of another, precedent factor. Commonly, this is said to be genetic. Another explanation draws on the fact that life events often have multiple stimulus properties. Thus, one stimulus is said to cause the psychological findings while another produces the physical one.

3.2 Somato → Psychic Causation

This model posits that the relationship follows entirely from the effects of somatic processes on the mind. This is the traditional biomedical view, which sees all disease as ‘physical’ in nature and origin. In a modification, a feedback loop is added going back from the mind to the body, acknowledging that psychological changes may have physical effects, but only secondarily. (If this is true, why cannot they do so primarily?)

3.3 Psycho → Somatic Causation

This is the counterpart of the preceding. It states that psychological responses to external events cause the somatic changes. Most commonly, stress or strong emotions are invoked as intervening mechanisms. Weiner (1977) noted that it may be the behavioral concomitants of the psychological processes which act on the body: ‘lifestyle factors’ such as smoking or substance abuse. Except for ‘specificity’ theories already described, the ‘choice’ of the resultant disease is usually left to other, biological factors. Again, secondary feedback is often added, bodily changes acting back on the mind.

3.4 Bidirectional Psycho ⇌ Somatic Causality

This is a combination of the last two models, allowing for causation in both directions and the feedback variations of each. Thus, longer causal chains are possible: ... psycho ... somato ... psycho ... somato ... (etc.); but these remain linear.

3.5 The Modern Psychosomatic Model

This is a holistic model that includes all the preceding models and their variants, indicating that each explains some psychosomatic processes. Beyond this, it suggests that several are usually operating simultaneously, and to this adds that there are usually *webs of relationships* involving multiple stimulus–response linkages. To explain this, it invokes data about the operation of the mechanisms which integrate human function.

4. Integrative Mechanisms

From the biological side, *the brain* is the organ of the mind; mentation is subject to the same physical influences as all other organ functions. Beyond this, it is the body's chief regulatory and coordinating organ. Inputs to the brain come from both the body and environment. Similarly, its output alters bodily functions and, via behavior, the external world. These inputs from inside and out converge on many of the same brain centers, and the resultant messages alter both psychological and somatic functions.

From a psychological perspective, input consists not merely of sensations but of information organized into meanings. Meanings have individual as well as shared components. The same event or somatic message has different nuances, at least, for each person. The origins of these meanings, as well as those shared in the culture, derive from life experiences: they are learned. Some visceral responses to psychological stimuli also are 'learned,' but the learning that establishes these psychosomatic linkages is not the familiar, conscious process, but *conditioning*, of both Pavlovian and operant types. Most of this occurs in early life. Manipulations during early development can alter brain anatomy, neurotransmitters, and hormone levels, and modify subsequent stress responses.

When important meanings are attached to inputs, emotions are activated. These further affect body functions. The original response becomes a stimulus, too. In studying organismic function, what is defined as stimulus, and what as response, is arbitrary; it depends on what part of the process is being examined and at what point in time. Nor is this a simple chain of stimulus-response pairs. A network of actions follow, in which some responses affect one organ, some another, and some both, while some of these influences are stimulating and others inhibitory; and the succeeding responses become additional stimuli of the same variegated nature. Because many life events have more than one meaning, several such networks may be in process simultaneously. In most instances, then, a stimulus sets off *a transacting web of responses and new stimuli with both psychological and somatic components*.

Ordinarily, inhibitory processes finally dominate, and the person returns to a resting state. Sometimes, especially if the stimulus persists, inhibition may fail, so that the reaction continues or recurs. These extended responses are homeostatically maladaptive, and likely to cause organ damage and disease.

The *endocrine and immune* systems also serve as integrators of organismic function, directly and in their inter-relationships with each other and the brain. Pituitary secretions, which control many other endocrines, are themselves subject to control by hormones and neurotransmitters from the hypothalamus. Psychological stress is associated with stimulation of adrenocortical hormones via the pituitary, as well as

neurally mediated adrenaline release from the adrenal medulla. The distinction between neurotransmitters and hormones is blurred: active neuropeptides and neuroamines are present in many peripheral organs. Hormones act back on the brain to inhibit their own overproduction, but also modify cognitive and emotional states.

Stress also affects both cellular and humoral immune functions, partly through the action of adrenal steroids; depression may have especially potent effects. Animals subjected to a variety of psychological stressors have increased morbidity and mortality to injected pathogens along with changes in immune functions. Hypnosis can cure warts and alter the hypersensitivity reaction to injected antigens. The immune system can 'learn' too: immune responses to previously inert substances can be conditioned.

5. Basic Psychosomatic Concepts

5.1 Health and Disease

Psychosomatic medicine rests on the fundamental physiological concept of homeostasis: the human organism exists in a state of constantly dynamic equilibrium subject to both endogenous shifts and perturbations induced by an ever-changing environment. These trigger homeostatic responses which keep the changes from exceeding viable limits, thereby maintaining organismic integrity and permitting continuing function in the environment. If the change is excessive or the compensatory adjustments become so, damage results. When damage significantly impairs adaptation, disease occurs: disease is failure of adaptation. At any given moment, every person is more or less healthy and more or less diseased.

Psychosomatic medicine enhances this concept to take account of our inherent psychological and physiological properties as well as the sociocultural and informational content of the physical world. Moreover, although 'disease' may seem to have a fixed definition, different times, places, and persons have provided differing criteria for deciding what disease exists and whether one is sick at all. (Epilepsy in ancient Egyptian royalty was a manifestation of divinity.)

5.2 The Nature of Disease; 'Psychosomatic Disease'

Many physicians and others dichotomize disease states as 'organic' (physical) or 'psychological' (sometimes extending these, respectively, as 'real' or 'imaginary'). Specificity theorists separated off 'psychosomatic' diseases in which psychogenesis played a uniquely central role. Psychosomatic medicine takes a wholly different view, conceptualizing *every disease in every*

person as physical AND psychological. There are no 'psychosomatic diseases' because *all diseases (and health) are psychosomatic*.

Even in the simple instance of accidental injury, it is the overall impact of the event on the person, not just the injured part, that defines the illness—including instances where behavioral factors played no causal role and no psychic 'trauma' results. The extent of pain, bleeding, swelling, etc., will be determined not only by the injury but by the victim's entire prior psycho-physiological state; and the victim's reactions to each of these effects, as well as his attitudes about the injured part and its functions, will alter that state. (For example, if anger arises, blood vessels may dilate, increasing bleeding.) The reactions of others, including health care providers, will also affect the psycho-physiological responses, as will the meanings of the treatment given. The somatic problems will be paramount medically, at least until late consequences require chronic care. Conceptually, however, even this extreme example of 'pure physical illness' can be fully understood only from a psychosomatic perspective. Optimal care will result only if health care providers are responsive to psychosocial needs throughout the course of treatment.

5.3 The Etiology of Disease

Similarly, the causes of all diseases are psychosomatic. Take malaria for example. The plasmodium, usually mosquito-borne, is prerequisite. But, what is to be said of the US soldier in the tropics who contracts malaria after failing to take prescribed prophylactic medication or to sleep beneath issued netting, or who fishes in a swamp designated 'off limits.' It is no less valid to categorize such behaviors as causative than the biological factors. Sociocultural factors, even political ones can enter in too. Examples include being a parentless child, or living in a culture ignorant of the mosquito's role or under a political system too poor or uncaring to provide public health programs. Any such factors or other psychological events can be involved in any given person. Often several are.

The choice of an infectious disease was deliberate. Traditional medicine views these as the prototype of external causation. However, as Dubos (1960) indicated, all people are incubators for various micro-organisms growing in their throat, intestine, etc.; and in these circumstances, why are people not sick all the time? The real question is why some people become sick with what they do, when they do. External factors explain only part of this. Host resistance variables, including differences in immunity, are necessarily involved too; and psychosocial factors affect these.

5.4 Pathogenesis; the Onset of Disease

Among the psychosocial causes of illness, *stress* is the best documented. Selye (1950) demonstrated that the

response to any noxious stimulus precipitated a nonspecific 'General Adaptation Syndrome' superimposed on the specific effects of the stimulus. He focused on the adrenocortical response, but it became clear that many other systems are involved in the syndrome, especially other hormones, the cardiovascular system, blood elements, including those active immunologically, and the psychological apparatus. Psychological stressors have proved as potent as physical ones. Indeed, some effects of physical stressors have proved to result from their psychological properties. Multiple clinical studies have documented that a variety of illnesses arise in the context of accumulated life stress.

Specific emotional states seem to operate in some circumstances. Anger and anxiety have been implicated, as well as depression, especially that associated with feelings of helplessness and hopelessness.

But disease is a process with a course that evolves variably, and often preceded by a period in which covert, even undetectable, lesions are developing; and a different superimposed process may be required for the disease to appear. Psychosocial factors have been implicated in each of these phases, but the specifics may differ from one to another, as Reiser (1975) clarified. For example, the gradual accretion of coronary atherosclerosis precedes thrombosis and infarction ('heart attack'). Anger and depression have been implicated in infarction, whereas nonspecific stress has been related to alterations in cholesterol metabolism which can lead to atherosclerosis. Of course, the events that provoke emotions differ among people, depending on their psychological makeup. Research provides the broad possibilities, but only individual study can delineate the pathogenic process in a given person.

The specific 'choice' of illness likely rests in genetic and other purely biological factors in many instances. Individually patterned psychosomatic links may be implicated in others. Either may be responsible for 'response specificity,' the tendency for individuals to have particular somatic responses to a variety of stressors. Moreover, discredited but not disproved psychological 'specificity' remains a possibility. The remarkable consistency among various specificity theorists regarding the psychological characteristics in people with certain disorders remains suggestive. Formidable methodological problems must be overcome to study this adequately, but such research is merited.

6. Clinical Applications

The application of psychosomatic medicine to patient care rests primarily on its basic concepts. As every patient is partly unique, only by identifying the issues specific to that individual can appropriate care be provided. The fundamental need both in diagnosis and treatment is establishing a trusting professional

relationship that allows patients to relax defenses and reveal personal information and feelings. Also required are interviewing skills to facilitate patients' communications of their life experiences, along with physical symptoms and the life context of the latter. This must be accompanied by sufficient knowledge of psychology and social science to identify the significance of these communications, and also of major psychosomatic research findings in order to identify correlations likely to be significant.

Treatment requires the ability to utilize this understanding of the patient in order to select the biological therapies that are most suitable for the individual, and to provide this in a manner that helps that patient use it best. On the psychological side, the need is to quantify and facilitate the extent of emotional release most helpful and tolerable to that patient, as well as the level of insight to pursue the relevant psychosomatic correlations.

All this seems complex, and it is. But the knowledge and skills can be acquired in the course of existing professional education, provided that psychosomatic medicine is integrated throughout. This is true for *all clinical specialties*. As noted initially, there is no specialty of psychosomatic medicine.

At the height of interest in psychosomatic medicine, a cadre of experts began to serve as educators of physicians, medical students, and others in academic health centers. Most were psychiatrists, although some internists and other specialists participated. This became a new subspecialty, named 'Consultation-Liaison Psychiatry,' as the service function became melded with education (see *Liaison and Consultation Psychiatry*). In the late twentieth-century, however, as psychiatry became more biological and positioned within the biomedical model, the psychosomatic element diminished. Given the widespread public dissatisfaction with the 'over(bio)scientific,' 'dehumanized' quality of medical care, this has been a significant loss.

But clinical psychosomatic medicine should not be confused with humanism or doing 'good'; it is applied *science*. Although its clinical practitioners act in a decent, humane fashion, it is not this but their skills and knowledge as applied scientists that make them effective.

7. Recent Developments

Substantial research has flourished under the rubrics of 'health psychology,' 'behavioral medicine,' and 'behavioral psychology' as well as psychosomatic medicine itself. Consequently, high-quality research has been advancing vigorously. Psychophysiological studies have begun to use highly sophisticated probe techniques to measure meaningful physiological parameters. A particularly new and useful approach

involves the application of randomized clinical trials that elucidate both causal factors and the effectiveness of interventions in a number of diseases. New research designs have incorporated more real-life situations, particularly in research on stress. Information particularly relevant to actual clinical and life situations has emerged, giving meaningful promise for the future.

Bibliography

- Alexander F 1950 *Psychosomatic Medicine*. Norton, New York
- Anderson N B, Armstead C A 1995 Toward understanding the association between socioeconomic status and health: A new challenge for the biopsychosocial approach. *Psychosomatic Medicine* **57**: 213–25
- Cannon W B 1915 *Bodily Changes in Pain, Hunger, Fear and Rage*, 2nd edn. Appleton, New York
- Cannon W B 1932 *The Wisdom of the Body*. Norton, New York
- Cobb S 1976 Social support as a moderator of life stress. *Psychosomatic Medicine* **38**: 300–14
- Dubos R 1960 *The Mirage of Health*. Allen and Unwin, London
- Dunbar H F 1935 *Emotions and Bodily Change*. Columbia University Press, New York (see also the 4th edn., 1954)
- Engel G L 1962 *Psychological Development in Health and Disease*. Saunders, Philadelphia, PA
- Engel G L 1977 The need for a new medical model: A challenge for biomedicine. *Science* **196**: 129–36
- Grinker R R 1953 *Psychosomatic Research*. Norton, New York
- Kiecolt-Glaser J K, Glaser R 1988 Methodological issues in behavioral immunology research with humans. *Brain Behavior and Immunology* **2**: 67–78
- Kleinman A, Becker A E 1993 'Sociosomatics': The contributions of anthropology to psychosomatic medicine. *Psychosomatic Medicine* **60**: 389–93
- Lazarus R S 1993 Coping theory and research: Past present and future. *Psychosomatic Medicine* **55**: 234–47
- Mason J W 1968 Organization of psychoendocrine mechanisms. *Psychosomatic Medicine* **30**(Suppl.): 365–808
- McEwen B S 1998 Protective and damaging effects of stress mediators. *New England Journal of Medicine* **338**: 171–9
- Moynihan J A, Ader R 1996 Psychoneuroimmunology: Animal models of disease. *Psychosomatic Medicine* **58**: 546–58
- Oken D 1989 Current theoretical concepts in psychosomatic medicine. In: Kaplan H I, Sadock B J (eds.) *Comprehensive Textbook of Psychiatry/V*, 5th edn. Williams and Wilkins, Baltimore, MD, Chap. 25.2
- Reiser M F 1975 Changing theoretical concepts in psychosomatic medicine. In: Reiser M F (ed.) *American Handbook of Psychiatry, Volume 4*, 2nd edn. Basic Books, New York
- Reiser M F 1984 *Mind, Brain, Body: Toward a Convergence of Psychoanalysis and Neurobiology*. Basic Books, New York
- Selye H 1950 *Stress*. Acta, Montreal, Canada
- Stainbrook E 1952 Psychosomatic medicine in the 19th century. *Psychosomatic Medicine* **14**: 211–27
- Weiner H 1977 *Psychobiology and Human Disease*. Elsevier, New York
- Weiner H, Hofer M A, Stunkard A J (eds.) 1981 *Brain, Behavior and Bodily Disease*. Raven Press, New York

D. Oken

Copyright © 2001 Elsevier Science Ltd.
All rights reserved.

12457

Psychotherapy and Hermeneutics

Well into the 1970s the atmosphere in the field of psychotherapy was characterized by a war of the schools. Psychoanalysis in its various forms, the humanistic movement, behaviorism, and cognitive therapy were the major players in a battle that included an ever growing number of contenders. The main assumption was that one theory was right and all the others were wrong; the question was which one would carry the day.

The statistical meta-analyses of psychotherapy outcome studies produced a result that called this basic assumption into question. It turned out that by and large all psychotherapeutic methods were helpful, and that the theory and technique applied was a very poor predictor of outcome. The question therefore became what metatheoretical framework could make sense of what seemed a conundrum: It simply didn't seem to matter how you worked. If anything it was important that client and therapist had a good rapport, and that the therapist had some convincing narrative framework that made sense of the client's suffering and problems (e.g., Orlinsky and Howard 1986).

This in turn squared well with positions taken by anthropologist Claude Lévi-Strauss ([1949] 1963) and psychiatrist Jerome Frank (1963). Both claimed that twentieth century Western psychotherapy was just a subcase of practices that were to be found in almost every existing culture. Lévi-Strauss's classic paper argued that the shamanic healer and the psychoanalyst did essentially the same thing: They integrated meaningless and therefore frightening phenomena into a symbolic system. Frank's model was that psychotherapy included three factors: a socially established status of healer, a myth or narrative that was applied to the patient's condition, and a ritual that structured the application of the myth to the patient's condition. Essentially the difference between Western and other forms of therapy only concerned the institutions that conferred status of healership, the myths that were applied, and the rituals that were enacted. These theoretical positions foreshadowed the empirical result of psychotherapy research. It turned out that a hermeneutic understanding of psychotherapy was one of the better candidates to present the profession with a new self-understanding.

1. The Nature and History of Hermeneutics

Hermeneutics originally denoted the discipline of interpreting texts. The place of origin of hermeneutics was the interpretation of the holy scriptures of the various religions, which entailed a variety of problems: how to derive laws from texts, and how to find out what the—presumably divine—intention behind the text was. There were several attempts to codify

hermeneutic rules that were guided by specific religious, political, and social needs. The assumption of classical hermeneutics was that there was one right interpretation; the question was how to reach it.

It was in the nineteenth century, with the rise of philology and the *Geisteswissenschaften* as academic disciplines that methodological self-reflection of hermeneutics achieved a different turn. The sharpening of historical consciousness made it evident that the model of the one meaning behind the text (determined by the author's intention) was problematic. Furthermore it became evident that interpretation varied according to the historical position of the interpreter.

In the twentieth century the groundwork for philosophical hermeneutics was laid by Martin Heidegger's *Being and Time* ([1927] 1962). In it Heidegger claimed that the idea of the human subject as a mind contemplating the world was flawed and misleading. Human beings were primarily involved in what he called 'being-in-the-world.' This mode of being involved a pre-reflective interpretation of the world within which subjects projected themselves into a future on the basis of an understanding of their past.

Interpretation was therefore not an activity confined to the reading of texts, but was essential to the human condition as such. There could be no such thing as meeting an object (or another subject) outside an interpretive framework. This preunderstanding was not created by individual subjects through conscious intention. Instead it was the horizon within which all action and understanding took place.

Heidegger's student Hans-Georg Gadamer can be credited with having written the major text of philosophical hermeneutics, *Truth and Method* ([1960] 1975), taking Heidegger's ideas as its foundation. Gadamer claimed that human beings always lived within a horizon of intelligibility, a preunderstanding that guided their whole mode of being. The understanding of texts was just a special case of the activity of understanding essential to human life.

The question of interpretation did not arise if two subjects moved in the same horizon of intelligibility. Understanding each other was then a matter of course as in everyday interactions with people and texts (like newspapers or bank statements). Interpretation becomes problematic only when the speaker's (or author's) and the listener's (or reader's) horizon or intelligibility diverges.

If they did, the reader needed to integrate the text into his or her own horizon of intelligibility, as there can be no understanding outside this context. But in doing so, the reader expands this very horizon of intelligibility. If for example contemporary readers try to understand a Platonic dialogue, they may find out that their horizon of intelligibility simply does not fit some of the assumptions needed to make the Platonic text intelligible. They may therefore try to reconstruct the horizon of intelligibility implicit in the Platonic text.

By doing so, they may come to the point of understanding Plato better than Plato understood himself. Plato (like any one of us) could not be fully aware of the horizons of intelligibility, the pre-understanding of the world that made his world and his life meaningful. Human beings are never completely transparent to themselves, as we are embedded in historical contexts that are not fully understood.

Historical distance therefore creates both the necessity for interpretation and the perspective that allows the interpreter to understand aspects of meaning that were not accessible to the author. At the same time the activity of interpretation also changes the interpreter. The integration of historically (or culturally) distant horizons of intelligibility makes the interpreter more conscious of his or her own point of view, and thus changes and widens it.

Philosophical hermeneutics provided an interpretive frame that made it clear why there could never be a final understanding of any text. Every historical period, and every cultural position, needs to integrate texts and cultural practices that are historically or culturally distant into its own horizons. The interpretive act is therefore much closer to a dialog between interpreter and text (or cultural practice) than to the unearthing of a pre-existing meaning.

2. Hermeneutics and Psychotherapy

The hermeneutic approach to psychopathology and psychiatry was already introduced at the beginning of the twentieth century primarily through Karl Jaspers (1913) and the existential school of psychotherapy. The latter (Ellenberger and May 1958) was primarily based on the direct application of Heidegger's philosophy to the understanding of psychopathology. Psychotherapy was essentially viewed as the elucidation of the patient's modes of being-in-the-world. The point of psychotherapy was therefore to interpret and change the patients' interpretation of their *Lebenswelt*.

The European existentialist tradition in psychotherapy lost its cohesiveness as a movement, and philosophical hermeneutics began to be reapplied to psychotherapy within psychoanalysis first in Europe: Paul Ricoeur ([1965] 1970) in France, Jürgen Habermas (1971) in Germany, and then in the USA, Donald Spence (1981) and Roy Schafer (1983). Their common denominator was to argue that psychoanalysis had misunderstood its own nature by formulating itself in terms of a natural science model.

Psychoanalysis, they argued, was not a theory that discovered causal laws that connected between childhood events and psychopathological symptoms. It was the activity of helping patients to make themselves intelligible to themselves (Strenger 1991).

Schafer and Spence used literary theory in addition to hermeneutics in their attempt to clarify the nature

of psychoanalysis (see *Psychoanalysis in Clinical Psychology*). In Schafer's model patients are seen as those whose narrative of themselves is overly narrow, simplified, and closed. This narrative (the term would correspond to Gadamer's 'preunderstanding') narrows down the patient's options of living, because life, human encounters, and actions are interpreted in overly rigid terms (e.g., 'every interaction is a power-struggle').

Psychoanalytic interpretation in this view is not the uncovering of a historical cause or pre-existing unconscious wish. It is the making explicit of the patients' interpretation of their world, their mind, and their interpersonal relationships. In other words: The object of psychoanalytic interpretation is the patient's self-interpretation. Its goal is to change this interpretation by making it conscious. The very act of making it conscious will lead to a change of the patient's self-interpretation.

In this manner Schafer achieved two objectives: first he freed psychoanalysis from the need of extraclinical validation of causal claims (e.g., unresolved Oedipal issues lead to psychoneurotic symptoms). Second, he made it possible to understand why analysts of different theoretical orientations would come to interpret the patient's mental life in different terms. In Gadamer's terms the analyst could only work from a particular horizon of intelligibility. Understanding the patient always meant the integration of the patient's self-understanding into a new horizon of intelligibility.

Quite quickly the application of the hermeneutic point of view began to spread to other psychotherapeutic approaches. Collections like *Hermeneutics and Psychological Theory* (Messer et al. 1988) showed that the hermeneutic point of view was, in fact, fruitfully applicable to almost every discipline in psychology. Anything from behavioral-cognitive through family therapy to social psychology could be reinterpreted in hermeneutic terms.

Since the 1990s the term 'hermeneutics' has not often been used explicitly in the psychotherapeutic literature. Nevertheless it could be said that its tenets have turned into something like a metatheoretical implicit consensus in the field. Most authors implicitly or explicitly assume that psychotherapy intervenes in the patient's meaning system. Furthermore it has become something like an implicit orthodoxy to accept that there are always many ways to construe the patient's experience.

The hermeneutic understanding of psychotherapy tends to go along with a constructivist understanding of psychotherapeutic activity: Psychotherapy does not uncover (or change) the single cause of symptoms, maladaptive character traits, feeling states, etc. It reinterprets the patient's actions, thoughts, and emotions in new terms. It is taken to be a matter of course that it is conceptually impossible to specify one correct such construal, since construals of the world are always perspectival (cf. Stolorow et al. 1987).

There are two major beneficial implications of the hermeneutic understanding of psychotherapy: first, it has led to a more ecumenical atmosphere in the field. The wars of the schools are largely over, and instead different approaches can enter fruitful dialogs. Second, it has led to greater sensitivity to the patient's right to a subjectivity of his or her own. If all psychotherapy is hermeneutic construction, it is not fruitful and legitimate to impose such construal on the patient. Current writers (e.g., Aron 1996, Strenger 1998) emphasize the necessity of turning the psychotherapeutic interaction into a more egalitarian and dialogical activity.

Meanwhile there have been extensive efforts to detect trends that favor some forms of therapy for particular disturbances (e.g., Roth and Fonagy 1996) that seem to reverse the original verdict that theory and technique do not make any difference whatsoever (e.g., obsessive-compulsive disorder favors a combination between cognitive-behavioral therapy and antidepressant medication; generalized, nonspecific character disorders favor interpersonal and psychodynamic approaches; see *Psychological Treatments, Empirically Supported*). Nevertheless these efforts have maintained the ecumenical trend in psychotherapy. The hermeneutic understanding of the nature of psychotherapy as an activity that generates meaning might well continue to serve as a framework that will enable the continuation of this discussion under the aegis of cooperation rather than competition between schools.

See also: Constructivist Psychotherapies; Hermeneutics, History of

Bibliography

- Aron L 1996 *A Meeting of Minds. Mutuality in Psychoanalysis*. Analytic Press, Hillsdale, NJ
- Ellenberger H, May R (eds.) 1958 *Existence*. Basic Books, New York
- Frank J D 1963 *Persuasion and Healing*. Johns Hopkins University Press, Baltimore, MD
- Gadamer H-G 1960 1975 *Truth and Method*. Seabury Press, New York
- Habermas J 1971 *Knowledge and Human Interest*. Beacon Press, Boston
- Heidegger M 1927 1962 *Being and Time*. Harper and Row, New York
- Jaspers K 1913 *Allgemeine Psychopathologie*. J. Springer, Berlin
- Lévi-Strauss C 1949 1963 The effectiveness of symbols. In: *Structural Anthropology*. Basic Books, New York, Vol. 1
- Messer S B, Sass L A, Woolfolk R L (eds.) 1988 *Hermeneutics and Psychological Theory*. Rutgers University Press, New Brunswick, NJ
- Orlinsky D E, Howard K I 1986 Process and outcome in psychotherapy. In: Garfield S L, Bergin A E (eds.) *Handbook of Psychotherapy and Behavior Change*, 3rd edn. Wiley, New York, pp. 311–84

- Ricoeur P 1970 *Freud and Philosophy*. Yale University Press, New Haven, CT
- Roth A, Fonagy P 1996 *What Works for Whom*. Guilford Press, New York
- Schafer R 1983 *The Analytic Attitude*. Basic Books, New York
- Spence D P 1981 *Narrative Truth and Historical Truth*, 1st edn. Norton, New York
- Stolorow R D, Brandchaft B, Atwood G E 1987 *Psychoanalytic Treatment, An Intersubjective Approach*. Analytic Press, Hillsdale, NJ
- Strenger C 1991 *Between Hermeneutics and Science. An Essay on the Epistemology of Psychoanalysis*. International Universities Press, Madison, WI
- Strenger C 1998 *Individuality, the Impossible Project: Psychoanalysis and Self Creation*. International Universities Press, Madison, WI

C. Strenger

Psychotherapy and Pharmacotherapy, Combined

1. Background and Introduction

Combined psychopharmacology and psychotherapy is a treatment form quite accepted and ubiquitous in modern psychiatric practice (Beitman 1983). It may have its origins in the writings of Fromm-Reichman wherein the role and responsibilities of the analytic treating therapist were differentiated from the managerial therapist, a feature of long-term psychotherapy in the 1940s (Fromm-Reichman 1947). Further explicated by Firman, the rationale of this division of labor was based on analytic principles: the 'psychotherapist' should ideally not be in any managerial relationship with respect to the patient so that the patient's fantasies about the therapist could be viewed and analyzed as material emanating from the patient's own self (Firman 1982). According to Firman, the usefulness of this relationship is two fold: the 'outsider' (managerial) therapist can serve as a negative split object upon whom the patient can focus her own negative features. This serves to allow the tenuous whole-object relationship with the treatment therapist to proceed without jeopardy. Firman viewed the splitting that would occur in such collaborative relationships to be in a therapeutic setting and thus under better and safer control for the patient (Balon 1999, Firman 1982).

Since then, the techniques used in analytic practice and with patients in long-term psychotherapy have been transferred to almost all types of psychotherapy. Fueling this have been newer and simpler types of psychotropic medication which physicians, other than psychiatrists, feel comfortable prescribing (Horgan

1985, Regier et al. 1978, Schurman et al. 1985) and systems of care which have advocated for lower costs for patient care. Additionally, numerous research studies confirm that the combination of psychopharmacology and psychotherapy treatments are more effective than each alone (Goldsmith et al. 1999, Karasu 1992, Koenigsberg 1994, Manning et al. 1992) adding incentives for therapists to seek consultation for their patients from physicians regarding medication.

In modern psychiatry, collaborative treatment wherein a psychiatrist or other physician provides the medication and manages medical issues while a non-physician therapist provides the psychotherapy is a common and growing treatment modality. While the field has not definitively defined this paradigm, the American Psychiatric Association published guidelines describing the responsibilities in a collaborative relationship, noting that the psychiatrist and therapist share responsibility for the care of the patient in accordance with the qualifications and limitations of each person's discipline and abilities (American Psychiatric Association 1980). Various terms have been used in the literature and in clinical care to describe this practice which include medication backup; split treatment; combined treatment; medical management; co-treatment; triangular treatment; collaborative treatment; medical backup, and parallel treatment. For the purposes of this section we will call this practice 'collaborative treatment.'

What has not been definitively researched, however, are the best methods for providing collaborative treatment. Recognizing the cognitive and emotional problems for all parties, Goldsmith and co-workers and others have sought to better explicate the potential difficulties and sources of conflict (Berg 1986, Bradley 1990, Chiles et al. 1984, Goldsmith et al. 1999, Kingsbury 1987). Others have described the positive aspects of this form of treatment (Balon 1999, Michels 1995).

The remainder of this article will focus on some of the highlights of collaborative treatment: the positive and negative aspects, legal and ethical issues, and suggestions for maximizing the collaborative process.

2. Positive and Negative Aspects of Collaborative Treatment

The combination of psychopharmacology and psychotherapy, or collaborative treatment, has some positive and negative aspects. As noted by Beitman (1983), collaborative treatment is based on a triangular relationship between psychiatrists, patients, and nonmedical therapists. In an ideal world, such a triangle should have just positive aspects, e.g., utilization of special talents. Triangular relationships, however, could also be a breeding ground for diffi-

culties or negative aspects, such as splitting. The negative aspects of collaborative treatment do not arise only from patients' pathology. As Brill (1977) pointed out, 'More and more nonmedical mental health professionals began to resist and resent working under the direction of a psychiatrist. They believe that their own professions and therapeutic skills have advanced to the point that independence and self-determination were justified.' He emphasized the emergence of territorial conflicts arising in collaborative 'team' treatment. The territorial or turf conflicts and concerns are a bilateral issue—psychiatrists, as much as nonmedical therapists, frequently view the 'other side' as a competitor rather than a collaborator.

Collaborative treatment could benefit the patient only if it is truly collaborative, when positive aspects are maximized and negative aspects minimized. Awareness of the possible positive and negative aspects of collaborative treatment is the first step in maximizing the collaborative relationship (see below).

2.1 Positive Aspects of Collaborative Treatment

2.1.1 Utilization of special talents of both therapist and psychiatrists. As Balon (1999) pointed out, both professionals may possess expertise in special areas, e.g., mood disorders, eating disorders. Firman (1982) also pointed out that a psychiatrist who is a skillful psychotherapist could also serve as a second therapist, or a negative split object, on whom the patient focuses many of his/her own disowned negative features.

2.1.2 Cost-effective utilization of all resources. This is an issue especially important to patients. Due to various economic pressures, most patients cannot afford psychotherapy by psychiatrists unless they have good mental health care benefits. In many cases, insurance companies either do not cover psychotherapy by psychiatrists or pay an amount that is not acceptable by psychiatrists. Psychotherapy provided by nonmedical psychotherapists is considered cheaper (a contentious issue; see Goldman et al. 1998) and frequently is covered by insurance companies. Most patients thus can still benefit from the combination of pharmacotherapy and psychotherapy.

2.1.3 Better opportunity for patient to select therapist with similar ethnic and/or gender background. Some authors (e.g., Foulks and Pena 1995) have advocated that matching a patient with a therapist or

physician of a similar ethnic or cultural background may facilitate the process of psychotherapy or pharmacotherapy. Others (Yamamoto et al. 1993) have warned that this matching could have its pitfalls and could be discriminatory at times. Nadelson and Zimmerman (1993) have recommended that a patient's request for a therapist or psychiatrist of a specific gender, ethnicity, or race or a patient's uneasiness with a particular therapist's or psychiatrist's gender, ethnicity, or race should be an issue explored and addressed in therapy.

2.1.4 Increased time and resources available for patients. The combination of a 20-minute medication review and 50-minute therapy session represent more resources than just one of these modalities and is an unquestionable benefit. The combination of resources is also advantageous during the temporary absence of a psychiatrist or therapist.

2.1.5 Increased amount of clinical information. There may be information that the patient feels more comfortable sharing with their therapist and vice versa. Thus, psychiatrist and therapist should set time aside to share information, impressions, and concerns and to support each other in their efforts to help patients (Pilette 1988).

2.1.6 Enhancement of compliance. As Paykel (1995) pointed out, psychotherapy may assist or possibly hinder, medication compliance. Similarly, collaborative treatment may also increase compliance with psychotherapy.

2.1.7 Professional and emotional support for both therapist and psychiatrist. As Pilette (1988) pointed out, such support is especially important when the patient is difficult to treat and also during times of crisis.

2.2 Negative Aspects of Collaborative Treatment

2.2.1 Inappropriate prescribing decisions without knowledge of the content of therapy. The prescribing psychiatrist may not always be aware of all the inner conflicts or external pressures felt by the patient. For instance, the psychiatrist may prescribe medication

for anxiety that arose after an important intrapsychic conflict was addressed in therapy with the non-medical therapist.

2.2.2 Potentially discrepant information given by the patient to each clinician. Frequently, the discrepant information provided by the patient to both treating professionals could lead to decisions with serious consequences. This issue is directly related to the next one—splitting.

2.2.3 Splitting of clinicians. This is probably the most frequently discussed negative aspect of collaborative treatment. Treatment provided by two persons provides a fertile ground for a patient to develop negative transference, to introduce problematic countertransference and to split the prescribing psychiatrist and therapist. Splitting may lead to non-compliance with any of the treatments.

2.2.4 Unclear confidentiality. As Pilette (1988) pointed out, is the collaborative treatment an occasion for free communication between a therapist and psychiatrist without the patient's consent? What if the patient asks that some important information not be related to the other treating professional? These issues are complicated and need to be discussed and resolved.

2.2.5 Clouded legal and clinical responsibility. It is not always clear who is legally responsible for various aspects of collaborative treatment. It is rather safe to assume, however, that the psychiatrist will most likely be found legally responsible for all aspects of clinical care including inappropriate or negligent supervision (Balon 1999).

2.2.6 Lack of reimbursement for collaboration. The fact that third party payors do not reimburse for collaboration somewhat hinders the collaborative treatment.

2.2.7 Various misconceptions of patient, therapist, and psychiatrist. These include issues such as therapists and psychiatrists ignoring the psychological meaning of medication, psychiatrists' belief that therapists should always agree with and support physicians' decisions, and patients' common percep-

tion that psychiatrists are only interested in prescribing medication (Goldsmith et al. 1999).

This is only a brief summary of positive and negative aspects of collaborative treatment. More thorough discussion of both positive and negative aspects is provided in the two introductory chapters of *Psychopharmacology and Psychotherapy. A Collaborative Approach* (Goldsmith et al. 1999, Riba and Balon 1999).

3. Legal and Ethical Issues

3.1 Legal

The major worry in collaborative treatment is that with two clinicians delivering the care, there will be a loss of information or understanding of what may be occurring for the patient. There are currently no standards in place for how frequently or for how long patients should be seen by physicians if collaborative treatment is provided. There is, therefore, a broad range of frequency of visits and communication patterns between the patient, therapist, and physician, which range from ideal to inadequate. Because medications are involved in collaborative treatment and historically have been the source of significant legal exposure, the issue of medical practice is an important topic for discussion. Negligence underlying malpractice liability is of two types: (a) carelessness in gathering necessary data about a patient and (b) negligence in assessing that information in determining appropriate treatment modalities, interventions, providing precautions, etc. (Macbeth 1999). Psychiatric medications, like all medications, are associated with side effects. Physicians who prescribe medications become liable if there is a deviation in the accepted uses of the medications and there results harm to the patient. It is critical for physicians to obtain informed consent from patients when prescribing medication and also to ascertain liability for patient suicide.

If patients are going to be seen infrequently by the prescribing physician, there must be a way for the physician to set up a communication pattern where side effects and suicidality are made known, to the extent possible, to the physician, on a regular basis. For example, if a patient were being seen by the physician every six months but the therapist is seeing the patient weekly, then the therapist and physician must discuss how issues and developments about the patient can be communicated.

In collaborative treatment, it must be determined how the treatment plan is shared between the clinicians; how emergencies including hospitalization will be managed; and who is directing the psychotherapy. The setting may play a large role in the extent to which the physician, therapist, and patient com-

municate. In many public settings, for example, the psychiatrist only sees the patient infrequently. The patient, however, has a case manager whose job it is to communicate to the various clinicians about the patient. In managed care settings, psychiatrists may find themselves in relationships with therapists whom they don't know and who are unaware of their skills, talents, experience, etc. There may be disincentives for the physician to see the patient often, even when more psychosocial problems accumulate for the patient. In collaborative treatment in certain settings, patients sometimes don't even know who the doctor is and often the physician has not had the chance or taken the opportunity to meet with the patient's family, since others (therapists, case managers, nurses, a previous physician) might have already done so.

There is, therefore, a range of potential legal problems for physicians who engage in collaborative treatment. Physicians are usually viewed as 'the deep pocket' when there is a lawsuit regarding a patient's treatment. The physician usually carries more insurance than the nonphysician therapist and is viewed by juries as 'being in charge' by virtue of education and experience. Even if the plaintiff's attorney does not believe that the psychiatrist was central to the matter, it is often likely that the psychiatrist will be part of the suit to extend the limits of insurance and to try to reach 'the deep pocket.' There is also the problem that occurs when there is more than one defendant. One defendant can try to make their case stronger by blaming the other defendant, thereby turning against one another. Another problem is that even if medications or the medical management provided by the physician is of minor importance to the reason for the suit, regardless of the realities, the physician is often viewed as being in charge while the therapist is viewed as subordinate to the physician.

Finally, physicians are increasingly signing contracts to provide services for patients in managed care. These contracts are usually written by the companies themselves and the fine print is not read carefully by the physician, causing them to be implicitly involved in relationships with therapists that are not collaborative in nature but more supervisory. This practice, therefore, often puts the physician at further risk.

There are some suggestions for clinicians to heed in order to decrease legal risk in collaborative treatment:

(a) Clinicians should understand the settings they are practicing in, including its operations and routines.

(b) Care should not be compromised because of the setting or because the patient is being treated in a collaborative model.

(c) Use good communication techniques with professionals with whom you are sharing responsibility for the patient. Document these discussions.

(d) If supervision is part of the collaborative relationship, then set regular supervision times and amend depending on the patient's condition and needs.

(e) Read the fine print in all managed care and other contracts that are signed.

For a more thorough discussion of these issues, please see the work of Macbeth (1999).

3.2 Ethical

When patients receive collaborative treatment because that is what is deemed best for their care, then there is not an ethical problem. An ethical and clinical dilemma however, may exist if such care is provided because of cost considerations (Lazarus 1999). While some patients clearly benefit from team approaches to care, others do not. Research is starting to help us discriminate between such patients and determine mental health outcomes in different practice and fee arrangements (Rogers et al. 1993, Wells et al. 1992). Goldman et al. (1998), for example, has shown that for many patients, psychotherapy and psychopharmacology provided by a psychiatrist alone is as effective and less costly than collaborative treatment.

There are four bioethical principles that need to be considered when one is providing collaborative treatment (Beauchamp and Childress 1994):

- (a) Beneficence: the covenant between professional and patient to do what is in the patient's best interest;
- (b) Nonmaleficence: the principle of doing no harm;
- (c) Autonomy: respect for the individual rights of patients, especially in the area of informed consent and treatment refusal; and
- (d) Justice: the fair treatment of patients.

As noted by Lazarus, when determining ethical dilemmas related to collaborative treatment, any two or more of the above principles may be compromised or in conflict (Lazarus 1999). In this era of cost containment and lack of standards regarding collaborative treatment, all of these principles are being tested. The American Psychiatric Association (APA 1993) and the American Medical Association (AMA 1994) have addressed ethical principles and have recognized the importance of collaborative treatment. Section 5 of the AMA principles states: 'A physician shall continue to study, apply and advance scientific knowledge, make relevant information available to patients, colleagues, and the public, obtain consultation, and use the talents of other health professionals when indicated' (AMA 1994).

It is a complex area where psychiatrists and other mental health professionals engaged in collaborative treatment can find themselves in ethical dilemmas. Some of this has to do with the settings and the rules within those settings; physician-therapist relationships that should be modified because of the patient's condition but may not be; need for differential types of supervision and monitoring between therapist and physician but which are not taken into consideration; fee schedules that cause disincentives for providers to

see patients more often, even when needed; lack of understanding by the patient of the nature of the collaborative relationship between the providers; etc.

In order for health professionals to provide psychotherapy and psychopharmacology in an ethical manner, Lazarus (1999) suggests:

(a) The physician and therapist keep themselves adequately and appropriately informed of the nature of the treatment and the progress of those patients cared for in this type of treatment.

(b) The physician and therapist assure themselves that the treatment is being carried out competently and adequately.

(c) The physician and therapist provide enough supervision, consultation, or collaboration to ensure that their ethical, medical, and legal responsibilities towards the patient are satisfied.

(d) The collaborative treatment is consistent with any local, state or national guidelines outlining psychiatrists' responsibilities to patients and in conformity with licensure requirements and the scope of practice of all the healthcare professionals involved.

4. Maximizing the Collaborative Relationship

Providing combined psychopharmacology and psychotherapy is a complex enterprise, much more difficult than a single therapist providing either the psychotherapy; psychotherapy and pharmacotherapy; or pharmacotherapy. There are, therefore, some general suggestions offered that might be useful.

It is incumbent upon the clinicians and patients involved in such treatment to be aware of the complicated nature of this form of care. While combining psychopharmacology and psychotherapy offers distinct and unique advantages, there are problems, as noted in the above section. Time and attention must, therefore, be given to the process of treatment as well as the content. Written contracts between clinicians have even been suggested as a way to delineate the specific duties and responsibilities of providers (Appelbaum 1991). Other authors have suggested that the patient be part of the three way contract and so be asked to sign such agreements for care (Chiles et al. 1984, Woodward et al. 1993). Specific guidelines for combined care based on the phases of treatment (initial, mid, termination) have been described by Rand (1999). Kahn (1990) has delineated the ideological differences that the psychotherapist and psychopharmacologist often have and the identified roles they must play in order to provide such treatment.

It is probably prudent for clinicians, both physicians and nonphysician therapists, to review their case mix from time to time. Goals would be to limit the numbers of such collaborative treatments based on the amount of time it takes to speak with other therapists; review and develop joint treatment plans; read other clinicians' notes; determine the background and treat-

ment orientation of the clinicians who are collaborating on patient care; work with patients' families and significant others; develop vacation and coverage schedules with collaborating clinicians; discuss such arrangements with patients; etc. Such thoughtful processes may lead to a decrease in malpractice exposure and ultimately more efficient and useful clinical care.

Attention must be paid carefully to the issue of confidentiality and communication between clinicians and with the patient. Consent to discuss treatment plans as well as content of sessions must be clarified. Since it is impossible to initially predict what will be important or relevant, the patient must be cognizant of these communication patterns and provide consent to clinicians at the beginning of treatment. Any barriers to such communication should be discussed with the patient as well as any relevant family members. It should be determined how the information will be transmitted, that is by telephone, electronic communication, mail, in person, etc. and how often. Communication must be regular, not just when there is a crisis (Vasile and Gutheil 1979). It is critical that the patient understands which type of problems should be discussed with which clinician. Further, patients should not be placed in the position of needing to relay information between clinician and the other.

Even with that said, it will sometimes be difficult to sort out. As an example, often what seems like uncomfortable medication side effects could be resistance to taking medications. What seems like a dynamic issue in therapy could in fact be a medical problem. Clinicians need to be supportive of the treatment plan; listen for resistances to either psychotherapy or pharmacotherapy; try to delineate the boundary of the treatment being provided by the individual clinician; and talk regularly with one another about the patient's progress. Because of financial remuneration patterns and modern contractual agreements between clinicians and payers, it is easier than ever for a clinician to find themselves responsible for large numbers of patients, many of whom are not being seen frequently or regularly. This can lead to problems with the inherent doctor-patient relationship that is the hallmark of providing care, whether it is pharmacotherapy, psychotherapy, or both. The clinician is well advised to review the ethical boundaries of the work situation and the nature of the tasks required.

Certain patient populations may be at higher risk in such combined therapies. For example, patients with personality disorders are notoriously difficult to treat given their problems with interpersonal relationships. As summarized by Silk (1999) and others, patients with Cluster B diagnoses are prone to splitting (Kernberg 1975); often do not tell the entire story equally to all clinicians (Main 1957); and tend to externalize problems (Silk et al. 1995). When there is

concomitant substance abuse, emotional lability, and self-mutilating behaviors, combined treatment with more than one clinician can be enormously difficult and complex.

Finally, psychiatry residency training programs and graduate programs in social work and psychology must develop education and training in the specific area of combined treatment. Teaching psychopharmacology alone or psychotherapy without integrated pharmacotherapy is no longer sufficient. There needs to be curricula developed and teachers who are trained themselves in how to deliver such care with emphasis on learning to communicate clearly with other professionals; thinking clearly about diagnosis and treatment strategies; and developing leadership abilities (Spitz et al. 1999).

5. Conclusion

Providing combined pharmacotherapy and psychotherapy is a growing and important part of psychiatric care. It is a complex dynamic treatment, fraught with potential pitfalls but offering many advantages. Importantly, it offers a unique opportunity for a biopsychosocial and fiscal model of care and interesting areas for future research and clinical development.

Bibliography

- American Medical Association Council on Ethical and Judicial Affairs 1994 *Code of Medical Ethics: Current Opinions with Annotations*. American Medical Association, Chicago, IL
- American Psychiatric Association 1980 Guidelines for psychiatrists in consultative, supervisory, or collaborative relationships with nonmedical therapists. *American Journal of Psychiatry* **137**: 1489-91
- American Psychiatric Association 1993 *The Principles of Medical Ethics with Annotations Especially Applicable to Psychiatry*. American Psychiatric Association, Washington, DC
- Appelbaum P S 1991 General guidelines for psychiatrists who prescribe medication for patients treated by nonmedical psychotherapists. *Hospital and Community Psychiatry* **42**: 281-2
- Balon R 1999 Positive aspects of collaborative treatment. In: Riba M B, Balon R (eds.) *Psychopharmacology and Psychotherapy: A Collaborative Approach*. American Psychiatric Press, Washington, DC, pp. 1-31
- Beauchamp T L, Childress J F 1994 *Principles of Biomedical Ethics*. Oxford University Press, New York
- Beitman B D 1983 The demographics of American psychotherapists: A pilot study. *American Journal of Psychotherapy* **37**: 37-48
- Berg M 1986 Toward a diagnostic alliance between psychiatrist and psychologist. *American Psychology* **41**: 52-9
- Bradley S S 1990 Nonphysician psychotherapist-physician pharmacotherapist: A new model for concurrent treatment. *Psychiatric Clinics of North America* **13**: 307-22
- Brill N Q 1977 Delineating the role of the psychiatrist on the psychiatric team. *Hospital and Community Psychiatry* **28**: 542-4

- Chiles J A, Carlin A S, Beitman B D 1984 A physician, a nonmedical psychotherapist, and a patient: The pharmacotherapy-psychotherapy triangle. In: Beitman B D, Klerman G L (eds.) *Combining Psychotherapy and Drug Therapy in Clinical Practice*. SP Medical and Scientific Books, New York, pp. 89–101
- Firman G J 1982 The psychiatrist-nonmedical psychotherapy team: Opportunities for a therapeutic synergy. *Journal of Operational Psychiatry* **13**: 32–6
- Foulks E F, Pena J M 1995 Ethnicity and pharmacotherapy: a component in the treatment of cocaine addiction in African Americans. *Psychiatric Clinics of North America* **18**: 607–20
- Fromm-Reichman F 1947 Problems of therapeutic management in a psychoanalytic hospital. *Psychoanalysis Quarterly* **16**: 325–56
- Goldman W, McCulloch J, Cuffel B et al. 1998 Outpatient utilization patterns of integrated and split psychotherapy and pharmacotherapy for depression. *Psychiatric Services* **49**: 477–82
- Goldsmith R J, Paris M, Riba M B 1999 Negative aspects of collaborative treatment. In: Riba M B, Balon R (eds.) *Psychopharmacology and Psychotherapy: A Collaborative Approach*. American Psychiatric Press, Washington, DC, pp. 33–63
- Horgan C M 1985 Specialty and general ambulatory mental health services: Comparison of utilization and expenditures. *Archives of General Psychiatry* **42**: 565–72
- Kahn D A 1990 The dichotomy of drugs and psychotherapy. *Psychiatric Clinics of North America* **13**: 197–208
- Karasu T B 1992 The worst of times, the best of times: Psychotherapy in the 1990s. *Journal of Psychotherapy Practical Research* **1**: 2–15
- Kernberg O F 1975 *Borderline Conditions and Pathological Narcissism*. Jason Aronson, New York
- Kingsbury S J 1987 Cognitive differences between clinical psychologists and psychiatrists. *American Psychology* **42**: 152–6
- Koenigsberg H W 1994 The combination of psychotherapy and pharmacotherapy in the treatment of borderline patients. *Journal of Psychotherapy Practical Research* **3**: 93–107
- Lazarus J 1999 Ethical issues in collaborative or divided treatment. Risk. In: Riba M B, Balon R (eds.) *Psychopharmacology and Psychotherapy: A Collaborative Approach*. American Psychiatric Press, Washington, DC, pp. 159–77
- MacBeth J E 1999 Divided treatment: Legal implications and risk. In: Riba M B, Balon R (eds.) *Psychopharmacology and Psychotherapy: A Collaborative Approach*. American Psychiatric Press, Washington, DC, pp. 111–58
- Main T F 1957 The ailment. *British Journal of Medical Psychology* **30**: 129–45
- Manning D W, Markowitz J C, Frances A J 1992 A review of combined psychotherapy and pharmacotherapy in the treatment of depression. *Journal of Psychotherapy Practical Research* **1**: 103–16
- Michels R 1995 Commentary: Medication use during psychoanalysis: A survey. *Journal of Clinical Psychiatry* **56**: 179
- Nadelson C C, Zimmerman V 1993 Culture and psychiatric care of women. In: Gow A C (ed.) *Culture, Ethnicity and Mental Illness*. American Psychiatric Press, Washington, DC, pp. 501–15
- Paykel E L 1995 Psychotherapy, medication combinations, and compliance. *Journal of Clinical Psychiatry* **56**(Suppl. 1): 29–30
- Pilette W L 1988 The rise of trepany treatment relationships. *Psychotherapy* **25**: 420–3
- Rand E H 1999 Guidelines to maximize the process of collaborative treatment. In: Riba M B, Balon R (eds.) *Psychopharmacology and Psychotherapy: A Collaborative Approach*. American Psychiatric Press, Washington, DC, pp. 353–80
- Regier D A, Goldberg I D, Taube C A 1978 The *de facto* US mental health services system: A public health perspective. *Archives of General Psychiatry* **35**: 685–93
- Riba M B, Balon R (eds.) 1999 *Psychopharmacology and Psychotherapy: A Collaborative Approach*. American Psychiatric Press, Washington, DC
- Rogers W H, Wells K B, Meredith L S et al. 1993 Outcomes for adult outpatients with depression under prepaid or fee-for-service financing. *Archives of General Psychiatry* **50**: 517–25
- Schurman R A, Kramer P D, Mitchell J B 1985 The hidden mental health network: Treatment of mental illness by nonpsychiatrist physicians. *Archives of General Psychiatry* **42**: 89–94
- Silk K R 1999 Collaborative treatment: Personality disorder patients. In: Riba M B, Balon R (eds.) *Psychopharmacology and Psychotherapy: A Collaborative Approach*. American Psychiatric Press, Washington, DC, pp. 221–77
- Silk K R, Lee S, Hill E M et al. 1995 Borderline personality-disorder symptoms and severity of sexual abuse. *American Journal of Psychiatry* **152**: 1059–64
- Spitz D, Hansen-Grant S, Riba M B 1999 Residency training issues in collaborative treatment. In: Riba M B, Balon R (eds.) *Psychopharmacology and Psychotherapy: A Collaborative Approach*. American Psychiatric Press, Washington, DC, pp. 279–306
- Vasile R G, Gutheil T G 1979 The psychiatrist as medical backup: Ambiguity in the delegation of clinical responsibility. *American Journal of Psychiatry* **136**: 1292–6
- Wells K B, Burnam M A, Rogers W H et al. 1992 The course of depression in adult outpatients: Results from the medical outcomes study. *Archives of General Psychiatry* **49**: 788–94
- Woodward B, Duckworth K S, Gutheil T G 1993 The pharmacotherapist-psychotherapist collaboration. In: Oldham J M, Riba M B, Tasman A (eds.) *American Psychiatric Press Review of Psychiatry*, Vol. 12. American Psychiatric Press, Washington, DC, pp. 631–39
- Yamamoto J, Silva J A, Justice L R et al. 1993 Cross-cultural psychotherapy. In: Gow A C (ed.) *Culture, Ethnicity and Mental Illness*. American Psychiatric Press, Washington, DC, pp. 101–24

M. B. Riba and R. Balon

Psychotherapy and Religion

It has been suggested that the reason psychotherapy is consistently linked with medicine and a medical model is an artifact of Freud's profession as a neurologist. Had he come from another discipline, what is considered to be therapeutic might be altogether different. If, for example, he had been a member of the clergy then perhaps the relationship between psychotherapy and religion that has persisted in the twentieth century might be less adversarial. The politics, economics, and

social forces that have been behind the formation of the relationship between religion and psychotherapy are beyond the scope of this article. The intent here is to present the parallel paths of religion, particularly that with Judeo-Christian roots, and psychotherapy as it has developed in the industrialized West. We will consider at which points these paths have crossed, where they have diverged, and, ultimately, what effect this has had on the practice of psychotherapy as it is understood at the beginning of the twenty-first century.

1. Establishing Common Language

Prior to our exploration of the paths of religion and psychotherapy, a common language for health and pathology must be established in order to accurately read the road markers. Miller and Thoresen (Miller 1999) have proposed a construct of health that includes the domains of suffering, function, and coherence. As they point out, suffering is a state that can be understood as a 'form of unhealth and a common reason for seeking help, from either spiritual or psychological professionals.' Second, functional ability also is a term that is understood in either discipline. Third, the domain of coherence or inner peace is also defined in the language of both disciplines, although perhaps more subjective and less easily observed than the other two.

Due to a lack of specific focus on behavior, these overlapping domains often add more confusion than clarity, as in the case of certain extreme behaviors diagnosed as psychotic when considered from a therapeutic standpoint. Within this context, this cluster of behaviors suggests that an individual's normal affective and cognitive functions have gone awry. Common manifestations include audio and/or visual hallucinations, delusions that may appear grandiose or paranoid in nature, hypersensitivity, and an apparent disregard for those personal care habits consistently linked with health. However, these same behaviors in the context of religious tradition might be deemed a mystical experience or even the pinnacle of spiritual health. William James conceptualized the mystical state as merely another form of consciousness, albeit rare, that could be stimulated either by internal states or external substance. This mirrors Hunt's (1985) sentiment that mysticism is a direct manifestation of normal but usually inaccessible semantic operations. Therefore, it would appear then that specific target behaviors are unlikely to be the most fruitful ground for establishing common definitions for health and pathology. Perhaps a more useful approach would focus not on the specific behavior itself, but on the individual's experience of it, which is completely in line with the Miller and Thoresen (Miller 1999) domains.

James described four aspects of the truly mystical experience that appear to be rare in the experience of psychosis. The first is presentience. Mystical transcendence is more similar with an intuitive sense of knowing than it is with knowledge derived from either thought or sensation. Additionally, this transcendental experience affords the individual a new awareness of 'self' and 'other.' Armstrong (1989) adds that true mystics then demonstrate that this inner identity has come to fruition in outer world behavior. In other words, to be 'an integrated inner person is to be a productive outer person,' which is rarely the experience of the person in the midst of a psychotic episode. The second characteristic has been termed a noetic quality, or that of a being knowing about itself. This knowledge can be similar to a successful outcome in psychotherapy, but rarely of the psychotic experience. The third quality is that the mystical experience is usually transient, even in the most devout and saintly. It apparently is impossible to hold on to the presence of the Absolute for any length of time. The final characteristic of the mystical that was suggested by James was that of passivity. While an individual can dispose oneself to incorporate a deeper understanding and knowledge of the Divine, this wisdom cannot be called up like a specter. James noted that it was not uncommon for mystical union to be methodically cultivated as an element of religious life, similar to the actions of medieval mystics. But, he added, it is possible to stumble into the mystical, perhaps if one is inherently of the appropriate nature. What is also important in this regard, is that both Eastern and Western traditions have suggested that the mystical experience is not pure if the person is unprepared for or unwilling to lead a better life based on the knowledge imparted. This is a vital distinction given the common and profound sense of unpreparedness of persons experiencing severe psychopathology.

As has been previously suggested, when these four attributes of the mystical are compared to psychotic experiences, one may find some common observable traits. However, in studies comparing the experiences of people with schizophrenia and mystics, Buckley (1981) found that while both share a sense of noesis, heightened perceptual abilities, communication with the divine, and exultation, only in psychosis was there the experience of thought disruption, and therefore an inability to incorporate the knowledge in an effective way. It is also true that individuals with schizophrenia can go for extended periods of time without any relief from these experiences, where the mystic, as noted above, sustains this experience only briefly. And, in spiritual people, mystical experiences correlate significantly with psychological well-being, self-actualization, and optimism. Conversely, psychotic episodes tend to be marked with more negative emotional affect.

To summarize, what is aberrant and what is healthy, regardless of whether one is following a psychological

or religious model, is not so much the events themselves but the experience of the events. Although psychospiritual experiences can be present in many different contexts, it would seem that both psychotherapists and religious practitioners can agree on an optimally healthy state for any individual. While cultural variation will no doubt exist, this target state would allow the person to function, for the most part, within the affective and cognitive norm, supported by and supportive of a community that could foster these goals. This common definition of health will allow better exploration of the paths of religion and psychotherapeutic practice, regardless of psychological sophistication or religious beliefs.

2. Historical Perspective

In the late 1800s there was little divergence in the paths of religion and psychotherapy. The period was replete with schools of thought that outlined methods generally considered to be faith healing or mind cures. These approaches had strong support from some of the most influential writers of the time, such as Alcott, Emerson, and Thoreau. Overwhelmingly positive in their beliefs about an individual's potential for rekindling the divine spark believed by some to be inherent in all of us, one can well imagine their appeal when contrasted to the methods and practices common to the nascent psychiatric science of the time. These approaches held that illness, whether mental or physical, was the direct result of one being out of sync with some universal law. In contrast, the faith healing interventions, whether called down by external sources or fostered through self-help, merely required the realignment of one's beliefs and actions to be more attuned with the ultimate reality and truth. The writings of William James, the Mental Hygiene movement, and the resurgence of interest in the use of the group as an optimal setting for the healing of the individual all were well in line with religious beliefs of the day. However, this openly parallel course did not last.

By the early century, psychoanalytic and behavioral theories of the time seemed to suggest that humans were either unsuspecting victims of various drive states or unformed blocks to be molded by external cues or forces. Humans had lost their status as the pinnacle of creation. In retrospect, it is interesting to note that this portrait of human beings as rather hapless characters, lacking control over some of their most basic qualities, was not so different from some of Fundamentalism's tenets about the inherent evilness of humankind in the absence of God's grace.

Convergence began again toward the middle of the twentieth century. In response to the effects of the Second World War and the coming of the nuclear age, Western thinkers, both religious and psychotherapeutic, again began to re-evaluate the nature of

humankind, its inherent qualities, and its potential for change. Important to both was a movement toward a more positive view of human nature with less of the pessimism that had marked some of psychotherapy's models in recent history. The existential shift that had affected so much of the Western world was incorporated into current thought. The nationalism that underlay much of the global conflict gave way to greater acceptance of 'that which is different' as being something other than 'that which is worse.' Prevalent again was the idea that humans were capable of virtuous acts, but this time the element of choice was worked strongly into the equation. And, while a number of religious leaders appeared to have understood that religion and psychotherapy were moving toward another confluence, many of today's psychotherapists question whether religion and psychotherapy can be reconciled. It is on this discrepancy that we now focus.

3. Current Perspective

The tradition of professional detachment regarding the spiritual beliefs of patients is reflected in current survey data indicating a profound disconnection between the religious beliefs and desires of the general population and mental health providers. For example, while 72 percent of the population view religion as an important influence on their lives, only 33 percent of psychologists respond similarly. In academic settings, psychologists are twice as likely as physicians to report that they have no religious preference (Shafranske 1996). Clinically, the issue of client religious beliefs typically is ignored or may be viewed as pathological. In contrast, survey data consistently reveal that religious beliefs are important to the general population, those presumably treated by mental health practitioners. Over 86 percent claim that they believe in God and 70 percent believe in a God who answers prayer (Barna 1992). Importantly, over half of responders state that they would prefer religious to secular mental healthcare regarding certain issues, especially marital and family problems (Privette et al. 1994).

There is reason to believe, however, that mainstream professional indifference to religious beliefs in psychotherapy has been overestimated. While mental health providers are less likely to report that religious beliefs are a major influence on their lives, 77 percent of them do claim to try to live their lives according to religious beliefs, which closely mirrors the general population (Bergin and Jensen 1990, p. 5). According to Miller and Thoreson, there is a growing recognition in the mental health field that religious beliefs in clients are correlated with longer life, and lower incidence of suicide, divorce, and alcohol abuse (Miller 1999). When the impact of religious beliefs is measured upon 'hard' variables, meaning those that are observable and easily

measured such as longevity, the correlation almost always indicates a protective effect of such beliefs. When the impact of religious variables is measured upon 'soft' variables, those that are not observable and easily measured such as rigidity, the results are mixed. Several researchers have noted that the anti-religious bias of mental health providers may lead to a lack of attention to the obvious benefits conferred by religious beliefs in important domains of human functioning, while overemphasizing the negative effects of religious beliefs in intrapsychic domains (Bergin et al. 1996).

Mental health providers clearly are moving toward a more egalitarian approach to the spiritual and religious beliefs of their clients, with momentum. For example, in 1992 the American Psychological Association included religious beliefs among the spectrum of diversity which is to be protected by psychologists, who are expressly forbidden to discriminate or tolerate discrimination on the basis of religious beliefs (Bergin et al. 1996). Additionally, a suggestion to add a new code to the *DSM* diagnostic formulation to indicate psychoreligious or psychospiritual problems (Lukoff et al. 1992) has been accepted. Perhaps more importantly, psychologists and other mental health providers are beginning to examine empirically the impact of religious beliefs upon the process of psychotherapy. Initial findings indicate that integrating religious beliefs into psychological treatments may offer superior results for anxiety (Propst et al. 1992) and depression (Razali et al. 1998) in controlled studies with random assignment, for clients with religious beliefs. The impact of incorporating a spiritual dimension to psychotherapy is perhaps nowhere more visible than in the treatment of addictions. The long tradition of incorporating spiritual renewal found in the 12-step approach has received empirical support in a growing number of correlational studies (Miller 1999) and rigorous research studies (Project MATCH 1997), regardless of client's pre-existing religious beliefs.

Another factor that may be influencing psychology's consideration of religious beliefs and practices is the large number of assessment instruments, with psychometric validation, currently available for the measurement of a wide variety of religious and spiritual constructs (see Hill et al. 1999). Many of these are relatively brief, but provide useful information on such topics as coherence, purpose in life, and spiritual well-being. The work of Pargament and Brant (1998) has addressed systematically an individual's coping strategies, their roots in either religious or spiritual beliefs, and where they are related to negative or positive health outcomes.

It appears, therefore, that the roads have once again begun to converge and that more practitioners of psychotherapy are recognizing the potential importance of assessing religious beliefs and incorporating them into practice. This is a marked change from the

belief that client's spiritual beliefs are pathological, irrelevant, or damaging to the therapeutic process (Shafranske 1996).

4. Future Directions

Ultimately, however, if the relationship between religion and psychotherapy is to continue to grow, there are several areas that will continue to challenge us. First, it is no longer the case that the vast majority of persons in industrialized Western countries are descendent from Anglo-European lineage. By the middle of this new century it is estimated that Whites will be a racial minority in the United States. Subsequently, both psychology and religion will have to incorporate this divergence. This will mean that the concepts of health and spiritual well-being probably will come to include philosophies prominent in Asian and African cultures and religions. These cultures have often looked at the inherent value of a single human being, the role of suffering, and the connection with the environment in a manner not typical of current psychological and Judeo-Christian tenets. Second, the role of women has changed radically since the early days of both Judeo-Christian religions and psychotherapy. In this case, feminist voices from both disciplines are likely to continue to express conflict with the goals of therapeutic intervention and mechanisms of change. Last, as scientific understanding of the working of the human brain and nervous system continues to progress, it will be important for both psychotherapy and religion to remain flexible enough to incorporate this new knowledge.

See also: Clinical Psychology in Europe, History of; Counseling and Psychotherapy: Ethnic and Cultural Differences; Psychotherapy: Ethical Issues; Psychotherapy, History of: Psychiatric Aspects; Psychotherapy Integration; Religion and Health; Religion: Psychiatric Aspects; Religion, Psychology of; Religiosity: Modern; Religiosity, Sociology of

Bibliography

- Armstrong R 1989 The psychology of inner and outer space. *Pastoral Psychology* 37: 161-4
- Barna G 1992 *What Americans Believe: An Annual Survey of Values and Religious Views in the United States*. Regal Books, Ventura, CA
- Bergin A E, Jensen J P 1990 Religiosity of psychotherapists: A national survey. *Psychotherapy* 27(1): 3-7
- Bergin A E, Payne I R, Richards P S 1996 Values in psychotherapy. In: Shafranske E (ed.) *Religion and the Clinical Practice of Psychology*. American Psychological Association, Washington, DC
- Buckley P 1981 Mystical experience and schizophrenia. *Schizophrenia Bulletin* 7: 516-21
- Hill P, Butman R, Hood R 1999 *Measures of Religiosity*. Religious Education Press, Birmingham, AL

- Hunt H 1985 Relations between the phenomena of religious mysticism (altered states of consciousness) and the psychology of thought: A cognitive psychology of states of consciousness and the necessity of subjective states for cognitive theory. *Perceptual and Motor Skills* **61**: 911–61
- Lukoff D, Lu F, Turner R 1992 Toward a more culturally sensitive *DSM-IV*. Psychoreligious and psychospiritual problems. *Journal of Nervous and Mental Disease* **180**(11): 673–82
- Miller W R (ed.) 1999 *Integrating Spirituality into Treatment: Resources for Practitioners*. American Psychological Association, Washington, DC, pp. 3–18
- Pargament K I, Brant C R 1998 Religion and coping. In: Koenig H G (ed.) *Handbook of Religion and Mental Health*. Academic Press, San Diego, CA, pp. 112–28
- Privette G, Quackenbos S, Bundrick C M 1994 Preferences for religious or non-religious counseling and psychotherapy. *Psychological Reports* **1**(2): 539–40
- Project MATCH Research Group 1997 Project MATCH secondary *a priori* hypotheses. *Addiction* **92**(12): 1671–98
- Propst R L, Ostrom R, Watkins P, Dean T, Mashburn D 1992 Comparative efficacy of religious and nonreligious cognitive-behavioral therapy for the treatment of clinical depression in religious individuals. *Journal of Clinical and Consulting Psychology* **60**: 94–103
- Razali S M, Hasanah C I, Aminah K, Subramaniam M 1998 Religious-sociocultural psychotherapy in patients with anxiety and depression. *Australia and New Zealand Journal of Psychiatry* **32**(6): 867–72
- Shafranske E P 1996 Religious beliefs, affiliations and practices of clinical psychologists. In: Shafranske E (ed.) *Religion and the Clinical Practice of Psychology*. American Psychological Association, Washington, DC, p. 52

V. A. Waldorf and T. B. Moyers

Psychotherapy, Brief Psychodynamic

Brief psychodynamic psychotherapy (BPP) describes a 'family' of psychotherapeutic interventions that share in common a short-term adaptation of psychoanalytically inspired psychotherapy. Other terms such as time-limited dynamic psychotherapy and short-term dynamic psychotherapy have also been used to describe this approach to treatment (Crits-Christoph and Barber 1991, Messer and Warren 1995).

The number of psychotherapies classified as brief psychodynamic has increased remarkably in recent years. As these numbers continue to grow, the clarity of features that constitute a psychotherapeutic system as brief dynamic becomes blurred. In an effort to provide a prototypical definition of BPP, we have gathered criteria from various authors, including those presented by Crits-Christoph et al. (1991), as follows.

(a) The theory of psychopathology explaining the cause and maintenance of a psychological disorder has a psychoanalytical basis (and thereby makes use of concepts related to such theories as Freudian, neo-Freudian, interpersonal, object relations, and self-

psychology). Among the most important constructs one should include unconscious motivations, the formative role of early years on character formation, the tendency to repeat early behavior, the pervasiveness of transference, the importance of anxiety as a signal of underlying conflict, and the role of defenses.

(b) The main treatment techniques employed by the psychotherapist are those described in the psychoanalytic literature such as making clarifications, interpretations, and confrontations and noting transference and countertransference. Despite sharing these techniques, most authors do not consider BPP a brief or condensed form of analysis but rather a distinct treatment of its own. In general, the use of advice and homework, often practiced in behavioral and cognitive psychotherapies, is not recommended.

(c) The length of treatment is planned, often predetermined, and always time-limited. It usually lasts from 12 to 40 sessions depending on the patient's problems.

(d) Because of its brevity, patients are selected for BPP. Patients have to be screened to ensure that they meet at least some of the requirements for this form of treatment. In the literature, the following requirements have often been mentioned: psychological mindedness, recognition that present problems are in part of a psychological nature, willingness to change, existence in the present or past of at least one good quality interpersonal relationship, the ability to circumscribe present complaints to one symptom, the capacity to interact flexibly with the interviewer, and adequate ego strength (including the ability to tolerate frustration and ambivalence, adequate reality testing, and the use of neurotic but not psychotic defenses).

(e) In order to maintain brevity, a focus for treatment is implemented, and issues not related to the focal point or central conflict are often considered as having secondary priority. In general, the focus of treatment is developed collaboratively either before therapy starts or immediately following the initiation of therapy. The therapist actively works to maintain the treatment focus on the agreed upon core issue.

(f) BPP is intended to alleviate specific symptoms and provide limited personality or character change. This is in contrast to psychoanalysis and long-term dynamic therapy where the goal is typically pervasive character change.

Because most brief dynamic psychotherapies do not fulfill all of the aforementioned requirements, these criteria cannot be considered necessary. However, the criteria can and should be viewed as a prototypic definition whereby the more criteria a therapy meets, the more likely it will be designated as brief psychodynamic psychotherapy.

1. Historical Background

Modern psychotherapy can be traced to Sigmund Freud's work that began in the late nineteenth century.

During that time, Freud developed a way to investigate the mind that ultimately evolved into a psychological intervention known as psychoanalysis. In psychoanalysis, the analyst attempts to bring repressed unconscious conflicts to the patient's awareness. In other words, by providing patients an understanding ('insight') of the unconscious aspects of their problems, patients have the opportunity to work them through and subsequently master these difficulties. Although Freud originally worked with patients for a small number of sessions, psychoanalytic treatment became increasingly longer over the years. Gradually, psychodynamically oriented psychotherapy, also referred to as psychoanalytically inspired psychotherapy, emerged from the modified use of psychoanalytic principles and therapeutic techniques.

Similar to psychoanalysts, long-term psychodynamic psychotherapists tend to be quiet, waiting for patients to free associate and observing how the patient-therapist relationship (or transference) develops, rather than actively pursuing psychologically meaningful material and addressing transferential material early in the treatment. Despite this similarity, however, psychodynamically oriented psychotherapy differs from traditional psychoanalysis in significant ways. For example, in contrast to psychoanalysts who see patients on the couch four to five times a week, long-term dynamic psychotherapists generally see patients face to face once or twice a week. Nevertheless, consistent with psychoanalysis, psychoanalytically inspired psychotherapy has tended to be a long-term therapy.

In addition to the long-term approach to dynamic therapy, there has also been a small group of therapists who have attempted to keep psychoanalytically inspired psychotherapy brief and concise. These clinicians continued the work originated by Ferenczi (1950), Rank (1936), and Alexander and French (1946), who recommended the maintenance of an active stance in therapy to hasten the exploration of unconscious material. In spite of their efforts, however, most psychodynamic clinicians continued to view brief dynamic therapy as inferior to the lengthier psychoanalytic treatment. It was not until Malan (1976) Mann (1973), Sifneos (1979), and Davanloo (1980) that brief psychodynamic psychotherapy was deemed a valuable treatment option. Malan's emphasis on the importance of careful patient selection through the screening of inappropriate referrals and through trial interpretations attuned therapists to the fact that BPP can be effective for a subset of the patient population. Malan and Sifneos were also among the first to stress the significance of defining and maintaining a therapeutic focus and the relevance of such focus for being an appropriate candidate for BPP. Malan explicitly illustrated some of the cornerstones of psychodynamic treatment through the descriptions of two triangles. The first triangle is the 'triangle of conflict.' Its apexes correspond to the labels 'defense,'

'anxiety,' and 'underlying feeling or impulse.' The second triangle is the 'triangle of person' and has the labels of 'relationships with current figures,' 'relationship with the therapist (representing transference),' and 'relationships with important figures from the past' (e.g., parents) for its apexes. According to Malan, the therapist's task is to expose the underlying feelings and impulses that the patient has been protecting via defense mechanisms and to elucidate the role of the defenses in reducing the anxiety that the feelings create. He posited that the patient's hidden feelings were originally experienced in relation to the parental figures at some time in the past and since then have frequently recurred with other significant figures in the patient's life including the therapist. During therapy, the patient must understand the hidden impulses underlying each of the relationships described in the triangle of person. Typically, insight into a current relationship (either with a significant other or the therapist) is achieved first and is then related back to the parental figures. Today, the writings of Malan (1976), Mann (1973), Sifneos (1979), and Davanloo (1980) are considered the four 'traditional' approaches to brief dynamic therapy (Crits-Christoph et al. 1991). Their work was the initial driving force for the development of modern BPP and fueled the debate that brief therapy can be as successful as longer-term therapies.

In the last 20 years, a new generation of brief dynamic therapists appeared on the clinical scene. Among the most important are Luborsky (1984), Horowitz et al. (1984), Strupp and Binder (1984), and Weiss et al. (1986). This new group of brief dynamic therapists can be distinguished from previous generations by their interest in the empirical status of their treatment approaches. Perhaps resulting from their interest in research, this new generation has written intricately detailed descriptions of their clinical approaches that are very helpful for training and monitoring clinicians in the adequate use of their techniques. In fact, many of these books are considered treatment manuals and have been used in empirical research addressing the efficacy of these therapeutic methods.

2. *Efficacy of BPP*

During the 1980s, researchers began to explore the efficacy of BPP relative to that of other psychotherapies. Representing the trend toward methodological rigor that has prevailed throughout the 1980s and 1990s, manualized BPPs have slowly been incorporated into research studies to provide the standardization of treatment that is needed for the studies to be credible (see *Psychological Treatment, Effectiveness of*). Despite this methodological improvement, very few high quality studies have been conducted to examine the efficacy of BPP.

Anderson and Lambert (1995) published a comprehensive meta-analysis of 26 BPP efficacy studies that indicated that although BPP is more effective than minimal or no treatment, it is neither superior nor inferior to other forms of therapy. In spite of the fact that these data indicate that BPP is an effective treatment, it is not often recognized as such. One possible reason for the lack of recognition is the small number of studies that have examined the efficacy of BPP for specific disorders. For example, only one study has examined the efficacy of BPP for depression (Gallagher and Thompson 1982) and that study only included elderly depressed patients in its subject pool. Hence it is hardly surprising that many therapists do not find the data supporting the efficacy of BPP sufficiently convincing.

In addition to the interest in examining the efficacy of brief dynamic therapy, researchers and clinicians have consistently been curious about the processes of change underlying BPP. Strong findings in the literature on process research indicate that both non-specific therapeutic factors, such as a positive therapeutic alliance, and specific therapeutic factors, such as accurate interpretations of the patient's core interpersonal conflicts, are associated with good therapeutic outcome.

3. The Future of BPP

During the last two decades, the popularity of BPP has grown. This shift can be attributed to the increased financial pressure from managed health care to shorten the average number of sessions in a treatment program, the patients' wish to receive help only for their specific complaints rather than for changing their personality, and the patients' wish to save money.

There is little reason to believe that the multiple causes influencing the trend in the late 1990s toward increased implementation of brief treatments including BPP will subside any time in the early part of the twenty-first century. Rather, it seems that efficiency and empirically demonstrated efficacy define future ideals. Consequently, it is likely that therapists will see an even larger increase in the use of brief therapy. In order to ensure that BPP will be one of the modalities recommended by the various governmental, insurance, and professional entities that are 'certifying' treatment modalities, there is an need for evidence of its efficacy for specific psychiatric disorders.

Without a doubt, ensuring that the quality of psychotherapy remains high is a major objective of all psychotherapists; however, other factors must also be considered with regard to the future status of brief dynamic psychotherapy. First, a large proportion of the training of psychiatrists and psychologists remains in the realm of long-term psychotherapy, although the trend toward training psychotherapists for short-term

interventions might be occurring. Nevertheless, increased training of clinicians in the domain of brief therapy is required for patients to receive maximum benefit from this form of treatment. Also, psychotherapists should strive to implement the BPPs designed for specific diagnostic patient populations in their treatment of these patients. This task may prove to be especially difficult since the descriptions of symptoms and syndromes of particular disorders according to psychodynamic theory do not always correspond to those listed in the most recent version of the American Psychiatric Association's *Diagnostic and Statistical Manual (DSM) of Mental Disorders*.

Modern psychotherapy, including BPP, is becoming increasingly more integrative and eclectic in its treatment approach (see *Eclectic Psychotherapy: Psychological Perspectives*). As a result of this greater eclecticism, examining the efficacy of the actual treatments used in clinical settings and defining a given treatment delivered by a specific clinician as BPP are becoming perpetually more difficult tasks. Psychotherapy, including BPP, must constantly evolve to fit with the fast paced nature and current trends of the modern world. Thus, maybe future practice of BPP will include more components from other psychotherapies.

See also: Eclectic Psychotherapy: Psychological Perspectives; Psychoanalysis: Current Status; Psychoanalysis in Clinical Psychology; Psychoanalysis: Overview; Psychological Treatment, Effectiveness of; Transference in Psychoanalysis

Bibliography

- Alexander F, French T M 1946 *Psychoanalytic Therapy*. Ronald Press, New York
- Anderson E M, Lambert M J 1995 Short-term dynamically oriented psychotherapies: a review and meta-analysis. *Clinical Psychology Review* **15**: 503–14
- Crits-Christoph P, Barber J P (eds.) 1991 *Handbook of Short-term Dynamic Psychotherapies*. Basic Books, New York
- Crits-Christoph P, Barber J P, Kurcias J S 1991 Historical background of short-term dynamic therapy. In: Crits-Christoph P, Barber J P (eds.) *Handbook of Short-term Dynamic Psychotherapy*. Basic Books, New York pp. 1–16
- Davanloo H (ed.) 1980 *Short-term Dynamic Psychotherapy*. Jason Aronson, New York
- Ferenczi S 1950 The further development of an active therapy in psychoanalysis. In: Ferenczi S, Suttie J I (trans.) *Further Contributions to the Theory and Technique of Psychoanalysis* 2nd edn. Hogarth Press, London, pp. 198–217 (original work published 1926)
- Gallagher D E, Thompson L W 1982 Treatment of major depressive disorder in older adult outpatients with brief psychotherapies. *Psychotherapy: Theory, Research and Practice*. **19**: 482–90
- Horowitz M, Marmar C, Krupnick J, Kaltreider N, Wallerstein R, Wilner N 1984 *Personality Styles and Brief Psychotherapy*. Basic Books, New York

- Luborsky L 1984 *Principles of Psychoanalytic Psychotherapy: a Manual for Supportive-Expressive (SE) Treatment*. Basic Books, New York
- Malan D H 1976 *The Frontier of Brief Psychotherapy*. Plenum, New York
- Mann J 1973 *Time-limited Psychotherapy*. Harvard University Press, Cambridge, MA
- Messer S B, Warren C S 1995 *Models of Brief Psychodynamic Therapy: a Comparative Approach*. Guilford Press, New York
- Rank O 1936 *Will Therapy*. Knopf, New York
- Sifneos P 1979 *Short-term Dynamic Psychotherapy: Evaluation and Technique*. Plenum Medical Book Co., New York
- Strupp H S, Binder J L 1984 *Psychotherapy in a New Key: a Guide to Time-limited Dynamic Psychotherapy*. Basic Books, New York
- Weiss J, Sampson H, Mount Zion Psychotherapy Research Group 1986 *The Psychoanalytic Process: Theory, Clinical Observation and Empirical Research*. Guilford Press, New York

J. P. Barber and R. Stratt

Psychotherapy: Case Study

'Case study' typically refers to the investigation of the onset and course of a disorder on the basis of personal reports. Frequently used synonyms are case study methodology, case study research, case report, and case description. The acronym ' $N = 1$ ' also pertains to case studies. Case study terminology is applied in an array of very different methodological approaches, ranging from nonstructured explorations to experimental testing of single-case hypotheses.

Description, documentation, and evaluation of a particular therapy are elements of 'controlled practice' (Petermann and Müller 2001), which in turn is based on patient-centered assessments made throughout the duration of therapy and on a well-organized documentation system. This approach allows for continuing control and optimization of the treatment process.

1. Example of a Case Study Within 'Controlled Practice'

At the outset of therapy the psychotherapist questions the patient in such areas as biography, current life situation, known problems, former experiences with psychotherapy, and the demand for therapy. The therapist documents answers by recording audiotapes and taking notes. A structured clinical interview for the purpose of classifying the disorder forms the foundation of the diagnosis. The client's current

problems and therapeutic aims are organized in what is called a functional analysis. The therapist devises a written plan for the therapy detailing, for example, time frame and partial goals. The patient and therapist then sign a contract outlining rules and guidelines to remain in effect throughout the therapy. Prior to the actual onset of therapy, psychotherapist and patient negotiate the benchmarks and indicators for a successful outcome, for instance, how frequently undesirable behaviors may occur on a daily basis. Throughout the course of therapy the patient is asked regularly to fill in a brief questionnaire on major outcome variables. Direct observation by the therapist accompanies the written information. Finally, all collected documents, including the final interview, are summarized, discussed, and reviewed in cooperation with the patient. The psychotherapist's appraisal of the intervention and its outcome is based on comparisons of several case studies, thus allowing for continuous improvement in clinical decision making.

2. Single-case: Different Methodological Approaches

Yin (1994) distinguishes between descriptive, exploratory, and explanatory case studies. Typically, in case studies questions concerning the 'why' and 'how' dominate over questions that address the 'what,' 'who,' and 'where.'

2.1 Descriptive Case Studies

Descriptive case studies contain descriptions of individual cases but dispense with interpretation. Indications for psychotherapy, the course of therapy, and its outcomes are summarized in descriptive case studies in a verbal format—case reports—which have important documentation and communication functions for therapists. The lucidity of descriptive case studies, in contrast to abstract theory, makes them particularly useful in psychotherapy education. A variant of the descriptive case study yields a detailed description of the therapeutic process. Documentation of crucial outcome parameters is a necessary condition for process monitoring and has practical implications for optimization of therapy. If documentation is kept in the foreground of therapy reporting, treatment integrity can be better controlled. 'Treatment integrity' refers to the degree to which actual therapeutic work is concordant with the therapeutic method chosen (e.g., psychoanalysis, behavioral therapy). The more precisely the method of treatment can be defined, for example as a manual (see *Clinical Psychology: Manual-based Treatment*), the more exactly the degree of concordance can be assessed.

2.2 Explanatory Case Studies

Explanatory case studies attempt to explain observable therapy-induced changes in behavior. Critical life events in the patient's biography and experimentally varied factors may serve as explanations. In general, planned (prospective) case studies must be distinguished from unplanned (mostly retrospective) case studies. Planned case studies require a clear design, standardized and replicable data collection, and statistical methods of analysis. These studies are time-consuming and labor-intensive and consequently are not used very often. This type of process research addresses the decisive procedural mechanisms in effect during psychotherapy (see *Psychotherapy Process Research*). Experimental designs are strictly controlled types of prospective analysis allowing for comparisons between suggested causes in a systematic manner (see *Single-case Experimental Designs in Clinical Settings*).

2.3 Exploratory Case Studies

Exploratory case studies are used to receive one's first impression and to generate hypotheses. Data are collected in interviews or from biographical material. Hypotheses are derived from data of the individual by developing a model of the causal and sustaining factors pertinent to a particular individual's behavior ('functional behavior analysis,' see *Behavior Analysis, Applied*). Functional analyzes may be classified as exploratory or explanatory case studies. They contain descriptions of the patient's problem(s) and statements concerning factors that trigger and maintain problem behavior(s).

3. History of the Single-case Methodology

In the early days of modern psychology, when Edward Bradford Titchener, William James, and Wilhelm Wundt rose to prominence, single-case analysis was a well-respected research strategy. Not uncommonly the researchers chose themselves as the subjects for investigation. Introspection and single-case experiments largely dominated psychological research in those pioneering times. The German philosopher Karl Jaspers introduced a special form of case study, called *Biografik* (biographic), which focused primarily on the course of individuals' lives. With the introduction of what were considered more objective data sources and analytical methods, introspection and single-case experiments—and, as a consequence, the single-case approach—were sidelined and eventually replaced. The reasons were many (Danzinger 1990). Psychology as a new field of research had to gain ground and hold out against well-established traditional disciplines such as medicine, physics, and biology. The decision to adopt an experimental methodology was made in the

hope of enhancing the reputation of psychological research. The questionnaire as an economic data collection tool was also making inroads into psychology. This methodology, adopted from the social sciences, allowed for the collection and analysis of large, standardized samples. Experiments could now be carried out involving substantial numbers of people.

The case study underwent a renaissance in the 1960s and 1970s. Awareness had arisen in psychotherapy research that certain questions could no longer be answered properly using group-based studies. Zubin (1950) had blamed the dire situation in clinical research on the lack of appropriate statistical tools for single cases, and called for the individual to be viewed as a distinctive universe. Kiesler (1966) complained about the absence of adequate research strategies in clinical psychology. He also challenged certain myths, including the 'patient uniformity assumption' (i.e., at the outset of therapy patients tend to be similar rather than different from one another) and the 'therapy uniformity assumption' (i.e., therapists treat individuals uniformly rather than variably).

New methods for the collection and analysis of data were developed in the 1960s. Goal-attainment scaling (Kiresuk and Sherman 1968), for example, opened up the possibility of collecting data that reflected individually defined outcomes. Innovations in statistics (e.g., Markoff chains and ARIMA models) now allowed for the analysis of data sampled from repeated measurements. Once again the case study was on its way to reclaim its reputation as a scientifically sound and respected methodology.

4. Single-case and Psychotherapy Research

Eysenck's (1952) provocative conclusion linking positive outcomes in the treatment of neurosis to spontaneous remissions sparked fervent clinical research seeking to prove the effectiveness of different treatments. The 'Dodo Verdict' (so called from the Dodo in *Alice in Wonderland*), according to which all therapies are equally successful, concluded the first phase of psychotherapy research. This view came under scrutiny during the second phase, when more differentiated analyzes were employed and differences in the indications for psychotherapy could be found for certain disorders. At the end of the second phase, knowledge of the actual mechanisms at work during different psychological therapies was still poor, a fact that eventually led to the third phase of clinical research, known as microanalysis. In microanalysis the interaction between therapist and patient became the primary focus. In an analogy to the use of a microscope, crucial moments during the therapy sessions were scanned with 'high methodological resolution' in order to uncover the crucial therapeutic mechanisms at work.

Gradually clinical practice research moved from a primary focus on outcomes to the model of process research. With each increase in the 'microscopic' resolution of details, single-case methodology gained in significance. Yet despite the fact that the theoretical rationale behind single-case study methodology was widely accepted in the 1980s, its impact on clinical practice remained limited. Not surprisingly, experts concluded: 'At present, clinical research has little or no influence on clinical practice' (Barlow 1981, p. 147).

5. Criticism of Single-case Methodology

Single-case methodology has always drawn criticism from different quarters. Some investigators felt that the presentation of qualitative case studies was too arbitrary; others have remained skeptical of the methodology in its entirety. Several problem areas with regard to case study research must be addressed:

- (a) objectivity in the collection of data,
- (b) degree of rigor in attempts to prove cause and effect, and
- (c) interpretation and generalizability of outcomes.

Case studies in clinical research usually deal with personal experiences (e.g., quality of life, well-being), which can be evaluated solely by means of self-reports. However, assessment of self-reported experience is problematic since any subjective report automatically compromises an assessment's validity. On the other hand, standardized measures are of limited use when tapping into a patient's subjective appraisals. This problem of control in clinical practice remains unresolved.

The validity of causality statements depends primarily on the degree to which contextual factors in data collection remain constant. Causal relationships are determined by testing hypotheses while excluding factors that may serve as alternative explanations. It is a classical problem of empirical research, however, that alternative influences can never be ruled out entirely in data analysis. While controlling for several potentially confounding variables increases internal validity on the one hand, it limits the generalizability of findings on the other. This problem is reflected in research terminology: the term 'proof' is avoided, while 'evidence' or 'tendency' are preferred, thus allowing for case-specific statements about a treatment's effectiveness while avoiding statements of more general applicability.

Given these factors, the following prerequisites must be met in order to appraise the quality of a case study:

- (a) Only validated or previously tested assessment tools should be employed,
- (b) The decision to use prospective vs. retrospective data collection must be explained, and
- (c) It must be decided to what extent contextual factors need to be controlled.

The accuracy of statements will increase with the precision and replicability of the data analysis as well as with the amount of detail provided on factors that may compromise findings.

6. Integrating Case Study and Evidence-based Clinical Practice

Legal and economic considerations increasingly require therapists to provide documentation of their interventions. Scientific evaluation is increasingly less limited to demonstrating the *efficacy* of certain therapeutic strategies and now must also show the *effectiveness* of the process of clinical practice as a whole.

The optimization of therapy serves both economic and humanitarian goals, as it is of utilitarian value for both patients and society in its entirety. In future the case study approach will form an integral part of what is known as 'evidence-based practice.' Particular difficulties include the problem of objectivity with regard to data collection and analysis, pressure on therapists to organize their interventions, and the integration of results with daily clinical practice. These demands on clinical practice control (see *Managed Care and Psychotherapy in the United States*) can only be met if data are analyzed in an exact and economical fashion. Recent software innovations may accelerate this trend.

See also: Single-case Experimental Designs in Clinical Settings

Bibliography

- Barlow D H 1981 On the relation of clinical research to clinical practice. Current issues, new directions. *Journal of Consulting and Clinical Psychology* **43**: 147–55
- Danzinger K 1990 *Constructing the Subject. Historical Origins of Psychological Research*. Cambridge University Press, New York
- Eysenck H J 1952 The effects of psychotherapy: An evaluation. *Journal of Consulting Psychology* **16**: 319–24
- Kiesler D J 1966 Some myths of psychotherapy research and the search for a paradigm. *Psychological Bulletin* **65**: 110–36
- Kiresuk T J, Sherman R E 1968 Goal attainment scaling. A general method for evaluating comprehensive community mental health programs. *Community Mental Health Journal* **4**: 443–53
- Petermann F, Müller J M 2001 *Clinical Psychology and Single Case Evidence. A Practical Approach to Treatment Planning and Evaluation*. Wiley, Chichester, UK
- Yin R K 1994 *Case Study Research. Design and Methods*. Sage Publications, Thousand Oaks, California
- Zubin J 1950 Symposium on statistics for the clinician. *Journal of Clinical Psychology* **6**: 1–6

F. Petermann and J. M. Müller

Copyright © 2001 Elsevier Science Ltd. 12475
All rights reserved.

Psychotherapy: Clinical Practice Guidelines

Since 1994, the process of delivering health and behavioral healthcare has been dramatically reshaped by the 'managed care revolution' (Strosahl 1994, 1995, see also *Managed Care and Psychotherapy in the United States; Managed Care*). Concerned by steadily escalating health care costs in the absence of convincing cost effectiveness data, both federal policy makers and the purchasers of health care have developed a variety of methods for containing health care costs. Most of these strategies (i.e., utilization review, precertification, case rates, capitation based financing) have focused on shifting the financial risk to vendors and providers, with the effect of restricting the services that are offered to consumers (Strosahl 1994). At the same time, health care reform has evolved from an overemphasis on supply side strategies (i.e., restricting access to services) to a cost and quality-oriented movement. This has stimulated a wide variety of system-level strategies for improving the quality of care while either reducing costs or holding them steady. Clinical practice guidelines evolved as one such strategy for quality improvement (Noonan et al. 1998). This article will examine the genesis of the practice guideline movement and different types of guidelines, discuss current parameters for evaluating the quality of guidelines, illustrate some of the major shortcomings of contemporary guidelines, and evaluate the empirical literature concerning the dissemination and impact of clinical guidelines.

1. The History and Context of Practice Guidelines

The most commonly accepted definition of a practice guideline is 'systematically developed statements to assist practitioners and patients make decisions about appropriate health care for specific clinical circumstances,' (Noonan et al. 1998). In this sense, practice guidelines have been developed and implemented informally at the agency and clinic level for years. Names such as treatment algorithms, clinical protocols, and practice standards have been used to describe such efforts. In contrast to contemporary practice guidelines, these early efforts were highly localized, often relied on clinical lore or the influence of a few key opinion leaders, did not systematically incorporate scientific evidence about the efficacy of clinical interventions, and addressed only the most controversial/problematic aspects of treating a patient with a particular condition. For example, many agencies and clinics have guidelines for assessing an acutely suicidal patient during the index crisis, primarily for the purpose of managing risk. At the same time, few guidelines have ever been developed for treating

suicidal behavior when it is not of crisis proportions, even though the bulk of outpatient treatment occurs with less acutely suicidal patients. Due to their parochial nature, early practice guidelines—some developed only miles apart and for the same clinical condition—differed markedly in their emphasis and clinical recommendations.

The resurgence of interest in practice guideline stems from the growing influence of outcomes-based medicine and the associated philosophies of population and evidence-based care (Noonan et al. 1998, Strosahl 1994). Based on a public health model, population-based care is concerned with the health status of the entire population, not just those patients with current psychological disorders. The goal is to build consistent processes of care that reduce variation in clinical outcomes for the population served, thereby increasing the cost effectiveness of the overall system of care. A major area of concern with both medicine and behavioral health services is that systems of care vary wildly in the types of services available and the services rendered to treat a particular condition. The goal of population care is to narrow this variability across systems and providers, ensuring better access to, and quality of, care for consumers.

The related notion of evidence-based care is noteworthy because it suggests that the best outcomes across a population are achieved by the consistent use of interventions that have been shown to work in randomized clinical trials (see *Psychological Treatments: Randomized Controlled Clinical Trials*). Many clinical experts and the purchasers of health services argue that there is excessive variation among behavioral health providers in the types and amounts of treatment that are delivered for common psychological problems. In discussing the decision-making behavior of physicians, it has been suggested that only 25 percent of health services are objectively based upon the medical status of the patient. Primary nonmedical factors related to health services include availability of needed services, patient's beliefs and preferences and the physician-patient relationship. Most providers are ill-equipped to meet the diverse array of psychosocial and medical needs posed by their patients. This 'mismatch' between what the provider can offer and what the patient expects is a major source of cost overruns in both health and behavioral health.

2. Types of Guidelines

Broadly conceived, there are two types of practice guidelines. The system guideline is widely used (Noonan et al. 1998). It addresses the parameters of population-based care by specifying who is eligible to enter the system of care (i.e., medically necessary treatment), what setting of care is appropriate (i.e., inpatient, residential, partial hospital, out-patient) and

the program to which a patient with a particular set of symptoms and functioning should be assigned (i.e., vocational rehabilitation, drugs and alcohol, dual diagnosis). Note that a system guideline functions to define *who is eligible* to receive services and *what level of care* is appropriate to address the patient's condition. It does not attempt to guide the decision making or practice patterns of providers within that level of care. However, system guidelines can and should be evidence based, evolving from the scientific literature concerning the type and intensity of treatment mostly likely to work for a patient with a given set of symptoms and specified functional status. System guidelines are currently the focus of a major developmental initiative by the Substance Abuse and Mental Health Services Administration of the United States Government. System guidelines can promote the goals of population-based care by specifying and 'minimum services menu' that is needed to properly address the diverse mental health needs of a population. System guidelines help identify gaps in programs or facilities that can then be brought to the attention of policy makers, funding agencies, and foundations. One of the chief criticisms of clinical practice guidelines by providers is that they recommend settings of care or types of programs that are not accessible. While system guidelines often receive less discussion, they are a critical component of a comprehensive population care system.

Clinical practice guidelines are a second and more widely discussed entity. Ideally, they weave scientific evidence into clinical practice (Clinton et al. 1994). There has been an unprecedented proliferation of clinical practice guidelines in the beginning of the twenty-first century. Some are discipline specific such as the series of practice guidelines published by the American Psychiatric Association (Herz et al. 1997). Others, such as the Agency for Health Care Policy Research Guideline for Treating Depression in Primary Care, focus on treatment recommendations for a particular care setting, such as primary care (Depression Guideline Panel 1993). Some guidelines, such as those published by the Academy of Adolescent and Child Psychiatry are cross-disciplinary and designed to be applied across settings (Bernstein et al. 1997). Managed behavioral health care companies have also developed proprietary guidelines as a mechanism for supporting utilization reviewers. Finally, there are a growing number of 'off the shelf' guideline packages that have been developed privately and marketed to the provider community. The promulgation of clinical practice guidelines is promising in that it testifies to the growing conviction among policy makers, administrators, and clinicians that using scientific evidence to guide clinical practice is the only way to assure consistent quality. However, there is growing concern that many existing practice guidelines are biased toward truncating care, promoting guild interests, or simply lacking scientific validity (Nathan 1998).

3. Functions of a Clinical Practice Guideline

Properly constructed, a practice guideline should provide easily accessed, scientifically and clinically sound decision support that can aid the provider and the consumer at key decision points in the assessment and treatment process (Practice Guideline Coalition 1997), including:

- (a) critical symptoms and/or functional attributes that make a patient eligible to be served under that guideline (i.e., diagnostic guidelines) and severity criteria that determine suitability for service setting (i.e., levels of care guidelines);
- (b) conditions and/or syndromes that have similar symptom patterns and methods for distinguishing these conditions (differential diagnostic guidelines);
- (c) relevant epidemiological and demographic information that helps providers identify different forms of clinical presentation related to geographic, racial, ethnic, or cultural factors;
- (d) clinical procedures shown to be effective with the condition, rates of remission when such procedures are properly administered, comparative effectiveness of different treatments, treatment side effects, and other known risks as well as the expected time frame for clinical/functional response;
- (e) commonly available and easily used self-report and/or interview methods for assessing clinical response, criteria for determining whether a procedure(s) is working and, if not, guidance in selecting an alternative procedure or treatment; and
- (f) relapse rates and maximum relapse risk periods, effective maintenance, and relapse prevention strategies.

Note that the concept of assisting the provider and consumer in selecting an appropriate intervention for a specific condition suggests that guidelines need to exist in a format that can support use by both the provider and consumer. In addition to the core clinical function of improving the quality and appropriateness of care to the individual consumer, the Institute of Medicine (Field and Lohr 1990) has identified other guideline functions, including:

- (a) decrease service costs produced by the provision of ineffective and/or unnecessary services,
- (b) assist consumers and their families in advocacy efforts,
- (c) assist payers in establishing rates and mechanisms for reimbursement,
- (d) encourage the collection of standardized data for the purpose of 'validating' a guideline, and
- (e) improve formal and post-graduate clinical training.

4. Methods for Constructing Practice Guidelines

As mentioned previously, early methods for constructing practice guidelines used either an informal or formal consensus method. Whereas informal con-

sensus might involve a group of providers agreeing to a care protocol at a lunchtime meeting, formal consensus methods require a structured development and ratification process (Woolf 1992). In the Delphi method, guideline development representatives determine rules about collecting and analyzing opinions from experts and distilling results into a set of recommendations. In iterative fashion, experts provide their opinions, then respond to the opinions of other experts until a convergence of opinion is obtained. In the Nominal Process method, a facilitator systematically elicits opinions from each group participant about the most important recommendations on each participant's list. This process continues until all recommendations from all lists have been cataloged, at which point the recommendations are distilled into a composite list. Each participant then rates this list in terms of priority, areas of maximum agreement are identified, and the resulting 'consensus' is put back to the group for ratification. The primary advantage of consensus-based methods is that guidelines can be rapidly generated, often ahead of the existing science. Innovations in clinical practice that have not been scientifically tested are more likely to be incorporated. The major, and often fatal, drawbacks are that the quality of consensus guidelines are entirely tied to the professional skills, qualifications, and biases of the group, tend to omit controversial topics where clinicians need the most guidance, produce bland recommendations that provide no advancement in clinical practice, or institutionalize locally favored practices that in fact are scientifically unsound. One of the best-known consensus-based guideline projects is the Expert Consensus Consortium, which has used the Delphi method to develop a wide variety of practice guidelines that are available nationally.

The limitations in consensus models for building guidelines have created a corresponding interest in using more objective, data-based methods. In contrast to consensus models, empirically-based methods stress that the quality of a guideline is directly related to its scientific underpinnings. For many years, the US Preventive Services Task Force has used the evidence-based guideline method. This approach involves conducting a comprehensive review of all scientific studies and reports pertaining to the condition under consideration, rating the amount and quality of the evidence, and then producing graded recommendations based upon the strength of the underlying scientific evidence. A 'Grade A' recommendation is based upon a number of well-conducted studies that meet specific and rigorous methodological criteria (i.e., sample size, randomization, well-specified, replicable interventions). In contrast, a 'Grade C' recommendation would have only limited evidence to support it, but still be seen as clinically important. Strict adherence to evidence-based methods tends to produce many fewer recommendations that are based on 'expert opinion.' However, providers and con-

sumers can be left with no guidance in very core areas because the science is not available, and this may limit the clinical usefulness of the resulting product. Another frequent criticism of evidence-based methods is that cutting-edge innovations in clinical practice are left out of guidelines because scientific validation of effective procedures may lag years behind the introduction of new procedures in a field.

Eddy (1996) proposed an Explicit Guideline Development method that involves conducting an analysis of condition base rates within a system of care, along with costs and outcomes associated with existing care processes. Using cost modeling data derived from a review of scientific data regarding alternative procedures, a balance sheet is created which outlines the projected clinical benefits, costs, and potential risks associated with implementing a guideline. Generally, this approach is designed to identify conditions that occur with a high enough frequency and/or associated cost to justify changes in a whole system of care. In essence, this method helps one determine whether a particular condition is appropriate for a clinical practice guideline. Normally, systems and agencies that employ the initial decision-making method will employ an evidence-based method for constructing the actual guideline.

5. Criticisms of Practice Guidelines

Considering that guidelines vary greatly in both structure and content, it is interesting that a consistent set of criticisms have been forthcoming from key stakeholders (Noonan et al. 1998, Practice Guideline Coalition 1997).

Consumers have expressed concerns that guidelines are used to artificially limit care, are developed with little or no consumer involvement, ignore, and/or minimize consumer preferences and limit consumer access to information critical to the treatment decision-making process.

Clinicians complain that guidelines are rigidly applied to limit services and reimbursement, are too complicated to be useful in daily practice, recommend procedures or services that the provider is not trained to administer, neglect the important role of clinical judgment, function to suppress innovative clinical practices, and promote guild or discipline specific interests.

Scientists claim many guidelines incorporate unsubstantiated expert opinion, improperly characterize or ignore the scientific literature, recommend cost-oriented interventions that produce short-term benefits only, and are not kept current with recent advances in science.

Managed health care executives complain that guidelines cost too much and take too long to develop, exceed the skill and training level of most providers,

are too guild oriented, are not 'user friendly' in daily practice, and ignore the issue of cost effectiveness.

Concerns about the varying quality of practice guidelines have led to several efforts to identify criteria for judging the quality of a guideline. The American Psychological Association (1995) has developed standards for evaluating the validity of guidelines. Generally, these standards require that guidelines incorporate valid and reproducible scientific data, be written clearly, recommend procedures that have proven clinical effectiveness in the field, allow for flexible decision making, provide clear and precise recommendations, and be 'refreshed' regularly to keep up with changes in the scientific literature. In addition, the guideline development process should be multidisciplinary and well documented.

6. Impact of Practice Guidelines

At this point in time, there are few scientific studies concerning the impact of clinical practice guidelines for behavioral health conditions. Most reports are anecdotal and suggest that properly implemented, guidelines can alter the process of delivering care particularly if provider involvement is sought. Much less is known, even at the case report level, about the extent to which practice guidelines improve clinical outcomes.

Studies of practice guidelines for physicians have shown that dissemination of a guideline is not an end in itself. Training in use of the guideline is often required, as is the 'translation' of the guideline to fit local resource and practice realities. In addition, physicians may benefit from concurrent feedback and enforcement mechanisms. Even under the best of implementation circumstances, physicians may show enhanced awareness of the guideline, yet fail to modify their medical practices (Lomas et al. 1989). In essence, medical practice guidelines have not been consistently shown to alter the process of care, improve health outcomes, or improve consumer satisfaction (Grimshaw and Russell 1993). This may be due to variations in the quality of practice guidelines, differences in dissemination and training methods, or reflect the immediate financial incentives influencing physician behavior.

7. Future Developments

It is clear that the next generation of practice guidelines will need to be less cumbersome to use, less expensive and time consuming to develop, and reflect a consensus of all the key stakeholders as to what constitutes scientifically sound, feasible, cost-effective treatment. The Practice Guideline Coalition (PGC) represents one such effort to bring the key stakeholders together

to form scientifically sound, user-friendly practice guidelines. The PGC is a coalition of approximately 80 scientific, clinical, managed care, and consumer groups, who have developed a common framework for the development, evaluation, and dissemination of practice guidelines. PGC guidelines have a very short development cycle, incorporate the views of scientists, clinicians, managed care representatives, and consumers, focus on core procedures that can be delivered by any qualified mental health provider and are designed for easy access in the context of daily practice (PGC 1997).

A second future theme will be the development of empirically validated methods for disseminating guidelines to the practitioner community. This very likely will involve the creation of training methods that are affordable and incorporate the results of early studies of guideline dissemination. The existing literature clearly suggests that it is one thing to put a guideline in front of a practitioner. It is quite another to have it impact the process or outcome of care. Dissemination and training represents perhaps the greatest challenge to the success of practice guidelines as a quality improvement tool.

Finally, there will be increasing pressure to scientifically validate practice guidelines in the field. The assumption underpinning interest in practice guidelines is that they will reduce unexplainable variations in the type of behavioral health care that is delivered, thus improving the overall quality. As suggested in the previous section, the evidence supporting this assumption is mixed, although the conceptual stance is intuitively appealing. Nevertheless, guidelines must be evaluated in field-based randomized studies that examine their impact on both processes and outcomes of behavioral health care. There is a need to demonstrate that guidelines produce their intended impact. Otherwise, the behavioral health community may be compelled to look for other methods for improving the quality of their care.

See also: Clinical Psychology: Manual-based Treatment; Managed Care; Managed Care and Psychotherapy in the United States; Medical Experiments: Ethical Aspects; Pecuniary Issues in Medical Services: Ethical Aspects; Psychological Treatments, Empirically Supported; Psychological Treatments: Randomized Controlled Clinical Trials; Psychotherapy: Ethical Issues; Psychotherapy: Legal Issues (United States)

Bibliography

American Psychological Association Task Force 1995 *Template for Developing Practice Guidelines: Interventions for Mental Disorders and Psychological Aspects of Physical Disorders*. American Psychological Association, Washington, DC

- Bernstein G A, Shaw K, Dunne J E 1997 Practice parameters for the assessment and treatment of children and adolescents with anxiety disorders. *Journal of The American Academy of Child Psychology* **36**: S49–S84
- Clinton J J, McCormick K, Besteman J 1994 Enhancing clinical practice: The role of practice guidelines. *American Psychologist* **49**: 30–3
- Depression Guideline Panel 1993 *Depression in Primary Care: Detection, diagnosis and treatment*. (Clinical practice guideline No. 5, AHCPR publication No. 93–0553.) US Department of Health and Human Services. US Public Health Service. Agency for Health Care Policy and Research, Rockville, MD
- Field M, Lohr K (eds.) 1990 *Guidelines for Clinical Practice: From Development to Use*. Institute of Medicine Committee on Clinical Practice Guidelines. National Academy Press, Washington, DC
- Eddy D 1996 *Clinical Decision Making: From Theory to Practice. A collection of Essays from SAMAs*. Jones & Bartlett, Boston
- Grimshaw J M, Russell I T 1993 Effect of clinical guidelines on medical practice: A systematic review of rigorous evaluations. *Lancet* **342**: 1317–22
- Herz M I, Liberman R P, Liberman J A 1997 Practice guideline for the treatment of patients with schizophrenia. *American Journal of Psychiatry* **154**(4): 1–63
- Lomas J, Anderson G M, Dominick-Pierre K, Vayda E, Enkin M, Hannah W 1989 Do practice guidelines guide practice? The effect of a consensus statement on the practice of physicians. *New England Journal of Medicine* **321**: 1306–11
- Nathan P E 1998 Practice guidelines: Not yet ideal. *American Psychologist* **55**: 290–9
- Noonan D, Coursey R, Edwards J, Frances A, Fritz T, Henderson M, Krauss A, Leibfried T, Manderscheid R, Minden S, Strosahl K 1998 Clinical practice guidelines. *Journal of the Washington Academy of Sciences* **85**: 114–12
- Practice Guideline Coalition 1997 *Mission Statement and Framework for Developing Practice Guidelines*. Available from Practice Guideline Coalition, Department of Psychology/296, University of Nevada-Reno, Reno, NV. 89557
- Strosahl K 1994 Entering the new frontier of managed health care: Gold mines and land mines. *Cognitive and Behavioral Practice* **1**: 5–23
- Strosahl K 1995 Three ‘gold-mine land-mine’ themes in generation of managed care. *Behavior Therapist* **18**: 52–54
- Woolf S H 1992 Practice guidelines: A new reality in medicine, Part II. Methods of developing guidelines. *Archives of Internal Medicine* **152**: 946–52

K. Strosahl and P. Robinson

Psychotherapy: Ethical Issues

When attempting to survey the landscape of ethical standards designed to deal with the practice of psychotherapy, one quickly encounters a plethora of formal codes fraught with, what at least initially appears to be, unrelated design, length, and content. Each of these codes reflects the official position taken by a professional body regarding ethical conduct on the part of the membership of their respective organiz-

ations (psychologists, psychiatrists, social workers, etc.). These differences, however, often create confusion among psychotherapists regarding just what is right or what is ethical. The problems created by these inconsistencies only become more serious when one is searching for ethical guidance on how to behave when dealing with ethical problems. Because of the inherent confusion that can frequently come from these codes, and the laws that come from them, perhaps the best way to address the topic of ethics in psychotherapy would be by turning away from the specific codes themselves to a study of the fundamental principles that form the foundation for ethical conduct in healthcare. Then, with an improved understanding of ethical theory one can move to a review of how these principles are applied to the practice of psychotherapy and to how they can give ethical guidance to the individual psychotherapist (Eberlein 1987).

A good source for information regarding ethical issues in healthcare can be found in the *Principles of Biomedical Ethics*, by Beauchamp and Childress (1994). These authors take the position that all ethical conduct in healthcare is based upon what they call a four-principled approach to biomedical ethics. They believe that the four primary moral principles that drive ethics in healthcare are: respect for autonomy, beneficence, non maleficence, and justice. Simply put, healthcare professionals are ethically bound to try to help those they treat (beneficence) while avoiding making things worse (non maleficence). In addition, healthcare practitioners should have a fundamental respect for the patient’s right to choose (autonomy) and should practice their profession in a system that is equitable and fair (justice). Other authors have argued for the addition of a fifth principle, that of fidelity, to this list (Kitchener 1984; Koocher and Keith-Spiegel 1999). As applied to healthcare, fidelity deals with the healthcare provider’s obligation to act in good faith, be honest, keep promises and to fulfill obligations to those they treat. Since these five fundamental principles are generally thought to compose the foundation of ethics in healthcare, they warrant closer scrutiny.

1. Respect for Autonomy

Respect for autonomy is tied to the belief that people have a right to hold views and make choices that reflect their own beliefs and that these beliefs do not need to be shared with, or approved of by, others. Thus, autonomy is a fundamental component of the concept of freedom and, as applied to healthcare, this freedom allows the patient, and not the provider, to control treatment.

Perhaps the best place to find evidence of the principle of autonomy as applied to the practice of psychotherapy is with informed consent to treatment. In its basic form, informed consent requires that the provider tell the patient about the risks and benefits of

a specific treatment prior to commencing the treatment itself. In addition, the person consenting must have the capacity to consent and must do so freely (Stromberg et al. 1992). In application, appropriate informed consent serves to level the field in treatment, and allows the patient to become the decision-maker when dealing with his or her well-being. It also places the obligation to fulfill the spirit of autonomy in psychotherapy upon the therapist, who must make sure the decision-maker has enough information to make an informed choice that is free from coercion.

Many authors have argued that good informed consent is not just limited to an exchange that occurs at the outset of therapy but is actually a concept that is applied throughout the treatment process (Pope and Vasquez 1998, Younggren 2000). That is, individuals who find themselves in therapy have a right to understand their treatment, its direction and any changes that occur in its direction. So as the direction of therapy shifts and as the types of interventions used change, the therapist must make sure the patient understands and agrees to these changes. Once again, it is the patient who must be in control of the decision-making process.

Satisfying the requirements of informed consent does not end with fulfilling these obligations to the patient. A good informed consent additionally requires honesty on the part of the psychotherapist regarding comprehensive treatment issues and related matters. If the information shared with the patient is not truthful or complete, then the whole structure of informed consent is threatened since one could not consent to engage in treatment without an honest explanation of that treatment by the therapist.

The requirement of honesty in informed consent raises a number of interesting questions for providers who work in today's complex healthcare system. For example, a therapist may work with a managed care company whose contract, in application, is more limited than the patient believes. Would it be ethical for this therapist not to inform his or her patients of that reality? Are not a patient's rights compromised by a therapist's decision to not tell the patient of the limitations of that insurance when, in fact, the patient might truly not choose to proceed under such circumstances? Is this purposeful lack of clarity unethical? It is clear that the principle of autonomy requires that the patient understand these types of realities prior to commencing treatment. However, fulfilling this requirement can create serious conflicts of interest for the psychotherapist because the detailed explanation of insurance limitations may be frowned upon by some managed care companies who see such conduct as noncooperative. It goes without saying the providers who are viewed as not being cooperative with the expectations of a managed care company run the risk of losing their provider contracts. All the same the ethical reality remains, that purposefully keeping realities about therapy away from a patient

truly violates one of the fundamental tenets of informed consent and that this conduct is unethical. Thus, to conform with the standards of good ethical conduct, therapists must engage in a description of psychotherapy with their patients that is not only detailed, understandable, and noncoercive, but also open, honest and free from conflicts of interest.

Also inherent in the principle of autonomy is a respect for the rights of patients to control the exchange of private information about themselves and to understand the limitations of that control. Once again, patients must, as a part of informed consent, understand how the psychotherapist is going to deal with their personal and, frequently, very private information. This includes an explanation of what information will be shared with insurance companies, other professionals and other family members, to name a few. The spirit of this principle even extends to minors whose rights, in reality, are limited legally. Psychotherapists who work with children have an ethical obligation to make sure that those children understand what is going to be shared with others and what their legal rights are. While children may not be of the legal age to consent, the philosophy of informed consent clearly demands that minors are asked to give age-appropriate informed consent and the child therapist who fails to fulfill this obligation with his or her patients violates the minor's autonomy rights.

Interestingly, the implementation of the principle of autonomy in the practice of psychotherapy does not allow the therapist to condone destructive behaviour that may result from patients' expressing their autonomy. Rights to autonomy exist in balance with other moral principles. Thus, ethical mental health professionals understand that destructive choices that damage oneself, like suicide, or damage another, like homicide or child abuse, become expressed exceptions to autonomy rights. In addition, there are settings, like prisons, where individuals have by definition had their freedom of choice restricted and, consequently, these individuals may not be allowed the full benefit of the principles of autonomy. Regardless of these types of exceptions, the general exceptions to autonomy should be clearly understood by all patients prior to beginning psychotherapy.

2. Beneficence

Fundamental to all healthcare is a belief that the provider is attempting to improve the patient's condition through the delivery of a professional service. This belief is also key to the practice of psychotherapy. Upon first assessment, the value of this principle seems so obvious that some would say it should stand without explanation. However, when applied to the psychotherapy loosely and without being carefully defined, the implementation of beneficence as a principle can become confusing and complex and can raise

many questions. These questions include the following: Is it an expression of beneficence to help a patient gain insight into a very painful, historical reality consistent with an insight-oriented model of therapy, with a risk that such insight will result in psychological distress? Where is beneficence found in an extensive, analytically-oriented psychotherapy that can have no concrete or well-defined goal? Is beneficence evident when the therapist's intentions are good but when the model of intervention has no scientific merit? The answers to the question are elusive and frequently cause extensive, passionate debate among professionals. Psychotherapists should, at a minimum, have a theoretical rationale that answers these and related questions.

Theoretical debates aside, there are some aspects of beneficence that should universally be found in the ethical practice of psychotherapy. Ethically-based psychotherapists who practice with an eye toward beneficence, practice therapies that are, at a minimum, arguably good treatments for the problems their patients bring to them. This type of therapist does not do the same type of therapy with each person who comes through the door. Implied in this is the corollary that the ethical therapist does not see the same problem in every patient he or she treats. The ethical psychotherapist plans therapy with a respect for the empirical and theoretical support that exists, showing that a specific intervention is the right thing to do for the specific problem being addressed.

Respect for beneficence also requires that the therapist engage in activities that improve the quality of their treatments. That is, they engage in consultation, continuing education and literature reviewing in order to stay current with the changing status of their profession. The ethical therapist also constantly combines the law, current literature, and community standards in using a model of intervention that has the highest probability of success and the ethical therapist is secure in his or her ability to justify their professional conduct. This style of practice assures the ethical psychotherapist that the intervention in a specific case is both consistent with the standard of care for a psychotherapist and in the best interests of the patient.

It goes without saying that practicing an outmoded model of intervention, in the face of clear evidence that another treatment could be more effective, would likely be a violation of the principle of beneficence in psychotherapy. There is little justification for a psychotherapist to apply a model of intervention that is less efficient, more costly and less likely to be successful when compared with another treatment, without a clear understanding or rationale for doing so. Once again, the reality of managed care constraints lends itself well to a discussion of the application of the ethical principle of beneficence to psychotherapy. That is, ethical psychotherapists do not allow the constraints placed upon them by insurance benefits or managed care contracts to lead them away from what

they believe would be the most effective intervention. The ethical therapist, at a minimum, must make sure that both the patient and the insurance company know that effectiveness of therapy is being administratively threatened by insurance constraints and all parties must know what the recommended treatment is. Most importantly, the ethical psychotherapist attempts to resolve any administrative insurance conflicts with a constant eye toward what is in the patient's best interest.

Finally, the ethical psychotherapist does not continue a treatment that is no longer benefiting the patient. When faced with an intervention that is leading nowhere, ethical constraints would demand a change in direction that would result in either a change in the type of therapy or, if absolutely necessary, the termination of treatment. To do otherwise is potentially exploitive of the patient, and denies the patient access to other, potentially more effective interventions: a clear violation of the principle of beneficence.

3. Non Malfeasance

It comes as no surprise that ethical psychotherapists are supposed to avoid doing bad or destructive things to and with their patients. In today's litigious society this has become a very powerful principle and is felt by some authors to take precedence over all other ethical principles (Marsh 1997, Harris 2000). It is obvious that therapy is not supposed to be a damaging process and it is the responsibility of the ethical therapist to minimize risk in order to avoid, at all costs, engaging in a process that is destructive. This prohibition against doing destructive things is very broad and not only applies specifically to the treatment aspects of psychotherapy but can also include administrative aspects of the treatment relationship.

Once again defining exactly what is destructive and damaging about psychotherapy can be quite subtle and elusive. As pointed out earlier, risk can frequently become a matter of theoretical debate among mental health professionals from differing theoretical orientations. Regardless of this lack of consistency and conceptual clarity, in general, therapists should constantly evaluate whether their interventions are potentially damaging to their patients. This is not only in the patient's best interests but also protects the therapist. However, this does not mean that ethical therapy must be void of risk, since there is probably some degree of risk in all treatments. What it does mean is that therapists do not engage in delivering treatments to patients that have risk without the patient understanding, as much as possible, the nature of those risks. If the term 'risk' comes from controversy in the field, and/or the lack of strong empirical support, it becomes all the more important that the patient receive a detailed explanation of these issues.

Non malfeasance principles would also prevent a therapist from engaging in extratherapeutic conduct with a patient that has no theoretical or empirical treatment value whatsoever, or that has even been shown to be potentially destructive. For example, a patient could never give informed consent to have sex with a therapist because research has already shown such conduct to be potentially very destructive to patients and to be of no treatment value. Even if the therapist informed the patient of these risks, agreement on the part of the patient would not reduce the serious nature of this type of ethical violation. A patient also could never give consent to allow the therapist to serve conflicting roles in a case regardless of the rationale for so doing. This prohibition against engaging in anything that is potentially destructive in psychotherapy includes patient exploitation, physical abuse, boundary violations, multiple relationships, and more. The seriousness of these types of violations is only made more obvious by the reality that this type of potentially dangerous conduct usually has nothing to do with psychotherapy and that it is frequently the result of designing the treatment relationship to meet the needs of the therapist, and not the patient.

4. Justice

The ethical principle of justice is the foundation for other very valuable healthcare concepts like fairness and equality. Justice should be evident whenever persons are due benefits from others because of their particular circumstance, as when they are ill or when they are struggling emotionally. The fulfillment of this principle is based upon the professional's respect for the fair and equal distribution of services regardless of societal differences like race, status, religious orientation, etc. (Kitchener 1984).

The principle of justice, as a concept applied to psychotherapy, takes on differing forms depending upon circumstance. Limited assets, limited time, limited staffing all are variables that have direct impact upon the implementation of justice in a treatment setting. Under some circumstances the lack of justice in the provision of professional services becomes clear; however, in others, it becomes clouded. For example, it would be obvious to most that a therapist who changes services based upon a patient's race would be unethical. In this circumstance, the lack of justice is easy to evaluate. However, what about when people cannot afford care because of poverty, societal unfairness or when benefits are limited or simply not available? How does the ethical psychotherapist pursue justice under these circumstances? Or, is the ethical therapist even responsible for addressing these types of issues when delivering care? Under these types of circumstances true justice becomes quite difficult to pursue.

In reality, things are not often just, and inequality is a part of the human circumstance. Suffice it to say that

ethical psychotherapists are those who work toward justice in treatment but are realistic in so doing. That is, they aim to do the best they can in psychotherapy, to be fairest to the persons they are treating. In fact, this is the psychotherapist's fiduciary obligation in psychotherapy.

In application, the ethical psychotherapist might pursue justice in therapy by providing a few extra termination sessions to the patient who has lost insurance benefits. He or she could offer a few pro bono visits to the patient who does not have the financial means to continue treatment or the fee for therapy could be reduced so that the patient could afford it. Or he or she could also volunteer to give time to those who do not have the means to pay for the services they receive. Regardless of the actual case, the ethical psychotherapist pursues justice in treatment and makes this one of the key principles in ethical decision-making.

5. Fidelity

The final principle of ethical decision-making is that of fidelity. Fidelity addresses a person's responsibility to be loyal and truthful in their relationships with others. It also includes promise keeping, fulfilling commitments, and trustworthiness (Welfel and Kitchener 1992). In healthcare, however, the implementation of this principle extends beyond the regular responsibilities of business or contractual fulfillment to the creation of relationship based upon trust: the trust the patient has that the professional will always operate in the patient's best interests and will always fulfill their professional obligations to operate in his or her best interests. Some authors have taken the position that fidelity, along with non maleficence, are the most legally salient moral principles (Bersoff and Koeppel 1993).

As applied to psychotherapy, fidelity forces the professional to orient toward the patient's needs and not his or her own. Thus, the therapist who allows a sexual relationship to develop in treatment would be violating this principle. Clearly such a relationship is designed to meet the therapist's needs and becomes exploitive. This principle also applies to patient abandonment since such conduct would violate the trust the patient places in the psychotherapist, trust that the therapist will see the patient through the treatment process. Finally, fidelity also applies to circumstances where professionals allow administrative factors, like benefits limitations, fee capitation, and contractual rebates to affect their treatment decisions.

Fidelity, as applied to healthcare, is a dynamic process: and the way in which a therapist maintains fidelity in treatment changes as treatment changes. Fidelity also blends with the other principles of ethical decision-making to require that the therapist be open

and honest in the treatment relationship and to make sure that the patient's expectations are consistent with what the therapist can provide. Fidelity requires psychotherapists to make the best choices they can for those that they treat.

6. Summary

The ethical psychotherapist is one who understands the fundamentals of ethical philosophy and knows how to apply them to individual cases. If a psychotherapist is capable of viewing ethical dilemmas from the perspective of ethical philosophy, decisions regarding the correct choice, if you will, become much easier to find. This approach is much more effective than a rule-based system. Rule-based guidance will never give the psychotherapist ethical security because the profession is complex and is fraught with theoretical debate and inconsistency. The search for rules, versus an understanding of principles that generate those rules, will only create a complex, awkward, and laboriously large system that would truly obscure a professional's ability to evaluate problems and to then choose the best course of action. Therefore, the answers to ethical questions can be found in the thoughtful implementation of the five principles of autonomy, beneficence, non maleficence, fidelity, and justice in psychotherapy.

See also: Assessment and Application of Therapeutic Effectiveness, Ethical Implications of; Ethics and Psychiatry; Ethics in Organizations, Psychology of

Bibliography

- American Psychological Association 1992 Ethical principles of psychologists and code of conduct. *American Psychologist* **47**: 1597–611
- Beauchamp T L, Childress J F 1994 *Principles of Biomedical Ethics*, 4th edn. Oxford University Press, New York
- Bersoff D N, Koepl P M 1993 *Ethics and Behavior* **3**: 345–57
- Eberlein L 1987 Introducing ethics to beginning psychologists: A problem solving approach. *Professional Psychology: Research and Practice* **18**: 353–59
- Harris E 2000 *Ethical Decision Making and Risk Management in Clinical Practice*. A workshop presented by the American Psychological Association Insurance Trust at the meeting of the American Psychological Association, Washington, DC
- Kitchener, K S 1984 Intuition, critical evaluation, and ethical principles: The foundations for ethical decisions in counseling psychology. *The Counseling Psychologist* **12**(3): 43–55
- Koocher G P, Keith-Spiegel P 1998 *Ethics in Psychology*, 2nd edn. Random House, New York
- Marsh, D T 1997 Ethical issues in professional practice with families. In: Marsh D T and Magee R D (eds.), *Ethical and Legal Issues in Professional Practice with families*, pp. 3–26 John Wiley, New York
- Pope, K S, Vasquez M J T 1998 *Ethics in psychotherapy and Counseling*, 2nd edn. Jossey-Bass, San Francisco

- Stromberg C D, Haggarty D J, Leibenluft R F, McMillan M H, Mishkin B, Rubin B L, Trilling H R 1988 *The Psychologist's Legal Handbook*. The Council for the National Register of Health Service Providers in Psychology, Washington, DC
- Welfel E, Kitchener K 1992 Introduction to the special section: Ethics education—an agenda for the 90s. *Professional Psychology: Research and Practice* **23**: 179–81
- Younggren J N 2000 Is managed care really just another unethical Model T? *The Counseling Psychologist* **28**(2): 253–62

J. Younggren

Psychotherapy, History of: Psychiatric Aspects

Psychotherapy—the systematic application of psychological principles to accomplish symptomatic or more substantial personality change—has its origins in two healing traditions—the magico-religious and the medico-scientific (Bromberg 1959). The idea that human experience is influenced by supernatural forces and that certain people have the power to intercede with these forces dates back to antiquity. Shamans and sorcerers have resorted to amulets, magical potions, and incantations for centuries. They have also conducted exorcism and induced altered states of consciousness.

With the Age of Reason, interest was aroused in the newly discovered phenomena of electricity and magnetism. Anton Mesmer (1734–1815), a brash physician, drew on the latter when propounding his theory of animal magnetism. Since the body contained 'magnetic fluid,' and this could become disturbed, by applying magnets to various parts a patient could be relieved of pain and other ailments. While Mesmer's work was ultimately denounced by a Royal Commission and he was forgotten, the observation that his methods appeared to help patients continued to attract attention. James Braid (1795–1860), a Manchester doctor, noted that he could induce a similar trance-like state by getting people to fix their gaze on a luminous object. Labeling the phenomenon hypnosis, he proceeded to demonstrate its effectiveness in treating a range of conditions.

The doyen of French neurology, Jean Martin Charcot (1835–93), considered hypnosis a neuro-physiological process, and set about its study in states like somnambulism and hysteria. His pupil Pierre Janet (1859–1947) suggested that the latter was brought about by weakening of a higher brain function, resulting in a constriction of consciousness. In this 'dissociated' state thoughts could not be integrated, and symptoms were beyond the reach of consciousness.

In an alternative explanation, Ambroise Liebeault (1823–1904), and Hippolyte Bernheim (1840–1919) in France posited a narrowing of attention in hypnosis

which rendered the patient vulnerable to the therapist's influence by suggestion. Far from invalidating hypnosis, the power of suggestion could be investigated scientifically and accepted as a bona fide therapy.

Freud (1856–1939) had studied briefly with Charcot in 1885–6 and became enthusiastic about new possible treatments for hysteria. Freud's mentor, a respected Viennese physician, Josef Breuer (1842–1925), had described the treatment of a young woman who suffered from an array of marked hysterical symptoms. Instead of suggesting disappearance of the symptoms, Breuer had encouraged Anna O (the pseudonym given to her) to talk freely about her life under hypnosis. She did so and over time began to share memories about her father's illness. Breuer noted that after each session, his patient became less distressed and her symptoms improved. Anna herself referred to this approach as her 'talking cure' or 'chimney sweeping'; Breuer termed it 'catharsis.' Whereas Charcot had described physically traumatic events as possibly causal, in his hysterical patients, Breuer and Freud's interest focused on psychological trauma such as humiliation and loss. Freud also became less keen on applying hypnosis.

Their volume, *Studies in Hysteria*, which included the case of Anna O, and four others treated by Freud himself, also contained ideas on defense mechanisms and an account of various techniques including suggestion, hypnosis, and catharsis. Most importantly, however, was the new concept of 'free-association': the patient was instructed to disclose whatever came into their mind, without any censoring, while the analyst adopted an attitude of 'evenly-suspended attention' in listening to and interpreting the material.

On the basis of his subsequent self-analysis and further clinical experience, Freud elaborated his theory of infantile sexuality. Sexual fantasies in young children centered around the triangular relationships of love and rivalry with their parents (the Oedipal complex). Promoting free association of this theme, as well as of dreams, slips of the tongue, and other unconsciously-based thoughts, feelings, and fantasies, Freud emphasized the centrality of the complex in the neuroses. While catharsis and insight (making the unconscious conscious) seemed pivotal to therapy, Freud soon realized that other factors also operated, particularly the role of transference, i.e., how feelings, thoughts, and fantasies stemming from childhood experiences and revived in current relationships are 'transferred' onto the analyst.

Interpretations were directed at this threefold nature of the analytic experience. The therapist's response to transference (countertransference) led to analysts having their own personal therapy, the aim being to minimize it, and thus to be thoroughly objective in attending to the patient's disclosures (Ellenberger 1970).

In summary then, psychoanalytic psychotherapy elaborated the following processes: the patient's un-

bridled disclosure of whatever enters his or her mind—free association; the transference of infantile and childlike feelings and attitudes to the therapist which were previously directed to key figures in the patient's earlier life; interpretation of the transference as well as of defenses the patient applies to protect him/herself and the resistance he/she manifests to self-exploration; and finally, the repeated working through of the discoveries made in treatment. The ultimate aim is insight with translation into corresponding change in behavior and personality.

1. Variations on a Theme

Freud revised his theories at numerous points in his long professional life. In addition, many colleagues extended the boundaries of psychoanalytic thinking. But differences of opinion soon surfaced, some radical. The nature of psychoanalysis then, especially Freud's intolerance of dissent, led notable figures to leave the fold and to evolve their own models on which corresponding schools of therapy were established. Jung and Adler were two foremost European dissidents; Karen Horney, Erich Fromm, and H. S. Sullivan were pioneering neo-Freudians in the US; and W. D. Winnicott and Melanie Klein were prominent in Britain (Mitchell and Black 1995).

Carl Gustav Jung is perhaps the most critical dissenter since he was clearly being groomed to take over leadership of the psychoanalytic movement. But he was also celebrated because of his own contribution. Specifically with regard to psychotherapy, he advanced the notion of 'individuation'—the aim was to discover all parts of oneself and one's creative potential. Jung was less concerned than Freud with the biological roots of behavior, especially infantile sexual development, but rather emphasized social and cultural factors. The role of transference was replaced by a more adult type of collaboration.

Furthermore, the unconscious for Jung is not merely the repository of the individual's history but also a wider social history, a phenomenon Jung labeled the collective unconscious. He arrived at this notion through a study of myths, legends, and symbols in different cultures, in different epochs. That these are shared by a variety of cultures is not fortuitous but reflects cosmic mythical themes or archetypes, a salient feature of man's collective history.

Alfred Adler (1870–1937), a Viennese physician, was like Jung a prominent dissenter. Adler joined Freud's circle in 1902 but broke away to form his own school of 'individual psychology' after nine years. An important tenet concerns individual development. We begin life in a state of inferiority, weak, and defenseless, for which we compensate by striving for power and by evolving a lifestyle to make our lives pur-

poseful. The pattern that emerges varies and may include such goals as acquisition of money, procreation, high ambition, or creativity. The drive for power and choice of lifestyle may, however, go awry in which case a neurosis results. A path is followed which leads to ineffective efforts to cope with the feeling of inferiority and assumption of a facade or false self.

Adler regarded therapy as a re-educative process in which the therapist, who serves as model and source of encouragement, engages in a warm relationship with the patient and enables him or her to understand the lifestyle he/she has assumed. Unconscious determinants of behavior are less crucial than conscious ones and the term 'unconscious' is used only descriptively to refer to aspects of the person that are not understood by him.

Freud's insistence on innate drives and infantile sexuality not only led to the schisms with Jung and Adler but also spurred a new generation of analytically oriented analysts to concentrate on interpersonal aspects of psychological experience. In the US, Sullivan pioneered the school of 'interpersonal psychiatry,' in which the therapist adopted the role of participant observer in treatment, the transference providing one opportunity to explore communication and its breakdown in the patient's interpersonal world.

We can consider the chief features in the approach of the neo-Freudians by looking briefly at Karen Horney (1885–1952). She became disenchanted with Freud's rigid focus on instinctual biological factors, arguing that cultural factors were more salient, as reflected in the differences in psychological development and behavior among different sociocultural groups. Indeed, behavior regarded as normal in one culture could be viewed as neurotic in another.

In line with her emphasis on culture, Horney advanced the role of parental love in the life of a young child. Children typically suffer anxiety, a consequence of feeling helpless in a threatening world. The child reared in a loving atmosphere succeeds in overcoming basic anxiety. By contrast, the deprived child comes to view the world as cruel and hazardous. An inevitable result is low self-esteem.

The task of therapy is to examine the patient's defective patterns in relating to others. In part this is achieved through study of what transpires between patient and therapist. But there is no emphasis on transference as occurs in Freudian analysis. Therapy aims to enable the patient to move *with* others, by engaging in relationships which are reciprocal and mutual. Another goal is greater self-realization, with freedom from determined modes of thought and action.

While the neo-Freudians were establishing a new pathway for psychoanalysis, a group was equally innovative in the UK and France. In the UK, Melanie Klein (1882–1960), a physician hailing from Berlin, concluded from her study of children's play that the Oedipal complex, as described by Freud, was a late

expression of primitive unconscious anxieties which children experienced in the first 18 months of life. Given that these states could be reactivated in adult life, Klein labeled them 'positions' rather than stages of psychological development. She also supported the move away from Freud's formulation of instincts and psychosexual energy toward one of object relations, namely the representations of the perception and experience of significant others. Similarly, transference was to be understood not only as a repetition of past or current relationships, but as evidence of relational patterns in the patient's current *internal* world.

In tandem with Klein, other analysts in Britain proposed ways in which significant early relationships, particularly between infant or young child and its mother, exerted a formative influence on the developing psyche. John Bowlby (1907–90), in particular, observed that children who had suffered a prolonged separation from their mother underwent a grief experience which predisposed them to a range of psychopathology in later life. Donald Winnicott (1896–1971), a pediatrician and analyst, proposed that a mother's attunement to her child, enabling him to feel psychologically 'held' and understood, was an essential ingredient for a stable sense of self. The American psychoanalyst Heinz Kohut advanced a similar concept, but highlighted the place of empathy whereby the mother functions as a 'self object' for the child.

Wilfred Bion, a student of Klein, suggested that the analyst functions as a 'container' into whom the patient projects their 'unprocessed' feelings and experiences; the analyst returns them later in a form accessible to him to work on. This experience is deemed to be crucial to the development of a person's ability to examine his or her own state of mind and that of others.

Michael Balint, an analyst at the Tavistock Clinic, worked along similar lines to Winnicott, focusing on the opportunity for the patient to 'regress' in the safety of the therapeutic relationship to a state of mind in which differentiating between patient and therapist may blur. The features of the encounter—its regularity, predictability, and the therapist's empathic attitude—creates circumstances for such regression which can then be used to examine and alter defense mechanisms.

These models have implications for practice in that the therapist expresses moment-to-moment empathic understanding of the patient's inner experiences and as this is repeated time and again the process promotes the evolution of a coherent identity.

Jacques Lacan (1901–81), the maverick French analyst, claimed a 'return to Freud,' his early work in particular, but through the prisms of semiotics and linguistics. These disciplines, he contended, do not see language as a value-free means to express ideas and convey meaning but as a system of signs that communicate symbolically tacit rules and power arrange-

ments. Language imposes on the child awareness of a separateness from its mother who represents a fantasized internal object (the imaginary mother), in whose eyes the child mirrors itself in an attempt to achieve a state of psychological unity. Lacan's controversial proposals regarding variable duration of treatment and trainees themselves determining their competence to qualify as psychoanalysts led to his expulsion from the International Psychoanalytic Association, albeit with a dedicated following.

Humanistic experientially oriented schools of therapy became popular in the 1960s and 1970s, mainly in the US. Their chief features were: challenging patients with the way they avoid emotionally significant matters in the 'here-and-now'; increasing awareness of nonverbal aspects of communication; facilitating emotional arousal; and providing a forum in which patients were encouraged to experiment with new behaviors.

Some of these schools have incorporated aspects of psychoanalytic theory. Transactional analysis, for instance, makes playful use of the concept of ego states, represented as parent, adult, and child; primal scream therapy is predicated on heightened arousal of the purported emotional trauma of birth; and psychodrama encourages people to enact their conflicts by personifying various roles in their psychological lives.

The most influential humanistic school has been the client-centered therapy of Carl Rogers, although it has been absorbed into psychotherapy generally and not always with appropriate acknowledgment. Its premise is that if the therapist creates an atmosphere that is nonjudgmental, empathic, and warm, people can realize their potential for self-expression and self-fulfillment.

Feminist-based psychotherapies, still evolving since the 1980s, reflect a range of opinion from those which condemn Freud as a misogynist who deliberately recanted his discovery of the actuality of childhood sexual abuse, to those who appreciate his linkage of life experience, innate drives, and corresponding defense mechanisms in influencing vulnerability to psychopathology (Elliot 1991, Appignanesi and Forrester 1992). The latter point of view, echoing a broader debate within psychoanalysis of the respective roles of internal and external reality, has led some feminist therapists to stress social reality, and others to stress physical reality.

2. The Development of Other Modes of Psychotherapy

Our focus, hitherto, has been on psychotherapy of the individual patient. The second half of the twentieth century saw the evolution of treatments conducted

with more than one person: groups of strangers and families. In this section, we provide a brief account of the development of these two modes.

Although therapy of stranger groups was experimented on early in the century, it was not until World War II that major developments ensued. The exigencies of war were the spur in that the group format proved highly economical for dealing with huge numbers of soldier-patients. The Northfield Military Hospital in the UK was a center of remarkable innovation, led by psychoanalytically oriented therapists like Wilfred Bion and S. N. Foulkes. Bion's (1961) influence pervaded subsequent work at the Tavistock Clinic, while Foulkes (1965) was the founding father of the Institute of Group Analysis (both in London).

In the US, the move to the study of group process was spearheaded by social psychologists, particularly Kurt Lewin. A national center was established to train people in human relations and group dynamics. Participants from diverse backgrounds studied group functioning in order to act more effectively in their own work settings. Paralleling the human potential movement in the 1960s, the group movement transferred its focus from group to personal dynamics, and before long the encounter group movement had evolved with its thrust of promoting greater self-awareness and personal growth. Encounter groups became widespread during the 1960s and 1970s but after an initial fervor, declined both in terms of membership and appeal.

Formal group therapy under professional leadership then assumed a more prominent role. The most popular model was fathered by Irvin Yalom who drew upon an interpersonalist approach, one originally molded by Harry Stack Sullivan who proposed that personality is chiefly the product of a person's interaction with other significant people. Yalom pioneered the study of therapeutic factors specific to the group rather than following the pattern of transposing psychoanalytic or other theories from the individual setting. His *Theory and Practice of Group Psychotherapy* (Yalom 1995) became exceedingly influential, attested to by the appearance of four editions.

Constrained perhaps by Western medicine's focus on the individual patient, psychiatry was slow to develop an interest in the family (Gurman and Kniskern 1991). Scattered through Freud's writings are interesting comments about family relationships and their possible roles in both individual development and psychopathology. His description of processes like introjection, projection, and identification explained how individual experiences could be transmitted across the generations in a family. Influenced by the work in the UK of Anna Freud, Melanie Klein, and Donald Winnicott, the child guidance movement developed a model of one therapist working with a disturbed child and another with the parents, most often the mother on her own. The two clinicians

collaborated in order to recognize how the mother's anxieties distorted her perception and handling of her child, which were added to the child's own anxieties.

Things took a different turn in the US. There, Ackerman (1958) introduced the idea of working with the family of a disturbed child using psychoanalytic methods in the 1950s. An interest in working with the family, including two or more generations, arose concurrently. Thus, Murray Bowen (1971) found that the capacity of psychotic children to differentiate themselves emotionally from their families was impaired by the consequences of unresolved losses, trauma, and other upheavals in the lives of parental and grand-parental generations.

Boszormenyi-Nagy and Spark (1984) also addressed this transgenerational theme by describing how family relationships were organized around a ledger of entitlements and obligations which conferred upon each participant a sense of justice about their position. This, in turn, reflected the experience in childhood of neglect or sacrifices for which redress was sought in adult life.

Bowen also introduced the principles of Systems Theory into his work but it was Salvador Minuchin, working with delinquent youth in New York, who highlighted the relevance of systems thinking to their interventions. The youngsters often came from emotionally deprived families, headed by a demoralized single parent (most often the mother) who alternated between excessive discipline and helpless delegation of family responsibilities to a child. Minuchin's Structural Family Therapy deploys a series of action-oriented techniques and powerful verbal metaphors which enable the therapist to join the family, and to re-establish an appropriate hierarchy and generational boundaries between the family subsystems (marital, parent/child, siblings).

Another major development took place in Palo Alto, California, where a group of clinicians gathered around the anthropologist Gregory Bateson (1972) in the 1950s. They noted that implicit in communication were *meta-communications*, which defined the relationship between the participants. Any contradiction or incongruence carried great persuasive, moral, or coercive force and formed part of what they labeled a 'double-bind'; they proposed this as a basis for schizophrenic thinking.

All these system-oriented views assume that the family is a system observed by the therapist. However, therapists are not value-neutral. They may take an active role in orchestrating change in accordance with a model of family functioning. Yet these models ignore therapists' biases as well as the relevance of their relationships with families. This probably reflected the determination of some American family therapists to distance themselves from psychoanalytic theory, and also led them to neglect the family's history, how it changed throughout the lifecycle, and the significance of past traumatic and other notable life events.

In response to these criticisms there was a move away from the problem-focused approach which had characterized most behavioral and communication views of psychopathology. The Milan school (1980) whose founders were psychoanalysts developed a new method of interviewing families in conjunction with observers behind a one-way screen formulating and then presenting to the family and therapist hypotheses about 'their' system.

Family therapists also began to consider that families might be constrained from experimenting with new solutions because of the way they had interpreted their past experiences or internalized explanatory narratives of their family, the expert's, or society at large. This led to a shift from considering the family as a social to a linguistic system.

The narrative a family conveys about their lives is a construction which organizes past experience in particular ways; other narratives are excluded. When a family with an ill member talks to a professional, conversations are inevitably about problems (a problem-saturated description). The participants ignore times when problems were absent or minimal, or when they successfully confined problems to manageable proportions. A different story might be told if they were to examine the context that might have led, or could still lead, to better outcomes.

A number of narrative approaches have applied these concepts. Philosophically, they align themselves with postmodernism, which challenges the notion of a basic truth known only by an expert.

Many criticisms of systems approaches have been leveled including:

(a) Disregard of the subjective and intersubjective experiences of family members.

(b) Neglect of their history.

(c) Denial of unconscious motives which affect relationships.

(d) Although family members are reciprocally connected, the power they exert on one another is not equal; this is highlighted in the violence against women and in child abuse.

(e) Inequality and other forms of injustice based on societal attitudes toward differences in gender, ethnicity, class, and the like, are uncritically accepted as 'givens'.

(f) Minimizing the role of the therapeutic relationship, including attitudes members develop toward the therapist and her feelings toward them.

This critique has led to growing interest in integrating systems-oriented and psychoanalytic ideas, particularly those derived from object-relations theory. One variant is John Byng-Hall's (1995) synthesis of attachment theory, systems-thinking, and a narrative approach.

A further criticism of systems-oriented approaches is that they minimize the impact of external reality such as physical disability or biological factors, in the etiology of mental illness, and sociopolitical phenom-

ena like unemployment, racism, and poverty. One result is the 'psycho-educational' approach, which has evolved in the context of the burden schizophrenia places on the family and the potential for its members to influence its course. This has led to a series of interventions including educating the family about the nature, causes, course, and treatment of schizophrenia; providing them with opportunities to discuss their difficulties in caring for the patient, and to devise appropriate strategies; and helping them to resolve conflict related to the illness, which may be aggravated by the demands of caring for a chronically ill person.

3. *The Scientific Era*

Systematic research in the psychotherapies was a low priority for many decades; instead, practitioners' interests focused on theory and technique. Investigation of the subject only took off in earnest in the early 1950s. A notable impetus was the critique by H. J. Eysenck (1952) in which he argued that the treatment of neurotic patients was no more effective than no treatment at all. Two-thirds of both groups showed improvement with time. In later reviews, Eysenck was even more damning: the effects of psychoanalysis in particular were minuscule. The attack led to much rancor and an extended battle between the psychoanalytic and behavioral camps (Eysenck featured prominently in the latter). Fortunately, a positive repercussion was the sense of challenge experienced by the analytic group. They had been riding high for several years, particularly in the US, and barely questioned whether their concepts and practice required scientific appraisal. Thus, although Eysenck's interpretation of the research literature was flawed and biased, he had stirred a hornet's nest.

Since the 1950s research has burgeoned and yielded much knowledge about whether psychotherapy works (outcome research) and how it works (process research). These developments were not without incident. Many therapists challenged the relevance of conventional research methodology to psychotherapy. A key argument was that the encounter between therapist and patient is unique, involving two people in a complex collaborative venture, and cannot be subject to the same form of scrutiny as occurs in the natural sciences. Moreover, the latter approach is necessarily mechanistic and reductionistic.

There is merit to the argument but actual research practice reveals that methodology can accommodate, at least in part, respect for the highly personal nature of psychological treatment. Sophisticated statistical procedures can also contribute to maintaining an appreciation that many characteristics in both patient and therapist are relevant, and they cannot be viewed as homogeneous groups. An illustration is the multi-

dimensional measurement of outcome. Instead of restricting this to one or two variables only, several can be examined concurrently which together encompass the patients' internal and interpersonal world, e.g., quality of life, self-awareness, authenticity, self-esteem, target problems, and social adjustment.

Another criticism of outcome research is examining effectiveness without adequate treatment being given. Duration of research-bound therapy in fact is often much briefer than in customary practice because of funding constraints. A similar objection relates to the follow-up period usually set. A robust test of effectiveness entails examining outcome one or more years following the end of therapy in order to judge clinical projects independent of the therapist.

4. *The Effectiveness of Psychotherapy*

Several reviews have been conducted since the 1970s. A sophisticated method—meta-analysis—which relies on a rigorous statistical approach, points to psychotherapy overall generating benefits for patients who have been suitably selected. Encouragingly, better designed studies show positive results more commonly than inadequate ones. In one trail-blazing meta-analysis, levels of effectiveness were computed from 375 studies in which some 25,000 experimental and 25,000 controlled patients were tested (Smith et al. 1980). The average patient receiving psychological treatment of one kind or another was better off than 75 percent of those who were untreated—a clear demonstration of the beneficial effects of treatment in general.

With tens of 'schools' available, many of them claiming a distinctive approach, the obvious question arises as to whether some are superior in producing benefit. This question of comparative effectiveness is complicated in that not only may 'schools' be compared but also specific procedures such as setting a time limit or not or conducting therapy individually or within a group. Notwithstanding these difficulties, considering research on comparative effectiveness is worthwhile. Several reviews show a consistent pattern—that 'everyone has won and must have prizes.' This was of course the judgment of a race handed down by the dodo bird in *Alice in Wonderland*. In the psychotherapy stakes, it appears that everyone has won too, a finding probably attributable to factors common to all therapeutic approaches. These factors, set out originally by Jerome Frank (1973), include a confiding relationship with a helping person, a rationale which provides an account of the patient's problems and of the methods for remedying them, instillation of hope for change, opportunity for 'success experiences' during treatment, facilitating emotion in the patient, and providing new knowledge, so

promoting self-awareness. These 'nonspecific' factors comprise a significant theme in psychotherapy; they probably serve as a platform for benefits from all treatments.

Any consideration of effectiveness brings up the issue of harm. In other words, treatment may be for better or for worse. After a long gestation, the concept of a 'negative effect' attracted widespread attention from the 1970s (Hadley and Strupp 1976). A growing sense of confidence perhaps permitted therapists to be more open to a potential harmful impact.

To distinguish between a patient becoming worse because of their intrinsic condition or following an adverse life event, a negative effect has been defined as deterioration directly attributable to treatment itself. Such a causal link is difficult to prove but genuine negative effects certainly do occur. The definitional difficulty, however, leads to estimates of prevalence ranging from rare to common. The type of therapy appears to influence the rate. Evidence points to worsening in about 5–10 percent of cases in psychotherapy generally. The reasons for deterioration are not well established although a common view is that patient factors probably contribute, particularly selecting people for treatments for which they are unsuited. In this situation, refining assessment, thus enhancing clinicians' ability to predict response to specific treatments, is deemed as helpful to reduce the casualty rate.

Another obvious facet is the therapist and/or technique. Here, inadequate training has emerged as salient, with poor judgment leading to inappropriate interventions. It has also been recognized that a proportion of therapists perform poorly because of their personality traits, no matter what their level of training or experience. In other words, the role of therapist does not suit all those who wish to practice. In a noteworthy study of different forms of encounter group, a substantial percentage of 'casualties' were produced by four leaders, who typically pummeled members to disclose personal information and express intense emotion.

During the half century of systematic research, attention has also been devoted to the processes that take place in the therapeutic encounter.

Process and outcome research are inevitably linked. If we can identify factors that promote or hinder effectiveness, we may be able to modify these and then note the result. The advent of recording techniques such as audio and video has facilitated studies of process and a wide range of observations of both verbal and non-verbal behavior, in both therapist and patient and their interaction, have been made. One illustration is the premise that a necessary condition for effective group therapy is group cohesiveness. It follows that anything enhancing cohesiveness could be advantageous. Compatibility between members has been proposed as pertinent and indeed been shown to relate to cohesiveness. Group therapy therefore could

conceivably be more effective if, in selecting members, compatibility were taken into account. For example, a patient incompatible with all his peers would presumably not be placed in that group.

Research on process has been shown to be as salient as that on outcome since only with the study of what occurs in treatment can therapists appreciate its inherent nature and the factors for optimal improvement of participants. Diligent process research has provided solid foundations for establishing hypotheses about outcome.

We have dealt in general terms with the theme of psychotherapy research. The second half of the twentieth century has been a fertile period and seen much achieved; to do it justice is beyond our remit. The interested reader is recommended to consult the four editions of the *Handbook of Psychotherapy and Behavior Change* (Bergin and Garfield 1994) which has served the field in distilling work done and providing a critique of its quality.

See also: Bowlby, John (1907–90); Freud, Sigmund (1856–1939); Janet, Pierre (1859–1947); Jung, Carl Gustav (1875–1961); Klein, Melanie (1882–1960); Psychiatry, History of; Psychoanalysis, History of; Psychological Treatment, Effectiveness of; Psychological Treatments, Empirically Supported; Psychotherapy and Religion

Bibliography

- Ackerman N W 1958 *The Psychodynamics of Family Life*. Basic Books, New York
- Appignanesi L, Forrester J 1992 *Freud's Women*. Basic Books, New York
- Bateson G 1972 *Steps to An Ecology of Mind*. Chandler Publishing Co., San Francisco
- Bergin A E, Garfield S L (eds.) 1994 *Handbook of Psychotherapy and Behaviour Change*, 4th edn. Wiley, New York
- Bion W 1961 *Experiences in Groups*. Tavistock, London
- Bromberg W 1959 *The Mind of Man: A History of Psychotherapy and Psychoanalysis*. Harper, New York
- Boszormenyi-Nagy I, Spark G M 1984 *Invisible Loyalties: Reciprocity in Intergenerational Family Therapy*. Brunner-Mazel, New York
- Bowen M 1978 *Family Therapy in Clinical Practice*. Jason Aronson, New York
- Byng-Hall J 1995 *Rewriting Family Scripts. Improvisation and Systems Change*. Guilford, New York
- Ellenberger H 1970 *The Discovery of the Unconscious*. Basic Books, New York
- Elliot P 1991 *From Mastery to Analysis, Theories of Gender in Psychoanalytic Feminism*. Cornell University Press, Ithaca, NY
- Eysenck H J 1952 The effects of psychotherapy. *Journal of Consulting Psychology* 16: 319–24
- Foulkes S H, Anthony E J 1965 *Group Psychotherapy: The Psychoanalytic Approach*. Penguin, Harmondsworth, UK
- Frank J D 1973 *Persuasion and Healing: A Comparative Study of Psychotherapy*. Johns Hopkins University Press, Baltimore, MD

- Gurman A S, Kniskern D P (eds.) 1991 *Handbook of Family Therapy, Vol. II*. Brunner/Mazel, New York
- Hadley S W, Strupp H H 1976 Contemporary views of negative effects in psychotherapy. *Archives of General Psychiatry* **33**: 1291–302
- Mitchell S A, Black M J 1995 *Freud and Beyond: A History of Modern Psychoanalytic Thought*. Basic Books, New York
- Selvini-Palzzoli M, Boscolo L, Cecchin G, Prata G 1980 Hypothesizing-circularity-neutrality: Three guidelines for the conductor of the session. *Family Process* **19**: 3–12
- Smith M L, Glass G V, Miller T I 1980 *The Benefits of Psychotherapy*. Johns Hopkins University Press, Baltimore, MD
- Yalom I D 1995 *The Theory and Practice of Group Psychotherapy*, 4th edn. Basic Books, New York

S. Bloch and E. Harari

Psychotherapy Integration

As an international movement, psychotherapy integration represents a revolutionary departure from the discipline's historical organization around single and competing theories. Integration is characterized by a desire to look across and beyond rival theories to see what can be learned—and how patients can benefit—from multiple ways of practicing psychotherapy. Although the movement faces several obstacles and has not yet been convincingly supported by empirical research, some form of integration will probably constitute the future of the discipline.

1. The Historical Context

The practice and training of psychotherapy has been historically organized around distinct and competing theoretical orientations, such as psychoanalysis, humanistic persuasions, behavior therapies, and family systems models. In the infancy of the field, therapy systems, like battling siblings, competed for attention and affection in a 'dogma eat dogma' environment. Clinicians traditionally operated from within their own particular theoretical frameworks, often to the point of being blind to alternative conceptualizations and potentially superior interventions.

Although this 'ideological cold war' may have been a necessary developmental stage for the discipline, psychotherapy since the mid-1980s has witnessed a decline in ideological struggle and the stirrings of rapprochement. The theoretical substrate of each system is undergoing intensive reappraisal, as psychotherapists acknowledge the inadequacies of any one system and the potential value of others. Distinctive of the present era is the tolerance for and assimilation of formulations that were once viewed as deviant, even heretical.

2. Integration Defined

As the field of psychotherapy has matured, integration has emerged as a dominating force. The dramatic growth of interest in psychotherapy integration during the 1990s has crystallized into a formal movement, or more dramatically, a 'metamorphosis' in mental health (London 1988).

Psychotherapy integration is characterized by a dissatisfaction with single-school approaches and a concomitant desire to look across and beyond school boundaries to see what can be learned—and how patients can benefit—from other ways of thinking about psychotherapy and behavior change. The ultimate outcome of doing so, not yet fully realized, is to enhance the effectiveness, efficiency, and applicability of psychotherapy.

Any number of indicators attest to the unprecedented interest in psychotherapy integration. Eclecticism or the increasingly favored term 'integration' is the modal theoretical orientation of English-speaking psychotherapists and is gradually increasing in prevalence throughout Europe. Leading psychotherapy textbooks increasingly identify their theoretical persuasion as integrative, and an integrative chapter is routinely included in compendia of treatment approaches. Handbooks on psychotherapy integration have appeared throughout the world, including several in Europe and two in the USA (Norcross and Goldfried 1992, Stricker and Gold 1993).

As with many relatively new movements, integration has been plagued by confusing vocabulary and multiple goals. Various labels are applied to the movement—convergence, rapprochement, treatment selection, combined therapies, prescriptionism, differential therapeutics, and eclecticism, among them—but *psychotherapy integration* has emerged as the overarching consensual term. Similarly, a plethora of integrations have been advanced: combining different therapy formats/modalities (such as individual and family therapy); blending self-help and psychotherapy; integrating pharmacotherapy and psychotherapy; synthesizing research and practice; integrating Occidental and Oriental perspectives; and so on. This article confines itself to the traditional and primary meaning of integration as the blending of diverse theoretical orientations or schools of psychotherapy.

3. Reasons for Recent Growth

Integration as a point of view has probably existed as long as philosophy and psychotherapy. In philosophy, the third-century biographer, Diogenes Laertius, referred to an eclectic school that flourished in Alexandria in the second century AD. In psychotherapy, Freud consciously struggled with the selection and integration of diverse methods.

Formal ideas on synthesizing the psychotherapies appeared in the literature as early as the 1930s (see Goldfried and Newman 1992 for a history), but it is only within the latter decade of the twentieth century that integration has developed into a clearly delineated area of interest. A number of interacting, mutually reinforcing motives have fostered the recent growth of integration (Prochaska and Norcross 1999):

(a) *Proliferation of therapies.* The field of psychotherapy has been staggered by over-choice and fragmented by future shock. Which of 400 plus therapies on the market should be studied, taught, or bought? This might also be called the Exhaustion Theory of Integration: the cynical but accurate observation that peace among warring schools is the last resort.

(b) *Inadequacy of single theories for all patients and problems.* No single system has cornered the market on utility. Underlying the ecumenical spirit is the stark realization that narrow conceptual positions do not begin to explain the evidence in psychotherapy. Clinical realities have come to demand a more flexible, if not integrative, perspective.

(c) *Ascendancy of short-term, problem-focused treatments.* The brief, problem focus of psychotherapy, particularly in the US, has brought formerly different therapies closer together. Integration responds to the pragmatic time-limited injunction of 'which therapy works better—and quicker—for this patient with this problem.'

(d) *Opportunities to observe and experiment with various treatments.* The introduction of treatment manuals, the availability of psychotherapy videotapes, and the establishment of specialized clinics for the treatment of specific disorders have all afforded exposure to other theories and therapies, and stimulated some practitioners to consider other orientations more seriously.

(e) *Recognition that commonalties contribute heavily to outcome.* The equivalent outcomes found among the empirically evaluated psychotherapies in many studies has led to increased emphasis on factors common to successful cases. Therapeutic success can best be predicted by patient characteristics and the therapeutic relationship, both transtheoretical features of psychotherapy, while only 15 percent of outcome variance is accounted for by specific techniques (Lambert 1992).

(f) *Identification of specific treatments for particular disorders.* While commonalties contribute heavily to therapeutic success, controlled research has identified several treatments of choice for particular disorders; for example, cognitive therapy and interpersonal therapy for depression, behavior therapy for specific phobias and childhood aggression, and exposure for PTSD. This has led to selective prescription of different treatments, and combinations of treatments, to individual patient needs.

(g) *Development of a professional network for integration.* Reflecting and engendering the burgeoning

field have been the establishment of interdisciplinary psychotherapy organizations devoted to integration, notably the Society for the Exploration of Psychotherapy Integration (SEPI).

4. Pathways to Integration

There are numerous pathways toward the integration of the psychotherapies. The four most popular routes are technical eclecticism, theoretical integration, change processes, and common factors. All four attempt to look beyond the confines of single theories, but they do so in rather different ways and at different levels.

Technical eclecticism is the least theoretical of the four, but should not be construed as either atheoretical or antitheoretical (Lazarus et al. 1992). Eclecticism seeks to improve our ability to select the best treatment based primarily on the evidence of what has worked for others in the past with similar problems and similar characteristics. Proponents of technical eclecticism use procedures drawn from different sources without necessarily subscribing to the theories that spawned them (see *Eclectic Psychotherapy: Psychological Perspectives*).

In theoretical integration, two or more therapies are synthesized in the hope that the result will be better than the constituent therapies alone. This entails a commitment, as the name implies, to a conceptual or theoretical creation beyond a technical blend of methods. The various proposals to integrate psychoanalytic and behavioral theories, such as the influential work of Wachtel (1977), illustrate this path, as do the grander schemes to meld the major systems of psychotherapy, such as the transtheoretical approach of Prochaska and DiClemente (1984).

A level of abstraction down from theory is that of the change process or change strategy. This intermediate level of abstraction between global theory and specific techniques is a heuristic that implicitly guides the efforts of experienced therapists and that is increasingly guiding the teaching of psychotherapy. Instead of smushing theories fraught with epistemological incompatibilities on the one hand, or melding techniques without an overarching theory on the other, proponents of change processes suggest this intermediate level as the ideal for conceptualizing effective behavior change (Goldfried 1980).

The fourth pathway to psychotherapy integration is the common factors approach, which seeks to determine the core ingredients that different therapies share in common, with the eventual goal of creating more parsimonious and efficacious treatments based on those commonalties. This search is predicated on the belief that commonalties are more important in accounting for therapy outcome than the unique factors that differentiate among them, as eloquently maintained by Frank (1973). Literature reviews have discovered that the most consensual commonalties

across theories are the development of a therapeutic alliance, opportunity for catharsis, acquisition, and practice of new behaviors, and clients' positive expectancies (Grencavage and Norcross 1990).

Integration, as is now clear, comes in many guises and manifestations. It is neither a monolithic entity nor a single operationalized system; to refer to the integrative approach is simply a fallacy. At the same time, these multiple integrative pathways are not mutually exclusive. No technical eclectic can totally disregard theory, and no theoretical integrationist can ignore technique. Without some commonalities among different schools of therapy, theoretical integration would be impossible. And even the most ardent proponent of common factors cannot practice 'non-specifically' or 'commonly'; specific methods must be applied.

5. Future Directions

Psychotherapy integration is a vibrant international movement that has begun to make encouraging contributions to the field. Transtheoretical dialogue and cross-fertilization fostered by the integrative spirit have produced new ways of thinking about psychotherapy and researching behavior change. This closing section outlines five probable directions of the movement.

First and foremost, psychotherapy integration, in some guise, will likely emerge as the psychotherapeutic Zeitgeist. A dispassionate panel of psychotherapy experts, for example, portend its escalating popularity. Indeed, integrative and eclectic therapies were two of the four orientations (along with cognitive therapy and family systems therapy) predicted to expand the most in the future (Norcross and Freedheim 1992).

Second, integrationists will increasingly attend to the potential obstacles and tradeoffs of the process of integration. These will include establishing ambitious training programs and processes that ensure student competence in multiple theories and interventions, an unprecedented task in the history of psychotherapy; acknowledging the underlying epistemological and ontological differences embodied by the different theories of psychotherapies, which precludes facile integration; and addressing the language problem that confounds understanding of disparate jargons and, in some cases, leads to active avoidance of theoretical constructs.

A third direction for the integration movement is to document through outcome research the effectiveness of various integrative treatments. The empirical evidence supporting integration remains largely anecdotal and preliminary (Lambert 1992, Kazdin 1996, Wolfe and Goldfried 1988).

Fourth, psychotherapy integration will evolve past its initial step of widening therapeutic horizons and embracing multiple techniques and theories. The next

step is to specifically know when and where to use these multiple techniques and theories in order to enhance therapeutic efficacy, a process also known as differential therapeutics or prescriptive matching. Psychotherapies will thus be matched to multiple client variables, including but not limited to diagnosis or problem.

In this matching process lies the fifth direction for psychotherapy integration: customizing the therapeutic relationship to the individual patient. While the historical emphasis has been on matching techniques to specific disorders, the integration movement is more and more committed to the prescriptive use of the therapeutic relationship. One way to conceptualize the issue, paralleling the notion of 'treatment of choice' in terms of techniques, is how clinicians determine the 'relationship of choice' in terms of interpersonal stances for individual clients. Again, this customization will occur on the basis of both diagnostic and cross-diagnostic considerations, such as the patient's treatment goals, arousal level, coping style, personality traits, and stage of change (Beutler and Clarkin 1990).

In this manner, the psychotherapy integration movement promises to be instrumental in customizing research-informed treatment methods and therapeutic relationships to individual patients and, in so doing, enhance the effectiveness, applicability, and efficiency of psychotherapy.

See also: Behavior Psychotherapy; Rational and Emotive; Constructivist Psychotherapies; Counseling and Psychotherapy; Ethnic and Cultural Differences; Experiential Psychotherapy; Feminist Psychotherapy; Group Psychotherapy, Clinical Psychology of; Interpersonal Psychotherapy; Person-centered Psychotherapy; Psychological Treatment Delivered Non-traditionally; Psychological Treatment, Effectiveness of; Psychological Treatments, Empirically Supported; Psychotherapy, Brief Psychodynamic; Psychotherapy Process Research

Bibliography

- Beutler L E, Clarkin J F 1990 *Systemic Treatment Selection: Toward Targeted Therapeutic Interventions*. Brunner/Mazel, New York
- Frank J D 1973 *Persuasion and Healing*, rev. edn. Johns Hopkins University Press, Baltimore, MD
- Goldfried M R 1980 Toward the delineation of therapeutic change principles. *American Journal of Psychotherapy* **35**: 991-9
- Goldfried M R, Newman C 1992 A brief history of psychotherapy integration. In: Norcross J C, Goldfried M R (eds.) *Handbook of Psychotherapy Integration*. Basic Books, New York
- Grencavage L M, Norcross J C 1990 Where are the commonalities among the therapeutic common factors? *Professional Psychology-Research and Practice* **21**: 372-8
- Kazdin A E 1996 Combined and multimodal treatments in child

- and adolescent psychotherapy: Issues, challenges, and research directions. *Clinical Psychology: Science and Practice* 3: 69–100
- Lambert M J 1992 Psychotherapy outcome research: Implications for integrative and eclectic therapists. In: Norcross J C, Goldfried M R (eds.) *Handbook of Psychotherapy Integration*. Basic Books, New York
- Lazarus A A, Beutler L E, Norcross J C 1992 The future of technical eclecticism. *Psychotherapy* 29: 11–20
- London P 1988 Metamorphosis in psychotherapy: Slouching toward integration. *Journal of Integrative and Eclectic Psychotherapy* 7: 3–12
- Norcross J C, Freedheim D K 1992 Into the future: Retrospect and prospect in psychotherapy. In: Freedheim D K (eds.) *History of Psychotherapy*. American Psychological Association, Washington, DC
- Norcross J C, Goldfried M R (eds.) 1992 *Handbook of Psychotherapy Integration*. Basic Books, New York
- Prochaska J O, DiClemente C C 1984 *The Transtheoretical Approach: Crossing the Traditional Boundaries of Therapy*. Dow Jones-Irwin, Homewood, IL
- Prochaska J O, Norcross J C 1999 *Systems of Psychotherapy: A Transtheoretical Analysis*, 4th edn. Brooks/Cole, Pacific Grove, CA
- Stricker G, Gold J R (eds.) 1993 *Comprehensive Handbook of Psychotherapy Integration*. Plenum, New York
- Wachtel P L 1977 *Psychoanalysis and Behavior Therapy: Toward an Integration*. Basic Books, New York
- Wolfe B E, Goldfried M R 1988 Research on psychotherapy integration: Recommendations and conclusions from an NIMH workshop. *Journal of Consulting and Clinical Psychology* 56: 448–51

J. C. Norcross

Psychotherapy: Legal Issues (United States)

1. The Term 'Legal Issues in Psychotherapy'

Legal issues in psychotherapy can be understood as the ways in which the law affects providers of mental health services (hereinafter, mental health professionals or MHPs, in all stages of their careers and in all aspects of their practices. This article discusses the importance of MHPs knowing the law, the laws pertaining to MHP business matters and responsibilities, and psychotherapy services that are affected by the law. While comprehensive, this article provides only a general introduction to the law from the perspective of the United States's legal system (for detail on the issues discussed herein see, Sales and Miller 2001; Sales et al. in preparation). The US experience has been seen as predictive of future trends in other countries (e.g., United Kingdom, Bell-Boule 1999). It also does not discuss the provision of forensic services, covered elsewhere in this volume

(see *Forensic Psychology; Expert Witness and the Legal System: Psychological Aspects; Courtroom Consultation: Psychology; Rehabilitation of the Offender; Treatment of the Repetitive Criminal Sex Offender: United States; Crime and Delinquency, Prevention of*).

2. Importance of MHPs Knowing the Law

Despite the law's pervasive influence, most MHPs do not know about or understand most of the laws affecting the operation of their practices, their services and their clients. This is due in part to the difficulty of understanding legal terminology, and in accessing the many different and applicable sources of law. This state of affairs is unfortunate because MHPs may not gain the benefits that the law may provide to them or their clients, and they may not be able to fully meet their clients' needs (e.g., Rothstein 1997). In turn, legal knowledge can increase MHPs' capability for handling their clients' distress (e.g., Shuman and Smith 2000). For example, MHPs who obtain legal information are better equipped to facilitate the appointment of a guardian for an abused child or a legally incompetent person (Tor and Sales 1996), and can help correct distorted beliefs about what may happen during their legal proceedings.

Not knowing the law is also potentially harmful in an era of increasing professionalization and litigiousness (e.g., Stromberg et al. 1988). MHPs run the risk of liability for professional practice that the law designates as inappropriate (e.g., divulgence of certain professional information when the law specifies it as confidential, or for not divulging certain professional information that the law requires to be reported).

3. The Law Affecting MHPs Business Matters and Responsibilities

Before engaging in the practice of psychotherapy, MHPs need to attend to their business matters and responsibilities, including: ensuring that they are properly licensed, deciding on the form or structure of their business, and implementing procedures for dealing with financial matters and confidential client information. Some of these business issues arise during the course of a MHP's practice (e.g., release of confidential information). The law typically affects all of these business issues.

3.1 Legal Credentialing and Licensure to Practice

In the United States, specific statutory and administrative laws govern each of the licensed mental health professions (e.g., Hogan 1979). The law of credential-

ing and licensure to practice is typically organized by discipline (e.g., psychiatrists, psychiatric nurses, psychologists, subdoctoral and unlicensed psychologists, school psychologists, social workers, professional counselors, chemical dependency/substance awareness counselors, school counselors, marriage and family therapists, hypnotists, polygraph examiners).

These laws have a number of different functions. First, they establish licensing boards for MHP professions. In addition, some States have created overarching boards that have oversight of MHPs from all disciplines, in order to standardize practices across disciplines. In general, overarching and discipline-specific boards operate similarly. Second, the laws specify licensure qualifications (i.e., education, experience, testing of competence, good character), and procedures and exceptions to licensure (e.g., for those unlicensed MHPs who are supervised by a licensed MHP).

Finally, the laws regulate MHPs' conduct, and prescribe sanctions for violations. The regulation of practice generally specifies what a MHP can and cannot do. As to what MHPs can do, the law will typically define the permissible scope of practice for licensed professionals (e.g., conduct psychotherapy, dispense medications). As to what MHPs cannot do, the licensure law will describe the powers of the agency or board to investigate and adjudicate violations of the law, specify reasons for sanctioning those found in violation of the law, and provide penalties for sanctioned professionals.

3.2 Legally-defined Business Arrangements

MHPs can organize their practices in a number of different forms, many of which are subject to legal regulation (see Sales and Miller 2001). Sole proprietorships involve a MHP being the only owner of his or her mental health practice, and usually do not require a legal application for a special business form. Certain MHPs may form professional corporations that may offer certain tax, insurance, and other practice benefits but require following legal procedures.

More types of business forms exist for those MHPs who choose to work with other MHPs. When two or more people enter into a 'for-profit' agreement as co-owners, a partnership is formed. These partners agree to combine their resources in order to undertake projects that would be financially difficult for one person, and to share responsibility for each other's acts. Recent developments in the law of some States allow for limited liability partnerships or corporations, in which there are a number of limited partners or shareholders whose financial commitment and liability is limited. Finally, MHPs may form a nonprofit or

not-for-profit organizations to provide mental health services, including education, research, and/or inpatient or outpatient therapy, in a community health center, private hospital, or other setting.

The structure and organization of MHPs' practices are also affected by the health care entity with which MHPs and their clients are affiliated. The law may control certain health care provider arrangements in the US (e.g., MHPs who work for managed care companies), and government-run health services in other countries (e.g., Great Britain) (see *Adolescent Vulnerability and Psychological Interventions; Managed Care and Psychotherapy in the United States*). For example, in the US, Health Maintenance Organizations (HMOs) provide services to patient-subscribers (an individual or group) in exchange for a predetermined fixed premium. The law encourages HMOs to provide persons with competitively priced, alternative healthcare delivery and requires HMOs to offer basic, comprehensive healthcare services. The law varies by State as to whether HMOs are also required to cover mental health services, and whether only certain professionals are allowed to provide services to HMOs.

The liability of HMOs is an area of the law currently in development, a matter of much controversy. An HMO can be sued, and possibly may be held liable, for harm caused by the acts of participating providers in the course of furnishing such services. Another potential source of liability may arise from an HMO's refusing to allow for certain services in a particular case (e.g., utilization review that result in patients not being admitted or treated, or being discharged early). This could be a particular problem for MHPs if HMO precertification reviewers refuse to approve hospitalization, and a distraught patient then assaults another or commits suicide.

3.3 Financial and Contractual Matters

There are economic and contractual components to the relationships of MHPs with their clients as well as with the entities that may pay for or reimburse them for services. These issues are raised when MHPs are forming a relationship with a client and/or third-party payor; considering which services to provide; wanting financial payment for their services; and doing their taxes.

Since the MHP-client relationship is economic and contractual as well as therapeutic, MHPs should prepare a formal written agreement in order to protect the rights and interests of both parties at the outset of services. The client has the right to receive confidential, competent, and appropriate services that are performed according to professional standards. In return, the MHP should receive the client's cooperation with the service's plan (e.g., treatment) and payment for

services rendered. The written contract, therefore, should be drafted according to local laws and clearly establish (a) the fee structure; (b) policies regarding charges and payment (e.g., charges per session, charges for missed sessions); (c) the client's commitment to paying the MHP; (d) the nature and scope of services to be provided; and (e) the general rule of confidentiality and its exceptions (see Sect. 3.4).

Such contracts are made with the informed consent of the client. Informed consent should be obtained before administering mental health services, disclosing information concerning the client to a third party, or taking any other action that may impact the client. Informed consent basically involves the process of discussions between a MHP and a potential client that leads to an informed choice by the client. In general, informed consent has three major parts: (a) *competency* of the person to be able to consent to services; (b) disclosure by the MHP of relevant and material *information* concerning the nature and risks of the proposed professional services; and (c) consent that is *voluntarily* given to the MHP. Even if not legally defined, the components of informed consent are generally agreed upon (see, e.g., Kitamura et al. 1999 for a Japanese example). The failure to obtain informed consent can lead to legal liability for malpractice.

In general, when people voluntarily contract with a MHP to receive services, they have a right to receive the services that they pay for. Conversely, competent adults have the right to refuse treatment that they voluntarily sought. Legal issues become more complicated with regard to persons who are unable to consent to treatment. In some cases, a legally competent person can execute an advance directive that identifies his or her treatment wishes if she or he becomes incompetent, and specifies whom the healthcare proxy decision-maker should be during this period. If no advance directive exists, a court can appoint a guardian (substitute decision-maker) or guardian *ad litem* (guardian over a specific issue such as making treatment decisions) after the person has been determined to be legally incompetent to make treatment decisions.

Because it is increasingly common that clients do not pay MHPs directly for services, MHPs may be reimbursed by third party payors, such as insurance agencies or government agencies. As discussed above, third party payors can legally affect the nature and length of services that MHPs can provide. However, laws are changing to prohibit both reimbursement denial for mental health services in certain cases and disparity in coverage between mental and physical health services.

Finally, the law also affects how MHPs must calculate their income and deductions to pay their taxes. In the US, MHPs are allowed to deduct expenses only once they have begun employment as a professional. For example, they may deduct expenses

from subsequent educational activities or courses that: (a) are actually related to their business; (b) help the MHPs maintain or meet minimum requirements of their jobs or performance; (c) involve improving or maintaining their individual skills; or (d) are required by law (e.g., continuing education requirements for licensure). The law is unsettled regarding deductions for personal psychotherapy and psychoanalysis. The general trend is to allow MHPs who receive mental health services themselves to deduct them as educational business expenses if the MHPs have begun their practice, and if it is a necessary part of their practice (e.g., because the experience improves their job skills).

3.4 Maintaining and Disclosing Information

The doctrine of confidentiality originated in professional ethics codes (see *Psychotherapy: Ethical Issues*). It was based on the belief that effective psychotherapy required a guarantee from the therapist that no information obtained in the course of patient evaluation or treatment would be given to others without the client's consent. Thus, any written or verbal information conveyed by the client to an MHP in the course of a professional relationship can be considered a confidential communication (DeKraai and Sales 1984).

Maintaining confidential client records is not only necessary for the conduct of a good and ethical psychotherapy practice, but licensure law generally requires it. The law may specify a minimum period of time for which records must be kept, and the necessary procedures to follow to maintain the confidentiality of client records. As a general rule, MHPs should take reasonable precautions (i.e., precautions that similar MHPs in your area would take) to protect patient records. These procedures typically ensure that the records are preserved in the case of MHP relocation, separation from a group practice, retirement, or death. MHPs, who generally own their records, are also legally required to make records available to their clients upon request. However, laws may vary as to whether the request needs to be in writing, if fees can be charged for the service (e.g., for time spent copying records), and if there are limits on what type of information and in what format the requested information can be provided.

However, MHPs' duty to maintain client confidentiality is not absolute. Thus, it is important that MHPs inform clients of the limits of confidentiality at the initiation of services. Several types of penalties may be imposed for violations of confidentiality. First, as noted above, licensure law provides that the Board may suspend or revoke the license of any person betraying professional confidences, or it may impose other penalties such as probation. The same offending

actions may also subject the MHP to criminal penalties or a civil lawsuit by the client.

There are a number of threats to the privacy of psychotherapy, including exceptions to confidentiality and competing duties. The law may require MHPs to divulge confidential client information to legal authorities: (a) under search and seizure law, (b) according to a subpoena, and (c) if the client sues the MHP. Generally, confidentiality protections do not apply in legal proceedings unless a privileged communications law protects the information. Confidential information may also be disclosed to third parties if the client gives his or her written consent to a release of information.

Reporting laws impose a competing, affirmative duty on MHPs to disclose specific information obtained in therapy. These laws are based on the premise that the duty to maintain private information revealed in therapy is outweighed by society's need to protect the safety of its members, including the client. For example, confidential information may be divulged if the client presents a clear and imminent danger to himself or herself, and refuses to accept further appropriate treatment. Other laws that require confidential disclosures include dangerous person reporting laws, and child abuse reporting laws. While commentary on such reporting laws has been critical (e.g., laws are based on faulty premise that MHPs can accurately predict dangerousness; see *Clinical Psychology: Validity of Judgment*), research has not shown that the laws are antitherapeutic, and courts continue to uphold the laws.

4. MHPs Psychotherapy Services that Are Affected by the Law

Because MHPs provide psychotherapy services to a variety of individuals and groups with different needs in different settings, this section discusses how the law directly and indirectly affects MHPs' psychotherapy practices with certain populations.

4.1 Psychotherapy Services for Individuals with Special Needs

Individuals with special needs include people with mental illness, who abuse alcohol and other substances, with developmental disabilities, who receive treatment and care because they have been victims of domestic violence or other crimes, and those individuals who are incarcerated in jail or prison. Knowing the laws allows MHPs to provide clients with the services to which they are legally entitled.

First, US State law provides for the care of people

with mental illness in a number of ways (e.g., Brakel et al. 1985). Some people may be eligible to receive inpatient and outpatient mental health services through their health care plans. The law also provides for voluntary civil admission and involuntary civil commitment of mentally ill persons to state operated facilities. Generally, voluntary commitment may be permitted for those people with a variety of mental, emotional, or personality disorders, while involuntary commitment is only authorized for those people who meet a set of more restrictive standards listed in the law (e.g., Perlin 1994). The law may explicitly favor voluntary treatment over involuntary commitment because involuntary commitment generates substantial costs, and involves a loss of liberty to the person being committed. In such cases, many legally-mandated procedures must be followed to ensure that the person's rights are protected. For example, regular or long-term commitment generally requires a petition for a court-ordered commitment, a court review of the petition, an order for commitment, and an appeals process. In addition, the committed patient has numerous rights during treatment, including the rights to be free from excessive or unnecessary medication, and to not be subjected to certain procedures (e.g., experimental research; shock treatment) without informed consent, and/or a court order. MHPs can evaluate the person for admission and provide therapeutic services within the facility.

The law also provides for the voluntary treatment and involuntary commitment of those persons who are seriously disabled by alcoholism, and/or those persons with major substance abuse problems (i.e., drug-dependent persons or addicts). The law in many areas has changed from a punitive model to a continuum of care model, involving treatment and accountability (e.g., drug treatment courts). MHPs may be part of a multidisciplinary evaluation and treatment team under these laws, and may provide services (see *Relapse Prevention Training; Alcoholics Anonymous; Psychotherapy Process Research; Substance Abuse in Adolescents, Prevention of*) in a variety of settings (inpatient, outpatient, residential).

Third, the law recognizes that people with developmental disabilities (DD) have the capacity to be productive citizens, many of whom can achieve at least some or complete independence (e.g., Stefan 2000). Thus, some laws specifically provide for and protect rights of people with DD, including the rights to be given a nutritionally-adequate diet, and regular and sufficient medical care, and to not be physically or chemically restrained or isolated, except under strict regulations in emergency situations. Also, the law provides various inpatient and outpatient services (including evaluation and treatment by MHPs) for such persons (see *Mental Retardation: Clinical Aspects; Autistic Disorder: Psychological*). However, because such services are often not required, the provision of services is heavily dependent on available

funding, and upon a government's commitment to provide a certain level of services.

Fourth, some laws provide services, including those of MHPs, for victims of crimes, including the crime of domestic violence. For example, Victim Assistance Boards and Victim-Witness programs help victims put their lives back in order after the crime, and help them to testify at trial. Beyond counseling abused partners, MHPs may also become involved in these cases by identifying an abusive or battering relationship. MHPs can encourage victims to obtain legal orders to stop the abuse, and help clients to develop safety plans for their and their children's safety (see *Childhood Sexual Abuse and Risk for Adult Psychopathology*).

Finally, some laws support and require mental health services in jails and prisons because they are recognized as a vital part of the overall healthcare program for incarcerated persons. While MHPs may provide services as employees or consultants to these institutions, very little law specifically governs their activities. In some States, the law provides specialized services for sex offenders as part of their sentence, or through a more controversial, and more regulated, civil commitment treatment program. Some States require sex offender evaluation and treatment providers to be certified and to be subject to standards of professional conduct unique to the certification (see *Treatment of the Repetitive Criminal Sex Offender: United States; Sex Offenders, Clinical Psychology* of).

4.2 Psychotherapy Services with Children and Families

MHPs who work with children and families often find themselves courted by the legal system to aid it in administration, as when they provide evaluations and psychotherapy to court-referred juveniles and adults. Juvenile and adult clients can be referred for treatment either before trial to evaluate and/or restore their ability to participate in the process (e.g., competency to stand trial), or after trial/adjudication as part of their disposition (e.g., juvenile determined to be delinquent/in need of supervision and services) (see *Transportation, Psychology of; Rehabilitation of the Offender; Psychotherapy: Clinical Practice Guidelines; Conduct Disorder*). Court cases can also clarify laws and regulations regarding the psychotherapeutic, educational, and other services which children (especially those who are disabled, gifted and talented, in need of guardianship/conservatorship, or in foster care or government custody) should be receiving.

Finally, MHPs find themselves indirectly affected by the law when their child and family clients become involved in legal entanglements that involve mental status issues (see generally, DeKraai et al., 1998), such as proceedings that concern the termination of marriage or parental rights. For example, legal child

custody arrangements can affect not only the therapeutic needs of the child (see, *Psychological Therapies: Emotional Processing; Childhood Depression; Clinical Assessment: Interview Methods*) and adult (see *Anxiety Disorder in Children*) family members, but also the procedures that the psychotherapist must follow (e.g., the law may determine whether the MHP can evaluate and treat children at the request of a non-custodial parent).

5. Future Legal Developments

Despite the complexity of current legal regulations relating to psychotherapeutic practice, the law will continue to grow and change for a number of reasons. First, no longer do consumers unquestioningly accept MHP recommendations about treatment, or MHP assertions about the causes for unfavorable or unexpected treatment effects. Consumers are beginning to view psychotherapeutic services as similar to many other business arrangements. They want providers to be accountable for their work, and are no longer afraid to ask the courts and legislatures to impose new legal oversight and obligations to protect consumer interests. Second, increased legal regulation is likely to occur because many societies continue to fear, or at least stigmatize (e.g., Perlin 2000), the consumers of psychotherapeutic services. These perceptions will continue to fuel increased legal supervision of mental health services, which will likely be intrusive to the psychotherapeutic process.

Third, with MHP disciplines continuing to evolve, and as consumer concerns over cost-effectiveness in health care increase, new forms of practice are emerging both within and across disciplines. The inevitable result is the need for more laws to regulate these new forms of business practice and the unexpected consequences that sometimes ensue from them. Finally, because lawmakers are constantly looking for better ways to respond to societal needs (e.g., violent juvenile crime), we will continue to see new civil and criminal laws that partially rely on the services of MHPs. Each law brings new opportunities, requirements, and obligations for practitioners. What all this means for MHPs is that there will be increasing pressure on them to become conversant with the laws that affect both their practices and the consumers of their services

See also: Health Care: Legal Issues; Socioeconomic Status and Health Care

Bibliography

Bell-Boule A 1999 Psychotherapy and the law. *International Journal of Psychotherapy* 4(2): 193-203

- Brakel S, Parry J, Weiner B 1985 *The Mentally Disabled and the Law*. The American Bar Foundation, Chicago
- DeKraai M, Sales B 1984 Confidential communications of psychotherapists. *Psychotherapy* 21: 293–318
- Dekraai M, Sales B, Hall S 1998 Informed consent, confidentiality, and duty to report laws in the conduct of child therapy. In: Kratochwill T, Morris R (eds.) *The Practice of Child Therapy*, 3rd edn. Allyn and Bacon, Boston, pp. 540–59
- Hogan D 1979 *The Regulation of Psychotherapists*, Vols. 1–4). Ballinger Publishing, Cambridge, MA
- Kitamura T, Kitamura F, Mitsuhashi T, Ito A, Okazaki Y, Okuda N, Katoh H 1999 Image of psychiatric patients' competency to consent to treatment in Japan: I. A factor analytic study. *International Journal of Law and Psychiatry* 22(1): 45–54
- Perlin M 1994 *Law and Mental Disability*. Michie Law Publishers, Charlottesville, VA
- Perlin M 2000 *The Hidden Prejudice: Mental Disability on Trial*. American Psychological Association, Washington, DC
- Rothstein L 1997 *Disabilities and the Law*, 2nd edn. West Group, St. Paul, MN
- Sales B D, Miller M O (series eds.) 2001 *Law and Mental Health, Professionals Series*. American Psychological Association Books, Washington, DC [A series of volumes that comprehensively reviews and integrates the law that affects mental health professionals in the United States. <<http://www.apa.org/books/lmhptest.html>>]
- Sales B, Miller M, Hall S (in preparation) *Clinical Practice Law*. American Psychological Association, Washington, DC
- Shuman D, Smith A 2000 *Justice and the Prosecution of Old Crimes: Balancing Legal, Psychological, and Moral Concerns*. American Psychological Association, Washington, DC
- Stefan S 2000 *Unequal Rights: Discrimination Against People With Mental Disabilities and the Americans with Disabilities Act*. American Psychological Association, Washington, DC
- Stromberg, C, Haggarty D, Leibenluft R, McMillian M, Mishkin B, Rubin B, Trilling H 1988 *The Psychologist's Legal Handbook*. The Council for the National Register of Health Service Providers in Psychology, Washington, DC
- Tor P, Sales B 1996 Guardianship for incapacitated persons. In: Sales B, Shuman D (eds.) *Law, Mental Health, and Mental Disorder*. Brooks/Cole Publishers, Pacific Grove, CA, pp. 202–18

S. R. Hall and B. D. Sales

Psychotherapy Process Research

Psychotherapy process research is the scientific study of what happens in psychotherapy. It involves the precise, systematic, and controlled observation and analysis of the events that constitute psychotherapy. The basic elements of psychotherapy are simple—a patient, the distress or illness experienced by the patient, a therapist, and the treatment that the therapist provides—but psychotherapy occurs in a wide variety of forms and settings, involves a wide diversity of practitioners, and is conducted according to widely divergent theories about the causes and cures of psychological illness and distress. Since psycho-

therapy process research takes all these variations as its subject, a brief overview of the field it studies may serve as an introduction.

1. Psychotherapy

The forms of psychotherapy most commonly include individual therapy (one therapist with one patient), couple therapy (one therapist with two married or emotionally committed patients), family therapy (one or more therapists with members of a multigenerational family), and group therapy (one or two therapists with several patients who were previously strangers).

The settings in which psychotherapy is practiced include community mental health centers, private professional offices, university or college counseling centers, inpatient or outpatient psychiatric facilities, social work agencies, specialized schools, correctional facilities, consulting firms, and church-affiliated counseling programs.

The practitioners of psychotherapy are not members of a single profession, but instead are individuals who have typically been trained in one of the following professions: psychiatry, clinical or counseling psychology, clinical social work, psychiatric nursing, marital counseling or other counseling specialties, or the religious ministry. The level of involvement of different professions in psychotherapeutic practice, and their relative standing, varies from country to country.

Psychotherapeutic practitioners also differ in terms of the theoretical models they use to guide their work with patients. Broadly, these include analytic and psychodynamic theories, behavioral theories, cognitive and cognitive-behavioral theories, experiential and humanistic, and systemic or family-systems theory. Traditionally, these theories have formed the bases for rival 'schools' of psychotherapy. More recently, many therapists have adopted an eclectic or integrative attitude, seeking to broaden their effectiveness by combining the strengths of several models.

According to the theory followed, the treatment setting and format, and the professional identity of the therapist, these varied mental health practices may be known as psychotherapy, therapeutic counseling, psychoanalysis, behavior modification, personal training, growth facilitation, or some other term. Their practitioners may be referred to as psychotherapists, counselors, psychoanalysts, or simply by the name of the profession in which they were trained (e.g., psychologist, psychiatrist, social worker). The persons who seek help from these practitioners may be called patients, clients, counselees, or analysands. For simplicity in discussion, I refer to all those treatments as psychotherapies, to the persons who practice them as psychotherapists, and to the persons who experience them as patients.

2. Psychotherapy Research

Scientific research on the psychotherapies emerged as a field of inquiry in the middle of the twentieth century, following World War II, through the work of teams led by investigators such as C. Rogers, J. Frank, D. Shakow, M. Lorr, and others (Freedheim 1992, Chaps. 9–12, Russell 1994, Chap. 7). In this field, a distinction soon grew up between studies that focused mainly on treatment effectiveness, or the impact of therapy on patients, which were called ‘outcome’ research, and studies concerned primarily with the interactions of patients and therapists during the course of therapy, which were called ‘process’ research. Even from the outset, however, some studies included measures both of process and outcome, and in the years that followed process–outcome research has continued to be a vital focus of inquiry (Orlinsky et al. 1994, Orlinsky and Howard 1986).

Although this simple distinction between process and outcome served adequately for three decades, an increasing volume and complexity of studies and findings led eventually to the proposal of more finely differentiated frameworks for research. Based on extensive reviews of process–outcome, one of these (Orlinsky and Howard 1986) offered an overview of the field based on two distinctions: first, between therapeutic process (the observable events of psychotherapy) and its functional contexts (the social, cultural, and psychological circumstances in which therapy takes place); second, between context states that exist before and after the process events being studied, which influence and are influenced by them.

The influences of prior contextual states on events in therapy constitute determinants of therapeutic process. These would include, most immediately, the personal problems and psychological characteristics of the patient, and the therapeutic model, professional skills, and personal characteristics of the therapist. However, they also (if indirectly) include the organizational characteristics of the treatment setting, the nature and state of the mental health service delivery system, current societal pressures and stressors, and prevailing cultural standards of individual effectiveness, maturity, and well-being.

The reciprocal impacts of therapeutic events on subsequent contextual conditions comprise the consequences of therapeutic process. Researchers have been most directly concerned with the impacts of therapeutic events on the patient’s life and personality (i.e., clinical ‘outcome’). However, this broader framework also directs attention to the impact of therapeutic events on the professional and personal life of the therapist, and to the aggregate influence of patients’ and therapists’ experiences on cultural values and social institutions.

Between the contextual determinants of psychotherapy and its contextual consequences lies the set of psychological interactions and transformations that

comprise the psychotherapeutic process. Researchers who would observe and analyze therapeutic process must, like photographers, first make a number of decisions about framing and focusing their studies. Framing a psychotherapy process study entails decisions about the location of the phenomenon (where to look), the temporal unit of observation (how long to look), and the observational perspectives to be employed (from which angle to look). Focusing the study involves making decisions about the facets and phases of therapeutic process to assess (which ‘variables’ or changes to measure).

3. Process Research Frames

3.1 Location

While therapeutic process may be defined abstractly in terms of psychological interactions and transformations, process researchers have differed over where, concretely, they can be best observed. Many researchers view therapeutic process as evidently consisting of the interactions and communications that take place between patient and therapist during the regular meetings in therapy sessions. On this view, therapeutic process includes all of the events that can be observed and recorded during therapy sessions. However, other researchers define therapeutic process in terms of the changes which patients undergo as they improve. This view directs observations primarily to the patient’s experience and activity in and outside of session, and regards much that occurs during therapy sessions as therapeutically inert or inessential. To distinguish them clearly, these contrasting but somewhat overlapping concepts will be referred to as therapy session processes and therapeutic change processes, respectively. Investigators who frame their studies in terms of session process generally take a descriptive, theoretically nonpartisan, approach to defining therapy. By comparison, investigators who frame their research in terms of change processes usually hold, and seek to test, some particular theoretical conception of therapy (e.g., Dahl et al. 1988, Rice and Greenberg 1984).

3.2 Temporal Units

The concept of process refer most generally to activities and events that proceed or evolve over time. However, some activities and events proceed or evolve so rapidly, or so slowly, that they normally cannot be noticed. Examples of the former are interactions that physicists observe among subatomic particles; or, more familiarly, the rapid projection of still photographs that we do not notice separately but experience sequentially as ‘moving’ pictures. Examples of

the latter include the gradual cooling of romantic affairs into stable companionate relationships through years of shared routine; or, more remotely, the imperceptible erosion of mountains by millennia of wind and rain. The point is commonplace but basic and relevant.

Observations of process must be framed within the temporal span of human consciousness, but there is still a range of noticeable rates at which events occur. In observing therapeutic process, the researcher may choose to examine very brief events, like changes in facial expression or vocal intonation; or somewhat longer events, like the statement by one person of a coherent line of thought involving several sentences, or dialog between patient and therapist on a particular topic that includes successive statements by each; still longer events, like the quality of whole therapy sessions; or even longer events, like the opening, middle, and closing phases of an extended course of treatment. The temporal units or frames of reference that researchers select for study inevitably determine the kind of therapeutic events they can observe, and the descriptions of therapeutic process they give (Orlinsky et al. 1994, pp. 275–6).

3.3 Observational Perspectives

Any complex event can be viewed from several angles, and each observational perspective has the potential to reveal a different aspect of the event. In social interactions, including the interactive events of psychotherapy, the observational perspectives include those of the participants (called 'participant observers') and of immediate or remote bystanders (non-participant observers). In psychotherapy, there are two distinct participant observational perspectives, that of patients and that of therapists, who experience the events of therapy directly but from different vantage points. (The simplest case is that of individual therapies, which constitute the most frequent form of treatment; the same principle becomes more complex in practice when the treatment format includes more than one patient or more than one therapist, as in group or family therapies.)

To the patient's and therapist's natural observational perspectives, psychotherapy research has added that of nonparticipant observers by using clinical raters who are not part of the events being observed, and whose view of therapeutic process is generally mediated by audio or video recordings of therapy sessions or transcriptions of the verbal content of sessions. (Nonparticipant observations are sometimes also made by mechanical means, e.g., by using clocks to time speech durations.) Participant observers have a 'first hand' and 'inside' but also fleeting view of the events they experience. They have direct awareness of their own thoughts, feelings, and intentions, and

immediate impressions of their own and others' behaviors, but these are typically organized around their goals in action. By contrast, nonparticipant observers just gain a 'second hand' knowledge of therapeutic process, and can only infer the covert internal states of the participants, but they have the advantage over participant observers of being able to review the recorded events many times, without the situational pressures to respond that participants experience.

Researchers may, and have, used one, two, or all of these observational perspectives in their studies, depending on their interests and the questions they investigate. However, since patients, therapists, and outside observers have different perspectives on the events of therapy, they are bound to see somewhat different aspects of process. It should not seem strange to researchers, or an indication of methodological weakness, if observers who view therapeutic phenomena from different perspectives sometimes produce discrepant observations, since those discrepancies may validly reflect the inherent complexity of social phenomena.

4. Process Research Focus

When undertaking a study of therapeutic process, researchers understand that they are dealing with a set of events that is too complex to be observed in its entirety, and must simplify the focus of the study if it is to be successful. However a study is framed in terms of session process or change process, temporal units, and observational perspectives, the researcher must focus its observations on a few specific facets of therapeutic process: the 'variables' that are to be qualitatively assessed or quantitatively measured. For the simplification to be meaningful, these variables must be defined in terms of some conceptual model of therapeutic process, and to be widely accepted by other researchers that model must be generally accepted within the research community.

4.1 Research Variables

Traditionally, the variables examined in psychotherapy process studies have been differentially defined according to whether they focus on the instrumental-adaptive (goal-directed, or task) aspect of therapist and patient behavior, or the social-emotional (expressive-interpersonal, or relational) aspect of therapist and patient behavior. Therapist instrumental or task-oriented behaviors involve the technical interventions, methods, or procedures that various clinical theories indicate therapists should employ to help patients. Research variables include the nature, timing, skillfulness, and appropriateness with which therapeutic

techniques are used. Examples of specific technical interventions in psychoanalytic therapy are use of free associations to interpret dreams, interpretation of patient resistances, and transference interpretations. A typical therapist intervention in client-centered therapy would be empathic reflection of feeling; in behavioral therapy, relaxation training to assuage phobic anxiety; in gestalt therapy, the 'two-chair' technique to explore and resolve internal conflicts; and so on. Patient instrumental-adaptive or task-oriented variables generally focus on the patient's response to technical interventions, such as cooperativeness or resistance, as well as specific effects of interventions such as enhanced insight following an interpretation, deeper self-exploration and 'experiencing' in response to therapist empathy, or decreased anxiety when facing previously frightening stimuli as a result of relaxation training. Major research methods focused on instrument aspects of therapeutic process include those described in Greenberg and Pinsof (1986, Chaps. 2–8, 14).

A second traditional type of process variable that has been studied by psychotherapy researchers is defined in terms of the social-emotional or expressive-interpersonal aspect of the therapist and patient behavior, and their respective contributions to the formation and development of the therapeutic relationship (see *Therapist–Patient Relationship*). Relationship variables generally focus on the ways that patient and therapist feel and behave toward one another: the manner or style of their relating, complementary to the content of their interactions. Again, specific qualities of relationships have been emphasized by different theoretical orientations. Researchers influenced by psychoanalytic thought have focused on two issues related to the clinical concept of transference: core-conflictual themes in relationships and therapeutic alliance. Others, influenced by client-centered theory, have focused on therapist-offered relationship conditions such as genuineness, empathy, and positive regard. Social psychological theories of interpersonal behavior have provided another conceptual approach to research on relationships, emphasizing the reciprocal influences between patient and therapist, focusing on dimensions of behavior such as acceptance vs. rejection and control vs. autonomy. Useful research methods for studying therapeutic relationships are described by their proponents in Greenberg and Pinsof (1986, Chaps. 9–12, 15).

A more detailed, systematic classification that has emerged from repeated reviews of the research literature relating process to outcome (Orlinsky et al. 1994) distinguishes the following categories of process variables: therapeutic contract, therapeutic operations, therapeutic bond, participant self-relatedness, in-session impacts, and sequential flow. These are aspects of process that can be observed simultaneously in any unit of process, viewed from any observational perspective.

The *therapeutic contract* includes variables indicating the goals, tasks, and norms of patient and therapist roles, as defined by the therapist's treatment model; the number and qualifications of persons occupying each role; and the terms, setting, and schedule of their meetings. *Therapeutic operations* refer to the specific instrumental actions that patients and therapists perform in implementing the therapeutic contract, including the patient's presentation of a problematic condition (e.g., symptoms), the therapist's expert understanding of that condition (diagnosis), the therapist's ameliorative interventions (techniques), and the patient's response to those interventions (e.g., cooperation). The *therapeutic bond* is the personal connection that develops between the participants in therapy as they perform their respective roles as patient or therapist more or less in accordance with their therapeutic contract: their 'personal compatibility' (mutual feelings, and expressive attunement or rapport), and their 'teamwork capacity' (personal commitment to their roles, and interactive co-ordination). *Participant self-relatedness* reflects the intrapersonal aspect of patients' and therapists' involvement in the therapeutic process, including variables of self-awareness (e.g., openness vs. defensiveness), self-efficacy (e.g., confidence vs. embarrassment), and self-control (e.g., expressiveness vs. constriction). The *in-session impact* of these interdependent and reciprocally influential aspects of therapeutic process consists of the patient's 'therapeutic realizations' (e.g., hope, relief, understanding, courage) as well as the therapist's reinforcements (satisfactions and frustrations). The *sequential flow* of these process aspects within the therapy hour is indicated by variables reflecting session development, and the flow over longer periods of time is indicated by variables reflecting the course of treatment.

4.2 Research Methods

The events of psychotherapy occur in a series of meetings between patient and therapist called therapy sessions. The most directly observable aspects of those events consist of conversation. In some types of therapy, conversation may be supplemented by inviting patients to engage in certain activities or exercises designed to provide significant learning or corrective emotional experiences for the patient. Given the prominence of verbal and nonverbal communications, researchers have relied heavily on audio and video tapes of therapy sessions to record the events of therapy. Recording and replay technology have provided investigators with voluminous records for detailed examination from an external or nonparticipant observational perspective, but have also constrained them to concentrate on temporal units that have generally been briefer than whole sessions, and sometimes no longer than one complete statement or

interchange between a patient or therapist. Examples of relevant methods can be found in Greenberg and Pinsof (1986, Chaps. 2–8).

But conversation is just the outward side of therapeutic events. The inner side of process is comprised of patients' and therapists' experiences of themselves and one another during their sessions, and during the imagined conversations or fantasy interactions that they may have in the times between sessions (e.g., Greenberg and Pinsof 1986, Chap. 13). This aspect of therapeutic process has been studied mainly through the use of postsession questionnaires. These retrospective ratings utilize the participant observational perspectives of patients and therapists to provide research data, and necessarily take the whole session as the temporal unit of observation.

Chapters providing further discussion of process research methods can be found in Bergin and Garfield (1994), Greenberg and Pinsof (1986), Horvath and Greenberg (1994), Kiesler (1973), and Russell (1994).

5. Main Process–Outcome Findings

A half century of research on psychotherapy has cumulatively demonstrated that qualities of the therapeutic bond or relationship, observed during the course of therapy from varied observational perspectives, robustly predict treatment outcomes as assessed by patients, therapists, clinical judges, and psychometric measures, although there is still some debate about whether these relationship qualities produce that outcome or merely reflect its on-going attainment. Global ratings of relationship quality are demonstrably linked to outcome from all observational perspectives. In terms of specific relationship qualities, therapists are better able to distinguish the patient's contribution to therapeutic process, whereas patient as well as therapist process contributions are associated with outcome from the distinct observational perspectives of patients, clinical judges, and psychometric measures. The patient's personal investment, collaborativeness, expressiveness, emotional responsiveness, and positive feelings towards the therapist are particularly important, as is the therapist's empathy and credibility. This has led to a widespread view of the therapeutic bond or alliance as the most salient aspect of therapeutic process, and the 'common factor' that enables different types of psychotherapy to be more or less equivalent in effectiveness.

With regard to therapeutic operations, some types of technical interventions, when skillfully focused on patients' life problems, have also been shown persuasively to be associated with positive outcomes (e.g., interpretation, experiential confrontation, and paradoxical intention), although this too is evident from the observational perspectives of patients and clinical

judges rather than therapists. In terms of participant self-relatedness, assessments of patient openness (vs. defensiveness) have been very consistently associated with outcome, from all observational perspectives. Likewise, patients' positive in-session impacts ('therapeutic realizations') have consistently been found to predict positive treatment outcomes.

6. Process Research and Practice

Psychotherapy process research thus far has proven itself a scientifically successful field of inquiry. Initially, it had to overcome an exclusion rule, stated by Freud (1953, p.185), that 'the "analytic situation" allows the presence of no third party. ... An unauthorized listener who hit upon a chance one of them would as a rule form no useful impression; he would be in danger of not understanding what was passing between the analyst and the patient.' Thus in the earliest days, investigators were mainly concerned to demonstrate the feasibility of systematic and methodologically objective research on therapeutic process. Once that had been established (e.g., Rogers and Dymond 1954), another 15 years or so were devoted primarily to devising reliable measures and rigorous methods of process research, during which time some therapists who had taken an interest in the field were disappointed by the methodological preoccupations of researchers and the paucity of clinically applicable results (e.g., Talley et al. 1994, Chaps. 2, 4, 5). However, the rate of research on process and process–outcome patterns increased steadily from about 1970 onwards. As the 1990s approached, a considerable body of well-replicated findings had been amassed, from which implications for therapeutic practice could begin to be drawn (e.g., Talley et al. 1994, Chap. 6). Current pressures for accountability and quality assurance from managed care have forced therapists of all schools to become interested in outcome research, in order to demonstrate the effectiveness of their treatments. Yet the knowledge that most psychotherapists have of process and process–outcome research findings lags behind the times. As progress in process research continues, a major challenge will be to increase clinical awareness of the relevance and potential usefulness of these research findings.

See also: Assessment and Application of Therapeutic Effectiveness, Ethical Implications of; Behavior Psychotherapy: Rational and Emotive; Behavior Therapy: Psychological Perspectives; Cognitive and Interpersonal Therapy: Psychiatric Aspects; Cognitive Therapy; Experiential Psychotherapy; Person-centered Psychotherapy; Psychological Treatment, Effectiveness of; Psychotherapy, Brief Psycho-dynamic; Psychotherapy: Case Study

Bibliography

- Bergin A E, Garfield S L (eds.) 1994 *Handbook of Psychotherapy and Behavior Change*, 4th edn. Wiley, New York
- Dahl H, Kaechele H, Thomae H (eds.) 1988 *Psychoanalytic Process Research Strategies*. Springer-Verlag, Berlin
- Freedheim D K (ed.) 1992 *History of Psychotherapy: A Century of Change*. American Psychological Association Press, Washington DC
- Freud S 1953 The question of lay analysis. In: Strachey A (ed.) *The Standard Edition of the Complete Psychological Works of Sigmund Freud*. Hogarth Press and the Institute of Psychoanalysis, London, Vol. 20
- Greenberg L S, Pinsof W M (eds.) 1986 *The Psychotherapeutic Process: A Research Handbook*. Guilford, New York
- Horvath A O, Greenberg L S (eds.) 1994 *The Working Alliance: Theory, Research, and Practice*. Wiley, New York
- Kiesler D J (ed.) 1973 *The Process of Psychotherapy: Empirical Foundations and Systems of Analysis*. Aldine, Chicago
- Orlinsky D E, Grawe K, Parks B K 1994 Process and outcome in psychotherapy—noch einmal. In: Bergin A E, Garfield S L (eds.) *Handbook of Psychotherapy and Behavior Change*, 4th edn. Wiley, New York
- Orlinsky D E, Howard K I 1986 Process and outcome in psychotherapy. In: Garfield S L, Bergin A E (eds.) *Handbook of Psychotherapy and Behavior Change*, 3rd edn. Wiley, New York
- Rice L N, Greenberg L S 1984 *Patterns of Change: Intensive Analysis of Psychotherapy Process*. Guilford, New York
- Rogers C R, Dymond R F (eds.) 1954 *Psychotherapy and Personality Change*. University of Chicago Press, Chicago
- Russell R L (ed.) 1994 *Reassessing Psychotherapy Research*. Guilford, New York
- Talley P F, Strupp H H, Butler S F (eds.) 1994 *Psychotherapy Research and Practice: Bridging the Gap*. Basic Books, New York

D. E. Orlinsky

Puberty, Psychosocial Correlates of

1. Introduction

This article aims to review recent research on the ways in which the biological changes of puberty influence and are influenced by the social and psychological contexts in which young people live and develop. While the biology of puberty has presumably been relatively constant since the emergence of *Homo sapiens*, the context in which young people are developing has changed dramatically. The evidence for these effects and interactions will be briefly reviewed, with some attention to the processes thought to be involved. It is important to note that almost all of this research is descriptive and correlational rather than experimental, although more work is being done with animal models to clarify causal explanations.

The phrase 'pubertal youth' is likely to evoke images of a group of children at least some of whom are

beginning to show manifestations of physical maturity in terms of adult size and shape. Similarly, 'pubertal behavior' would conjure images related to interest in the other sex. A photograph often used in textbooks on adolescence in which somewhat taller girls are paired as dancing partners with later-developing boys depicts both 'pubertal youth' and 'pubertal behavior.' That some youngsters depicted in the photograph do not yet show physical changes of puberty demonstrates that some will engage in 'pubertal behavior' for social reasons. Further, the probabilistic nature of any complex developmental process implies that what generally occurs in a cohort of youngsters will not apply to some individuals.

The psychosocial correlates of puberty, together with the biological process of puberty itself, are usually considered to be a single biopsychosocial process (that is, a process involving biological, psychological, and social aspects) in development. In this encyclopedia, 'puberty' is used to refer to the biological aspects of the process rather than the biopsychosocial process. In this article, psychosocial correlates will be considered in two ways: (a) those that appear to influence the timing or nature of the biological process and (b) those that appear to be results or effects of the process. The term 'behavioral' will be used for both kinds of psychosocial correlates. It is assumed that influences from the social context (including structural influences) have effects through behavior.

2. Behavioral Influences on Puberty

2.1 Influences on Pubertal Timing

There is evidence that several behavioral influences affect the timing of puberty, with some altering physical appearance. These include eating behavior, exercise, stress, and sexual behavior. Starvation and other extreme stress may prevent the pubertal process from ensuing, presumably to prevent reproductive capacity in an organism unable to sustain it. Limited nutritional intake along with high physical demands on the body result in a delay of pubertal onset. For example, puberty is delayed by 1–2 years in ballet dancers, figure skaters, and gymnasts who practice intensely and restrict eating. Because prepubertal growth of the long bones continues, these youngsters may have longer limbs than would otherwise be the case (Eveleth and Tanner 1990).

Recent hypotheses have linked familial stress to acceleration or delay of puberty. For example, evolutionary theory is used to posit that children of parents under high stress will engage in behavior that will speed their exit from the family (Belsky et al. 1991). Parents under high family stress will engage in insensitive, inconsistent, and affectively negative

parental care. Young children in this circumstance then develop insecure attachments to parents and an opportunistic style of interacting with others. These children are more likely to develop behavior problems such as aggression or social withdrawal in childhood and risky sex or substance use behaviors in adolescence. While the exact process linking family relationships to pubertal timing has not been identified, it is argued that earlier pubertal timing is adaptive in this circumstance.

Delay of puberty has also been proposed to result from stressful family contexts. In this hypothesis, environmental stress increases adrenal hormones (i.e., the hypothalamic–pituitary–adrenal axis) which inhibit sex steroids (i.e., the hypothalamic–pituitary–gonadal axis) (Malo and Tremblay 1997). Research supporting this hypothesis has found that families with paternal alcoholism or with negative emotions expressed by parents were more likely also to have sons with delayed puberty. The presumptive mechanism is a stress response in the sons that interferes with hormonal mechanism for initiating puberty resulting in delayed pubertal onset.

Another line of work links prepubertal sexual abuse to earlier puberty (Trickett and Putnam 1993). The underlying mechanism may be pheromones that trigger pubertal onset, or some other triggering mechanism related to the stimulating effects of sexual contact.

This work is all relatively new and more work is needed to refine understanding of precipitating factors and mechanisms. It is possible that behavioral influences operating directly on the gonadal system accelerate pubertal onset whereas those that operate primarily through the adrenal system delay puberty. However, at this point there is no definitive research to clarify why some stressors accelerate while others delay pubertal onset.

In any case, it is clear that even timing of pubertal onset, an aspect of puberty with strong evidence of genetic influence, can be affected by behavior within the social contexts in which young people live and develop.

2.2 Influences on Other Aspects of Puberty

The example of exercise above is one in which effects have been seen on the adult shape that results from pubertal growth, in addition to delaying pubertal onset. Young ballet dancers who train many hours a day develop longer limbs—both arms and legs—compared with sisters and mothers. A similar effect has been observed among swimmers. While there surely are selection effects confounding the results for these elite athletes, the comparison with other family members provides some degree of control.

Although there are likely to be other examples of effects of behavior on pubertal outcomes, effects on

gross aspects of physical development are easiest to attribute. Even here, though, there is the continuing confound of experience during development that makes causal attribution to pubertal change more difficult to establish. For example, different shapes surely result from different eating patterns (e.g., overeating versus undereating), but in this discussion of puberty the focus is on those outcomes that are pubertal rather than those that would occur at any point in development. As complicated as these issues become with physical attributes, they are even more complicated with psychological or social outcomes.

One effect of context on puberty should be mentioned. The increased life span and generally better health and nutrition of many people today have led to an earlier age at puberty, clearly among more developed nations in the world and increasingly among developing nations (Garn 1992). This changes the context for puberty in that it occurs in otherwise less mature organisms. Some had thought that all aspects of development might proceed in concert, so that earlier puberty would be accompanied by earlier cognitive, psychological, and social development. This does not appear to be the case, probably because of the experience or training (sometimes called socialization) needed to become a mature human. Given the existence of psychologically or socially immature adults, we should not be surprised about physically mature but otherwise immature adolescents.

The fact that puberty occurs a few years earlier now than even in the mid-1900s, and is now closer to 10 years of age among girls than to 20 years as was more typical around 1900, has tremendous significance for our understanding of and also the experience of childhood. Biological maturity permitting procreation occurs at ages much earlier than those typically considered appropriate for taking on adult work and family roles. Most societies have not adjusted to this disjunction, either to protect children from becoming parents too early or to change socialization practices and social institutions to support earlier entry into adult roles. Instead, some societies have blamed the victim in this situation, suggesting that early child-bearing represents a moral failure on the part of the child or their families.

3. Pubertal Influences on Behavior

As with the earlier discussion of effects on puberty, pubertal timing is much more important than are other aspects of pubertal change as a factor influencing behavior. Both kinds of effects will be summarized.

3.1 Pubertal Timing Effects

Three hypotheses or models have been proposed for pubertal timing effects on psychosocial adjustment,

the outcomes that have been the focus of most research: the stage termination hypothesis, the deviance hypothesis, and the goodness-of-fit hypothesis (Petersen and Taylor 1980, Alsacker 1995). The stage termination hypothesis places pubertal change within the context of other developmental stages, and suggests that early puberty could cause premature termination of prepubertal development. The deviance hypothesis suggests that negative effects could accrue from being deviant for one's age and gender, as with earlier-developing girls and later-developing boys. The goodness-of-fit hypothesis posits that negative effects will result when the social context differs from the individual's developmental status, as with an aspiring dancer who may be developing on time but producing a body shape different from the prepubertal one valued for dance. There is some support for each of these hypotheses, although developmental stage has receded in importance as a psychological construct.

3.1.1 Body image. Early-developing girls tend to have poorer body images during adolescence, with increasing effects by the end of the adolescent decade, compared with girls developing later or on time relative to their peers. In contrast, early-developing boys had the most positive body image in early adolescence, with effects diminishing over adolescence (Graber et al. 1996). Although body dissatisfaction is most strongly related to short stature among boys and heavier weight among girls, preference for body types has a strong cultural component, with variations seen cross culturally (Silbereisen et al. 1989).

3.1.2 Internalizing behaviors. Early-developing girls and, to some extent, late-developing boys appear to be more vulnerable to a range of mental health indicators, including internalizing problems (e.g., depression, phobic disorders) (Graber et al. 1997). Research has also shown that pubertal timing effects can be compounded by other life changes occurring at around the same time, such as change to a new school context (Petersen et al. 1991).

3.1.3 Externalizing behaviors. Earlier pubertal timing among both boys and girls is related to more problems with externalizing behaviors (e.g., aggression, delinquency). This is a strong effect from almost all the studies of problem behavior, including substance use and abuse, with effects found on lifetime rates of substance dependence and disruptive behavior disorders (Graber et al. 1997). There is some evidence that this pubertal timing effect is especially

strong when early developers begin associating with older peers.

3.1.4 Sexual behaviors. Recent research studying hormones, pubertal timing, and sexual behavior demonstrates that for boys high levels of testosterone represent a marker for early puberty, and lead to earlier sexual behavior. For girls, high levels of testosterone also lead to earlier initiation of coitus but do not seem to be markers of early puberty. This research also demonstrates that social control (e.g., religious restriction of social behavior) can moderate biological effects (Halpern et al. 1997). Many other studies find that earlier developers are more likely to be sexually active, consistent with the pattern reported above for externalizing behaviors.

3.1.5 Summary of pubertal timing effects. For both body image and internalizing behaviors, the deviance hypothesis appears to predict the pattern of results; early-developing girls and, to some extent, late-developing boys are at greater risk of problems with these outcomes. For externalizing behaviors and sexual behavior, early development in both boys and girls leads to earlier and more extensive involvement in these behaviors. To the extent that these youngsters become involved with older peers, these results fit the goodness-of-fit hypothesis, but most studies find that early pubertal onset precedes peer choice. Hence, these results seem to suggest that a different model is operating, one more directly biological, although with social moderation of effects.

3.2 Pubertal Status Effects

As noted earlier, very few effects of pubertal status have been reported in the literature. The young person's stage of pubertal development seems to have little effect on behavior. The one area that is an exception is that of parent-child relationships. As the young person becomes pubertal, there is an increase in parent-child conflict and a decrease in warm interactions peaking midpuberty (or around the time of menarche for girls) and then declining (Paikoff and Brooks-Gunn 1991). The nature of conflict is mild bickering, frequently over chores or other expectations.

4. Conclusion

As this review demonstrates, becoming pubertal has modest significance for relationships with parents but otherwise has little consistent effect on behavior. Whether one becomes pubertal earlier or later than

one's peers is of much greater significance. Deviant development relative to the group of same-aged peers—that is, early development for girls and late development for boys—is related to internalizing behaviors, or the way one feels about oneself. Early development for both boys and girls is related to problems with externalizing behavior, or the way one feels about others. The latter finding suggests that in this time of earlier puberty for youngsters generally, those who are even earlier are especially at risk. Early-developing girls are at risk for both internalizing and externalizing problems while early-developing boys are primarily at risk for externalizing problems.

In general, puberty constitutes an important change in a young person's life. The effects of developing adult size, shape, and reproductive potential are significant. Even more important to the developing individual, however, is the meaning of these changes for the society in which the young person is developing. If these biological changes signal marginality in the society, the impact on the individual is likely to be negative. In contrast, if these changes are welcomed and even celebrated, as was the case in many traditional societies in which puberty was celebrated with rites of passage into adult society, the impact on the individual is likely to be positive. Therefore, societies can choose to have more positive, constructive behavior among their young people, or they can increase the likelihood of negative and life-damaging behaviors among young people.

There is much yet to learn about effects on and effects of puberty. Recent advances in genetics and neuroscience provide tremendous opportunities to link these advances to understanding of development, especially at puberty. Equally important is research illuminating the effects of social and cultural contexts on biological processes such as puberty and the ways in which the biological process of puberty interacts with individual psychology and social context. Especially needed are studies that establish the generality of current findings, or the limits of their generalizability, with broader populations.

Policy efforts are needed to apply what we already know about puberty. There is a need for greater understanding, and recognition in policy and practice, of the fact that puberty is occurring significantly earlier now for young people than it did in the current generation of parents and grandparents. Different effects for boys and girls also require some broader attention in practice if not policy.

See also: Adolescence, Psychiatry of; Adolescence, Sociology of; Antisocial Behavior in Childhood and Adolescence; Childhood and Adolescence: Developmental Assets; Cognitive Development in Childhood and Adolescence; Identity in Childhood and Adolescence; Social Competence: Childhood and Adolescence; Socialization in Adolescence

Bibliography

- Alsacker F D 1995 Timing of puberty and reactions to pubertal changes. In: Rutter M (ed.) *Psychosocial Disturbances in Young People: Challenges for Prevention*. Cambridge University Press, Cambridge, UK, pp. 37–82
- Belsky J, Steinberg L, Draper P 1991 Childhood experience, interpersonal development, and reproductive strategy: An evolutionary theory of socialization. *Child Development* **62**: 647–70
- Eveleth P B, Tanner J M 1990 *Worldwide Variation in Human Growth*, 2nd edn. Cambridge University Press, Cambridge, UK
- Garn S M 1992 Physical growth and development. In: Friedman S B, Fisher M, Schonberg S K (eds.) *Comprehensive Adolescent Health Care*. Quality Medical Publishing, St. Louis, MO, pp. 18–23
- Graber J A, Lewinsohn P M, Seeley J R, Brooks-Gunn J 1997 Is psychopathology associated with the timing of pubertal development? *Journal of the American Academy of Child and Adolescent Psychiatry* **36**: 1768–76
- Graber J A, Petersen A C, Brooks-Gunn J 1996 Pubertal processes: Methods, measures, and models. In: Graber J A, Brooks-Gunn J, Petersen A C (eds.) *Transitions Through Adolescence: Interpersonal Domains and Context*. Lawrence Erlbaum, Mahwah, NJ, pp. 23–53
- Halpern C T, Udry J R, Suchindran C 1997 Testosterone predicts initiation of coitus in adolescent females. *Psychosomatic Medicine* **59**: 161–71
- Malo J, Tremblay R E 1997 The impact of paternal alcoholism and maternal social position on boys' school adjustment, pubertal maturation and sexual behavior: A test of two competing hypotheses. *Journal of Child Psychology and Psychiatry* **38**: 187–97
- Paikoff R L, Brooks-Gunn J 1991 Do parent–child relationships change during puberty? *Psychological Bulletin* **110**: 47–66
- Petersen A C, Sarigiani P A, Kennedy R E 1991 Adolescent depression: Why more girls? *Journal of Youth and Adolescence* **20**: 247–71
- Petersen A C, Taylor B 1980 The biological approach to adolescence: Biological change and psychological adaptation. In: Adelson J (ed.) *Handbook of Adolescent Psychology*. Wiley, New York, pp. 117–55
- Silbereisen R K, Petersen A C, Albrecht H T, Kracke B 1989 Maturation timing and the development of problem behavior: Longitudinal studies in adolescence. *Journal of Early Adolescence* **9**: 247–68
- Trickett P K, Putnam F W 1993 Impact of child sexual abuse on females: Toward a developmental, psychobiological integration. *Psychological Science* **4**: 81–7

A. C. Petersen

Public Administration: Organizational Aspects

Public administration is the term used to define the formal procedural and organizational arrangements under which public employees serve a government, by

implementing and advising on policy, and managing resources. Organizational aspects refer to both the overall structures and relationships within public administrations. This could include: The organizations that make up a civil service, sometimes referred to as the machinery of government; internal organizational arrangements; and/or organizational behavior. Thus, organizational aspects can be studied in a broad sense, or within several defined fields. This contribution covers organizational aspects, widely interpreted.

Beyond public administration as a discrete body of knowledge, organizational aspects can be examined further through other theories and practices relating to, for example, political science, public policy, sociology, economics, and management. In this sense, each area of study has its own theories and concepts. Furthermore, each state has its own history, organizational form, and approach, although there are many universal, common elements, which have developed through the international transfer of ideas. New governments, formal review processes, focused research, and events have often stimulated notable change. Therefore, the area of public administration is a difficult area to research, and over the years studies have been largely descriptive, rather than experimental or empirical.

For the purpose of this article, organizational aspects of public administration are examined as four phases of development and evolution, representing important paradigm shifts, or prospective shifts. These are the premodern, modern, postmodern, and virtual/supranational phases. Critical issues relating to the theories, models, trends, and practices of organizational aspects of public administration during these times are explored.

1. Premodern Public Administration, Organizational Aspects (Pre-1850s)

The organizational aspects of public administration have been apparent from ancient times, through many civilizations, such as Greek, Roman, and Chinese dynasties, and have often been the subject of criticism and analysis. Indeed, the ancient Greek satirist and playwright Aristophanes was probably one of the earliest in the literate world to highlight the lack of accountability and fairness in relation to public office. The works of ancient philosophers, such as Plato and Aristotle, address pertinent issues of interest in this regard.

Military histories over many years from the ancient Greeks and Romans to Machiavelli, to Frederick the Great, and to Napoleon provide further insights into the evolution of organizational aspects of public administration. For example, in the 1700s, in the context of militarism and empire building, the Prussian leader, Frederick the Great, and his successor,

Frederic William III, structured an exemplary public administration. Furthermore, France's public administration, which had been evolving for centuries through constitutional reform following the French Revolution of 1789, was organized on the basis of similarly functioning districts, or *départements*. In the late 1700s to early 1800s, Napoleon oversaw the central appointment of officials charged with the administrative responsibility for the *départements*. He also developed bodies which undertook administrative review, and involved the separation of juridical and public administrative powers. The civil service was made up of a number of distinct corps supported by training schools.

In the context of royalty, an early patronage system of public administration developed in England, primarily, as a support function for the royal Anglo-Saxon courts. The later structuring of a distinct, centralized public administration was related to a concept of a parliamentary cabinet. Some senior government ministers had portfolio responsibilities and a small supporting staff. Separate departments eventually existed to provide administrative functions to the ministers in charge of the Treasury, the Exchequer, the Chancellery, and the State. Accountabilities of the ministers were divided, between the Parliament and the monarch. While ministers undertook many of the administrative duties because of a general lack of suitably qualified men [*sic*], a small, elite corps of public administrators was established with a resultant continuity of tenure from government to government. To a small extent, these public administrations served a legitimate purpose, but for the most part they existed to meet the needs and ends of the administrators themselves.

Some of the public administration ideas from Europe were also adopted in New World states. In the USA, in the context of a revolutionary federal constitution which came after the War of Independence, their early patronage system was based on political connections, rather than royalty or class. Not surprisingly, these arrangements meant that the administrative organizations of government became highly corrupted, as Woodrow Wilson observed (1887). The dissatisfaction with this type of system, and a promoted interest in reform processes, led to the eventual development of professional public administrations in many states.

1.1 Modern Public Administration, Organizational Aspects (1850s–1970s)

Organizational aspects of public administration continued to evolve because of constitutional and parliamentary reform processes and, in some cases, shifts from royal courts to representative governments. The industrial revolution, the rise of capitalism, and the

new bourgeoisie all contributed to pressures for professional civil services. Thus, in modern times, organizational aspects of public administration have been based largely on the need for professionalism and have often come about because of administrative review.

In Britain, for example, the Northcote–Trevelyan Report (1853) recommended major changes in civil service recruitment and other processes. Fundamental to the changes was the establishment of a central Board of Public Administration, which eventually became a Civil Service Commission in 1870. The Board was responsible for the competitive recruitment, based on merit, of professional civil servants, and their appointment to departments; it was largely independent of government. From an organizational aspect, the idea of a central and independent agency of government responsible for recruitment and other personnel matters was instituted from that time. The best educated men [*sic*] with knowledge of history, classics, language, and mathematics were appointed to the service. Three classes of civil servant also applied: the elite administrative classes; clerical support classes; and the technical or professional classes.

In the USA, European and British theories and practices continued to influence reform of public administration, especially through the Civil Service (Pendleton) Act of 1883. However, there was a determination, as Woodrow Wilson (1887) notes, to ‘Americanize’ the system. The state was regarded as an instrument of social cooperation, and a more bureaucratic public administration structure was needed to manage the state. However, there was always a fundamental rejection that the senior administration could perform in a non-partisan way. This meant that there were dual streams of career civil servants and political appointees at the high levels of the civil service. Therefore, it was recognized that there was a blurred delineation between the political and public administration environments. High-level entry requirements were embedded into the career civil service, and a political education was favored as the main qualification. Taylor’s (1912) theory of scientific management and related developments in the field were promoted as appropriate models of organization and work distribution.

Another major contributor to the evolution of ideas about public administration and organizational aspects was the German economist and sociologist, Max Weber. Weberian principles of bureaucracy have had a universal impact on the organizational aspects of public administration. These principles included the matching of fixed job responsibilities to defined, permanent positions; a hierarchy of official authority; the requirement for appointed, suitably qualified, and anonymous staff with continuity of tenure until retirement; recognition that high level appointments bring social status; and entitlement for a salary and a retirement pension (Weber 1946).

As an example of Weber’s force, his ideas were evident as far away as Japan. Public administration and organizational aspects were based largely on German bureaucracy and Weberian principles, with other local attributes. This included the preferential acceptance of law graduates into the administrative elite stream, which has meant that the critical decision processes have been rule and precedent bound. However, the developmental role of the state has led to the dominance of certain elite governmental organizations, such as the Ministry of International Trade and Industry.

Furthermore, major events and trends, such as world wars and the growth of global trade, have impacted upon organizational aspects of modern public administration. Consequently, public sectors have increased in certain periods to address major reconstruction initiatives. Cycles of nationalization of certain industries have also impacted upon organizational aspects. Political history is important, too, and organizational change at a micro level will usually occur when new governments are elected. Thus, there is a strong link between significant history and highly dynamic organizational aspects of public administration.

Thus, public administration has evolved as a distinct area of research and study in modern times. This is supported organizationally where, in concept, there is meant to be a clear separation of powers and responsibilities between political executives (governments); public administrations (civil or public services); and, the juridical arena (the law and courts). However, there are obvious accountabilities between governments and public organizations, as well, given that there is usually a correlation between the portfolio responsibilities of members of the government and the way public administrative arrangements are determined. For example, members of a government are allocated portfolio responsibilities pertinent to particular public policy fields, such as health, defense, or education. Typically, there is a defined public administration department dedicated to the implementation of government policy and/or service provision in the corresponding policy field.

Moreover, within public administrations, organization can take a number of forms. While categorization may vary from government to government, in general terms, five types of organization may be apparent. First, there are central agencies dedicated to coordinating and supporting government effort, responsible for areas such as financial and human resource management and management improvement. Second, there are departments charged with the responsibility for direct service provision to the public, across a broad range of policy fields. Third, there are government business enterprises (GBEs) which operate at a distance from government, in a more business-like way, usually providing some essential or commercial service. Fourth, there are the review

and regulatory agencies, such as auditors-general, ombudsmen [*sic*], and anti-corruption and whistle-blowing protection agencies. Fifth, there are the more peripheral boards and agencies, often semi-autonomous entities that conduct other, sometimes more obscure, aspects of government business.

In general, these organizational aspects still apply, but the public administration–bureaucratic paradigm which endured for over a hundred years has to a significant extent been rejected by governments in postmodern times. By the 1960s, especially in Anglo-American polities, it was apparent that the modern paradigm of public administration, as it had evolved, was no longer serving governments, as intended. Public administrations were seen by critics and reviewers to be too big, too inefficient, and too powerful in comparison to governments. These criticisms were also linked to the perceived failure of Keynesian economic principles, which had underpinned governments' strategic policy decisions since the 1930s.

1.2 Postmodern Public Administration, Organizational Aspects (1979–Present)

It took until the late 1970s and early 1980s for fundamental changes to be implemented in a major paradigm shift toward a market-based model of public administration. Conservative political leaders, Prime Minister Thatcher in Britain and President Reagan in the USA were the first politicians demonstrably to sustain such an approach, which led to adoption in other Anglo-Saxon polities. Similar earlier attempted changes had been evident in Latin America, where free-market concepts of the Chicago School were influential for a relatively short time. The market-based model, in varying forms, has been adopted in over 100 countries, worldwide, with some convergence of ideological thought between traditionally conflicting political parties. The market-based model supplanted communism in Eastern Europe in the late 1980s as a guiding ideology and has led to revolutionary change in organizational form in other largely totalitarian states.

Supranational governance structures, involving organizations such as the International Monetary Fund (IMF), the World Bank, the Organization for Economic Cooperation and Development (OECD), the Group of 7–8 leading trading nations, the World Trade Organization (WTO), and the European Union (EU) have all supported and promoted the market-based model. This may account for the active transfer and adoption processes in place from polity to polity. The EU, as the most comprehensive organizational form for supranational governance, operates with its own parliament and public administration system.

Several common themes based partly on public choice, and other less well defined neoclassical economic theories of rational choice are characteristic of the market-based model. They all have implications

for public administration and organizational aspects. First, for example, there is a move away from a mixed economy to a firm belief in markets as the primary and more efficient economic mechanism. To this end, deregulation, relating to trade, finance, and monetary policies has taken place.

Second, if markets are to prevail, then the size of governments and their public sectors needs to be reduced. Included in the radical downsizing of public sectors is commercialization, and/or the privatization of GBEs where it is deemed inappropriate for governments to continue to operate in these areas. Major utilities and commercial businesses traditionally involving governments, in areas of activity such as telecommunications, power generation, distribution and transmission, water, banking, insurance, postal services, and airlines may be targeted.

Third, and most importantly, for the public sector, the strategic focus is on organizational efficiency, the creation of internal market-style competitive conditions and the more purposive application of private-sector business techniques to public management. These changes have involved a shift from a focus on processes of administration and policy, however flawed, to organizational management. A business, or corporation, metaphor may be used to define organizational form.

As a subset of public choice theory, principal–agent theory focuses on the relationships between governments and their top bureaucrats. Public bureaucrats are perceived to be budget-maximizing, self-interested agents, not trusted by governments. Therefore, governments need to develop executive structures where principal–agent theory can be applied directly so that governments can compel their administrative elite to act in the government's interests. The establishment of formal senior executive services, with performance incentives and fixed-term contracts may result.

Organizationally, the acceptance of the market-based model by governments has seen other major changes. These include, first, substantial machinery of government changes including the separation of policy, regulatory, commercial, and operational services and programs into discrete organizations. Second, the purchaser/provider split, where governments are the contract managers and service provision is contracted out. Third, more openly political appointments at the top of the public service are apparent in previously nonpartisan services. Frequently, this has involved the centralization of power away from independent personnel bodies, which have been abolished. These senior executive managers are expected to support publicly governments' electability. In such circumstances, the possibility of ministers accepting responsibility for departmental managerial failure, within the construct of ministerial responsibility, considerably diminishes.

The rational choice logic, which underpins the market-based model, is open to question, given the

presupposition of perfect markets and exchange processes. Such a conceptual framework is unrealistic for the political environment of public administration. Rewards for certain actions and behaviors of agents, required by the principals, can be achieved through both formal and informal power structures. This is an important organizational aspect of the market-based model, where parallel systems of exchange, reward, and relationships operate as a matter of course. While the formal systems may promote a constancy of action and process, the informal system can support implicitly institutionalized power relationships.

Whether the paradigm shift has resulted in better public administration and organizational aspects is still to be determined as the key economic indicators of performance, predicated within this new paradigm, suggest a mixed result. There has certainly been a shift from idiosyncratic or confined interpretations of public administration, relevant to particular nations, regions or systems, to a more universal construct. Nevertheless, there are many variations on the market-based theme depending on governments in power and individual leaderships at any given time.

From an organizational aspect, one of the weaknesses of the market-based model is that it focuses strongly on economic issues and individualism, rather than on the collective responsibility of the state to its citizens. In this sense, social and related policy fields may be diminished in their importance. Some governments are attempting to redress this imbalance with more attention to social, environmental, and other imperatives, relating to the development of human and social capital. Furthermore, the continuing growth of supranational structures in terms of greater organizational form, and advances in information technology, now provide new challenges for governments, and another likely significant paradigm shift. The organizational shape and size of public administrations into the future remains uncertain.

1.3 Virtual and Supranational Public Administration, Public Administration (Now and into the Future)

Two developing areas of interest will undoubtedly have major organizational impacts on public administration in the future. These are, first, more virtual governments, and, second, the power of supranational organizations. With developments in information technology and the related rise of the global new economy, governments are adding more virtual organizational forms to public administrations. For example, especially with customer service initiatives, knowledge management becomes increasingly important. Free information lines, supported by computer technology, provide coordinated details of government services. Similarly, one-stop shops allow citizens to attend to a range of government requirements, such as license renewals and the payment of bills, at a single

location. Comprehensive information can be obtained from government Internet websites, and civil servants can be contacted, more informally and directly, by e-mail. E-commerce, relating to procurement, contracting, and the lodging of documents may now apply. Within governments, more virtual communication also takes place involving typed, aural, and visual media. Undoubtedly, this will continue to impact upon organizational aspects of public administration in the future.

Second, the continuing development of supranational governance structures, in the context of globalization, has led to both cogent and conspiratorial debate in recent years about future organizational aspects of public administration. The Parliament of the EU as a structure challenging the sovereignty of national member governments, is an example of how future organizational aspects may evolve as other trading coalitions consolidate their form. However, responses to globalization, especially negative ones, also prompt support for localism, which emphasizes the role of individual communities.

With all these complex local and global pressures, it is likely that organizational aspects of public administration will experience quite dramatic change in the first decades of the twenty-first century. Whether the more traditional structures of agencies, departments, trading enterprises, and review bodies, with the complex relationships between key actors, will remain as the primary organizational aspects is difficult to predict. Even if they do remain, the more virtual aspects of public administration are expected to endure and grow, to the point where there will be far more virtual public administration than there is at the time of writing.

See also: Administration in Organizations; Bureaucracy and Bureaucratization; Civil Service; Organization: Overview; Organizations and the Law; Organizations: Authority and Power; Organizations, Sociology of; Parliamentary Government; Planning, Administrative Organization of; Public Administration, Politics of; Public Bureaucracies; Public Sector Organizations; Weber, Max (1864–1920)

Bibliography

- Aucoin P 1995 *The New Public Management CANADA in Comparative Perspective*. Institute for Research on Public Policy, Ottawa, Canada
- Barberis P (ed.) *The Civil Service in an Era of Change*. Dartmouth, Aldershot, UK
- Barratt Brown M 1995 *Models in Political Economy*. Penguin Books, London
- Beloff M 1959 *The American Federal Government*. Oxford University Press, London
- Boston J, Martin J, Pallot J, Walsh P (eds.) 1991 *Reshaping the State*. Oxford University Press, London

- Campbell C, Wilson G 1995 *The End of Whitehall*. Blackwell, Oxford, UK
- Clegg S 1989 *Frameworks of Power*. Sage, London
- Eldridge C C 1978 *Victorian Imperialism*. Hodder & Stoughton, London
- Galbraith J K 1994 *The World Economy Since the Wars*. Mandarin, London
- Halligan J, Power J 1991 *Political Management in the 1990s*. Oxford University Press, Melbourne, Australia
- Hood A 1998 Public officials, government and the public trust. *Australian Journal of Public Administration*. 57(1), March: 98–114
- Hughes O 1994 *Public Management and Administration*. St. Martin's Press, New York
- Hutton W 1995 *The State We're In*. Jonathan Cape, London
- Institut Internationale d'Administration Publique 1996 An introduction to French administration. La Documentation française. Institut Internationale d'Administration Publique, Paris
- Kingdom J 1991 *Government and Politics in Britain*. Polity Press, Cambridge, UK
- Koh B 1989 *Japan's Administrative Elite*. University of California Press, Berkeley, CA
- Laffin M, Painter M (eds.) 1995 *Reform and Reversal: Lessons from the Coalition Government in New South Wales 1988–1995*. Macmillan, South Melbourne, Australia
- Lynn L E Jr 1996 *Public Management as Art, Science and Profession*. Chatham House, NJ
- Northcote–Trevelyan Report 1853 *Report on the Organisation of the Permanent Civil Service*. Appendix B in Fulton, Lord (Committee under the Chairmanship of) 1968 *The Civil Service*. Vol. 1. HMSO, London
- Ormerod P 1994 *The Death of Economics*. Faber and Faber, London
- Peters B G 1996 *The Future of Governing: Four Emerging Models*. University Press of Kansas, St Lawrence
- Pollitt C 1993 *Managerialism and the Public Services*, 2nd edn. Blackwell, Oxford, UK
- Pusey M 1991 *Economic Rationalism in Canberra*. Cambridge University Press, Cambridge
- Rainey H 1991 *Understanding Public Organizations*. Jossey-Bass, San Francisco, CA
- Savoie D 1994 *Thatcher, Reagan, Mulroney*. University of Toronto Press, Toronto, Canada
- Self P 1993 *Government by the Market?* Macmillan, London
- Shafritz J, Hyde A 1992 *Classics of Public Administration*. Brooks/Cole, Pacific Grove, CA
- Spann R N 1979 *Government Administration in Australia*. Allen & Unwin, Sydney, Australia
- Sturgess G 1993 Can parliament cope? Towards a new era in public accountability, *Canberra Bulletin of Public Administration*, 7, September: 45–52
- Taylor F 1912 *Scientific Management* (Reprinted in Shafritz J, Hyde A 1992 *Classics of Public Administration*). Brooks/Cole, Pacific Grove, CA
- Wanna J, O'Faircheallaigh C, Weller P 1992 *Public Sector Management in Australia*. Macmillan, South Melbourne, Australia
- Weber M 1946 *Bureaucracy*. Reprinted in Shafritz J, Hyde A 1992 *Classics of Public Administration*. Brooks/Cole, Pacific Grove, CA
- Wilenski P 1986 *Public Power and Public Administration*. Hale and Ironmonger, Sydney, Australia

- Wilson W 1887 *The Study of Administration*. Reprinted in Shafritz J, Hyde A 1992 *Classics of Public Administration*. Brooks/Cole, Pacific Grove, CA

J. Johnston

Public Administration, Politics of

From its self-conscious beginnings in the late nineteenth century until the years around World War II, the practice of public administration in America, conceived then, as now, as the organization of government and the management of its activities, was regarded as a rational, technical enterprise properly pursued entirely apart from the give and take of politics. Working against the backdrop of civil service reform, those who sought initially to define the field in both a descriptive and prescriptive sense were certain that public administration could only be accomplished by strict adherence to principles of neutral competence and efficiency rather than partisanship, compromise, and preference. In this view, representative political institutions, whose function is to weigh competing interests and make choices for the society, set the tasks of the bureaucracy. In his seminal essay of 1887, *The Study of Administration*, Woodrow Wilson wrote of the possibility of a 'science of administration,' maintaining that 'administrative questions are not political questions' (Wilson 1941). In the crucible of the Great Depression of the 1930s and the war that followed, however, challenges to the politics-administration dichotomy began to take form, and by 1950 John Gaus (1950) could declare that 'a theory of public administration means in our time a theory of politics also.' Today, there is no controversy: no one would deny that public administration is a decidedly political process.

1. How Public Administration is Political

There are two general ways in which public administration may be regarded as a political process. One is that by their actions public administrators make value choices that allocate benefits and costs authoritatively; these choices are likely therefore to be contested. The other is that public administrators, far from being passive agents of representative political institutions, act strategically, seeking to maximize the influence of their agencies by competing with other actors for resources and by building alliances.

1.1 Administration as the Pursuit of Values

The organization of government and the implementation and evaluation of public policy are carried out

in the context of certain normative standards or values that guide both management strategies as well as day-to-day actions. The origin of these standards lies normally with the legislative or judicial branches of government or the elected executive. They include, among others, injunctions to pursue the business of government with economy, efficiency, equity, responsiveness, transparency, and accountability.

Many of these standards are in conflict with one another, however: economy and equity may not be attainable simultaneously, for example, if the former stresses cost-savings at the expense of broad distribution of benefits. Nor is efficiency, which may require swift and confidential action, always compatible with transparency, which may entail open decision-making and citizen participation. Furthermore, each of these standards is susceptible to interpretation as it relates to any given concrete situation. Since representative political institutions may not provide specific guidance about priorities or offer ways to resolve ambiguities as these standards are applied in the field, public administrators must necessarily make choices and render interpretations that inevitably create advantages for some interests and deny them to others. Thus, as Carl Friedrich (quoted in Lynn 1996) once observed, 'Public policy is being formed as it is being executed ... In so far as particular individuals or groups are gaining or losing power or control in a given area, there is politics.' To the extent that politics is the process of deciding who gets what, then public administration is a political activity.

1.2 Public Administrators as Strategic Actors

Through the aggrandizement of resources and the solidification of alliances, managers of bureaucratic agencies strengthen their ability to exercise influence in the shaping and implementation of public policy. Selznick's study of the Tennessee Valley Authority (1949) was among the first to suggest the importance of establishing external allies (such as client groups, interest groups, the news media, or members of Congress) as a key to the survival and prosperity of a bureaucratic agency. Subsequent observers have seen the efforts of bureaucratic agencies to build a resource base and foster external support as creating new competitors for power with legislative bodies in government (Rourke 1987). Another perspective, however, has emphasized the mutually reinforcing relationships that bureaucratic agencies at the national level tend to forge with interest groups and the legislative committees critical to their funding and authority. These relationships are variously referred to as 'iron triangles' or, to suggest the greater fluidity with which these alliances evolve, 'issue networks' (Heclo 1979).

2. The Public Management Perspective

The political nature of public administration has taken on sharper focus with the emergence in the last quarter of the twentieth century of the subdiscipline of public management. Whereas traditional public administration was concerned primarily with the organization, design, and operation of bureaucratic institutions, public management is more interested in how managers actually manage public policy (Lynn 1996). Kettl (1996) attributes the rise of public management as a field of study and practice in part to research on policy implementation that began in the 1970s. These investigations suggested that policy success and failure lay not so much with the laws that established programs, or even with the design of the organizations responsible for implementation, as with the skills and capacities of the managers. Some were more politically capable than others. Lynn (1996), for example, writes of 'the politics of management,' by which he means the ability of managers of public agencies to set a course, maintain credibility with overseers, marshal authority and resources, and position the organization advantageously in the political environment. Thus, the public management perspective focuses explicitly on strategy and interorganizational relations rather than the administration of processes internal to bureaucratic agencies.

3. The Politics of Public Administration in Action

The politics of public administration are revealed in the routine processes of government organization and operations, such as reform, personnel administration, and policy implementation.

3.1 Reform

Reformers have made repeated attempts during the twentieth century to transform American government to make it more effective and more responsive to public desires. These efforts have often focused on reorganization schemes whose effect was to redistribute power within government. For example, the creation of an executive budget at the national level in the United States in the 1920s and the establishment of an Executive Office of the President in the postwar period both served to strengthen the presidency in relation to Congress. In more recent times, the movement to 'reinvent' government has sought to shift power both from Congress to the bureaucracy and from the higher reaches of the bureaucracy to mid-level managers and line functionaries (Kettl 1998). Not only has the reinvention movement aimed at 'empowering' government functionaries at the lower rungs of the bureaucracy, but it has also attempted to

recast citizens as customers rather than supplicants. Since customers may exercise their right not to buy, to 'exit,' the successful market actor must be attentive to their desires. The reinvention movement at the end of the twentieth century is, thus, partly an effort to transform the culture of public administration by conferring on citizens seeking service from government rights and powers—respect, attentiveness, quality—similar to those enjoyed by customers in the private market (Osborne and Gaebler 1992).

3.2 Personnel Administration

Hiring public employees, defining job requirements, motivating workers to perform, and appraising their achievements are standard tasks of the public administrator. As Levine et al. (1990) have pointed out, public executives seek to achieve a diverse set of values in personnel administration, including instrumental achievement, social equity, merit, and employee rights. Often these may be in conflict with one another. This is nowhere more apparent than in the debate over affirmative action and descriptive representation in the public workforce.

Holden (1996) has observed that no theory of public administration can fail to take account of competition among ethnic groups in a plural society. Public employment has long been a prize over which ethnic and racial groups have contended, for work in government has provided not only a measure of economic security but also advancement. The need to respond to the political demands of groups underrepresented in public jobs may be seen as clashing at times with the exercise of strict merit standards. Public administrators face the dilemma of resolving this value conflict as they seek to manage their agencies, posing clear political choices. Elected officials also struggle with this issue, and those who resolve it by committing to affirmative action illustrate how political considerations may merge with and shape the public administration of personnel management.

3.3 Policy Implementation

One of the most important factors accounting for the political nature of the administrative process is the degree to which government functionaries necessarily exercise discretion in implementing laws and administrative directives. When state highway department planners choose to establish a right-of-way for a new road through a poor neighborhood rather than a middle-class area, or decide to locate an interchange near one suburban commercial development rather than another, it is clear that the implementation of highway policy is not simply a matter of the application of engineering and administrative con-

siderations; management choices have distributive consequences. They create winners and losers.

Discretion may be more or less circumscribed. Schneider and Ingram (1997) point out that the allocation of discretion in any given policy area varies from the situation in which the laws and rules establish unambiguous goals, means, and target populations to a situation in which the law is so general as to leave all of these in great measure to the judgment of government employees in the field. At the lower levels of the organization so-called 'street-level bureaucrats' (Lip-sky 1980), such as police officers, welfare case workers, health and safety inspectors, and other government employees who deal on a face-to-face basis with the public, often labor under conditions that conform more to the latter end of the spectrum than the former. The choices of these bureaucrats are fraught with political meaning: police officers who stop automobiles driven by young black men more often than those driven by whites simply in the belief that blacks are more likely to be criminals, or caseworkers whose sympathies lead them to facilitate applications by elderly clients for food stamps but not for single mothers, serve as key actors in the allocation of policy benefits and costs.

Discretion may even lead to the sabotage of laws with which bureaucratic officials do not agree. For example, in the mid-1970s the secretary of the US Department of Agriculture resisted implementing the Women, Infants and Children supplemental food program, claiming that such assistance was both duplicative of other programs and ineffective. Delays in the implementation of the program led to congressional claims that the department was breaking the law.

4. The Administration–Politics Synthesis

To reformers in the early twentieth century, for whom the business corporation provided the preferred template for the administrative state, politics was understood as the competitive struggle for partisan advantage. This struggle served its purpose in the selection of leaders and the adoption of policy, but in an ideal society, administration, the carrying out of political decisions as well as the housekeeping functions of government, such as street paving, would be separated from partisan conflict and placed in the hands of disinterested professionals, beholden neither to party nor politician, but to rational principles of management. Yet in the century since these ideas took form, students of the administrative process came to see that the operational details of organization and management were themselves deeply political.

Public administration has come to be understood as the continuation of the political process—the authoritative allocation of values—by government actors

in government settings other than the elected institutions. To the extent that public administrators make choices in organizing their bureaus, in choosing their personnel, in establishing operational priorities, and most important, in applying the laws, they distribute costs and benefits across groups and interests in the society. In that sense, public administration is a political activity.

See also: Accountability: Political; Administration in Organizations; Bureaucracy and Bureaucratization; Management Accounting; Organizations, Sociology of; Planning, Administrative Organization of; Public Administration: Organizational Aspects; Public Bureaucracies; Public Sector Organizations

Bibliography

- Gaus J 1950 Trends in the theory of public administration. *Public Administration Review* 10: 161–68
- Heclo H 1979 Issue networks and the executive establishment. In: King A (ed.) *The New American Political System*. American Enterprise Institute, Washington, DC
- Holden M 1996 *Continuity & Disruption*. University of Pittsburgh Press, Pittsburgh, PA
- Kettl D 1996 Introduction. In: Kettl D, Milward H B (eds.) *The State of Public Management*. Johns Hopkins University Press, Baltimore, MD
- Kettl D 1998 *Reinventing Government: A Fifth Year Report Card*. Brookings Institution, Washington, DC
- Levine C, Peters B G, Thompson F 1990 *Public Administration: Challenges, Choices, Consequences*. Scott, Foresman, Glenview, IL
- Lipsky M 1980 *Street Level Bureaucracy*. Russell Sage, New York
- Lynn L 1996 *Public Management as Art, Science and Profession*. Chatham House, Chatham, NJ
- Osborne D, Gaebler N 1992 *Reinventing Government*. Addison Wesley, Reading, MA
- Rourke F 1987 Bureaucracy in the American constitutional order. *Political Science Quarterly* 102: 217–32
- Schneider A, Ingram H 1997 *Policy Design for Democracy*. University of Kansas Press, Lawrence, KS
- Selznick P 1949 *TVA and the Grassroots*. University of California Press, Berkeley, CA
- Wilson W 1941 The study of administration. *Political Science Quarterly* 56: 481–506

P. Eisinger

Public Broadcasting

The history of broadcasting, beginning in the early 1920s, has seen three main types or models of broadcasting organizations. One is the commercial model, in which advertising is used to provide

revenues, with programs seen as vehicles for providing an audience to which products can be sold. A second model is one in which the broadcasting organization is funded by the state, usually so that the state can project its political message and thus influence the public. The third model is known in the USA as public broadcasting but in most of the rest of the world as public service broadcasting. The most famous and influential example of this model is the BBC. It is this model with which this article is concerned, particularly by defining it in ways that will allow the reader to see the fundamental difference between it and the other two models.

1. The General Concept and Background of Public Broadcasting

In a public system, television and radio producers acquire money to make programs. In a commercial system they make programs to acquire money. However simple, this little epithet articulates the divergence of basic principles, the different philosophical assumptions, on which broadcasting is built. History and experience fashioned inside public broadcasting a definable canon, a set of principles and practices which constitute the core theses around which the institution has been formed and which have guided its performance.

The particular institutional structures and forms of funding may vary from place to place, but public broadcasting is above all else an ambition, a belief that the sheer presence of broadcasting within all our lives can and must be used to nurture society, to proffer the opportunity that society and its inhabitants can be better served than by systems which primarily seek consumers for advertisers or subjects for the state.

Some of the most powerful visions of the purpose of broadcasting emerged within unusual and trying circumstances. Consider, for instance, the cultural histories of the occupations of Germany and Japan in the late 1940s and the formulation of Allied policy for broadcasting in the rebuilding of those societies. Here was testament to the idea of broadcasting as primarily a social rather than an economic process, as something with moral, cultural, intellectual, and creative purpose and not just as a source of mild comment and moderate pleasure. The charters that established broadcasting in Japan and in Germany, by the Allied occupying powers, were replete with the public service ideal. If broadcasting was to comment, it should do so with a flourish. If it was to amuse, it should do so with elan. If it was to educate, it should do so with real professionalism. It was believed by the Allied leadership, including the USA, that the life of the mind of a society was far too precious and important to be left to what they took to be the vagaries of a commercial system or the oppressive force of state control.

It could be argued that such policies were creatures of the moment, as massive destruction demanded enormous reconstruction, of which communications would inevitably be part. But what was required was the restoration not just of highways, buildings, plants, but also of the shattered imaginative lives of whole populations. The architects of postwar Germany and Japan sensed correctly that healthy, diverse cultural institutions were a prerequisite to a functioning liberal democracy. Broadcasting was thus to be used as a key part of the cultural and social regeneration of those societies. In that lies the real clue to the nature and purpose of great public broadcasting: that it makes best sense when it represents a national and moral optimism within a society, when it suggests through the diversity and quality of its programs that societies be better than they are, better served, better amused, better informed, and, thus, better societies.

A period which is widely regarded within the advanced industrial societies as a high water mark of public service broadcasting, is the BBC in the early 1960s. A key figure from those years was Sir Arthur fforde, possibly the greatest of the chairmen of the BBC. In 1963 he wrote: 'By its nature broadcasting must be in a constant and sensitive relationship with the moral condition of society.' He felt that the moral establishment had failed modern society and that broadcasting was a way in which that failure could be rectified. He added that it 'is of cardinal importance that everyone in a position of responsibility should be ready to set himself or herself the duty of assuring, to those creative members of staff... that measure of freedom, independence and elan without which the arts do not flourish' (quoted in Tracey 1983, p. 235).

That idea of providing a protective layer within which the imaginative spirit might create, lay at the heart of the BBC version of public service broadcasting which flourished in the postwar years and which constituted a model that other nations sought to emulate. Ian Jacob, Director-General of the BBC from 1952 to 1959, refined the notion. In 1958, in an internal document called 'Basic Propositions', he described public service broadcasting as:

... a compound of a system of control, an attitude of mind, and an aim, which if successfully achieved results in a service which cannot be given by any other means. The system of control is full independence, or the maximum degree of independence that Parliament will accord. The attitude of mind is an intelligent one capable of attracting to the service the highest quality of character and intellect. The aim is to give the best and the most comprehensive service of broadcasting to the public that is possible. The motive that underlies the whole operation is a vital factor; it must not be vitiated by political or commercial consideration.

This is one of the best attempts to capture in words a concept and view of broadcasting which continues to circulate in the world of cultural politics. Yet even here the vision, the articulation, is limited. Jacob's words

imply that we understand the nature of public service broadcasting not by defining it, but by recognizing its results, rather as one plots the presence of a hidden planet or a subatomic particle not by 'seeing' it, but by measuring the effects of its presence. The Pilkington Committee, on the future of British broadcasting, said as much in 1962 when it observed: 'though its standards exist and are recognizable, broadcasting is more nearly an art than an exact science. It deals in tastes and values and is not precisely definable' (Home Office 1962, p.13, para. 34).

One obvious characteristic of efforts to define public service broadcasting is the elegant vagueness that is studied and deliberate. In effect what is suggested is that one thing that has to be understood about defining public service broadcasting is that in any abstract sense one cannot.

The beast that lurks in the shrubbery of those discussions is that whatever the definitional uncertainties, in that public service broadcasting can be experienced and recognized but never properly captured by language, someone has to decide on what is 'good' and 'bad,' and there should be a guiding hand, what has been referred to as 'the custodians, the caretakers' of culture. As traditionally understood by public broadcasters the whole idea of custodianship was totally unproblematic, as was understanding just who the custodians were, the characteristics they would possess and the locations in which they would be found. Such presuppositions appeared only natural throughout the history of broadcasting because that history was embedded within a social order in which hierarchy was also assumed: hierarchies of social status and cultural judgment. And in a curious kind of way the point of hierarchy is to reproduce itself, since the fundamental belief of the hierarchical system is, and has to be, that such arrangements have worth and merit. The plausibility of such a thesis becomes very much dependent on delivering evidence that attests to both worthiness and meritoriousness in such a way as to drown out the noise of any emergent countervailing thesis.

A key justification for the custodial role in most societies where public service broadcasting was established was that since the radio frequency used for transmissions was a limited natural resource, someone had to ensure that its use served the public good, the whole community. The cultural geology of this decision had, however, a deeper level to it, based on nineteenth century assumptions about the ways in which the arts and humanities could elevate the human condition. In an essay on the formation of modern literary analysis, Steiner wrote that behind the study of literature in the nineteenth century lay:

a kind of rational and moral optimism... a large hope, a great positivism... The study of literature was assumed to carry an almost necessary implication of moral force. It was thought self-evident that the teaching and reading of the great poets and prose writers would enrich not only taste or style

but moral feeling: that it would cultivate human judgment and act against barbarism (Steiner 1985, pp. 77–8).

Reading this account one could quite properly substitute the word ‘broadcasting’ for ‘literature’ and have a powerful explanation of what the creation of the model of public service broadcasting was all about: a relocation of a nineteenth century humanistic dream of human betterment. The fear that drove that dream was of ‘the mob,’ the pervasive belief among cultural, religious, and political elites that there was indeed a dark side to the human soul that was, when let loose, dangerous and devastating to the flesh as well as the spirit. Whatever the apparent elitism, who is to say that they were wrong nestling, as they did, between the first great war and a looming second. There remained, however, a residual faith, tied to the whole condition of the Enlightenment, humanism, and belief in progress, that popular culture need not be debauched but could in fact transcend itself. Consider these key passages from the Pilkington Report:

Television does not, and cannot, merely reflect the moral standards of society. It must affect them, either by changing or by reinforcing them... Because the range of experience is not finite but constantly growing, and because the growing points are usually most significant, it is on these that challenges to existing assumptions and beliefs are made, where the claims to new knowledge and new awareness are stated. If our society is to respond to the challenges and judge the claims, they must be put before it. All broadcasting, and television especially, must be ready and anxious to experiment, to show the new and unusual, to give a hearing to dissent. Here, broadcasting must be most willing to make mistakes; for if it does not, it will make no discoveries (Home Office 1962, pp. 19–20, paras. 50–3).

The suggestion here is not that public broadcasters are all hoping and dreaming that their programs will transform people from cultural and intellectual slobes into something of which one can more readily approve, but rather that objectively some such argument must be the public broadcaster’s last line of defense. The language is of standards, quality, excellence, range. The logic is of social enrichment, that in however indefinable a manner this society is ‘better’ for having programs produced from within the framework of those social arguments termed public service broadcasting, compared to those programs produced within an environment in which commerce or politics prevail.

One cannot, however, escape the charge that those sentiments rest on a set of ill-explored assumptions about the sociological organization of modern culture. How are we better? What are the mechanics? And, vitally, where is the evidence? Indeed, what would such evidence even look like?

2. The Public Service Idea: Eight Principles

What then are the definable principles of public service broadcasting? Eight principles are outlined below.

2.1 Universality of Availability

Public broadcasting has historically sought to ensure that its signals are available to all, that no one should be disenfranchised by distance or by accident of geography. The imperative that guides this principle is not that of maximizing customers in a market but of serving citizens in a democracy, that if one defines one’s audience as the citizens of a country, then logically one has to reach them all. To a remarkable extent in country after country this principle has been made real.

2.2 Universality of Appeal

Public broadcasting seeks to provide programs that cater to the many different tastes and interests that constitute a society’s life. The public broadcasting community understands that each of us, at different moments, is part of a majority and a minority. In seeking to provide programs for a wide range of tastes and interests, public broadcasting does so with an eye cocked to the need to ensure that whether the program is pitched at the many or the few it is done so with real quality. Public broadcasting does not expect that it can please all the people all of the time—indeed it sees in that approach precisely the kind of populism that nurtures cultural mediocrity, as quality is sacrificed on the altar of maximizing the audience size.

It is an important element of this principle that public broadcasting serves not only tastes and interests that are readily apparent, but also those that are dormant and latent—that may be part of the potential we all possess but that circumstance may not have allowed us to develop.

2.3 Provision for Minorities, Especially Those Disadvantaged by Physical or Social Circumstance

It is commonplace to characterize the medium of television as essentially serving ‘the mass.’ Certainly public broadcasting understands the vast capability of one medium to reach enormous numbers of people. It sets its face, however, against the logic of commercial systems to see people as no more than statistics in skins, with a definable value captured in the most desirable rates, demographic buys, and cost per thousand. As suggested in the second principle, public broadcasting views the public as a rich tapestry of tastes and interests, each of which, insofar as possible, should be served.

In particular it recognizes that there are whole subcultures of minority social experiences crying out for attention. People of different color, language groups, and religious preferences all have vital needs for expression in the political and social discourse of the nation. Public broadcasting is dedicated to a dual role here—on the one hand to give access to such groups, to provide them with the opportunities to

speak to one another and to voice the issues as they see them, and on the other to provide coverage of their histories, interests, and concerns for the public at large.

2.4 *Serving the Public Sphere*

Some television programs are successful because they get a fair-sized audience, make some money, and sometimes even exemplify the craft of popular television. Other programs are successful because they reach out and touch a small, particular, but powerful audience. Some programs are successful because the craft of the program maker is used to speak to us all. They touch us, move us, make us laugh and cry and cheer. They speak to us because they speak for us. Like all great art, they help us make sense out of life, to see and understand things with a fresh eye, and give us a burning sense of the collective, of belonging to the nation-as-community. It is an increasingly vital principle of the work of public broadcasting that it recognizes its special relationship to a sense of national identity and broad community. Any nation is a patchwork of localities and regions, but it also is a nation, heterogeneous and homogeneous to a remarkable degree at one and the same time. Public broadcasting's very nature is then to nurture the public sphere as a means of serving the public good. It understands that while within civil society individuals pursue their own private self-interests, it is within the public sphere that they function as citizens. It is a fundamental principle then that public broadcasting must motivate the viewers as citizens possessing duties as well as rights, rather than as individual consumers possessing wallets and credit cards.

2.5 *A Commitment to the Education of the Public*

The most outstanding example of public broadcasting's commitment to the audience-as-citizen is the long-time provision in almost all systems of educational programming at every level. Public broadcasting knows that political and social literacy, as well as of course literal literacy, is an essential prerequisite to the healthy working of a democratic order. Above all else, the commitment to this principle requires that it treat its audience as mature, rational beings capable of learning and growing in many ways. Thus much of public broadcasting has retained its commitment to institutional services. Daytime school broadcasting and formal learning services of all kinds continue to play a role in most national services.

2.6 *Independence from Vested Interests*

It is a simple but key principle of public broadcasting that its programs can best serve the public with excellence and diversity when they are produced from

within a structure of independence. Programs funded by advertising necessarily have their character influenced in some shape or form by the demand to maximize the garnering of consumers. Programs directly funded by the government, and with no intervening structural heat shield, inevitably tend to utter the tones of their master's voice.

The whole history of public broadcasting has been dominated by the idea that it can best serve the nation when it remains distanced from any particular commitment to any particular power structure inside the nation. Of particular importance to this principle is the ability of public broadcasting to support a cadre of independent-minded program makers, who are thus well able to speak with authentic tones and to offer that singularity of vision allied to creativity and passion that has traditionally produced some of public television's finest moments. It follows that the political and economic architecture of this principle is such as to support the making of programs that are good in their own terms, whatever their intended audience. In the making of programs for public broadcasting, there should be no ulterior purpose or motive. It is axiomatic to this principle that the funding of public broadcasting should be such, in total amount and in the absence of any strings attached, as to encourage rather than negate the independence enjoyed.

2.7 *Encouraging Competition in Good Programming Rather Than Competition for Numbers*

This principle is central to public service broadcasting and essentially involves a commitment to making programs which, whatever their intended audience, are of high quality. The overwhelming brunt of the evidence leads to the conclusion that the most important aspect of such structuring relates to the forms of finance. Where commercial sources of revenue are dominant, or even present, or where there is direct subvention from government, the program maker's eye is almost inevitably diverted away from what should be the main focus, the inherent quality of the program he or she is making.

2.8 *The Rules of Broadcasting Should Liberate Rather Than Restrict the Program Maker*

While all broadcasting will inevitably be governed by certain prescriptions—'educate,' 'inform,' 'entertain,' 'balance,' 'fairness'—and certain broadly drawn restrictions—obscenity, national security—the essence of the legislative foundation by which it is empowered should sustain a liberal function for the program maker. The legislation should create secure living space, arena for action, for broadcasters with all kinds of interests in possible programs and possible varieties of audience, rather than leaving the field to those who are interested chiefly in delivering maximum audiences most of the time. The legislation should also ensure

that the higher echelons of broadcasting contain individuals who understand its potential and who themselves care for the importance of the creative work of their staff. Part of that understanding would be the need for experiment and innovation in broadcasting, the need to provide a focus for a society's quarrel with itself, the recognition that mistakes will be made but as such may signify the health of the system.

Perhaps above all else, such leadership should be helped to understand that experiment, innovation, quarrel, and mistake are likely to come from the younger program maker, without whom the system is in danger of institutional arteriosclerosis.

See also: Broadcasting: General; Broadcasting: Regulation; Mass Media, Political Economy of; Mass Media, Representations in; Media, Uses of; Public Relations in Media; Public Sphere and the Media

Bibliography

- Avery R L (ed.) 1993 *Public Service Broadcasting in a Multi-channel Environment: The History and Survival of an Ideal*. Longman, New York
- Briggs A 1985 *The BBC: The First Fifty Years*. Oxford University Press, Oxford, UK
- Bullert B J 1997 *Public Television: Politics and the Battle Over Documentary Film*. Rutgers University Press, New Brunswick, NJ
- Home Office 1962 *Report of the Committee on Broadcasting (Pilkington Report)*. HMSO, London
- Jarvik L A 1997 *PBS, Behind the Screen*. Forum, Rocklin, CA
- Ledbetter J 1997 *Made Possible By: The Death Of Public Broadcasting in the United States*. Verso, New York
- Skene W 1993 *Fade to Black: A Requiem for the CBC*. Douglas and McIntyre, Vancouver, BC
- Smith A, Paterson R (eds) 1998 *Television: An International History*. Oxford University Press, New York
- Stewart D 1999 *The PBS Companion—A History of Public Television*. TV Books, New York
- Stuart C (ed.) 1975 *The Reith Diaries*. Collins, London
- Tracey M 1983 *A Variety of Lives: A Biography of Sir Hugh Greene*. Bodley Head, London
- Tracey M 1998 *The Decline and Fall of Public Service Broadcasting*. Oxford University Press, Oxford, UK

M. Tracey

Public Bureaucracies

Throughout history public bureaucracies have served as a source of continuity and stability in the life of the state. This role has sometimes led to their becoming a significant obstacle to innovation in government policies. More recently, however, these bureaucracies have also become useful agents of change in many societies.

The power that bureaucrats exercise comes mainly from the expertise they command both as individuals

and when linked together in organizations. This power has come under increasing challenge in the industrialized democracies. Other institutions—governmental and nongovernmental—have begun to acquire their own expertise, and public bureaucracies are less able to cloak their operations in secrecy. Moreover, many public agencies that formerly performed regulatory or service delivery functions have been abolished, and private or quasi-public organizations have begun to perform tasks that were once within the exclusive preserve of government agencies. By the end of the twentieth century, it had become clear that public bureaucracy was not the growth industry it had appeared to be in the aftermath of World War II, fifty years earlier.

1. Bureaucracy: the Permanent Government

Although perhaps the least venerable, public bureaucracies rank among the oldest of all existing government institutions. Such bureaucracies played a central and continuing role in the operation of two of the largest and the most effective of the empires that emerged in the early stages of human history, the Chinese and the Egyptian, as they did in the imperial systems of Rome and Byzantium. Each of these was a political system that endured for centuries.

As history evolved and as nation-states emerged as the dominant form of political community in Western society, public bureaucracies were still able to demonstrate as formidable a capacity for holding on to power as their ancient forbears. Even when other political institutions, such as elected executives and legislatures, began to emerge as new centers of government authority, these public bureaucracies still managed to maintain a leading position in crafting and carrying out state programs and decisions.

This continuity in the exercise of power enabled public bureaucracies to be a source of stability in the life of a nation even when fundamental changes were taking place in the nature of the regime they served. Alexis De Tocqueville has given us a memorable description of bureaucratic organizations serving in this way amidst the turbulence of French politics at the end of the eighteenth century: 'For though in each successive revolution the administration was, so to speak, decapitated, its body survived intact and active. The same duties were performed by the same civil servants... These men administered the country or rendered justice in the name of the King, then in that of the Republic, thereafter in the Emperor's' (De Tocqueville 1955).

2. Bureaucracy and Change

Since De Tocqueville wrote, the role that public bureaucracies can play as a stabilizing force in societies beset by a time of troubles has sometimes been seen in a less favorable light. It is argued, for example, that

their durability as government institutions has enabled bureaucracies to function as an obstacle to change in the social order, long after it has become clear that such change is urgently needed. During the twentieth century, some bureaucracies, especially in the military sector, have played this kind of obstructive role in the emerging nations of Africa and Asia. In its most extreme form, resistance to change on the part of a public bureaucracy may even lead a government agency to sabotage programs placed under its control that it opposes.

2.1 *Bureaucracies as Change Agents*

In the modern world, however, it is also possible to find public bureaucracies that serve not as barriers to innovation but as agents of change in their own societies. When Social Democratic political parties came to power in Europe after World War II, it was widely anticipated that the reforms they intended would be undercut by the conservative bureaucracies in their own country that would be responsible for putting these programs into effect. This kind of sabotage did not in fact occur. Such bureaucracies carried out the changes proposed by the Labour Party when it took control over the British government after the parliamentary elections of 1945 with the same diligence they had shown when administering the programs of the Conservative governments under which they had previously served.

Moreover, one of the enduring legacies of that great period of social and economic reform in the USA in the 1930s, which came to be known as the New Deal, was the creation of new bureaucratic organizations such as the Rural Electrification Administration (REA) and the Tennessee Valley Authority (TVA) that were designed to serve neglected sectors of the American population. In the 1960s a similar reform movement gave birth to agencies at all levels of government in the USA that were authorized to protect women and minority groups from discrimination in employment and other spheres of everyday life. Of course, even when a bureaucracy is created to serve as an agent of change, it can still find itself under attack from its own supporters on the grounds that it is insufficiently zealous in pursuing this objective. As a result, the stereotype of bureaucracy as an obstacle to change tends to endure, even in settings where an organization is breaking new ground.

The effort to use public bureaucracies as instruments of social reform is the product of a general movement toward democracy that has taken place throughout the Western world. As this has occurred, victory in regularly scheduled elections has become the chief source of empowerment, and the political parties competing in such electoral contests have every reason to see public bureaucracies as vehicles for promoting policy changes or carrying on programs that will ultimately provide the political support necessary to

win or maintain control of the government. This convergence of interest between party and bureaucracy has energized the effort in modern democratic societies to use bureaucratic institutions as instruments for breaking with the past rather than preserving it.

3. *Bureaucratic Expertise*

Throughout history the fundamental source from which public bureaucracies derive their power is the possession of a body of knowledge or skill indispensable to the successful performance of the government in which they operate. In many cases, especially during the early stages in their historical development, these organizations held a monopoly of such expertise. In addition to the influence it gave them within the councils of government, it enabled them to win a certain measure of deference within the community at large. To be sure, this deference was usually coupled with hostility, since ordinary citizens commonly resent the power exercised by bureaucrats in democratic as well as authoritarian societies. Such antagonism toward bureaucracies is perhaps the chief way in which these disparate systems of government continue to resemble one another in the modern world.

Over the course of modern history societies have differed in their willingness to allow expertise to become the signature characteristic of their public bureaucracies. The UK, modeling itself after the example of ancient China, pioneered in this respect, and various segments of the British Empire followed this example even after they won independence. In the more egalitarian culture of the USA, proposals to establish such a Mandarin-style bureaucracy met with strong initial resistance in the nineteenth century as betrayal of the country's democratic traditions. A similar outcry can still be triggered today in US politics by proposals that the country create an institutional counterpart to France's *'Ecole Nationale d'Administration,'* a school for the training of the French administrative elite.

3.1 *Competition among Experts*

One development that has greatly diminished the advantage public bureaucracies derive from their command over bodies of expertise in specialized areas of government policy has been the gradual migration of such expertise to other institutions of government. In the USA, for example, Presidents now have a large number of experts of their own clustered in various units within the Executive Office of the President. While Prime Ministers and Cabinet members in the parliamentary systems that prevail elsewhere in democratic societies do not have similar access to such organized cadres of experts, they are often able to turn elsewhere to satisfy their need for expert assistance.

3.2 *The Rise of Think-tanks*

Another important source of expertise widely available to elected officials today lies within the private sector in the form of private research organizations known as 'think-tanks' that have emerged and begun to focus on current issues of government policy in the Western democracies. The research findings of these organizations sometimes attract almost as much, if not more, public attention than the reports emanating from the public bureaucracies whose work they monitor. In earlier days these think-tanks tended to provide intellectual support for programs and policies that led to an expansion in the size and power of public bureaucracies, especially in the areas of the social services and economic regulation. More recently, however, think-tanks with a conservative orientation have begun to emerge, and they have provided intellectual ammunition for groups critical of public bureaucracies and the programs and policies these organizations pursue.

4. *Bureaucratic Secrecy and Open Government*

The emergence and expansion of new forms of mass communication has tended to make public bureaucracies increasingly vulnerable to scrutiny and criticism from institutions and groups outside the government. Pressure from such media organizations has been chiefly responsible for the enactment in a number of modern societies of 'freedom of information' statutes which require public bureaucracies to provide information in their files to private citizens or groups upon request, except for certain specified categories of data. Max Weber long ago saw the secrecy in which the public bureaucracies of his day normally operated as a major source of the power they exercised. As a study edited by Itzhak Galnoor has shown, the scope of such secrecy has been greatly narrowed in both Europe and the USA today.

5. *Alternatives to Public Bureaucracies*

Critics of the role of public bureaucracies in government, or of the programs they administer, have long argued that they are in fact immortal institutions, which, once established, are impossible to terminate. Certainly it is true that such bureaucracies are tenacious organizations, sinking deep roots into the fabric of the societies they serve.

Still, recent developments in modern societies make it abundantly clear that public bureaucracies are far from immortal. During the past two decades a wide variety of regulatory agencies have been put to death in both the USA and Europe, casualties of a bipartisan consensus that they mainly serve not the public but the interests of the industries they regulate. Public bureaucracies have also been victims of the growing belief that private ownership is more efficient

than state operation. A striking example of this trend occurred in Canada in 1996, when Parliament transferred the vital task of carrying on the country's air traffic control system from the government's transport department to a nonprofit firm created by the legislature.

Along with these developments, there has been a continuing expansion in both Europe and the USA of organizations in the nonprofit sector that administer programs that were once thought to lie within the exclusive domain of government. These nongovernmental agencies, or 'semi-public' bureaucracies as they might be called, supplement rather than supplant the work of public bureaucracies in such areas as health, welfare, and education, and much of their revenue actually comes from the government itself. One such organization in the international sphere—Medecins Sans Frontieres—received the Nobel prize for Peace in 1999 for its work in providing medical care and moral support for population groups in distress around the world.

See also: Administration in Organizations; Administrative Law; Bureaucracy and Bureaucratization; Bureaucratization and Bureaucracy, History of; Civil Service; Public Policy Schools; Public Sector Organizations; Think Tanks

Bibliography

- De Tocqueville A 1955 *The Old Regime and the French Revolution*. 1st edn., Doubleday, Garden City, NY
 Galnoor I (ed.) 1977 *Government Secrecy in Democracies*. Harper and Row, New York
 Gidron B, Kramer R M, Salamon L M (eds.) 1992 *Government and the Third Sector*. Jossey-Bass, Inc., San Francisco
 Huddleston M W, Boyer W W 1996 *The Higher Civil Service in the United States: Quest for Reform*. University of Pittsburgh Press, Pittsburgh, PA
 Kaufman H 1976 *Are Government Organizations Immortal?* Brookings Institution, Washington, DC
 Rourke F E 1984 *Bureaucracy, Politics, and Public Policy*. Little, Brown, Boston

F. E. Rourke

Public Commemoration

Public commemoration is a matrix of activity, through which social groups express 'a collective shared knowledge ... of the past, on which a group's sense of unity and individuality is based' (Assmann 1995) The group that organizes the commemoration inherits earlier meanings attached to the event, as well as adding new meanings. Their activity is crucial to the presentation and preservation of commemorative forms. When such groups disperse or disappear,

commemoration loses its initial force, and may fade away entirely.

Public commemoration is therefore a process with a life history. It has an initial, creative phase; followed by a period of institutionalization and routinization. Such markings of the calendar can last for decades, or be abruptly halted. In most instances, the date and place of commemoration fade away with the passing of the social groups which initiated the practice.

Commemoration operates on many levels of aggregation and touches many facets of associative life. While commemorative forms were familiar in the ancient and medieval period, they have proliferated in more recent times. Consequently, the subject has attracted much discussion in social scientific literature. We therefore concentrate on public commemoration in the epoch of the nation state, primarily in the nineteenth and twentieth centuries.

In the modern period, most commemorative events were marked distinctively and separately from the religious calendar. There has been, though, some overlap. Armistice Day, 11 November, in countries remembering the end of the 1914–18 war, is close enough to the Catholic feast of All Saints on 2 November; in some countries with a large Catholic population, the two days occupy a semisacred space of public commemoration. The day marking the end of the Second World War in Europe, 8 May, is also the Saint's day of Joan of Arc. Those engaging in commemorative acts on that day may be addressing the secular celebration or the Catholic one; some celebrate the two together.

Commemoration is an act arising out of a conviction, shared by a broad community, that the moment recalled is both significant and informed by a moral message. Moments of national humiliation are rarely commemorated, although here too there are exceptions of a hortatory kind. 'Never again' is the hallmark of public commemoration on the Israeli Day of Remembrance for victims of the Nazi persecution of the Jews. Where moral doubts persist about a war or public policy, commemorative moments are hard to fix. That is why there is no date commemorating the end of the Algerian War in France, or the end of the Vietnam War in the USA. There was no moral consensus about the nature of the conflict; hence there was no moral consensus about what was being remembered in public, and when was the appropriate time to remember it (Prost 1999).

1. Commemoration and Political Power

Much of the scholarly debate about commemoration concerns the extent to which it is an instrument of the dominant political elements in a society. One school of opinion emphasizes the usefulness to political elites of public events establishing the legitimacy of their rule

(Nora 1984–92). Some such events are observed whoever is in power—witness Bastille Day in France or Independence Day in the USA. But other events are closely tied to the establishment of a new regime and the overthrow of an older one: 7 November was the date in the calendar marking the Bolshevik revolution and establishing the Communist regime in power in Russia. That date symbolized the new order and its challenge to its world-wide enemies. May Day similarly was a moment when labor movements, and Labor parties in power, publicly demonstrated their place in history.

This top-down approach proclaims the significance of commemoration as a grammar of national, imperial or political identity. Anzac Day, 25 April, is celebrated as the moment when the Australian nation was born. It commemorates the landing of Australian and New Zealand troops as part of the British-led expeditionary force sent to Turkey in 1915. The fact that the landing was a failure does not diminish the iconic character of the date to Australians. It is the day, they hold, when their nation came of age (Inglis 1992).

By no means are all commemorative activities associated with warfare. The birthdates of monarchs or deceased presidents are marked in similar ways. Queen Victoria's birthday, 24 May, was Empire Day in Britain; now (2000) it is celebrated as Commonwealth Day. The creation of such commemorative dates was part of a wider movement of what some scholars have termed 'the invention of tradition.' That is, at the end of the nineteenth century, new nation states and preeminent imperial powers deepened the repertoire of their ceremonial activity. Such flourishes of the majesty of power were then immediately sanctified by a spurious pedigree. To display ceremonies with a supposed link to ancient habits or forms located in a foggy and distant past created an effective cover for political innovation, instability or insecurity (MacKenzie 1986, Ranger and Hobsbawm 1986).

This functionalist interpretation of commemoration has been challenged. A second school of scholarship emphasizes the ways that public commemoration has the potential for dominated groups to contest their subordinate status. However much political leaders or their agents try to choreograph commemorative activity, there is much space for subversion or creative interpretation of the official commemorative script. Armistice Day, 11 November, was, for different groups, a moment for both the celebration and the denigration of military values. Pacifists announced their message of 'Never again' through their presence at commemorative ceremonies; military men used these moments to glorify the profession of arms, and to demonstrate the duty of citizens, if necessary, to give their lives for their country in a future war. The contradictions in these forms of expression on the same day were never resolved (Gregory 1994, Winter 1999).

This alternative interpretation of the political meaning of commemorative activity emphasizes the multi-vocal character of remembrance. From this point of view, there is always a chorus of voices in commemorations; some are louder than others, but they never sound alone. Decentering the history of commemoration ensures that we recognize the regional, local, and idiosyncratic character of such activities.

Very occasionally, these dissonant voices come together. Between 1919 and 1938 in Britain, there was a two-minute silence observed throughout the country. Telephonists pulled the plugs on all conversations. Traffic stopped. The normal flow of life was arrested. Then the Second World War intervened, and such disruption to war production was not in the national interest. Thereafter the two-minute silence was moved to the Sunday nearest 11 November. But in the two decades between the Wars, it was a moment of national reflection. Mass Observation, a pioneering social survey organization, asked hundreds of ordinary people in Britain what they thought about during the silence. The answer was that they thought not of the nation or of victory or of armies, but of the men who were not there. This silence was a meditation about absence. As such it moved away from political orchestration into the realm of family history. To be sure, families commemorated their own within a wider social and political framework. But the richest texture of remembrance was always within family life. This intersection of the public and the private, the macro-historical and the micro-historical, is what has given commemoration in the twenty-first century its power and its rich repertoire of forms.

2. *The Business of Remembering*

Commemoration is and always has been a business. It costs money; it requires specialists' services; it needs funding and, over time, re-funding. There are two kinds of expenditure we can trace in the history of commemoration. The first is capital expenditure; the second is recurrent expenditure.

The land for such sites must be purchased; and an appropriate symbolic form must be designed and then constructed to focus remembrance activities. The first step may require substantial sums of public money. Private land, especially in urban areas, comes at a premium. Then there are the costs of architects' fees, especially when a public competitive tender is offered inviting proposals from professionals. Finally, once the symbolic form is chosen, it must be constructed out of selected materials and finished according to the architect's or artist's designs.

When these projects are national in character, the process of production is in the public eye. National art schools and bodies of 'experts' have to have their say. Standards of 'taste' and 'decorum' are proclaimed. Professional interests and conflicts come into play.

Much of this professional infighting is confined to national commemorative projects, but the same complex stepwise procedure occurs on the local level too, this time without the same level of attendant publicity. Local authorities usually take charge of these projects, and local notables can deflect plans towards their own particular visions, whatever public opinion may think about the subject.

Most of the time, public funding covers only part of the costs of commemorative objects. Public subscriptions are critical, especially in Protestant countries where the concept of utilitarian memorials is dominant. In Catholic countries, the notion of a 'useful' memorial is a contradiction in terms; symbolic language and utilitarian language are deemed mutually exclusive. But the Protestant voluntary traditions has it otherwise. In Protestant countries, commemorative projects took many forms, from the sacred to the mundane: in Britain there are memorial wards in hospitals, memorial scholarships in schools and universities, alongside memorial cricket pitches and memorial water troughs for horses. In the USA and in Australia there are memorial highways. The rule of thumb is that private citizens pick up most of the tab for these memorial forms. The state provides subsidies and occasional matching grants, but the money comes out of the pockets of ordinary people. The same is true in Britain with respect to a very widely shared form of public commemoration: the purchase of paper poppies, the symbol of the Lost Generation of the First World War. These poppies are worn on the lapel, and the proceeds of the sale go to aid disabled veterans and their families.

Recurrent expenditure for commemorative sites is almost always paid for by taxpayers. War cemeteries require masons and gardeners. The Imperial (now Commonwealth) War Graves Commission looks after hundreds of such cemeteries all over the world. The cost of their maintenance is a public charge. Private charities, in particular Christian groups, maintain German war cemeteries. Once constructed, memorial statues, cemeteries, or highways also become public property, and require public support to prevent them from decomposing. They are preserved as sites of commemorative activity.

Much of this activity is directed towards inviting the public to remember in public. This means directed the public towards particular sites of remembrance. Some of them are nearby their homes. In Britain and France there are war memorials in every city, in every town, and in every village; it is there that Armistice Day ceremonies are held annually. Churches throughout Europe of all denominations have memorial plaques to those who died in war. Special prayers were added to the Jewish prayer book to commemorate the victims of the Nazis in the Second World War, and later, those who died on active service in the Israeli army.

Remembrance in local houses of worship or at war memorials required that the public travel a short

distance from their homes to sites of remembrance. But given the scale of losses in the two World Wars, and the widely dispersed cemeteries around the world in which lie the remains of millions of such men and women, the business of remembrance also entails international travel. Such voyages start as pilgrimage; many are mixed with tourism (Lloyd 1998). But in either case, there are train and boat journeys to take; hotel rooms to reserve; guides to hire; flowers to lay at graves; trinkets and mementos to purchase. In some places, museums have arisen to tell more of the story the pilgrims have come to hear and to share. There too money is exchanged along with the narratives and the symbols of remembrance.

This mixture of the sacred and the profane is hardly an innovation. It is merely a secular form of the kind of pilgrimage, for example, that made San Juan de Compostela in Spain the destination of millions of men and women in the middle ages who came to honor the conventionally designated resting place of the remains of one of the original Apostles. Pilgrimage to war cemeteries is public commemoration over long, sometimes very long, distances. Today (2000) thousands of Australians make the journey to Gallipoli, on the border between Asia and Europe, to be there for the dawn service on 25 April (as we have already noted), the day of the Anzac landing in 1915. Where does pilgrimage stop and tourism take over? It is impossible to say, but in all cases, the business of remembrance remains just that—a business.

3. Aesthetic Redemption

Public commemoration is not only a set of political gestures and material tasks. It is also an art form, the art of arranging and interpreting signifying practices. This field of action can be analyzed on two different levels: the first is aesthetic; the second is semiotic. The two are intimately related.

Some national commemorative forms are distinctive. Others are shared by populations in many countries. The figure of Marianne as the national symbol of the French Republic could not be used in Germany or Britain. The German Iron Cross, on commemorative plaques, denotes the location and the tradition in which commemoration is expressed. Germany's heroes' forests or fortresses are also imbricated in Teutonic history.

At times the repertoire of one country's symbols overlap with that of others, even when they were adversaries. After the First World War, the first industrialized war fought among fully industrialized nations, many commemorative forms adopted medieval notation. Throughout Europe, the revolutionary character of warfare was marked by a notation of a backward-looking kind. Medieval images of heroic and saintly warriors recaptured a time when combat was between individuals, rather than the impersonal

and unbalanced duel between artillery and human flesh. The war in the air took on the form and romance of chivalry. On the losing and the winning sides, medievalism flourished.

On war memorials, the human form survived. In some instances, classical images of male beauty were chosen to mark the 'lost generation'; others adopted more stoical and emphatically untriumphalist poses of men in uniform. In most cases, victory was either partially or totally eclipsed by a sense of overwhelming loss. Within this aesthetic landscape, traditional Christian motifs were commonplace. The form of the grieving mother—*Stabat Mater*—brought women into the local and national constellation of grief.

In Protestant countries, the aesthetic debate took on a quasireligious character. War memorials with crosses on them offended some Protestants, who believed that the Reformation of the sixteenth century precluded such 'Catholic' notation. Obelisks were preferable. In France, war memorials were by law restricted to public and not church grounds, although many local groups found a way around this prescription. In schools and universities, the location of such memorials touched on such issues. Some were placed in sacred space—in chapels—semisacred space—around chapels—or in secular space. Public thoroughfares and train stations also housed such lists of men who had died in war.

Twentieth-century warfare democratized bereavement. Previously armies were composed of mercenaries, volunteers, and professionals. After 1914, Everyman went to war. The social incidence of war losses was thereby transformed. In Britain, France and Germany, virtually every household had lost someone—a father, a son, a brother, a cousin, a friend. Given the nature of static warfare on the Western front, many—perhaps half—of those killed had no known grave. Consequently, commemorative forms highlighted names above all. The names of the dead were all that remained of them, and chiseled in stone or etched on plaques, these names were the foci of public commemoration on both the local and the national scale.

This essential practice of naming set the pattern for commemorative forms after the Second World War and beyond. After 1945, names were simply added to Great War memorials. This was partly in recognition of the links between the two twentieth-century conflicts; partly it was a matter of economy. After the Vietnam war, naming still mattered, and First World War forms inspired memorials, most notably Maya Lin's Vietnam Veterans' Memorial in Washington. Her work clearly drew on Sir Edwin Lutyens's memorial to the missing of the Battle of the River Somme at Thiepval, inaugurated in 1932.

By the latter decades of the twentieth century, artistic opinion and aesthetic tastes had changed sufficiently to make abstraction the key language of commemorative expression. Statues and installations

thereby escaped from specific national notation and moved away from the earlier emphasis upon the human figure. The exception to the rule is Soviet commemorative art, which resolutely stuck to the path of heroic romanticism in marking out the meaning of what they called the Great Patriotic War. In many instances in Western Europe, but by no means all, forms which suggested absence or nothingness replaced classical, religious or romantic notions in commemorative art.

This aesthetic shift had important social and political implications. Great War commemorative forms had sought out some meaning, some significance in the enormous loss of life attending that conflict. There was an implicit warning in many of these monuments. 'Never again' was there ultimate meaning. But 'never' had lasted a bare 20 years. Thus after the Second World War, the search for meaning became infinitely more complex. And the fact that more civilians died than soldiers in the Second World War made matters even more difficult to configure in art.

Finally, the extreme character of two elements of the Second World War challenged the capacity of art—any art—to express a sense of loss when it is linked to genocidal murder or thermonuclear destruction. Both Auschwitz and Hiroshima defied conventional notations of 'meaning,' although some individuals continue to try to rescue redemptive elements from them both. The construction of a national Holocaust memorial at the Brandenburg gate in Berlin, the construction of which began in 1999, is a case in point (Young 1999). The debate over this and similar projects does not abate, since such events seem to force all participants—public officials, the public, survivors, architects, artists—to the limits of representation and beyond.

4. *Ritual*

Public commemoration is an activity defined by the gestures and words of those who come together to remember the past. It is rarely the simple reflection of a fixed text, a script rigidly prepared by political leaders determined to fortify their position of power. Inevitably, commemoration overlaps with political conflicts, but it can never be reduced to a simple reflection of power relationships.

There are at least three stages in the history of rituals surrounding public commemoration. The first we have already dealt with: the construction of a commemorative form. But there are two other levels in the life history of monuments which need attention. The second is the grounding of ritual action in the calendar, and the routinization of such activities; the third is their transformation or their disappearance as active sites of memory.

One case in point may illustrate this trajectory. The date of 1 July 1916 is not a national holiday in Britain;

but it marks the date of the opening of the British offensive on the river Somme, an offensive which symbolized the terrible character of industrial warfare. On that day the British army suffered the highest casualty totals in its history; on that day a volunteer army, and the society that had created it, were introduced to the full terrors of twentieth-century warfare. To this day, groups of people come to the Somme battlefields to mark this day, without national legislation to enable them to do so. There are locally-defined rituals. In France, 11 November is a national holiday, but not in Britain. Legislation codifies activities the origins and force of which lie on the local level.

Public commemoration flourishes within the orbit of civil society. This is not true in countries where dictatorships rule; Stalinist Russia smashed civil society to a point that it could not sustain commemorative activity independent of the party and the state (Merridale 1999). But elsewhere, local associations matter.

And so do families. Commemorative ritual survives when it is inscribed within the rhythms of community and, in particular, family life. Public commemoration lasts when it draws about overlaps between national history and family history. Most of those who take the time to engage in the rituals of remembrance bring with them memories of family members touched by these vast events. This is what enables people born long after wars and revolutions to commemorate them as essential parts of their own lives. For example, children born in the aftermath of the First World War told the story of their family upbringing to grandchildren born 60 or 70 years later. This transmission of childhood memories over two or sometimes three generations gives family stories a power which is translated at times into activity—the activity of remembrance (Winter and Sivan 1999).

This framework of family transmission of narratives about the past is an essential part of public commemoration. It also helps us understand why some commemorative forms are changed or simply fade away. When the link between family life and public commemoration is broken, a powerful prop of remembrance is removed. Then, in a short time, remembrance atrophies and fades away. Public reinforcements may help keep alive the ritual and practice of commemoration. But the event becomes hollow when removed from the myriad small-scale social units that breathed life into it in the first place.

At that moment, commemorative sites and practices can be revived and reappropriated. The same sites used for one purpose can be used for another. But most of the time, commemorative forms live through their life cycle, and like the rest of us, inevitably fade away. This natural process of dissolution closes the circle on public commemoration. And rightly so, since it arises out of the needs of groups of people to link their lives with salient events in the past. When that

need vanishes, so does the glue that holds together the social practice of commemoration. Then collective memories fade away.

We have reached, therefore, a quixotic conclusion. Public commemoration is both irresistible and unsustainable. Time and again people have come together in public to seek meaning in vast events in the past and try to relate them to their own smaller networks of social life. These associations are bound to dissolve, to be replaced by other forms, with other needs, and other histories. At that point, the characteristic trajectory of public commemoration—creation, institutionalization, and decomposition—comes to an end.

See also: Collective Identity and Expressive Forms; Collective Memory, Anthropology of; Collective Memory, Psychology of; Media Events; Pilgrimage; Public Sphere and the Media; Ritual

Bibliography

- Assmann J 1995 Collective memory and cultural identity. *New German Critique* 65: 125–34 (esp. 132)
- Gregory A 1994 *The Silence of Memory*. Berg, Leamington Spa, UK
- Inglis K 1992 World War One memorials in Australia. *Guerres Mondiales et Conflits Contemporains* 167: 51–8
- Lloyd D 1998 *Battlefield Tourism: Pilgrimage and the Commemoration of the Great War in Britain, Australia and Canada, 1919–1939*. Berg, Oxford, UK
- MacKenzie J (ed.) 1986 *Imperialism and Popular Culture*. Manchester University Press, Manchester, UK
- Merridale C 1999 War, death and remembrance in Soviet Russia. In: Winter J, Sivan E (eds.) *War and Remembrance in the Twentieth Century*. Cambridge University Press, Cambridge, UK, pp. 61–83
- Nora P (ed.) 1984–92 *Les Lieux de Mémoire*. Gallimard, Paris
- Prost A 1999 The Algerian War in French collective memory. In: Winter J, Sivan E (eds.) *War and Remembrance in the Twentieth Century*. Cambridge University Press, Cambridge, UK pp. 161–76
- Ranger T, Hobsbawm E (eds.) 1986 *The Invention of Tradition*. Cambridge University Press, Cambridge, UK
- Winter J 1999 *Sites of Memory, Sites of Mourning: the Great War in European Cultural History*. Cambridge University Press, Cambridge, UK
- Winter J, Sivan E 1999 Setting the framework. In: Winter J, Sivan E (eds.) *War and Remembrance in the Twentieth Century*. Cambridge University Press, Cambridge, UK, pp. 1–40
- Young J E 1999 Berlin's Holocaust Memorial: a report to the Bundestag Committee on Media and Culture. *German Politics and Society* 17(3): 54–70

J. Winter

Public Good, The: Cultural Concerns

From the inception of political theory to twenty-first century reflections on democracy and the welfare state, the ideal of public good has consistently been

linked to the sociomoral resource of civic spirit. This is so insofar as the ideal of the public good claims civic spirit from a society's citizens, while the presence of civic spirit is an indispensable premise for any orientation toward the public good.

Through such a circular linkage, the ideal has gained strategic importance as an element of political rhetoric catalyzing a process of normative self-limitation, due to conditions at play in the democratic public sphere.

1. The Public Good in the History of Political Ideas

From time immemorial, the idea of the public good has been a fixed quantity in debates concerning legal, economic, and political forms of social cooperation. For example, in ancient Egypt, Pharaoh's duty to see to the welfare of his subjects was considered an element of *Ma'at*, the world-order connecting human beings to the gods; and in Islamic legal doctrine, the normative standard for the legal-religious community, *maslaha*, can be interpreted as similar to the idea of the public good. The concept's occidental career began in ancient Greece. Plato distinguished the public good from private profit, linking the former to the idea of justice (Plato 1993, 420 b–c, 421 a–c). Aristotle followed Plato in this respect (Aristotle 1990, III./VII./1); but in contrast to Plato he did not equate the just state with an authoritarian pedagogic utopia. Rather, he understood the public good to involve a political limitation on the right to rule, demarcating a difference between the proper constitutional form, which acknowledges such a good, and the improper, serving the ruler's interests (Aristotle 1990, III./IV./7). The political thinking of Roman republicanism focused on reconciling individual interests with the public good, the legal system here taking on special meaning. But even at this early period, lamenting at a loss of orientation around the public good—this seen as the source of a growing danger to the republic—is a predominant theme.

The redemptive orientation of the early Christian fathers transcended the antique approach to the public good, depoliticizing it in the process. But within scholasticism, the dogma of the subordination of earthly existence to the divine order again received a ruler-limiting and hence political accent. In his *Summa Theologiae*, Thomas Aquinas defined a regime as unjust tyranny when the ruler does not strive for the public good but for his private benefit (1952, Pars IIa, IIae, q.42, a.2). Subjects can here thus evoke the public good, placed higher than the ruler's interests, in opposition to the ruler; and the ruler can appeal to the public good to justify his actions. In this manner, a struggle emerged regarding semantic supremacy over the concept of public good—the concept being capable of both limiting and legitimizing political rule. If the ruler could claim to be enacting the public good,

nothing stood in the way of his rule; to be sure, the latter was then limited through involvement of the estates—such involvement being inextricably bound up with the idea of public good in the medieval world of political ideas (Eberhard 1993).

Whereas into the early modern period, self-interest and the public good represented 'asymmetrical counterconcepts' in Reinhart Koselleck's sense, i.e., 'binary concepts with claims to universality ... involving a mutual recognition of the parties involved' (Koselleck 1985), the Reformation saw the onset of a relativization process. Set in motion by the 1555 Peace of Augsburg, the increased individualization in estate-bound society was accompanied by a normative transformation, its import captured in Leonhard Fronsberger's initial (at the time singular) praise of self-interest. This already reflected an important paradigm shift that would unfold over three decades, replacing the old European ideal of political virtue, in the sense of the citizen's voluntary orientation toward the public good (Münkler 1991). Instead, rational self-interest now emerged as a core political-theoretical concept. Its immediate context was a market society in the course of self-differentiation, with the state now simply expected to set the conditions for maximized economic well-being of the citizenry in a *civil* society. By turning the inherited conceptual opposition of public good and self-interest upside down, Bernard Mandeville's *The Fable of the Bees* (1929; its first, pamphlet form appearing in 1795 as *The Grumbling Hive or Knaves turn'd Honest*) played a key part in this transformative process. 'Private vices, public benefits' was Mandeville's credo. The end effect of each individual pursuing his private good is the greatest possible advantage for the polity, so that one, particular maxim prevails: 'The worst of all the Multitude Did something for the Common Good' (Mandeville 1929). Admittedly, this construct is based on citizens being in a position to pursue their own interests rationally; and Mandeville remained skeptical regarding his own theory's implicit premise. In fact, he demanded state intervention to check the loss of civic spirit—particularly feared on account of short-sightedness and ignorance when it came to the needs of following generations.

In respect to this general problem, Adam Smith was essentially more optimistic. Bringing Mandeville's idea to a head, he saw a mechanism at work that produces a common benefit in direct proportion to efforts at its opposite. The citizen, he argued, 'generally, indeed, neither intends to promote the public interest, ... he intends only his own gain, and he is in this, as in many other cases, led by an invisible hand to promote an end which was no part of his intention' (Smith 1981). With this argument, Adam Smith in a way achieved liberalism's semantic coup vis-à-vis the public good. Through the definition of the public good as a wondrous result of well-understood self-interest—a predictable product of the individual's egoistic maxi-

mization of advantage—the asymmetric conceptual opposition between self-interest and public good, already relativized by Mandeville, was enduringly destroyed (Bohlender 1998). Reflecting the advent of modern contractual theory, this way of thinking can lead to the sort of relativizing of the idea of public good that we find in Kant: a complete focus on formal legal principles, and the treatment of the sociomoral qualities and intentions of both rulers and the ruled as a secondary matter. Kant, however, was not the final word in the long conceptual history of the public good, for speaking up against him, Hegel had already insisted that 'the so-called "general good," the welfare of the state, i.e., the right of mind actual and concrete, is quite a different sphere, a sphere in which abstract right is a subordinate moment like particular welfare and the happiness of the individual' (Hegel 1942, Sect. 126 Remark).

In the twentieth century, the notion of public good was subject to steady abuse: first by the Nazis, who used the empty phrase 'the common interest precedes self-interest' as propaganda for its collectivist 'you are nothing; your *Volk* is everything' ideology, then by the Bolsheviks. As a consequence, a refusal substantively to define the public good and how it is to be realized, once and for all, has emerged as an identifying trait of proponents of freedom and democracy. Rather, through use of democratic procedures, society needs constantly to arrive at a consensus regarding the concept—a consensus subject to revision. This does not preclude the pluralistic process of opinion formation being oriented toward normative concepts of public good with universal, i.e., ubiquitous and extra-temporal claims to validity: the natural-law claims, for instance, staked by Catholic advocates of social transformation. In the context of, among other things, a global ecological crisis, such universal claims have been further developed into a concept of a worldwide public good.

2. Public Good and Civic Spirit

This overview of the conceptual history of 'public good' points to two basic features. First, it represents a functional formal concept with effects that vary considerably according to the context of usage and semantic competence. It can, for instance, serve for both the legitimation and limitation of rule. The renunciation by democratic societies of an *a priori*, essentialist understanding of the public good can thus be seen as a consequence of the insight into its semantic indeterminacy having become manifest in the course of time.

Second, precisely when thematized from a democratic perspective, the concept of public good needs to be viewed in connection with another central theme of political theory, namely, civic virtue or civic spirit, which Mandeville had already addressed in his skep-

tical fable. Against the contractualistic tendency to place an economic model of order in the foreground, Montesquieu, for example, had stressed that, more than is the case with any other form of state, democracies must see to it that their citizens identify with the public good. For a republic, he indicates, needs 'a continuous preference of the public interest over one's own' (Montesquieu 1989, IV, 5)—and this without the state force that Mandeville speaks of. The emphasis on this voluntariness underscores the importance of a minimum of civic spirit (or civic virtue) as an indispensable premise for any orientation toward the public good. The ideal of public good sets the norms for the behavior of a commonwealth's free citizens; it needs, however, the preliminary presence of some degree of civic virtue as a willingness to act according to normative ideals in the first place. As such a preliminary, motivational specification for any orientation toward the public good, civic spirit emerges as a prepolitical basis for the political, only capable of being inadequately guaranteed by democratic politics, which may not control (or even guide) the citizenry's motivations (Böckenförde 1991, Goodin 1992).

In this sense, civic spirit is a *nonrenewable* sociomoral resource (as the metaphor of social capital suggests, by alluding to the economic understanding of 'capital') and hence it ought to be neither undertaxed nor overtaxed. Undertaxing occurs when, for instance, the solidarity every society needs increasingly becomes a form of institutional machinery (Gunnar Myrdal), creating an impression among individual citizens that the state is itself removing all required acts of solidarity from their own purview. Overtaxing is apparent when the political-social unit subject to the norms set by the public good exceeds a certain size, the imagination of belonging and reciprocal interdependence required for public good thus no longer being capable of production. This is all the more likely when such extension is a successive process making use of moral appeals and excessive public-good rhetoric with visible semantic development. Old European republicanism already saw the erosion of a society's sociomoral resources being threatened here; and communitarianism confirms this in referring to the need for supportive habits, the experience of personal identification, and one's own acts of solidarity (Bellah et al. 1985, 1991, Walzer 1983). But even Adam Smith's trust in the invisible hand was tied to the premise of a limited, overseeable marketplace: one guaranteeing that consumers could themselves sanction an egoistic benefit-maximization spurred forward at their own cost.

A game-theoretical analysis using postulates of rational choice backs up this assumption of a link between, on the one hand, a group subject to public-good norms having a comprehensible size, and, on the other hand, civic virtue (or civic spirit) as the readiness to follow these norms. Take, for example, a group of

people who visit an expensive restaurant, having agreed together to share the costs equally regardless of the different prices of individual orders: the factors determining the choice between self-interest and cooperation are the group's size, communicative capacity, and duration. In this manner, we can perceive the expectation of cooperation as a decisive element in the calculation of individual benefit—thus rendering civic spirit an explicable quantity of political-social action (Glance and Huberman 1994).

But not only democracy-theoretical factors connect the two basic features of the political idea of the public good—the fact that it represents a functional formal concept, and its immanent ties to civic spirit. The ongoing debate over the future of the Western welfare-state model underscores an interesting variation on the historically familiar semantics of *public good* and *civic spirit*. It is evident, for instance, in an argument regarding distribution: policies misperceived as a guarantee of welfare by the state produce a situation in which egoistic benefit-maximization can no longer further the public good, but—caught up in an illiberal system of administrative care—simply results in an exploitation of social output. According to this argument, a welfare state oriented towards a false ideal of public good effects something like an amputation of the invisible hand conjured up by Adam Smith; it does so by gradually diminishing the productive readiness of its members—hence eroding its own sociomoral basis, the free and responsible engagement of its citizens. Without the sociomoral resource of civic spirit, the development of active and productive self-interest—the self-interest that promotes the public good—is paternalistically thwarted. What remains is passive and unproductive self-interest: the sort damaging the public good.

3. The Public Good as a Political Concept

The meaning of the public good as a political concept—one through which interest-conflicts are consistently played out as interpretive struggles—emerges not least of all in this rhetorical mode of a differing public-good semantics. With the satisfaction of one's own demands being represented as a general advantage, precisely this theme has the merit of masking a variety of self-interests. As has been shown in an empirical study of the orientation of different German industrial organizations toward the public good, participating in such interpretive struggles has now basically become compulsory. For those failing to offer a 'declaration of public good' risk finding their positions cast under the suspicion of group-egoism and irrationality. Lack of responsibility and self-inflicted damage then emerge as the discrediting political-semantic labels (Vobruba 1992, Luhmann 1990, 2000). Once the connection between public good and civic spirit is established, political rhetoric can

even have its effect as self-fulfilling prophecy. This is the case when an impression of general decline in civic spirit is awakened among the citizenry, in turn prompting a politically-induced disappointment, defined by Albert O. Hirschman as the basis for withdrawal into the self-interested, private sphere: a disappointment cyclically alternating with periods of engagement for the public good (Hirschman 1982).

We can thus discern one particular continuity between the medieval and contemporary approach to the public good: the effort at a monopoly over its definition as constituting a central element of political conflict. Now as before, the theme can be invoked both as a basis for political demands and proposals for reform and, as it were, a 'governmental concept' meant to defend individual claims of a various interest groups. The latter function involves the government's task being depicted as to protect the public good from damaging special interests—even if such a protective function in a pluralistic democracy itself requires validation. In this respect, the breadth of the concept's variations—taking the form, as well, of common good, common weal, and public interest, each with its own fine shade of meaning—offers a rich range of possibilities for political communication.

Nevertheless, accompanying these different rhetorical functions, the idea of the public good consistently retains what—broadly speaking—constitutes a normative dimension: this as a result of the self-limitation necessarily imposed on anyone claiming to orient his political behavior toward the public good. Such a limitation furnishes a check on any purely strategic exploitation of the rhetoric of public good. Those invoking the concept in public debate reveal great expectations regarding the total effect on the welfare state of the policies they are enacting or defending. (This remains the case even when the invocation serves to rebuff claims and demands tendered by various social groupings.) At the same time, they are publicly aligning themselves with a standard against which they and their policies are to be measured due to conditions at play in the democratic public sphere. In this sense, invoking the public good involves a self-limitation by society's political agents.

To this extent, its status as a functional formal concept contradicts neither the public good's normative dimension, resulting from the obligation of honesty in democratic politics, nor its political relevance. This leads once more to both the extreme attitudes toward the public good mentioned above. On the one hand, the radically individualistic view sees free-market mechanisms without bureaucratic restriction or attempts at political steering as the best way to promote the public good. Beside the fact that such a radical view must ignore the linkage between the public good and the sociomoral resource of civic spirit we have been discussing, the use of public-good rhetoric not only by politicians, but also by representatives of society's other functional systems, indicates

that the ideal of the public good is an inevitable category for any public action, particularly within the political system. At the same time, the conditions of pluralistic democracy exclude any substantialistic and durably valid definition of the public good *a priori*, as characterizing totalitarian ideologies. In a pluralistic democracy, the public good is necessarily an 'open' formula, comprising a valid norm only if it can be freely accepted by all of the individuals who are affected by it. The application of this Habermasian idea to the ideal of the public good points to the limits of its validity, as well as its importance as a core concept of political thought and practice.

See also: Altruism and Prosocial Behavior, Sociology of; Citizenship, Historical Development of; Citizenship: Sociological Aspects; Civic Culture; Civil Society/Public Sphere, History of the Concept; Individualism versus Collectivism: Philosophical Aspects; Public Goods: International; Virtue Ethics

Bibliography

- Aristotle 1990 Politics. In: *Aristotle in Twenty-three Volumes*. Harvard University Press, Cambridge, MA, Vol. 21
- Bellah R N, Madsen R, Sullivan W M, Swindler A, Tipton S M 1985 *Habits of the Heart. Individualism and Commitment in American Life*. University of California Press, Berkeley, CA
- Bellah R N, Madsen R, Sullivan W M, Swindler A, Tipton S M 1991 *The Good Society*. Knopf, New York
- Böckenförde E W 1991 *State, Society and Liberty. Studies in Political Theory and Constitutional Law*. Berg, New York
- Bohlender M 1998 Government, Commerce und Civil Society. Zur Genealogie der schottischen politischen Ökonomie. In: Kaelble H, Schriewer J (eds.) *Gesellschaften im Vergleich*. Peter Lang, Frankfurt, Germany, pp. 115–47
- Eberhard W 1993 Herrscher und Stände. In: Politisches Denken in der Zeit der Reformation. In: Fetscher I, Münkler H (eds.) *Pipers Handbuch der politischen Ideen. Vol. 2: Mittelalter: Von den Anfängen des Islams bis zur Reformation*. Piper, München, Germany, pp. 467–551
- Glance N S, Huberman B A 1994 The dynamics of social dilemmas. *Scientific American* **270**: 76–82
- Goodin R E 1992 *Motivating Political Morality*. Blackwell, Cambridge, MA
- Hegel G W F 1942 *Hegel's Philosophy of Right*. Clarendon, Oxford, UK
- Hirschman A O 1982 *Shifting Involvements. Private Interest and Public Action*. Princeton University Press, Princeton, NJ
- Koselleck R 1985 *Futures Past. On the Semantics of Historical Time*. MIT Press, Cambridge, MA
- Luhmann N 1990 *Political Theory in the Welfare State*. W de Gruyter, Berlin
- Luhmann N 2000 *Die Politik der Gesellschaft*. Suhrkamp, Frankfurt, Germany
- Mandeville B 1929 *The Fable of the Bees: Or, Private Vices, Public Benefits*. Clarendon, Oxford, UK
- Montesquieu C S 1989 *The Spirit of Laws*. Cambridge University Press, Cambridge, UK
- Münkler H 1991 Die Idee der Tugend. Ein politischer Leitbegriff im vorrevolutionären Europa. *Archiv für Kulturgeschichte* **73**(2): 379–403

- Plato 1993 *Republic*. Oxford University Press, Oxford, UK
Smith A 1981 *An Inquiry into the Nature and Causes of the Wealth of Nations*. Liberty Classics, Indianapolis, IN
Thomas Aquinas 1952 *S. Thomae Aquinatis Summa Theologiae*. Marietti, Torino, Italy
Vobruba G 1992 Wirtschaftsverbände und Gemeinwohl. In: Mayntz R (ed.) *Verbände zwischen Mitgliederinteressen und Gemeinwohl*. Verlag Bertelsmann Stiftung, Gütersloh, Germany, pp. 80–121
Walzer M 1983 *Spheres of Justice. A Defense of Pluralism and Equality*. Basic Books, New York

H. Münkler and K. Fischer

Public Goods: International

International public goods are goods that are accessible to all at international level, the consumption of which by one individual does not reduce the possibility of consumption by other individuals. However, there are few public goods of this pure type. Instead, nearly all are characterized by mostly unlimited accessibility and simultaneous rival consumption, i.e., the use of the resource by one individual reduces the possibilities for others to consume this good. These types of goods are called common-pool resources. Where there is free accessibility and rival consumption, a collective action problem or social dilemma arises: how can the individual user be prevented from using the limited resource and jeopardizing the long-term utility of all as a result of maximizing their own short-term individual interest? A number of institutional possibilities exist to secure the provision of public goods and common-pool resources. However, the possibilities for collective action available at national level—such as legal prescriptions imposed by majority decision in order to force individual actors to comply with a scheme of collective action—are not all applicable at international level because of nation–state sovereignty. In what follows, public goods will first be defined, then different paths to the development of institutional answers to collective action problems will be presented, before different institutional modes of provision of public goods and common-pool resources and their possible implications will be discussed.

1. Definition of Public Goods

Public goods and common-pool resources—as stated above—may be defined by the analytic economic criteria of universal accessibility and (non-)rival consumption. These attributes may be properties that are inherent in the goods themselves, or they may be properties that are attributed to goods through societal

consensus and/or political decision. In the older economic literature the central attributes of public goods—nonexcludability and nonrivalry (Samuelson 1954, Musgrave and Musgrave 1976)—have been regarded as sources of market failure, from which the need for state intervention has been derived. The problem with common-pool resources and public goods lies in their limited excludability. As a result, prices cannot be charged and hence private provision through the market is not feasible. More recently, economic theory (Cornes and Sandler 1996) has argued in a more differentiated way that an exception is the case where a particular actor has an incentive to provide a public good irrespective of the free-riding behavior of other beneficiaries. The same holds true when, in the case of a common-pool resource, the stakes of provision are high enough to induce co-operation among actors (Cornes and Sandler 1996). Only in cases where the private provision of common goods fails are governments called upon to remedy the market failure (Bator 1958). Another aspect from which public goods can be viewed is the phenomenon that some market activities generate external effects. Positive, as well as negative, external effects are not part of the actor's economic calculation and as a result they are under- or over-produced from an aggregate view. In the traditional perspective externalities are seen as by-products of economic activities and internalization of externalities traditionally has been viewed as a government task (Pigou 1920). However, in the 1960s a new interpretation started to dominate the economic literature. According to this view, Pareto-relevant externalities are internalized through market forces if property rights are well defined and transaction costs are sufficiently low (Coase 1960, Buchanan and Stubblebine 1962).

While economic theory views the properties of a public good as inherent attributes of such a good from which a specific institutional form of provision is deduced, political science goes beyond this notion to insist additionally on a processual view of common goods. It argues that what constitutes such a good and how it is provided for institutionally depends on specific situational and technological conditions, and, moreover, is determined in a political and social process. First, it is not clear in many cases what constitutes a public good or a common-pool resource in so far as there are no absolute and objective criteria for defining them in a general sense. Instead, what constitutes a public good depends in part on the specific situational context of provision and consumption of the good, and its technical properties.

Thus, the importance of a specific situational context for the definition of a good as a public good in terms of nonexcludability and nonrival consumption is illustrated by the example of the protection offered by the ozone layer vis-à-vis radiation for sunbathers, which is a public good in the classical sense. One sunbather enjoying the sun does not impinge upon the

protection offered to other sunbathers. In this sense, the ozone layer is a public good, accessible to all with nonrival consumption. In another sense, however, it may be perceived as a common-pool resource. Seen in terms of the relationship between producers and sunbathers, and given that scientific research has identified a problem of scarcity in the ozone layer, certain productive activities, such as the production of refrigerators with chlorofluorocarbons, diminish the protective quality of the ozone layer and jeopardize the protection offered to all. This is rival consumption. In this specific context, the ozone layer constitutes a common-pool resource which is accessible to all and is depletable (Snidal 1986).

Technological innovation may force a reconsideration of the definition of a public good in cases where a good, which had hitherto been regarded as universally accessible and nondiminishable, has been rendered potentially excludable. A case in point is broadcasting. This good used to be accessible to all and of nonexcludable nature, but as a result of technological change, access to the good can now be limited to those paying for the service via a telematic coding system. Put in different terms: the transaction costs for establishing property rights were reduced drastically by technological progress so that the gains from excluding free riders and the precise allocation of user costs now exceed the costs of establishing property rights. Second, the definition of what constitutes a public good and what does not, may not necessarily be determined by 'objective' characteristics, but instead by political and social definition. There are huge cultural variations across polities and time as to what is considered a public good or common-pool resource and what is a private good. Thus, universal and free healthcare may be formulated as a political goal and consequently be considered as a common-pool resource in one polity. But in another, the same service may be regarded as a private good that each individual has to buy on the market. In the same way, policy makers in one society may decide that its provision should be left entirely to the buying power of the individual through the market. Hence, what constitutes a public good or common-pool resource can be answered either in terms of analytic economic criteria or in terms of a process of social and political definition. Once a good has been identified as a public good or a common-pool resource, the institutional mode of provision of the good has to be determined.

2. The Provision of Institutional Answers to a Collective Action Problem

Different paths may lead to an institutional answer for a stable solution for providing public goods at international level. The solutions may be designed intentionally, as is the case of international agreements. They may develop without being planned, as in

a case of spontaneous coordination, for example, in the case of the standardization of products. Either way, once in place, they develop an institutional dynamism of their own and over a period of time they may transform the processes in the institution and, in turn, transform the preferences of the actors involved.

The intentional formulation of an institutional solution to a public-good problem may presuppose an acute awareness of an existing social dilemma and a need for collective action where there is cross-boundary problem interdependence. For example, in the case of an imminent cross-border environmental problem, an international conference will be convened that seeks to develop a format of cooperation. After a lengthy process of bargaining, institutional answers for solving the problem will be found and agreed. All those involved in the negotiations will seek to maximize their own interests, for example, by minimizing the implicated costs. In other words, institutional solutions to collective action problems are interpreted in terms of their distributional impacts on the actors involved (Knight 1995). In the great majority of cases, the negotiators do not all enjoy the same degree of power, and in the event of an agreement not being reached, the stronger actors may take up a fall-back position. While a solution may be desirable for them, they may not consider it vital. In other words they have more resources for negotiating an institutional solution. This route to an institutional solution, i.e., interest-oriented negotiation with a view to the distributional results, generally applies where the interests at stake are clear-cut and substantial. A case in point is the negotiation of CO₂ emission limits at international climate conferences. All the states engaged in the bargaining process, to find an institutional solution for reducing emissions, view the process in terms of its distributional impacts on the market position of domestic industries, and so forth. However, there are situations where collective action problems need to be solved and the actors involved do not really appreciate the nature of the problem at hand, much less which interests are at stake. Consequently, the preferences of the actors involved are uncertain. In such situations deliberation, exchange of arguments, and persuasion will prevail as the actors strive to reach, at best, a common perspective on the problem at hand together with possible solutions; or as a minimum, to establish clarity as to the different stakes involved (Knight and Johnson 1994). Only then does interest-oriented bargaining follow and asymmetries of interests and power come into play. Cases in point are complex environmental problems or the problem of providing more transparency and certainty in financial markets.

The second avenue for developing an institutional solution to a collective action problem is that of the spontaneous emergence, rather than the intentional development of a solution in a planned social interaction at international level. Here, the strategic

uncertainty as to the preferences of other actors promotes behavior likely to resolve a social dilemma. It makes sense to shape other actors' perceptions of one's preferences and to give signals of willingness to coordinate and be trustworthy. These signals enable like-minded actors to locate others who are equally so, and who can in turn build up a reputation for responding cooperatively to trust (Bates 1988). Because actors prefer coordination as one of the outcomes, as opposed to noncoordination, they will use whatever salient information (focal point) they have in order to achieve that coordination (Schelling 1960). In time, other actors will follow suit, and a social convention is established which provides information about the future actions of others (Knight 1995). The focal point may be relatively contingent, but henceforth serves as a point of reference for future coordinating actions (Calvert 1986, Schelling 1960). Third, irrespective of the path along which an institutional solution develops, be it design or spontaneous emergence, once established, it gains a dynamism of its own. The institutional process then begins to develop its own rules beyond the formal and informal rules originally instituted, or interprets the original rules in the light of the day-to-day need to make decisions, thereby transforming and developing those rules incrementally (March and Olsen 1989). This process may change the preferences of the actors involved. As well as the path of development of an institutional solution to secure the provision of a public good or a common-pool resource, the specific context in which it is shaped and the institutional arrangements in which it is embedded have been discussed in the literature.

3. The Context and Institutional Mode of Provision

What is the general context and what are the more specific institutional and instrumental conditions in which the provision of a public good and a common-pool resource occurs? The overall question is: under which conditions do particular institutional arrangements apply particular instruments to produce favorable results for providing public goods at international level? With respect to the overall context, generally it is argued that the greater the number of participants involved in an attempt to resolve a social dilemma and the greater the heterogeneity of the actors involved, the greater the difficulty in arriving at a solution. In this sense, the prospects of an efficient solution to collective action problems for agreements at international level, with its heterogeneous actors and large numbers of participants, would appear to be bleak. However, as has been pointed out (Ostrom 1990, Martin 1994, Snidal 1984), this relationship is not as simple as it first appears. Heterogeneity may constitute a precondition for an actor to take the lead in organizing collective action because they have a

specifically strong incentive for this action to be successful. Thus, the US has repeatedly acted as a hegemonic actor in cooperating with Western public and private actors when it comes to solving collective action problems such as sharing banking information to enable countries to uncover money-laundering (Schmidt 1995). Furthermore, the problem of coordinating an unwieldy number of actors is often alleviated by organizing the participants into groups according to commonly-shared criteria. Each group is represented by a speaker, as in the case of climate negotiations, thereby reducing the number of actors sitting at the bargaining table. Moreover, the extent to which homogeneity facilitates negotiations in collective action problems depends upon the nature of the homogeneity: if all actors are intent on having the same resource, homogeneity may render agreement difficult. If, by contrast, the preferences of actors differ over time and in intensity, they may be made complementary through issue-linkage (Martin 1994).

Beyond the general context conditions, various institutional arrangements and instruments are applied in the provision of public goods and common-pool resources at international level which determine the degree of success in solving the collective action problem. The institutional arrangement denotes the structure and the nature of the relationship among the actors involved; the instruments that are applied define how the desired behavior of the target actor is to be elicited. The basic types of institutional settings for providing public goods are hierarchy, negotiation, societal self-regulation, and markets (Mayntz 1997, Scharpf 1999). Hierarchy is an arrangement in which a central actor (government) has an electoral mandate that permits it to legislate specific provisions, such as the statutory requirement to use a fully regulated catalytic converter, with noncompliance being sanctioned in order to protect the quality of ambient air. Thus, legal obligation is backed up by a bureaucratic apparatus to ensure implementation. In an international context, hierarchy as an institutional arrangement for providing public goods and protecting common-pool resources can only be applied to a very limited extent. However, even in international negotiations among sovereign states seeking to solve a collective action problem, a certain amount of initiative needs to be taken by one or a group of actors if public goods are to be provided jointly. As a result, we find power being yielded by central actors in varying degrees in order to facilitate cooperation. Thus, a central body may consist of member-actors who only make framework decisions that subsequently are made more specific by actors at the decentral level, or the central body may delegate power to members to implement the decisions. However, it is also conceivable that the outcome of autonomous actors cooperating voluntarily is strict implementation by command-and-control once a decision is agreed. Depending on whether the decision is vaguely, as

opposed to precisely, formulated the ensuing latitude for such actors is either great or little. Power to sanction noncompliance may again be vested in a central body or, alternatively, in decentral bodies (Martin 1994).

Another important institutional setting in which solutions to collective action problems at international level are found is negotiations in bipartite, tripartite, or multipartite actor constellations. However, it can often be linked to elements of various forms of hierarchy. Thus, if negotiations are not successful in bringing about a solution to a problem, actors may move out of the arena and turn to a court or a tribunal for a ruling in order to overcome the conflict. In anticipating this possibility, actors may be more willing to compromise. Thus, in the European dispute over social policy, if social partners do not negotiate an agreement successfully, the Commission may propose legislation for submission to the Council of Ministers. The 'shadow of hierarchy' (Scharpf 1997) speeds up negotiations. Some actors are particularly apt at brokering agreements in negotiations at international level in order to bring about a common solution. In the European context, the Commission has played a pre-eminent role in hammering out such agreements by acting as a policy entrepreneur and broker in different arenas, bargaining with various actors simultaneously in order to develop solutions acceptable to all, for example, by proposing package deals, compensation payments, etc. (Grande 1995).

Yet another institutional arrangement is the self-organization and self-regulation of actors. Self-regulation in providing a public good or common-pool resource implies that the same actors who make joint decisions about how to provide the common good also implement these decisions. The actors concerned build on the motivation, expertise, and resources of the actors immediately affected in order to provide public goods (Mayntz and Scharpf 1995). These forms of self-governance are also linked frequently to elements of hierarchy, such as the possibility of appeal to a third-party tribunal (Ostrom 1990).

Finally, markets may serve as an institutional context for the provision of public goods and common-pool resources under specific conditions, for example, when a group of actors have such high stakes that they engage in the provision. It has been increasingly claimed that quasimarkets should be created to provide public goods and common-pool resources (Le Grand 1991). Thus, many Western countries have promoted the liberalization of state monopolies in the provision of network utilities. However, a pure form of market as an institutional arrangement for providing public goods and common-pool resources can rarely be found either. Rather—as, for example, in the case of the provision of network utilities—markets are regulated at two levels: in terms of market creation to guarantee competition on the one hand, and at the level of market correction as regards 'service public'

obligations to guarantee the accessibility and affordability of these services. Here—as with other institutional arrangements—two or three different types of institutional arrangements for providing public goods or common-pool resources are combined rather than existing separately as pure forms.

A further, more contextual institutional mechanism for solving collective action problems is the development of 'social capital' and trust among the actors concerned (Taylor 1996, Hardin 1993). The underlying notion is that you can trust the others involved to cooperate in the provision of a public good. Trusting others is reasonable because it is in their interest to be trustworthy in a specific situation in which a social dilemma is to be overcome. One's own trust depends on the interest of the trusted (Hardin 1999).

The instruments of command-and-control, bargaining, setting incentives, persuasion, and information are often closely linked to the institutional settings in which they are applied. Thus, command-and-control is often applied in a hierarchical context where a government authority requires the regulatees to behave in a particular way and where sanctions are applied in cases of noncompliance. However, a majority decision can also be made in favor of applying incentives. Hence, hierarchy may be linked to instruments other than command-and-control. In the institutional setting of negotiation in bi-, tri-, and multipartist settings, bargaining, together with issue-linkage, compensation payments, and compromises are the main instruments used to reach agreement, while deliberation and argument also play an important part. The negotiation process may, however, also result in the application of a strict prescription of behavior linked with the levying of fines. Incentives are set in the institutional environment of the market. Trust would be linked to information and persuasion.

4. Implementation

Agreement on a joint institutional solution to provide a public good does not automatically guarantee its implementation. The path from decision-making to the final realization of the measure is a long one, and many obstacles may prevent it being a smooth process (Pressman and Wildavsky 1973, Mayntz 1983, Windhoff-Héritier 1980). At international level, typically we are faced with the situation that the definition of the goods and institutional solutions for providing the public good or common-pool resource occur at international level, while implementation takes place at national level. This may raise particular problems of compliance. Thus, for example, if an international environmental agreement stipulates the extent and time-span for a reduction in CO₂ emissions, this does not mean that the signatory states, in cooperation with industry, will necessarily honor their obligations.

Research on implementation concludes that solutions for providing public goods that are redistributive in nature have little chance of being realized, whereas solutions with distributive outcomes, where all stand to win, meet with fewer difficulties. This applies as long as the redistributive solutions do not rely on hierarchy/command-and-control that prevents actors who are adversely affected by the decision from trying to change the contents of the decision in the course of implementation; and as long as those who are to benefit from the potential solution do not mobilize in order to exert pressure for a speedy implementation process (Windhoff-Héritier 1980). However, since hierarchical guidance and command-and-control can only be used to a limited extent in international governance, effective implementation depends much more on setting incentives which induce actors, over whom there is no hierarchical control, to comply with the agreed solution. Alternatively, it depends on societal mobilization where the actors profiting from the agreement press those who resist implementation into compliance. The same effect may be achieved if an agreed provision is deeply embedded in existing societal value systems. Furthermore, by using reputational mechanisms parties which do not comply are 'named and shamed' into compliance. The noncompliance of a state would result in the loss of reputation.

5. Outcomes and Impact of the Institutional Provision of Public Goods

A final question relates to the outcomes and impacts of specific institutional provision of a public good at international level. What are the policy outcomes of such a solution and how does it affect existing governance structures? The first question addresses the problem-solving capacity of an institutional solution, i.e., the performance of the institutional setting that has been created and the instruments applied in solving the problem. Thus, the question may be raised: does an international environmental agreement help protect endangered natural resources? Since institutional solutions are largely negotiated with a view to their potential distributional outcomes, their *de facto* distributive effects must be identified. Who gains and who loses on account of a specific institutional solution, and what does this mean in terms of the stability of the chosen solution? The second aspect relates to the fact that specific institutional solutions to public good problems across boundaries affect existing governmental and governance structures. Thus, the formal decision-making powers and the *de facto* power of national actors are changed by specific institutional solutions to problems. A case in point is the attempts of the European Union to solve collective action problems in environmental policy that have challenged the competencies of subnational actors in the member

states. Similarly, the *de facto* involvement of Euro-associations in policy shaping has implied a loss of political power of national associations. Finally, the question must be raised as to how the overall structure and functions of the traditional state change in an environment where public goods are provided by forms of international governance. Does it imply the demise of the traditional state (Majone 1996, Grande and Eberlein 1999) in direction of a the regulatory state?

See also: Action, Collective; Global Governance; International and Transboundary Accords, Environmental; International Organization

Bibliography

- Bates R H 1988 Contra contractarianism: Some reflections on the new institutionalism. *Politics and Society* 16: 387–401
- Bator F M 1958 The anatomy of market failure. *Quarterly Journal of Economics* 72(August): 351–79
- Buchanan J M, Stubblebine W C 1962 Externality. *Economica* 29: 371–84
- Calvert R L 1986 *Models of Imperfect Information in Politics*. Harwood Academic, Chur, Switzerland
- Coase R H 1960 The problem of social cost. *Journal of Law and Economics* 3
- Cornes R, Sandler T 1996 *The Theory of Externalities, Public Goods, and Club Goods*, 2nd edn. Cambridge University Press, Cambridge, UK
- Grande E 1995 Regieren im verflochtenen Verhandlungssystemen. In: Mayntz R, Scharpf F W (eds.) *Gesellschaftliche Selbstregulierung und politische Steuerung*. Campus, Frankfurt, Germany
- Grande E, Eberlein B 1999 Der Aufstieg des Regulierungsstaates. Zur Transformation der politischen Ökonomie der Bundesrepublik Deutschland. [The Rise of the Regulatory State. The Transformation of the Political Economy of the Federal Republic of Germany]. Presentation March 1999 at the Max-Planck-Projektgruppe Recht der Gemeinschaftsgüter, Bonn, Germany
- Hardin R 1993 The street-level epistemology of trust. *Politics and Society* 21(4): 505–29
- Hardin R 1999 *Liberalism, Constitutionalism, and Democracy*. Oxford University Press, Oxford, UK
- Knight J 1995 Models, interpretations and theories: Constructing explanations of institutional emergence and change. In: Knight J, Sened I (eds.) *Explaining Social Institutions*. University of Michigan Press, Ann Arbor, MI
- Knight J, Johnson J 1994 Aggregation and deliberation: On the possibility of democratic legitimacy. *Political Theory* 22: 277–96
- Le Grand J 1991 *Equity and Choice: An Essay in Economics and Applied Philosophy*. Harper Collins Academic, London
- Majone G 1996 *Regulating Europe*. Routledge, London
- March J G, Olsen J P 1989 *Rediscovering Institutions: The Organizational Basis of Politics*. Free Press, New York
- Martin L L 1994 Heterogeneity, linkage and commons problems. *Journal of Theoretical Politics* 6(4): 473–93
- Mayntz R 1983 Zur Einleitung: Probleme der Theoriebildung in der Implementationsforschung. In: Mayntz R (ed.) *Implemen-*

- tation politischer Programme II: Ansätze zur Theoriebildung [The Implementation of Political Programmes II: First Attempts to Develop a Theory]. Westdeutscher Verlag, Opladen, Germany
- Mayntz R 1997 *Soziologie der öffentlichen Verwaltung*, 4th edn. [Sociology in Public Administration]. Mueller Verlag, Heidelberg, Germany
- Mayntz R, Scharpf F W (eds.) 1995 *Gesellschaftliche Selbstregulierung und politische Steuerung* [Societal Self-regulation and Political Steering]. Campus, Frankfurt, Germany
- Musgrave R A, Musgrave P B 1976 *Public Finance in Theory and Practice*, 2nd edn. McGraw-Hill, New York
- Ostrom E 1990 *Governing the Commons. The Evolution of Institutions for Collective Action*. Cambridge University Press, Cambridge, UK
- Pressman J L, Wildavsky A B 1973 *Implementation: How Great Expectations in Washington Are Dashed in Oakland*. University of California Press, Berkeley, CA
- Pigou A C 1920 *The Economics of Welfare*. Macmillan, London
- Samuelson P A 1954 The pure theory of public expenditure. *Review of Economics and Statistics* 36: 387–9
- Scharpf F W 1997 *Games Real Actors Play: Actor-centered Institutionalism in Policy Research*. Westview Press, Boulder, CO
- Scharpf F W 1999 *Regieren in Europa: effektiv und demokratisch?* [Governing Europe: Effective and Democratic? Oxford University Press, Oxford, UK]. Campus Verlag, Frankfurt, Germany
- Schelling T C 1960 *The Strategy of Conflict*. Harvard University Press, Cambridge, MA
- Schmidt V A 1995 The new world order, incorporated: the rise of business and the decline of the nation-state. European University Institute working paper 5
- Snidal D J 1984 *Interdependence, Regimes and International Cooperation*. University Microfilms International, Anne Arbor, MI
- Taylor M 1996 Good government: on hierarchy, social capital and the limitations of rational choice theory. *The Journal of Political Philosophy* 4(1): 1–28
- Windhoff-Heritier A 1980 *Politikimplementierung: Ziel und Wirklichkeit politischer Entscheidungen* [Policy Implementation: Aims and Reality in Political Decision-making]. Hain, Königstein/Taunus, Germany

A. Héritier

Public Health

Calling to mind Pasteur, his students and successors, Frederic Gorham (1921) wrote, 'In fifty short years, the world has been transformed by their discoveries.' A 'blissful optimism' surrounded public health. In the following decades, the advent of DDT, sulfa drugs, penicillin, and more recent antibiotics would endow public health doctors with unshakeable self-confidence. In contrast, the rapid expansion of health care systems during the 1960s signaled a decline for public health until the 'comeback' of infectious diseases, the

prevalence of chronic illnesses, the increasing number of 'complex emergencies' (famines, wars, 'ethnic cleansing,' natural catastrophes), the growth of inequality, the development of resistance against proven drugs, the proliferation of technological risks, the economic recession and the spread of poverty—all proved the need for stronger structures and political support, especially for monitoring and controlling public health.

In the nineteenth century, sanitarians saw poverty as a product of disease. In the twentieth century, social medicine reversed the relation by affirming that destitution and exploitation bred disease. Although there is hardly any doubt that malaria, to take one example, is a cause and not an effect of underdevelopment (in Kenya and Nigeria, it costs from two to six percent of the GNP), can we not see the aggressive return of poverty in industrial societies since the 1980s or the ongoing impoverishment of the Third World as substantial evidence that increasing poverty is the major factor in current health crises?

Trends in health policy closely depend on the relation between the requirements of social justice and the objectives of political authorities. Debate about the utopia of peaceful cooperation among citizens and among nations along with the political decision to provide for national security—of which 'health security' is a part—have marked public health's history. What is original at the end of the twentieth century is that public health has been drawn off the sidelines of charity, its traditional quarters, and placed center field in world politics and nation-state strategies.

1. Health as a Local Concern

Under exceptional circumstances, public health rises to the rank of a duty for centralized states. Such has been the case when the purpose was nation building or the protection of the national territory and its population. But under usual circumstances, public health has been a local matter.

Owing to these deep local roots, the philosophical grounds for public health have been the science and art of preventing illness and of furthering the health and well-being of all thanks to an organized community effort. According to C. E. A. Winslow (1923), public health involves far more than granting health authorities police powers; it is 'not a concrete intellectual discipline, but a field of social activity.' This is still an accepted notion. In 1936, certain thinkers maintained that medicine was not a branch of science but that, if a science, it was then a social science. This position was decisive not only because it sparked debate about the history of health (as part of social history or of the history of science?) but also because it foreshadowed another debate, namely, the current controversy between public health as a social science

and epidemiology as a biomedical science. Public health is a practical science halfway between the laboratory and the public administration. Public health doctors must be, all at once, experts in the field of prevention, judges in matters of community health, and leaders. More than just clinicians, they are (in Winslow's words) scientists, administrators, and statesmen.

At present, there is no consensus about public health, about its duties, jurisdiction, and qualifications. Historically, these have seldom been clearly defined in relation to medicine, bacteriology, epidemiology, and laboratories—a sensitive question in tropical medicine. Starting in 1948, and even more as of 1965 in the United States, the gap between biomedical research (National Institute of Health), epidemiology (Centers for Disease Control), and public health (US Public Health Service) has never stopped widening.

Public health has been understood as a government duty—by revolutionaries in 1848 or the architects of state medicine in Great Britain. This is still a generally accepted idea, even in the United States despite opposition in the past not just from doctors but from hygienists themselves. Despite a chronic staff shortage, public health in the USA still provides a wide array of services coordinated by federal, state, and local governments. As the Secretary of Health, Education and Welfare declared in 1979, these health services represent 'a governmental presence at the local level.' Public health's history has been marked by swings back and forth between a narrow focus on issues and a broad public duty, between preventive and curative medicine, between hygienists and doctors.

Noting the epidemiological transition underway during the 1920s, Winslow wanted public health to restore the broken link between preventive and curative medicine, and to infuse students and practitioners with a 'spirit of prevention.' At that time, medical schools in Scandinavia and, of course, the Soviet Union offered courses in social medicine. In contrast, the 1933 declaration of the Committee of Administrative Practice of the American Public Health Association (of which Winslow had been the first president) deliberately split the field of health in two. Offices of hygiene were assigned the tasks to fight against contagious diseases and protect children's health, as reflected in the new definition of the 'basic six' duties of local health institutions (statistics, sanitation, contagious diseases, laboratories, maternal and child care, health education) and in the national plan ('Local health units for the nation') published by the aforementioned committee in 1940 and 1945.

Public health still has far to go to resolve its ambiguous assignment. Basing their argument on the variety and multiplicity of health determinants, the authors of WHO's 1999 *Report* stated that public health doctors should measure the effects on health of factors lying outside the field of medicine. But just as

soon, they admitted that these doctors had quite limited qualifications, jurisdiction, and credibility outside their field, and that they could better concentrate their energy in the health field.

2. *The People's War for Health*

Public health cannot deny that it leaned for a long time toward social reform—specifically, toward sociology at a time when the latter was more a form of social and political analysis than the academic discipline it has since become. Even before World War I, Newsholme in Brighton, Calmette in working class areas in northern France, Goldberger and Sydenstricker in the mining towns of South Carolina and, following the war, Stampar in the Balkans, Rubinow, Sigerist, and the Milbank Memorial Fund in the USA, René Sand a little later in Belgium, a current was proclaiming itself the heir of Virchow (or rather of the threesome Frank–Virchow–Grotjahn) in Germany, social medicine's homeland. Eighty years later, the advocates of social medicine, public health, and human rights still lay claim to this heritage.

For Warren and Sydenstricker in 1918, overcrowding and insalubrious housing were the major determinants of inequality in health. Illness and health had social, economic, as well as biological causes; Sigerist vastly widened this range of health determinants by including cultural causes. Public health—a national health program—should set the goal of universal protection so as to 'equalize social classes in relation to health' in the words of the Pasteurist Étienne Burnet in 1935. More recently, an interpretation prevailing in some quarters following Alma-Ata (1978) has sought to 'put medicine where it belongs' by seeing public health as reaching far beyond illness treatment and prevention. But how to implement this 'social medical contract,' which called for redefining democracy?

The idea of social medicine was not at all unanimously approved. Typical of differing opinions was the debate about what caused malaria: an illness bred by poverty and poor hygiene, as some claimed, or a clinical and entomological problem that, as others asserted, called for eradicating the vector rather than poverty? The very phrase 'social medicine' was cause for debate. Did it not risk making people believe public health should be based on the medical profession, whereas its only true allies were local authorities and the people?

The 1929 depression had a part in this shift toward a more sociological conception of public health. Studies in nutrition or on the links between unemployment and health took sides with a movement opposed to eugenics. Social medicine was but a clinical medicine aware of the social determinants of a patient's state of health. This awareness would inevitably call for 'socializing' health systems and opening

health care to everyone. But 'social' should not be confused with 'socialist.' The three terms 'social-socialized-socialist' tended to correspond to a political geography that might, on closer examination, turn out to be little more than a series of cultural gradients: Europe tended to be more socialist as one moved from areas using horse-power to areas using horses! The Russian 'example' must be set in this light. The Webbs, Winslow, Sigerist, like delegates at the 1978 Alma-Ata Conference, were enthusiastic about Soviet medicine (or even about community medicine in Russia since Alexander II's reign). Even though Moscow had proclaimed the doctrine of the social origins of disease, this guiding Eastern light could not outshine other sources of luminosity. In Socialist Berlin and Vienna—not to mention East Harlem—health centers and dispensaries were springing up that provided maternal and child care as well as birth control.

Once 'war Communism' came to an end, social medicine in the former USSR took a radical turn into state medicine. As conveyed to the USA under Franklin D. Roosevelt, it served, for a while, as a political alternative to the mercantile spirit of the American Medical Association. In the absence of a Federal health plan, Oklahoma and Maryland, with the backing of the Farm Security Administration, farmer organizations, labor unions, and medical cooperatives, actively defended rural interests. In its radical versions, social medicine focused on zones with high rates of poverty, malnutrition, and child mortality—areas abandoned by hospitals, which, located in cities, were too far away and too expensive. In the late 1960s once again, Tanzania gave priority to rural areas as it built a decentralized system, covering the whole country, of health stations and dispensaries, and appointed village health agents. In 1972, the proportion of funds appropriated there for rural compared with urban health services was 20:80. Eight years later, it stood at 40:60. But WHO did not reorient its primary health care strategy until 1989 in order to cope with massive rural flight toward urban areas.

3. *Alma-Ata, the Peak of Convergence*

Following World War II, social medicine experienced a golden age. The coming to power of the socialist Cooperative Commonwealth Federation in Saskatchewan, Canada in 1944, and the emergence of the New Democratic Party at the national level in 1961 opened the way for setting up, between 1962 and 1971, a national health insurance based on values of solidarity and social justice. Even in the United States, a hierarchical organization of health centers existed in New York City. In contrast, social medicine was losing wind in Europe even as it was attaining official recognition by international organizations.

Social medicine—'promotive, preventive, curative, and rehabilitative,' to borrow Sigerist's words—was crowned in Alma-Ata at the joint WHO–UNICEF conference on September 12, 1978. Potentially subversive ideas had been carefully defused by replacing certain phrases (e.g., 'health centers' with 'primary health care' or 'barefoot doctors' with 'community health workers') in the 'Health for all by the year 2000' declaration already approved by WHO's 30th General Assembly in May 1977. This 1978 Alma-Ata Conference represented a peak for the theory of convergence, itself the outcome of thawed relations between the two blocs, as the former Soviet Union sought to cooperate with the West in science and technology while still exporting its model of development toward the Third World. Shortly after the conference, WHO's regional committee for Europe would refer to this theory as 'Equity in care.' Borne by the expanding social sciences and by development theory (after all, 'Health for all by the year 2000' was intended to be an indispensable part of development), this theory stipulated that differing social systems would ultimately converge into a single 'human society' for satisfying all basic needs (food, housing, health care) after eliminating poverty. Current in the 1960s and 1970s, this ideology might be seen as a new form of the functionalism that had thrived 30 years earlier at the League of Nations. Food, housing, social insurance, paid holidays, cooperatives, leisure activities, education, and especially public health, all this would shrink the field of politics as welfare expanded. Would national administrations not be integrated into a single international system that would free Geneva from the domination of sovereign states?

The Alma-Ata conference defined the public for primary health care in terms of high morbidity and mortality rates and a socioeconomic status with risks for health. Health education, a safe water and food supply, maternal and child care (including birth control), and vaccination (against tuberculosis, diphtheria, neonatal tetanus, polio, and measles) were declared priorities along with universal access to health services and the necessity for coordinating health with education and housing. Primary health-care was to be based on the most recent scientific developments, which were to be made universally available—what Aneurin Bevan in 1946 had called 'universalizing the best.' Adopted by more than 100 states in 1981, the 1978 Alma-Ata Declaration sought to reflect criticisms both of the Western medical model and of the population explosion in poor lands. In particular, it systematically demolished the 'vertical' programs that WHO had, until then, undertaken with quite inconsistent results (success against smallpox but a relative setback in combating malaria). Based on participatory democracy, 'peripheral' services were to lie at the core of the new strategy, but were also to be integrated into 'districts' around a first referral-level hospital (140 beds, internal medicine, gynecology,

obstetrics, a laboratory and blood bank, and the capacity for performing minor surgery and providing outpatient care). The aim was not to juxtapose primary care with a still centralized system but to combine the advantages of hospitals and of health centers so as to facilitate both prevention work and access to medical and social services for a poor, often scattered, population.

To its credit, primary health care has made major, though inconsistent, progress. In Colombia, the 1500 health centers set up as of 1976 were covering 82 percent of the population by the early 1980s. Between 1978 and 1982, the percentage of the vaccinated population rose from 23 percent to 41 percent for polio, 36–71 percent for tuberculosis, 21–50 percent for measles, and 22–37 percent for diphtheria, tetanus, and pertusis. If, by 1989, 90 percent of babies less than one year old had been vaccinated against polio in Cuba, Botswana, Brazil, Nicaragua, China, Tunisia, Egypt, Korea, and Saudi Arabia, in India this rate did not reach 20 percent in the two-thirds of the poorest strata of the population, where infant mortality is the highest.

True, the methods used were sometimes counter-productive. As we know, the Expanded Program on Immunization (launched in 1974 and broadened in 1987 to include polio) had better results in places with a solid health infrastructure. In continuity with the Rockefeller Foundation's campaigns against yellow fever in Latin America, and those against 'sleeping sickness' in the Belgian Congo or French Equatorial Africa, or the campaigns against smallpox organized by WHO in association with Indian and Bengali authorities in 1962 and 1967, international organizations still preferred mass vaccination programs conducted like military campaigns. Coercion was all the more willingly used insofar as vaccinators blamed the superstitious population for any opposition.

The installation of 'community health agents' elected by villagers, would, it was believed, overcome local opposition. This lightweight health auxiliary, a sort of 'barefoot doctor,' would bridge the gap between the health system and rural dwellers. These experiments were disappointing: this community health agent often turned out to be a 'transitional last resort' who fell victim to the envy aroused by power or fell into the 'voluntary help' trap. True, public health provided certain governments (in Cuba, Sandinist Nicaragua, or Nyerere's Tanzania) with means for mobilizing the population and nation building. But five years after the signing of the Alma-Ata agreement, only experimental or small-scale projects were in the works, mainly as a function of circumstances, when a local leader showed concern for his constituents' welfare.

Under the pressure of users, prevention work in health centers often deviated toward more curative purposes. Like state medicine in the Soviet Union, social medicine in the Third World represented a

limited injection of 'collective vaccine' in an 'individualistic' village setting. Except for Scandinavia, where national identities are strongly imbued with concern for health and with eugenic ideas, Europe, too, has tended to resist this type of medicine.

Primary health care was originally intended to better distribute public expenditures by avoiding congestion in big urban hospitals—an objective that the Harare Conference, organized by the WHO, reaffirmed in 1987 but that has been a resounding failure. At the Bamako meeting organized by WHO and UNICEF in 1987, African ministers of Health had to admit this, but then went on to ratify a proposal for furthering decentralization and making health care users pay part of the costs. Applied in 33 countries by 1994 (28 of them in sub-Saharan Africa), the Bamako Initiative looks like an antidote to Alma-Ata.

4. Transnational Networks and the Territory's Revanche

Symbolized by the publication of the 1993 World Bank report, *Investing in Health*, the waning of WHO—which as recently as 1973 had been said to be the 'world's health conscience'—signals a turning point. In Africa, from 60 to 70 percent of health budgets come from outside sources: non-governmental organizations (NGOs), the World Bank, or the International Monetary Fund. The World Bank's centralized approach allows little room for a local or national diagnosis. Zimbabwe, where expenditures on prevention work had increased significantly from 1982 to 1990, was, in 1991, summoned to reduce its public health investments by 30 percent in line with the Bank's recommendations for shifting funds toward 'selective' and more productive primary health care activities and toward prevention work for reducing maternal and infant mortality (such as growth-monitoring, rehydration, breast-feeding, and immunization). The Bank's recommendation to start making users pay may well reduce the utilization of health services and thus worsen the prevalence of endemic diseases such as malaria or AIDS. Another drawback is as follows: the need to obtain fast results in order to please donors keeps governments who receive aid from acquiring the ability to manage in the long term.

The World Bank's growing role in financing the health field can but amplify its intellectual role and lead to neglecting WHO's abilities and experience. Unlike Geneva, whose relations with national administrations are still limited to (weak) ministries of Health, the Bank has direct access to the centers of power (ministers of National Planning or Finances). It may even, as happened in Bangladesh, take the place of the state in heading health policy. As happened with the interventions by the Rockefeller Foundation or the Milbank Memorial Fund in France during the

1920s or in several central European and Balkan lands, the organization of public health in low-income countries comes to be based on an imported state. Since the 1980s, however, the big foundations have given up their place in the field to other non NGOs specializing in humanitarian aid. The number of NGOs rose from 1,600 in 1980 to 2,970 in 1993. A trend on this scale is a sign of the times, namely: the proportion of funds reserved for emergency actions in war zones has increased constantly since 1990. In Asia, Africa, or the Balkans, public health is, once again, seen against the backdrop of warfare.

Welfare and warfare were, indeed, intimately connected between the two world wars. Interventionism thrived with World War I and the 'Spanish flu' epidemic. As a consequence of growing government interventionism between 1914 and 1918, the first health ministries sprung up in Europe, New Zealand, and South Africa. Following the Great War, there was a general awareness of the role of public health in reconstruction and the dangers of mass epidemics (such as cholera or typhus) for political stability. World War I stimulated reformers like Dawson, who can rightfully be considered to be a forerunner to the British National Health Service (1948). But by doing away with taboos surrounding the respect for life, the war also opened wide the doors for the excesses committed by a racial form of hygienics.

Throughout the inter-World War period, a tragic dialectics prevailed in the international health order (at least in Europe), a constant swinging back and forth between the triumph of networks of transnational experts and the revanche of the territory, i.e., the nation or 'race.' Hitler's regime did not hesitate to use public health for a policy genocide. Its fight against typhus in Poland led to 'the final consolidation of the notion of a quarantine' in the Warsaw ghetto (Browning 1992). This was proof of the potential lethality of the health sciences.

As the twenty-first century begins, war is, once again, menacing health. It has even been said that 'today's armed conflicts are essentially wars on public health' (Perrin 1999). Civilian populations are the first war casualties. Since the wars in Abyssinia and Spain during the 1930s (well before Rwanda and the first Chechnya war), 80 percent of the total 20 million persons killed and 60 million wounded in various conflicts have been civilians. Three out of five have been children. The most intense war zone in Africa lies at the heart of the continent, which is also the epicenter of new diseases (Ebola) and epidemics (HIV/AIDS). But civilians fall casualty not just as a result of war's toll on health, as it sets off mass migrations and disrupts food supplies. Between 1982 and 1986, fighting destroyed more than 40 percent of Mozambique's health centers. By interrupting routine activities related, in particular, to maternal and child care, war increases mortality and the number of premature births. It also reduces vaccination coverage,

not to mention its long-run effects, such as outbreaks of malaria as people seek refuge in infested areas. According to a 1996 UNICEF estimate, war-related shortages of food and medical services in Africa, added onto the stress of flight from battle zones, accounted for 20 times more deaths than actual fighting. On the one hand, the renewed threat of infectious diseases (in particular, the interaction between HIV and tuberculosis in poor countries) and, on the other, the consequences of the epidemiological transition and the aging of populations in wealthy lands, this is the 'double burden' that, according to the WHO, public health will have to carry in the twenty-first century.

Prevention work now represents a very small percentage of public health expenditures: 2.9 percent in the United States, a statistic similar to that in most other industrialized lands. However, this percentage includes only prevention activities such as screening for disease or vaccinating. It is not to be confused with 'prevention education,' such as campaigns against smoking. 'Health for all by the year 2000' is still a far cry from reality, even in wealthy nations.

The somewhat vague phrase 'health promotion,' officially used in the British Health Ministry's 1944 report, refers, all at once, to health education, taxation as a health policy instrument, and the prohibition against advertising alcoholic beverages or tobacco. This phrase was revived in the 1970s by the Canadian Health Ministry to support the thesis that a population's state of health depends less on medical progress than on its standards of living and nutrition. The chapter devoted to the fight against nicotine addiction in the 'Health for all' program has been followed up on, since it does not cost the government anything and can even, as in Australia, help pay for 'health promotion' thanks to cigarette taxes. To its advantage, it also garners broad support for public health. The keynote of the 1986 Ottawa Charter, the phrase 'health promotion' indicates how closely prevention work in wealthy countries hinges on the system of medical care. The state itself participates in this trend to make individuals responsible for health. Witness the campaigns focused on tobacco, alcohol, or certain foodstuffs with the objective of changing 'lifestyles'—a far cry from the 'old' public health which depended on the dispensary and other facilities as the basic means of education.

Consequent to discussions in the late 1970s about women's role in the sociology of health, the awareness dawned that more persons are involved in health care in the often neglected 'informal' sector than in public or private health services. Everywhere around the world, family, kin, and neighbors bear the largest burden of care for the sick. Evidence of this awareness is, on the right, the defense of the family's role and, on the left, the activism of patients' associations (especially homosexual associations in the fight against AIDS). Owing to its ability to mobilize various sorts of

actors with different opinions on anything other than health issues, thus 'civil' society is partly taking the state's place in both wealthy and poor countries. As this emergence of local actors once again proves, applying decisions in prevention work is not at all a purely administrative matter.

Have governments been put to rout? But public health raises problems directly related to the state's powers. Grounding its arguments in new bacteriological concepts, New York City's Health Service had no qualms at the start of the twentieth century about locking up a 37-year-old Irish cook ('Typhoid Mary' Mallon) for having infected 32 persons. Given the tragic affairs of blood transfusions in France and 'mad cow disease' in the UK, public health now seems to have become a fully fledged national security issue. The most recent evidence of this is the 'principle of precaution,' which—the ideology of national sovereignty tempered with expertise—would entail political decisions for strengthening public health. State security increasingly means taking into account not only political and military threats but also public health and environmental risks. As the concept of security expands into a 'shared human security,' monitoring transmissible diseases is becoming a major concern for public authorities.

On the threshold of the twenty-first century, a global public health strategy would entail relating the fight against transmissible diseases and environmentally related risks to the management of emergencies. But was this not also the issue 100 years earlier? The twentieth century started with the French–English dispute about cholera, which centered around the quarantine issue. The conflict between national independence and public health has persisted during controversial affairs all the way down through the century. Seen from Europe, the century seems to be ending with a clash about 'mad cow disease' between Great Britain, on the one hand, and Germany and France, on the other. By reviving the old Churchill myth of 'Britain alone against Europe,' the beef war (1996–2000) fully illustrates that public health, far from being an ordinary administrative matter, can turn into a fully fledged political controversy and field for government actions.

See also: Health Care Delivery Services; Health Economics; Health in Developing Countries; Cultural Concerns; Health Policy; Public Health as a Social Science

Bibliography

- Browning C R 1992 Genocide and public health: German doctors and Polish Jews, 1939–1941. In: *The Path to Genocide*. Harvard University Press, Cambridge, MA
- Burnet E 1935 Médecine expérimentale et médecine sociale. *Review of Hygiene* 57: 321–42

Fee E, Brown T (eds.) 1997 *Making Medical History: The Life and Times of Henry E. Sigerist*. The Johns Hopkins University Press, Baltimore, MD

Gorham F P 1921 The history of bacteriology and its contribution to public health work. In: Porcher Ravenel M (ed.) *A Half-Century of Public Health*. American Public Health Association, New York

Perrin P 1999 The risks of military participation. In: Leaning J, Briggs S M, Chen L C (eds.) *Humanitarian Crises: The Medical and Public Health Response*. Harvard University Press, Cambridge, MA

Viselsteart A J 1982 Compulsory health insurance and the definition of public health. In: Numbers R L (ed.) *Compulsory Health Insurance: The Continuing American Debate*. Greenwood Press, Westport, CT

Warren B S, Sydenstricker 1918 The relation of wages to the public health. *American Journal of Public Health* 8: 883–7 [1999 *American Journal of Public Health* 89: 1641–3]

Webster C (ed.) 1995 *Caring for Health: History and Diversity*. The Open University Press, Buckingham, UK

Winslow C E A 1923 *The Evolution and Significance of the Modern Public Health*. Yale University Press, New Haven, CT

L. Murard and P. Zylberman

Public Health as a Social Science

Public health can be defined as:

the combination of science, skills and beliefs that are directed to the maintenance and improvement of the health of all people through collective or social actions. The programs, services, and institutions involved emphasize the prevention of disease and the health needs of the population as a whole. Public health activities change with changing technology and values, but the goals remain the same: to reduce the amount of disease, premature death and disability in the population (Last 1995, p. 134).

Health was defined by the World Health Organization (WHO) almost 40 years ago as a state of complete mental, physical, and social well-being, and not merely the absence of disease or infirmity (WHO 1964). From this definition, health is clearly more than just a physical state and is inextricably linked to personal well-being and potential, and incorporates other dimensions of human experience, such as quality of life, psychological health, and social well-being. Public health is different from many other fields of health as it focuses on preventing disease, illness and disability, and health promotion at the *population* level. *Populations* are groups of individuals with some similar characteristic(s) that unify them in some way, such as people living in rural areas, older people,

women, an indigenous group, or the population of a specific region or country.

Contemporary public health is particularly concerned with the sections of the population that are typically under- or poorly served by traditional health services, and consequently, most disadvantaged and vulnerable in terms of health. Importantly, public health is therefore concerned with the more *distal* (i.e., distant from the person) social, physical, economic, and environmental determinants of health, in addition to the more *proximal* determinants (i.e., closer to the person), such as risk factors and health behaviors.

Another distinction is often made between primary, secondary, and tertiary levels of prevention in public health. *Primary prevention* reduces the incidence of poor health by means such as vaccinations to prevent disease. *Secondary prevention* (or early detection) involves reducing the prevalence of poor health by shortening its duration and limiting its adverse social, psychological, and physical effects, such as different types of screening to identify a condition in the early stages. *Tertiary prevention* involves reducing the complications associated with poor health and minimizing disability and suffering, such as cardiac rehabilitation programs for patients after a heart attack. The boundaries between different levels of prevention are often indistinct and blurred, and a specific intervention strategy might simultaneously address the three different levels of prevention in different individuals.

The application of behavioral and social science theories and methods to improve health and prevent disease and disability at a population level can occur at multiple levels and in many different settings throughout society. Many public health initiatives are aimed at the individual or interpersonal level, and are delivered through the primary care setting, workplaces, and schools. Such programs typically have a low *reach*, in that only a limited number of people are involved in the programs and/or have access. However, these programs have high levels of *exposure* for the individuals in that they are quite intensive. Other programs can be organized and implemented through whole organizations or systems. Finally, community-wide or societal-level interventions may involve media, policy, or legislative action at a national or governmental level, and have a very large *reach*; however, program *exposure* is typically quite low.

The practice of public health is therefore very broad and can include the delivery of many different types of programs and services throughout the whole community. It can include efforts to reduce health disparities, increase access to health services, improve childhood rates of immunization, control the spread of HIV/AIDS, reduce uptake of smoking by young people, implement legislation and policies to improve the quality of water and food supplies, reduce the incidence of motor vehicle accidents, reduce pollution, and so on. In fact, the combined health of the population is more directly related to the quality and

quantity of a country's total public health effort than it is to the society's direct investment in health services *per se*.

This article discusses the extent to which theories and models from the behavioral and social sciences inform our understanding and practice of public health. The article provides a brief overview of the major public health problems and challenges confronting developed and developing countries; the contribution made by social and behavioral factors to many of these health problems; and how social and behavioral theories and models can be usefully applied to the development, implementation, and evaluation of public health programs.

1. Global Patterns of Health and Disease: The Major Public Health Challenges

Disease burden includes mortality, disability, impairment, illness, and injury arising from diseases, injuries, and risk factors. It is commonly measured by calculating the combination of years of life lost due to premature mortality and equivalent 'healthy' years of life lost due to disability. The past 100 years have seen dramatic changes in health and the patterns of disease burden. Most countries have experienced an increase in *life expectancy*, with greater numbers of people expected to reach late adulthood. In the last 40 years, life expectancy has improved more than during the entire previous span of human history, increasing from 40 years to 63 years during 1950 to 1990 in developing countries (World Bank 1993). The number of children dying before five years has similarly reduced by 50 percent since 1960 (World Bank 1993).

Developed countries in particular have seen a decline in the proportion of deaths due to *communicable and infectious diseases* such as smallpox and cholera. However, this decline has occurred simultaneously with a marked increase in *noncommunicable diseases*, particularly the *chronic diseases*. In the early 1990s, the chronic diseases of ischemic heart disease and cerebrovascular disease were identified as the leading causes of mortality, together accounting for 21 percent of all deaths in the world (WHO 1999). In 1998, 81 percent of the disease burden in high-income countries was attributable to noncommunicable diseases, in particular cardiovascular disease, mental health conditions such as depression, and malignant neoplasms, with alcohol use being one of the leading identified causes of disability (WHO 1999). Other major noncommunicable causes of death include road traffic accidents, self-inflicted injuries, diabetes, and drowning.

In developing countries, current major public health problems continue to include infectious diseases, undernutrition, and complications of childbirth, as well as the increasing prevalence of chronic diseases. Maternal conditions, HIV/AIDS, and tuberculosis

Table 1
Changes in ranking for the most important causes of death from 1990 to 2020

Disorder	Ranking		Change
	1990	2020	
Ischemic heart disease	1	1	0
Cerebrovascular disease	2	2	0
Lower respiratory infections	3	4	−1
Diarrheal diseases	4	11	−7
Perinatal disease	5	16	−11
Chronic obstructive pulmonary disease	6	3	+3
Tuberculosis	7	7	0
Measles	8	27	−19
Road traffic accidents	9	6	+3
Trachea, bronchus, lung cancers	10	5	−5
Malaria	11	29	−18
Self-inflicted injuries	12	10	+2
Cirrhosis of the liver	13	12	+1
Stomach cancer	14	8	+6
Diabetes mellitus	15	19	−4
Violence	16	14	−2
War injuries	20	15	+5
Liver cancer	21	13	+8
HIV	30	9	+21

Source: Murray and Lopez, *Lancet* 1997, **349** (9064): 1498–504.

are three of the major causes of disease burden for adults in developing countries (WHO 1999). Among children, diarrhea, acute respiratory infections, malaria, measles, and perinatal conditions account for 21 percent of all deaths in developing countries, compared with 1 percent in developed countries (WHO 1999). Developing regions account for 98 percent of all deaths in children younger than 15 years; the probability of death between birth and 15 years ranges from 1.1 percent in established market economies to 22 percent in sub-Saharan Africa (Murray and Lopez 1997). Importantly, many of these conditions are preventable; at least two million children die each year from disease for which vaccines are available (WHO 1999).

Accidents and injuries—both intentional and non-intentional—are a significant and neglected cause of health problems internationally, accounting for 18 percent of the global burden of disease (WHO 1999). In high-income countries, road traffic accidents and self-inflicted injuries are among the 10 leading causes of disease burden (WHO 1999). In less developed countries, road traffic accidents are the most significant cause of injuries, and war, violence, and self-inflicted injuries are among the leading 20 causes of lost years of healthy life (WHO 1999).

The Global Burden of Disease Study has projected future patterns of mortality and disease burden (Murray and Lopez 1997). Life expectancy is anticipated to increase notably in females to 90 years, with a lesser increase for males. Table 1 presents the

changes in ranking for the most important causes of death. Worldwide mortality from perinatal, maternal, and nutritional conditions is expected to decline. Noncommunicable diseases are expected to account for an increasing proportion of disease burden, rising from 43 percent in 1998 to 73 percent in 2020. HIV will become one of the 10 leading causes of mortality and disability, and this epidemic represents ‘an unprecedented reversal of the progress of human health’ (Murray and Lopez 1997, p. 1506). The leading causes of disease burden will include ischemic heart disease, depression, road traffic accidents, cerebrovascular disease, chronic obstructive pulmonary disease, lower respiratory infections, tuberculosis, war injuries, diarrheal disease, and HIV. Important health risk factors will include tobacco, obesity, physical inactivity, and heavy alcohol consumption. By 2020 tobacco is expected to cause more premature death and disability than any single disease, with mortality projected to increase from 3 million deaths in 1990 to 8.4 million deaths.

2. *The Behavioral and Social Epidemiology of Health and Disease*

There are a multitude of factors that can be used to explain the changing patterns of morbidity, mortality, and the spread of diseases, both globally and between regions and countries. Some of these factors are related to population growth and aging and social patterns,

while others are more related to changes in the local and global environment. An impressive amount of epidemiological evidence collected over the past 50 years has also identified the importance of a range of behavioral determinants and modifiable risk factors such as physical inactivity, smoking, and poor nutrition. The social and behavioral sciences are key disciplines underpinning public health as they have generated theories and models which provide a link between our understanding of these modifiable determinants of disease and health and strategies for disease prevention and health promotion.

The modern era for this evidence base began with the publication of the landmark study that investigated the relationship between British doctors' smoking habits and disease (Doll and Hill 1964); the first reports from the Framingham Study that identified a number of 'risk factors' for cardiovascular disease (Kannel and Gordon 1968); and the United States Surgeon General's Report on Smoking and Health (US State Surgeon General's Advisory Committee 1964). The Lalonde report (1974) and the Alameda County study (Belloc and Breslow 1972) both identified the importance of wider environmental factors such as social support, social networks, and socioeconomic factors as influences on health. In 1980, the United States Centers for Disease Control released a report estimating that 50 percent of mortality from the 10 leading causes of death in the United States could be traced to lifestyle factors (Centers for Disease Control 1980). McGinnis and Foege (1993) assessed the relative contribution of important and potentially alterable health behaviors and social factors to mortality and morbidity in the USA, and concluded that for approximately half of all deaths in 1990, the actual cause of death was attributable to health behaviors. In 1996, the report from the US Surgeon on Physical Activity and Health stated that physical activity reduced the risk of premature mortality in general, and of coronary heart disease, hypertension, colon cancer, and diabetes mellitus in particular (US Department of Health and Human Services 1996).

This rapidly expanding evidence base has been used by many countries for setting public health priorities and recommended goals and targets for addressing national public health challenges and reducing premature mortality and the burden of disease via changing lifestyle and related risk factors. Lifestyle factors targeted to influence population health include physical inactivity, diet, smoking, alcohol and other drug use, injury control, sun protective behaviors, appropriate use of medicines, immunization, sexual and reproductive health, oral hygiene, mental health, and many others.

Behavioral and social researchers have contributed to our understanding of the broad array of ecological, social, and psychological factors that influence lifestyle and health behaviors, and in turn, the etiology and pathogenesis of disease and health. In a recent edited

series of monographs, Gochman (1997) provides an extensive overview of the range of environmental, cultural, family and personal factors that have been shown to influence lifestyle and health behaviors.

Moving more *upstream* from individual behaviors, there is also increasing interest in the ways in which the broader environment influences health, in particular the specific role played by social and economic factors. The Ottawa Charter (WHO 1986) recognized the importance of adopting a more *social view of health* by stressing that in order for public health programs to be effective they must incorporate the following elements: the development of personal skills; the creation of environments which are supportive of health; the refocusing of health and related services; an increase in community participation; and finally, the development of public policy which is health enhancing. The trend towards trying to understand more about the social, economic, and environmental determinants of health is probably one of the most fundamental developments to occur in mainstream public health over the past 20 years.

Recently there has been a very explicit research focus on socioeconomic *health inequalities*, social disadvantage, and poverty as major influences of health risk factors, disease, and reduced life expectancy. Poverty is undoubtedly one of the most important, single causes of preventable death, disease, and disability; the major variations and disparities in health observed both within and between countries can be explained to a large measure by social and economic disparities. Two key reports from the United Kingdom, the Black Report (Department of Health and Social Services 1980) and the Acheson Report (Acheson 1998), have emphasized the importance of these issues and ameliorative strategies.

Future social and behavioral epidemiological research into public health will need to address more adequately the multiplicity of factors contributing to health and disease. This will involve not only studying factors at the level of the individual, such as their beliefs and behavior, but also the full array of economic, social, environmental, and cultural factors as well.

3. Theories and Models for Prevention and Change

Social and behavioral theories and models not only help to explain health-related behavior and its determinants, they can also guide the development of interventions to influence and change health-related behavior and ultimately improve health. Theories can help understand adherence to a health regime, identify the information needed to develop an effective intervention program, provide guidelines on effective communication, and identify barriers to be overcome in implementing an intervention. While some theories

are most usefully applied to understanding the role and influence of personal-level factors, others are more useful for understanding the importance of families, social relationships, and culture, and how these can be influenced in a positive fashion to promote improvements in health.

Models of individual health behavior such as the Health Belief Model and the Theory of Reasoned Action/Planned Behavior, as well as models of interpersonal behavior such as Social Cognitive Theory, have been used widely in public health and health promotion over the past 20 years and are reviewed in various publications (e.g., Glanz et al. 1997). Staging models, such as the Transtheoretical Model, have often been used to improve the potency of health promotion programs (e.g., see Glanz et al. 1997). A combination of models or theories can often assist in the development of more individually tailored and targeted strategies (see *Health Behavior: Psychosocial Theories*).

While these theories and models focus primarily on achieving changes in health behaviors, other models are more concerned with communication and diffusion of effective programs, so as to benefit the health of a larger population. McGuire's communication-behavior change model and Rogers' diffusion of innovations model are examples of these types of theories (see Glanz et al. 1997). Health behaviors, and the broad array of environmental and social influences on behavior, are usually so complex that models drawing on a number of individual theories to help understand a specific problem in a particular setting or context are usually required to develop, implement, and evaluate effective public health responses. For example, Green and Kreuter's PRECEDE-PROCEED model, social marketing theory, and ecological planning approaches are models and approaches that are used commonly in public health and health promotion (see Glanz et al. 1997).

However, within the behavioral sciences, much less attention has been given to models or theories that understand change strategies and programs within the broader environment, groups, organizations, and whole communities. This is despite increasing research evidence indicating that social and environmental factors are critically important in understanding individual attitudes, behavior, and health. The design of programs to reach populations requires an understanding of how social systems operate, how change occurs within and among systems, and how large-scale changes influence people's health behavior and health more generally. A combination of theories and models of individual behavior can be used to shift the focus of intervention strategies beyond a clinical or individual level to a population level by their application via a range of settings in the community, such as the workplace and primary health care (Oldenburg 1994). This increases the reach of intervention programs beyond a relatively small number of individuals. Many

of the social and behavioral theories or models that have been used extensively to understand and describe change at an individual or small group level can also inform our understanding of the change and diffusion process at an organizational or population-based level.

4. *Wider Implementation of the Social and Behavioral Science Evidence Base in Public Health*

With the development, implementation, and evaluation of a series of major community-based cardiovascular prevention trials during the 1970s and 1980s, widespread use was made of multilevel intervention or change strategies underpinned by behavioral and social science theories and research for the first time. Importantly, these strategies did not focus exclusively on change in the individual. Instead, preventive strategies using multiple approaches were directed at the media, legislation, and the use of restrictive policies, and involved specific settings such as schools, the workplace, health care settings, and other key groups in the community, with the aim of reducing population rates of risk behaviors, morbidity, and mortality.

The North Karelia Project in Finland (Puska et al. 1985) was initiated in response to concern that the community, at that time, had the highest risk of heart attack worldwide. Strategies included tobacco taxation and restriction, televised instruction in skills for nonsmoking and vegetable growing, and extensive organization and networking to build an education and advocacy organization. After 10 years, results indicated significant reductions in smoking, blood pressure, and cholesterol, and a 24 percent reduction in coronary heart disease mortality among middle-aged males. The Stanford Three Community Study (Farquhar and Maccoby 1977) demonstrated the feasibility and effectiveness of mass media-based educational campaigns and achieved significant reductions in cholesterol and fat intake. The Minnesota Heart Health Project (Mittelmark et al. 1986), the Pawtucket Heart Health Program (Lefebvre et al. 1987), and the Stanford Five-City Project (Farquhar et al. 1990) used interventions aimed at raising public awareness of risk factors for coronary heart disease, and changing risk behaviors through education of health professionals and environmental change programs, such as grocery store and restaurant labeling.

Worksite intervention trials, such as the Working Well Trial (Sorensen et al. 1996) and an Australian study with ambulance workers (Gomel et al. 1993), have utilized interventions aimed at favorably altering social and psychological factors known to influence health-related behaviors, including increased awareness of benefits, increased self-efficacy, increased social support, and reduced perception of barriers. Strategies in this setting can include promotion of nonsmoking policies, increasing the availability of healthy foods in

canteens and vending machines, group education sessions, and access to exercise facilities. School-based interventions provide an established setting for reaching children, adolescents, and their families to address age-pertinent health issues such as smoking, teenage pregnancy, and substance use. Strategies have targeted social reinforcement, school food service, health-related knowledge and attitudes, local ordinances, peer leaders, and social norms.

5. Evaluation of Public Health Intervention Trials

In evaluating the effectiveness of such trials, Rose (1992) states that it is important to acknowledge the differences between clinical and more population-based approaches. Clinical approaches, while more likely to result in greater mean changes in individuals, are less significant at a population level because fewer individuals are involved. In contrast, the small mean changes achieved by population-based approaches, if applied to a large number of people, have the potential to impact much more significantly on population rates of disease and health.

One of the difficulties in evaluating population-based approaches is the length of time required to demonstrate changes in population rates of health outcomes. While researchers can readily measure changes in a behavioral outcome—such as the rate of smoking cessation—following a widespread and well-conducted program, the time required for such behavioral changes to impact upon long-term morbidity and mortality—such as lung cancer—is much more problematic. Nevertheless, many of the above identified behavior-based trials demonstrate the value of social and behavioral theories and the feasibility of activating communities in the pursuit of health.

6. Diffusion and Dissemination of the Evidence Base

The ultimate value of this research must be determined by the extent to which the knowledge generated is widely disseminated, adopted, implemented, and maintained by 'users,' and ultimately, by its impact on systems and policy at a regional, state, and/or national level (i.e., *institutionalization*). While considerable effort and resources have been devoted to developing effective interventions, relatively little attention has been given to developing and researching effective methods for the *diffusion* of their use. It is important to acknowledge that the availability of relevant research findings does not in itself guarantee good practice. Such knowledge transfer involves the development of formal research policies, formalized organizational structural support, appropriate and targeted funding, formal monitoring of activity, and ongoing training.

In essence, achieving the effective diffusion of innovations, both within the general community and in organizational settings, involves change, and the change principles which underpin the diffusion process are not so different to those previously identified for understanding change at the individual, organizational, or community levels (see Oldenburg et al. 1997). At the level of the individual, family, or small group, uptake of a health promotion innovation typically involves changes in behaviors or lifestyle practices which will either reduce risk factors or promote health. At an organizational level, such as the workplace, school, or the health care setting, successful uptake of an innovation may require the introduction of particular programs or services, changes in policies or regulations, or changes in the roles and functions of particular personnel. At a broader community-wide or even societal level, the change process can involve the use of the media and changes in governmental policies and legislation, as well as coordination of a variety of other initiatives at the individual and the settings level.

However, in considering the principles that underpin the diffusion of innovations at a population level, further complexity arises from the need to consider change occurring at multiple levels, across many different settings, and resulting from the use of many different change strategies. This requires the application of multiple models and theories in order to develop frameworks with sufficient explanatory power.

7. Looking to the Future

It is difficult to be certain about the public health challenges that lie ahead in this new millennium. In all countries, there is a need to reduce preventable disease and premature death, reduce health inequalities, and develop sustainable public health programs.

Much of the social and behavioral science research that has been conducted with individuals is still relevant in the field of public health, as are many of the models and theories that have been developed and tested over the past 30 years. To move to a population-based focus, we need innovative means of implementing such knowledge and methodologies so as to increase the reach of such programs to more people, particularly those individuals and sections of society who are most disadvantaged.

Adopting a preventive approach to health, acknowledging and acting on the importance of social and environmental factors as major determinants of population health, and reducing health inequalities provides a substantial challenge to public health in the coming years. Furthermore, limited resources place an added emphasis on evaluating the affordability and sustainability of public health interventions. Public

health should be at the forefront of developing and researching effective intervention methods and strategies for current and future health challenges, with global sharing of information in terms of monitoring health outcomes and development of the evidence base. However, this research should not focus solely on developing strategies and testing their efficacy and effectiveness. There is an urgent need for research that examines larger-scale implementation and diffusion of these strategies to whole communities and populations, through the innovative use of a variety of psychological, social, and environmental intervention methods, including evidence from the social and behavioral sciences.

See also: Behavioral Medicine; Epidemiology, Social; Health Behavior: Psychosocial Theories; Health Behaviors; Health Education and Health Promotion; Health Interventions: Community-based; Health Psychology; Mental Health: Community Interventions; Public Health; Socioeconomic Status and Health; Socioeconomic Status and Health Care

Bibliography

- Acheson D 1998 *Independent Inquiry into Inequalities in Health*. The Stationary Office, London
- Belloc N B, Breslow L 1972 Relationship of physical health status and health practices. *Preventive Medicine* **1**: 409–12
- Centers for Disease Control 1980 *Ten Leading Causes of Death in the United States 1977*. US Government Printing Office, Washington, DC
- Department of Health and Social Services 1980 *Inequalities in Health: Report of a Research Working Group Chaired by Sir Douglas Black*. DHHS, London
- Doll R, Hill A B 1964 Mortality in relation to smoking: Ten years' observations of British doctors. *British Medical Journal* Vol.(i): 1399–410
- Farquhar J W, Wood P D, Breitrose H 1977 Community education for cardiovascular health. *Lancet* **1**: 192–5
- Farquhar J W, Fortmann S P, Flora J A, Taylor C B, Haskell W L, Williams P T, Maccoby N, Wood P D 1990 Effects of community-wide education on cardiovascular risk factors: The Stanford Five-City Project. *Journal of the American Medical Association* **264**: 359–65
- Glanz K, Lewis F M, Rimer B K (eds.) 1997 *Health Behavior and Health Education: Theory, Research, and Practice*. Jossey-Bass, San Francisco, CA
- Gochman D S (ed.) 1997 *Handbook of Health Behavior Research*. Plenum, New York
- Gomel M, Oldenburg B, Simpson J, Owen N 1993 Worksite cardiovascular risk reduction: A randomised trial of health risk assessment, education, counselling, and incentives. *American Journal of Public Health* **83**: 1231–8
- Kannel W B, Gordon T 1968 *The Framingham Study: An Epidemiological Investigation of Cardiovascular Disease*. US Department of Health Education and Welfare, Washington, DC
- Lalonde M 1974 *A New Perspective on the Health of Canadians: A Working Document*. Government of Canada, Ottawa, ON
- Last J M 1995 *A Dictionary of Epidemiology*. Oxford University Press, New York
- Lefebvre R C, Lasater T M, Carleton R A, Peterson G 1987 Theory and delivery of health programming in the community: The Pawtucket Heart Health Program. *Preventive Medicine* **16**: 95
- McGinnis J M, Foege W H 1993 Actual causes of death in the United States. *Journal of the American Medical Association* **270**: 2207–12
- Mittelmark M B, Luepker R V, Jacobs D R 1986 Community-wide prevention of cardiovascular disease: Education strategies of the Minnesota Heart Health Program. *Preventive Medicine* **15**: 1–17
- Murray C J L, Lopez A D 1997 Alternative projections of mortality and disability by cause 1990–2020: Global Burden of Disease Study. *Lancet* **349**(9064): 1498–504
- Oldenburg B 1994 Promotion of health: integrating the clinical and public health approaches. In: Maes S, Leventhal H, Johnston M (eds.) *International Review of Health Psychology*. Wiley, New York pp. 121–44
- Oldenburg B, Hardcastle D, Kok G 1997 Diffusion of innovations. In: Glanz K, Lewis F M, Rimer B K (eds.) *Health Behavior and Health Education: Theory, Research, and Practice*. Jossey Bass, San Francisco, CA
- Puska P, Nissinen A, Tuomilehto J, Salonen J T, Koskela K, McAlister A, Kotke T E, Maccoby N, Farquhar J W 1985 The community-based strategy to prevent coronary heart disease: Conclusions from ten years of the North Karelia Project. *Annual Review of Public Health* **6**: 147–93
- Rose G 1992 *The Strategy of Preventive Medicine*. Oxford University Press, Oxford, UK
- Sorensen G, Thompson B, Glanz K, Feng Z, Kinne S, DiClemente C, Emmons K, Heimendinger J, Probart C, Lichtenstein E 1996 Worksite based cancer prevention: primary results from the Working Well Trial. *American Journal of Public Health* **86**: 939–47
- US Department of Health and Human Services 1996 *Physical Activity and Health: A Report of the Surgeon General*. US Department of Health, Washington, DC
- US State Surgeon General's Advisory Committee 1964 *Smoking and Health*. US Department of Health, Washington, DC
- The World Bank 1993 *World Development Report 1993: Investing in Health*. Oxford University Press, New York
- World Health Organization 1986 The Ottawa Charter for Health Promotion. *Health Promotion International* **1**(4): iii–v
- World Health Organization 1964 *World Health Organization Constitution*. WHO, Geneva
- World Health Organization 1999 *World Health Report: Making a Difference*. WHO, Geneva

B. Oldenburg

Public Interest

Scholars debate the question of the public interest somewhat less than they once did, but it remains an important issue. As is true of other central concepts, the passage of time does not always lead to an increase in acuity, although it does lead to changes in emphasis

(Friedrich 1962, Lewin 1991). Much that might be considered under the rubric of the public interest, moreover, continues to be discussed in terms of associated ideas such as justice, responsibility, and community. This ambiguity complicates a topic that is itself elusive.

1. *What is the Public Interest, Conceptually?*

The issue of the public interest is best organized under four headings. First is the formal or conceptual question. This includes problems such as the meaning of public as opposed to private, self or common interest, and the meaning of interests as distinguished from other goods. 'Public' sometimes is taken as anything that is not private, so that regard for the public interest means any regard for interests or groups that are not strictly or exclusively one's own (Lewin 1991). Less broadly, a public often is said to be those affected by private groups, in the sense that people who are neither labor nor management are affected by strikes. Such publics, moreover, sometimes are taken to refer to those potentially and not just actually affected as, say, any potential passenger might be inconvenienced by a strike of air traffic controllers.

(Someone can, therefore, have both a private and a public interest in the same matter, and 'self' interest need not be identical to private interest.) In this sense we are all members of many publics. Often, however, the public interest is taken to be what is in the interests of 'all' of us, presumably as citizens within some boundary, although the questions of scope and authority are often curiously undiscussed. There are not many publics in this view but, finally, one public (Friedrich 1962, Barth 1992). The public is the group that makes law. Of course, its interests may be the sum of those of the many affected publics into which it can be divided.

Many discussions of the 'public' in these terms have in mind economic matters, or the point of view of academic economics (Friedrich 1962). A public is seen as an affected or lawmaking group of individuals that seeks satisfaction in economically calculable terms. Publics want things that market failures cause them to receive inadequately. It is serving such publics that defines the responsibilities of professionals and of businesses 'affected with the public interest.' One strength of this conception is that since the 1950s it has brought to bear an increasingly sophisticated understanding of economically rational action, both strategic and straightforward.

There are, however, limits to this perspective. They are evident in the narrowness of the economic view of the public when it is compared to other notions of what is potentially common in common goods. For, what is common is not limited to what each independently and equally can use or enjoy, as we use markets. The common also can involve things of

which we are only parts, sometimes equal or identical and sometimes not, as a face is common to its different parts or a team or family common to its members. The ties that characterize what is common, moreover, are not limited to what covers or blankets us universally, but include attaching, placing, fitting, ordering, attracting and other links integral to images of politics from Plato through Hegel. Practical and economically based notions of the public interest will, of course, often take these distinctions for granted, but deeper exploration of the concept of the public interest cannot.

Just as 'public' is often narrower than common, so too are 'interests' conceptually different from goods more broadly conceived. Discussions of the public interest often refer to Adam Smith and, therefore, to a view of what is good as what satisfies desire in the somewhat flat, calculable, and exchangeable manner of post-Hobbesian or Lockean bourgeois life. Interests are linked to what is merely interesting, not to goods as objects of passion, longing, pride or nobility or to the character and institutions that support and direct these. Public interests are what several of us happen to desire, or means to satisfaction that we generally want. Part of the conceptual task in understanding the public interest is to consider the different ways of experiencing goods so that we can grasp how interests are goods conceived in a limited manner that is linked to the premises of liberal individualism.

2. *What is the Public Interest, Substantively?*

A second question about the public interest is what it is substantively. This issue involves both scope and content. Some scholars seek to limit questions of the public interest only to certain kinds of public goods. Others use it to stand for national goals, purposes or goods more comprehensively. In practice, too, speaking of the public interest generally differs from considering the public's interest in assuring the provision to all of some specific good such as telephone service.

When the scope of the public interest is limited to economic concerns, so too is the content of what is said to comprise it; for some the public interest does not even include protecting rights. To limit the concept in this way risks confusion when 'public interest' retains its broad use in practice and even in theory; to use it broadly, however, seems to involve every question of human purpose and to require some other concept to take the place of the idea used narrowly.

Just as our understanding of the public interest conceived narrowly has benefited from developments in economic understanding, so too has substantive discussion about the public interest conceived broadly undergone notable developments since the 1950s. The triumph of liberal democracies in the Cold War has directed the concerns of the left away from socialism

simply to expanding the place of egalitarianism and community within liberal democracy itself. A wish among those who reject the revived concept of natural rights to nonetheless ground the place of rights and defend them from simple utilitarian calculations has become widespread. Concern with the place of virtue and nobility in liberal democracies and, indeed, concern with these phenomena simply has been renewed (Rawls 1971, Strauss 1952). All these trends have enriched and re-energized discussions of the public interest.

3. *Can the Public Interest be Implemented?*

The third question is how the public interest might be implemented concretely and used as a guide for action. Some thinkers argue that the 'public interest' is too vague to be used other than rhetorically or that it is merely a name that sums up a country's political and economic life at some moment frozen in analytical time (Friedrich 1962). More charitably, one might concede that it means a certain level of economic satisfaction and political freedom but that these are best reached through private and political competition, that is, when no one aims at them directly.

When we conceive the public interest broadly, however, it is not always so difficult to discern how it might guide action. Constitutions, for example, cannot be formed well without explicit attention to more than private concerns: Charles Beard's worm's eye view of the framing of the American Constitution has been debunked by more thoughtfully elevated judgments (Kesler 2000, Mansfield 1991). Once constitutionally secured, individual rights and liberties must be exercised individually, but steps to protect these rights publicly must be attended to explicitly by jurists. National defense is a public interest that guides specific legislative and executive action.

To say that the public good often receives explicit attention is not to say that it goes without debate, but to call it debatable is not to call it meaningless. The deliberations and result of the US' constitutional convention prove the contrary. Similarly, if we take our private interest to include moral and intellectual excellence and not only economic opportunity and satisfaction we would not say that the disputability of these ends makes them pointless. On the contrary, their subtlety may make our attention to them all the more intense.

If we conceive the public interest more narrowly as assuring some equitable or safe provision and distribution of certain goods then it is again clear that the public interest can be a concrete guide. Are expanding access to communications and increasing public health served better by explicit government regulation with a defined public interest in mind of such and such a level of drug safety and computer ownership, or are they better served by discovery and application fueled largely by private desires for wealth, knowledge, and

mastery, with little regulatory adjustment of the private application of resources? Is improved education secured better by unfunded parental choice, by funded choice, or by public provision or standards? Explicit attention to health, education and responsible choice as public interests seems necessary to help secure their sufficient provision and fair distribution, if only to ward off excessive regulation and allow private efforts to flourish.

4. *Is the Public Interest Followed?*

The fourth question is whether anyone in practice pays attention to the public interest or whether private interest in fact dominates even in public matters. It is argued that the actions of voters, representatives, and bureaucrats all can be explained as attempts to enhance private goods; this analysis can be expanded to professionals, journalists, and intellectuals who do public talking, if not public business, for a living. Even if the public interest could guide us, in this view it does not.

Some recent research in the US and Western Europe indicates that voters do not simply or usually vote their immediately selfish economic concerns, that the behavior of representatives cannot always or primarily be explained by their wish for re-election, that bureaucrats do not always want to increase their agency's size, and that groups and individuals sometimes do seek benefits for publics from which they and their members do not receive special advantages (Lewin 1991, Wilson 1989). Of course, a link between private and public interest is to be expected and largely desired in liberal democratic regimes, for the responsibility we encourage in citizens and officials is meant to benefit not just others but oneself as well (Kesler 2000).

Obviously, whether political action actually attempts to serve public interests requires continued study. Only the excessively naive would believe that unbridled devotion to a clearly conceived common good directs each and every political choice. But only the false or foolish sophisticate would believe the public interest to be as fleeting and airy as the summer snow.

See also: Communitarianism: Political Theory; Democracy; Environment and Common Property Institutions; Interest: History of the Concept; Political Culture; Political Representation; Political Sociology; Public Interest Law; Public Opinion: Political Aspects; Public Reason; Representation: History of the Problem

Bibliography

Antitrust and the Economics of the Public Interest 1997 Symposium: Economics of Antitrust Enforcement. *Antitrust Bulletin* 42(Spring)

- Barth T J 1992 The public interest and administrative discretion. *American Review of Public Administration* 22: 289–300
- Blitz M 2000 Liberal freedom and responsibility. In: Boxx W T, Quinlaven G (eds.) *Public Morality, Civic Virtue and the Problem of Liberalism*. Eerdmans, Grand Rapids, MI
- Box R C 1992 The administrator as trustee of the public interest: Normative ideals and daily practice. *Administration & Society* 24: 323–45
- Dewey J 1927 *The Public and Its Problems*. Holt, New York
- Friedrich C J (ed.) 1962 *The Public Interest*. Atherton, New York
- Goodin R E 1996 Institutionalizing the public interest: the defense of deadlock and beyond. *American Political Science Review* 90: 331–43
- Kesler C (ed.) 2000 *Alexander Hamilton, John Jay and James Madison, The Federalist Papers*. Random House, New York
- Lewin L 1991 *Self Interest and Public Interest in Western Politics* (trans. Lavery D). Oxford University Press, Oxford, UK
- Mansfield Jr. H C 1991 *America's Constitutional Soul*. Johns Hopkins University Press, Baltimore, MD
- Rawls J 1971 *A Theory of Justice*. Belknap Press of Harvard University Press, Cambridge, MA
- Strauss L 1952 *Natural Right and History*. University of Chicago Press, Chicago
- Wilson J Q 1989 *Bureaucracy*. Basic Books, New York

M. Blitz

Public Interest Law

Public interest law is a branch of legal practice committed to representing traditionally unrepresented groups and interests. It went through a period of rapid growth in the USA in the 1970s and at the beginning of the twenty-first century is an important, though small component of the legal system in the USA and many other countries. Public interest law is often advanced through test case litigation designed to not merely effect the outcome of a particular dispute, but to influence public policy. In tandem with the litigative work, public interest lawyers frequently engage in advocacy before government administrative agencies and legislative bodies.

Public interest law can be distinguished from legal services that focus on representing the personal legal needs of individual clients. The outcome of a public interest lawsuit is likely to reshape public policy or affect the interests of a substantial population. The class action device is sometimes used in such cases, although similar results can be had by representing an organization or, in appropriate cases, even by representation of a single individual on a matter where substantial issues of legal principle are involved. Moreover, a lawsuit in which a doctrinal innovation is made is likely to be followed by strategic efforts to elaborate, or expand, the new principles. For example, the principle that an involuntarily confined mental

patient is entitled to receive adequate treatment, which has far-reaching consequences for mental health delivery systems, was established in the case of an individual mental patient (*Rouse vs. Cameron*).

Subsequently, the principle was developed in class actions in other jurisdictions, and expanded to cover mentally retarded people. The right to refuse treatment was explored in other settings. Legislative, budgetary, and administrative activities also were catalyzed by the judicial decisions.

During the last third of the twentieth century there were significant shifts in public policy in the USA in critical areas such as race relations, environmental protection, and womens rights. In each of these shifts in public policy, public interest lawyers played an important role in tandem with citizen organizing and political action. When legislatures and other agencies of government were closed to citizen advocates, courts were often available. Moreover, the courts frequently, through their decisions in public interest cases, precipitated action in administrative agencies, legislatures, and the executive branch, which would otherwise not have come about.

1. Antecedents

Although the name public interest law was not coined until 1969, its antecedents existed as a set of practices for much of the twentieth century. Historically, public interest law in the USA grew out of four movements, some of which were concerned with broad societal issues, others of which were more concerned with the individual's access to basic legal representation.

The first of the precursors to modern public interest law is the decades-long campaign to redress racial discrimination and racist institutional structures. In this arena, the National Association for the Advancement of Colored People (NAACP) and its Legal Defense and Education Fund (LDEF) were the paradigm. The LDEF was established in the mid-1930s, and a brilliant group of civil rights lawyers led by Thurgood Marshall (later a Supreme Court Justice) devised a strategy to bring a succession of test cases through the federal courts to the Supreme Court that would gradually undermine the legal foundations for racial segregation. The sequence of cases led to the landmark decision in 1954 of *Brown vs. Board of Education*, which ended formal, legally mandated segregation in American public schools. In subsequent decades, a series of lawsuits followed to apply the principles of *Brown* in specific situations and to carry them forward into other areas where racial discrimination was practiced such as employment and housing.

The second pillar upon which public interest law rests is the movement for freedom of expression centered around the American Civil Liberties Union, which grew out of the National Civil Liberties

Bureau's defense of anti-war advocates during the World War I years. Third, modern-day consumer advocacy, which is central to much public interest law and is embodied in the various organizations founded by Ralph Nader, traces its lineage back to Supreme Court Justice Louis Brandeis and progressive-era anti-trust legislation. Those years, at the start of the twentieth century, saw the emergence of the idea that the American government had a right, and often a duty, to regulate the marketplace and ameliorate some of the injustices thrown up by a *laissez-faire* economic structure. Brandeis developed the idea that a 'private attorney general' could go to court to enforce the law if government failed to act.

The last is the legal aid movement that began in the late nineteenth century in New York as a way of providing charity-based legal representation for people too poor to pay private attorneys. Legal aid expanded in the post World War I period in the wake of Reginald Heber Smith's 1919 book *Justice and the Poor*. In the 1960s, as part of President Lyndon Johnson's War on Poverty, a major federally funded legal services program supplemented legal aid. With the expansion of federal anti-poverty programs in the 1960s, many lawyers within legal services developed public interest law strategies and highly specialized knowledge, litigating on such issues as welfare policy, health care for the poor, and the availability of subsidized housing. They used class actions and other litigation strategies to achieve important policy shifts.

Since the 1960s, public interest law has broadened its scope to include such areas as environmental law, an area where public interest advocacy groups have been particularly successful in building a framework of effective judicial precedents, federal statutes, and regulations. There have also been public interest law efforts addressing gender-based and sexual preference-based discrimination, and legal work on behalf of prisoners. In the 1970s, public interest groups litigated suits that delayed the construction of the Trans-Alaska oil pipeline and forced re-design; this challenged, and eventually halted, the domestic use of the pesticide DDT; and that won stricter oversight of nuclear power plants. In the following decades, public interest law centers specializing in environmental issues have developed into large membership organizations with increasingly sophisticated technical expertise, and a range of coordinated strategies involving litigation, administrative advocacy, grassroots organizing, political advocacy, and public relations.

2. 1969: The 'Public Interest Law Firm'

Sarat and Scheingold (1998) describe public interest law as using 'legal skills to further a vision of the good

society.' Most legal work is, in theory, value-neutral, and the attorney representing a fee-paying client without regard to the merits of his claim. Public interest law is, at its core, a systematic effort to bring new voices into policy formulation, and to use the law to attain certain public policy objectives, such as the abolition of legally authorized segregation or the implementation of better environmental protection. Lawyers who accept lower levels of remuneration in exchange for having the freedom to pursue social goals and to represent only clients whose causes they believe have genuine merit practice it.

Most public interest lawyers in the USA work in non-profit public interest law firms, granted tax-exempt status by the Internal Revenue Service on the condition that they are paid a salary by the firm instead of fees by their clients, that they take on cases broadly defined as having implications for the public good, and that their firm operates under a Board of Trustees. The firms' efforts are supplemented by the work of lawyers in private firms who are willing to take on a limited number of *pro-bono* cases; and by law students who receive academic credit for their work in law school clinics.

The Internal Revenue Service's recognition of the charitable status of public interest law firms has allowed foundations to channel money to these centers. Philanthropic organizations such as the Ford Foundation, the Rockefeller Brothers Fund and the Edna McConnell Clark Foundation donated millions of dollars to public interest law groups. Foundations were particularly important in providing seed money to the public interest movement in the 1970s. Subsequently, many of them became reluctant to fund public interest law activities, especially litigation, because of the controversy such activities engender. Nonetheless, many progressive foundations have continued to support regional organizations to provide funds for local and regional public interest law activities.

The first of the new generation of public interest law firms, and one representative of the genus, is the Center for Law and Social Policy (CLASP) established in 1969. It was founded by four attorneys (including the author) experienced in the distinctive legal culture of Washington, DC. They had backgrounds in corporate law practice and the executive and legislative branches of the federal government. They believed that the effective legal advocacy on behalf of corporations often subverted the mechanisms of government; that the process was impoverished by the absence of effective advocacy on behalf of citizen interests consumers, environmentalists, the poor, the victims of discrimination and others. They organized a bipartisan group of eminent bar leaders, led by Arthur Goldberg, a former Supreme Court Justice and cabinet member, to serve as a Board of Trustees for CLASP. Within a year, CLASP had won landmark victories in two major environmental lawsuits. Early CLASP

clients included Ralph Nader, the Wilderness Society and the American Public Health Association.

The CLASP model inspired other public interest law firms, and over time, spun off public interest law groups which served other constituencies among them the Mental Health Law Project (renamed the David L. Bazelon Center for Law and Mental Health) and the National Womens Law Center. It also was a co-founder of the Project on Corporate Responsibility which pioneered in the use of stockholder initiatives to promote socially significant issues, such as environmental protection, consumer interests, and civil rights, and which sponsored Campaign GM in the early 1970s, an effort to elect a slate of independent directors to the Board of General Motors.

Historically, public interest law centers have focused on issues that can be broadly defined as progressive or liberal, acting as a counterbalancing force to corporate power and serving the needs of the poor and powerless. Public interest lawyers frequently provided a counterweight to corporate advocacy. As such, it was essential that they render top-flight legal services and that they be backed up by the specialized technical expertise that was vital in many complicated cases.

Progressive public interest law thrives on both publicity and broader citizen participation. Often litigation launched on behalf of individuals or groups by public interest lawyers is one element in broad campaigns involving political lobbying, sophisticated media campaigns, and sometimes direct action.

3. The Evolution of Public Interest Law

In 1961, Rachel Carson's book *Silent Spring* was published, putting environmental pollution firmly into American public consciousness. Five years later, Ralph Nader's book *Unsafe at Any Speed* (1966) did the same for unsafe automobiles. Coming at a time of dramatic social change and political unrest, the two books helped create two new constituencies in American politics: the environmental and the consumer. In the following decades, both issues have joined civil rights, women's rights, and poverty issues at the core of public interest law. When public interest lawyers brought these issues to the courts, they were framed in a way that required decision by judges and other policy makers.

The 1970s were a decade of growth for public interest law, and by 1975 there were more than 90 public interest law firms employing over 600 lawyers. But, although the number of progressive public interest law firms continued to increase in the 1980s, and to spread beyond their traditional bases of the Northeastern USA and California, they did so at a slower rate.

The political climate during the Reagan and Bush presidencies (1981–93) was hostile to many of the constituencies represented by such groups; and their

administrations made a concerted effort to counter public interest law activity, through cutting the federal legal services program, limiting government grants to 'advocacy' organizations and appointing federal judges hostile to the interests represented by public interest lawyers.

That the public interest movement weathered these years as well as it did, and was able to enter a new period of growth in the 1990s, reflects the sophisticated fund-raising and publicity techniques it developed, as well as the institutionalization of this branch of law that had been regarded only a decade earlier as a somewhat marginal voice of conscience to the legal profession. Indeed, so successful have some Washington public interest firms become that critics argue that they have become policy 'insiders' and lost touch with the grassroots groups, which they originally set out to serve. As the courts have become less receptive to public interest litigation, public interest lawyers have turned more often to other forms to legislative advocacy, coalition building, initiatives and referenda, and public education. Litigation remains an important dimension of public interest law activity, but it is only one element in developing advocacy strategy.

3.1 The Different Forms of Public Interest Law Practice

Although the main vehicle for public interest law is the nonprofit public interest law firm in which salaried lawyers are employed full-time to handle public interest cases, other forms also exist. Many public interest law firms, such as the Natural Resources Defense Council, have become membership organizations and rely heavily on dues. Some began as affiliates of existing membership organizations, such as the Sierra Club Legal Defense Fund (later spun off and renamed the Earth Justice Legal Defense Fund). Some, like the Institute for Public Representation at Georgetown Law Center, function as clinics in law schools. Some, like the Environmental Defense Fund, integrate scientists and other experts into their staff operations.

Some private lawyers have established public interest practices, without tax exemption, and rely on court-ordered attorney's fee awards in successful cases. The private public interest law firm may also combine public interest law with traditional, financially more profitable cases on behalf of fee-paying clients. What is characteristic of the public interest law firm in its various permutations is the sense of social mission and the prominent role that lawyers play in establishing policy, even in areas where technical expertise is critically important.

Many large corporate firms encourage their lawyers to do a certain amount of 'pro bono work,' although conflict of interest considerations disqualify them from some kinds of cases. Some firms, such as Skadden, Arps in New York, have set up their own public

interest fellowship program. The Skadden Fellowship Foundation has awarded 25 such fellowships each year since 1988 to newly graduating law students who are paid to work on public interest matters for a year in connection with a public interest law firm.

Federally funded legal service offices, although facing financial cuts and stringent restrictions on their advocacy strategies (e.g., no class actions), still sometimes, though increasingly rarely, litigate public interest cases.

3.2 *New Forms of Financial Support*

Many law schools defer or cancel the tuition loans owed by graduates who choose public interest law as a career, and have also set up public interest legal foundations, which raise money among students and alumni to support public interest law work by recent graduates and student interns. The National Association of Public Interest Law, based in Washington, DC, awards student fellowships in the field, and over 120 law schools now belong to the Public Service Law Network (PSL Net), which steers students toward work in approximately 10,000 law-related public service organizations throughout the world. In the case of large membership-based groups such as the Environmental Defense Fund and the Natural Resources Defense Council money is also raised through membership dues. Government grants are sometimes awarded for specific projects. Some regulatory laws, such as the Clean Water Act, have built-in provisions that allow public interest law firms to be awarded fees in successful cases to be paid by the losing party.

3.3 *The Divergent Politics*

Traditionally, public interest law has promoted progressive causes and the rights of the powerless. But a parallel development has been the development of conservative public interest law firms with an entirely different, generally pro-business agenda. From the mid-1970s a number of aggressively conservative public interest legal centers have emerged in the USA. These law groups the Pacific Legal Foundation, the National Legal Center for the Public Interest, the National Chamber Litigation Group and others represent interests, such as timber companies, that they define as being under attack in a world of environmental protections. However, because of their link to affluent business interests through leadership and financing, it is by no means accepted that these groups meet the traditional criteria for public interest law. Other groups, which represent conservative social positions, such as opposition to abortion and affirmative action, are sometimes grouped with public interest law firms.

3.4 *The Internationalization of Public Interest Law*

Although public interest law is most widely practiced in the USA, there has been a gradual expansion into other countries. Many groups, especially in the Third World, are funded by Western foundations and non-governmental organizations. During the apartheid era in South Africa, Geoffrey Budlender toured the USA interviewing public interest lawyers on their strategies. Then, funded by the Carnegie and Ford Foundations, he returned to South Africa to participate in the founding of the Legal Resource Center (LRC). Although fighting in a system that had appropriated to itself emergency powers, the LRC litigated over the Pass Laws and issues such as internal banishment. In post-apartheid South Africa, several lawyers who practiced public interest law during the apartheid years have risen to positions of power: Arthur Chaskelson, for example, one-time chair of the LRC board is chief judge of the country's constitutional court.

In Bangladesh, the Madaripur Legal Aid Association represents poor rural Bangladeshis. Likewise, in Nepal, the Women's Legal Services Project works with abused women in a society that is overwhelmingly patriarchal. In India, lawyers from Action for Welfare and Awakening in Rural Environment (AWARE) have litigated against the oppression of lower caste bonded laborers in a culture still heavily influenced by the caste system.

Public interest law has also begun in Latin America; the Latin American Institute of Alternative Legal Services works with over 6,000 lawyers, paralegals and social workers throughout the continent, bringing to public attention issues such as environmental protection and the treatment of street children and indigenous peoples. In Eastern Europe, public interest law groups funded by financier George Soros' Open Society Institute, the German Marshall Fund, the Ford Foundation and others, have worked to solidify freedom of expression, and to counteract racism against groups such as the *Roma* (Gypsies).

Public interest lawyers are also active in the UK. They have established a Law Centres movement, and worked on immigration issues, trying to halt deportations and to secure refugee status for more asylum-seekers.

In recent years, one promising arena for public interest law has been on a transnational level: international human rights law and the attempts by labor unions to have labor protections introduced into treaties such as NAFTA and the WTO.

4. *Summary*

Public interest law is an assertion of a commitment within the justice system to fairness, open process, and diversity of viewpoints. It has significantly shaped public policy debates in the USA, even during the

Reagan and Bush presidencies. Their administrations helped to make access to the courts a less significant tool for citizen organizations. In response to these developments, many public interest lawyers developed a range of alternate strategies; working more directly in the legislative and administrative processes; filing suits in state courts; and relying more heavily on citizen actions.

The turn of the federal courts in a conservative direction was matched by increasing sophistication and resources in the advocacy effort of large corporations. No longer content to rely on the skills of legal advocates to shape public policy, corporations drew more heavily on public relations professionals. They relied more heavily on political action committees and the financing of political candidates and campaigns, and on massive advertising campaigns to shape public opinion.

Despite these challenges, public interest law in the USA continued to thrive, affecting policy decisions, and providing a measure of balance in the legal system. There are many more law school graduates seeking public interest law work than there are jobs; and there are many situations in which effective public interest advocacy, were it available, would lead to better policy outcomes. The development of public interest law in other countries holds significant promise for fair processes and broader public participation in critical policy decisions ranging from environmental protection to the protection of racial minorities. Globalization creates new needs and opportunities for public interest lawyers to function in transnational settings.

See also: Democracy; Discrimination; Equality and Inequality; Legal Aspects; Justice and Law; Justice, Access to; Legal Representation of the Poor; Law and Democracy; Law and Society; Sociolegal Studies; Law as an Instrument of Social Change; Law: Overview; Political Lawyering; Public Interest; Race and the Law

Bibliography

- Aron N 1989 *Liberty and Justice for All: Public Interest Law in the 1980s and Beyond*. Westview Press
- Carson R 1961 *Silent Spring*
- Council for Public Interest Law 1976 *Balancing the Scales of Justice: Financing Public Interest Law in America*. Council for Public Interest Law
- McClymont M, Golub S (eds.) 2000 *Many Roads to Justice: The Law Related Work of Ford Foundation Grantees Around the World*. The Ford Foundation
- Nader R 1966 *Unsafe at Any Speed*
- Sarat A, Scheingold S (eds.) 1998 *Cause Lawyering: Political Commitments and Professional Responsibilities*. Oxford University Press

Weisbrod B A et al. 1978 *Public Interest Law: An Economic and Institutional Analysis*. University of California Press, CA

C. Halpern

Public Management, New

A term coined in the late 1980s to denote a new (or renewed) stress on the importance of management and 'production engineering' in public service delivery, often linked to doctrines of economic rationalism (see Hood 1989, Pollitt 1993). The apparent emergence of a more managerial 'mood' in several (mainly but not exclusively English-speaking) countries at that time created a need for a new label. The new term was intended to denote public service reform programs that were not confined to the 'new right' in a narrow sense, but also came from labor and social-democratic parties and in that sense could be considered as part of what was later labeled a 'third way' agenda.

New Public Management is sometimes (understandably) confused with the 'New Public Administration' movement in the USA of the late 1960s and early 1970s (cf. Marini 1971). But though there may have been some common features, the central themes of the two movements were different. The main thrust of the New Public Administration movement was to bring academic public administration into line with a radical egalitarian agenda that was influential in US university campuses at that time. By contrast, the emphasis of the New Public Management movement a decade or so later was firmly managerial in the sense that it stressed the difference management could and should make to the quality and efficiency of public services. Its focus on public service production functions and operational issues contrasted with the focus on public accountability, 'model employer' public service values, 'due process,' and what happens inside public organizations in conventional public administration. That meant New Public Management doctrines tended to be opposed to egalitarian ideas of managing without managers, juridical doctrines of rigidly rule-bound administration and doctrines of self-government by public-service professionals like teachers and doctors.

However, like most divinities, the core of New Public Management is somewhat mystical in essence, despite or perhaps because of the amount that has been written about its central content. Different authors give various lists of its key traits (e.g., Hood 1989, Pollitt 1993). Some have identified different styles of public-sector managerialism over time (see Ferlie et al. 1996). How far the small-government economic-rationalist agenda that went together with more stress on public-sector management in the 1980s and 1990s was integral to those managerial ideas is

debatable (see Barzelay 2000). But it is hard to separate these elements historically, since the advent of a new generation of public-sector managerialism coincided with concern by numerous OECD governments to reduce the power of public service trade unions, increase regulatory transparency and tackle perceived inefficiencies of public enterprises. A commonly-cited view of New Public Management's central doctrinal content is Aucoin's (1990) argument that it comprised a mixture of ideas drawn from corporate management and from institutional economics or public choice. To the extent that Aucoin's characterization is accurate, it suggests New Public Management involves a marriage, if not exactly of opposites, at least of different outlooks, attitudes, and beliefs that are in tension. Savoie (1995) argues that the central doctrinal theme of public-sector managerialism is the idea of giving those at the head of public organizations more discretionary decision space in exchange for direct accountability for their actions.

Despite the label, many of the doctrines commonly associated with New Public Management are not new. Jeremy Bentham's voluminous philosophy of public administration developed in the late eighteenth and early nineteenth century provides the *locus classicus* for many supposedly contemporary ideas, including transparent management, pay for performance, public service provision by private organizations, and individual responsibility. The idea that more effective public services could be obtained by judicious application of private-sector management ideas is also a theme going back at least to the US city-manager movement of the late nineteenth century (cf. Downs and Larkey 1986). It was advanced early in the twentieth century by figures like Taylor (1916) and Demitriadi (1921). The idea that public services can be improved by giving some autonomy to managers operating at arms length from political standard-setters was often invoked in the nationalized public enterprise era. It was central to Beatrice and Sydney Webb's early twentieth-century Fabian idea of the proper way of organizing the 'social' tasks of government. Some have argued the contemporary doctrine of creating 'managerial space' in public services harks back to the US Progressive-era doctrine of a politics-administration dichotomy and independent regulators (cf. Overman 1984).

Like feminism or environmentalism, New Public Management is both a social movement and a subject of academic study. Indeed, during the 1990s, New Public Management has become a major academic industry across the world, filling bookshelves and websites with writings and conference proceedings using the term in their titles. It has its advocates and its critics, its analysts, morphologists and epistemologists, its evaluators and case-historians. Advocates stress the value to citizens and consumers to be gained by enlightened managers moving beyond what is claimed to be an outdated 'bureaucratic paradigm,'

paying more attention to how to satisfy citizen demands and to service delivery through organizations other than traditional public bureaucracies (see Barzelay and Armajani 1992, Osborne and Gaebler 1992, Jones and Thompson 1999). A sophisticated defence of public-sector managerialism is Moore's (1995) exposition of a 'strategic triangle' (of political possibility, substantive value, and administrative feasibility) within which skillful and entrepreneurial public managers can 'add value' to public services. Critics of public-sector managerialism stress the virtues of traditional Weberian bureaucracy for rule-of-law or public accountability (cf. Goodsell 1994) or see managerialism as a 'wrong problem problem' (Frederickson 1996) diverting governments' attention from hard policy choices. Some critics of New Public Management doctrines see them as too heavily based on business-school and private-sector management perspectives, and some public-choice scholars have offered 'rent-seeking' explanations of contemporary public sector reforms (see Dunleavy 1992).

Morphologists and analysts of New Public Management have several concerns. They include exploring different forms and types of public-sector managerialism, identifying how managerialism varies cross-nationally, and explaining the observed commonalities and differences (cf. Savoie 1995, Aucoin 1995). A key debate (to date inconclusive because of poor benchmarking of historical points of origin) concerns how far or in what ways contemporary public management reforms represent convergence on some global paradigm. Some stress the international commonality of public-sector management reform themes while others stress the different political motivations that prompt reforms (Cheung 1997) or argue that New Public Management represents an 'Anglo-Saxon' reform agenda that does not resonate in other contexts (Derlien 1992, Kickert 1997). A related explanatory enterprise is represented by a search to understand why management-reform ideas became popular when and where they did (cf. Hood and Jackson 1991). Epistemological debates concern the appropriate disciplinary framework for reasoning about contemporary public sector management and the rival claims of different schools or approaches to intellectual ownership of the subject, particularly in the US context of competing schools of public policy, public administration and political science (see Lynn 1996, Barzelay 2000).

Evaluators and case-historians of public-sector managerial change track the processes of reform and explore their effects. Much of this work is pitched at the level of single agencies (such as Jones and Thompson 1999), but there have been some attempts to evaluate government-wide changes (see Schick 1996). Evaluation of the effects of new managerial techniques has been patchy, but several scholars have identified 'managerial paradoxes' of one kind or another. One of those paradoxes is Maor's (1999)

observation that the development of a more managerial approach to public service produced more, not less, politicization of the senior public service in six countries. A second is Gregory's (1995) controversial claim that orthodox managerial approaches foster a 'production' approach to public services that leads to several unintended effects, including downgrading of responsibility and what he termed 'careful incompetence.' A third is the claim, redolent of Tocqueville's paradox of administrative development in post-revolutionary France, that contemporary public management may in fact involve more rather than less 'rules-based, process-driven' bureaucracy, as a result of increasing oversight and regulation and continuing stress on compliance-based rather than result-based evaluation (see Light 1993, Jones and Thompson 1999, Hood et al. 1999, Pollitt et al. 1999).

In spite of the scale and growth of the New Public Management 'industry,' or perhaps because of it, the term New Public Management has probably outlived its analytic usefulness. The term is ambiguous because the agenda of public sector reform has moved in some respects beyond the traits identified by scholars of public management in the 1990s reflecting the various cultural cross-currents that have swept through managerial debate. The term is also too crude for the fine-grained distinctions between different sorts and themes of managerialism that academic scholars need to make as the study of public services develops and the public sector reform movement becomes professionalized. Just as Eskimos are said to have many different terms to distinguish different types of snow, we need more words to describe the cultural and technical variety of contemporary managerialism. So it is not surprising that there have been numerous attempts to proclaim a move beyond New Public Management (e.g. Minogue et al. 1998). Nevertheless, in spite of its oft-proclaimed death, the term refuses to lie down and continues to be widely used by practitioners and academics alike.

See also: Administration in Organizations; Bentham, Jeremy (1748–1832); Planning, Administrative Organization of; Public Administration: Organizational Aspects; Public Administration, Politics of; Public Bureaucracies; Public Sector Organizations

Bibliography

- Aucoin P 1990 Administrative reform in public management: paradigms, principles, paradoxes and pendulums. *Governance* 3: 115–37
- Aucoin P 1995 *The New Public Management*. IRPP, Montreal, Canada
- Barzelay M 2000 *The New Public Management*. University of California Press, Berkeley, CA
- Barzelay M, Armajani B J 1992 *Breaking Through Bureaucracy*. University of California Press, Berkeley, CA
- Cheung A B L 1997 Understanding public-sector reforms: Global trends and diverse agendas. *International Review of Administrative Sciences* 63(4): 435–57
- Demitriadi S 1921 *A Reform for the Civil Service*. Cassell, London
- Derlien H-U 1992 Administrative reform in Europe—Rather comparable than comparative. *Governance* 5(3): 279–311
- Downs G W, Larkey P D 1986 *The Search for Government Efficiency*. 1st edn. Temple University Press, Philadelphia, PA
- Dunleavy P 1992 *Democracy, Bureaucracy and Public Choice*. Harvester Wheatsheaf, Aldershot, UK
- Ferlie E, Pettigrew A, Ashburner, Fitzgerald L 1996 *The New Public Management in Action*. Oxford University Press, Oxford, UK
- Frederickson G H 1996 Comparing the reinventing government movement with the new public administration. *Public Administration Review* 56(3): 263–70
- Goodsell C T 1994 *The Case for Bureaucracy*, 3rd edn. Chatham House, Chatham, NJ
- Gregory R 1995 Accountability, responsibility and corruption: Managing the 'public production process'. In: Boston J (ed.) *The State Under Contract*. Bridget Williams Books, Wellington, New Zealand
- Hood C 1989 Public administration and public policy: Intellectual challenges for the 1990s. *Australian Journal of Public Administration* 48: 346–58
- Hood C, Jackson M 1991 *Administrative Argument*. Aldershot, Dartmouth, UK
- Hood C, Scott C, James O, Jones G W, Travers A 1999 *Regulation Inside Government*. Oxford University Press, Oxford, UK
- Jones L, Thompson F 1999 *Public Management*. JAI Press, Stamford, CA
- Kickert W J 1997 Public management in the United States and Europe. In: Kickert W J M (ed.) *Public Management and Administrative Reform in Western Europe*. Edward Elgar, Cheltenham, UK, pp. 15–38
- Light P Q 1993 *Monitoring Government*. Brookings Institution, Washington, DC
- Lynn Jr. L E 1996 *Public Management as Art Science and Profession*. Chatham House, Chatham, NJ
- Marini F 1971 *Toward a New Public Administration*. Chandler, Scranton, PA
- Maor M 1999 The Paradox of Managerialism. *Public Administration Review* 59(1): 5–18
- Minogue M, Polidano C, Hulme D (eds.) 1998 *Beyond the New Public Management*. E Elgar, Cheltenham, UK
- Moore M H 1995 *Creating Public Value*. Harvard University Press, Cambridge, MA
- Osborne D, Gaebler T 1992 *Reinventing Government*. Addison-Wesley, Reading, MA
- Overman E S 1984 Public management. *Public Administration Review* 44: 275–8
- Pollitt C 1993 *Managerialism and the Public Services*, 2nd edn. Blackwell, Oxford, UK
- Pollitt C, Girre X, Lonsdale J, Mul R, Summa H, Waerness M 1999 *Performance or Compliance?* Oxford University Press, Oxford, UK
- Savoie D J 1995 *Thatcher, Reagan, Mulroney*. Pittsburgh University Press, Pittsburgh, PA
- Schick A 1996 *The Spirit of Reform*. State Services Commission and NZ Treasury, Wellington, New Zealand

Taylor F W 1916 Government efficiency. *Bulletin of the Taylor Society* 2: 7–13

C. Hood

Public Opinion: Microsociological Aspects

Although the insight that public opinion is a powerful force goes back to premodern social thought, the task of modeling how public opinion evolves through the interdependent choices of individuals was not taken up until the last quarter of the twentieth century. Through the ensuing theorizing, the term has acquired a more precise meaning than it holds in everyday language. In choice-based, microsociological theory, a group's 'public opinion' now refers to the distribution of the preferences that its members express publicly. By contrast, 'private opinion' signifies the corresponding distribution of genuine preferences. Where individuals choose to misrepresent their actual desires in public settings, these two distributions may differ.

1. Theoretical Foundations

The theoretical literature on public opinion addresses diverse phenomena that have puzzled social thinkers. They include its resistance to changes in social structures and policy outcomes; its capacity for immense movement if ever this resistance is overcome; its sensitivity to the ordering of social shocks; and its imperfect predictability.

Most variants of the theory draw on a class of constructs known as 'threshold,' 'critical mass,' 'bandwagon,' or 'cascade' models. Originally designed to explain epidemics, these were subsequently applied most influentially by Schelling (1978) and Granovetter (1978) to various social phenomena that exhibit periods of stability punctuated by sudden change, for example, stock market bubbles, bank runs, cultural fashions, corporate conservatism, and academic fads. The common feature of these phenomena is that they are shaped by interdependencies among individual decisions.

Public opinion harbors two distinct interdependencies. Wherever individuals lack reliable information of their own, they look to others for clues about the reality they are seeking to grasp. Such 'free-riding' on publicly available knowledge makes people's genuine wants interdependent. Also, in articulating preferences and conveying knowledge, individuals frequently tailor their choices to what appears socially acceptable. This conformism, which is motivated by the wish to earn acceptance and respect, creates interdependencies among the wants that individuals convey publicly.

2. The Public Opinion of a Poorly Informed Group

The essential contributions of the theory may be sketched through a model drawn from Kuran (1995), which contains additional details. Certain initial assumptions will be relaxed further on.

2.1 Thresholds and Expressive Equilibria

Consider a 10-person society that must decide a certain issue, say, the government's crime-fighting budget. Two alternatives present themselves: low (*L*) and high (*H*). Each individual will support whichever option appears socially optimal. Everyone will want this budget increased if crime seems to be rising but decreased if crime appears to be falling.

By assumption, no one has perfectly reliable private information. Consequently, every individual's judgment will depend partly, if not largely, on the apparent judgments of others. This does not mean that individual beliefs must converge. Because of differences in private information, any given common signal about the distribution of preferences may make some people favor *H* and others *L*.

For any individual, the minimum share that must consider *H* optimal for him or her to concur constitutes what is called that individual's responsiveness threshold. This threshold is a number between 0 and 100. Listing all 10 thresholds in ascending order yields society's threshold sequence (*T*). Here is an example:

T: (0, 10, 30, 30, 40, 50, 50, 60, 70, 80)

These individuals may be labeled *a*, *b*, ..., *j*, with *a* considered female, *b* male, and so on, in alternating fashion.

Given the heterogeneity reflected in this sequence, the perceived division between the supporters of *H* and those of *L* can influence dramatically the realization of public opinion. For a demonstration, suppose that initially only 10 percent of society is thought to favor *H*. Whatever the source of this perception, *a* and *b* will favor *H*, because their thresholds lie at or below 10; and the remaining eight members of the group will favor *L*, because their thresholds lie above 10. The resulting public opinion will thus lean heavily against *H*, which will enjoy the support of only 20 percent of the 10 individuals. The outcome will not impel anyone to switch positions, so it represents an expressive equilibrium. The knowledge that only two of the 10 individuals favor *H* will make exactly that share do so.

This particular expressive equilibrium is not the only possible self-sustaining outcome. If initially society was believed to be divided evenly between the two options, the first seven members of the sequence (all

except those with thresholds above 50) would have favored *H*. Observing that 70 percent favor *H*, the two individuals with thresholds of 60 and 70 would then have joined the majority, raising this support to 90 percent. Finally, completing the learning process, this second adjustment would have impelled the lone hold-out to join in. As with the above case of a divided public opinion, the resulting unanimity would have been self-reproducing.

2.2 Properties of Public Opinion

This pair of illustrations highlights several characteristics. Public opinion is path dependent, in that changing its historical starting point might alter the outcome. It is also sensitive to small events, which are intrinsically insignificant happenings that an analyst seeking a retrospective explanation would probably overlook. Whereas an initial perception of 20 percent support for *H* would be self-confirming, a perception of 30 percent would lead to a consensus in favor of *H*. The difference in these two outcomes is vastly disproportionate to the difference in their starting conditions, whose determinants would not necessarily register on a historian's viewscreen. Finally, public opinion may be unrepresentative of the available private information. Where the outcome is a consensus, it conceals the existence of substantial private information favorable to the unanimously rejected option.

2.3 Informational Cascades

The iterative process through which a self-reproducing public opinion takes shape generally is known as an informational cascade (Bikchandani et al. 1992). Under certain conditions, an informational cascade will enhance social efficiency. For example, where the first people to articulate their views are those with the most accurate information, the consequent interdependent learning will help to dampen private ignorance. However, such a fortuitous outcome cannot be guaranteed. Instead of educating the poorly informed, the informational cascade might serve to un-educate the most knowledgeable. This finding points to yet another characteristic of public opinion. It is sensitive to the order in which people express themselves.

Insofar as the evolution of public opinion influences policy decisions, it may weaken the political system's effectiveness in addressing social concerns. For example, it might induce the government to pour resources into solving a problem widely considered serious only because the pertinent social learning was spearheaded by the ignorant. Yet, the resulting allocation would not be undemocratic. Whatever the reason why public opinion has come to favor the

allocation, and however wasteful an outsider might consider the instituted policy, the members of society would only be exercising the freedom to make their own mistakes. Moreover, the government would simply be responding to the popular will.

3. Reputational Determinants of Public Opinion

Though analytically useful, the above account provides an undersocialized view of human nature. In practice, people do not express their potentially controversial views mechanically, without considering how their social standings might be affected. As social beings who wish to be liked and accepted, they may refrain from expressing beliefs or preferences that are likely to offend powerful groups. Some evidence lies in Asch's (1957) line experiment, in which subjects contradicted the evidence of their own senses to avoid isolation and criticism, and in Noelle-Neumann's (1984) pre-election surveys, in which the publicly acknowledged party preferences of respondents, though not necessarily their private sentiments, varied according to their perceptions of which party was running ahead. Additional evidence comes from Scott's (1985) ethnographic research, which finds that peasants may profess loyalty to their political regime, yet also work to undermine it in settings where they can do so with impunity.

The activists who direct society's pressure groups seek to exploit the reputational concerns that such studies highlight. They do so partly by treating people differently according to their public expressions and partly by providing incentives for criticizing, belittling, perhaps even ostracizing, overt opponents of the groups they are promoting. Insofar as such tactics are successful, individuals tailor their public positions to the prevailing social demands, and public opinion is thereby affected.

3.1 Private and Public Preferences

Depending on the nature of the information distorted to obtain social rewards and escape social punishment, such misrepresentation is called either preference falsification or knowledge falsification. On any given issue, preference falsification occurs when an individual's publicly expressed preference, or simply public preference, differs from the corresponding privately held preference, or private preference. Similarly, knowledge falsification entails a discrepancy between the communicator's private knowledge and public knowledge.

Whereas a person's private preference always reflects the relevant private knowledge, this person's public preference and public knowledge need not be

mutually consistent. In practice, however, the two choices tend to be tightly coordinated in order to make the former credible. Someone who pretends to favor one particular option, but then proceeds to give reasons for another option's superiority, will hardly come across as sincere; the person's intended audience will sense that they are concealing something.

3.2 Reputational Cascades

For further insights into how individual expressions both mold public opinion and are molded by it, reconsider the threshold sequence of Section 2.1, adding the assumption that if the 10 individuals were polled anonymously, H would be favored five-to-three, with two abstentions. Specifically, the five members with thresholds of up to 40 would support H , the two with thresholds of 50 would be indifferent, and the three with the highest thresholds would support L . These preferences expressed under the cover of anonymity are private preferences. Thresholds depend partly on private preferences, because submitting to others' demands causes psychological discomfort, albeit of a magnitude subject to variation across contexts and individuals. A second determinant of thresholds is sensitivity to social pressures. Under this section's more realistic scenario, then, a threshold reflects not only the nature and depth of an individual's private preference but also that individual's preparedness to pay the price of joining an underprivileged expressive minority. Put differently, a threshold captures an individual's tradeoff between the benefits of truthful self-expression and those of reputationally advantageous preference falsification. If person c 's threshold is 30, the sum of their expected satisfaction from all sources, including truthful self-expression and both favorable and unfavorable reputational consequences, is maximized by expressing a preference for H whenever the share of its supporters appears to be at least 30. Otherwise, their expected satisfaction is maximized by publicly favoring L .

This threshold sequence produces, as already noted, two stable equilibria that represent highly different realizations of public opinion. Under the current assumptions, the selection among these equilibria occurs through a reputational cascade. In the course of a reputational cascade, people conform to the apparently popular choice not because they see wisdom in numbers but merely to maintain a favorable reputation. In so doing they make it increasingly difficult for others to remain nonconformist.

3.3 The Possibility of Inefficiency

Like the social learning mechanism described earlier, this social adaptation mechanism driven by reputa-

tional concerns makes public opinion path dependent, unrepresentative of the available private information, and sensitive to both small events and the order of individual choices. Additionally, it may make public opinion unrepresentative of private opinion. In the illustration of Section 3.2, neither of the possible outcomes matches what an anonymous opinion survey would identify as the 'popular will.' Whereas three members privately favor it, H could achieve either unanimous support or the support of just a two-person minority. Either way, social policies designed to accommodate public opinion would be responding to signals distorted by preference falsification. As such, these policies could be inefficient.

3.4 Availability Cascades

There are contexts in which public opinion is shaped by either purely informational or purely reputational cascades. As a rule, however, the outlined informational and reputational mechanisms feed on each other, or at least interact. The composite mechanism whereby public opinion evolves through both social learning and conformism aimed at securing social acceptance is known as an availability cascade (Kuran and Sunstein 1999).

4. Feedback from Public Discourse to Private Opinion

Reputational cascades treat private preferences as fixed. For their part, informational and availability cascades recognize the plasticity of private preferences, but only during the transition to a self-sustaining public opinion. In reality, this plasticity is not limited to periods of disequilibrium. Whether public opinion is at rest or in flux, the private knowledge that undergirds private preferences is ordinarily shaped and reshaped by public discourse, which consists of the suppositions, observations, and arguments that are communicated publicly (Coleman 1990).

Preference falsification affects this private learning through two channels. As both informational and availability cascades posit, the state of public opinion, whatever its inconsistencies with private opinion, molds individual interpretations of reality. In addition, the knowledge falsification that accompanies preference falsification sows ignorance and confusion by corrupting information in the public domain through the substitution of facts that people know to be false for ones they consider true. Over time, knowledge falsification may even make once-common objections to a publicly popular option disappear from society's collective memory.

Under diverse conditions, population renewal will facilitate such an outcome. If public discourse excludes

criticisms of the fashionable option, its drawbacks will tend to be forgotten as society's composition changes through births and deaths, or simply entries and exits. This is because new cohorts will not be exposed to the unfiltered private knowledge of previous cohorts but only to the reconstructed knowledge that was considered safe to communicate. The previous cohorts may have disliked the option yet refrained from challenging it. The younger cohorts will preserve it not only because they are afraid to object but also, perhaps mainly, because the impoverished public discourse of the past has blunted their capacity to criticize the inherited social order and imagine better alternatives. Insofar as past preference falsification removes individual inclinations to want something different, current preference falsification will cease to be a source of public opinion's stability. People will support established arrangements genuinely rather than to make a good impression (Kuran 1995).

A complementary mechanism leading to intellectual impoverishment and the absence of dissent is driven by individual decisions to choose 'exit' over 'voice.' When people withdraw from a troubled society to avoid the costs of speaking up, the society's ability to institute reforms diminishes (Hirschman 1970).

The foregoing dissent-dampening mechanisms have universal relevance. Customs, norms, laws, and other institutions that keep societies poor, organizations ineffective, and inequalities unchallenged often owe their stability to the paucity of overt demands for reform. By the same token, other institutions that promote growth, efficiency, and equality are held in place through pressures that diminish both the desire and the willingness to work toward their overthrow. As Tocqueville (1945) observes, these mechanisms even contribute to the stability of traits that give nations and cultures their unique identities.

5. The Imperfect Predictability of Public Opinion

If public discourse were the only determinant of private knowledge, once a public consensus had taken shape, it would become frozen. In fact, private knowledge has other determinants, including the material consequences of prevailing policies. By sowing doubts about established structures, these other factors can induce changes in private preferences and, ultimately, also in public preferences. However, the unraveling of an established consensus need not track the escalation of private opposition to the status quo. As long as the prevailing social pressures keep preference falsification optimal for all potential dissenters, private opposition will intensify without any change in public opposition. Yet, as the undercurrent of opposition swells, a shock that affects the right people's thresholds in the right manner will stimulate

switches in certain public preferences. Even without further shocks, these switches will then encourage individuals with slightly higher thresholds to join the growing chorus of dissent. Once put in motion, then, open opposition can feed on itself, with each addition generating further additions until much of society publicly identifies with the reformist camp.

Because preference falsification had concealed the opposition brewing under the surface, this revolutionary transformation will not have been anticipated. Even so, the massive change in public opinion may easily be explained with the benefit of hindsight. The revolution's success will have lowered the personal risk of acknowledging past acts of preference falsification. It will also have created incentives for people who had supported the status quo ante sincerely to pretend that they had simply been waiting for a safe opportunity to speak out for change. Examples of largely unanticipated revolutions include the French Revolution of 1789, the Iranian Revolution of 1978, and the East European Revolutions of 1989. Each of these earth-shaking transformations entailed a massive shift in public opinion that stunned its beneficiaries, victims, and observers.

6. Methodology

A distinction of the choice-based, microsociological analysis of public opinion is that it admits circular relationships generated by feedback effects. For one thing, it links the social pressures that determine the extent to which private and public opinion diverge to the very collective process that shapes public opinion. For another, it recognizes that public opinion shapes the private preferences that are among its own determinants. By virtue of these characteristics, the microsociological approach differs methodologically from social scientific traditions in which either culture or the motives of individuals are fixed. Another salient distinction of this approach lies in its multidimensional conception of individual motives. In recognizing that individuals make tradeoffs between their expressive and reputational needs, it stands out from analytical traditions, including most variants of the public choice school, that treat individuals as invariably sincere (Downs 1957). At the opposite extreme, it stands out from self-presentation theories that view individuals as one-dimensional agents interested only in making a good impression (Goffman 1973).

Nothing in the arguments surveyed here presuppose that certain cultures promote truthfulness and others mendacity. Because people everywhere express themselves with an eye toward balancing diverse objectives, the identified social processes may be observed in any society—economically developed or advanced, democratic or undemocratic, large or small. Where public opinion shows variations across societies—as it often

does markedly—the variations do not stem from differences in the underlying basic mechanisms. Rather, they reflect historical contingencies that generated persistent differences in institutions, social pressures, and private knowledge.

See also: Collective Beliefs: Sociological Explanation; Democracy; Influence: Social; Lazarsfeld, Paul Felix (1901–76); Media Effects; Opinion Formation; Political Discourse; Political Sociology; Polling; Popular Culture; Public Opinion: Political Aspects; Public Sphere and the Media; Rational Choice Theory in Sociology; Tocqueville, Alexis de (1805–59)

Bibliography

- Asch S E 1957 Studies of independence and conformity: I. A minority of one against a unanimous majority. *Psychological Monographs* 70: 416
- Bikchandani S, Hirshleifer D, Welch I 1992 A theory of fads, fashion, custom, and cultural change as informational cascades. *Journal of Political Economy* 100: 992–1026
- Coleman J S 1990 *Foundations of Social Theory*. Belknap Press, Harvard University Press, Cambridge, MA
- Downs A 1957 *An Economic Theory of Democracy*. Harper, New York
- Goffman E 1973 *The Presentation of Self in Everyday Life*. Overlook Press, Woodstock, NY
- Granovetter M 1978 Threshold models of collective behavior. *Am. J. Soc.* 83: 1420–43
- Hirschman A O 1970 *Exit, Voice, and Loyalty: Responses to Decline in Firms, Organizations, and States*. Harvard University Press, Cambridge, MA
- Kuran T 1995 *Private Truths, Public Lies: The Social Consequences of Preference Falsification*. Harvard University Press, Cambridge, MA
- Kuran T, Sunstein C R 1999 Availability cascades and risk regulation. *Stanford Law Rev.* 51: 683–768
- Noelle-Neumann E 1984 *The Spiral of Silence: Public Opinion—Our Social Skin*. University of Chicago Press, Chicago
- Schelling T C 1978 *Micromotives and Macrobehavior*. Norton, New York
- Scott J C 1985 *Weapons of the Weak: Everyday Forms of Peasant Resistance*. Yale University Press, New Haven, CT
- de Tocqueville A 1945 *Democracy in America*. Knopf, New York

T. Kuran

Public Opinion: Political Aspects

1. Scope and Methods

Public opinion has been usefully defined by V. O. Key Jr. as ‘those opinions held by private persons which governments find it prudent to heed’ (Key 1961). Scholars of public opinion attempt to describe and

account for the politically relevant preferences and beliefs of ordinary citizens, and to assess the political impact of those preferences and beliefs. They investigate a wide range of subject matter, including broad ideologies, specific policy preferences, evaluations of political parties and candidates, assessments of self-interest and collective interest, attributions of political responsibility, and attitudes toward social groups and political institutions (Kinder 1998).

The study of public opinion draws upon and overlaps the study of political psychology, but tends to focus on collective preferences and beliefs in their broad political context rather than on individual mental processes (see *Opinion Formation*). For example, the most notable work in the field begins with a psychological account of ‘how citizens acquire information and convert it into public opinion,’ but ‘pays vastly more attention to the social sources of mass attitudes—in particular, the availability of information in elite discourse—than to the largely autonomous operation of people’s minds and psyches’ (Zaller 1992, p. 3).

Obviously, which opinions held by private persons governments will find it prudent to heed, depends significantly upon the political setting. It should not be surprising that systematic attention to the workings of public opinion has been most common in modern liberal democracies, where the preferences and beliefs of ordinary citizens are supposed to be broadly and routinely consequential, whether through elections of public officials, voting in referenda, interest group activities, or other mechanisms. By the same logic, the proliferation of new democracies in Eastern Europe, Latin America, and Asia since the 1980s may be expected to stimulate unprecedented scholarly attention to public opinion in those parts of the world.

Contemporary studies of public opinion are profoundly shaped by insights and concerns stemming from the rise of mass politics in the nineteenth century and from the intertwined evolution of modern democratic states and societies. For example, Zaller’s (1992) influential model of the dynamics of opinion change builds upon Bryce’s (1888) distinction between the ‘active class’ of political entrepreneurs and the ‘passive class’ of ordinary citizens who have ‘been told what to think, and why to think it,’ and upon Lippmann’s (1922) analysis of the role of the mass media and political elites (including government bureaucrats) in shaping ‘the pictures in our head’ that govern our reactions to complex and distant public affairs.

While long-standing insights and concerns are readily recognizable in modern scholarship on public opinion, empirical research in the field has been transformed since the 1940s by the development and proliferation of increasingly ambitious and sophisticated opinion surveys. Semistructured in-depth interviews, focus groups, and laboratory experiments continue to generate valuable insights, but their

relatively small scale and problems of external validity make them better suited to the study of individual political psychology than of collective public sentiment. By comparison, an opinion survey with several hundred randomly selected respondents can provide a reasonably accurate representation of the preferences and beliefs of millions of ordinary citizens—public opinion on a scale which governments may well find it prudent to heed.

2. *Pioneering Analysis*

The pioneering analyses of public opinion based upon survey data were conducted in connection with US presidential elections of the 1940s (Berelson et al. 1954) and 1950s (Campbell et al. 1960). Berelson et al. concluded that election campaigns mostly activate pre-existing political preferences rooted in ethnic, sectional, class, and family traditions and reinforced by social interactions with like-minded acquaintances (see *Campaigning: Political; Voting: Class*). Campbell et al. emphasized the central importance of long-standing psychological attachments to the political parties (see *Party Identification*), while also acknowledging the significance of such short-term forces as Dwight Eisenhower's personal popularity, which overcame but did not supplant the Democratic advantage in party identification dating from the New Deal era. Both of these studies portrayed public opinion as surprisingly thin and unsystematic, arguing, for example, that 'many people know the existence of few if any of the major issues of policy' (Campbell et al. 1960, p. 170) and that political preferences 'are characterized more by faith than by conviction and by wishful expectation rather than careful prediction of consequences' (Berelson et al. 1954, p. 311).

These uncomplimentary characterizations of public opinion were refined and elaborated in what remains the single most influential (and controversial) work in the field, Philip Converse's essay on 'The nature of belief systems in mass publics' (Converse 1964). Converse's analysis was based on three primary sorts of evidence. First, he showed that survey respondents answering open-ended questions only rarely made spontaneous use of systematic ideological concepts or language. Second, he showed that respondents' political views were only weakly constrained by conventional notions of 'what goes with what': liberals on one issue often took conservative positions on other issues, even within the same policy domain (see *Issue Constraint in Political Science*). Third, repeated interviews with the same respondents demonstrated a great deal of instability in their opinions, and the pattern of that instability suggested that it was caused more by random fluctuation than by systematic responses to political events. On the basis of this evidence, Converse (1964, p. 245) concluded that 'large

portions of an electorate do not have meaningful beliefs, even on issues that have formed the basis for intense political controversy among elites for substantial periods of time.'

3. *Civic Competence*

The civic competence of ordinary citizens has remained a central issue of scholarly debate in the decades since Converse's devastating analysis. One line of revisionist argument accepted the accuracy of Converse's characterization of public opinion in the USA of the 1950s, but denied its relevance to other times and places. Most notably, Nie et al. (1976) and others argued that the political upheavals of the 1960s had produced a 'new American voter' more sophisticated, more ideological, and more attuned to political issues than the one portrayed by Campbell et al. (1960) and Converse (1964) (see *Voting: Issue*). While this thesis has been subjected to vigorous methodological critiques—and has come with the passage of time to seem at least as time-bound in its own right as the portrayal it aimed to supplant—it focused attention on the potential responsiveness of public opinion to such powerful contextual factors as social change and the nature of political rhetoric.

A newer wave of revisionist scholarship has acknowledged that most citizens in most circumstances are relatively inattentive to politics and uninformed about public affairs, but has interpreted these facts as evidence of 'rational ignorance' in Anthony Downs's sense of avoiding costly information gathering in a realm in which individual investment can have only negligible collective consequences. This interpretation is nicely conveyed by the titles of three notable works of the early 1990s: *The Reasoning Voter* (Popkin 1991), *Reasoning and Choice* (Sniderman et al. 1991), and *The Rational Public* (Page and Shapiro 1992). These and other works argued that ordinary citizens make sense of the political world despite their lack of detailed information about ideologies, policies, and candidates. In some cases, the emphasis is on how citizens use 'information shortcuts' and 'gut reasoning' (Popkin 1991) to deal efficiently with political complexity. In others, the emphasis is on the tendency of statistical aggregation to iron out individual errors and idiosyncrasies in mass publics, building upon an argument first set forth by Condorcet in the eighteenth century. This 'extenuationist' literature (to use a label coined by Robert Luskin) has in turn inspired attempts to measure the extent to which information shortcuts and aggregation actually do mitigate the political effects of 'rational ignorance.' For example, Bartels (1996) estimated that individual vote choices in a series of US presidential elections fell about half way between random behavior and hypothetical 'fully informed' choices (imputed to each voter by observing

the choices of better-informed voters with similar demographic and social characteristics). Deviations from 'fully informed' behavior at the individual level were mitigated but not eliminated at the aggregate level: on average, Democrats did almost two percentage points better and incumbent Presidents did almost five percentage points better than they would have if all voters were, in fact, 'fully informed.'

Normative concerns about the quality of public opinion have also led to the development by James Fishkin and others of 'deliberative opinion polls' designed to combine the efficiency and representativeness of conventional opinion surveys with the informative and deliberative aspects of a seminar or town meeting. Rather than simply reporting their existing opinions, respondents are brought together with policy experts, public officials, moderators, and each other to consider political issues in depth. The aim is to find out—and then to publicize—'what the public *would* think, if it had a more adequate chance to think about the questions at issue' (Fishkin 1991, p. 1).

4. Political Consequences of Public Opinion

Political scientists have been more successful in characterizing and accounting for public opinion than in specifying its role in the broader political process. The path-breaking study by Miller and Stokes (1963) of representation in the US Congress provided a model for much subsequent research examining the relationship between public opinion and elite attitudes and behavior. However, the basic research design employed in these studies—cross-sectional analysis of the impact of constituents' opinions on the opinions or behavior of individual legislators—is arguably ill-suited to answering questions about the political impact of public opinion in the aggregate, especially in legislative systems marked by strong party discipline or proportional representation, or when crucial policy decisions are made by executives or bureaucrats rather than by legislatures.

An alternative research design focuses on co-variation between aggregate public opinion and policy outcomes across political units or in a single political unit over time. The analysis of 'dynamic representation' by Stimson et al. (1995) is an ambitious example of this approach, relating variations in the liberalism of America's 'public mood' over a 35-year period to subsequent variations in the behavior of Presidents, Congresses, and courts. A more specialized literature treats fluctuations in presidential popularity not only as a product of political, economic, and international events (as in the pioneering work of Mueller (1973)), but also as a factor in accounting for the varying ability of Presidents to get things done (see *Presidential Popularity*).

Efforts to specify the broad political causes and consequences of public opinion will no doubt be facilitated by the continuing accumulation of survey data over time and space. For example, the cumulative data file of the American National Election Studies project now includes more than 40,000 individual survey respondents in more than two dozen national elections spanning half a century, while the Eurobarometer series includes biannual surveys since 1973 in each of the member states of the European Union. By merging these enormous survey data sets with aggregate-level data reflecting variations in national economic conditions, electoral laws, party platforms and campaign appeals, policy initiatives, media coverage of public affairs, and many other contextual factors, analysts should be able to illuminate important connections between mass opinion and elite behavior, as well as the role of social and political institutions in shaping those connections.

On the other hand, the availability of richly detailed survey data has probably distracted scholarly attention from crucial aspects of public opinion that are not readily operationalized in surveys. For example, Converse's (1964, pp. 245–6) 'issue publics'—the various minorities of the public whose 'activated' opinions on particular issues of interest are 'expressed in the writing of letters to the editor, the changing of votes, and the like'—have seldom been subjected to detailed scrutiny. Even more notably, the elusiveness of what Key (1961) referred to as 'latent opinion'—the preferences and beliefs politicians *expect* to confront in future elections, after today's policies have had their effects and tomorrow's political opponents have had their say—has discouraged political scientists from giving it attention commensurate with its undoubted political significance. In these cases, among others, scholars will have to be more ingenious and eclectic in collecting and analyzing relevant data if they are to place the study of public opinion more squarely in the mainstream of political science.

See also: Attitude Measurement; Attitudes and Behavior; Issue Constraint in Political Science; Issue Evolution in Political Science; Party Identification; Political Advertising; Political Communication; Polling

Bibliography

- Bartels L M 1996 Uninformed votes: Information effects in presidential elections. *American Journal of Political Science* 40: 194–230
- Berelson B R, Lazarsfeld P F, McPhee W N 1954 *Voting: A Study of Opinion Formation in a Presidential Campaign*. University of Chicago Press, Chicago
- Bryce J 1888 *The American Commonwealth*. Macmillan, New York

- Campbell A, Converse P E, Miller W E, Stokes D E 1960 *The American Voter*. Wiley, New York
- Converse P E 1964 The nature of belief systems in mass publics. In: Apter D E (ed.) *Ideology and Discontent*. Free Press of Glencoe, Glencoe, IL
- Fishkin J S 1991 *Democracy and Deliberation: New Directions for Democratic Reform*. Yale University Press, New Haven, CT
- Key V O 1961 *Public Opinion and American Democracy*. Knopf, New York, 1st edn.
- Kinder D R 1998 Opinion and action in the realm of politics. In: Gilbert D, Fiske S, Lindsey G (eds.) *Handbook of Social Psychology*, 4th edn. McGraw-Hill, Boston
- Lippmann W 1922 *Public Opinion*. Macmillan, New York
- Miller W E, Stokes D E 1963 Constituency influence in Congress. *American Political Science Review* 57: 45–56
- Mueller J E 1973 *War, Presidents and Public Opinion*. Wiley, New York
- Nie N H, Verba S, Petrocik J 1976 *The New American Voter*. Harvard University Press, Cambridge, MA
- Page B I, Shapiro R Y 1992 *The Rational Public: Fifty Years of Trends in Americans' Policy Preferences*. University of Chicago Press, Chicago
- Popkin S L 1991 *The Reasoning Voter: Communication and Persuasion in Presidential Campaigns*. University of Chicago Press, Chicago
- Sniderman P M, Brody R A, Tetlock P E 1991 *Reasoning and Choice: Explorations in Political Psychology*. Cambridge University Press, New York
- Stimson J A, MacKuen M B, Erikson R S 1995 Dynamic representation. *American Political Science Review* 89: 543–65
- Zaller J R 1992 *The Nature and Origins of Mass Opinion*. Cambridge University Press, New York

L. M. Bartels

Public Policy Schools

1. What is Policy Analysis?

Before describing public policy schools, it is first necessary to describe the field of policy analysis. Although neither academics nor practitioners of policy analysis have been able to agree on a common formal definition, almost all would agree that policy analysis involves the use of formal reasoning to formulate advice for the solution of problems faced by governmental and nonprofit sectors. Of course, political leaders, decision-makers, managers, and administrators have relied on advisors over recorded time. Scholars and practitioners in many fields, particularly in the professions such as law, medicine, public health, social welfare, and urban planning, claim to incorporate policy analysis as part of their task. How then do public policy schools and their graduates differentiate themselves from the mass of advisors? They do so by using formal analytic techniques to make informed guesses about the outcomes of alternative

choices, as they would be implemented on the ground in the real world.

The modern policy analysis movement had its genesis in the United States and has spread very slowly, if at all, to other countries. For that reason, the history of public policy as it evolved into an academic curriculum is a history of the application of policy analysis in the United States.

The staffing effort of the Second World War and the subsequent Cold War with the Soviet Union brought the policy analysis movement into existence. Although many historians of the policy analysis movement point to Daniel Lerner and Harold Lasswell's call for a policy sciences field that would apply reasoned advice to all stages of the policy process (deLeon and Steelman 1999, Lerner and Lasswell 1951), in practice the field developed with the application of what might be called the optimization sciences—particularly micro-economic analysis and the techniques of operations research—to policy choices. These techniques were first used in this way during the Second World War when the need for war mobilization led to a planned economy significantly replacing the market economy. After the war the new techniques found a home at a newly formed home for defense analysis (the RAND Corporation). Ultimately these techniques, and many defense analysts, found a home in offices or ministers of defense and then diffused throughout government. The analytic offices that were set up became an established source of policy advice using the techniques of benefit-cost analysis, cost-effectiveness analysis, operations research, and policy evaluation. These techniques became the heart of the new policy analysis curriculum.

Since the 1970s in the United States, public policy schools have become major centers for graduate education for service in the public and nonprofit sectors in the United States. At many prestigious US universities they have replaced schools and programs of public administration. Their number has grown from under 10 in 1970 to over 60 in 2000. Many existing public administration programs, moreover, have incorporated parts of the public policy curricula, and in some cases have adopted public policy concentrations.

The growth in public policy schools and programs has been associated with the creation of an academic field complete with an association—the Association for Public Policy Analysis and Management (APPAM)—and a journal—the *Journal of Policy Analysis and Management* (JPAM). Of equal if not greater importance has been the spread of the intellectual approach of policy analysis as graduates of policy schools have replaced traditional public administrators in high-level staff positions at the US federal, state, and local governments.

As graduates of policy programs have replaced students of accounting and public administration in various governmental staff agencies—particularly

budget offices—the policy analysis movement and approach have been institutionalized within the governmental service. This institutionalization was reinforced by the slow economic growth and resulting budgetary constraints of the 1970s and 1980s which forced decision-makers to continually face tighter constraints and tradeoffs. New institutions, such as the Congressional Budget Office of the US Congress, were largely staffed by those with policy analysis training obtained either in schools of public policy or Ph.D. programs in economics.

These new staff organizations reflected the norms and values of their dominant staff professions—the graduates of economics departments and policy analysis schools and programs. Chief among these was a belief that, all things being equal, markets are the best way to allocate society's goods and services. That a decision to use other allocation mechanisms was justified only on grounds of equity or rights (the distribution of benefits and costs and/or the existence of some activities that society believed should not be freely exchanged), or when one could show that markets had failed and that governmental intervention would (not could) lead to a more economically efficient allocation than was the case under the failed market. This reliance on markets separated the new policy analysis movement from the traditional planning approach and built within the policy analysis framework a conservative bias when it came to justification of public sector activity.

The centrality of the norms of economics to the new policy analysis movement led to a belief that analysis of policy requires the laying out of the tradeoffs inevitable in the creation and implementation of policy. These tradeoffs are reflected in the centrality of two ideas of economics—opportunity costs and the margin. The policy analyst relies on opportunity cost to inform the decision-maker that no option is free, that one can always use the resources for other activities. The concept of the margin allows the economist or policy analyst to tell the decision-maker what she/he can obtain from adding or subtracting one unit of resources. The use of these concepts separates the economist and policy analyst from the political analyst, who often seeks to suppress tradeoffs so as to increase the membership of political coalitions.

A third effect was a conscious attempt to present decision makers with analyses of policy options that make choices concrete by clearly defining problems, setting out possible solutions, and specifying the criteria to be used to judge those solutions so that those decision-makers can evaluate the solutions in a comparative, predictable manner. Although students of policy analysis often wrote about qualitative as well as quantitative indicators, in practice practitioners of the new field found that their comparative advantage vs. other advisors was in the use and presentation of quantitative measures—particularly those in a benefit/cost format.

The reliance on analytic techniques and on policy options separated economists and policy analysts from traditional American public administrators. Since the progressive era and the early academic writings of Woodrow Wilson on public administration (Wilson 1887), students of public administration had operated under the belief that it was possible to separate the creation of policy from its efficient administration. Although by the 1960s most public administrators knew that this dichotomy between politics and policy on the one hand and administration on the other was a myth, much, if not most, of the public administration approach still centered on the efficient running of bureaus and departments. Consequently, to a great extent, the educational core of masters in public administration programs reflected this approach with required courses on the functional skills needed by a line manager—budgeting, organizational behavior, personnel management, and the role of an unelected civil servant in representative government.

2. What Are Public Policy Schools and How Do They Differ From Other Programs?

The first public policy schools and programs were created in the US in the late 1960s. With considerable philanthropic support, 1971 had founded nine schools or programs in the United States. In a pattern that has continued to this day, several existing programs quickly adopted the new approach and curricula of the policy school movement.

The advocates and sponsors of these new schools and programs sought to apply the new optimizing techniques of policy analysis to the problems facing nations in the late 1960s—particularly the problems facing urban areas. At this time it was still widely believed that government could solve social problems if the right policies were found, created, adopted, and implemented. Thus, the initial educational effort sought to support positive public sector activity to solve problems, a view that has much less support today. The creators of the new programs also sought to do for education towards public service what the changes in business administration education had done for business administration in the mid-1950s—replace a curriculum centered on institutional description and anecdotes with one grounded on analytics.

2.1 Similarities Across the Schools

Although one can point to many differences among the original nine policy schools and programs—one had a heavy emphasis on undergraduate education, a second only offered a Ph.D.—viewed from afar their similarities far outweighed their differences. With the exception of the one that only offered a Ph.D., all

focused their graduate education efforts on a Masters degree that was viewed as a terminal professional degree rather than a step toward a doctorate. In most cases they labeled this primary degree the masters in public policy (MPP).

2.2 Concentration on Analytics and Professionalism

The basic analytic public policy core set of courses has remained quite similar since the founding of the nine original schools. It centers on a combination of three yearlong sequences of analytic skill offerings—in economics, quantitative methods, and political and organizational analysis—one or more workshops in which students undertake policy analyses for real, as against hypothetical, clients, and a required summer internship. Some programs also required courses in individual and/or policy ethics. Others required courses in legal analysis. Initially the economics sequence centered on microeconomics, although several of the programs required a microeconomics–macroeconomics sequence. Although centered on courses on descriptive and inferential statistics, many of the initial required curricula also included material drawn from decision analysis, game theory, and operations research.

The political and organizational analysis sequences were included so that students trained in technical skills would also graduate from the programs with the political and managerial skills needed to enact, implement, and administer policies. Of all the required sequences in the policy analysis curriculum, political and organizational analysis has remained the most problematical. From their inception public policy programs rejected the public administration separation of politics from administration. Instead it was assumed that to be an effective analyst one had to be an effective political actor and manager. But unlike economics with its instrumental approach, the academic discipline of political science was largely descriptive and as such was less useful for a professional curriculum. This posed and still poses a great challenge to the public policy faculty charged with the political and managerial core of the policy curriculum. Attempts have been made at a number of institutions to take the findings of political science and public administration research and, through the use of decision forcing case studies, to build a prescriptive and strategic approach to management. Students were challenged to go from answering the descriptive question, ‘What are the political forces or organizational tendencies which affect the kind of policy we can formulate?’ to the more instrumental question, ‘What do I, as analyst or manager, need to do, as I shape my policy prescriptions or implement the chosen policy, to ensure that the political forces and organizational tendencies do not succeed in frustrating

or changing desirable policy in undesirable ways?’ (Fleishman 1990, pp. 743–4).

Many observers have felt that this attempt has been only partially successful, however. To overcome this deficiency some scholars have focused on identifying examples of ‘best practice’ that might form the basis of improved political and organizational behavior (Bardach 1994, 2000). Others, criticizing the lack of scientific rigor in this approach, have urged their colleagues to create a body of knowledge in which statements of appropriate political and organizational actions are subject to formal hypothesis testing (Lynn 1994).

The professional nature of public policy programs is reflected in the fact that, unlike the traditional pattern in which professional masters students simply take graduate courses offered by the economics and political science departments, the required policy analysis courses were and are especially designed for a professional program. The goal in most cases is to recast traditional material in a policy centered, problem solving framework and to present the material in an instrumental rather than descriptive manner. As indicated, political science approaches are largely descriptive, and the new programs, therefore, had to take this material and turn it on its head by presenting it in a fashion that would allow future practitioners to use it as a lever to achieve policy and managerial goals—what might be called political and managerial engineering.

One constant across nearly all the original courses was the focus on domestic policy. Over time international material slowly entered the curricula, but this occurred largely through applications of the analytic skills of the curricula (such as benefit/cost analysis) to international situations, such as whether a water project should be built in another country. In recent years several policy schools have moved into the area of economic development, but almost all have avoided the traditional subjects of international relations and country analysis.

2.3 Differences Between Public Policy and Public Administration Programs

The new public policy programs sought to replace the existing public administration educational model of public administration. In addition to their focus on analytic skills, the new programs differed from their public administration counterparts in their large required set of core courses, the absence of required concentrations, and the greater number of economists and statisticians on the faculty.

Existing public administration programs had a small set of core required programs (four to five courses) and lasted a year to a year and a half. The new public policy programs were two-year programs with

a large number of required core courses (at least a full year's worth at a minimum and in some cases up to two-thirds of a standard two-year course load). In most cases students were required to take these courses in a set sequence. The average public administration program (MPA) offered many specialties or concentrations (with a mean of five in 1984). (Ellwood 1984). Initially at least, public policy masters (MPP) programs almost totally focused on the analytic techniques taught in their large cores. In 1971, in the United States, the average MPA program faculty member's doctorate was either in political science or public administration (21 and 20 percent in 1971). Economists represented only 9 percent. In the new public policy schools and programs economists represented a plurality (31 percent in 1971) and in some cases a majority of the faculty with political scientists being the second most represented discipline (23 percent) (Ellwood 1984).

3. Changes and Challenges to Public Policy Schools

Masters programs in public policy currently dominate graduate education for public and nonprofit service in the United States. Graduates from the programs are now found in key decision-making positions throughout government and the nonprofit sector. However, recent developments provide a challenge to the future of the approach and the field.

3.1 The Decline in the Belief that the Public Sector Can Solve Problems

The public policy school initiative was fostered by the belief that better public policies could lead to the solution of public problems—particularly those of the poor and of urban areas. Over the past three decades this belief has attenuated if not disappeared in the United States and elsewhere. To paraphrase Ronald Reagan, in the United States government is now seen as the problem, not the solution. Public service itself has become less attractive. The private sector increasingly has come to be seen as the creator and locus of heroes. Private sector remuneration has grown at a far faster pace than public or nonprofit sector pay.

The effect of all these changes has been to draw graduates of public policy programs away from careers in the public and nonprofit sectors toward jobs in the private sector. One recent study found that between 1973 and 1993, in the United States, the percentage of MPP graduates taking jobs in the public sector declined from 76 to 49 percent (Light 1999). At the largest US MPP program this percentage dropped to 28 percent for one year in the late 1990s before rebounding to the mid-40 percent range. This trend

poses this question, will policy programs continue to attract the 'best and the brightest' as they have over the past three decades?

3.2 The Difficulty of Management and Leadership Education

As indicated above, from the beginning of the policy school movement it was understood that high technical skill was but a necessary condition for an effective policy analyst. The sufficient condition was the acquisition of political, leadership, and managerial skills. The inclusion of a required sequence of political analysis and management in most programs was meant to impart these skills. But this effort has always been the most problematic endeavor of public policy education. Academics simply know less and have a more difficult time imparting what they know to practitioners in these areas than they do when it comes to technical skills. Moreover, from the beginning, management skills were grafted on to a curriculum focusing on the skills needed by staff analysts. In contrast to masters of business administration (MBA) programs in business schools, where the entire curriculum (or at least the first year of required courses) was seen as management education, policy schools sought to condense courses in accounting, finance, marketing, and operations into one or two required offerings. The result has never been totally satisfying.

3.3 The Tradeoff Between Breadth and Depth

The original public policy curricula focused on imparting the skills that would be needed by policy analysts who, acting as advisors to decision-makers, would create policies that would hopefully solve society's problems. Over time policy schools have moved away from this narrow focus on the training of staff to a broader focus of training future managers and leaders. Moreover, the initial narrow focus on domestic policy has also broadened. Today the policy schools seek to serve the needs of students with a wide set of career goals. Some want to become elective officials, some political activists, some social activists, some career civil servants, some managers of nonprofit organizations, and some private sector consultants. The percentage seeking careers in technical policy staff positions has become smaller and smaller. As the programs have tried to meet the needs of this increasingly diverse set of students, they have to some extent lost the coherence and depth of the required set of skills that made up the core of their curricula.

As they have tried to serve future elected officials, political activists, public and nonprofit managers, as well as policy analysts, some policy schools have dropped such topics as decision analysis and oper-

ations research from their core set of requirements. Others have moved to the creation of tracks or concentrations, with some providing more quantitative, analytic rigor. Some have approved these changes as responding to the needs of a new set of students while others fear that public policy by trying to be relevant to all will lose its distinctive purpose.

3.4 The Intellectual Challenge

Professional education always faces the challenge that it is simply an applied or less rigorous version of the material offered by the traditional academic professions. Frequently this charge is misplaced. The need to apply theory to actual situations often leads to new breakthroughs—sometimes theoretical breakthroughs. Such was the case with the creation of the modern theory of finance in business schools.

It was hoped that implementation theory or principal-agent theory would be the public policy equivalents of the business schools' finance theory. But this has not occurred. Implementation theory has never got beyond its inductive beginnings and principle-agent theory is limited when applied to the multiple-goal, multiple-agent world of the public sector (Moe 1984). This has meant that public policy education has remained derivative and applied. For the long-term health of the public policy movement within the academy, overcoming this weakness is perhaps the greatest challenge to the field.

See also: Policy Knowledge: Advocacy Organizations; Policy Knowledge: Epistemic Communities; Policy Knowledge: Nonprofit Institutions; Policy Knowledge: Universities

Bibliography

- Bardach E 1994 Comment: the problem of best practice research. *Journal of Policy Analysis and Management* 13(2): 260–8
- Bardach E 2000 *A Practical Guide for Policy Analysis: the Eightfold Path to More Effective Problem Solving*, Part III, Chatham House Publishers, New York
- deLeon P, Steelman T A 1999 The once and future public policy program. *Policy Currents: Newsletter of the Public Policy Section of the American Political Science Association* 9(2): 1–9
- Ellwood J W 1984 A Morphology of Graduate Education for Public Service in the United States. Report for the Study of Trends and Innovations in Graduate Education for Public Service. National Association of Schools of Public Affairs and Administration
- Fleishman J L 1990 A new framework for integration: policy analysis and public management. *American Behavioral Scientist* 33(6): 733–54
- Lerner D, Lasswell H D 1951 *The Policy Sciences*. Stanford University Press, Palo Alto, CA

Light P C 1999 *The New Public Service*. The Brookings Institution Press, Washington, DC

Lynn L E 1994 Public management research: the triumph of art over science. *Journal of Policy Analysis and Management* 13(2): 231–59

Moe T M 1984 The new economics of organization. *American Journal of Political Science* 28(4): 739–77

Wilson W 1887 The study of administration. *Political Science Quarterly* 2: 197–222

J. W. Ellwood and E. Smolensky

Public Psychiatry and Public Mental Health Systems

Public psychiatry in the US encompasses several different elements: buildings, such as state mental hospitals; programs, such as community mental health centers; and insurance mechanisms, such as Medicaid and Medicare, that enable poor or elderly recipients (respectively) to receive care in private facilities. Public mental health systems are found at all levels of government: the federal government runs one of the largest health care systems in the world, the Veterans' Administration Medical System, all individual states run state psychiatric hospitals and systems, and local governments (e.g., cities and counties) run hospitals and clinics. In this article, both public psychiatry and public mental health systems will be discussed simultaneously, starting with their historical context and moving on to a more detailed examination of their elements.

1. The History of Public Psychiatry and Public Psychiatric Systems

For many years after the US was colonized, there were no services for the mentally ill. In time, local communities built jails, workhouses, and almshouses or poorhouses, intended for 'rogues and vagabonds,' the 'idle and disorderly,' and the needy, respectively, to which mentally ill persons were sent if violent, disruptive, or simply wandering. While the first facility for the mentally ill, the 'Publick Hospital' in Williamsburg, Virginia, opened in 1752, it was not until the 1840s that Dorothea Dix, a pioneering advocate for the mentally ill, sought higher governmental responsibility for this population, and as a consequence, state mental hospitals and systems flourished (Talbot 1978). The federal government entered the scene in the eighteenth century with the creation first of Public Health Service Hospitals for the Merchant Marine,

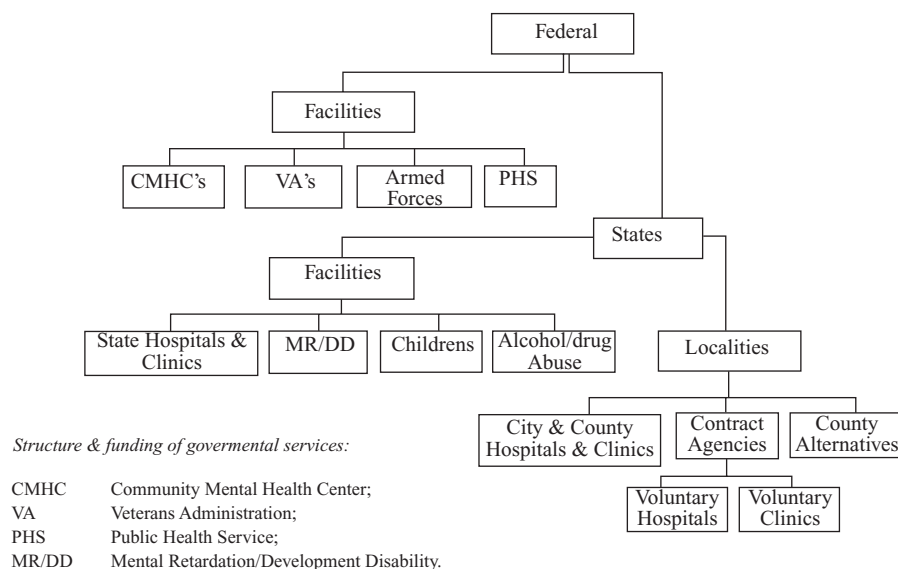


Figure 1
Structure and funding of governmental services

then Veterans Hospitals and finally, in 1963, Community Mental Health Centers. In some ways, the federal government's commitment to returning the treatment and care of the mentally ill back to their communities represented a coming full-circle from the jails, workhouses, and almshouses or poorhouses of colonial days, but now this responsibility was coupled with the technology, programs, and staffing necessary to provide effective care nearer to home, family, and work. Thus, by the turn of the twentieth century there was a rich variety but often confusing multiplicity of mental health services sponsored by different governmental levels and agencies extant in this country (Fig. 1).

2. City and County Care

As mentioned above, the very first public institutions in which the mentally ill were cared, were county or city jails, workhouses, and almshouses or poorhouses. While these 'houses' existed for a number of decades (Rothman 1971), in the 1880s the mentally ill began to be cared for and treated in separate psychiatric hospitals, established by individual states. However, to this day, some cities (e.g., New York, New Orleans, San Francisco) and some counties (e.g., Cook (Chicago), Milwaukee, San Diego) have city and county general and/or psychiatric facilities. In addition, as also was mentioned above, the federal legislation in the 1960s creating community mental health centers plus the devolution of the state's single authority to

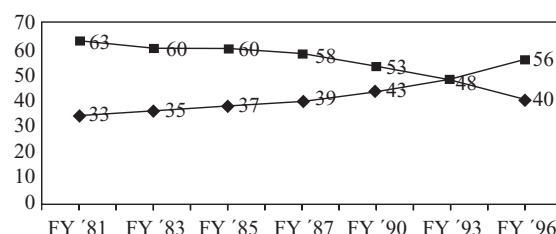


Figure 2
State Mental Health Agency controlled expenditures for state mental hospital inpatient (◆) and community-based (■) services: 1981–1996

counties in the 1980s (Hogan 1999) have ensured a prominent role for local entities in taking public responsibility for the mentally ill at the close of the twentieth century. In terms of numbers, at their height in 1955, city and county mental health facilities cared for only 100,000 persons, compared to the 460,000 persons residing in state hospitals. Today there are fewer than a dozen city or county facilities remaining, caring for 10,000 patients. Recently, the Robert Wood Johnson Foundation initiated a project in nine cities to 'devolve' mental health services from state to city 'quasi-public authorities' (Goldman et al. 1990). Because of these and other initiatives to move from (state) institutional to (local) community care, the majority of moneys allocated by each state now goes to community-directed mental health care (Fig. 2).

3. State Mental Health Services

As mentioned above, the first facility for the mentally ill, the 'Publick Hospital' in Williamsburg, Virginia, opened in 1752. Called 'public(k),' the hospital was actually private and it was not until the 1820s and 1830s that state governmentally supported and run psychiatric hospitals came into being. While Massachusetts and New York pioneered in opening new facilities to accommodate for the need for treatment and care of the mentally ill, other states lagged behind until the 1840s when Dorothea Dix, a Sunday-school teacher, sought first federal and then state initiatives to care for all the mentally ill in America. From this point to 1955, state mental hospital systems grew; first slowly, and then with the waves of new immigrants, more quickly (Talbot 1978). By 1955, about 460,000 Americans were housed in such facilities (Fig. 3).

The role of the state mental hospital throughout its first century was comprehensive and inclusive; it provided almost everyone almost all the care and treatment then imaginable and available, including housing, food, clothing, supervision, social services, medical, psychiatric and dental services, and leisure time activities. If the number of admissions grew, new hospitals were built. Alternatives to state hospitals began to appear in the US starting in 1855 beginning with the Farm of Sainte Anne's in Illinois, and eventually encompassed cottages on the grounds, 'boarding out,' 'aftercare,' or outpatient clinics, traveling clinics, crisis intervention, satellite clinics, day, night, and weekend hospitals, social and vocational rehabilitation, home care, and halfway houses. However, the creation of all these 'alternatives' did not reverse or halt the increasing population treated in

state facilities—until 1955. In that year, deinstitutionalization of these large facilities began as the result of a combination of technological (a new, the first, antipsychotic medication) and philosophical (that care in the 'community,' nearer family, work, and friends, was better than that in 'asylums') initiatives, soon to be joined by legal (the rights to treatment, to refuse treatment, and to be treated in less restrictive settings) and economic (medical insurance via Medicaid and Medicare for treatment in the community and income support via Supplemental Security Income for housing in the community) forces. In the next 30 years, the population in state hospitals fell 80 percent. Thus, at present, there are fewer than 90,000 residents in state facilities. It must be noted, however, that state governments were not only responsible for state mental hospitals (Fig. 1), but for 'aftercare' and outpatient clinics, as well as facilities and clinics for children and persons suffering from alcohol and drug abuse and mental retardation/developmental disorders, and these services have continued to this day.

4. Federal Mental Health Services

4.1 Veterans' Hospitals

While pensions, cash payments, and land grants were awarded to veterans of the Revolutionary War by grateful states and the federal government, it was not until the American Civil War (1861–1865) that national veterans' homes were established (Godleski et al. 1999) and not until after World War I, in the 1920s, that health benefits were offered to all veterans.

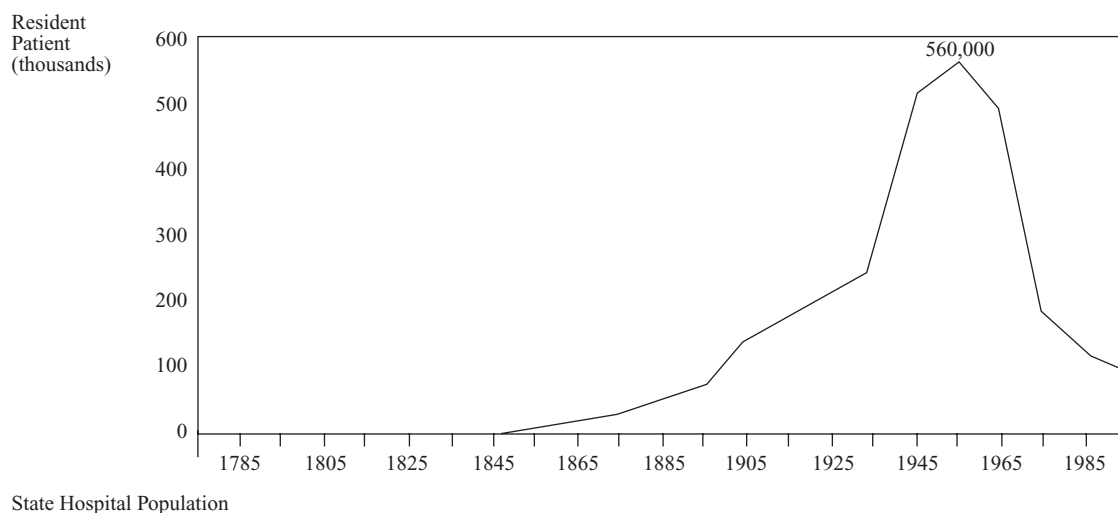


Figure 3
State hospital population

By 1930, 4.7 million veterans were being served by 54 hospitals and in 1995 the 'Veterans Administration' system of care included 172 medical centers, 375 outpatient clinics, 133 nursing homes, 39 'domiciliary' care facilities, and 202 'Vet Centers' (largely urban drop-in centers).

4.2 Military Psychiatric Services

Physicians have always accompanied American forces during war. The recognition of the role of stress in combat was first documented in America during the Civil War when 'combat heart' was often diagnosed. In World War I, psychiatrist Thomas Salmon visited the battle lines in France before America's entry and it was his observation of the difference in the practices and outcomes of the French (treatment near the lines, expectation of recovery; with low resultant disability) and English (evacuation; with high disability) in treating 'shell-shock' victims that predicated the American military's practice henceforth. In World War II and in Korea, these lessons had to be re-learned but by Vietnam, American psychiatrists were inducted into the Armed Forces in significant numbers and assigned to both (a) hospitals in combat and evacuation areas and (b) line troop divisions, where they practiced 'preventive psychiatry' (Talbot 1966). Currently, to serve a 'peacetime military' of almost 1 million men and women, there are approximately 300 psychiatrists on active duty.

4.3 Public Health Service Hospitals

In 1798, Congress established the US Marine Hospital Service—the predecessor of today's US Public Health Service—to provide health care to sick and injured merchant seamen (History of the Office of the Surgeon General 1999). Eventually, 25 hospitals were established across the country (Mullen 1989). In 1981 these hospitals began being phased out and at present the Surgeon General's Office and the Public Health Service, while active in articulating policy (on smoking, the treatment of schizophrenia, etc.) and employing 6,000 persons who serve as a 'uniformed cadre of health professionals on call 24 hours a day, 7 days a week in the event of a public health emergency,' does not run nonresearch related clinical services.

4.4 The National Institutes, Administrations, and Centers

The first element in the federal array of research laboratories and agencies to be established was the Laboratory of Hygiene in 1887; the first psychiatric

component was the Division of Mental Hygiene of the US Public Health Service in the 1930s. The National Cancer Institute was the first of the now 25 national institutes now comprising the National Institutes of Health (NIH) to be established for research; the National Institute of Mental Health (NIMH) coming into being in 1946. The research mission was broadened in the 1950s and 1960s to include teaching and services (e.g., community mental health centers) as well as alcohol and substance abuse treatment and prevention. Soon there were Institutes on Drug Abuse (NIDA) and Alcohol Abuse (NIAA) and in 1968 the Health Services and Mental Health Administration was established as an umbrella organization to house all of them. In 1973, NIMH rejoined the NIH and later came under a new 'umbrella,' the Alcohol, Drug Abuse and Mental Health Administration (ADAMHA). Finally, in 1992, prevention, services, and services research were separated off into a new agency—the Substance Abuse and Mental Health Services Administration (SAMSHA), while the NIMH, NIDA, and NIAA remained in the NIH as 'research institutes.'

4.5 The Community Mental Health Center

As mentioned above, since colonial days, localities had always provided some services. But, in 1963, the shift from localities to states to federal auspices came full circle, and it was the federal government that legislated the establishment of what was intended originally by President John F. Kennedy's administration to be about 750 community mental health centers (CMHCs), each of which was to serve 75,000–150,000 persons through the provision of 10 essential services (inpatient, outpatient, partial hospital, emergency, consultation and education, diagnostic, rehabilitative, precare and aftercare, training and research, and evaluation) (Glasscote et al. 1969). While the 1963 legislation had to be amended in 1975 to include care for children and the elderly and those suffering from alcohol and drug abuse as well as transitional housing, follow-up care, and screening for courts and agencies prior to hospitalization, the model of the CMHC was a powerful one in the US for most of the last half of the twentieth century. At present, while the US cannot claim to have either comprehensive or universally available, federally initiated community mental health centers, there are 675 in the nation (Ray and Finley 1994).

4.6 Federal Entitlements for Community Care

The blind and needy aged have been provided federal 'Social Security' to enable them to make ends meet

since 1935, and in 1950 persons suffering from other disabilities, including mental disorders, were included under two additional programs: Supplemental Security Income (SSI) and Social Security Disability Income (SSDI). In 1966, two other large 'public' populations, the poor and the aged and disabled were guaranteed medical insurance through two federal entitlement programs, Medicaid and Medicare, respectively. What no one had realized at the time was the fact that with these four massive federal entitlement programs, the mentally ill could 'afford' to live and receive treatment from 'private' clinics and general hospitals rather than 'public' state hospitals. This further added to the technological, philosophical, and legal forces moving psychiatric care and treatment in America from institutional to community care; albeit publicly financed. In 1997, 13.4 percent (37,514 million) of the population was covered by Medicare and 13 percent (35,028 million) by Medicaid; by 1994, 16.5 percent (42,878 million) of the population was covered by Social Security, 10 percent (26,281 million) by SSI, and 1.5 percent (5,893 million) by SSDI (US Bureau of the Census 1998).

4.7 Community Support and Case Management

Starting in 1977–1978, the NIMH provided grants to community agencies to initiate community support programs. Such a system was to include: identification of the population, assistance in applying for entitlements, crisis-stabilization, psychosocial rehabilitation, supportive services, medical and mental health care, back-up support to families, involvement of concerned community members, protection of client rights, and case-management (Shaftstein et al. 1978). Like the CMHC, this model was compelling and moved community services to adopt such strategies and services even in the absence of federal moneys. Most widely undertaken have been case management programs for high-intensity and average utilizers. It is impossible to estimate how many community support and case management systems there are in America today, but it is in the tens of thousands.

4.8 The Indian Health Service

Members of Federally recognized Indian tribes and their descendants have been eligible for services provided by the Indian Health Service (IHS) since 1787. The IHS, part of the US Public Health Service, Department of Health and Human Services, operates a comprehensive health service delivery system for approximately 1.5 million of the nation's two million American Indians and Alaska Natives and in 1999 its annual appropriation was approximately \$2.2 billion.

There are more than 550 federally recognized tribes in the US whose members live mainly on reservations and in rural communities in 34 states, mostly in the western US and Alaska. IHS services are provided in three ways; by the IHS directly, through tribally contracted and operated health programs, and by purchasing care from more than 2,000 private providers. As of March 1996, the federal system consisted of 37 hospitals, 64 health centers, 50 health stations, and five school health centers. In addition, 34 urban Indian health projects provide a variety of health and referral services. The IHS clinical staff consists of approximately 840 physicians (approximately 80 of whom are psychiatrists), 100 physician assistants, and 2,580 nurses. As of March 1996, American Indian tribes and Alaska Native Corporations administered 12 hospitals, 116 health centers, three school health centers, 56 health stations, and 167 Alaska village clinics.

5. The Correctional System

The correctional system in America cuts across all governmental levels: federal, state, and local (city/county). It may well house more mentally ill persons than any other 'governmental health system.' There are about 10 million adults booked into US jails each year and more than 1.5 million persons currently incarcerated. The US Bureau of Justice estimates a rate of mental illness in local and state jails of over 16 percent and in federal prisons of over 7 percent; thus over 250,000 mentally ill persons are in prison and jail on any one day (Ditton 1999). It is not known how many psychiatrists are employed in the correctional system nor how many services are rendered.

6. Public Managed Care

In the 1930s the Kaiser Corporation, a large industrial concern, formed the first employer-run group practices in America to reach its employees working in rural areas. These tightly run group practices became known as health maintenance organizations (HMOs) and in the 1970s were promoted by the Nixon Administration. HMOs, along with other forms of tightly managed forms of service provision, have grown and expanded from their original base serving employed persons through industry or 'private' insurance to serving 'public' populations, e.g., the aged, disabled, and poor who are insured by Medicare and Medicaid. The popularity of such 'managed care' as a method of cost-control and provision is apparent from the fact that, in 1992, the percentage of Americans covered by managed-care programs exceeded those covered by 'traditional' (e.g., Blue Cross/Blue Shield) insurance

plans. In the 1990s, the federal and state governments have actively sought to enroll their publicly funded populations (e.g., those receiving Medicaid and Medicare) in such programs. Estimates of the percentage of Medicare recipients enrolled in managed care programs are at present over 10 percent nationally; regarding Medicaid, at present every state in the union has begun a 'managed Medicaid program,' has applied for federal government permission to do so, or is actively discussing it.

7. The Future

The future of public psychiatry and public mental health systems is unknown. While some systems (e.g., county and state systems) have endured for almost the entire three centuries of our young country's existence, others (such as public managed care) are growing at a rate that will surely diminish the importance and even existence of these traditional systems. What is known, however, is that until poverty and severe mental illness no longer exist there will be public psychiatrists and public mental health systems.

See also: Mental Health: Community Interventions; Mental Health Programs: Children and Adolescents

Bibliography

- Ditton P M 1999 Mental health and treatment of inmates and probationers. *Bureau of Justice Statistics*, NCJ 174463
- Glasscote R M, Sussex J N, Cumming E, Smith L H 1969 *Community Mental Health Center: An Interim Appraisal*. Joint Information Service of the American Psychiatric Association and the National Association for Mental Health, Washington, DC
- Godleski L S, Vadnal R, Tasman A 2001 Psychiatric services in the veterans' health administration. In: Talbott J A, Hales R E (eds.) *The Textbook of Administrative Psychiatry: New Concepts for a Changing Behavioral Health System*. American Psychiatric Press, Washington, DC
- Goldman H H, Morrissey J P, Ridgely M S 1990 Form and functions of mental health authorities at RWJ Foundation program sites: Preliminary observations. *Hospital and Community Psychiatry* **41**: 1222–30
- History of the Office of the Surgeon General 1999 <http://www.surgeongeneral.gov/osg/sghist.htm>
- Hogan M F 2001 State and county agencies. In: Talbott J A, Hales R E (eds.) *The Textbook of Administrative Psychiatry: New Concepts for a Changing Behavioral Health System*. American Psychiatric Press, Washington, DC
- Indian Health Service Internet Home Page 1999 <http://www.ihs.gov>
- Mullen F 1989 *Plague and Politics: The Story of the United States Public Health Service*. Basic Books, New York
- Ray C G, Finley J K 1994 Did CMHCs fail or succeed? Analysis of the expectations and outcomes of the community mental

health movement. *Administration and Policy in Mental Health* **21**(4): 283–93

- Rothman D J 1971 *The Discovery of the Asylum: Social Order and Disorder in the New Republic*, 1st edn. Little, Brown, Boston
- Shaftstein S S, Turner J E, Clark H W 1978 Financing issues in providing services for the chronically mentally ill and disabled. In: Talbot J A (ed.) *Ad Hoc Committee on the Chronic Mental Patient. The Chronic Mental Patient: Problems, Solutions, and Recommendations for a Public Policy*. American Psychiatric Association, Washington, DC
- Talbott J A 1966 Community psychiatry in the army. *Journal of the American Medical Association* **210**(7): 1233–7
- Talbott J A 1978 *The Death of the Asylum: A Critical Study of State Hospital Management, Services, and Care*. Grune and Stratton, New York
- US Bureau of the Census 1998 *Statistical Abstract of the United States*, 18th edn. Washington, DC

J. A. Talbott

Public Reason

'Public reason' implies a contrast to 'nonpublic' or 'personal reason.' Identifying a distinctly public or shared notion of reason is only important—indeed is only intelligible—when the concept of reason is understood in ways which allow that individuals, employing good reasoning, may come to divergent conclusions. This article begins by examining the lines of inquiry that have led many to question whether all fully-informed and fully-competent inquirers would come to the same conclusions about science and morals. According to some philosophers and social scientists, the very concept of reason has broken apart, leaving different individuals and groups with different conceptions of rationality, and so of reasons for belief and action. The article then considers the different ideals of public reason that have been advanced to overcome worries about the 'fragmentation of reason.'

1. The Diversity of Reason

1.1 The Enlightenment

Before considering different conceptions of public reason, it is important to grasp what the very idea of public reason presupposes. According to the conception of reasoning that dominated the European Enlightenment of the seventeenth and eighteenth centuries, reason is inherently universal, and so shared and public. On this Enlightenment conception, reason is a shared capacity of all human beings, and norms of

good reasoning are universal. Any premise p that is true for one person is necessarily true for all others; if the inferential rule ' $(p \cdot (p \rightarrow q)) \rightarrow q$ ' is valid for one person, it is necessarily valid for all. The true and valid results of one person's reasoning are thus necessarily true and valid for all. According to this standard Enlightenment view, the use of common human reason not only produces consensus on scientific beliefs, it also leads to convergence of moral and political opinions. To be sure, people might disagree on matters of science or ethics, but that would be due to mistaken beliefs or irrationality: some have arrived at the wrong answer. The process of Enlightenment was the only remedy for this—the increasingly better use of reason to uncover truths about the natural and the social world. The ideal or model was that of Newtonian physics: just as our common reason had uncovered the laws of matter and motion, so too could it be expected to uncover the laws of human nature, society, morals, and politics. Each field was awaiting its Newton.

1.2 *Relativism and Social Science*

It is a familiar story how the systematic study of cultures with radically different norms from those of Europe gave rise to cultural relativism. When confronted with the very different mores of the Zuni and Kwakiutl Indians and Dobu Islanders of Papua New Guinea, Ruth Benedict (1934) endorsed a relativistic view—their norms were justified for them, just as Western norms were for Europeans. As anthropologists began to study non-European cultures in more depth, and became increasingly more sophisticated and self-reflective in their techniques, disputes arose about the proper method for anthropologists to interpret the belief systems of other cultures. For example, was a belief in magic or a deity to be understood as simply a false belief on which the 'natives' relied? John Stuart Mill (1974 [1843], pp. 766–7) believed that an investigator could best understand the magical beliefs of other cultures by attributing to them erroneous beliefs and invalid inferential rules. Others, adopting a *principle of charity*, have insisted that the best interpretation of a culture minimizes the number of false beliefs attributed to its members (Davidson 1973). Thus the best interpretation—which makes sense of the 'native's' metaphysical theories, religious convictions, and their beliefs about nature—might seek to show that a belief in spirits is, after all, rational given their world view. Thus the second step in the relativist attack on the Enlightenment's ideal of reason was to endorse *relativism concerning what beliefs are rational* in different cultures. The last step in the relativistic project is to apply tolerance to the idea of reason itself. Can Western anthropologists properly interpret other cultures if they apply their Western conception of reason in their interpretation? Is the very idea of reason culturally relative?

Independently, psychological studies of human reasoning led to doubts about whether everyone shares the same norms of reasoning. The work of, among others, Richard Nisbett, Lee Ross, Paul Slovic, and Amos Tversky showed that normal adults often do not employ the norms of reasoning long advocated by philosophers as correct. For example, many normal adults adhere to the 'gambler's fallacy.' According to standard probability theory, the odds of independent events occurring do not depend on whether that type of event has occurred in the past: if the odds of a 'six' being rolled by a fair dice is one-in-six, this probability in no way depends on whether no sixes have been rolled or all sixes have been rolled in the past. Yet many normal reasoners believe that, after a long run of sixes, the odds of another six are less than one-in-six, or if none recently have been rolled, many are convinced that a six is 'due.' Based on a variety of studies of such ordinary reasoners, some philosophers advocate 'normative cognitive pluralism.' According to Stephen Stich, '*Normative cognitive pluralism* is not a claim about the cognitive processes people do use; rather it is a claim about *good processes*—the cognitive processes that people ought to use. What it asserts is that there is no unique system of cognitive processes that people should use, because various systems of cognitive processes that are very different from each other are equally good' (1990, p. 13).

1.3 *Pluralism and Indeterminacy*

Social science thus provided one source of the 'fragmentation of reason.' Of course there was, and still is, lively debate within social scientific and philosophical circles about whether any such relativistic view of reasoning is justified, and if so to what extent individuals employ different, but equally good, norms of reasoning. Nevertheless, some form of cognitive relativism is widely accepted. Concurrently, developments in the philosophy of science were challenging the idea of common, human, reasoning in the very citadel of the Enlightenment—science itself. Whatever doubts may be entertained about the application of Western reason to other cultures, or departures from expert norms by ordinary reasoners, surely science is still the model of rational discourse tending to convergence of opinion leading to the truth. To many, Thomas Kuhn's analysis of scientific practice demonstrated that, rather than a shared project of pursuit of the truth, judged by rules of reason held by all, scientific inquiry takes place within 'paradigms' that determine not only what constitutes research problems but what constitutes success in a research program. *Within* a group of scholars sharing a paradigm the Enlightenment's ideal of common norms of reason and consensus on the truth will be approximated; and insofar as 'normal science' is dominated by a single paradigm in each field, there often is a common mode of reasoning. But in times of scientific crisis,

when paradigms compete, some have insisted that there is no rational, impartial way to adjudicate among the competing theories.

The crux of Kuhn's analysis is that because data underdetermines theories, scientists draw on values in adjudicating among theories. 'To a greater extent than other sorts of components of the disciplinary matrix,' writes Kuhn, 'values may be shared by men who differ in their application. Judgments of accuracy are relatively, though not entirely, stable from one time to another and from one member to another in a particular group. But judgments of simplicity, consistency, plausibility, and so on often vary greatly from individual to individual ... Even more important, in those situations where values must be applied, different values, taken alone, would often dictate different choices. One theory may be more accurate but less consistent or plausible than another' (Kuhn 1970, p. 185). The crucial claim is that there is no uniquely rational way to order these various desiderata—simplicity, consistency, plausibility—and different orderings endorse different scientific theories upholding competing truth claims. Equally well-informed scientists employing their reasoning in perfectly legitimate ways can arrive at different judgments about what is true. Fred D'Agostino has generalized the analysis. In any field of inquiry, when there exists both a plurality of criteria and no impartial way to order the criteria, the criteria are apt to be indeterminate in their application. Different orderings produce different outcomes—reasons for belief. According to D'Agostino, the criteria only yield a determinate result when an ordering is provided, but impartial reason cannot give us that. Much of post-modernism is based on this key idea. Reason is inherently perspectival because of the inherent plurality of relevant considerations for which there is no rational, impartial ranking.

1.4 Reasonable Pluralism

Kuhn's analysis of scientific practice and D'Agostino's generalization rely on the idea of indeterminacy: reason undetermines choice between various theories and perspectives. A more modest view maintains that our powers of reasoning are inconclusive on many complex matters of science, morality, and politics. Claim C is characterized by reasonable pluralism if some perfectly reasonable agents do, while others do not, have good reasons for accepting C. For a controversy to be characterized by reasonable pluralism it is not simply the case that people actually disagree about the merits of C, but that it is reasonable for one person to assert C and another not-C. This is a far less radical claim than was examined in the last section: it is not argued that the question 'C or not-C?' is inherently *indeterminate* but only that present beliefs about C are *inconclusively* justified. According to John

Rawls, our disputes about value seem subject to reasonable disagreement because our understanding of what is good and valuable is especially subject to what he has called the 'burdens of judgment.' According to Rawls, reasonable judgments so often are at odds because:

- (a) the evidence is often conflicting and difficult to evaluate;
- (b) (as in Kuhn's example) even when we agree on the relevant considerations, we often weigh them differently;
- (c) because our concepts are vague, we must rely on interpretations that are often controversial;
- (d) the manner in which we evaluate evidence and rank considerations seems to some extent the function of our total life experiences, which of course differ;
- (e) because different sides of an issue rely on different types of normative considerations, it is often hard to assess their relative merits;
- (f) in conflicts between values, there often seems to be no uniquely correct answer (Rawls 1993, p. 57).

Because these matters are so complex and uncertain, different people will reach different, competing, credible, or reasonable conclusions. In the face of complexity, our powers of reasoning do not produce convergence.

2. Conceptions of Public Reason

2.1 Public Rationality and Public Reasons

Important movements in social science, the philosophy of science, and moral and political philosophy thus have combined to cast grave doubt on the Enlightenment's confidence that the free exercise of human reason leads us to agree. In the face of these challenges to the ideal of common human reason, theories of distinctly public reason have been advanced to show how people can *share reasons* (Korsgaard 1996, Chap. 10). However, this idea is ambiguous. It may constitute a response to relativistic social science (Sect. 1.2), which challenges the very idea of a shared conception of human reason. Understood thus, the ideal of public reason is an ideal of *public rationality*, or a shared conception of rationality. The challenge posed by reasonable pluralism (Sect. 1.4) is more modest. Advocates of reasonable pluralism maintain that even if we apply shared standards of reason, our deliberations will not yield shared reasons for belief. In a weaker sense, then, the ideal of public reason is an ideal of *public reasons*, or how we can arrive at shared reasons to believe and act. A theory of public reason may advance a theory of public rationality and/or public reasons.

Three main conceptions of public reason stand out: (a) the epistemic, (b) the consensual, and (c) the political.

2.2 Epistemic Theories of Public Reason

In philosophy, 'epistemology' refers to the theory of knowledge or, more broadly, the theory of justified belief. Epistemic theories of public rationality, then, maintain that at least some common norms of reasoning can be justified to everyone. The Enlightenment view was that *all* valid norms of reasoning can be justified to everyone—reason was inherently public. Epistemic theories of public rationality need not, however, advance such a sweeping claim. Importantly, an epistemic theory of public rationality may adopt a modest version of the 'relativism of reasons' (Sect. 1.2), according to which in *some* cases person Alpha may have a fully justified norm of reasoning N while Beta accepts a fully justified norm of reasoning M, where it is the case that N and M are inconsistent. In cases where only N and M are relevant to deliberation, a relativism of rationality will manifest itself; Alpha may have reasons based on N that Beta will deny on the basis of M, and both will be justified. However, to establish a thoroughgoing relativism of reasoning that undermines the possibility of public reason, it is not enough to establish that this can *sometimes* occur; it must be shown that *no* common norms of rationality can be justified to all—that it is *always* (or at least usually) the case that for any justified norm of reasoning held by Alpha, some person Beta will be justified in accepting a competing norm. That is the sweeping relativistic claim rejected by epistemic theories of public rationality (Gaus 1996).

Advocates of epistemic public rationality can reply to both the anthropological argument for the relativism of reasons, and psychological studies uncovering the diversity of norms employed in reasoning (Sect. 1.2). In reply to the *anthropological argument*, advocates of public reason can embrace Martin Hollis' view that an anthropologist must build up an interpretation of an alien language by locating a 'bridgehead.' The assumption underlying the bridgehead strategy is that others generally perceive what we perceive and tend to say about it the sorts of things we would say. Relying on these assumptions, the anthropologist begins by translating everyday, basic perceptual sentences such as 'Yes, this is a brown cow' and 'No, it is not raining right now.' Now, Hollis insists, this bridgehead not only includes translations of such basic beliefs, but basic logical rules, since 'what a sentence means depends on how the beliefs which they express are connected, and that to justify a claim to have identified a belief one must show the belief is connected to others.' Consequently, in the main, the logic employed by those in different cultures must either be very similar to our own or unintelligible. Axioms such as ' $p \rightarrow p$ ' ' $\sim(p \sim p)$ ' and ' $(p \cdot (p \rightarrow q)) \rightarrow q$ ' are more than simply *our* norms of reasoning. They are constituents of any system that we can identify as a way of reasoning (Hollis 1970a, pp. 214–15, 1970b, p. 232.) If 'natives' reason logically at all, then they

reason as we do. The point, then, is that the context that sets the stage for the possibility of relativism—that we confront others, different from us but of whom we can make some sense—presupposes widespread shared norms of inference as well as beliefs. Thus the very possibility of mutual intelligibility sets a limit on the extent to which we can understand others as employing cognitive processes different from our own.

In reply to psychological studies showing the diversity of norms employed by normal reasoners, advocates of public reason can rely on evidence supporting the hypothesis that people employ a mental logic—a system of mental inference rules—in their actual reasoning. Indeed, most psychologists studying actual reasoning suppose some version of the mental logic theory. To be sure, mental logic may depart from the standard propositional calculus. Psychologists studying English reasoners, for instance, have found that English speakers do not interpret connectives in a way consistent with standard logic. Particularly troubling are interpretations of 'if, then' statements (as well as 'or' connectives), which are not typically used in ways that correspond to standard analyses in logic. In response to these findings, psychologists such as Lance J. Rips and Martin D. Braine have developed sets of natural inferential rules that correspond to the way in which speakers of English employ connectives. Rips has gone beyond this to develop rules modeling causal reasoning as well.

Mental logic views strongly incline toward the hypothesis that these logics are in some way natural or innate, though to precisely what extent is open to investigation. There may well be a complicated interplay between language and logic, such that English speakers tend to employ some inferential rules not utilized by others. A plausible hypothesis is that the capacity to develop a mental logic is innate, though the precise developmental path taken depends on a variety of factors, including one's native language. If so, however, this is sufficient to undermine the plausibility of a radical relativism of norms of rationality. The set of possible mental logics—laws governing thought—is apt to be pretty severely constrained.

2.3 Consensus Theories of Public Reason

Epistemic accounts of public reason uphold our ability to reason together because they insist that we do, indeed must, share basic rules of inference. In contrast, consensual accounts of public reason conceive it as arising out of a shared social life and/or shared discourse (see *Shared Belief*). A number of otherwise diverse philosophers such as Ludwig Wittgenstein, Jürgen Habermas, Richard Rorty, Susan Hurley, and Philip Pettit all point to the idea that reasoning is a social phenomenon. In order to be a deliberator, one must be a member of a community in which one's beliefs and rules are intelligible to others. One comes

to understand what is a justified or unjustified belief, and what is a good or a bad way of reasoning, by sharing the norms and standards of communities, including communities of inquirers. Thus some insist that 'ultimately, there is only one criterion by which beliefs can be judged valid, and that is that they are based on agreement reached by argumentation'—we converge on them (this is Habermas's 1991, p. 14 gloss on Rorty's 'particular version of discourse theory'). Habermas draws an intimate connection between mutual understanding (*Verständigung*) and agreement (*Einverständnis*). It would seem that to understand others we must in some way arrive at intersubjective agreement about norms of rationality and good reasons. Consequently, the phrase of 'public reason' becomes almost redundant: something can only be identified as reason because it is the object of public, interpersonal, agreement.

Wittgenstein and his contemporary followers such as Hurley (1989) and Pettit (1993) suggest a broadly similar view. Reasoning itself presupposes a 'community'—common rules of thought. Wittgenstein (1958) argued that one could not have a private rule that one only followed once—for in such a case we could not know whether the rule was actually followed. More generally, it has been argued that one cannot have a purely private rule, for one needs to be able to distinguish when one has correctly followed the rule and when one has gotten it wrong. But, assert Wittgensteinians, the difference between correct and incorrect application of a rule cannot be located in individual belief systems, but only by the individual comparing her application of the rule to those of her fellows: each must look to others to check and correct her rule-based performances. Since thinking is necessarily rule driven, the analysis applies to reason itself. Thus it seems that the very idea of rule-following—and, so, of reasoning—supposes convergence with others. If so, the search for reason is a search for consensus.

In one respect this view, somewhat surprisingly perhaps, returns to the basic idea of the Enlightenment—all reason is public reason. But whereas the Enlightenment was confident in asserting this because of the widespread faith in a universal, common, human faculty of reason that could accurately represent reality, consensual theorists see reason as arising out of a shared public life. To put it crudely, rather than reason producing consensus, consensus generates reason. Thus the publicness of such reason, unlike that of the Enlightenment's, is consistent with a cultural relativistic conception of rationality. The public, intersubjective agreement, reached by one culture—which is constitutive of their very conception of rationality—may be very different from that upheld by another.

Because consensual accounts of public reason stress the constitutive nature of public consensus to the very idea of reason, and because public consensus may be

promoted by political institutions, more radical consensual theorists such as D'Agostino give political institutions a constitutive role in determining *what reason is* (1996, Chap. 9). This is a radical view indeed—however, it is not entirely novel, perhaps going back as far as Hobbes. For Hobbes ([1651] 1948), the conflict in the state of nature arises from conflicting private judgments; people's private reasoning yields conflicting judgments of right and wrong, as well as matters of fact, and this leads to the less intellectual conflict that characterizes the state of nature. Hobbes's solution is to appoint an 'arbitrator.' This 'judge,' says Hobbes, provides 'public reason' to which private reason 'must submit.' The arbitrator proclaims what each has reason to do, and so defines a single, coherent conception of reason. Because, in Hobbes' view, we have authorized the judge (i.e., sovereign) to define public reason for us, the sovereign's pronouncements constitute a shared public reason on which there is consensus.

2.4 Public Reason qua Political Reason

D'Agostino's radical view allows the political to shape the rational. A more modest proposal with which this radical view is easily confused maintains that the notion of public reason is a purely political idea, in the sense that it concerns only how we reason together *about politics*. This understanding of public reasons lies at the heart of what is commonly known as 'political liberalism.'

Rawls's political liberalism is a response to the problem posed by reasonable pluralism (Sect. 1.4). For Rawls the crucial problem is that citizens in democratic societies entertain a plurality of what he calls reasonable 'comprehensive doctrines'—overall philosophies of life centered on religious, philosophical, or moral beliefs. This pluralism seems a permanent feature of modern societies. 'Political liberalism assumes that, for political purposes, a plurality of reasonable yet incompatible comprehensive doctrines is the normal result of the exercise of human reason within the framework of free institutions of a constitutional regime' (Rawls 1993, p. xvi). The reasonable pluralism of comprehensive views renders such views unacceptable as bases for the justification of political power. Rawls endorses the 'liberal principle of legitimacy,' i.e., that 'our exercise of political power is fully proper only when it is exercised in accordance with a constitution the essentials of which all citizens as free and equal may be reasonably expected to endorse in the light of principles and ideals acceptable to their common human reason' (p. 137). Thus, because there exists a reasonable plurality of comprehensive doctrines, basing the justification of political power on any of them violates the liberal principle of legitimacy. This leads Rawls to seek a political conception that 'all affirm' and that is 'shared

by everyone.' Such a conception would be supported by the diverse reasonable comprehensive doctrines that characterize our democratic societies. It would generate public reasons that would fit into our many reasonable, but irreconcilable, comprehensive views. Some version of the principle of political legitimacy motivates the search for public reasons in contemporary liberal political philosophy: if the exercise of coercive political power relies on nonpublic reasoning, it is oppressive and illegitimate in relation to those who do not share those reasons.

Rawls's core argument—and so his entire political liberalism—apparently depends on the contrast between comprehensive views and the shared, political, conception. Whereas comprehensive conceptions produce diverse reasons that conflict, the political point of view provides public reasons—reasons that all reasonable comprehensive views can endorse.

Habermas (1995) suggests that Rawls posits an '*a priori*' distinction between the political and non-political spheres. At least at times Rawls seems to suggest that the political point of view can be conceptually distinguished from moral, religious, and philosophical matters; whereas the former identifies a common point of view that can be affirmed by all, when human reason is applied to moral, religious, and philosophical issues it leads to reasonable disagreement. It is difficult, however, to accept the claim that some single conception of the political is shared by all reasonable persons in modern democratic societies. Political theorists have often pointed to the 'political' as an 'essentially contested' concept—one that is open to divergent, reasonable, interpretations (Connolly 1983, Chap. 1). If citizens of modern democratic societies disagree about the concept of the political, it cannot serve the function of identifying public reasons that overcome reasonable pluralism. Like the moral and philosophical, the nature of the political itself seems a matter of reasonable disagreement.

An alternative interpretation of Rawls' view is to take the political conception as *constructed out of* that which is shared. On this reading the *nonpolitical* is, by definition, those matters on which the free use of reason by citizens in democratic societies leads to different, reasonable conclusions. It is, by its very nature, the realm of reasonable pluralism. In contrast, one might *define* the political as those matters on which reason converges in democratic societies, and so necessarily generates constitutional principles that satisfy the principle of liberal legitimacy. The political is thus characterized as the overlapping consensus of those diverse comprehensive views characterizing modern societies. The problem for this interpretation is that Rawls himself indicates that the free use of human reason leads us to reasonable disagreement about conceptions of justice and constitutional essentials. The 'political' qua 'shared perspective' is limited to the abstract concept of a 'liberal political order' (Rawls 1993, p. 241). 'By this [Rawls says] I mean

three things: first, it specifies certain basic rights, liberties, and opportunities (of the kind familiar from constitutional democratic regimes); second, it assigns a special priority to these rights, liberties and opportunities, especially with respect to claims of the general good and of perfectionist values; and third, it affirms measures assuring all citizens' adequate all-purpose means to make effective use of their basic liberties and opportunities' (p. 223). Thus stated, the content of public reason seems sparse indeed. Unfortunately, if a more detailed articulation of this idea is provided, Rawls indicates that the use of human reason again leads to disagreement. Apparently modern democratic societies agree on only the most abstract of public—political—reasons.

3. Conclusion: Reasoning in a Plural Society

Puzzles about the possibility of shared, public rationality and reasons derive from a variety of sources: social science, the philosophy of science, and the moral pluralism of modern societies. Although developments in all these fields serve to cast doubt on the Enlightenment's ideal of a common universal reason, they pose different challenges. The findings of anthropology call into question universal norms of rationality; psychology questions whether so-called universal norms are actually employed in human reasoning. Kuhnian philosophy of science suggests an irreducible indeterminacy of the application of our general standards, even in science itself. Worries about reasonable pluralism focus on the possibility of shared moral and political norms in societies that disagree deeply on questions of value and the nature of the good life.

Just as the challenges arise from multiple sources, so do theories of public reason. *Epistemic theories* dispute that modern social science casts serious doubt on the necessity of shared norms of rationality in the interpretation of other cultures and in individual deliberation. Although some divergence in justified norms of reasoning no doubt occurs—and can be made sense of—radical divergence is implausible. *Consensus theories* of public reason maintain that to make sense to each other we must think alike—we must employ common rules of thought. The application of such common rules, say neo-Wittgensteinians, only makes sense in the context of shared social practices. Reason arises out of shared social life; reason is inherently public because it is necessarily a product of social consensus. Finally, *political liberalism* seeks to establish that, amid the diversity of moral and philosophical views that characterize modern societies, a common political point of view can generate shared public reasons to regulate coercion.

The ideal of public reason, then, is as diverse as the sources of our doubts about the possibility of shared ways of reasoning in contemporary societies. In all its manifestations, however, the ideal of public reason

seeks to maintain the heritage of the Enlightenment—the free use of human reason will lead us to converge on the best justified beliefs.

See also: Bioethics; Examples from the Life Sciences; Economics and Ethics; Ethics and Values; Prisoner's Dilemma, One-shot and Iterated; Utilitarian Social Thought, History of; Utilitarianism: Contemporary Applications; Welfare

Bibliography

- Benedict R 1934 *Patterns of Culture*. Houghton Mifflin, Boston, MA
- Braine M D S 1978 On the relation between the natural logic of reasoning and standard logic. *Psychological Review* **85**: 1–21
- Connolly W E 1983 *The Terms of Political Discourse*, 2nd edn. Princeton University Press, Princeton, NJ
- Davidson D 1973 Radical interpretation. *Dialectica* **27**: 314–28
- D'Agostino F 1996 *Free Public Reason*. Oxford University Press, New York
- D'Agostino F, Gaus G F 1998 *Public Reason*. Ashgate, Aldershot, UK
- Gaus G F 1996 *Justificatory Liberalism: An Essay on Epistemology and Political Theory*. Oxford University Press, New York
- Habermas J 1990 *Moral Consciousness and Communicative Action*. MIT Press, Cambridge, MA
- Habermas J 1995 Reconciliation through public reason: remarks on John Rawls's political liberalism. *Journal of Philosophy* **92**: 109–31
- Hobbes T 1948 [1651] *Leviathan*. Basil Blackwell, Oxford, UK
- Hollis M 1970a The limits of irrationality. In: Wilson B R (ed.) *Rationality*. Blackwell, Oxford, UK, pp. 214–20
- Hollis M 1970b Reason and ritual. In: Wilson B R (ed.) *Rationality*. Blackwell, Oxford, UK, pp. 221–39
- Hurley S 1989 *Natural Reasons: Personality and Polity*. Oxford University Press, New York
- Kahneman D, Slovic P, Tversky A (eds.) 1982 *Judgments Under Uncertainty: Heuristics and Biases*. Cambridge University Press, Cambridge, UK
- Korsgaard C M 1996 *Creating the Kingdom of Ends*. Cambridge University Press, Cambridge, UK
- Kuhn T S 1970 *The Structure of Scientific Revolutions*, 2nd edn. University of Chicago Press, Chicago
- MacIntyre A C 1988 *Whose Justice? Which Rationality?* University of Notre Dame Press, Notre Dame, IN
- Mill J S 1974 [1843] A System of Logic. In: Robson J M (ed.) *The Collected Works of John Stuart Mill*. University of Toronto Press, Toronto, ON, Vols. 7 and 8
- Nisbett R, Ross L 1980 *Human Inference: Strategies and Shortcomings of Social Judgment*. Prentice-Hall, Englewood Cliffs, NJ
- Pettit P 1993 *The Common Mind: An Essay on Psychology, Society and Politics*. Oxford University Press, New York
- Rawls J 1993 *Political Liberalism*. Columbia University Press, New York
- Rorty R 1991 *Objectivity, Relativism and Truth*. Cambridge University Press, Cambridge, UK
- Rips L J 1983 Cognitive processes and propositional reasoning. *Psychological Review* **90**: 38–71

Stich S P 1990 *The Fragmentation of Reason*. MIT Press, Cambridge, MA

Wittgenstein L 1958 *Philosophical Investigations*, 3rd edn. Macmillan, London

G. F. Gaus

Public Relations in Media

Of all the many factors that shape the news appearing in the media, the rise of source professionals or public relations experts is one of the most dramatic. Known under a variety of titles, from public relations practitioners, communications managers, and political consultants, to pollsters, spin doctors, and lobbyists, their job is to manage the news on behalf of their employers, ensuring their visibility, boosting their image, or repairing their reputation. These cultural intermediaries also help to feed the growing appetite for stories generated by the twenty-first century's proliferating media and, as such, have a significant influence on both the agenda and content of current journalistic output (for further references see *Political Advertising*; *Political Communication*).

It was the establishment of universal suffrage that gave the biggest fillip to the spread of modern day public relations activity in the public sphere. For, just as the newly enfranchised citizens began to seek information about their country's power bases—over which they could now exercise some control—their opinions began to assume an impact that they did not have before. Public opinion became worth swaying by those who needed popular support. This coincided with increasing literacy that initially promoted the spread of newspapers and subsequently the development of the broadcast media, all of which were hungry for news items.

In early definitions of the role of the professional source, Bernays—one of the profession's founding fathers—commented that the job of those brought in to court this newly influential public opinion was to isolate ideas and develop them into events that could be used as news items by the media (1923, p. 171). Writing in 1922, Lippmann, observed that a new age of opinion management had arrived and the creation of consent among the governed had begun to develop beyond anything seen before in the political organization of human affairs. Public relations thus emerged in the early decades of the twentieth century particularly to link the political actor as a source of news with the media who would produce it, in a relationship that met the needs of both (McNair 1998, p. 146).

The role of public relations in swaying public opinion has since been regarded as crucial to those

seeking to gain or maintain political power. Consequently, as an instrument in manipulating a powerful public opinion consensus, the professional source operating within the political arenas of liberal democratic states has attracted increasing attention.

1. Source–Media Relations

News that influences public opinion is heavily affected by the nature of the relationship between source and media. Throughout the 1970s and for much of the 1980s, two contrasting sociopolitical paradigms prevailed. One was the Marxist view, which argued that media content was governed by elite sources representing the dominant classes of society with the result that their ideology permeated the media to the virtual exclusion of alternative views. This dominance was achieved with the compliance of journalists who readily accepted, and often sought, their input because these elite sources were perceived as authoritative and even objective. Furthermore, the dominant class learnt to deliver stories that conformed to the criteria of newsworthiness and the regularity of their news supply could be relied upon by the news media. In contrast, the pluralist view maintained that there was a wealth of groups and interests all competing for influence. The media were thus acting as a marketplace of ideas, none predominating to the exclusion of others.

However, recent research on public relations has delved into the question of access and the degree of openness and closure of the communications system, revealing a more complex picture than either of these early paradigms suggested. To what extent does the dominant class, characterized by the corporate and state elite, control the media agenda through direct access to the detriment of non-institutional and poorly resourced groups? This new research on news sources and promotional culture (see Wernick 1991), concentrates on the public relations activities of corporate and institutional sources in the public sphere and also on the strategies of opposition and resource-poor groups as they seek to gain access to the media.

The political economy perspective, with its emphasis on economic power and media ownership, argues that corporate and institutional/state sources have huge resource advantages that enable them to dictate the current of news by restricting or enabling information flows (Glasgow Media Group 1976, Herman and Chomsky 1988, Gandy 1982, for further reference see *Mass Media, Political Economy of*). Similar to this position is Hall et al.'s (1978) radical structural culturalist approach which concentrates on identifying the legitimacy conferred on state and corporate sources by journalists. Media statements are grounded in objective and authoritative statements from sources accredited by virtue of their institutional power and position. News production inevitably leads to a systematically structured over accessing to the media of those in powerful and privileged institutional po-

sitions. The result of structured preference conferred by journalists to the views of these elites is that they become the primary definers of topics and the primary interpreters of all subsequent debate on the issues.

Challenges to this account of source-journalist behavior as being overly deterministic come from the recently termed radical pluralist. Whilst acknowledging the structuralist view that elite sources hold strategic advantages and wield substantial economic and political power over the news agenda, Schlesinger and Tumber point out that 'because the conception of primary definition resolves the question of source power on the basis of structuralist assumptions, it closes off any engagement with the dynamic processes of contestation in a given field of discourse' (1994, p. 21). The concept of primary definition ignores: the conflict between elite sources therefore presenting the media with more than one message; the role of 'leaks' and off-the record briefings by non attributable sources; the non-retention of access over time as new forces and their representatives emerge; the vulnerability and possible discreditation of 'accredited' sources; and finally, journalists' challenges to official sources (1994, pp. 18–20). In this formulation primary definers are therefore not structurally predetermined, but earn this role by providing journalists with regular, reliable, and authoritative story leads.

These considerations have rethought the behavior of sources and led to the introduction of new methodologies in conducting empirical studies. Instead of relying solely on internalist accounts that rely on either journalists' descriptions of their relationships with sources or a reading of media content, recent research has expanded externalist or source centered approaches by examining the media strategies and tactics of news sources (see Ericson et al. 1989, Deacon and Golding 1994, Hallin 1994, Schlesinger and Tumber 1994, Miller 1994, Anderson 1997). All these studies with their various approaches emphasize the way all groups, official and nonofficial, whether state, corporate, voluntary, or pressure need to foster favorable public opinion through the establishment and protection of identity and image. New pluralist accounts offer the view that 'public relations is less capital dependent than other forms of public communication and non-official sources can employ strategies that require little institutional legitimacy' (Davis 2000, p. 51) thereby conferring on them the potential for gaining access to the media that traditional radical accounts did not acknowledge.

2. The Growth of Public Relations in the Media

As a business activity, public relations began to gain in importance from the end of World War I particularly in the USA and later in Europe. Corporate identity work grew up in the early 1960s and, as a professional practice, gained wider acceptance in the following decade during which there was an upsurge in

acquisitions, mergers, and divestitures. In commercial climates such as these, both predatory companies and their corporate prey have found that it has paid to heighten their image. And, as they have merged, both nationally and internationally, there has been a need to establish and maintain a single identity by which the corporate whole can be recognized (Meech 1996, pp. 65–6). Privatization programs through which previously state-owned companies seek shareholders from among the general public, also boosts the need for the companies involved to seek regular, favorable coverage in the media. This is generated through public relations professionals, both in-house and from hired agencies, whose specialisms now include media relations, public affairs, issues management, and lobbying. The lead shown by the corporate world has been enthusiastically embraced by both political parties and politicians.

2.1 Proactive Public Relations

The work of these public relations professionals can be described in two broad categories. First, there are proactive public relations through which, in particular, ‘pseudo events’ are used to create news that will be attractive to journalists (Boorstin 1962). These artificially created ‘happenings’—such as press conferences and photo opportunities—are staged purely to present journalists with material that is tailored to their needs. But their lack of authenticity and rationality has been criticized for promoting style over substance, and performance over policy, thereby turning news into a promotional form for politics and business.

Public relations, particularly lobbying on behalf of corporate and other sectional interests, has risen in importance. The size and development of the public relations industry in the United States and the United Kingdom mean that British and American firms dominate the market in most member states of the European Union, and their actions are increasing in scope and intensity. Public relations specialists and lobbyists are also diversifying by targeting not just the national and regional governments in the member states, but also the EU institutions, along with trade associations, corporations, and interest groups (Miller and Schlesinger 1999).

2.2 Reactive Public Relations

The second category can be described as reactive public relations. Sometimes called crisis management, this form of news manipulation springs into action to counter real news events that are potentially damaging. For politicians the rise in scandal news has led to a preoccupation with new strategies and tactics to counter negative images. Within the corporate

world, as globalization has developed, multinational companies, face problems arising in one small part of their global operation turning into an international crisis within minutes thanks to ever-faster and more ubiquitous electronic links and the consequent speed with which media agencies can flash news around the world. During the 1980s and 1990s, a spate of corporate public relations offensives was launched to refute criticisms over safety procedures and corporate responsibility in response to a number of well-publicized disasters and accidents. Swift action was also deemed necessary in the wake of takeover and other financial scandals, and to allay public fears about the real or potential dangers of contaminated or faulty products. High profile cases illustrate that an efficient crisis management plan, with an emphasis on skilful news management, can be effective in lessening the damage such catastrophes can wreak on a company. Indeed, those that have managed crises well in public relations terms have even managed to benefit from them. They have achieved this by attracting public sympathy in the immediate aftermath of a disaster, and then by winning people’s trust and respect for the way in which it has been subsequently handled. In contrast, a corporate crisis that is badly handled in public relations terms can cause irreparable damage leading to board resignations, a plunge in share value, and even corporate takeover (for further reference see *Scandal: Political*).

The message that even the largest companies ignore at their peril is that no organization or individual is immune to negative media coverage, especially now that effective public relations techniques are being increasingly understood and adopted by campaigning organizations. These relatively small bodies have found that lack of resources need not prevent them, too, from manipulating the news and, as a result, they have scored a number of public relations successes at the expense of government and big business. Since the 1960s, there has been a significant increase in organized lobbying with the emergence of several single-issue pressure groups particularly within the green and consumer movements. These organizations have pioneered new ways to harness media coverage to coincide with campaigns in a way that has compelled politicians and other powerful groups to take them seriously. Thus, through the creation of pseudo events, the targeting of specialist correspondents, and other techniques to meet journalists’ need for an endless flow of new material, these pressure groups indicate that non-official and resource poor organizations can also make significant interventions in media discourses despite lacking definitional power and cultural capital (see Bourdieu 1979). The ability of resource poor groups to gain access to the media, though, requires favorable conditions and the requirement to establish credibility and maintain reliability. Occasional entry into the public arena is possible but not secure in the long term (Goldenberg 1975).

3. Impact on Democracy

Public relations has emerged as a key specialism in the conduct and management of national and international political affairs and also corporate image. But its impact on democracy has often been viewed with alarm. For instance, the rise of the spin-doctor on the political scene has led to accusations of distorting news to an undesirable extent. The argument is that if the electorate requires objective information, then the aim of professional sources to manage this information represents a threat to democracy. This concern has been heightened by changes in the working conditions of journalists, which have increased their reliance on the input of sources. As many media organizations have evolved into businesses whose prime concern is the bottom line, cost cutting has been a focus. Thus, at the same time as external organizations are investing growing amounts in public relations activity, media companies are reducing editorial budgets, compelling journalists to produce more for less. With less time and money to produce their own stories, journalists are obliged to rely more heavily on information subsidies supplied to them by sources. Consequently, now that journalists frequently have little time to check facts or to mount any independent investigation, professional sources have assumed more power in placing their stories in the media, complete with their particular spin. And as a result, the journalist's role, both as gatekeeper and as watchdog, is diminished.

Despite a reformulation of source media relations and public relations activity there is general acknowledgment that source access is still far from equal. Although non-institutional and resource poor organizations have become skilled at manipulating the news media, they are still at a disadvantage lacking the cultural capital of corporate and state sources in the ability to limit or enable access to information and, neither possessing the political, legal, and financial power to apply pressure to journalists. Furthermore, despite conflict between elite sources, these still have an automatic legitimacy that other sources have to earn.

Powerful institutions, 'driven by the promotional dynamic to attain new heights of sophistication in trying to manage the news' (Schlesinger and Tumber 1994, p. 275) are, therefore, likely to remain the most common suppliers of material to journalists, with elite sources continuing to invest heavily in activities to sway public opinion in their favor. And, whatever concerns it might provoke, public relations is now seen as a necessary element of both the modern political process and of corporate management.

See also: Advertising Agencies; Advertising: Effects; Mass Media: Introduction and Schools of Thought; Mass Media, Representations in; Media Ethics; Media Imperialism; Media, Uses of; Scandal: Political

Bibliography

- Anderson A 1997 *Media, Culture and the Environment*. UCL Press, London
- Bernays E 1923 *Crystallizing Public Opinion*. Boni & Liveright, New York
- Boorstin D 1962 *The Image or What Happened to the American Dream*. Weidenfeld and Nicholson, London
- Bourdieu P 1979 *Distinction*. Routledge, London
- Ericson R V, Baranek P M, Chan J B L 1989 *Negotiating Control: A Study of News Sources*. Open University Press, University of Toronto Press, Toronto
- Davis A 2000 Public relations, new production and changing patterns of source access in the British national media. *Media Culture and Society* 22: 39–60
- Deacon D, Golding P 1994 *Taxation and Representation*. John Libby, London
- Gandy O H Jr 1982 *Beyond Agenda Setting: Information Subsidies and Public Policy*. Ablex, Norwood, NJ
- Glasgow University Media Group 1976 *Bad News*. Routledge and Kegan Paul, London
- Goldenberg E 1975 *Making the Papers: The Access of Resource-poor Groups to the Metropolitan Press*. Lexington Press, Lexington, MA
- Hall S, Crichter C, Jefferson T, Clarke J, Roberts B 1978 *Policing the Crisis: Mugging, the State, and Law and Order*. Macmillan, London
- Hallin D C 1994 *We Keep America on Top of the World: Television Journalism and the Public Sphere*. Routledge, London
- Herman E S, Chomsky N 1988 *Manufacturing Consent: The Political Economy of the Mass Media*. Pantheon, New York
- Lippmann W 1922 *Public Opinion*. Harcourt, Brace and Co., New York
- McNair B 1998 *The Sociology of Journalism*. Arnold, London
- Meech P 1996 Corporate identity and corporate image. In: L'Etang J, Pieczka M (eds.) *Critical Perspectives in Public Relations*. International Thomson Business Press, London
- Miller D 1994 *Don't Mention the War: Northern Ireland, Propaganda and the Media*. Pluto Press, London
- Miller D, Schlesinger P 1999 The changing shape of public relations in the European Union. In: Heath R (ed.) *The Handbook of Public Relations*. Sage, London
- Schlesinger P, Tumber H 1994 *Reporting Crime: The Media Politics of Criminal Justice*. Clarendon Press, Oxford, UK
- Wernick A 1991 *Promotional Culture: Advertising, Ideology and Symbolic Impression*. Sage, London

H. Tumber

Public Sector Organizations

'Public sector organizations' are entities that have been formed to manage the policy and operating requirements that enable a government to achieve its goals of public governance. The term 'public governance' has been defined as the management of a nation through the use of political power (Rainey 1991), or more simply put, answers the question: 'How should government govern, and what should it do' (Peters

1996, p. 19)? Such a broad definition of public governance permits a wide range of activities to be undertaken by public sector organizations as the administrative arm of government.

The role and structure of public sector organizations depend on the choices made by governments over time and the traditions widely recognized within the system of government operating in a country. Public sector organizations can be structured as central agencies of government, service providers, government business enterprises (GBEs), regulatory bodies, watchdog agencies, and as quasi-government bodies. While many of these institutions have remained comparatively unchanged, the emergence of Next Steps agencies in the United Kingdom represents a significant example of a new approach to the structure of public sector organizations. New Steps agencies operate under a form of contract that sets out the scope of their responsibilities and accountability to both government and clients, while related policy issues remain under the control of government.

In the 1930s, Mary Parker Follett suggested that there were significant 'similarities between the business of government and the government of business' (Corkery 1989, p. 193). However, it took until the last two decades of the twentieth century for reform of public sector organizations to explore this notion through the application of private sector practices to the public sector. This reform process coincided with a shift away from a politico-economic ideology involving the pursuit of full employment, a focus on demand management within an economy, and intense government intervention, characterized by the principles expounded by John Maynard Keynes. In its place emerged the so-called, neoclassical model of economic management. Pioneered by Milton Friedman, this economic ideology centers upon the capacity of an economy to meet potential demand (supply-side economics), and the regulation of economic activity through the acceptance of market forces in place of government intervention. The neoclassical model is closely associated with the principles of smaller government pursued at first instance, by national leaders Thatcher, Reagan and Mulroney of the United Kingdom, the USA, and Canada respectively.

By the early 1990s, the concept of reform of public sector organizations had gained a significant international following. Often referred to as the New Public Management (NPM), the drive for reform was accompanied by emergence of the range of public sector organizations, an example of which is outlined above. The NPM was accompanied by a wave of privatizations in a number of jurisdictions, and by a rhetoric that espoused a new approach to the management of public sector organizations in terms of 'doing more with less,' 'letting the managers manage,' and 'pursuing greater efficiency and effectiveness.' This rhetoric, often described as 'managerialism,' supported the application of private sector principles

to the management of public sector organizations and translated into a variety of radical reforms in public organizations designed to support the philosophy of smaller government. These reforms included: reducing the number of public employees through 'downsizing,' outsourcing a variety of activities previously undertaken by public sector organizations, introducing budget reductions to enforce financial restraint within public sector organizations, and selling government assets and entities to reduce the budgetary problems experienced by many governments.

Over this period, research into public sector organizations has examined the consequences of these reforms. The research has divided along lines that either support market-oriented reforms or seek to justify a return to more traditional models of public sector organizations. However, some recent research highlights the importance of strong citizen support for the public sector and attempts to put to rest the view that the 'public-sector service is inherently inferior to private-sector service' (Kernaghan 2000, p. 7). Organizational theory research in this area has also explored the consequences of adopting private sector organizational practices in the public sector. It is likely that the divergences of opinion about the relative merits of organizational performance in each sector will remain a source of intense debate and a matter of personal opinion.

1. *The Rise of the Public Sector Organization*

A theory of a public sector organization might be first found in the development of the democratic city-state described by Plato in *The Republic*. His concept arose from the notion that people are 'not independent of one another, but need the aid and co-operation of others in the production of the necessities of life' (Copleston 1962, p. 251). Livy (15 BC) also described a public sector that was involved in sophisticated risk management negotiations with the private sector, whose help was needed to save part of the Roman army stationed in Spain. However, evidence from Chinese scholars suggests that public sector organizations existed in that country in the Xia Dynasty, many centuries earlier than in Europe.

The origins of public sector organizations in the English-speaking world can be traced to the arrival of William the Conqueror on British shores in 1066. His administration led to the development of public institutions that supported the power and financial strength of the monarch, through the development of the Common Law and the creation of the Domesday Book, which recorded the ownership of land throughout the kingdom which then encompassed England. At this time, the accountability of these institutions was to the monarch, rather than the polity, demonstrating the nature of the public sector that existed during this period of feudal society.

The emerging role and structure of public sector organizations can be seen in their support of empires in the late Middle Ages and the transfer of traditional public sector roles and structures from the dominant power to a subordinate country. This transfer led to some commonality of ideals in public sector organizations that can be traced back to the influence of, or reaction against, periods of colonial rule of a nation. Bureaucratic traditions also became well-established features of public sector organizations, a factor that has been a source of criticism of public sector organizations and the subject of considerable theoretical and practical evaluation.

In 1776, Adam Smith had suggested the traditional role of government should be limited to the protection of a nation from internal and external violence, the administration of a fair system of justice, and the development and maintenance of major public works. These principles found favor in the early nineteenth century and, more recently have attracted approval from those supporting contemporary neoclassical views. The Northcote–Trevelyan report in 1853 eventually led to a radical and influential restructure of the British civil service and provided the foundation of modes of public sector organization in the twentieth century.

It could be argued that in many jurisdictions public sector organizations enjoyed the peak of their influence in the economic boom years of the 1960s and 1970s. However, by the 1980s, the pressure for reform had begun as the changing economic ideology of governments and emerging budgetary pressures forced governments to re-examine the role of their public sector institutions.

2. The Nature of a Public Sector Organization

It is the complex relationship between public sector organizations and government, which distinguishes this type of organization from those in the private and nonprofit sectors, regardless of the recent attempts to merge public with private sector management practices. Public sector organizations can be distinguished from other sectors by their accountability to the principles of public governance adopted by a government. It can be claimed that this accountability gives rise to an overriding management (including ministerial) culture that punishes mistakes and guarantees that few risks will be taken. In a democratic state, there is an expectation that the public sector is also accountable to the government, which is itself answerable to the electorate. In a more authoritarian state, the public sector can be expected to be accountable to the government alone, although in this environment, public opinion may remain an important countervailing force.

The issue of accountability to government and in turn, civil society, highlights another distinctive fea-

ture of public sector organizations: the decision-making complexity arising from the need to cater for the diversity of demands faced by public sector organizations. The challenges to the decision-making process emerge, first, from the comprehensive legal requirements that bind the form and operation of public sector organizations. In addition, the demands of civil society, the intensity of competing interests within an economy, and the unpopular role of a number of public entities, all contribute to the decision-making complexity experienced by public sector organizations. For example, as societies have become more diverse and interest groups more influential, their competing demands make the role of both government and public sector organizations, more difficult (Olson 1982). In the face of these tensions, there are suggestions that the emerging challenges of globalization require ‘the emergence of new forms of government’ (Metcalf 2000, p. 122) able to conceptualize and respond to the increasingly varied and competing demands of public sector management.

Public scrutiny of the policies and operations of public sector organizations is also widespread. Formal scrutiny on behalf of government may occur through a systematic process of audit as well as periodic review by politicians and the judiciary. However, informal review by the media represents a major tension for public sector organizations whose performance may be the subject of intense media review at any time, for political purposes and ‘in the public interest.’ There are suggestions that this pressure makes the performance and decision processes of public sector organizations slower and more conservative as the managers of these organizations work to avoid public criticism of their organization and the government.

The decision-making processes used in the public sector may be seen as slower and more dilatory than other sectors because of the bureaucratic and accountability structures of government organizations. This perception was supported by Drucker who, in the late 1960s suggested that ‘any attempt to combine governing with ‘doing’ on a large scale paralyses the decision-making capacity’ (1969, p. 233). While this criticism may arise from a frustration with the speed with which public sector organizations move towards many decisions, it occurs because of a failure to recognize the impact of bureaucratic and political demands on the decision processes within government and public sector organizations.

The comparative absence of competition between public sector organizations and between public and private sector organizations has also been a distinguishing feature of public sector organizations, although there are exceptions to this rule, especially in the case of some GBEs that operate in a competitive business environment. As Osborne has commented: ‘Businesses typically exist in fiercely competitive markets, most public organizations don’t’ (1994, p. 137). Business performance is also built upon comparatively

concrete targets such as overall profitability, return on assets, return on owners' funds, inventory turnover, and the maintenance of cash flow.

Non-traditional performance measures such as assessing customer loyalty, the level of customer satisfaction, and relationship management have also been developed by some organizations seeking to enhance their market standing. However, there are a number of other performance criteria, such as timeliness, ability to operate within budget expectations, and perceived quality of implementing government policy, which might be substituted for profit in public sector organizations. A customer service focus has been a feature of the reform of public sector organizations in the 1980s and 1990s, and remains high on the agenda of the US and UK governments.

The operating environment of a public sector organization has often been regarded as being in some way inferior to private sector organizations. This view, which has long been promoted in literature (Parkinson 1958) and the media, may have influenced public opinion against members of public sector organizations, a view that several governments are actively attempting to redress. The changing nature of the service requirements of many public sector institutions has also created pressure to move away from the generalist model of management. There is an increasing expectation that managers of these entities, in addition to their existing management skills, may find they now need some of the skills possessed by many private sector managers. These capabilities include a working knowledge of sophisticated aspects of historical cost accounting, information technology, national and international finance, contract negotiation, sales and marketing, human resources management, and assessment of customer satisfaction.

3. Public Sector Organizations and the New Public Management

The causes of the drive to reform public sector organizations are multiple. They can be traced to the political pressures faced by governments in the early 1980s, which began to find that budgetary constraints emerged, full employment had become an apparently unattainable goal, inflation was more difficult to manage than expected, and the ideology of economic management had begun to change. The model of demand management enunciated by Keynes came to be replaced by the more rational, market forces model, or supply-side economics, an ideology that also found favor in developing new policies for defining the relationship between public sector organizations and their respective governments and polities. However, reform can also be attributed to the need for politicians to be seen to be doing something constructive to please both the polity generally as well as the financially powerful private sector.

As noted earlier, the most contentious reform of public sector organizations focuses on the reengineering of the public sector within the philosophies espoused by those promoting the concept of the NPM. At its extreme, the emergence of the NPM has seen a shift in the focus of government from seeking policy advice and implementation, to a drive for a more 'efficient and effective' public sector based on performance principles more commonly associated with the business sector. The accompanying rhetoric includes the goals of 'economic competitiveness, doing more with less, building faith in government and strengthening community and civil society' (Gore 1999, p. 1). Across the Atlantic, the goals of the British Government have been expressed as targets in five key areas of public service, namely:

- forward looking policy making;
- public services focused on needs of users;
- quality public services;
- the use of technology to support a modernized public service; and
- a move to value, rather than denigrate, public service (Her Majesty's Government 1999).

From an organizational point of view, these emerging trends have resulted in a number of significant changes to many public sector organizations. Because of the range of ways in which public sectors are structured, these trends vary in their importance. In many public sector organizations, attaching private sector salary levels to senior executives who are appointed under contract for a specific term has tended to increase the politicization of public sector organizations and has changed their traditional relationship with the government.

The contracting out of activities previously undertaken by public sector organizations has had the potential to reduce their knowledge of particular markets and affect their ability to enter contractual arrangements. One of the main areas of risk is ensuring that the legal and operational sophistication possessed by private sector organizations, is matched by those acting on behalf of the public sector. Contracts are required which take account of possible contingencies arising during the life of a contract. Because these contracts exist between the public sector and private sector, in some jurisdictions there are increasing numbers of cases where the transparency usually associated with public contracts is increasingly hidden behind the barrier of 'commercial in-confidence,' perhaps to the detriment of the polity through the diminution of public review.

The outcomes of the reform process have been mixed and are subject to widely varying interpretations. The reform process is not an event; it represents a continuum of change in public sectors in most countries, where the pressures of globalization, technological change, reduction in government debt, and emergence of aspects of the neoclassical approach have taken hold. As a result, it is not possible to

evaluate the precise cost-benefit of public sector reforms presently, nor is such an evaluation likely to be very reliable in the future. Some public sector organizations have reduced their regulatory control within the economy, while others have found it necessary to consider a re-regulation process. The contracting-out of activities within some public institutions has provided both benefits and costs to government and civil society, an experience that is likely to be mirrored within private sector organizations that have undergone similar reengineering processes. Major experiments in social change such as that attempted in New Zealand have led to complex outcomes for both public and private sector organizations, the merits of which remain widely debated.

4. Future Directions

As the twenty-first century begins, it is clear that public sector organizations will continue to experience the tensions that arise between the demands of government, business, and civil society. The basic role of public sector organizations may not change dramatically, although governments may find themselves under intense pressure to meet the demands of rapidly aging societies, to reduce the inequities apparent from contemporary versions of globalization and to manage the consequences of emerging e-business trends.

Current trends to downsize public sector organizations as part of a dedication to the goals of smaller government will force these entities to work towards further financial and operating efficiencies, to achieve higher service quality standards, and to emulate private sector organization structures. However, it is unlikely that these organizations will gain any relief from the responsibilities of public governance.

The tensions between organizational theory and public sector practice can be expected to continue to be a source of academic review and public scrutiny as specific events cause governments and public sector organizations to reexamine their role. For example, re-regulation of certain markets, a requirement for public investment in previously privatized businesses, and a range of potential responses to the combined effects of e-business and globalization are examples of major challenges for public sector organizations. These demands are likely to increase as the structure of these organizations continues to be affected by the combined effects of downsizing, outsourcing, budgetary restrictions, and the decentralization of organization structures inherent in the moves towards electronic government.

See also: Bureaucratization and Bureaucracy, History of; Organizations and the Law; Organizations, Sociology of; People in Organizations; Public Goods: International; Public Interest; Public Management, New; Quangos in Politics;

Bibliography

- Barney J, Hesterley W 1996 Organizational economics: Understanding the relationship between organizations and economic analysis. In: Clegg S, Hardy C, Nord W (eds.) *Handbook of Organization Studies*. Sage, London
- Copleston F 1962 *A History of Philosophy: Greece and Rome*. Image Books, New York, Vol. 1, p. 1
- Corkery J 1989 Challenges of the next twenty years: Future directions for training. *Australian Journal of Public Administration* **48**: 185–95
- Drucker P 1969 *The Age of Discontinuity*. Heinemann, London
- Friedman M 1962 *Capitalism and Freedom*. The University of Chicago Press, Chicago
- Gore A 1999 Transforming governments in the 21st century. Global forum on reinventing government. (<http://www.21stcentury.gov/conf/gore1.htm>) Washington, DC
- Her Majesty's Government 1999 *Modernising Government*. The White Paper. (<http://www.official-documents.co.uk/document/cm43/4310/4310.htm>) Her Majesty's Stationery Office, London
- Kernaghan K 2000 *Rediscovering Public Service: Recognizing the Value of an Essential Institution*. Institute of Public Administration of Canada, Toronto, Canada
- Keynes J M 1964 *The General Theory of Employment Interest and Money*. Macmillan, London
- Mauk T 1999 The death of bureaucracy. *Public Management (US)* **81**(7): 4–8
- Metcalf L 2000 Linking levels of government: European integration and globalization. *International Review of Administrative Sciences* **66**(1): 119–42
- Nutt P 1999 Public-private differences and the assessment of alternatives for decision making. *Public Administration Research and Theory* **9**(2): 305–50
- Olson M 1982 *The Rise and Decline of Nations*. Yale University Press, New Haven, CT
- Osborne D 1994 Reinventing the business of government: An interview with change catalyst David Osborne. *Harvard Business Review* May–June: 132–43
- Parkinson C N 1958 *Parkinson's Law or the Pursuit of Progress*. Murray, London
- Peters B G 1996 *The Future of Governing: Four Emerging Models*. University Press of Kansas, Lawrence, KS
- Rainey H 1991 *Understanding and Managing Public Organizations*. Jossey-Bass, San Francisco, CA
- World Bank 1992 *Governance and Development*. World Bank, Washington, DC

G. Callender

Public Sphere and the Media

To analyze the media from the perspective of the public sphere is to focus upon the media's relationship with democracy. The concept of the public sphere has been used to describe both the historical development of modern, post-Enlightenment forms of representative democracy and the role of the media and public

opinion within them, and has also served as a normative concept with which to pass judgement on the current performance of the media from the perspective of the creation and maintenance of democracy.

In recent years, the public sphere perspective has become dominant in thinking about the media and their political role, by providing, in a return to the central questions of enlightenment political thought, a bridge between Marxist theories of ideology and the liberal 'free press' tradition. It provides a critique of Marxist theories of ideology by revalidating the specificity of the political, by giving due weight to the emancipatory potential of liberal bourgeois concepts of free assembly and debate, and by shifting attention from worker to citizen. On the other hand, it provides a critique of an abstract, property-based 'free press' model by stressing both the systemic social barriers to entry and the manipulative distortions to which its discourse is subject.

This return has focused on concepts of the public sphere, civil society, citizenship, and identity. It has in particular pitted those, such as Habermas, who see the emancipatory project of the enlightenment as unfinished business and wish, in a movement of immanent critique, to hold liberalism to its emancipatory ideals, against a range of communitarians and postmoderns. The latter, in the name of various versions of identity politics, see liberal, rights-based political theories as inherently repressive and their emancipatory and universalizing tendencies as bogus (Gray 1995).

The public sphere approach has provided, against the background of the turn away from Marxism associated with the collapse of 'actual existing socialism,' an alternative to theories of dominant ideology or hegemony as an explanation for the coincidence, within what used to be called 'bourgeois democracies,' of growing social inequality and political apathy on the one hand with the relative stability and increasingly consensual, nonideological nature of representative party politics on the other. Its emphasis on discursive practices and on communicative action as central to democratic practice and legitimization fitted well with the wider 'linguistic turn' in the human and social sciences. At the same time, its use of the spatial metaphor of sphere, and its stress on the necessary institutional foundations for the realization of those citizen rights to free expression and debate central to democratic theory, addressed the questions arising from both the perceived empirical reality in the mature democracies of an increasingly media-saturated, image-driven, public relations-oriented politics and the task of constructing democracy in both state and civil society in newly democratized countries with little or no historical traditions of democracy to call upon.

The concept of the public sphere was introduced into media analysis by Jurgen Habermas (1989) in *The Structural Transformation of the Public Sphere*. In Habermas's original formulation, the term was designed to describe a historical process whereby in

the shift from premodern feudal, absolutist rule to forms of modern representative democracy public opinion, based upon publicly available information and debate, developed as both a check on, and as a source for the legitimacy of, government.

The public sphere was both a set of institutional spaces—newspapers, lecture halls, coffee houses, etc.—and a set of discursive rules. Drawing upon a Kantian heritage that links freedom to a personal autonomy grounded in the exercise of public reason, Habermas and his followers have stressed the role of the public sphere as a site within which the formation of public opinion, and the political will stemming from and legitimized by such opinion, is subject to the disciplines of a discourse, or communicative ethics, by which all views are subjected to the critical reasoning of others. At the same time, a democratically legitimate public sphere requires that access to it is open to both all citizens and all views equally, provided only that all participants are governed by the search for general agreement. This general model is then applied as a normative test against which the performance of contemporary media, in terms of political effects and democratic potential, can be judged in terms of either the critical rationality of their discourse or in terms of the range of views or speakers granted access.

The public sphere is added as a fourth sphere to Hegel's tripartite division of society into family, civil society, and state. The public sphere is distinguished from the family, the sphere of intimate privacy, although it is the creation of this sphere, particularly under the influence of the literary public sphere that creates both the norms of noncoercive, literate, interpersonal discussion and the associated subjectivities that form the foundation of the public sphere. The public sphere is distinguished from civil society, and especially the market, where, following Hegel, citizens pursue their private interests, since although the private economy provided the resources for the public sphere, and citizens participated in the public sphere as private persons (i.e., access was not governed by socially conferred status or the possession of public office), the purpose of debate within the public sphere was common agreement on matters of general interest. On the other hand, it is distinguished from the state since it was there that citizens held the state to account and guided its actions.

Habermas then argued, following in part the general Frankfurt School argument of the Dialectic of Enlightenment, that there was a pervasive crisis of democratic legitimacy in the modern world because public spheres had been squeezed almost out of existence by both market and state. On the one hand, the market had increasingly come to dominate the institutions and practices of public communication through the commodification of information and opinion for sale on a mass consumer market and through the associated manipulative use of advertising. On the other hand, the state, ironically in

response to political pressures from below, had become increasingly both economically interventionist and manipulative of public opinion. The result was a public opinion and related politics increasingly dominated by advertising and public relations, what Habermas termed the refeudalization of the public sphere—a shift back from a politics of rational, participatory debate to a politics of display.

The public sphere approach has been criticized on three main grounds. First that its procedural rules are too rationalist—that the persuasive use of rhetoric can never, and indeed should never, be excluded from political communication—and that to impose such discursive norms in practice excludes from the public sphere as illegitimate not just a range of culturally specific discursive forms but those who do not possess the cultural capital required to mobilize those discursive forms. Here we find the criticism that the model of procedural rationality being deployed is in fact a model of a certain intellectual practice and thus excludes, in a movement of symbolic violence, those who are not members of the social group who are the carriers of that practice, whether conceived as intellectuals or as the white male bourgeoisie, and that it leads to a privileging of certain genres of media. These rationalist assumptions may take news and overt political coverage to be ‘serious’ while taking entertainment-oriented programs to be a sign of ‘dumbing down’ or ‘refeudalization.’ This neglects the role that entertainment plays in the formation of publics and as a site for the development of an understanding of issues of public importance.

Second, the public sphere model of procedural rationality is criticized for drawing the distinction between public and private in such a way as to exclude both key groups of citizens, e.g., women, or key matters of potential public political concern, e.g., the regulation of domestic, intrafamilial, or sexual relations. It is in this sense that the famous slogan ‘the personal is political’ mounts a challenge to the way in which public sphere theory has drawn the distinction between the private and public spheres. And this in turn will affect the ways in which we think about the ethical issues relating to invasions of ‘privacy’ by the media and to what is or is not an appropriate matter for public discussion in the media.

Third, it is criticized on the grounds that it has valued general agreement around a set of universal values or norms, however rationally discursively arrived at, which, its critics argue, derive in turn from a liberal model of proceduralism, abstract individual rights, and ethical neutrality which prioritizes the just over the good and thus denies difference and thus the inevitability of perpetual normative conflict within modern societies. This in its turn, so critics argue, leads to an overcentralized model of the public sphere incompatible with the politics of identity in multicultural societies, whereas what is needed is a decentralized model of multiple public spheres ex-

pressive of each distinct collective identity, culture or form of life.

This debate concerning the proper relationship between social communication and politics can be traced back to the Enlightenment and the basic paradox that modernity posed; the paradox of what Kant called the ‘unsocial sociability’ of human beings. We might think of this as the inherent tension between liberty and fraternity. In essence, modern democracy is about how we handle the relationship between individual freedom and moral agency on the one hand and the necessary and unavoidable social norms and structures within which alone such freedom can be exercised on the other. As Rousseau put it in *The Social Contract*, ‘the problem is to find a form of association ... in which each, while uniting himself with all, may still obey himself alone, and remain as free as before.’

The political problem to which the public sphere tradition of analysis addresses itself is then two-sided and much of the problems with current uses of the term public sphere, and criticisms of it, derive from a confusion about which of the two sides of the dilemma are being addressed.

Here we are brought back to the Hegelian concern for *Sittlichkeit* and to critiques of Habermas’s public sphere approach from a communitarian and neo-Aristotelian perspective. Here the debate over the public sphere has to be seen as part of the wider debate between advocates and defenders of the theory and practice of rights-based liberalism and the various proponents of communitarianism and identity politics (Benhabib 1992, Guttman 1994).

The communitarian critique leads to the opposite solution to that of social contract theory and the liberalism based upon it. Now political values are social before they are individual. Politics is embedded in, and ideally expresses, a set of pre-existing social values, or a way of life, and the role and legitimacy of the state, or the public realm, is then to foster and uphold those communal values and defend that way of life. The citizens find their identity in, and give their loyalty to, not a set of abstract rights, but to a way of life, an ethos, that embodies a set of moral values. From this perspective liberalism, and the particular forms of democracy that it supports, is but one way of life among many possible alternatives. From this perspective, the role of the media is then to foster and defend this shared way of life, sometimes described, in the nationalist model of communitarianism, as national identity or culture, rather than to serve as a value neutral space for the critical rational debate over values.

The communitarian position on the public/private divide, and the approaches to both politics and the media that stem from it, is in a sense contradictory. The core of identity politics is the call for the recognition in the public realm of values hitherto deemed private, both in the sense of being excluded

from the public gaze and from public debate, but also in the sense of stemming from private group interests and identities rather than from a generally shared interest and identity, and thus for the acquisition of rights that recognition of these values as public entails, while at the same time drawing for its evaluative arguments upon a range of sources which must exclude the very concept of the public, and its liberal valuation of rights and the recognition and equal treatment of the diversity of private interests, that they are demanding.

Following Benhabib, we can distinguish between the integrationist and participatory versions of communitarianism. The integrationist strand seeks *Sittlichkeit* in an attempted return to the moral certainties and social unities of traditional societies. Here everything is in a sense public; the values that motivate people and give them their social anchorage are derived from and shape a whole shared way of life. In a theocracy there is no room for the individual or the private. Indeed, this strand in communitarianism criticizes the liberal tradition precisely for creating, and philosophically justifying as the highest good, social arrangements that separate political or public spheres of action and value on the one hand from private spheres of action and value on the other, to the impoverishment, in their view, of both (Macintyre 1998).

On the other hand, the participationist strand wishes to build its way out of the alienation and formalism of liberalism by both accepting the conditions of modernity which are liberalism's starting point—namely posttraditional societies and reflexive individuals—while at the same time arguing for a refounding of political communities on the universalization of the discourse ethic and its practices. The more postmodern end of communitarianism, which links to so-called new social movements and the politics of identity, makes this attempt by, in effect, advocating the fragmentation of societies into small-scale or specialized communities of interest or identity, each with its own public sphere. They tend to see more universalistic and unified definitions of the public sphere as repressive of difference and thus antidemocratic. What is at issue here is both whether within a polity we should be looking at one or multiple public spheres and how we think of the relation between public and private. An integrationist communitarianism and the politics of recognition that stems from it demand a unified public sphere and place continuous pressure on the existence of a meaningful private sphere. Its aim is a one-to-one fit between a unitary set of values, a single public sphere (if this term any longer has meaning here) and a single polity. If diverse communities exist within a single territory or polity, the aim becomes one of political and cultural fragmentation, not co-existence and the toleration of diversity. It is clearly incompatible with the exercise of public reason in the Kantian sense and with media practices

and institutional forms in harmony with such an ideal.

The case of participatory communitarianism and the identity politics that stems from it is more complex. Here the problem arises from the ambivalence of the liberal value of tolerance vis-à-vis the public/private divide. On the one hand toleration can be taken to mean the acceptance by public authorities of a range of practices and beliefs by defining them as private and outside the realm of public regulation, e.g., religious observance. Such toleration rests, as the communitarian critics rightly point out, on a prior judgement that the practice or belief in question is not in any sense a threat to the public weal or interest. This form of toleration can be rejected by certain advocates of the politics of recognition because, it is argued, it implicitly downgrades, and thus in some sense fails to recognize fully, the importance or centrality of the practice or beliefs in question. On the other hand, we can also understand toleration as giving public recognition to and bringing into the public realm people, practices, or beliefs to which by so doing we signal that we give equal value to those already so recognized and with which we are prepared to live and argue within a shared culture and polity. Thus we need to distinguish between two kinds of demands for recognition and for multiple public spheres. One set of demands seeks to extend the private, so that their group identities remain unthreatened by the risk of corrosion that participation in the critical discourse and compromise of the public sphere carries.

The second set demands a widening of the definition of what is public to let them into what remains a common arena for critical public debate and decision making with the acceptance of the duties and risks that such entry carries with it. In the former case, what is being demanded in the name of difference is the dissolution of any shared culture, polity, and set of rights and associated obligations. Here the question is how much of such private fragmentation can any society or polity sustain while remaining viable, and whether in fact we will only regard such demands as democratically legitimate where the group identity claimed is itself subject to the disciplines of the discourse ethic. In the latter case, the liberal polity is being asked, rightly, to live up to its ideals and to accept that those values that it considers central to its way of life must always be held provisionally and be subject to the review of critical discourse. Here we need to distinguish between a truly private realm, the irreducible site of individual autonomy, and the multiple public spheres within which we necessarily live and have our being, defined by Benhabib as any occasion or situation where people share views on matters of common interest, multiple public spheres that all contribute to the formation of our identities and which may be more or less democratic in the sense of meeting the requirements of the discourse ethic. The media, ranging from the small-scale, special-interest

media of magazines, newsletters, Internet bulletin boards up to the space for public debate and exposure of views offered by the mainstream mass media, are integral to these multiple public spheres and should be judged in each case on the basis of the identities and practices that they foster. However, these multiple public spheres then need to be distinguished from the political public sphere where discussion is necessarily aimed at the common agreement necessary for concerted action within a unified polity. Research and debate within and around the public sphere perspective now focuses on three issues. The first is the institutional question of the relationship between the political public sphere and the nation-state. Both the actual institutional development of the modern mass media, the public sphere, and representative democracy, and thinking about them, have taken place within a nation-state framework. It is argued that current globalizing trends, both within society at large and within the media, are undermining this nation-state foundation of both democratic politics and its associated political public sphere. The question then is whether this is a process of cultural and discursive emancipation and the beginnings of the construction of a global public sphere or, on the contrary, is a further tightening of the screw of refeudalization which places the key centers of power outside the structures of democratic accountability and rational critical debate. On the one hand, the trends towards globalization within the media can be exaggerated. For reasons of audience taste and interest, most media output, and the organizations that provide it, remain obstinately linked to local, often national, markets, cultures, languages, and political structures. There are few examples of successful transnational newspapers or broadcasting news services. Even in the field of entertainment their spread is limited. This is itself, as the example of the European Union shows, a major barrier to the construction of a transnational public sphere. On the other hand, it can be argued that current developments on the World Wide Web, particularly its use for mobilization, debate, and advocacy by NGOs, are a welcome start in the construction of a global public sphere. The question here is how representative and how public these new Net-based interest groups and associated debates are. What certainly is the case is that the development of global media has begun to create a situation where nation-states, both within their own borders and in their inter-State relations, are having increasingly to take account of global public opinion (e.g., the cases of Tienanmen Square or the Balkans).

Second is the discursive question of how we can both analyze and normatively judge from the perspective of democracy the wide range of rhetorical forms used and identities appealed to within the range of modern media. Here the debate around the public sphere reprises an older debate about mass culture.

The question at issue is whether current developments in media forms, genres, and their associated audiences/publics is a liberatory process of demassification and cultural pluralism on a global scale or, on the contrary, whether it is a process of 'dumbing down' which, in its appeals to irrationality and social fragmentation, undermines the rational/universalistic norms upon which citizenship, and thus democracy, rest. In the press, attention is focused on tabloidization, i.e., the stress on the scandalous and the sensational and the coverage of politics according to the norms of show business; in broadcasting on the shift in the balance of the programming mix in favor of entertainment at the expense of news, current affairs, and political coverage and debate, a shift associated, at least in Europe, with the retreat of public service broadcasting in the face of increasingly intense commercial competition for audiences and revenues. Analysis and debate have focused in particular on the spread of audience participation shows of the Jerry Springer type, of 'docu-soaps' and of a general soundbite coverage of politics and public issues. In this debate there are no easy answers. First, empirically the trends are by no means as clear cut as many critics claim. It is as easy to argue for a rise in standards. Second, it is all too easy to fall into familiar elitist or populist stances, neither of which take due account of the complexity, both actual and historical, of the construction of the identity of citizens or of the rhetorical problems of communicating with the generality of such citizens. There is also often a danger of stretching beyond breaking point the proper and necessary boundaries of politics and the public sphere.

Third, there is the question of the role of the mediators. Public sphere theory finds it difficult to break free from a participatory model of direct democracy. Representation always introduces a potentially distorting element into the communication process between the private citizen and political whole. The public sphere is conceived as a structure for the aggregation of the opinions of individual private citizens through a public debate to which all have equal access. The ideal model is the Athenian agora. The scale, social complexity, and specialization of the modern world make such a model unrealistic. It has clearly never fitted the actual forms and practices of the mass media. It therefore leads to unrealistic demands being made upon them and to an excessively dismissive critique when these demands are not met. It has in particular led to exaggerated claims being made for the democratic potential of the Internet on the grounds precisely that cyberspace is a new agora without mediators or mediation. The challenge, therefore, is to think of a legitimate role for mediators within the public sphere—for journalists, TV producers, bulletin board moderators, etc.—which is neither one of pure transparent 'objective' mediation nor one of the mere pursuit of instrumental pro-

fessional or economic self-interest. This in its turn links to the wider question of the social and political role of intellectuals and experts within modern democratic polities.

See also: Critical Theory: Contemporary; Critical Theory: Frankfurt School; Legal Issues: Public Opinion; Mass Communication: Normative Frameworks; Mass Media: Introduction and Schools of Thought; Mass Media, Political Economy of; Mass Society: History of the Concept; Public Opinion: Microsociological Aspects; Public Opinion: Political Aspects

Bibliography

- Benhabib S 1992 *Situating the Self: Gender, Community and Post-modernism in Contemporary Ethics*. Polity, Cambridge, UK
- Calhoun C (ed.) 1992 *Habermas and the Public Sphere*. MIT Press, Cambridge, MA
- Gray J 1995 *Enlightenment's Wake: Politics and Culture at the Close of the Modern Age*. Routledge, London
- Guttmann A (ed.) 1994 *Multiculturalism*. Princeton University Press, Princeton, NJ
- Habermas J 1989 *The Structural Transformation of the Public Sphere*. Polity, Cambridge, UK
- Macintyre A 1998 *Whose Justice? Whose Rationality?* Duckworth, London
- Peters J 1993 Distrust of representation: Habermas on the public sphere. *Media, Culture and Society* 15 No. 4: 541–72
- Schudson M 1998 *The Good Citizen: a History of US Civic Life*. New York
- Splichal S 1999 *Public Opinion*. Rowman and Littlefield, Lantion, MD

N. Garnham

Public Sphere: Eighteenth-century History

Any reflection trying to understand how the notion of public sphere was constructed in the eighteenth century is necessarily referred to the classical book published by Jürgen Habermas in 1962 (Habermas 1962). According to Habermas, the 'political public sphere' was construed in the eighteenth century (in some places sooner, in others later) as a space for discussion and exchange removed from the control of the state. The new 'public sphere in the political realm,' founded on the exercise of public criticism was thus opposed to the 'sphere of public authority' identified with the exercise of state power. For Habermas, this new intellectual and political 'space' was distinct from the court, which belonged within the domain of state power, and from the common

opinions of the people who had no access to enlightened critical debates. This is why it is possible for him to qualify it as 'bourgeois.'

1. Definitions of the Public

There were different ways of conceiving the construction of the new political public sphere of the eighteenth century. Was it issued from the literary controversies of the seventeenth century and rendered possible by the diffusion of new forms of intellectual sociability? Or was it the extension of the modalities of print communication proper to the Republic of Letters to a greater number of individuals which made possible the transformation of dispersed private persons into a unified public community?

1.1 Habermas's Political Public Sphere

For Habermas such a 'public' supposed, on the one hand, that the various participants were not acting as agents or subjects of the ruler and, on the other hand, that they were considered as equal. The political public sphere thus ignored both the obedience required by the exercise of state power and the distinctions of 'orders' and 'estates' imposed by the traditional social hierarchies.

No limit could subject the critical exercise of reason, and no domain was to be forbidden to it. Overcoming the division instituted by the cartesian methodical doubt between opinions which could legitimately be submitted to examination and obligatory credences and obediences, the members of the new public sphere considered that no domain of thought, including the political and religious ones, was to be removed from critical and reasonable judgement.

This judgment was exercised within the different forms of sociability which had been previously erected as a tribunal of aesthetic criticism: the salons, the cafés, the clubs. In the seventeenth century, this new public challenged the traditional monopoly of the court, the official academies, or the connoisseurs in such matters. It was enlarged during the eighteenth century thanks to the multiplication and diffusion of the periodicals, and the success of new forms of association: the book clubs, the reading societies, the masonic lodges. But it supposed also the exclusion of all those who lacked the competence which made possible the participation into a critical community welcoming only the persons who had sufficient wealth and culture.

1.2 Kant's Public Use of Reason

Habermas's perspective was based on one of the most famous text of the eighteenth century, Kant's *What Is*

Enlightenment? published as an article in the periodical *Berlinische Monatsschrift* in 1784. But it also distorted it. In his text Kant proposed a new way of conceptualizing the relation of the public to the private. He equated the public exercise of reason with judgment produced and communicated by private individuals acting 'as scholars before the reading public' and he defined the public as the sphere of the universal in opposition to the private considered as the domain of particular and domestic interests (which may even be those of an army, a church, or a state).

He inverted, thus, the accepted meanings of the two terms by associating the private use of reason with the exercise of an office or with the duties toward the state and by locating the public use of reason within a community without limitations or boundaries. If the private use of reason can be restrained legitimately in the name of public ends without prejudice to the progress of enlightenment, the liberty of the public use of the reason was absolute and has to be respected by the prince himself.

In this same text, Kant shifted the way in which the legitimate limits put on critical judgment should be conceived. Such limits were no longer defined by domains forbidden to methodical doubt; they laid in the position of the subject who exercises his reason. He could be constrained legitimately when he was executing the duties of his charge or of his status, but he was necessarily free when he acted as a member of the 'society of world citizens.'

That universal society was unified by the circulation of written works that authorized the communication and discussion of thoughts without the necessity of physical presence of the different interlocutors. Kant systematically associated the public use of reason with the production or reading of written matter by private individuals. The public was not construed on the basis of new forms of intellectual sociability such as clubs, cafés, literary societies, or lodges, because those associations were modalities of particular, circumscribed, and domestic entities. For him, only written communication was admissible as a figure for the universal.

Kant's conception of the public use of reason was drawn for the notion of the Republic of Letters as it was constructed through correspondence and the circulation of manuscript or printed texts from the late seventeenth century onwards (see Goldgar 1995). Founded on the free engagement of its members, on equality among its interlocutors, and on reciprocity and disinterestedness, the Republic of Letters which united the learned and the scholars provided a model and a support for thinking of the new public sphere. At the same time, it showed the distance separating the universality of the 'society of world citizens' and the restricted social composition of the 'reading public.' In Kant's time, this reading public was not the whole of society by any means, and those capable of writing were even less numerous. It is only when everyone will

be able to exercise critically one's reason that humanity will experience 'an enlightened age.'

1.3 The French Philosophers: Public Opinion vs. Popular Opinions

Kant held the distinction between the public and the people to be temporary, transitory, and characteristic of a century which was a 'age of enlightenment' but not yet an 'enlightened age.' But for many thinkers of the eighteenth century, particularly within the framework of the French Enlightenment, the two notions constituted an irreconcilable dichotomy (see Ozouf 1989, pp. 21–35). Public opinion was defined precisely in contrast to the opinions of the multitude. Public opinion, set up as a sovereign authority and a final arbiter, was necessarily stable, unified, and founded on reason. The universality of its judgments and the constraining self-evidence of its decrees derived from that unvarying and dispassionate constancy. On the other hand, popular opinions were multiple, versatile, and inhabited by prejudice and passion. Such a qualification reveals a strong persistence of older representations of the people considered as subject to extremes, inconstant, and foolish.

When the concept of public opinion did emerge in France around 1750 as the superior authority to which all particular opinions must bow, the distinction between the public and the popular became essential. A new political culture thus took shape which transferred the seat of authority from the will of the king alone, who decided without appeal and in secret, to the judgment of an abstract entity embodied in no institution, which debated publicly and was more sovereign than the sovereign who was obliged to seek to win its approval and support.

This gave acuity and urgency to a fundamental question: how could the people, eliminated from the exercise of critical judgment by their lack of competence, be 'represented' in the new political space? A first answer to this question assigned this function to one or another of the theories of political representation confronted after the mid-eighteenth century. Following Keith Baker, three of the theories were fundamental: the absolutist theory, which made the person of the king the only possible representative of a kingdom divided into orders, estates, and bodies; the judiciary theory which instituted the Parliaments as interpreters of the consent or remonstrances of the nation; and the administrative theory, which attributed the rational representation of social interests to municipal or provincial assemblies founded not on privilege but on property (see Baker 1990).

An alternative model of representation that removed the notion institutional setting—monarchical, parliamentary, or administrative—was offered by the category of public opinion. Substituting the self-

evidence of unanimity to the uncertainties of particular and popular opinions, detached from any form of exercise of governmental and state authority, the concept delegated the representation of the whole nation to those who were able to express its reasonable decrees: the men of letters.

2. Public Opinion: Voice or Tribunal?

Although everyone recognized the existence of public opinion and postulated its unity, there was no unanimous consent on two issues: first, who were its true spokesmen and, second, how was one to evaluate the self-evidence of its judgments? In eighteenth-century France, the answers to such questions were expressed through the metaphors used to designate this new entity: public opinion as a voice to be heard; public opinion as a tribunal which had to be persuaded.

2.1 From Theater to Tribunal

In 1775, in his *discours de réception* before the *Académie française*, Malesherbes forcefully expresses the idea, by then commonly accepted, that public opinion was to be considered a court of justice more imperious than any other. It is a tribunal 'independent of all powers and that all powers respect,' composed by the men of letters who are 'amid the public dispersed, what the orators of Rome and Athens were in the middle of the public assembled.'

There are several arguments contained in this comparison. First, it invested the new 'judges' with an authority that ordinary judges did not have. Their competence knew no bounds and their jurisdiction no limits; their freedom of judgment was guaranteed because they were in no way dependent upon the power of the ruler; their decrees had the force of self-evident propositions. With such a definition of public opinion, the men of letters were invested with the judiciary legitimacy of all traditional powers and with veritable public office.

Reference to this dispersed judiciary function had another meaning, however. It aimed at establishing a connection between the universality of judgments and the dispersal of persons, and at constructing a uniform opinion which, unlike that of the ancients, had no physical location in which it could construe and experience its unity. As for Kant later, it was the circulation of written matter and particularly the printed one which made it possible to envisage the constitution of a unified public in a nation in which people were necessarily separated from each other and formed their ideas individually. By associating the public nature of the written word, vastly increased by the invention of the printing press, with the supreme authority of the judgments pronounced in the new political sphere, Malesherbes (as many others) converted the plurality of reflections and criticisms that

emerged from solitary reading into a collective and anonymous conceptual entity.

From the seventeenth century to the eighteenth century there had been a radical shift in the manner of conceiving the public. In the age of baroque politics the traits that defined the public were the same as those which characterized theatrical audiences: heterogeneous, hierarchized, and formed into a public only by the spectacle in which the most powerful effects were produced by hidden machines and secret maneuvers.

When the concept of public opinion did emerge, it effected a dual rupture. It opposed the art of dissimulation and secrecy by appealing to transparency. Before the tribunal of opinion all causes were to be argued without duplicity. But all citizens were not (or not yet) adept at joining together to form enlightened opinion. If the audience which mingled in the theaters potentially was composed of men and women from all social levels, the public which served as the tribunal to judge literary merits, religious matters, or political issues was more selected and homogeneous. When opinion was thought of more as actor than as acted upon, it became public, and by that token it excluded many people who lacked the competence to establish the decrees which it proclaimed.

2.2 From Audience to Public

Constituting the public as an entity the decrees of which had more force than those of the established authorities supposed several operations. They can be exemplified by the memoirs published in great number by both French lawyers and litigants from 1770 onward. They took the judicial comparison literally and exposed before opinion affairs which were examined by a court of justice. The cases subjected to the secret procedures of justice were thus transformed into public debates, shifting the place in which judgment had to be pronounced (see Maza 1993).

The most fundamental strategy required by such an operation consisted in endowing particular cause with general and exemplary value. The debt that a court noble refused to pay to his bourgeois creditors became an ideal occasion for denouncing unjust privilege, just as the arbitrary imprisonment of a young gentleman was an opportunity to criticize the *lettre de cachet*. To achieve such goals supposed, first, that the secrecy of judicial procedure had to be broken by the circulation of printed memoirs on the largest possible scale, and, second, that a dramatic style or a first-person narrative had to replace the customary legal prose. Universalizing the particular, making public what had been secret, and 'fictionalizing' discourse were the techniques that lawyers used to appeal to opinion and, in doing so, to proclaim themselves the authorized and legitimate interpreters of that opinion.

The traditional, discrete, and exclusive relationship that bound individuals to the king, the guarantor and guardian of domestic secrets, gave way to a totally

different situation in the public exposition of private conflicts. From that point of view, judicial memoirs are the exact inverse of the *lettres de cachet* accorded by the sovereign in response to requests from families interested in concealing familial disorders which sullied their honor. The memoirs displayed what the *lettres* concealed; they expected from the judgment of opinion what the *lettres* hoped to gain from the omnipotence of the monarch; they converted into public issues the scandals that the *lettres* were charged with burying.

It is in this sense that the new public space was built on conflicts proper to the private sphere, assigning political significance to simple familial or conjugal strife. In a parallel manner, the clandestine pamphlets stigmatized the aristocracy, the courtiers, the queen, and finally the king by revealing their corrupt mores. People became accustomed to seeing individual affairs transformed into general causes. This procedure was not restricted to pamphlet literature but also lay behind the lawsuits instituted by rural communities against their landlords or by journeymen workers against their masters. It was the process of 'privatization' in which individuals conquered autonomy and freedom removed from state authority which permitted the very existence of a new public sphere.

Even if, or because, it was defined as a conceptual entity, and not in sociological terms, the notion of public opinion which invaded the discourse of all segments of society—political, administrative, judicial—in the second half of the eighteenth century operated as a powerful instrument for social division and legitimization. The mobilization of the category founded the authority of all who, by affirming that they recognized its decrees alone, set themselves up as mandated to pronounce its judgments. Universal in its philosophical essence, the unified, enlightened, and sovereign public was far from universal in its social composition. The public sphere, emancipated from the domain in which the ruler held sway, thus had nothing in common with the shifting opinions and blind emotions of the multitude. Between the people and the public there was a fundamental difference.

3. *Prohibited Books, Subversive Words*

According to Darnton, this new public was detached from the traditional authorities by a wider and wider dissemination of philosophical ideas. In this process the role played by the great classics of the Enlightenment was perhaps less important than the texts that the eighteenth-century book trade called 'philosophical books': critical pamphlets, scandalmongering chronicles, anticlerical satires, pornographic works. The circulation of this clandestine and seditious literature, which depicted the despotic corruption of the monarchy or the depraved mores of the court,

engendered a change in thought and in collective representations, henceforth detached the subjects from their old loyalties: 'Sedition was hatching. It was instilled in people's minds [...] We know for certain that it was communicated by a formidable instrument: the book' (see Darnton 1991, 1995).

But other historians like Farge have challenged this interpretation which sees the increase in seditious writings in the last two or three decades of the *ancien régime* as the matrix of a desacralization of the monarchy. She affirms: 'Popular opinion did not emerge from the cumulative reading of pamphlets and placards; it was not unilinear and did not base arguments on the sum of what is read' (see Farge 1994, p. 35).

Statements of disapproval or hatred of the sovereign, attacks on his acts, and expressions of a desire to kill him did not arise from the flourishing clandestine literature. Nor did they start with Damiens's attempt to assassinate Louis XV, which, rather than unleashing a proliferation of regicide discourse, persuaded the authorities of the reality of Jansenist and Jesuit plots and set them in pursuit of such literature. For Farge, 'the failed murder of Louis XV had a ready-made public opinion; it tells us more about the monarchy's reactions than about any new and original turn of popular thought' (p. 174).

The symbolic and affective disinvestment which transformed relations with authority when it was deprived of all transcendence was manifested but not caused by the wide diffusion of 'philosophical books.' The erosion of founding myths of the monarchy, the desacralization of royal symbols, and the distancing of the person of the king formed a body of representations which was 'already there,' 'ready-made' when the public sphere arose conceptually and sociologically.

4. *Women and the Public Sphere*

What was the role played by women in this new public sphere? Joan Landes' thesis opposes radically the 'symbolic politics of the emerging bourgeois public sphere (which) was framed from the outset by masculinist interests and assumptions' with the forms of sociability dominated by women in the *ancien régime*, for example, the salon society conceived as 'an alternative sphere of cultural production inside absolutism' (Landes 1988, p. 40). This is the reason why the exclusion of women from political rights during the French Revolution is inscribed by Landes within the logics of the enlightened condemnation of the role that women played within the *absolutist* public sphere: 'the structures of modern republican politics can be construed as part of an elaborate defense against women's power and public presence' (p. 203).

Against such a perspective Goodman argues that women played a central role both in the collective

project of the Enlightenment and in the construction of a new political and critical public space. For her, 'the French Enlightenment was grounded in a female-centered mixed-gender sociability that gendered French culture, the Enlightenment, and civilization itself as feminine' (Goodman 1994, p. 6). In this perspective the salon society did not belong to the old regime absolutist public sphere but had a fundamental role in shaping a 'discursive space' opened to public and political criticism. Between 1750 and 1775 women were essential in the government of such society because they kept within the limits of civility and politeness the tensions which necessarily arose from the free confrontations of opinions and the competitions for intellectual leadership: 'Enlightenment salons were places where male egos were brought into harmony through the agency of female selflessness' (p. 104).

According to Goodman (1994, p. 280), it is during the 1770s that women began to be excluded from the new public sphere. On the one hand, the harmony of Republic of Letters was then destroyed by a series of ferocious discords starting with controversies about Physiocracy. On the other hand, new forms of association (among them, literary societies called *lycées* or *musées* and masonic lodges) either excluded women or gave them a subordinate position. The return to 'masculine self-governance' prepared women's expulsion from the revolutionary political space: 'When the literary public sphere was transformed into the political public sphere in 1789, it had already become masculine; the 'democratic' republic of 1792 would reflect the limitations and exclusions of the Republic of Letters of the 1780s.'

It is possible to discuss such an interpretation. The Freemasonry did not ignore women's presence and gave them an important role in the 'lodges of adoption.' In these lodges many women were involved in the new values coined by the late Enlightenment: the philanthropic activities, the interest for political issues, the friendship between men and women (Burke and Jacob 1996, pp. 513–49). Besides that, it is clear that even between 1750 and 1775, the salon society was always very minority and limited to a restricted number of participants. They cannot be considered as the center of the Republic of Letters.

The collective construction of the new public sphere cannot be reduced either to a unique process or to a sole place. It arose from the intertwining between conceptual novelties and social practices, the alliance between speech and print, and the common (but dissimilar) participation of men and women.

See also: Absolutism, History of; Civil Society, Concept and History of; Enlightenment; Intellectual History; Kant, Immanuel (1724–1804); Public Sphere: Nineteenth- and Twentieth-century History; Space and Social Theory in Geography

Bibliography

- Baker K M 1990 *Inventing the French Revolution: Essays on French Political Culture in the Eighteenth Century*. Cambridge University Press, Cambridge, UK
- Burke J M, Jacob M C 1996 French freemasonry, women, and feminist scholarship. *Journal of Modern History* 68(3): 513–49
- Chartier R 1991 *The Cultural Origins of the French Revolution*. Duke University Press, Durham and London
- Crow T E 1985 *Painters and Public Life in Eighteenth-century Paris*. Yale University Press, New Haven and London
- Darnton R 1991 *Édition et sédition. L'univers de la littérature clandestine au XVIIIe siècle*. Gallimard, Paris
- Darnton R 1995 *The Forbidden Best-sellers of Pre-revolutionary France*. Norton, New York
- Farge A 1994 *Subversive Words: Public Opinion in Eighteenth-century France*. Polity Press, Cambridge, MA
- Goldgar A 1995 *Impolite Learning: Conduct and Community in the Republic of Letters 1680–1750*. Yale University Press, New Haven
- Goodman D 1994 *The Republic of Letters: A Cultural History of the French Enlightenment*. Cornell University Press, Ithaca
- Habermas J 1962 *Strukturwandel der Öffentlichkeit. Untersuchungen zu einer Kategorie der bürgerlichen Gesellschaft*. Luchterhand, Neuwied (English trans 1989 *The Structural Transformation of the Public Sphere: An Inquiry into a Category of Bourgeois Society*. Polity Press, Cambridge, MA)
- Kant E I 1795 *Foundations of the Metaphysics of Morals and What Is Enlightenment (1784)*. Bobbs-Merrill, Indianapolis
- Landes J B 1988 *Women and the Public Sphere in the Age of the French Revolution*. Cornell University Press, Ithaca
- Maza S 1993 *Private Lives and Public Affairs: The Causes Célèbres of Prerevolutionary France*. University of California Press, Berkeley, CA
- Ozouf M 1989 *L'homme régénéré. Essais sur la Révolution française*. Gallimard, Paris
- Roche D 1988 *Les Républicains des Lettres. Gens de culture et Lumières au XVIIIe siècle*. Fayard, Paris

R. Chartier

Public Sphere: Nineteenth- and Twentieth-century History

The closely related concepts of public sphere and public space are rooted in ancient Greek distinctions of public and private and developed in early modern Europe through changing ideas of citizenship, civil society, and republican virtue (see *Civil Society/Public Sphere, History of the Concept*). They refer first and foremost to settings for political discussion among citizens about the order of their lives together, and secondarily to a much wider range of public culture including art, literature, and popular discourse on a host of subjects.

1. Publicness as a Modern Concern

New ideas about public discourse were complemented by development of new communications media, especially those dependent on print; rising literacy and education levels; growth of the state; and

expansion of popular political participation. In this process, the distinction of public and private took on new importance and complexity. On the one hand, the realm of public interaction expanded; cities were the primary setting for this, especially cosmopolitan trading and capital cities. Public spaces appeared literally with coffee houses, parks, theaters, and other places where people who were not bound by private relations gathered and communicated. They also grew metaphorically with printed sermons, pamphlets, newspapers, books in vernacular languages, journals that reviewed them, and other media of public communication. On the other hand, the state also expanded and with it the range of *res publica*, public things which included property held in common and matters of concern to the whole polity. Publicness took on a dual sense, referring both to openness of access and interaction and to collective affairs as managed by the state. The public referred both to the collective subject of democracy—the people organized as a discursive and decision-making public—and as its object—the public good.

The two dimensions were linked in the notion that political debate among responsible citizens was a way to arrive at sound understanding of common affairs. This depended on notions of participation in debate that developed in the realms of science and religion (Ezrahi 1990, Zaret 1999) and literature (Habermas 1962, Hohendahl 1982) as well as politics. Processes of rational-critical debate were held to form educated public opinion as distinct from other forms such as the 'representative publicity' of monarchs appearing before their subjects or the 'mere opinion' of uneducated masses. Interest in such public opinion grew alongside civil society as a self-organizing realm of social relations, and especially with the rise of democracy.

Complementing the growth of the public realm were new senses of the private. The 'privacy' of the family was more sharply affirmed. Women and children were increasingly sequestered in private homes, especially among the bourgeoisie; the modern idea of the individual included the notion of nurturance and maturation in private as preparation for action in public, but this applied in the first instance to men. At the same time, economic activity was increasingly moved out of the household. In one sense, then, going to work meant going out into public, being exposed to the public gaze. In another sense, however, property relations continued to be understood as private in the sense that they were to be managed by individual persons and not the state. The eventual rise of business corporations further complicated the distinction, since these held property as artificial private persons but operated as collective, public actors, especially when shares of ownership were openly available on the market rather than closely held within families.

As the example of corporations and private property suggests, the distinction of public and private was sometimes difficult to sustain. This undermined the

classical notion of the public sphere: 'The model of the bourgeois public sphere presupposed strict separation of the public from the private realm in such a way that the public sphere, made up of private people gathered together as a public and articulating the needs of society with the state, was itself considered part of the private realm' (Habermas 1962, pp. 175–6). As corporations became public actors while still claiming the status of private property, states began to intervene ever more into civil society and even intimate life. These trends joined with the rise of mass media, and especially professions and institutions devoted to the manipulation of public opinion through mass media (advertising, public relations) to undermine the conditions for the effective operation of the public sphere as a source for educated public opinion.

2. History as Decline

For many later analysts, the eighteenth century was a sort of 'golden age' of modern public life (see *Public Sphere: Eighteenth-century History*). Hannah Arendt (1958) theorized 'public' in terms of creative action, the making of a world shared among citizens, and saw the founding of the USA as a crucial example. Jürgen Habermas (1962) idealized eighteenth century English parliamentarianism, newspapers, and coffee house conversation. He presented the public sphere as a realm of civil society in which private citizens could communicate openly about matters of public concern, transcending their particular statuses and addressing the state without becoming part of it. Others have focused on Parisian clubs in the Revolution of 1789 (Amann 1975) or the literary journalism of the German Enlightenment (Hohendahl 1982).

Idealization of the eighteenth and sometimes the early nineteenth century public sphere underwrites explicit or implicit narratives of decline. In Habermas's classic *Structural Transformation of the Public Sphere*, for example, nineteenth and twentieth century public discourse is analyzed in terms of the loss of rational-critical capacity attendant on expansion of scale and the rise of public relations management that incorporated the public into the realm of administered society (see also Horkheimer and Adorno 1944). Such accounts reveal that the notion of public is not only basic to democracy, but linked to an apparent contradiction in democratic ideals. At least in its most prominent modern, liberal forms, democracy seems to promise both a politics based on reason rather than power, and the inclusion of all adults in the political process. The latter, the classic accounts of the public sphere suggest, undermines the former. The nineteenth and twentieth century history of democratic politics is taken to demonstrate this. Public opinion is increasingly 'managed' on the basis of research and manipulation rather than developed in open, rational-critical discourse.

Already in the nineteenth century itself, political theory of many persuasions emphasized the fragility and limitations of the liberal democratic conception of the public. Tocqueville, most famously, argued as early as the 1830s that the democratization of society tended to eliminate the *intermediary* public bodies which had traditionally refined opinion and furnished the individual a collective social identity outside the state. Engaged, politicized publics composed of distinct views and interests could be reshaped over time into mass publics—passive, conformist, and atomized before the state. Tocqueville's fear of the unmediated state would resonate with generations of critics of mass society.

Central to the analyses of historical decline in the public sphere is the idea that as its scope expanded, it was unable to sustain a high level of rational-critical debate. This happened with the rise of various sorts of opinion managers, from pollsters to public relations specialists, and with the institutionalization of political parties as products of negotiations among interest groups rather than arenas for debate about the public good in general. A key phase in this, accordingly to the critics, was the institutionalization of class politics during the late nineteenth and early twentieth centuries. Habermas treated this in terms of negotiation among interests, as distinct from reasoned debate about the public good. Negt and Kluge ([1964] 1993), however, famously emphasized the existence of a proletarian public sphere, rooted in distinctive experiences, and productive of reasoned internal discourse. Indeed, the German Social Democratic Party, like successful workers organizations elsewhere, created a parallel public sphere to that dominated by the bourgeoisie (Eley 1992, see also Katznelson and Zolberg 1987). This faltered principally because of a nationalization of popular politics, not class politics *per se* (Mosse 1975, 1987). Such nationalization was often accompanied by mass spectacle and reliance on mass media. Both did frequently collapse the distinction of public and private, especially in emotive appeals to national unity. Equally important was the rise of two-party electoral systems. Rather than being able to participate actively in public life, late twentieth century citizens of democratic societies were reduced to being invited to respond to plebiscites (Lowi 1985, Ginsburg 1986).

Historical research on the public sphere has questioned both the accuracy of the late eighteenth century ideal and the narrative of subsequent decline. Schudson (1998), for example, has argued that active participation of citizens in public discourse and politics has ebbed and flowed without linear trend. The ideal of the good citizen as an active participant in the public sphere has long been contrasted with the failings of actual citizens. A substantial line of research on 'civic culture' shifted the emphasis away from the problematic of public debate and towards a more general idea of participation and incorporation into

civic life, for example, in voluntary associations (see, in particular, Almond and Verba 1966 and critical analysis in Somers 1993, 1995). This, most researchers argued, grew through the nineteenth and early twentieth centuries. More recently, however, critics of late twentieth and early twenty-first century democracies have emphasized a decline in civic participation (e.g., Putnam 2000).

3. *Participation and Exclusion*

Although openness was basic to the ideology and theory of the public sphere, various forms of exclusion were basic to actually existing publics in the nineteenth and twentieth centuries. Women, for example, had been prominent in late eighteenth century salon society and elite public spheres, but were widely excluded from nineteenth and twentieth century public space; this was mirrored by a commensurate bias in political thought (Landes 1988, Elshtain 1993). Feminist struggles made gender a basic political issue in new ways (Ryan 1990, 1992). In addition, women were prominent in other movements, such as abolitionism in the USA, and may have figured less in subsequent historical accounts than in actual nineteenth century social movements. This may be partly because of different styles of address in the public sphere, including a greater emphasis on narrative and less on argumentation (Lara 1999). Similarly, Blacks were largely excluded from the US public sphere, even after abolition. This encouraged a similar duality: struggles for inclusion and development of parallel public discourse marked by distinctive communicative styles (Diawara 1995).

This has occasioned more general inquiry into 'counterpublics' which contest the apparently neutral dominance of more mainstream publics (Fraser 1992, Warner 2001). This argument is rooted in Negt and Kluge's ([1964] 1993) account of the proletarian public sphere. Members of religious communities, ethnic groups, and the women's movement have been among the many who have constructed counterpublics. During the classic heyday of the late eighteenth and early nineteenth century British public sphere, thus, many artisans and workers were denied participation in the public sphere. They were not simply and unambivalently members of a proletarian public sphere, though they did develop their own media and organizations and to some extent constitute a counterpublic (Gilmartin 1999). They claimed the right to participate in the dominant, unmarked public sphere and challenged those who introduced restrictive measures to make it a specifically 'bourgeois' (or more generally, propertied) public sphere (Jones 1983). The same people who excluded those with less wealth from the public sphere nonetheless claimed it in unmarked form as simply *the* British public.

Accounts of the public sphere by both contemporaries and historians have been tacitly linked to the idea of nation, since nation defined the range of citizens entitled to participate in public affairs and the basis for legitimacy of modern, especially democratic, states (Calhoun 1997). In liberalism's triumphant early phase, experiments in republicanism and constitutionalism placed these issues squarely on the table. They asked how new political structures were to acquire legitimacy and how, within those structures, the plurality of opinion was to be represented. Defenses of free speech and the press, too, signaled an appreciation of the importance of venues in which views could be widely and, in principle, freely circulated. At the same time, the rhetoric of nationalism typically presented the nation not as creature of public discourse but as pre-existing unity, commonly on ethnic or cultural grounds. This is true even, if ironically, in the USA, the nation quintessentially created through the public action of its citizens. A tension running through American history has pitted the capacity for creation of a new political order exemplified by the founders against claims that the defining national essence is already in place and needs only defense from the loyal.

Notions of a vital republican public sphere have been influential especially in American historical studies, visible in the work of J. G. A. Pocock, Bernard Bailyn, Gordon Wood, and more recently, Michael Warner. Here the focus, in Hannah Arendt's terms, is on the capacity of action in public not only to define individual identity but to create the world citizens share in common. Republicanism, implied several things: (a) a concerted effort to reach agreement about the public good, as distinct from the attempt to find the net balance of private, individual goods—the essence of utilitarianism; (b) a sense of politics as a work in progress regarding the appropriate structures and limits of the political, as opposed to a hallowed set of procedures for achieving particular ends—such as the cult of the constitution; and (c) a belief that individuality and identity are only fully expressed and enjoyed through engagement in public life, rather than constituted beforehand in the private realm. These writers emphasize the ephemeral nature of republicanism and the rapid emergence in the nineteenth century of a variously liberal, national and representative (rather than participatory) model of the public sphere. In American Studies, as in Frankfurt School critical theory, this decline is commonly associated with the rise of mass culture—and the blurring of divisions between reasoned political discourse and other social processes of opinion formation.

4. Cities and Media

Transformations of public life were also shaped by the ways in which communications media and the physical public spaces of cities changed through the nineteenth

and twentieth centuries. Richard Sennett (1977) *The Fall of Public Man*, for example, contrasts a vibrant and highly varied eighteenth and nineteenth century urban life (especially among the bourgeoisie) to a reduction in both the occasions for interaction across lines of difference and the psychological ability to manage open relations with strangers well. Citizens retreated, Sennett suggests, into both privacy and the conformity of mass culture.

Urban public spaces anchor face-to-face interaction, which was basic to the classical image of the public sphere, conceived of largely in terms of direct interaction. This is important for a sense of serendipitous contact among people of diverse backgrounds. Parks, cafes, and theaters all inform this. Many of Europe's cities, especially older ones, were distinctive in their pedestrian character and their scale. From the mid-nineteenth century, however, this began to shift. Hausmann's post-1848 construction of the Paris boulevards and destruction of some of the local *quartiers* is taken as emblematic (e.g., by Harvey 1985). Both suburbanization and less pedestrian-friendly, larger-scale designs for newer cities have changed the character of public interaction in Europe, largely in ways familiar earlier in the US. Commentators throughout the nineteenth and twentieth centuries have been concerned for the implications of this for democracy. As Mumford (1938, p. 483) wrote, 'One of the difficulties in the way of political association is that we have not provided it with the necessary physical organs of existence: we have failed to provide the necessary sites, the necessary buildings, the necessary halls, rooms, meeting places ...'

This has been linked to a huge increase in scale. One crucial result is that directly interpersonal relations organize proportionately less of public life; mediations of various kinds have become increasingly important (Calhoun 1988, Garnham 1992, Keane 1991, Thompson 1995). This is partly a matter of communications media *per se*; it is also a matter of growing importance for bureaucracies and other formal organizations. The eighteenth and early nineteenth century public sphere was closely tied to the spread of print media. The nineteenth and early twentieth centuries were the heyday of great urban newspapers; after this, media that transcended locality became increasingly important. First radio and then television fundamentally altered the public sphere. They contributed to a shift in what was publicly visible as well as in how public discourse was organized (Meyrowitz 1985, Thompson 1995). New media shared both information and emotionally powerful images widely. Broadcast media are often charged with furthering the debasement of reason by substituting powerful images for sustained analysis, by appealing to a lowest common denominator in audiences, by blurring the lines between entertainment and critical discourse, and by centralizing control over messages in the hands of a few corporations. Some caution is in order about

generalizations from medium, though, as newspapers produced their own mass forms, with broadly similar if perhaps less extreme effects.

Crucially, the changes in scale and media forms of the twentieth century challenge the classical eighteenth century model of the public sphere, which was grounded in readership on the one hand and directly interpersonal discourse on the other. Any account of contemporary public life must deal with a dramatic increase in indirect, mediated relationships. Transformations of media continue, of course, and the implications of the Internet and computer-mediated communication are important topics of contemporary investigation (Rasmussen 2000). Relatedly, the dominant approaches to the public sphere have emphasized domestic politics within nation-states. Especially during the last years of the twentieth century, however, communications media and the rise of transnational organizations have facilitated growth of an international public sphere (Thompson 1995, Held 1995, Calhoun 2002). To what extent transnational public opinion can steer global society directly, and to what extent only through the agency of state structures, remains a crucial question.

Finally, though this article concentrates on the public sphere in Western European and American history, it is important to note that the nineteenth and twentieth century also saw dramatic and transformative growth in public discourses elsewhere. In late Imperial and early Republican China, for example, the category of 'public' became influential and both directly interpersonal and mediated public spheres were created (Wakeman 1993, Rankin 1993). This was basic to development of the idea of citizen, as distinct from subject, to constitution of notions of public welfare, and to cultural transformations such as the introduction of new vernacular writing. It also contributed to the growth of both nationalist and class politics in the twentieth century. Further transformations followed, attendant on changes in media, politics, and culture, and creating important linkages within the Chinese diaspora as well as with the West.

An unusually vital print-mediated public sphere developed in India. The early history of this is closely interwoven with the history of British colonialism and opposition to it. The history of Indian public life has been an important setting for examining the role of both international power relations and issues of domestic diversity. Not least, questions like the claims of nonreaders in a public sphere dominated by print media have surfaced and shaped not only Indian but international discussions. Broadcast media have played an important role recently in the constitution of a new Hindu nationalism (Rajogopal 2001). In many settings, including in the European countries and the USA as well as both India and China, the development of a public sphere has been closely related to the development of national consciousness, identity, and political contestation.

5. Conclusion

Growth and transformation of public spheres has been a major theme in nineteenth and twentieth century history. It is crucial to the history of democratic politics, including not only electoral and parliamentary regimes, but also broader struggles in civil society and beyond national borders. A dominant historical theme is the inclusion of a broader range of participants in public discourse. This has been coupled with analyses of the implications of growing scale, including mismatch between local—especially urban—public spaces and enlarged range of potential participants, dynamics of inclusion and exclusion, and changing roles of media. A prominent line of analysis suggests a concomitant decline in the rational-critical capacity of public discourse. This is challenged by both research on the limits of earlier public spheres and the strengths of later counter-publics. The importance of the public sphere is in any case not limited to argumentation but includes also the formation and spread of culture, as in the development of national identity.

See also: Civic Culture; Civil Society, Concept and History of; Civil Society/Public Sphere, History of the Concept; Communication and Democracy; Mass Society: History of the Concept; Media, Uses of; Participation: Political; Policy History: State and Economy; Political Communication; Political Discourse; Privacy: Legal Aspects; Public Sphere and the Media; Public Sphere: Eighteenth-century History; State and Society

Bibliography

- Almond G, Verba S 1966 *The Civic Culture*
Amann P 1975 *Revolution and Mass Democracy: The Paris Club Movement in 1848*. Princeton University Press, Princeton, NJ
Arendt H 1958 *The Human Condition*. University of Chicago Press, Chicago
Bellah R 1985 *Habits of the Heart: Individualism and Commitment in American Life*. University of California Press, Berkeley, CA
Calhoun C 1988 Populist politics, communications media, and large scale social integration. *Sociological Theory* 6(2): 219–41
Calhoun C 1997 Nationalism and the public sphere. In: Weintraub J, Kumar K (eds.) *Public and Private in Thought and Practice*. University of Chicago Press, Chicago, pp. 75–102
Calhoun C 2002 Constitutional patriotism and the public sphere: Interests, identity, and solidarity in the integration of Europe. In: DeGreiff P, Cronin C (eds.) *Transnational Politics*. MIT Press, Cambridge, MA
Diawara M (ed.) 1995 *The Black Public Sphere*. University of Chicago Press, Chicago
Eley G 1992 Nations, publics and political cultures: Placing Habermas in the nineteenth century. In: Calhoun C (ed.) *Habermas and the Public Sphere*. MIT Press, Cambridge, MA, pp. 289–339
Elshtain J B 1993 *Private Man, Public Woman*, rev. edn. Princeton University Press, Princeton, NJ

- Fraser N 1992 Rethinking the public sphere: A contribution to the critique of actually existing democracy. In: Calhoun C (ed.) *Habermas and the Public Sphere*. MIT Press, Cambridge, MA, pp. 109–42
- Garnham N 1992 The media and the public sphere. In: Calhoun C (ed.) *Habermas and the Public Sphere*. MIT Press, Cambridge, MA
- Gilmartin K 1999 *Print Politics: The Press and Radical Opposition in Early Nineteenth-Century England*. Cambridge University Press, Cambridge, UK
- Ginsburg B 1986 *The Captive Public: How Mass Opinion Promotes State Power*
- Habermas J 1984 1988 *Theory of Communicative Action*. Beacon Press, Boston, Vols. 1 and 2
- Habermas J 1991 *The Structural Transformation of the Bourgeois Public Sphere: An Inquiry into a Category of Bourgeois Society* [trans. Burger T]. MIT Press, Cambridge, MA
- Harvey D 1985 *Consciousness and the Urban Experience*. Johns Hopkins University Press, Baltimore, MD
- Held D 1995 *Democracy and the Global Order*. Polity, Cambridge, UK
- Hohendahl P U 1982 *The Institution of Criticism*. Cornell University Press, Ithaca, NY
- Horkheimer M, Adorno T 1944 *Dialectic of Enlightenment*. Continuum, New York
- Jones G S 1983 *Languages of Class: Studies in English Working Class History, 1832–1982*. Cambridge University Press, Cambridge, UK
- Katznelson I, Zolberg A (eds.) 1987 *Working-Class Formation: Nineteenth-Century Patterns in Western Europe and the United States*. Princeton University Press, Princeton, NJ
- Keane J 1991 *The Media and Democracy*. Polity, Cambridge, UK
- Landes J 1988 *Women and the Public Sphere in the Age of the French Revolution*. Cornell University Press, Ithaca, NY
- Lara M P 1999 *Moral Textures: Feminist Narratives in the Public Sphere*. University of California Press, Berkeley, CA
- Lowi T 1985 *The Personal President*. Cornell University Press, Ithaca, NY
- Meyrowitz J 1985 *No Sense of Place*. Oxford University Press, New York
- Mosse G L 1975 *The Nationalization of the Masses*. Cornell University Press, Ithaca, NY
- Mosse G L 1987 *Masses and Man: Nationalist and Fascist Perceptions of Reality*. Wayne State University Press, Detroit, MI
- Mumford L 1938 *The Culture of Cities*. Secker and Warburg, London
- Negt O, Kluge A [1964] 1993 *The Public Sphere and Experience*. University of Minnesota Press, Minneapolis, MN
- Putnam R 2000 *Bowling Alone: The Collapse and Revival of American Community*. Simon and Schuster, New York
- Rajogopal A 2001 *Politics after Television: Religious Nationalism and the Reshaping of the Indian Public*. Cambridge University Press, Cambridge, UK
- Rankin M 1993 Some observations on a Chinese public sphere. *Modern China* 19(2): 158–82
- Rasmussen T 2000 *Social Theory and Communication Technology*. Ashgate, London
- Ryan M 1990 *Women in Public: Between Banners and Ballots, 1825–1880*. Johns Hopkins University Press, Baltimore, MD
- Schudson M 1998 *The Good Citizen. A History of American Civic Life*. Free Press, New York
- Sennet R 1977 *The Fall of Public Man*. Knopf, New York
- Somers M 1993 Citizenship and the place of the public sphere: Law, community, and political culture in the transition to democracy. *American Sociological Review* 58: 587–620
- Somers M 1995 What's political or cultural about political culture and the public sphere? Toward an historical sociology of concept formation. *Sociological Theory* 13: 113–44
- Thompson J 1995 *Media and Modernity*. Stanford University Press, Stanford, CA
- Tocqueville A de [1840] 1844 *Democracy in America*. Scribners, New York
- Wakeman F 1993 The civil society and public sphere debate: Western reflections on Chinese political culture. *Modern China* 19(2): 102–38
- Warner M 2001 *Publics and Counterpublics*. Zone Books, Cambridge, MA
- Yeo E J 1998 *Radical Femininity: Women's Self-Representation in the Public Sphere*. Manchester University Press, Manchester, UK
- Zaret D 1999 *Origins of Democratic Culture*. Princeton University Press, Princeton, NJ

C. Calhoun

Publishing as Medium

Publishing, in its most general sense, refers to the act of making a work of words or images known to the public. Book publishing encompasses a range of activities through which an author's manuscript is turned into a printed and bound book which can be made available to an audience of readers. These activities are carried out by networks of organizations and individuals who coordinate production, and the promotion and distribution of books to their potential publics.

1. History and Description

In the ancient world, there were a variety of ways in which the reproduction and circulation of texts occurred. Evidence from China, India, Greece, and Rome suggests that publishing was carried out under the auspices of religious, political, and commercial bodies; in some places, individuals copied and traded books among themselves. During the Middle Ages, publishing in Europe was centered in the monasteries where manuscripts were reproduced primarily for religious purposes. Following the twelfth century, scribes associated with European universities took over this task, serving the clergy as well as selling books to new readers among officials, the nobility, professionals, and merchants. But with the appearance of the printing press in 1450, book publishing and bookselling became organized into an industry, its members striving to sell books for profit to a growing and increasingly far-flung audience.

Before the seventeenth century in Europe, and the nineteenth century in North America, the book trade tended to remain undifferentiated and unstandardized. While conditions varied in different regions, the functions of publisher, printer, and bookseller were often melded into a single individual who printed and sold books at the request (and payment) of authors. Nevertheless, the trade in books would rarely have been enough to support the printer who often published a newspaper as well as books, and who derived a large portion of his income from printing bills, stationery, and other jobs. As the market for books grew, a more complex division of labor gradually emerged, with different segments of the industry responsible for publishing, printing, and selling. Along with this, many publishers and booksellers began to specialize in specific types of books, such as texts or medical works.

Within the modern book industry, the publisher is central to the entire process of getting books from author to reader. Publishing firms select, edit, and prepare manuscripts for production, and also assume the financial risks involved in bringing a book into print. It is the publisher who interacts and contracts with the author (or the author's representative, the agent), and it is also usually the publisher who determines a sales strategy and how a book will be marketed. Following the pattern of other industries, a separate wholesaling sector has taken over from publishers some of the functions of warehousing and filling orders. A more differentiated retail sector has also developed, as books are sold through both exclusive bookstores and retailers whose primary business is not in books, through educational, employment, and voluntary organizations, and through mail order and electronic venues.

Since the early twentieth century, the publishing industry has been characterized by a degree of specialization that corresponds in large part to the market for particular kinds of books as well as to channel of distribution. The most visible form of publishing is in trade books, those titles directed at a general audience. However, both textbooks and professional titles compose a significant part of the overall book market, and in many developing countries textbooks make up the single largest segment. Other important segments include scholarly and religious books. As these various divisions between book sectors developed during the twentieth century, different organizational and sales practices also appeared within their corresponding publishing houses. These divisions now appear to be breaking down as merger activity has brought together dissimilar publishing houses under a single roof.

2. *Publishing as a Field for Analysis*

The world of book publishing has been viewed in rather contradictory ways: as a rarefied environment of literary connoisseurs, and as a profane site where

expression is reduced to a commodity; as an industrial backwater, and as a central player in the spread of capitalism. Perhaps these contrasting images are related to the way in which publishing, as a field for analysis, belongs to no single group. Those who work in the industry, journalists, and scholars from several disciplines have all been influential in shaping ideas about publishing's significance.

2.1 *Historical Perspectives*

Befitting lives spent among writers and books, numerous publishers, both prominent and obscure, have penned reminiscences of their activities. To some degree, this set the tone for accounts of the book business since the nineteenth century. Many such accounts are really more histories of publishers than of publishing, describing the rise and fall of various publishing houses and the personalities who directed them.

A different kind of scholarship emerged with the tradition known as 'the history of the book,' which many date to the publication of Febvre and Martin's (1958) *L'Apparition du Livre*. Febvre, who was a prominent member of the *Annales* school, and his co-author described the early publishing and bookselling scene in Europe, while also attempting to account for the book's impact on the spread of various ideas and cultural movements. For some, such as Darnton (1982), the field of research that has evolved since Febvre and Martin's book is equivalent to the social and cultural history of print communication. In this view, the object is to understand the role books have played in society at various times and in various places, and to understand how books have interacted with other social developments and processes. Darnton's proposal is also methodological in nature, suggesting that the book be studied in terms of a communication circuit that runs from author to publisher to printer to shipper to bookseller and finally, to reader.

Although the program outlined by Darnton has been influential among book historians, other historians have taken a different approach by emphasizing the history of printing and the subject of bibliography, that is, the study of books as physical objects. Bibliography is concerned with describing books and identifying their attributes, such as paper, type, and bindings, so as to be able to distinguish different editions or printings. This kind of examination can also be used to provide information about particular publishers and manufacturers, or about authorial intention (see, for example, McKenzie 1986). Frequently, the underlying purpose of such studies also differs from that emphasized by the history of the book. The intent is to make the book itself the central focus of attention rather than using the book as a means for discovering other social patterns.

2.2 Social Science Perspectives

While historians have long recognized the book as an important area of study, social science research on book publishing has been intermittent and surprisingly sparse considering social scientists' interest in other forms of mass media. Instead, some of the most detailed descriptions and critiques of publishing have been written by people directly associated with the publishing industry. Of course, books and book publishing tend to be central to the academic's work, and so, despite the relative lack of empirical research, have not escaped academic scrutiny altogether. The state of the publishing world is taken to be a matter of public concern, and scholars and other public intellectuals frequently address a general audience when discussing the significance of developments in publishing for print culture. Throughout the twentieth century, such discussions have often reflected the assumptions of mass culture theory. Commentators question whether publishing practices that cater to a mass market leave room for the quality books that sell in moderate numbers, and whether increasing commercialism and centralization will bring about a standardization of the available literary fare (see, for example, Dutscher 1954).

As this suggests, the curiosity of publishing tends to lie in its dual nature as a commercial enterprise on the one hand, and as a producer of literary objects associated with artistic, educational, and religious values on the other. Indeed, the perceived difficulty of reconciling these two dimensions of publishing—often described as the conflict between culture and commerce—is probably the central problem analysts of the book industry have confronted.

Beginning with work by Escarpit (1958), a relatively small group of sociologists has tried to dissect the workings of the publishing industry. These studies draw on organizational theory (sometimes called the production of culture perspective) which, especially during the 1970s and 1980s, was being applied to various media firms and industries. This perspective begins with the assumption that publishers are similar to other kinds of bureaucratic organizations, facing the same problems of how to reduce uncertainty and risk (Hirsch 1972). Scholars in this tradition have provided specific accounts of how books are acquired and distributed, and how publishing strategies vary according to size and type of firm. Common to much of this work is a stress on the gatekeeping role that publishers play, meaning the ability to determine what ideas and literary forms will make it into print and thus be able to enter the marketplace of ideas.

Congruent with trends in mass media and popular culture studies, more recent researchers have reacted against theories of mass culture and hegemony which posit audiences as easily manipulated and degraded by mass marketed literature. In this view, scholarship needs to grant the reader a more privileged place in

analyses of publishing, in terms of both acknowledging the reader's active role in choosing and finding meaning in reading material, and in understanding how conceptions of the audience shape the work of book professionals. Moreover, this theoretical stance questions the very distinction between popular and quality books. Instead, some suggest that the fear of commercialism has been primarily a cover for the discomfort of cultural elites who see their authority over cultural matters challenged (Long 1985–86, Radway 1997). Similarly, it has been argued that publishers play an important role in the production of literary value (Verdaasdonk and van Rees 1991).

3. Themes in Publishing Research

Changes occurring within the publishing industry during the last few decades of the twentieth century shaped many of the questions posed in publishing research. Renewed interest in publishing has been concurrent with its integration into large, multinational, entertainment conglomerates. Researchers have been especially attentive to the consequences of concentration, the significance of publishing as a global phenomenon, and the convergence between books and other entertainment and information media.

3.1 The Dynamics of Concentration

The American situation perhaps best illustrates certain trends in the industry. During the post-World War II era, most major New York publishing houses were transformed from private or family-owned companies into public corporations. Various social changes contributed to this development. With the post-War education boom, many publishers desired to meet a new demand for printed matter by expanding their existing educational divisions or founding new ones. This coincided with the aging of heads of publishing houses who wished to protect their heirs from inheritance taxes. Issuing public stock appeared to be an attractive proposition, a way to both raise capital and guarantee the long-term future of the firm.

Along with the move towards corporatization, publishing houses became favored acquisition targets by electronics firms in the 1960s. Although these buyers later divested their publishing properties, merger and acquisition activity was renewed in the mid-1970s. These years saw the formation of conglomerates with holdings in film, broadcasting, music, newspapers, and magazine publishing as well as books. Book publishers would not add much in terms of revenues, but entertainment conglomerates hoped that book content could be reproduced across a variety of media, and cross-promotional strategies could be

employed. Similar waves of merger and acquisition activity were seen in the next two decades. By this time, foreign-owned companies were often doing the acquiring, including media giants such as Bertelsmann, News Corporation, and Pearson. The result has been the consolidation of the nation's largest publishers into a handful of companies which also include American-owned Viacom, AOL Time Warner, and Disney. Concentration at the level of distribution happened somewhat later. While chain booksellers began to grow in the 1960s with the establishment of standardized outlets in shopping malls, their decisive dominance of the bookstore market did not occur until the 1990s with the appearance of the superstore.

The effects of concentration have been a constant source of debate. Along with longstanding concerns about declining quality and the homogenization of reading material, these developments raise questions about changes in the organizational culture of the book business. Attempts to rationalize the publishing process, invest heavily in promotion, and exploit new markets have altered the balance of power within the industry, diminishing the autonomy of editors while giving increased clout to accounting, marketing, and subsidiary rights divisions. At the same time, publishers have grown dependent on a few large bookseller customers who subsequently play a more authoritative role in the industry.

Complicating the issue of concentration is that while a declining number of publishers account for a large share of book sales, the absolute number of publishers, along with titles published, has steadily risen. Some argue that these various trends result in much diversity at the margins of the industry, and a less elitist conception of books and publishing at the center. For others, the concentration of control of the book business means that a small number of for-profit companies have undue power over what individuals can and will read.

3.2 Globalization and National Identities

The role that the circulation of books plays in maintaining or undermining specific identities has long been an important theme in publishing research. The relatively new concept of globalization is readily applied to the publishing industry. On the one hand, trends in ownership mean that multinational companies with far-flung holdings are now common among the major players. But the integration of international markets is the primary way in which publishing can be conceived as a global phenomenon. The selling of rights to publish or distribute books in foreign territories has become one of the most important sources of revenue for major publishers. The trade in books goes in many directions, but disproportionately travels from Europe and especially the USA to the rest of the world.

This raises both theoretical and policy questions about national sovereignty and the maintenance of national identities. In Canada, for instance, the overwhelming presence of first British and then American books has hampered the growth of a national publishing industry. While government policy has helped to preserve Canadian-controlled firms, foreign-controlled publishers in Canada remain more profitable. Also, many Canadian publishers can survive only because they have the exclusive right to sell American books, the popularity of which is attested to by their dominance of best-seller lists in the English-language market. Even in Quebec, where French publishers play a strong role, American translations are plentiful. Many argue that this situation has helped to further Americanize Canadian culture, diffusing American tastes, styles, and values through the medium of popular books.

Economic agreements between nations to eliminate free trade barriers have in the past often excluded cultural products, with books in particular singled out for exception. There are significant obstacles to maintaining a viable publishing industry in less developed countries, and as the Canadian situation demonstrates, even highly developed nations find it difficult to make the economics of publishing work without protective measures. With the USA and the European Commission trying to eliminate the cultural exception in trade negotiations, these issues will probably move more to the forefront of publishing research.

3.3 Unyoking Content from Print

The international trade in books has been greatly dependent on the state of copyright law, and a significant body of research has examined how copyright and its breach have affected the publishing industry. However, the development of new technologies along with publishing's integration into entertainment conglomerates have raised new issues concerning the control of literary property. Content is now easily transportable across media, and entertainment firms attempt to create synergies, as it has been termed, between book, film, television, music, magazine, and electronic divisions. This introduces a new level of instability into the publishing process. On the one hand, traditional arrangements for the control of intellectual property are challenged when content is so mobile, with publishers and authors in conflict over who will own the new forms that emerge. On the other hand, new technologies, including electronic books and on-demand publishing, where single copies are printed as needed, are utilized by distributors, authors, and software companies as well as publishers. Such developments suggest that what it means to engage in publishing may be more difficult to define in the future.

See also: Audiences; Internet: Psychological Perspectives; Literacy and Illiteracy, History of; Literacy and Orality; Literary Texts: Comprehension and Memory; Mass Communication: Empirical Research; Mass Communication: Normative Frameworks; Mass Communication: Technology; Mass Media: Introduction and Schools of Thought; Mass Media, Political Economy of; Textbooks

Bibliography

- Altbach P G, Hoshino E S (eds.) 1995 *International Book Publishing: an Encyclopedia*. Garland Publishing, New York
- Coser L A, Kadushin C, Powell W W 1982 *Books: the Culture and Commerce of Publishing*. Basic Books, New York
- Darnton R 1982 What is the history of books? *Daedalus* 111: 65–83
- Dutscher A 1954 The book business in America. *Contemporary Issues* 5: 38–58
- Escarpit R 1958 *Sociologie de la Litterature*. Presses Universitaires de France, Paris [1971 *Sociology of Literature*. Frank Cass, London]
- Febvre L P V, Martin H-J 1958 *L'Apparition du Livre*. Éditions Albin Michel, Paris [1990 *The Coming of the Book: the Impact of Printing 1450–1800*. Verso, London]
- Hirsch P M 1972 Processing fads and fashions: an organization-set analysis of cultural industry systems. *American Journal of Sociology* 77: 639–59
- Horowitz I L 1991 *Communicating Ideas: The Politics of Scholarly Publishing*, 2nd edn. Transaction, New Brunswick, NJ
- Lane M, Booth J 1980 *Books and Publishers: Commerce Against Culture in Postwar Britain*. Lexington Books, Lexington, MA
- Long E 1985–86 The cultural meaning of concentration in publishing. *Book Research Quarterly* 1: 3–27
- McKenzie D F 1986 *Bibliography and the Sociology of Texts*. British Library, London
- Powell W W 1985 *Getting Into Print: The Decision-making Process in Scholarly Publishing*. University of Chicago Press, Chicago
- Radway J A 1997 *A Feeling for Books: The Book-of-the-month Club, Literary Taste, and Middle-class Desire*. University of North Carolina Press, Chapel Hill, NC
- Verdaasdonk H, van Rees C J 1991 The dynamics of choice behavior towards books: the role of cultural products in cultural behavior. *Poetics* 20: 421–37

L. J. Miller

Punishment, Comparative Politics of

1. The Minimally Punishing Democratic State

The most striking development in twentieth-century penal politics was the emergence in the advanced democracies of the minimally punishing state. The

development was striking because it occurred during one of the largest crime waves in recorded history—a tidal wave ignored by all democratic political systems for one or even several decades. The modern democratic state was able to dismiss the demand for a state response to the crime wave constituted by the annually mounting numbers of thousands, even millions of crime reports. Most democratic governments developed surprising capacities *not* to punish in spite of the victim's entitlement to punishment implied by penal codes. (No penal code contains a warning label to victims that the punishments it threatens will be applied at most to three percent of those arrested and one percent of those who violate its terms.) Attrition rates for criminal cases became extremely high in all democratic systems, from the apparently most punitive to the least. A decision seems to have been taken in most democracies that the state will never punish (or at least imprison) much more than 100 per 100,000 people at any time, no matter which or how many crimes are committed. It has been possible for democratic states *not* to enforce the criminal law on a massive scale, although enforcing the penal code against violence and predation is popular among voters.

2. The Convergence on Minimal Punishment

2.1 The Convergence

The facts of convergence on minimal punishment levels are clear. In 1940, most of the currently advanced democracies had capital punishment; by 2000, only two of them still inflicted the death penalty and only on a much reduced scale. In 1960, there were in the advanced democracies on average 11 prisoners (for all offenses) per reported robbery; in 1990, only three, and, without Japan, only one. In 1960, the range in punishment lay between New Zealand with 30 prisoners per robbery and Norway with 24, on the one hand, and the USA with only two; by 1990, the range (excluding Japan) lay between three prisoners per robbery in California and Austria, and fewer than 0.5 in Italy and The Netherlands. Between 1960 and 1990, the democratic state largely relinquished the threat of death punishment and reduced to nearly token levels the threat of prison punishment.

2.2 Elite Conversion to Leniency

The decision to ration punishment rested on what had become, by the 1960s, a gap between the lenient mentalities of political elites and the more traditionally punitive, if also ambivalent, mass public. This gap was best explained by Norbert Elias, who traced what he

called 'the civilizing process'—the diffusion of anti-violence mentalities among and by the middle and upper classes since the Middle Ages, especially since the monopolization of legitimate violent punishment by the state. (Elias unfortunately did not theorize the strong decivilizing processes on full display in the twentieth century, including the decivilizing results of overcivilizing the state: Elias 1982.)

2.3 The Emergence of Penal Sentimentalism

The medieval elite–mass gap widened with the top-down spread of the ideas of the eighteenth-century Enlightenment and of nineteenth-century Romantic, liberal, and socialist movements—humanitarianism, human perfectibility, and the social responsibility for crime. Gradually created was a powerful group of aristocratic and bourgeois reformers who changed elite attitudes toward the lower classes, the poor, the disorderly, and eventually even minorities. The revision of (elite) attitudes toward the poor, the sick, the aged (as well as vagrants, drunks, addicts, juveniles) which eventually brought about the welfare state was accompanied by a closely related revision of attitudes toward criminals—which brought about the minimally punishing state. Both the poor and criminals were seen as irresponsible victims of society—to be helped, not chastened. And the penal sentimentalism of enlightened seculars soon also came to characterize the elites still faithful to Christianity but less and less faithful to traditional hard-line Pauline concepts of punishment, either in this world or the next.

2.4 Leniency in Spite of Marxian Capitalism

The growing leniency of elites in nineteenth- and twentieth-century capitalist democracies was directly contrary to Marxist predictions; indeed, Marx had to be 'turned on his head' to explain the diminution of punishment under democratic capitalism and its exasperation under communism. Ironically, one of the few serious Marxist works on penality, Rusche and Kirchheimer's *Punishment and Social Structure* (1968), was also one of the first and few accounts of 'the general tendency toward leniency' under capitalism; it purported to use Marxist concepts (the relationship between capitalism, labor supply, and punishment) to explain penal evolution generally, and the trend toward penal leniency, in particular. It argued grandly that 'Every system of production tends to discover punishments which correspond to its productive relationships... The use or avoidance of specific punishments, and the intensity of penal practices (are) determined by social forces, above all by economic and fiscal forces' (Rusche and Kirchheimer 1968). But to account for the trend toward leniency under

capitalism, Rusche and Kirchheimer were forced to be eclectic, rather than dialectic, to explain leniency in terms only incidentally related to evolving productive relationships. They invoked instead the rising standards of living for the lower classes; a 'consequent' decline in criminality; the development of 'penal liberalism' and social insurance ('welfare statism'); declining conviction rates; and successful penal reform campaigns by bourgeois professionals. (This particular package of explanations would not, of course, explain why, a decade after World War II, when crime became rampant, crime-based punishment rates under capitalism continued to decline (Rusche and Kirchheimer 1968).)

2.5 Leniency in Spite of Social Diversification

In contrast to Marx, but with the advantage of greater hindsight, Durkheim correctly predicted (in 1902) growing leniency under democratic capitalism, but he attributed leniency to a supposed fracturing of the collective conscience due to the division of labor (Durkheim 1902/1978). Actually, social diversification has not destroyed consensus on the evils of violence and predation. The factor most pertinent to the decline of punishment, in fact, was not social and economic but political: the ability of elites (as theorized by Michels, Mosca, and other 'democratic elitist' theorists) to ignore rank and file opinion in otherwise democratic structures.

2.6 Leniency in Spite of Democratic Competition

Convergence on leniency was indeed paradoxical—given democratic competition and the persisting, even growing, crime-induced popularity of punishment among voters. The paradox is possibly explained by:

- (a) party elites' ability to manipulate electoral and political agendas, to collude in order to suppress conflict over penal policy;
- (b) the ability of the Left for decades to ignore constituent demands for security;
- (c) conservative complicity in the antipunishment project;
- (d) antipunishment bias in the ideological-group lobby and the ideology production system (ideological or advocacy groups vastly outweighed interest groups);
- (e) stratification of the media, with elite media reinforcing antipunishment ideological hegemony and rewarding elites for antipunishment action;
- (f) the ability of elites to freeze and even to reduce prison capacity during an historic crime wave;
- (g) declining detection rates (though more and more suspects were being produced by the police);

(h) the formation of a powerful international antipunishment movement enabling lenient domestic elites to invoke international norms, praise, and interventions on behalf of their domestic agendas;

(i) the ability of lenient political elites sharply to reduce prosecution and prison-sentencing rates, even in compulsory-prosecution systems and in inquisitorial justice systems, with low evidentiary barriers to conviction; and

(j) the stigmatization within elite circles of responses to demands for punishment as 'populist', or worse.

Thus penal policy has often become one of many policy areas in which public influence, especially in Europe, has often been weak—in spite of the public's interest, known preferences, and voting power.

Convergence on leniency also occurred in the advanced democracies in spite of periods of high unemployment, increasing ethnic diversity, a drugs pandemic, periods of unremitting terrorism, the flourishing of organized crime, and a sevenfold increase in serious crime (using robbery as an indicator).

3. The Differential Development of State Capacity Not to Punish

But if all democracies have become minimally punishing states, some have become less minimally punishing than others. The token-punishment state has been politically most viable in Europe and least viable in the USA. The factors behind the convergence on leniency operated with differential effect in Europe and the USA, resulting in somewhat differential state capacity *not* to punish. Among 90 national and subnational jurisdictions (all in advanced democracies), the strongest predictor (in cross-sectional analysis) of per capita incarceration rates is the serious crime rate (using murder or robbery rates). The USA has had higher per capita prison rates because it has had much serious crime to be punished: it is (marginally) harder not to punish in high-crime jurisdictions than in low-crime jurisdictions.

3.1 Does Party Count?

One of the best predictors of per capita prison rates, after crime, is the strength of the Left since 1960. Although its historical record for leniency in penal matters is quite checkered (Robespierre, Lenin, Stalin, Mao, Pol Pot, and Castro come to mind), the Left—at least after World War II and in the democracies—became the major political force behind the antipunishment movement. The Left became solidly against capital punishment, sympathetic to criminals, defensive of prisoner rights, and hostile to spending on prisons, especially after the myth of therapeutic prisons was shaken (but far from destroyed). Ironically, the Left was slower than the Right to heed the

demands of its own constituents—the predominantly lower-class and minority victims of crime—but the Left had major stakes in trying to prove that the welfare state could replace the punitive state in preventing crime. As it became clearer in the 1970s and 1980s that criminals thrived in even the most perfected welfare states (Sweden and The Netherlands, for example), that working class voters could be even angrier about crime (especially minority crime) than conservative voters, that the crime wave denied for so long by criminologists was real, and that left-voting middle-class penal liberals were not likely to desert to the Right, the rhetoric of the Left and eventually some of its policy stands began to correspond more closely to widespread public sentiments of insecurity and ineffective justice (Tham 1999).

Moreover, in spite of the correlation between Leftism and low per capita punishment rates, the creation of the minimally punishing state has required, as noted above, the complicity of conservatives. The lowest imprisonment levels in their respective postwar state histories occurred under Reagan in California, Agnew in Maryland, and Wallace in Alabama; the lowest levels of federal imprisonment occurred under Nixon; the lowest imprisonment levels occurred under religious party hegemony in The Netherlands and Italy, and under Catholic auspices in Ireland; the leading historical examples of decarceration (Rutherford 1984) occurred under Liberal and Tory auspices in England and Liberal Democratic Party auspices in Japan. The role of conservatives in the antipunishment movement is, of course, an old story. Capital punishment abolition, after all, was carried out by enlightened despots in the eighteenth century; by the czar and other monarchs in the nineteenth century; and by nineteenth-century bourgeois liberal oligarchs in The Netherlands, Italy, and Portugal. Abolition after World War II was proposed in West Germany by the extreme Right and supported by the Christian Democrats in West Germany and Italy. Abolition in France in 1981 was supported by Gaullist and Giscardian elites. Restoration of capital punishment in the UK and Canada has repeatedly been blocked by Tory cabinet ministers.

4. Punishment and State Formation

We should not, however, exaggerate the differences among the advanced democracies. Even in the USA public tolerance for penal sentimentalism lasted well into the 1970s and early 1980s. And the subsequent backlash that doubled and tripled US prison populations only brought the prisoner–robbery ratio (1.8 in 1970, 1.6 in 1990, 3.5 in 1997) into line with European ratios—about the same as Sweden in 1970 or Austria in 1990, and nowhere near the 11 prisoners per robbery average for the advanced democracies in 1960.

The modern state was built by substituting state for private punishment. Bentham could write without fear of contradiction: 'No punishment, no government; no government, no political society' (Bentham 1831, p. 7). Many democracies of the twentieth century set out to disprove Bentham by showing that a state could protect its citizens and preserve order with few sanctions involving life or liberty. They succeeded in preserving the state's monopoly of *legitimate* violence—and proving the electoral viability of the minimally punishing state—but often at the cost of allowing the level of *illegitimate* violence to return to near medieval (i.e., prestate) levels.

See also: Crime and Class; Crime and Ethnicity (Including Race); Crime: Sociological Aspects; Criminal Justice, Sociology of; Criminal Law and Crime Policy; Drugs: Illicit Use and Prevention; Imprisonment: Sociological Aspects; Prison Life, Sociology of; Prisons and Imprisonment; Punishment: Social and Legal Aspects

Bibliography

- Bentham J 1831 *Jeremy Bentham to his Fellow Citizens of France on Death Punishment*. Heward, London
- Cusson M 1990 *Croissance et décroissance du crime*. 1st edn. Presses Universitaires de France, Paris
- Durkheim E 1902/1978 Two laws of penal evolution. In: Traugott M (ed.) *Emile Durkheim on Institutional Analysis*. University of Chicago Press, Chicago pp. 153–80
- Elias N 1978–82 *The Civilizing Process*. Blackwell, Oxford, UK
- Enguéléguélé S 1998 *Les Politiques Pénales*. L'Harmattan, Paris
- Garland D 1990 *Punishment and Modern Society*. Clarendon Press, Oxford, UK
- Hood R G 1996 *The Death Penalty: A Worldwide Perspective*. 2nd rev. and updated edn. Clarendon Press, Oxford, UK
- Jankovic I 1984 Socialism and the death penalty. *Research in Law, Deviance and Social Control* 6: 109–37
- Jeschek H-H 1983–4 *Die Freiheitsstrafe und ihre Surrogate im deutschen und ausländischen Recht*. Nomos, Baden-Baden, Germany
- Lynch J P 1987 *Imprisonment in Four Countries*. Bureau of Justice Statistics, Washington, DC
- Rusche G, Kirchheimer O 1968 *Punishment and Social Structure*. Russell and Russell, New York
- Rutherford A 1984 *Prisons and the Process of Justice: The Reductionist Challenge*. Heinemann, London
- Tham H 1999 Lag & ordning som vänster projekt? Socialdemokratin och kriminalpolitiken. In: 8 *Reflektionen om Kriminalpolitik*. Swedish Council for Crime Prevention, Stockholm
- Törnudd P 1993 *Fifteen Years of Decreasing Prisoner Rates in Finland*. Oikeuspolittinen tutkimuslaitos, Helsinki, Finland
- Young W, Brown M 1993 Cross-national comparisons of imprisonment. In: Tonry M (ed.) *Crime and Justice: A Review of Research*. University of Chicago Press, Chicago, Vol. 17, pp. 1–43

Zimring F, Hawkins G 1991 *The Scale of Imprisonment*. University of Chicago Press, Chicago

R. C. Fried

Punishment: Social and Legal Aspects

1. Punishment in Modern and Postmodern Society

As recently as the 1970s, this article would have almost certainly situated punishment in modern society as the culmination of an evolution toward greater rationality and humanity. This vision of penal progress influenced both reformers and scholars of punishment.

The past was represented in Enlightenment tracts against bloody scaffolds and dark dungeons. The present was embodied in reform oriented prisons rooted in scientific studies of human behavior. The prison represented a space of public control (as against the tradition of private 'dungeons') in which the state could be held accountable for its treatment of offenders. It offered a penalty, the deprivation of freedom, uniquely suited to the core values of modernizing societies (mobility, responsibility, personal gain). Imprisonment was also a punishment that could be set precisely to establish a scale of deterrence, or varied in practice to allow individualized treatment.

The prison did not altogether replace earlier modes of punishment (Spierenberg 1984). Corporal punishment, fines, and the death penalty, the mainstay of penal technique before the nineteenth century, remained, although in diminished and modernized form. In the twentieth century, new forms of community sanction, probation and parole, were introduced, designed to elaborate the prison's capacity for surveillance while minimizing its tendency to isolate the offender from the normalizing influences of the community at large.

The limitations on twentieth century penology were long seen in the context of a future that saw the elimination of residues of private vengeance seeking, and the perfection of scientific techniques for resocializing 'delinquent' subjects. This program, known as 'correctionalism' or 'rehabilitation', grew from nineteenth-century approaches to moral reform anchored in religion, especially Protestant strands of Christianity. Beginning in the late nineteenth century, this was secularized (if only in some respects) into a social scientific penology that promised a range of interventions from psychological therapy to community action to secure the normalization of criminal offenders. Even dystopian visions, like Anthony Burgess's *A Clockwork Orange*, assumed that authorities would acquire the capacity for transforming the behaviors and motivations of deviant individuals.

Social theorists from Durkheim (1964) to Foucault (1977) saw penal progress as signaling a fundamental break in social organization (Garland 1990). Durkheim suggested that more 'primitive' societies were unified by themes of shared identity expressed in strict penal laws and reinforced in rituals of severe punishments. More modern societies were unified by a complex division of labor expressed in civil law and reinforced in varied individual relationships. Durkheim's theory suggested that the proportion of penal to civil law in a particular society formed a rough and ready index of its modernization. Max Weber, writing in the same period, saw in the evolution of prisons and penal law the process of rationalization that he had identified as the crucial pattern of social change in modern societies (Garland 1990). At the other end of the twentieth century, Foucault (1977) saw the prison as an exemplar of a new technology of power that became more generalized in the nineteenth century, that is the disciplinary organization of space for the purpose of governing potentially unruly masses of people. Although rejecting the humanitarian impulse as the primary motivation behind the rise of the prison, Foucault saw punishment as moving toward forms of exercising power that created fewer economic costs and political resistances. Where Durkheim emphasized the normative integration produced by the division of labor, Foucault saw modern society as evolving subtle forms of surveillance and coercion to establish conformity. Thus, while Foucault rejected the substantive normative optimism regarding 'progress' in punishment and society, he recognized the modern approach as distinctive and representative of an increase in the capacity and security of power.

The narrative of penal progress began to unwind in the 1970s (Rotman 1995). Ideas of rehabilitation and deterrence that dominated modern thinking about punishment for more than two centuries have not disappeared, but rather are viewed more skeptically and compete with a broad array of conceptions, some openly directed at vengeance. These developments have not been unique to punishment, but resonate with a growing skepticism about science, rationality, and reformist government that some have identified with the idea of postmodernity (Beck 1992). While modern punishment evidenced a persistent aspiration to progress beyond itself and to become a form of education, therapy, or pastoral care, postmodern punishment has embraced punishment as a good in itself and as a model for governing in other fields like education, healthcare, and welfare (Simon 1997).

These changing cultural sentiments are reflected in an increase in penal severity throughout many societies since the 1970s (Stern 1998). The United States has been far and away the leader in this new pattern. At the end of the twentieth century, some three percent of adult residents of the United States were in some form of correctional custody, up from one percent in 1980 (Maguire and Pastore 1998). The post-World War II

trend toward abolishing the death penalty has also become much more complicated since the 1970s. The post-1989 collapse of Communist governments in Europe and Asia, and the end of Apartheid in South Africa unleashed a wave of abolitions. In the same time period, however, a number of countries that had temporarily ceased executions resumed them (Hood 1996). The United States is again the leading example. After a brief judicially imposed moratorium on executions in the early 1970s (Zimring and Hawkins 1986) the death penalty has returned to use in 38 of 50 states and overall numbers of those executed have returned to the levels of the 1950s.

2. The Origins, Meanings, and Functions of Punishment

Punishment, as that term is used here, refers to all legally regulated practices intended to denounce, harm, control, or reform specific individuals convicted of violating a criminal law or comparable norm. This is a broader definition than that used by many courts (for instance, in the United States), where many kinds of disadvantages or burdens imposed on those convicted of crimes are not considered punishments, including forfeiture of property linked to crime and requirements for certain types of offenders to register their address with police. At the same time, it is a more limited definition than might follow from popular understandings of punishment which can include illicit acts of vengeance and self imposed forms of discipline (Hudson 1996).

At the end of the twentieth century a wide variety of penal practices were in use, including the death penalty, imprisonment, fines, and various forms of supervised release under conditions of limited freedom in the community. Societies vary enormously with respect to the relative proportion of these punishments. Popular sentiment has been growing for a return to penal practices of the past, such as flogging and symbolic markings aimed at shaming the offender.

There is also a growing set of penal practices outside the formal limits of the criminal law. These include forfeiture of property linked to criminal activity, confinement of those deemed mentally abnormal and dangerous (e.g., sex offenders), deportation of aliens deemed a threat to security, and mechanisms to notify private individual and community organizations of the presence of convicted sex offenders in their proximate vicinity. Forms of civil punishment date back to the Middle Ages but have become more common in recent years.

For interpreting trends in penal practices it is useful to distinguish two further terms: 'penology' and 'penality.' Penology is often defined as the 'science of punishment' and can refer to the rigorous study of penal techniques by professionals in both scholarly and institutional settings. While this reflects the

prestige of science in modern society, it is too narrow to capture the range of nonscientific discourses about punishment, both earlier and today. Penology here refers to any organized and systematic discourse about punishment. 'Penalty' is a term that has been applied usefully to thinking about the entire constellation of penal practices and penologies in operation in a society at a particular time (Garland and Young 1983).

Any particular practice of punishment or penology has a genealogy that can in principle be traced back to an origin. Likewise, we can speak of the meaning or function intended for a specific penal practice or projected by a particular penology. It is safe to assume, however, that any society's penalty will reflect a wide range of different penal practices, institutions, and penologies with quite different origins, meanings, and functions. As the philosopher Nietzsche pointed out at the turn of the nineteenth century, 'the concept of "punishment" presents, at a very late stage of culture ... not just one meaning but a whole synthesis of "meanings"' (Nietzsche 1994, p. 57). This is surely even more true a century later (O'Malley 1999).

Since the nineteenth century, philosophical, legal, and social science writers have 'cannonized' four 'appropriate' functions of punishment: deterrence, prevention (or incapacitation as it is more commonly referred to today), rehabilitation or reform, and retribution. Deterrence as a systematic theory of punishment or penology originates with the influential Enlightenment penal thinker Beccaria (1996) and was elaborated by Jeremy Bentham (Hudson 1996). Deterrence views punishments as evils that can only be justified by their effect on the rational calculus of the free and responsible individuals. The ideal of deterrence in this perspective is to threaten just enough 'pain' or 'disutility' in response to the proscribed act that a rational chooser who considers both the 'pleasure' or 'utility' of the crime and the likelihood of being caught and convicted will choose to forbear. Deterrence presumes a common perception of the disutility of punishment and what might be thought of analogically as an acoustically consistent social space in which subjects, as it were, receive deterrent signals from the state. Despite considerable criticism over the decades of both of these core presumptions, deterrence has proved one of the most enduring penologies whose logic can be found in recent trends like mandatory sentences.

Reform or rehabilitation means using the punishment process to identify and to transform pathological features in the character or motivation of an offender that have led to the offense, with the goal of reintegrating the offender into the community as a productive (or at least nonoffending) participant. Like deterrence, rehabilitation is committed to the possibility of reducing crime by inducing the offender to change his or her pattern of behavior. But while deterrence presumes that offenders share with non-offenders a capacity for rational calculation of utility

that the penalty seeks to invoke, rehabilitation assumes that offenders are deviant in their subjective features and seeks to normalize them.

Beginning in the early nineteenth century, reform was associated first with religious penitence and later with rigorous military-like discipline, surveillance and control (Rothman 1990, Foncault 1977). The penitentiary style prisons constructed from the late eighteenth century on were designed to support reform through isolation, compulsory contemplation, and intense surveillance tied to rewards for cooperation and sanctions for lack of cooperation. The cellular structure helped frame the individual as the crucial focus of penal art and lent itself to the comparative evaluation upon which modern criminological science has been based.

Twentieth-century rehabilitation took its cue from the rise of scientific criminology in the same period that emphasized the roots of crime in pathological individual anomalies. While much of the initial scientific criminology focused on seemingly irreparable biological anomalies (which lent itself to the eugenic aspect of prevention), its ethos and later much of its substance fostered the quasiexperimental approach that guided twentieth-century rehabilitation with its emphasis on individual diagnosis and treatment.

Prevention or incapacitation received systematic treatment in the late nineteenth and early twentieth century in response to the rise of scientific criminology. Punishment, on this theory, works by making it impossible for a given offender to commit crimes while he or she is confined. In this sense prevention achieves its crime control effects by removing the offender from the community either temporarily or permanently. Earlier in the twentieth century, this was framed primarily in eugenic terms. Imprisonment or execution prevented offenders from passing on criminal traits to future generations. Eugenically oriented preventive penology suffered the same cultural decline that eugenics did generally during the second half of the twentieth century, due to its perceived link to the Holocaust and the generally genocidal strategies of the Nazis (Duster 1990). In the 1980s, however, in response to the rise of crime rates that started in the 1960s and the seeming failure of rehabilitation methods, penologists became interested in the incapacitative effect of prisons. On this theory, the crimes an offender would have committed had he or she been free constitute the benefits of prevention, regardless of whether the offender resumes criminal behavior upon release (assuming that imprisonment does not actually lead to an increase in the frequency or severity of crime).

The first three of the four common legitimate functions of punishment are often referred to as 'utilitarian' because they are justified primarily by their hopes for reducing crime. A fourth function, retribution, generally is associated with the idea that punishment serves a moral function by inflicting

suffering on those who deserve to suffer for having violated the moral rights of another. The core issue in retribution from that perspective is that of proportionality, that is, making sure that the punishment is proportionate to the moral wrongfulness of the criminal act. Retribution is often confused with revenge since both emphasize the suffering of the offender and do not emphasize the goal of reducing crime. But while revenge highlights the individual right of the victim to carry out a reprisal, retribution highlights the moral values of society as a whole and views crime as an affront to those values.

Retribution received its most famous and influential statement by the philosopher Immanuel Kant. Kant was so adamant that punishments be targeted at (and only at) moral desert that he suggested it would be morally necessary to carry out all punishments even if organized society was about to dissolve and had no future concern for the criminal behavior of offenders (Kant 1974). During much of the nineteenth and twentieth centuries retribution was taken to be a powerful sentiment in society, but it received little elaboration from penologists more interested in pursuing utilitarian leads. Only in the last third of the twentieth century, with the collapse of enthusiasm for rehabilitation, did a new penological discourse on retribution revive under the slogan 'just deserts,' (Von Hirsch 1976). With severe punishment politically popular in many countries there has been a renewed effort by philosophers and legal theorists to provide new foundations for retribution.

These four oft-repeated functions or justifications of punishment refer necessarily to the idea of legitimate or appropriate punishment. Many other functions of punishment have been articulated by critics and by sociological observers of penalty. For example, sociologists in the Marxist tradition have suggested that the key function of punishment is to help regulate labor markets for the stability of capitalist enterprise (Rusche and Kirchheimer 1968), or to produce obedience to and even loyalty toward exploitive and authoritarian governments (Thompson 1975, Hay et al. 1975). Sociologists in the Durkheimian tradition suggest that punishment has the function of increasing the sense of solidarity among members of community. Others have even suggested that the function of punishment is to produce a class of delinquents convenient for manipulating society from below through the control of informants and provocateurs (Foucault 1977). Such functions are necessarily latent, since making them an explicit purpose of punishment would delegitimize penal institutions. A third kind of function also excluded from modern penal thought is not quite so hidden: punishment should serve as an instrument for private vengeance for the victim or catharsis for the public. As shall be addressed later, 'populist punitiveness' (Bottom 1995) has become a major feature of contemporary, post-modern penalty.

It is tempting to treat the history of modern punishment as a shift in dominance among the four justifications. Deterrence seemed to prevail for much of the nineteenth century, followed by rehabilitation, then a resurgence of retribution, and by the end of the twentieth century, by prevention. While helpful in capturing moments of shifting power within the overall effort to coordinate and control penalty at any one time, this approach is likely to disguise the significant practical overlap between the established functions and to cut off observation and analysis of different logics entering the contemporary field of penalty. Likewise, the general discussion of the four functions treats them as eternal verities rather than as a cluster of responses to a specifically modern framing of the problem of punishment since the eighteenth century. Enlightenment critics attacked monarchical penalty as both excessive and ineffective. They set out to develop penal techniques that were simultaneously appropriate to the new status of individuals as moral if not political equals while effectively checking crime. Each of the four major functions of punishment offers a vision of how punishment can provide crime control effectively or at least in a way that is consistent with the political status of the individual as a free subject.

3. The Triumph of Modernist Penology

In the aftermath of the Enlightenment and the age of Revolution, penology was largely a debate between proponents of deterrence, primarily lawyers and politicians who saw the whole population as the primary subject of penal sanctions, and proponents of normalization who saw the individual offender as the subject. By the 1920s establishment opinion was well set in favor of increasing the rehabilitative capacity of the penal system, even at the expense of deterrence. The ubiquity of three particular penal practices reflects the spread of this approach. Juvenile justice was the first great step in a global pattern toward reform and rehabilitation followed by the most advanced societies in the late nineteenth and early twentieth centuries (Rothman 1980, Garland 1985). In creating a distinct penal system for youth, these societies reflected both the influence of scientific ideas on penal practice (the recognition of adolescence as a distinct phase of life was being promoted by the rising field of psychology), and the belief that at least some offenders should be handled in a nonharmful way. Juvenile courts, the heart of these juvenile systems, in turn showcased two other features that would become synonymous with modernist penology, probation and the indeterminate sentence.

The crucial legal invention of probation was the 'probation officer,' an agent of the court who presumably gathered knowledge of the offender's individual circumstances by conducting close research into the offender's biological, psychological, and

sociological circumstances. As a punishment for both minor and serious offenses, probation subjected offenders to the supervision of a probation officer, charged to direct their rehabilitation and empowered to enter their homes, to search them, to interview their neighbors and co-workers, and to arrest them for violations. The indeterminate sentence meant that most of those convicted of serious crimes like robbery or assault were given a formally quite lengthy sentence with the expectation that an administrative agency, usually known as the 'parole board,' would exercise discretion to release the inmate from prison and allow him or her to serve some remaining years under probation-like supervision in the community (by parole agents) (Simon 1993).

The establishment of probation and indeterminate sentencing programs reflected a willingness to de-emphasize both deterrence and retribution as penal values. An indeterminate sentence (e.g., 5 years to life for a burglary) left the potential offender free to dream that he or she would get out in five. Juvenile justice, probation, and parole all lowered deterrence by diminishing the severity of the expected sanction, since all three involved reducing the amount of time that would otherwise be served in confinement. Moreover, in all three institutions, the legal authority to punish was grafted onto an elaborate decision-making mechanism incorporating not only legal but social and psychological facts.

By World War II most industrial societies operated some version of this modernist penology. The Soviet Union under dictator Joseph Stalin adopted a particularly humanist version of this penology (albeit undercut by the penal violence of the state security organizations). Of the three systems engaged in World War II only the Axis powers repudiated modernist penology, adopting penological ideas consistent with fascism central concerns with racial unity and populist excess (Evans 1996). The victory of the Allies in World War II secured a place for modernist penology in the legal and political structure of the post war world. For the next 25 years the rehabilitative logic of modernist penology enjoyed unprecedented ideological support and investment. As the European economies struggled to rebuild and the United States enjoyed surging affluence, the underlying presumption of modernist penology, that society had a significant stake in the adequate adjustment of even its most recalcitrant members, had broad currency.

Modernist penology also enjoyed unprecedented legal support after World War II due to the genocidal use of penal-like sanctions by the Axis powers. As the victorious Allies began to construct an international legal order to assure that such actions would never be repeated (an aspiration that remains unfulfilled), modernist penology became a constitutive element. With its focus on the whole circumstance of the individual and on rehabilitating rather than eliminating those the state defines as enemies, modernist

penology reaffirmed core elements of human rights ideology and offered a normative check within the penal organs of the state against that organ becoming an adjunct to genocidal crimes of the state. For example opposition groups powerfully resisted the death penalty in the post-World War II international legal order. Abolition and limitation of capital punishment has been a central issue of concern since the Enlightenment, but the Nazi atrocities underlined the potential for execution to become mass state policy.

4. The Crisis of Modernist Penology

The dominance of modernist penology began to falter across Europe, Australia and North America during the 1970s. Skepticism about the capacities of scientific penology to control crime, combined with greater demands for government accountability, shattered the consensus that punishment was a kind of technical state service aimed at resocializing offenders (Garland 2001). Punishment as a painful harm to the guilty subject, an embarrassing residue to modern penology, returned as a positive product of the system.

The delegitimation of modernist penology spread initially by the political left that had traditionally supported it. This was an outgrowth of the same human rights ideology that also enshrined modernist penology. The emphasis on equal treatment, especially in the framework set by racial discrimination in the United States, made suspect the tremendous discretion enjoyed by administrative decision makers under modernist penology.

In the 1970s these kinds of considerations led to a wave of critical discourse on modern punishment comparable in quality and consistent in ethos with the Enlightenment writers on punishment. The United States was the primary center for the resurgence of retributive thought that came to be known as the 'just deserts' movement. Widely read books by Von Hirsch (1976), the American Friends Service Committee (1972), Frankel (1973), and Fogel (1975), among many others, deeply criticized modernist penology and offered in its place a more transparent and presumably more limited state use of coercion to punish or to incapacitate those who act clumsily and dangerously. The central principles of just deserts were sentences proportionate to the seriousness of the offense, equal punishment of those committing similar offenses, and transparent decision making that would prevent the possibility of the total loss of liberty faced by minor criminals.

By the 1980s and 1990s, the philosophy of just deserts largely had been superseded by two quite different penologies that are postmodernist and post-Enlightenment. One is a discourse of punishment as risk management; it represents some continuity with the modernist quest for effective crime control, but in terms that reject the key values of modernism and the

Enlightenment. The other is a populist punitiveness (Bottom 1995) or authoritarian populism (Hall et al. 1978) that valorizes victims and the therapeutic potential of vengeance and cruelty (Scheingold 1984, 1991, Sarat 2001).

5. The New Penology and Populist Punitiveness

As modernist penology began to unravel in the 1970s, scholars have begun to describe the emergence of an inchoate successor penology, a 'new penology' (Cohen 1985, Feeley and Simon 1992), or 'managerialism' (Bottom 1995). This new penology is most apparent in its contrast with the central tendencies of modernist penology. Rejecting the historic modern commitment to the individual as the target of penal justice, the new penology deals primarily in aggregates, dangerous classes, high risk populations and the like. The new penology replaces the objective of transforming criminal deviance with that of managing it. The new penology drops the focus on the social context and interest in penalty in favor of the system itself as the primary source of normative order.

These elements can be seen in such recent penal practices as preventive detention of sex offenders, mandatory minimum sentences (Tonry 1995), and super max prisons (King 1999). These practices combine a discourse of risk management and a skepticism about the potential for transforming offenders with a systematic preference for decision-making criteria rooted in forms of knowledge internal to the system itself.

While the 'new penology' describes the way professionals within the penal system understand their mission, there is a large gap between the managerialism of this discourse and increasingly emotive populist discourse of punishment which emphasizes vengeance, shaming, and the therapeutic benefit of punishment for victims. Throughout the era of modernist penology popular consciousness about punishment was presumed to lag behind the progressive logic of professionals. With the exception of aberrant regimes like the National Socialists in Germany in the 1930s and 1940s, most modern governments treated penal policy as something that should be shaped by professional concerns with only grudging concessions to popular demands. What is distinctive about the turn in penal policy since 1980, especially in the United States, is the open embrace of populism as an acceptable logic of punishment. This is exemplified by the embrace of the death penalty by virtually all political leaders in the United States and the enactment of expressly emotional laws like 'Three-Strikes' (Sechrest and Sichor 1996).

See also: Criminal Law and Crime Policy; Death Penalty; Deterrence; Deterrence: Legal Perspectives;

Deterrence Theory: Crime; Juvenile Justice: International Law Perspectives; Prisons and Imprisonment; Punishment, Comparative Politics of; Sanctions in Political Science

Bibliography

- American Friends Service Committee 1972 *Struggle for Justice*. Hill and Wang, New York
- Beccaria C 1996 *Of Crime and Punishment*. Trans. Grigson J. Marsilio Publishers, New York
- Beck U 1992 *The Risk Society: Towards a New Modernity*. Trans. Ritter M. Sage Publications, London
- Bottom A 1995 The philosophy and politics of punishment and sentencing. In: Clarkson C, Morgan R (eds.) *The Politics of Sentencing Reform*. Clarendon Press, Oxford, UK
- Cohen S 1985 *Visions of Social Control*. Polity Press, Cambridge, UK
- Cummins E 1994 *The Rise and Fall of California's Radical Prison Movement*. Stanford University Press, Stanford, CA
- Durkheim E 1964 *The Division of Labor in Society*. Trans. Simpson G. Free Press of Glencoe, New York
- Duster T 1990 *Backdoor to Eugenics*. Routledge, New York
- Evans R J 1996 *Rituals of Retribution: Capital Punishment in Germany, 1600–1987*. Oxford University Press, Oxford, UK
- Feeley M, Simon J 1992 The new penology: Notes on the emerging strategy of corrections and its implications. *Criminology* 30: 449–74
- Fogel D 1975 *We are the Living Proof: The Justice Model for Corrections*. Anderson, Cincinnati
- Foucault M 1977 *Discipline and Punish: The Birth of the Prison*. Trans. Sheridan A. Pantheon, New York
- Frankel M E 1973 *Criminal Sentences; Law Without Order*. Hill and Wang, New York
- Garland D 1985 *Punishment and Welfare*. Gower, Brookfield, VT
- Garland D 1990 *Punishment and Modern Society: A Study in Social Theory*. University of Chicago Press, Chicago
- Garland D, Young P 1983 Toward a social analysis of penalty. In: Garland D, Young P (eds.) *The Power to Punish: Contemporary Penalty and Social Analysis*. Heinemann, London
- Garland D 2001 *The Culture of Control: Crime and Social Order in Contemporary Society*. University of Chicago Press, Chicago
- Hall S, Critcher C, Jefferson T, Clarke J, Roberts B 1978 *Policing the Crisis: Mugging, the State, and Law and Order*. MacMillan, London
- Hay D, Linebaugh P, Rule J G, Thompson E P, Winslow C 1975 *Albion's Fatal Tree: Crime and Society in Eighteenth-Century England*. Pantheon, New York
- Hood R G 1996 *The Death Penalty: a World Wide Perspective*. Oxford University Press, New York
- Hudson B A 1996 *Understanding Justice: An Introduction to Ideas, Perspectives and Controversies in Modern Penal Theory*. Open University Press, Philadelphia PA
- Kant I 1974 *The Philosophy of Law* (trans. Hastie W 1877). Reprinted 1974, A M Kelley, Clifton, NY
- King R 1999 The rise and rise of the Supermax: an American solution in search of a problem. *Punishment and Society* 1: 163–86
- Maguire K, Pastore A L 1998 *Sourcebook of Criminal Justice Statistics 1997*. Bureau of Justice Statistics, Washington, DC

- Nietzsche F W 1994 *On the Genealogy of Morals*. Trans. Diethe C. Cambridge: Cambridge University Press, UK
- O'Malley P 1999 Volatile punishments: Contemporary penalty and neo-liberal government. *Theoretical Criminology* 3: 175–96
- Rothman D J 1980 *Conscience and Convenience: The Asylum and its Alternatives in Progressive America*. Little Brown, Boston
- Rothman D J 1990 *The Discovery of the Asylum: Social Order and Disorder in the New Republic*, rev. edn. Little Brown, Boston
- Rotman E 1995 The failure of reform: United States, 1865–1965. In: Marris N, Rothman D (eds) *The Oxford History of the Prison: The Practice of Punishment in Western Society*. Oxford University Press, New York
- Rusche G, Kirchheimer O 1968 *Punishment and Social Structure*. Russell and Russell, New York
- Sarat A 2001 *When the State Kills: Capital Punishment and the American Condition*. Princeton University Press, Princeton, NJ
- Scheingold S 1984 *The Politics of Law and Order: Street Crime and Public Policy*. Longman, New York
- Scheingold S 1991 *The Politics of Street Crime: Criminal Process and Cultural Obsession*. Temple University Press, Philadelphia
- Sechrest D, Sichor D (eds.) 1996 *'Three Strikes' and Public Policy*. Sage, London
- Simon J 1993 *Poor Discipline: Parole and the Social Control of the Urban Underclass, 1890–1990*. University of Chicago Press, Chicago
- Simon J 1997 Governing through crime. In: Friedman L M, Fisher G (eds.) *The Crime Conundrum: Essays on Criminal Justice*. Westview, Boulder, CO
- Spierenburg P 1984 *The Spectacle of Suffering: Executions and the Evolution of Repression from a Pre-Industrial Metropolis to the European Experience*. Cambridge University Press, Cambridge, UK
- Stem V 1998 *A Sin Against the Future: Imprisonment in the World*. Northeastern University Press, Boston
- Thompson E P 1975 *Whigs and Hunters: The Origin of the Black Act*. Pantheon, New York
- Tonry M H 1995 *Malign Neglect—Race, Crime, and Punishment in America*. Oxford University Press, New York
- Von Hirsch 1976 *Doing Justice: The Choice of Punishments*. Hill and Wang, New York
- Zimring F E, Hawkins G 1986 *Capital Punishment and the American Agenda*. Cambridge University Press, New York

J. Simon

Copyright © 2001 Elsevier Science Ltd.
All rights reserved.

Q

Qualitative Methods, History of

This brief history traces the development of socio-historical methodologies employed in relation to data that are either archival or based on direct social observation and interaction, insofar as these methods do not depend on quantitative analysis. The history of qualitative methods can be explored in three respects: (a) institutionalization; (b) how objects of inquiry have been construed; and (c) logics and analytic strategies. Recently, debates in the wake of the 'postmodern' turn have created new possibilities for understanding qualitative research, and research more broadly, as a terrain of inter-discursive communication with properties of 'integrated disparity.'

1. Institutional Emergences of Qualitative Methods

Sociohistorical inquiry emerged long before the rise of the modern academic disciplines. In China, statecraft early on was informed by typologies of social circumstances and transitional possibilities of action specified by the *I Ching*. In the fifth century before the modern era, ancient Greece had Herodotus, as much an observer of the contemporary as a historian. And during the fourteenth century, in the Arab world, Ibn Khaldun wrote in a way that anticipated comparative and historical sociology. These and other relatively sophisticated practices of inquiry coexisted with much more basic efforts of social groups to document their existences through the development of archives, king lists, annals, and chronicles. Thus, the arrangements through which sociohistorical knowledge has been produced are longstanding and diverse.

However, contemporary qualitative inquiry is especially a product of modernity, and of critical and postmodern reactions to the research practices spawned of modernity. Initially, qualitative research was inextricably linked with the gradual emergence of modern national societies and their colonial expansion, most concertedly during the nineteenth and twentieth centuries (Vermeulan and Roldán 1995). Intellectually, the rise of social and humanistic disciplines can be traced to the European Enlightenment of the eighteenth century. Thinkers such as Giambattista Vico and Johann Gottfried Herder increasingly began to regard social and historical processes as subject to understanding in their own terms rather

than products of the will of God alone, and Edward Gibbon developed analytic history by use of concrete social explanations (i.e., in his *Decline and Fall of the Roman Empire*).

In the nineteenth century, history developed under two aligned tendencies: efforts to differentiate the professional discipline from popular historical writing, and attempts to harness sociohistorical inquiry to the 'liberal' project of constituting the State, citizens, and the economy as a viable assemblage for social development. Thus, Leopold von Ranke sought to establish accounts of 'what really happened' on the basis of 'scientific' validation of facts drawn from archival sources, while simultaneously positing the history of political and religious elites as the subject matter that would reveal the meaning of history as a whole, that is, as a manifestation of the will of God. Historians in the United States for the most part embraced a modernist quest for 'objectivity' inspired by Ranke (Novick 1988), while leaving the ideological underpinnings of Ranke's scientific historicism largely uninspected. By contrast, historical sociology developed through the middle of the twentieth century mostly in relation to grand evolutionary social theories, at points aligned with anthropological ethnology, which was focused on constructing racialized world evolutionary sequences of social-group development. Only after World War II, and especially in the 1960s, did historical sociology develop a methodologically sophisticated array of comparative and theoretical strategies for the analysis of history (Barnes 1948, Smith 1991).

As for field research, it emerged in relation to modern colonial expansion and capitalist industrial development. Among Europeans, especially from the sixteenth century onward, knowledge gained from travel became a basis of colonial power sustained by exploratory expeditions, practical sciences, missionary work, and territorial penetration of colonial civil services. Sometimes authors combined traveling for a trading company or state with writing for a popular audience that wished to visit other lands vicariously. But epistemological principles of social knowledge were also at stake. Samuel Johnson's latter eighteenth-century trip to western Scotland implicitly demonstrated that cultural differences would require forms of knowledge different from the then emerging rationalized social quantification (Poovey 1998).

'Exploration' journals became increasingly systematic not only in describing the flora and fauna of the 'discovered' world, but also in inventorying the lan-

guages and practices of its peoples. Organizations such as the Smithsonian Institute (founded in 1846) coordinated efforts to gather data and circulated inventory questionnaires that could be filled out by non-professional travelers. Yet travelers' accounts—often saturated with racist and ethnocentric assumptions—largely served as grist for evolutionary ethnology and philosophical ('armchair') anthropology, for example, in Emile Durkheim's work on religion and James Frazer's *The Golden Bough*.

By the late nineteenth century, ethnology and philosophical anthropology still dominated the disciplinary space where modern ethnography was beginning to emerge. Lewis Henry Morgan's field study of the Iroquois, published in 1851, was largely a springboard to his ethnology. Only with the work of Franz Boas and Bronislas Malinowski did inquiry begin to abandon the ethnological search for trans-group anthropological principles of development, and emphasize analysis of social groups in their contemporaneous existences (a shift much bemoaned by the theorists). Boas is a crucial transitional figure. His late nineteenth-century observation and interviewing of indigenous peoples such as the Kwakiutl was initially intended to advance his studies of perception and speech, to the point that he sometimes became impatient with ceremonial potlatches that wasted research time! Nevertheless, Boas pioneered efforts to create valid information based on direct, systematic field research, and he trained students such as Edward Sapir and Margaret Mead.

During and after World War I, Malinowski even more strongly embraced sustained participant observation oriented toward characterizing a group's present way of life. His Trobriand Islands fieldwork, published as *Argonauts of the Western Pacific*, proposed a methodology borrowed from sociology, placing empirical observations within a functionalist framework, yet it also consolidated the modern ethnographic role as expert offering an authoritative account aimed at an emic understanding of the 'other's' point of view. This tension—between an objective theoretical framework and emic analysis—became a hallmark of modern anthropology, for example, in the French Durkheimian tradition, carried forward by Claude Lévi-Strauss (and in Britain by Alfred R. Radcliffe-Brown), and even in the way Clifford Geertz sometimes combined 'thick description' with social-systems rhetoric.

Ethnographic study of the exotic 'other' was hardly the only impetus to institutionalization of field research. Famously, Alexis de Tocqueville voyaged to the United States in the 1830s to learn more about the American prison system. During the late nineteenth and early twentieth centuries, sociology was influenced by similar practical quests for knowledge about social conditions, for example, in studies of social issues sponsored by the *Verein für Sozialpolitik* in Germany, by socialists in England, and among social workers

such as those at Chicago's Hull House in the United States.

Building on the latter tradition, during the early twentieth century, sociology's strongest institutional development of qualitative methodology came at the 'Chicago school' founded by Robert Park, who took inspiration from Boas and Malinowski to encourage field research. The city of Chicago became an urban laboratory for the direct observational study of social phenomena. Chicago faculty and students produced a stream of monographs such as *The Hobo* (1923) and *The Gold Coast and the Slum* (1929), and their approach found common cause with observational and in-depth interviewing field studies of communities, such as *Middletown* (1929) and the *Yankee City* studies (1941–59), work itself influenced by anthropological ethnographies (and sometimes carried out by anthropologists). In the 1950s, Herbert Hyman formalized interviewing as a methodology, and Chicago students such as Howard S. Becker, Jr., continued the pragmatic and symbolic interactionist Chicago tradition that had been initiated by George Herbert Mead and developed by Everett Hughes. During the 1960s, phenomenological sociology, inspired by Alfred Schutz, and ethnomethodology, promoted by Harold Garfinkle, emerged as alternative approaches that problematized objectivity, realism, and the social activities through which qualitative knowledge is created. Added to this mix during the same decade were the constructivism of Peter Berger and Thomas Luckmann, the dramaturgical approach of Erving Goffman (himself trained at Chicago), and a nascent bridge between field research and history via the emerging interest in oral history. These developments, together with the 1960s retooling of historical sociology, set the stage for an efflorescence of qualitative research in the last decades of the twentieth century.

Yet, just when qualitative research flourished, it also was beset by differentiation and methodological crisis. Some symbolic interactionists, ethnomethodologists, and comparative and historical sociologists sought to affirm the scientific legitimacy of their enterprises by formalizing research methods and strategies, while other researchers took seriously the rejection of scientism consolidated from the 1960s onwards by critical theorists such as Jürgen Habermas, hermeneuticists such as Paul Ricoeur and Charles Taylor, philosophers of history (most notably Hayden White), feminist epistemologists such as Dorothy Smith and Sandra Harding, anthropologists like Marshall Sahlins and James Clifford, and the wave of poststructuralist and postmodern 'turns' heralded by Michel Foucault and Jacques Derrida. Ethical considerations about privacy, asymmetry of field research, and exploitation of subjects—raised especially by feminists and within anthropology—cast long shadows on the high modernist quest for objectivity and validity in qualitative analysis. During the last two decades of the century, these developments, along with

the rise of world systems and globalization theories, led to the formulation of new hybrid qualitative methodologies that linked the local and the global, ethnographic and historical analysis, by such researchers as Michael Burawoy and Philip McMichael. The old dream of the ethnologists, of theorizing case studies in relation to general social processes, was reborn, but in a dispensation that eschewed ethnological ethnocentrism.

2. The Object of Inquiry, a Brief History

One enduring problem of qualitative research concerns how to construe the object of inquiry, that is, what is studied. At the end of the eighteenth century, Immanuel Kant raised the philosophical problem of selection, namely, which of myriad events revealed the 'guiding thread' of history. In the nineteenth century, under the influence of teleologies such as those of Hegel and Ranke, historians were generally content to regard their studies as centered on 'natural' objects that had their own coherences in the world itself. However, as Hayden White (1973) has shown, the nineteenth-century European historians established rhetorical coherence through one or another genre of emplotment that offered a metanarratological structure into which to fit events. Outside of history (and increasingly within it), an alternative tendency developed the object of analysis via theory. Thus, Karl Marx and Frederick Engels theorized about capitalism as a dynamic totality changing through its own contradictions over time, and used their perspective to thematize social events and processes in places as diverse as India, the United States, and Manchester, England. Within academia, the Marxian approach found its counterpart in the program of German historian Karl Lamprecht, who sought to displace Ranke's elites as the object of historical analysis with the study of diverse social groups and their interaction.

The Kantian problem of the relation between object and selection was more directly addressed during the German *Methodenstreit* toward the end of the nineteenth century, especially through the effort of Wilhelm Rickert to posit the objective value of truth in a way that would yield a shared way of looking at the world, and by extension, a shared object. Among mainstream historians during the twentieth century, a sort of ontological historicism emerged, in which shifts in styles (for example, the emergence of social history) depended on shifting the object of historical analysis, rather than its methods. The major alternative to historicism, which Max Weber spelled out at the beginning of the twentieth century, depended on two central assertions:

(a) first, that because values could not be warranted scientifically, the objects of sociohistorical inquiry could not be scientifically determined, even though,

once determined, scientific methods could be applied to their study; and

(b) that the objects of sociohistorical inquiry are not natural kinds, but rather, infinitely varying socio-historical phenomena that become objects of inquiry through cultural interest in them.

Alternatives to Weber's approach can be traced in history to the Annales School and Marxism, and in the social sciences, methodologically to the objectivism of Emile Durkheim, and analytically to Georg Simmel. Durkheim had provided the general warrant for objectivist approaches in his 1895 *Rules of Sociological Method*. And, in a way that is not widely recognized, Durkheim's structuralism had its equivalent in Simmel's early twentieth-century identification of social 'forms' (such as 'sociability'), a central inspiration for a number of qualitative symbolic interactionists (Hall 1999, Chap. 4). Within structural Marxism, the perspective of the proletariat was assumed to offer an objective standpoint for the analysis of capitalism, though during the first half of the twentieth century this position became increasingly problematic at the hands of Georg Lukács and critical theorists Max Horkheimer and Theodor Adorno. If the object of inquiry for structuralist Marxism was a totality, for the Annales school, in the formulation of Fernand Braudel, the totality was the grid of objective temporality, onto which multiple levels of change—ecological, structural, *histoire événementielle*—could all be mapped (Hall 1980, Burke 1990).

Ethnographic field researchers during the first half of the twentieth century generally treated their cases as naturally occurring 'primitive' societies subject to 'salvage anthropology.' Yet, the supposed isolation of the natural case became increasingly problematic with the post-World War II emergence of first modernization and then world-system theory approaches, both of which drew into question the autonomy of indigenous social groups. A parallel tension developed within sociological domains of field research. At the moment of high modernism, symbolic interactionists like Herbert Blumer would posit the empirical world 'out there' as the object of 'naturalistic research.' Yet, interactionists shared with phenomenologically oriented researchers and ethnomethodologists the sense that the empirical object of research was 'constructed,' i.e., that the meaningful world was constituted by its participants' practices.

By the beginning of the twenty-first century, the objectivist-relativist divide in constituting objects of qualitative inquiry persisted, but the cultural and Foucauldian turns in history, the linguistic turn in anthropology, and increased understanding of how ethical choices shape data from field research all created conditions in which answers to the question 'what is a case?' (Ragin and Becker 1992) had become considerably more nuanced and sophisticated. In particular, by 1978, Arthur Stinchcombe argued against comparing cases as total entities one against

another, and in favor of searching out the deeply analogous social processes that might be at work in cases that, on the face of it, might not even be considered as members of the same set. Overall, in today's post-positivist climate, reflexivity rules. Even objectivists recognize that theoretical orientations and concepts have consequences for how cases are construed, while dialogic relativists, some of whom doubt the existence of any straightforwardly accessible empirical world to be studied, have had to come to grips with the problem of how adequately to 'represent' the (constructed) realities of their encounters with the people about whom they write.

3. *Developments in Logics of Qualitative Analysis*

Philosophers have debated the logics of human reasoning for millennia. But the most compelling specification directly relevant to modern sociohistorical inquiry came in John Stuart Mill's (1843/1950) philosophy of scientific method, which specified two methods of non-experimental research. The method of agreement entailed isolating crucial similarities in a set of diverse cases, yielding the claim that if two aspects co-occur under otherwise diverse circumstances, they are likely to be somehow connected with each other. On the other hand, the indirect method of difference extended the logic of control-group experiments to nonexperimental settings, by searching for contingent co-related variations. In effect, two (or more) findings derived from parallel uses of the method of agreement could be placed side by side. If a particular propensity were found only in the presence of another particular propensity, and never when the second propensity was absent, Mill reasoned that one was 'the effect, or the cause, or an indispensable part of the cause' of the other.

The application of Mill's logic to sociohistorical inquiry met with opposing visions of qualitative analysis that became crystallized by issues raised during the late nineteenth-century *Methodenstreit*. Whereas the logic of Mill yielded a nomothetic search for general principles, an alternative, idiographic view suggested that the study of human affairs necessarily dictates emphasis on the individually distinctive, indeed unique, aspects of social phenomena. Whatever the resolution to this question, a second issue concerned whether the special character of sociohistorical phenomena made them necessarily the subjects of a different kind of knowledge than the explanatory search for universal causal laws, namely, a cultural science based on understanding social meanings and actions. Three broad logics of qualitative methodology—interpretive strategies, ideal-type analysis, and formal analysis—can be traced in relation to these contestations.

First, as the twentieth century dawned, Wilhelm Dilthey proposed a historicist hermeneutics that

would seek to understand history as embedded in the life meanings of the people who participate in it. This basic approach has its parallels in both ethnography and sociological field research. Elaborations of analytic logics were especially concentrated in sociology. Alfred R. Lindesmith (1947) developed the method of 'analytic induction' as a procedure for successively refining hypotheses in relation to cycles of field research, an approach that Glaser and Strauss (1967) subsequently reinterpreted as 'grounded theory.' By the 1960s, 'participant observation' (Platt 1983) and in-depth interviewing became increasingly institutionalized as legitimate approaches to inquiry, with their own epistemological rationales, challenges (notably how to construe and know subjectivity of the other person), research procedures, and handbooks, such as those of Buford Junker, Severyn Bruyn, and John and Lyn Lofland.

Second, working at the same time as Dilthey, Max Weber consolidated a *Verstehende* approach that was both comparative and interpretive. He did not deny the possibility that sociohistorical inquiry might discover general regularities of the sort Mill sought to identify, but he regarded individualizing knowledge as the central value interest of sociohistorical inquiry. Therefore, he combined explanation and understanding, by using generalized ideal types that would be 'adequate on the level of meaning' and thus serve as meaningfully coherent analogs to the patterned character of empirical cases, yet nevertheless allow for comparison beyond individual particularities. Weber's methodology became central to the renaissance of historical sociology consolidated in the 1960s (Bendix and Roth 1971). In the 1990s, Geoffrey Hawthorn and David William Cohen, among others, developed Weber's logic of the 'mental experiment' into 'counterfactual analysis.'

Third, Mill's methodological effort became the touchstone for subsequent formalizations of qualitative logic. The most important elaboration has been the Qualitative Comparative Analysis pioneered by Charles Ragin (2000), which builds on Mill's principles by using set theory and Boolean algebra to identify multiple configurational patterns within an array of systematically described cases. In a different development, with the late twentieth-century turn toward the study of discourse (partly inspired both by ethnomethodology and post-structuralism), scholars such as Larry Griffin theorized ways in which narrative accounts might become subjected to formal analysis that would yield descriptions of discrete social processes. These and related developments of formalization—as well as techniques of 'triangulation' that compared findings gained from diverse methodologies—increasingly blurred the lines dividing qualitative from quantitative analysis. Yet with logical development also came critique, such as that of Stanley Lieberman (in Ragin and Becker 1992), who questioned whether formalization alone could resolve logical

difficulties of comparative studies with small numbers of cases.

4. Postmodern Reflexivity in the Shared Domain of Qualitative Methods

By the end of the twentieth century, there had been substantial differentiation and development of qualitative methods, as well as experimentation in both research role relationships in field research and the form and rhetoric of texts reporting qualitative research more generally. Yet, these methodological shifts typically transpired within relatively autonomous schools where one or another approach was alternatively embraced, reinvented under a new flag, or syncretically revised in relation to alternatives. The differences between inductive and deductive historical sociologists, or between mainstream symbolic interactionist field researchers and those who had taken the postmodern turn, often loomed as important as the disjuncture between quantitative and qualitative research more generally.

Under these conditions, a fruitful development has been the specification of typologies that describe alternative practices of inquiry in terms of the shared or contrasting cultural logics of their research methods, rather than assuming an equivalence between particular qualitative methodologies and various schools or genres (e.g., symbolic interactionism, cultural history). Theda Skocpol and Margaret Somers, and Charles Tilly developed typologies of historical and comparative sociology, and John R. Hall (1999) used discursive analysis to elaborate a general typology describing eight core cultural logics of research practice. In these terms, the three methodologies most favored in field research and historical sociology are historicism, contrast-oriented comparison, and analytic generalization. These methodologies are part of a wider array of possibilities, including practices such as 'configurational history' that are often employed without benefit of formalization, and other practices, such as 'universal history,' that persist despite postmodern condemnation of 'metanarratives.' Typologies of qualitative analysis might seem to harden the boundaries between alternatives. However, the outcome of discursive analysis is the opposite, for it reveals a condition of 'integrated disparity' among research practices that share cultural logics in ways not previously understood, and it clarifies the conditions and potential for translation of research knowledge produced through radically alternative research practices (Hall 1999, Chap. 9). Thus, although qualitative research is shot through with contemporary tensions between science, relativism, and constructivism, qualitative researchers may find ripe conditions for a new era of sophistication about the character of qualitative research that will have

implications for research practices more widely. But to consider those possibilities moves beyond history into the realm of as-yet contingent possibility.

See also: Analytic Induction; Ethnomethodology: General; Field Observational Research in Anthropology and Sociology; Grounded Theory: Methodology and Theory Construction; Health Research, Qualitative; Interactionism: Symbolic; Interpretive Methods; Macromethods; Interpretive Methods: Micromethods; Qualitative Methods in Geography; Science and Technology Studies: Ethnomethodology; Symbolic Interaction: Methodology

Bibliography

- Barnes H E 1948 *Historical Sociology*. Philosophical Library, New York
- Bendix R, Roth G 1971 *Scholarship and Partisanship: Essays on Max Weber*. University of California Press, Berkeley, CA
- Burawoy M et al. 1992 *Ethnography Unbound*. University of California Press, Berkeley, CA
- Burke P 1990 *The French Historical Revolution: The Annales School, 1929–89*. Polity, Cambridge, UK
- Glaser B G, Strauss A L 1967 *The Discovery of Grounded Theory*. Aldine, Chicago, IL
- Hall J R 1980 The time of history and the history of times. *History and Theory* 19: 113–31
- Hall J R 1999 *Cultures of Inquiry: From epistemology to discourse in sociohistorical research*. Cambridge University Press, Cambridge, UK
- Lindesmith A R 1947 *Opiate Addiction*. Principia Press, Bloomington, IN
- Mill J S 1843 1950 *John Stuart Mill's Philosophy of Scientific Method*. Hafner, New York
- Novick P 1988 *That Noble Dream: The 'Objectivity Question' and the American Historical Profession*. Cambridge University Press, Cambridge, UK
- Platt J 1983 The development of the 'participant observation' method in sociology: origin myth and history. *Journal of the History of the Behavioral Sciences* 19: 379–93
- Poovey M 1998 *A History of the Modern Fact*. University of Chicago Press, Chicago, IL
- Ragin C C 2000 *Fuzzy-Set Social Science*. University of Chicago Press, Chicago
- Ragin C C, Becker H S (eds.) 1992 *What is a Case?* Cambridge University Press, Cambridge, UK
- Smith D 1991 *The Rise of Historical Sociology*. Temple University Press, Philadelphia, PA
- Stinchcombe A L 1978 *Theoretical Methods in Social History*. Academic Press, New York
- Vermeulan H F, Roldán A A (eds.) 1995 *Fieldwork and Footnotes: Studies in the History of European Anthropology*. Routledge, London
- White H 1973 *Metahistory: The Historical Imagination in Nineteenth-century Europe*. Johns Hopkins University Press, Baltimore, MD

J. R. Hall

Copyright © 2001 Elsevier Science Ltd.
All rights reserved.

12617

Qualitative Methods in Geography

Contemporary human geography is characterized by a diversity of topic areas, theoretical questions, and methodological plurality. A central concern with spatiality is constant, but a diversity of geographical imaginations (Gregory 1994) suggests a growing number of approaches appropriate to investigating how 'geography matters' in how social, economic, and political life unfolds. Among these, qualitative methodology provides an alternative paradigm of knowledge creation to that of quantitative techniques. With its admission of subjective experience and focus on meaning, qualitative methodology provides a variety of ways through which to analyze and interpret landscape, region, location, space, place, and the complex intersections of social and spatial relations. It has been critical to the re-imagining of the discipline as contemporary philosophical and theoretical debates have been integrated into empirical inquiry.

1. A Paradigm of Inquiry and Collection of Methods

Rather than accepting the possibility of a unitary, objective explanation of observed phenomena, qualitative research approaches acknowledge and seek the multiple and partial versions of reality that are constituted in the course of social life. Embracing a variety of approaches and field methods, a common interest in qualitative research is to enhance understanding through description and concepts grounded in the perspective of those studied. The meanings brought to bear by individuals and groups in ordering and making sense of their everyday social life and its spatiality are privileged in analysis. Methods of inquiry include various ethnographic approaches, including participation in and observation of the worlds of those studied and nonstandardized forms of interviewing with individuals or groups; text and visual analysis; and a growing repertoire of other non-numerical ways of observing, describing, and interpreting the world.

The shifting understandings of relationships between people and places and society and space are important in circumscribing how specific qualitative research strategies are employed. Interest in epistemological issues, particularly through the critical arm of the discipline, has fostered debate about the production of geographical knowledge and further shaped the choice of methods. There is a close association between theoretical developments, philosophies underlying inquiry, societal change, and the increasing use of qualitative methods in geography. As geographers seek to understand complex socio-spatial relations from different subdisciplinary and theoretical perspectives, such methods are used in various ways

according to the questions and purpose of research. They can provide descriptive data in mixed strategy designs to enhance findings from quantitative data, but more commonly are associated with postpositivist analyses that center the interpretation of meanings in explanation of narrative and visual data. Inductive analyses of narrative data, informed by concepts of theory, focus on generating concepts grounded in categories emerging from the data. Interpretation of various types of visual data, such as paintings, film, photographs, maps, and historical documents, similarly are sensitized by the theoretical and philosophical lenses of the 'reader.'

2. Qualitative Methodology in a Context of Philosophical and Theoretical Turns

Since the challenge of interpretative understanding to the dominance of quantitative methods in the 1960s, intellectual turns have brought new questions and new topics to the agenda of geographical inquiry, which have required a rethinking of appropriate methods. Geographers from a number of subdisciplines have explored qualitative approaches in their research as the influences of, first, humanism and then post-modernism, feminist scholarship, and cultural studies have problematized conventional methods of inquiry and categories of analysis.

Humanistic geography, with its philosophical inspirations from idealism, phenomenology, and existentialism, first challenged the strong quantitative turn of the 1960s (Ley and Samuels 1978). Interest in human agency, the inclusion of people's intentionality, a focus on lived experience, and the quest of understanding how meanings were produced in the context of interacting social and geographical worlds, marked the value of qualitative research strategies. Participant observation studies produced ethnographic accounts that permitted a linking of the social with the geographical in ways that opened up inquiry to numerous topic areas that privileged the perspective of the human subject. Interpretation of landscapes included concerns with sense of place, while case study approaches informed by a theoretical framework of social constructionism explored the embeddedness of human action and perceptions in place and provided models of inquiry for the 1980s. The qualitative methods text of Eyles and Smith (1988) signaled the establishment of qualitative research as an appropriate and useful paradigm of knowledge production within the discipline of geography.

Paralleling these developments, the influence of feminist scholarship resulted in further embracing of qualitative strategies in research. Noting the absence of women in geographical representations of the world, feminist geographers set out to include the neglected 'half' of the population in the mappings and

narratives of the discipline. In making the lives and geographies of women visible, qualitative methods, as in other disciplines, were privileged in their ability to uncover the everyday lives of women and the spatialities of their domestic and paid work (McDowell 1992). Bringing subjective realities of women into geographical explanation exposed the masculinist partiality of geographical inquiry and permitted the construction of concepts reflecting the geographies of women's lives. Linked with principles of research with an emancipatory aim, qualitative approaches remain important in feminist methodology as a vehicle, together with the contribution of quantitative methods, in exploring the spatiality of gendered social relations (Jones et al. 1997).

Postmodernism, the ongoing development of feminist theorizing and the later influence of cultural studies brought a 'cultural turn' to human geography (Massey et al. 1999). The importance of discourse in constructing human subjectivity and representing difference, whether reflected in the organization of space, social life, or in visual representations, was an important aspect of a complex literature on culture that influenced methodological approaches. Throughout human geography qualitative methods enabled inquiry to reveal the lifeworlds of marginalized 'others,' to engage critically with various media as cultural representations, to expose the multiple, partial meanings of landscapes and texts, and to disrupt many taken-for-granted categories of analysis. Immigrants, the ill, the disabled, the colonized, and other marginalized groups emerge as others whose subjectivities and geographies are being included in a broadening geographical knowledge. Ethnographies of workplaces and interviews with elites, analysis of contemporary visual representations of social diversity, and the inclusion of historical texts that include women's experiential accounts are examples that further reflect a concern with competing knowledges and representations of the world. These journeys into the lives of others has led to expansion and innovation in field techniques, including the use of life stories, focus groups, autobiographical writing, photography, film, and the adaptation of interviewing techniques and formats to meet the needs of particular research projects.

3. Constructing Knowledge Through Qualitative Research: Methodological Issues

With the increased use of qualitative research strategies, geographers have engaged with a methodological literature in the social sciences discussing the difficulties that emerge when conducting research with the 'other' of studies. Reflection centers on the complexity of constructing knowledge through reliance on the intersubjective relationship between researcher and researched and the interpretation of data

through the self as a socially positioned researcher who is part of the world they study. Issues of rigor in research design and practice have been addressed with attempts made to translate the concepts of quantitative research into those more amenable to qualitative methodology, but most discussion has focused on the close association between practical and theoretical dilemmas in research, and their location in epistemological concerns.

The particular attentiveness of feminist scholars to the micropolitics of research encounters, together with the acknowledgment of the situatedness and partiality of all experience and knowledge, have become the concerns of many qualitative researchers in the discipline. Gaining access to previously subjugated knowledges is shown to be a process located within a layering of power-laden social relations that situate the face-to-face research encounter within the wider politics and processes of 'the field,' theoretical and philosophical tensions, and the institutional values and processes of universities and funding bodies. Specificity of place is shown to mediate how this intricate layering plays-out in the microdynamics of particular fieldwork situations.

Particular issues vary in different types of inquiry, but the exercise of reflexivity by researchers has problematized the whole enterprise of knowledge production. Whether in the context of reading landscape as text, rewriting colonial geographies, the demedicalization of medical geography, the shifting concerns of feminist inquiry, interviewing elites in economic geography, or exposing processes of exclusion of marginalized groups in society, researchers consider the different ways their own identities and contextuality influence their analyses.

That the findings of research are theory laden and interpreted through the researcher's subjectivity has brought attention to the presentation and writing up of research. The representation of the multiple voices of the researched is seen as a politically charged and theoretically tricky enterprise. Both in relation to narrative and visual data interpretation, the text produced and its claims to knowledge are problematized. The authority of the author is challenged, in that a particular representation is one of several possibilities. Following from this, the reading and consumption of research writing are viewed as crucial dimensions of the research process and the development of knowledge, rather than separate activity.

Attempts to disrupt hierarchical research relations, including issues concerning the politics and ethics of qualitative research have moved increasingly to center stage in methodological debate. For whom research is conducted, how findings are used and related to theory production, and concern over the possible exploitative nature of the relationship with research participants have become important considerations in research conduct. Participatory and emancipatory models of research are being explored, accompanied by thinking

through how academia and activism are and may be connected. The ethical issues involved in such attempts to 'empower,' whether research is conducted at home, or in developing countries, have become an important focus of discussion.

4. Continuing Contributions

Expansion of the use of qualitative methods in geography is reflected in special issues of geographical journals and new texts on research methods that focus on qualitative methods or include qualitative approaches as important, rather than subordinate research strategies. Their value is recognized as theoretical developments have signaled the need to find new ways to discover and analyze data appropriate to current topics and questions. The various methods of qualitative research have opened up different ways of 'reading' and 'seeing' landscape, space, place, and the relationships between people and environments, so continually reshaping the nature and scope of geographical inquiry. The ongoing theorization of space, that has emphasized its relational, social, and recursive dimensions, has been accompanied by qualitative studies that are inclusive of human experience, perceptions, and meaning. Such studies have brought into view the diversity of local, everyday geographies, while at the same time locating individual and group experience within complexly linked local, national, and global relations.

The use of qualitative and quantitative methods is sometimes presented in terms of competing paradigms, sitting uncomfortably with each other. Theoretical turns and epistemological concerns may suggest that one be privileged over the other in the quest for valid knowledge in geographical inquiry, but different questions require different approaches. Both quantitative and qualitative work can and do contribute to furthering the understanding of spaces, places, locations, and how variously positioned individuals and groups live within and experience them. Qualitative geographical inquiry, however, has exposed the complexity of gaining access to and representing the situated knowledges of research participants, that precludes the development of rigid protocols or 'ways of doing' qualitative research. Different ways of telling, listening and knowing require flexibility in research design and approaches. Ways of doing qualitative research need to be situated within and be sensitive to the many situations that comprise the academic geographer's field. It may be anticipated that further innovation of methods and flexible models will remain the hallmark of research concerned with the interconnections of people, place, and space.

See also: Humanistic Geography; Phenomenology; Philosophical Aspects; Qualitative Methods, History of; Reflexivity; Method and Evidence

Bibliography

- Area* 1999 **30** (whole issue)
Barnes T J, Duncan J S 1992 *Writing Worlds: Discourse, Text and the Representation of Landscape*. Routledge, London, UK
Environment and Planning D: Society and Space 1992 **10**(5) (whole issue)
Eyles J, Smith D M 1988 *Qualitative Methods in Human Geography*. Blackwell, Oxford, UK
Geoforum 1999 **30**(4) (whole issue)
Gregory D 1994 *Geographical Imaginations*. Blackwell, Cambridge, Oxford, UK
Hay I 2000 *Qualitative Research Methods in Geography*. Oxford University Press, Melbourne, Vic
Jones III J P, Nast H J, Roberts S M (eds.) 1997 *Thresholds in Feminist Geography: Difference, Methodology, Representation*. Rowman and Littlefield, Lanham, MD
Ley D, Samuels M S 1978 *Humanistic Geography*. Maaroufa Press, Chicago
Massey D, Allen J, Sarre P (eds.) 1999 *Human Geography Today*. Polity Press, Cambridge, UK
McDowell L 1992 Doing gender: feminism, feminists and research methods in human geography. *Transactions of the Institute of British Geographers* **17**: 399–416
The Professional Geographer 1994 **46**(1) (whole issue)
The Professional Geographer 1999 **51**(2) (whole issue)

I. Dyck

Quality Control, Statistical: Methods

1. Introduction

Modern quality improvement by statistical methods began in the 1920s through efforts of W. A. Shewhart (see Shewhart 1931) of the Bell Telephone Laboratories. Rather than wait until manufactured items arrived at an inspection center at the end of the production line, he conceived of the idea of occasionally sampling during the manufacturing process and making corrections immediately, if needed. To monitor the need for possible adjustments, certain characteristics—called statistics—of each sample were recorded. If any of these statistics had unusual values that indicated something was wrong with the process, corrective action was taken before too many more defective items were produced. Accordingly, the quality of the final output was improved.

While Shewhart started this modern quality improvement by trying to reduce the number of defects in manufacturing, many organizations soon recognized that there was much waste throughout their operations. There were simply many activities that added nothing to satisfactory outcomes. This is not only true in manufacturing but in most service areas as well, including hospitals, banks, and educational institutions. Estimates of how much waste and worthless activities exist range from 25 to 50 percent of the total effort, although there are some organizations that

probably exceed that bound of 50 percent. In any case, most industries and businesses could make huge improvements in quality products and services if they only recognized how poorly they are doing. Often such recognition is found by taking measurements, collecting data, and turning these data into information through statistics.

There are many philosophies and techniques about improving quality of products and services. Some of these have been formalized by considering the ISO 9000 or Baldrige awards (see Ledolter and Burrill 1999). Many companies have earned these awards and have had continued success; but some have then gone on to fail because they have not had a continuous commitment by top management to maintaining quality products and services. Bob Galvin, former CEO of Motorola, truly believed that 'quality improvement is a daily, personal, priority obligation,' and his commitment made Motorola what it is today.

At the end of the twentieth century it is worth making note of one quality guru who almost lived that entire twentieth century. W. Edwards Deming was a statistician who was born October 14, 1900 and died December 20, 1993 and was giving short courses in quality improvement ten days before his death. Deming went to Japan after World War II and taught the Japanese how to make quality products. For his work there, he was awarded the Emperor's Medal, and the Japanese established the Deming Prize to be awarded each year to a company or individual contributing the most to quality improvement.

There are others too; certainly Joe Juran and Japan's Kaoru Ishikawa (see Ishikawa 1982) must be mentioned. However, it was Deming's approach to optimizing the entire system that was a major point. In addition, his thoughts on 'profound knowledge' are extremely valuable, particularly the important part of it about understanding variation and statistical theory.

Understanding variation is a key to quality improvement. Many of Deming's 14 points (Deming 1986), while not mentioning statistics, centered on reducing variation. Deming believed that barriers between departments, between management and workers, and among workers, must be broken down to improve communication and the ability to work as a team.

The lines of communication, all the way from suppliers to customers, must be open to help reduce variation and improve products and services. For example, he argued that a company should not only buy on price tag but have a few reliable suppliers (possibly one) for a single part because that will reduce variation. That is, many different suppliers would obviously increase variation. Moreover, he argued that you should become partners, friends if you like, with your suppliers. You learn to trust the other; in that way, you can use methods like 'just in time,' in which your inventory can be kept reasonably small, to keep costs down.

Deming also preached constancy of purpose. If management ideas tend to change too much, employees really do not know what to do and cannot do their best. That is, they are mixed up, increasing variation. It is better for them to receive a constant signal from their employer, a signal that only changes if research dictates ways of improving. More training and education for everyone associated with a company also decreases variation by teaching how to make the product more uniform. Workers must know that they should not be afraid to make suggestions to improve the process.

Being team members will make it easier for workers to speak up without fear of reprisal. Deming also noted that requiring quotas does not help quality. A foreman who has a quota of 100 per day will often ship out 90 good and 10 bad items just to make the quota. Clearly, it would reduce the variation and satisfy the customer better if only 90 good ones were shipped.

This leads to the point that a final inspection of products does not really improve the quality. With such a procedure, you can only eliminate the bad ones and send on the good ones. Improvements in the design of the products and manufacturing processes are needed. If these are done well, often with the help of statistical methods, that final inspection can be eliminated. That is, improvements should be made early in the process rather than try to correct things at an end inspection by weeding out the bad items.

Deming's points are still extremely important in the quality movement today. Understanding variation and reducing it still 'makes the doors fit better.' However, most of Deming's ideas also apply to service activities, ranging from market surveys to financial institutions (see Latzko 1986). Since so many more persons are employed in the service areas than in manufacturing, there are many more opportunities for quality improvement there. Why are some restaurants, clothing stores, law offices, banks, accounting firms, and hospitals better than others? Of course, those involved must know the technical aspects of their work, but the really excellent managers know that quality begins by treating their employees and customers right. The cost of a few dissatisfied persons is difficult to assess, but it is usually much greater than most think because these people will tell many others about their unpleasant experiences. Most of this quality improvement is common sense but it is not commonly used.

2. Statistical Methods for Quality Improvement

Let us begin our statistical methods by starting with one of the Shewhart charts, the mean chart. We sample from the output of a process. The frequency of sampling of the output of the process is really determined by the number of items being produced. If

there are a large number produced in a given day, the sampling might occur every hour. On the other hand, few in number might suggest sampling once per day is enough. Often these samples are small, and a typical size is $n = 5$, particularly when the measurement taken is something like length, weight, or strength. For example, in the manufacturing of paper, there is interest in its 'tear strength' because most often the manufacturer does not want paper to tear easily.

All of the underlying process have some mean μ and standard deviation σ . Hopefully the level μ is acceptable; and, in most cases, it is desirable to reduce the spread, as measured by σ , to make the output as consistent as possible. Hence, from each sample of $n = 5$ items, two statistics are computed: one to measure the level (middle) and one to measure the spread. We first concentrate on measure of the middle and that statistic is denoted by $\bar{x}$ and called the sample mean. Theory shows that once μ and σ of measurements of the underlying process are known (possibly estimated), the statistic will be between $\mu - 3\sigma/\sqrt{n}$ and $\mu + 3\sigma/\sqrt{n}$ unless there has been some change in the process producing these items. That is, if is not within the control limits $\mu \pm 3\sigma/\sqrt{n}$, then the process is investigated and usually some corrective action taken. The usual measures of spread are the sample range R and the sample standard deviation σ , and they have control charts similar to that of $\bar{x}$, and are called R - and σ -charts.

There are two other Shewhart control charts that must be described. The first of these is when we observe the number of flaws, say c , on a certain unit; for illustration, count the number of deaths on a certain highway each week. Here c can equal one of the non-negative integers: 0, 1, 2, 3, 4, and so on. A Poisson probability model is used for the statistic c , and these probabilities can be determined from the mean μ of the number of flaws. Again μ is often estimated using past data. However, once it is determined, then, for the Poisson model, the standard deviation of c equals. Thus, $\mu \pm 3\sqrt{\mu}$ gives the control limits for the c -chart. Future values of a c are plotted sequentially on this chart, taking some action if one is outside the control limits: if above $\mu + 3\sqrt{\mu}$, find out why and hopefully learn how to improve the process; if below $\mu - 3\sqrt{\mu}$, again find out why and hopefully learn how to continue this gain in quality.

The last control chart arises in a situation in which the items are measured as being 'good' or 'bad.' If n of these items are observed, let D equal the number of defective items in the n items. The binomial probability model is appropriate for the distribution of D . Here n is much larger than 5, often equal to as much as 100. Of course, $\hat{p} = D/n$ is an estimate of p , the fraction of defective items being produced. The binomial theory states that $\hat{p} = D/n$ has mean p and standard deviation $\sqrt{p(1-p)/n}$.

Once some estimate of the fraction defective, p , is found from past data, the control limits

$p \pm 3\sqrt{p(1-p)/n}$ are calculated for the p -chart. Future values of $\hat{p} = D/n$ are plotted on this control chart, and appropriate action is taken if one value is outside the control limits. Again, very small values of D/n would indicate some sort of improvement, and we would try to maintain that gain. When the statistics are between the control limits, which are usually estimated from the data, the process is said to be in statistical control.

The manufacturing team must make the decision whether or not the actual items from an 'in control' process are satisfactory. That is, they might find that the items are not satisfactory even though the process is in statistical control; then some adjustments must be made. However, once this is done, Shewhart control charts have proved to be extremely valuable in the past and have helped maintain the gain in the quality improvement of manufactured products. In addition, these control charts are now being used in many service areas, for example, studying errors made in recording the amounts on checks by banks.

Since most major manufacturing companies are requiring their suppliers to use statistical methods that help improve their products, the need for acceptance sampling to check upon the quality of incoming products has diminished. That is, partnerships with suppliers have been established, and hence the trust among the various parties is usually high.

Nevertheless, there are enough cases in which the incoming lot of items is poor; accordingly a large number of companies still use acceptance sampling. To describe acceptance sampling, suppose a lot has $N = 5,000$ items. Say that these materials are acceptable if two percent or less is defective.

This number, 2 percent = 0.02 is called the acceptable quality level. Understand that 0.02 is used only as an example, and some might insist that the incoming lot is only acceptable if there are almost zero defectives. In the latter case, this requires 100 percent inspection, which has some disadvantages, particularly if the item must be destroyed to inspect it, like 'blowing a fuse.' So most manufacturers are willing to accept lots with a small percentage of defects; needless to say, this percentage varies depending upon the importance of the item under consideration. To continue with the example, say a sample of $n = 200$ is taken from the lot of $N = 5000$, and it is decided to accept the lot if the number of defectives, D , in the sample is less than or equal to $Ac = 7$, the acceptance number.

Now clearly, as with any sampling acceptance plan, there are two types of errors that can be made. Say p is the fraction defective in the lot. If $p = 0.02$ and $D > 7$, then the lot is rejected when it should be accepted. This is an error of the 'first type.' The 'second type' error occurs if p is large, say $p = 0.06$ and $D = 7$, so that the lot is accepted when it should be rejected. It is desirable to have the probabilities of these errors as small as possible. Using a Poisson

probability approximation, it is found in this illustration that the probabilities of the two errors are $a = 0.051$ and $b = 0.090$, respectively. If these probabilities are not small enough, then the sample size n must be increased and the acceptance number Ac changed.

Typically, the manufacturer and the customer require certain specifications associated with the product. These are usually given in terms of a target value, and the lower specification limit, LSL, and the upper specification limit, USL. These 'specs' are set, often by the engineer or customer, so that if the product is within specs, it will work satisfactorily. One of the best ways to see if the products are within specs is to construct a histogram concerning the measurement of interest. If the product is satisfactory, the lower and upper values associated with the histogram should be within specifications. If several items fall outside of specs, then seemingly the process, without proper adjustments, is not capable of producing satisfactory products. If adequate improvements cannot be made with the current machinery, possibly 100 percent inspection must be used to eliminate the bad items, or else new machinery should be installed.

Clearly, economic characteristics of the situation are needed to make a final decision on the appropriate action. While a histogram seems to be a satisfactory way to measure the capability of the process, a large number of companies are resorting to various measures called capability indices. Possibly the most popular one is Cpk. Say μ and σ are the mean and the standard deviation, respectively, of the measurements associated with a process which is in statistical control.

The index Cpk is equal to distance of μ to the closest specification divided by 3σ . For example, suppose $LSL = 48.1$ and $USL = 49.9$ and $\mu = 49.1$ and $\sigma = 0.2$, then $Cpk = 0.8/3(0.2) = 1.33$, because $USL = 49.9$ is the closest spec to $\mu = 49.1$ and the distance between them is 0.8.

The reason that such a measure seems good is that if the underlying distribution is normal, then 99.73 percent of the items are within $\mu \pm 3\sigma$. If $Cpk > 1$, then the closest spec is more than 3σ away from μ . Most companies using Cpk require that the Cpk be greater than 1.33 or even 1.5, so that μ is 4σ or even 4.5σ away from the nearest spec.

However, there are some difficulties associated with the use of Cpk, or any of the capability indices. One is that it is assumed that the underlying distribution of measurements of an item is normally distributed, when often it is skewed with possibly a very long tail. But even worse, usually m and s have been estimated using some sample mean and standard deviation.

Knowing these are only estimates with some error structure, the computed Cpk using these estimates may be in error as much as 0.2 or 0.3, even with samples of sizes as large as $n = 50$ or 100. For illustration, the computed Cpk using estimates might be 1.19 while the real Cpk might be 1.39 or 0.99.

Frequently, the management might be very satisfied with a computed Cpk of 1.39 and not one of 1.28, if it is using the 1.33 cutoff; and yet there may not be any 'real' difference between the two values. Management should be aware of this possibility in making certain judgments and decisions.

Only some basic, but important, statistical methods for quality improvement have been touched upon. There are many more quantitative methods that can help, and these are considered in the references. These range from simple ideas like flowcharts, Pareto diagrams, and cause-and-effect diagrams, through CUSUMS and exponential weighted moving average charts to more complicated techniques in the design of experiments, including Taguchi's methods. Most of these techniques are explained in such books as those of Montgomery (1996) or Ledolter and Burrill (1999). Accordingly only a few remarks are made about the simple ones, and the references should be used for the more complicated statistical techniques.

A flowchart of the total process shows the steps that are needed to complete the process. Often by constructing a flowchart, the persons involved can see where certain steps can be eliminated or worthwhile shortcuts taken. A Pareto diagram is named for an Italian economist; it is essentially a bar graph in which the height of each bar represents the percentage of defects caused by a certain source. The possible sources are ordered so that the corresponding bars are in order of heights, with the tallest being on the left. Often one or two (possibly three) sources account for more than 85 percent of the defects; hence the Pareto diagram indicates clearly where the most effort should be placed to achieve improvement. The 'cause and effect' diagram is often referred to as an Ishikawa 'fish bone' diagram, because it looks like one. The main horizontal line points to the 'trouble', which is labeled to the right side of the horizontal line. Four to six diagonal lines leave the horizontal line, slanting to the 'northwest' or 'southwest' with labels at their ends that might be major sources of the trouble, like people, materials, machines, environment, etc. Then little 'bones' are attached to each of these with appropriate labels. It is amazing how completely these can be filled in with four or five knowledgeable persons 'brainstorming' for 30 minutes or so. All labels on the bones should then be considered for possible sources of the trouble, and often such a diagram calls attention to sources that might otherwise have been overlooked.

More advanced techniques like CUSUMS and exponential weighted moving average charts that can be used in place of or along side of Shewhart charts have proved most worthwhile. However, the designing of new products, or troubleshooting to find factors that are causing defects, are best served through statistical design of experiments. Some of these methods, in particular 2^k and 2^{k-p} designs, have been exceptional tools; they are considered in the book by Box et al. 1978. Overall, quality improvement has been greatly

enhanced by statistical theory and methods. The major contributors, including Shewhart, Deming, Juran, Ishikawa, Taguchi, and many more, deserve the appreciation of many in the quality movement. Clearly, from various experiences that have not been successful, much more remains to be done, and statisticians should play a major role in future efforts in improving the quality of products and services.

See also: Quality Control, Statistical: Reliability and Life Testing

Bibliography

- Box G E P, Hunter W G, Hunter J S 1978 *Statistics for Experimenters*. Wiley, New York
- Deming W E 1986 *Out of the Crisis*. Massachusetts Institute of Technology, Cambridge, MA
- Hogg R V, Ledolter J 1992 *Applied Statistics for Engineers and Physical Scientists*. Macmillan, New York
- Ishikawa K 1982 *Guide to Quality Control*. Asian Productivity Organization, Tokyo
- Latzko W J 1986 *Quality and Productivity for Bankers and Financial Managers*. Marcel Dekker, New York
- Ledolter J, Burrill C W 1999 *Statistical Quality Control*. Wiley, New York
- Montgomery D C 1996 *Introduction to Statistical Quality Control*. Wiley, New York
- Shewhart W A 1931 *Economic Control of Quality of Manufactured Product*. Nostrand, New York

R. V. Hogg

Quality Control, Statistical: Reliability and Life Testing

1. Reliability: What Does it Mean? Its Relation to Quality

The term reliability has become a part of our daily vocabulary. For example, the word unreliable describes the untrustworthy behavior of an individual, an item, or a procedure. In most situations the terms reliable and unreliable are conversational, and convey an intuitive feel. However, in dealing with situations involving modern machines and technologies, the need to be more scientific is on the increase. This has resulted in efforts to come up with sharper definitions of reliability, definitions that lend to quantification. But why must one quantify, and how should the quantification be done?

An answer to the first part of this question lies in the thesis that the notion of reliability is useful for certain kinds of decision making under uncertainty, and that the normative approach to decisions is based on quantifiable inputs (see Lindley 1985). Thus, in a

broad sense, quantified measures of reliability are inputs to normative decision making. The answer to the second part of the question depends on how reliability is defined, and this depends on how it is to be used. Thus, in psychological testing, reliability reflects consistency of results provided by a test instrument, and so quantification is based on a notion of consistency. By contrast, reliability used in the engineering sciences pertains to an attribute of functional quality, and this indeed is its connection with quality. Functional quality is an ability to satisfy a set of requirements, and reliability comes into play when the requirement is either survival, or an ability to perform. It is in this context that reliability is of interest. Here, the quantification of reliability becomes a subtle issue, because the approach used to quantify leads one to a philosophical debate, a debate that has plagued generations of thinkers. More about this will be said later. For now it suffices to say that reliability is a quantified measure of uncertainty about a unit's survival for a prescribed time, or its ability to perform as desired. The unit can be biological or physical; when biological, reliability takes the label 'survival analysis,' a topic that has become the backbone of biostatistics.

2. Background: Interest in Reliability—Past, Present, and Future

Reliability is a quantification of a certain type of uncertainty. From a historical perspective, the quantification of this type of uncertainty dates back to Huygens, who in 1669 constructed what is now known as a 'survival curve.' Huygens' work was based on Graunt's data on vital statistics. Since the first 'survival table' was published by Edmund Halley in 1693, one may claim that statistical methodology, and statistical data analysis, have their origins in reliability. During the early 1930s, the problems of commerce and insurance sustained an interest in this topic, whereas between the 1950s and the 1970s, quantified measures of reliability were needed for engineering design, for government acquisitions in aerospace and defense, for enforcing regulation in drug approval, and in matters of public policy, such as the safety of nuclear power generators.

Landmark achievements during this period are the papers by Davis (1952) on the analysis of equipment failure, by Epstein and Sobel (1953) on life testing for the exponential distribution, by Kao (1956) for acceptance sampling based on the Weibull distribution, and by Grenander (1956) on the analysis of mortality measurement. The 1960s also heralded the birth of the mathematical theory of reliability, with Barlow and Proschan's (1965) classic book summarizing the knowledge to date. This work essentially defined the subject and provided avenues of research for years to come. Also noteworthy are the book by Gnedenko et

al. (1969) which summarized the Soviet work in the area, and Marshall and Olkin's (1967) fascinating paper on modeling dependencies.

During the 1980s and the early 1990s, pressures of consumerism and competitiveness forced manufacturers to use quantified measures of reliability for specifying warranties particularly with regards to the credibility of software (see, for example, Singpurwalla and Wilson 1999). The late 1990s also began to see concepts of reliability used in business, economics, and finance, and in matters of political science, particularly disarmament and national security.

The continued importance of reliability is anticipated in the twenty-first century because of a growing reliance on the infrastructure of communication systems, transportation networks, distributed power grids, and civil engineering structures at earthquake prone zones. Major initiatives by government and industry are underway under the label of 'infrastructure protection,' and quantified measures of reliability are a key factor of these. Finally, an area of growing concern wherein reliability is poised to play a signal role pertains to the credibility of nuclear stockpiles in the face of test ban treaties (cf. Armageddon Moves Inside The Computer, *The Washington Post*, November 28, 1998).

3. Foundational Issues in Reliability: Is a Paradigm Shift Needed?

Reliability, as a quantification of a certain uncertainty, leads to philosophical issues that have been the subject of debate. These issues are prompted by several questions that arise naturally, the first of which is: whose uncertainty? Is it that of an individual or a group of individuals, or is uncertainty inherent and therefore universal? The second question is if uncertainty is to be quantified, what should the quantification be based upon? Should it be subject matter knowledge, collective judgement, hard data alone, or a combination of all the above? The third issue pertains to the strategy for quantification. Should it be probability and the calculus of probability or should it be alternatives to probability like possibility theory, belief functions, confidence limits, tests of hypotheses, etc.?

Much of the literature on reliability fails to acknowledge these issues; rather, the focus has been on methodology—not principles. All the same, a major portion of the written material on reliability regards uncertainty as being inherent or universal, and subscribes to the principle that uncertainty can be quantified only by means of hard data. Furthermore, it takes the position that probability and its calculus need not be the only way to quantify uncertainty. Consequently, the mathematical theory (cf. Barlow and Proschan 1965, 1975) treats all probabilities as being known, whereas the statistical literature emphasizes 'life-testing' (a way to generate hard data), confidence

limits, tests of hypotheses, etc. Indeed, statistical reliability theory's major export, the Military Standard 781C, wholeheartedly subscribes to the above approaches for describing uncertainty and for decision making. A more recent example is the book by Meeker and Escobar (1998). Whereas a use of this strategy may be meaningful in the context of mass phenomena, like the output of a continuous assembly line, its use for one of a kind items, like computer software or limited size populations, is ill-founded. Furthermore, approaches that rely solely on hard data tend to be ignored by the users. This in turn has undermined the role of statistics and calls for a paradigm shift for addressing the modern problems of reliability that have been raised by the engineering community. Does this mean that one should forsake the calculus of probability in favor of alternatives like possibility theory and fuzzy-logic that many are attracted to?

The answer to the above question is a cautious no! Despite some shortcomings of the probability calculus, there does not appear to be any need to forsake it. However, a paradigm shift is indeed necessary, but this has to do with the interpretation of probability, not an alternative to probability! To address the modern problems of reliability, problems involving one of a kind items for which there is little or no failure data but a lot of informed testimony, it is the interpretation of probability that matters. Specifically, one needs to subscribe to the subjective or the personal view of probability in its purest form, as advocated by de Finetti (1964) and by Savage (1972). In addition to facilitating reliability assessment of one of a kind items, the subjective view offers other advantages. First, it helps explain several results in reliability theory that have been obtained via purely mathematical derivations (see, for example, Lynn and Singpurwalla 1997, in Block and Savits (1997), and Kotz and Singpurwalla (1999). Second, it facilitates the development of normative decision making procedures for acceptance sampling and drug testing (see, for example, Lindley and Singpurwalla 1991 and Etzioni and Kadane 1993). Most important, the subjective view of probability enables the treatment of dependencies in system reliability assessment in a natural manner (cf. Lindley and Singpurwalla 1986). The matter of dependence has plagued reliability analysts for quite some time, because, it is often the case that independence is assumed in place of dependence, with the consequence that assessed reliabilities do not agree with actual experiences (cf. US Nuclear Regulatory Commission 1975, National Research Council 1998).

Thus, to summarize, the view is held (cf. Singpurwalla 1988) that the twenty-first century will see a paradigm shift in the methods for assessing reliability. This paradigm shift will entail the interpretation of reliability as a personal probability, and not the adoption of a new set of rules for describing and combining uncertainties.

4. Key Notions, Terminology, and Overview of Developments

In what follows, the position that reliability is a personal probability about the occurrence of certain types of events is adopted. Since the most common event of interest is survival of an item, under specified conditions, for a duration of time τ , $\tau \geq 0$, the reliability of the item is defined as

$$P(T \geq \tau) \stackrel{\text{def}}{=} R_T(\tau)$$

where P denotes probability, and $T \geq 0$, stands for the item's life-length. The interval $[0, \tau]$ is called the mission time, this terminology reflecting reliability's connections to aerospace. The quantity $R_T(\tau)$, as a function of $\tau \geq 0$, is called the reliability function, and if the item is a biological unit, then this function is called the survival function, denoted by $S_T(\tau)$. Clearly, $R_T(\tau)$ decreases in τ , going from one at $\tau = 0$, to zero, as τ increases to infinity. The life-length T could be continuous, as is usually assumed, or discrete when survival is measured in terms of units of performance, like miles traveled or rounds fired.

Let $f_T(\tau)$ be the derivative of $-R_T(\tau)$ with respect to $\tau \geq 0$, if it exists; the quantity

$$f_T(\tau) = -\frac{d}{d\tau} R_T(\tau)$$

is the probability density of $R_T(\tau)$ at τ . When $f_T(\tau)$ exists for (almost) all values of $\tau \geq 0$, then $R_T(\tau)$ is absolutely continuous, and $f_T(\tau)$ is called the failure model. Much literature in reliability pertains to ways of specifying failure models. Classic examples are the exponential with a scale parameter $\lambda > 0$, i.e.,

$$f_T(\tau | \lambda) = \lambda e^{-\lambda\tau}$$

and the Weibull with a scale (shape) parameter λ ($\alpha > 0$, i.e.,

$$R_T(\tau | \lambda, \alpha) = e^{-\lambda\tau^\alpha}$$

Other examples of failure models are surveyed in Singpurwalla (1995). The expected value of T

$$E(T) \stackrel{\text{def}}{=} \int_0^\infty \tau f_T(\tau) d\tau < \infty$$

is called the *mean time to failure* (MTTF). For the exponential model

$$E(T | \lambda) = \lambda^{-1}$$

A concept that is specific and unique to reliability is the failure rate function or the hazard function. For an

absolutely continuous $R_T(\tau)$, the failure rate function $h_T(\tau)$, $\tau \geq 0$, is

$$h_T(\tau) \stackrel{\text{def}}{=} \frac{f_T(\tau)}{R_T(\tau)}.$$

The failure rate function derives its importance from two features, one interpretative and the other, technical. Specifically, since

$$h_T(\tau) d\tau \approx P(\tau \leq T \leq \tau + d\tau | T \geq \tau)$$

the failure rate at τ is (approximately) the probability of an item's failure in $[\tau, \tau + d\tau)$, were the item surviving at τ . Clearly, for items that age with time, $h_T(\tau)$ will increase with τ , and vice versa for those that do not. This is the interpretative feature. The technical feature pertains to the fact that if

$$H_T(\tau) \stackrel{\text{def}}{=} \int_0^\tau h_T(s) ds$$

then

$$R_T(\tau) = \exp(-H_T(\tau))$$

thus, knowing $h_T(\tau)$ is equivalent to knowing $R_T(\tau)$ and vice versa. The quantity $H_T(\tau)$ is known as the cumulative hazard at τ , and $H_T(\tau)$ as a function of τ is known as the cumulative hazard function. Biostatisticians like Kalbfleisch and Prentice (1980) have used a continuously increasing stochastic process, like the gamma process, to describe $H_T(\tau)$ for items operating in a random environment. The connection between $H_T(\tau)$ and the Lorenz curve of econometric theory (see Gastwirth 1971) has been noted by Chandra and Singpurwalla (1981). This connection suggests that concepts of reliability have relevance to econometrics vis-à-vis measures of income inequality and wealth concentration.

It is easy to verify that when

$$R_T(\tau | \lambda) = e^{-\lambda\tau}$$

$\lambda > 0$, $\tau \geq 0$, $h_T(\tau | \lambda) = \lambda$, and vice versa. Thus, under an exponential model, the failure rate is the reciprocal of the MTTF and vice versa. However, this reciprocal relationship holds only for the exponential, and not for all other distributions as practitioners often assume. An attractive feature of the Weibull distribution is that by choosing α to be greater (smaller) than one, the failure rate function can be made to be increasing (decreasing) with τ . When $\alpha = 1$, the Weibull becomes an exponential. Because of this flexibility, the Weibull distribution has become one of the most widely used models in engineering reliability and survival analysis.

The failure rate function has become a cornerstone of the mathematical theory of reliability. Specifically,

all models whose failure rate increases (decreases) monotonically have been classified into one group called the IFR (DFR) class (for increasing (decreasing) failure rate), and properties of this class have been studied. All properties are in relation to the exponential which is both IFR and DFR, since increasing (decreasing) in IFR (DFR) is taken to be non-decreasing (nonincreasing). The famous 'bath-tub curve' of reliability engineering pertains to a distribution whose failure rate is initially decreasing, then becomes a constant, and finally increases, just like an old fashioned bath-tub. This functional form is appropriate for describing the life-length of humans, and large systems of many components.

The mathematical theory of reliability has many interesting results, several of which are intuitive, but some not. One such unintuitive result states that a mixture of exponential distributions (which have a constant failure rate) has a decreasing failure rate. Another counterintuitive result states that the time to failure distribution of a parallel redundant system of components having exponentially distributed life-lengths, has an increasing failure rate, but is not necessarily monotonic. Both these results appear in Barlow and Proschan (1975), but the arguments used to prove them are purely technical. Their intuitive import is apparent only when we adopt the subjective view of probability; Barlow (1985) makes this point clear.

The reliability assessment of systems of components is relatively straightforward if the component's life-lengths are assumed to be independent, the component's reliabilities are assumed, and the system is not a network. When the component reliabilities are unknown, the life-lengths are dependent. This situation becomes even more complicated when the system is a network. Lynn et al. (1998) proposed a Monte Carlo approach for treating such problems. Dependence in system reliability can be treated by multivariate failure models of the type introduced by Marshall and Olkin (1967). However, the number of parameters of such models grows exponentially with the size of the system, so that even for moderate size systems a use of multivariate models becomes an onerous task. One way of overcoming this difficulty has been to assume that the component life-times are associated and then to argue that reliability assessments assuming independence provide bounds for the assessed reliabilities if dependence was incorporated. This strategy may be suitable for small systems, but with large systems the lower (upper) bound tends to zero (one), so that the bounding is effectively meaningless. There is a pressing need for new multivariate models with a small number of parameters; an example is in Singpurwalla and Youngren (1993).

Thus far, the discussion has been restricted to the case of a single index of measurement, namely time or some other unit of performance, such as miles. In many applications, both engineering and biomedical,

the survival of an item is indexed by two (or more) scales. For example, automobiles under warranty are indexed by both time and miles. In the biomedical scenario, the onset of disease is recorded with respect to age and also the amount of exposure to a hazardous element. Failure models involving more than one scale are, therefore, germane and initial progress on this topic is currently underway. Such models are known as failure models with multiple scales; it is important not to confuse these models with multivariate failure models. The former pertain to a single unit, whereas the latter to multiple units.

5. Life Testing: Its Role in Reliability Assessment

Life testing is a scientific and controlled way of obtaining data. For those who do not subscribe to the personalistic view of probability, data are mandatory for reliability assessment, and this includes the selection of a model. The data could be from a life test or from the field. For those who subscribe to the personalistic view, the actual data is desirable but not essential. Desirable because the role of the data is to update ones initial assessment of reliability. This is achieved by using Bayes's law. The initial assessment of reliability is done by subjectively specifying a model and a prior distribution for parameters of the model, the prior also being subjective. The term subjective specification includes the use of expert testimonies (see, for example, Kadane and Wolfson 1998). The role played by the data, be it from a life test or the field, is to update the prior distribution to the posterior distribution.

Most life testing experiments are statistically designed experiments wherein several identical copies of the item whose reliability is being assessed are observed under conditions that replicate those of actual use. The life test is conducted until all items under observation fail or until a prespecified number fail. When the latter is done, the life test is said to be censored. The test can also be such that it is terminated after a fixed amount of time, irrespective of how many items fail in that time; such a life test is said to be a truncated test. A truncated test is terminated as soon as the last surviving item on the test fails. A life test can also be a combination of censoring and truncation, i.e., the test terminates at a fixed time or after a fixed number of failures, whichever occurs first.

When life tests are time consuming, it is common to make the conditions of the test harsher than those encountered in actual use. Such tests are called accelerated life tests, and tests with continuously increasing stress are those in which acceleration is done continuously (see De Groot and Goel 1979). Whereas accelerated life tests produce early failures, their disadvantage is that acceleration may induce changes in the material properties of the item, so that

the failure behavior is different. Accelerated life testing has a parallel in the biological context; it goes under the name of dose-response experiments.

Whereas censoring and truncation can be well controlled in laboratory tests, actual data may have arisen from either scenario, and the analyst may not know from which. This is known as the stopping rule problem (see, for example, Roberts 1967). Interestingly, for a subjectivist, the stopping rule is immaterial (provided that it is noninformative). However, for those who do not subscribe to the Bayesian point of view, knowledge of the stopping rule is fundamental. In the absence of such knowledge they cannot proceed. A very readable account of the stopping rule problem is in Lindley and Phillips (1976). It is because of these and the other arguments mentioned before that reliability in the twenty-first century has to be not only Bayesian, but also subjective Bayesian!

See also: Causal Inference and Statistical Fallacies; Quality Control, Statistical: Methods; Reliability: Measurement; Survival Analysis: Overview

Bibliography

- Barlow R E 1985 A Bayes explanation of an apparent failure rate paradox. *IEEE Transactions Reliability* **R-34**: 107–8
- Barlow R E, Proschan F 1965 *Mathematical Theory of Reliability*. Wiley, New York
- Barlow R E, Proschan F 1975 *Statistical Theory of Reliability and Life Testing*. Holt, Rinehart and Winston, New York
- Block H W, Savits T H 1997 Burn-in. *Statistical Science* **12**(1): 1–13
- Chandra M, Singpurwalla N D 1981 Relationships between some notions which are common to reliability theory and economics. *Mathematics of Operations Research* **6**(1): 113–21
- Davis D J 1952 An analysis of some failure data. *Journal of the American Statistical Association* **47**(258): 113–50
- Epstein B, Sobel M 1953 Life testing. *Journal of the American Statistical Association* **48**(263): 486–502
- Etzioni R, Kadane J B 1993 Optimal experimental design for another's analysis. *Journal of the American Statistical Association* **88**(424): 1404–11
- de Finetti B 1964 Foresight: Its logical laws, its subjective sources. In: Kyburg Jr. H E, Smokler H E (eds.) *Studies in Subjective Probability*. Wiley, New York
- Gastwirth J L 1971 A general definition of the Lorenz curve. *Econometrica* **39**(6): 1037–9
- Gnedenko B V, Belyayev Y K, Solov'yev A D 1969 *Mathematical Methods of Reliability Theory*. Academic Press, New York
- Grenander U 1956 On the theory of mortality measurements, Part II. *Skand Akt* **39**: 125–53
- De Groot M H, Goel P K 1979 Bayesian estimation and optimal designs in partially accelerated life testing. *Naval Research Logistics Quarterly* **26**: 223–35
- Kadane J B, Wolfson L J 1998 Experiences in Elicitation. *Journal of the Royal Statistical Society, Series D (The Statistician)* **47**(1): 3–19
- Kalbfleisch J D, Prentice R L 1980 *The Statistical Analysis of Failure Time Data*. Wiley, New York
- Kao J H K 1956 A new life-quality measure for electron tubes. *IRE Transactions Reliability and Quality Control* **PGRQC-7**

- Kotz S, Singpurwalla N D 1999 On a bivariate distribution with exponential marginals. *Scandinavian Journal of Statistics* **26**: 451–64
- Lindley D V 1985 *Making Decisions*, 2nd edn. Wiley, London
- Lindley D V, Phillips L D 1976 Inference for a Bernoulli process (a Bayesian view). *American Statistics* **30**: 112–9
- Lindley D V, Singpurwalla N D 1986 Multivariate distributions for the life lengths of components of a system sharing a common environment. *Journal of Applied Probability* **23**(2): 418–31
- Lindley D V, Singpurwalla N D 1991 On the evidence needed to reach agreed action between adversaries with application to acceptance sampling. *Journal of the American Statistical Association* **86**(416): 933–7
- Lynn N J, Singpurwalla N D 1997 Comment: 'Burn-in' makes us feel good. *Statistical Science* **12**(1): 13–9
- Lynn N J, Singpurwalla N D, Smith A 1998 Bayesian assessment of network reliability. *SIAM Review* **40**(2): 202–27
- Marshall A W, Olkin I 1967 A multivariate exponential distribution. *Journal of the American Statistical Association* **62**: 30–44
- Meeker W Q, Escobar L A 1998 *Statistical Methods For Reliability Data*. Wiley, New York
- National Research Council 1998 *Improving the Continued Airworthiness of Civil Aircraft. A Strategy for the FAA's Aircraft Certification Service*. National Academy Press, Washington, DC
- Roberts H V 1967 Informative stopping rules and inferences about population size. *Journal of the American Statistical Association* **62**(319): 763–75
- Savage L J 1972 *The Foundations of Statistics*, 2nd rev. edn. Dover Publications, New York
- Singpurwalla N D 1988 Foundational issues in reliability and risk analysis. *SIAM Review* **30**(2): 264–82
- Singpurwalla N D 1995 Survival in dynamic environments. *Statistical Science* **10**(1): 86–103
- Singpurwalla N D, Wilson S P 1999 *Statistical Methods in Software Engineering*. Springer, New York
- Singpurwalla N D, Youngren M A 1993 Multivariate distributions induced by dynamic environments. *Scandinavian Journal of Statistics* **20**: 251–61
- US Nuclear Regulatory Commission 1975 Reactor Safety Study—An Assessment of Accident Risks in US Commercial Nuclear Power Plants. WASH-1400 (NUREG-75/014) October

N. D. Singpurwalla

Quality of Life: Assessment in Health Settings

1. Quality of Life

Since the 1980s or so there has been an increasing realization that traditional biologically based endpoints such as morbidity and mortality alone do not represent adequately the potential outcomes of medical interventions. Health status measurement has evolved to allow insight into patients' experiences in such areas of function as mobility, mood, life satisfaction, sexuality, cognition, and ability to fulfil

occupational, social, and family roles. Quality of life (QoL) has emerged as a broad term to describe this domain of measurement. The QoL construct may be viewed as a paradigm shift since it shifts the focus of attention from symptoms to functioning and establishes the primacy, or at least the legitimacy, of the patient perspective. QoL measures have many applications in medicine and healthcare. They are used to describe the subjectively perceived health and social status of given populations, to compare interventions and to assess the costs and benefits of treatments and health policies (Spilker 1996). Despite the increasing popularity of QoL, there is dissent about the meaning of the term, how it should be measured, and indeed whether it should be measured at all (Hunt 1997). In a critical review, Gill and Feinstein (1994) found that investigators defined what they meant by the term QoL in only 15 percent of the 75 articles reviewed. As Humpty Dumpty said to Alice: 'when I use a word, it means exactly what I want it to mean, nothing more and nothing less!'

2. Health-related QoL

Because of the amorphous and multidimensional nature of QoL, most researchers in medicine and healthcare concern themselves with a subcomponent of QoL called health-related QoL (HRQoL). This is distinguished from QoL as a whole, which would also include adequacy of education, housing, income, and perceptions of the immediate environment. Health status and HRQoL measures increasingly are used in studies addressing the costs and benefits of services and treatments (Patrick and Erickson 1993). The US Congress established the Agency for Health Care Policy and Research to undertake research on the effectiveness of medical care in terms of its impact on patient outcomes including HRQoL. The US Food and Drug Administration encourages the collection of HRQoL data for new drugs and, in the UK, health status and the experiences of patients and carers are included explicitly in outcome assessments.

3. Definition of Health-related QoL

Patrick and Erickson (1993) defined HRQoL as:

the value assigned to the duration of life as modified by the social opportunities, perceptions, functional states and impairments that are influenced by disease, injuries, treatments or policy.

Assessment of HRQoL usually focuses on physical function, psychological state, social function, somatic symptoms, sexual function, occupational function, and, occasionally, on financial state. Many existing measures frequently are criticized for including only concepts at the negative or illness end of the continuum while neglecting concepts at the more positive end.

4. Assessment of Health-related QoL

There are a number of approaches to the assessment of HRQoL. Most of these involve some form of subjective multidimensional assessment preferably completed by the respondent whose QoL is under scrutiny since levels of agreement among doctors and between doctors' and patients' judgments are usually low (Sprangers and Aaronson 1992). Measures include generic instruments such as health profiles and utility measures, disease-specific questionnaires, and individually focused measures.

4.1 Generic Measures

Generic measures, such as the SF-36, The Nottingham Health Questionnaire, the McMaster Health Index Questionnaire, and The Sickness Impact Profile are broadly-based questionnaires and can be applied in a wide range of conditions and populations. The main limitation of generic questionnaires is that they are broad measures of health status. Consequently, they may not provide sufficient information on specific aspects of a given disease which are likely to be important in assessing disease impact and the outcome of treatment. It is common practice to use a generic measure together with a disease-specific measure. It is likely that the use of generic instruments will diminish as more sophisticated disease-specific measures are developed.

4.2 The World Health Organization QoL Instrument (WHOQOL)

The concept of HRQoL owes much to the original World Health Organization (WHO) definition of health as a state of complete physical, mental, and social well-being and not merely the absence of disease. The WHO established a multinational group to develop a measure that would allow for cross-cultural comparisons of QoL (WHOQOL 1993). The definition underlying the measure was:

QoL is defined as the individual's perception of their position in life in the context of the culture and value systems in which they live and in relation to their goals, expectations, standards and concerns. It is a broad ranging concept affected in a complex way by a person's physical health, psychological state, level of independence and their relationships to salient features of their environment.

The WHOQOL is a somewhat unusual measure of HRQoL because, in addition to measuring physical and psychological health and social relationships, it also measures spirituality and it deals with facets of the environment such as financial resources and home environment. In developing the WHOQOL, the authors unexpectedly found that the basic factors inherent in QoL did not differ substantially across cultures (Power et al. 1999).

4.3 Utility Measures

Utility measures are generic measures derived from economics and decision theory and represent patients' preferences for different health states in the form of a single summary score. Their key characteristics are that they are based on preference judgments and usually are measured along a continuum from death (0.0) to perfect health (1.0), although scores less than 0 representing states worse than death are possible. Utility scores represent both health status and the value of that health status to the patient. They are used in cost-utility studies in which the cost of an intervention is related to the utility value afforded by the intervention. Various approaches to measuring utilities have been developed. In the Kaplan Index of Well-being Scale, case descriptions were compiled to illustrate combinations of functional levels, symptoms or problems and these cases were rated by random samples of the public who gave their preference ratings which were then used to determine utility values. In the somewhat controversial QALY (quality-adjusted life years) approach, improvements in the length and quality of life are amalgamated into a single index. In the standard-gamble approach, patients are asked to assess the risks they would tolerate for certain medical interventions. In the time-trade-off method, respondents are asked to estimate their relative preference for quality vs. quantity of life.

The utility approach assumes that individuals are capable of reporting their preferences and that they can predict accurately these preferences in relation to some future putative health scenario. It is further assumed that the preferences would not change if the patient were actually to experience such a health scenario. An additional problem relates to who should provide the utility values—the general public, health-care providers, and/or patients and their families. One response to these problems has been the development of the Euroqol (Euroqol Group 1990). The aim of the Euroqol is to provide a standardized, nondisease specific survey instrument for describing HRQoL and to generate a single health index for each health state. The measure, however, suffers from a number of logistical and methodological limitations (Bowling 1995).

4.4 Disease-specific Measures

The second approach to QoL assessment focuses on aspects of health status that are specific to the area of primary interest. It is assumed that responsiveness will be increased by including only important aspects of HRQoL that are relevant to the patients being studied. The measure may be specific to a disease (such as cancer or arthritis), to a population of patients (frail elderly), to a certain function (sexual function or sleep), or to a particular problem (such as pain). Specific measures focus on areas that are likely to be

explored routinely by clinicians and are likely to be more responsive to small but clinically significant changes. The modular approach adopted by the European Organization for Research on Treatment of Cancer (EORTC) provides a core generic measure for use in a variety of cancers supplemented by disease-specific modules for use in particular types of cancer (see *Cancer: Psychosocial Aspects*).

4.5 Limitations of Health-related QoL

The restriction of QoL assessment to HRQoL and the use of health-status measures to assess HRQoL is not entirely satisfactory. Most of the questionnaires used are essentially measures of health status in a new guise and these place an overwhelming reliance on the assessment of functional capacity. As such, while purporting to incorporate patients' perspectives, they represent implicitly a medical model that stresses the ability to perform everyday tasks and fulfil social and occupational roles. Such measures ignore the meaning and importance of such tasks and roles for the individual and often preserve the supremacy of professional judgements leading to the suppression of what is supposed to be under scrutiny (Joyce et al. 1999, Hunt 1997).

5. Individual QoL Measures

A number of researchers have argued that QoL depends on the unique interpretation and perceptions of the individual and that questionnaires represent an oversimplification of what is a complex, multi-dimensional, subjective phenomenon. QoL is seen as a uniquely personal perception, denoting the way that individuals feel about their health status and/or nonmedical aspects of their lives. It can be measured suitably only by determining the opinions of patients and by supplementing (or replacing) the instruments developed by 'experts.' This hermeneutic approach would seek to understand individuals as self-reflective beings who are responsible for their own actions and who are the best judges of their QoL. This approach has resulted in the development of a number of individualized measures of QoL such as the Schedule for the Evaluation of Individual QoL or SEIQoL (Joyce et al. 1999).

6. Future Trends and the Need for Theoretical Models

Assessment of patient QoL facilitates improved clinical intervention, assists in treatment comparisons, and should prove increasingly important in the identification of services and facilities and in resource allocation. However, much of the research, to date,

has been underpinned by somewhat tautological operational definitions of QoL. There is a growing need for theoretical models that capture the psychological reality and complexity of QoL. For example, most current measures fail to address adequately the dynamic nature of QoL. Patients confronted with a life-threatening or chronic disease are faced with the necessity to accommodate to their illness. An important mediator of this adaptation process is 'response shift' which has been defined by Schwartz and Sprangers (2000) as changing one's internal standards, values, or one's conceptualization of QoL. Response shift is but one example of a psychological process that underpins the essentially dynamic and phenomenological nature of QoL. Research in this area must incorporate psychological variables into theory construction in order to produce a body of research that more validly reflects the true nature of QoL.

QoL has, heretofore, been considered solely as an outcome measure in health research. However, exciting developments in psychoneuroimmunology and evidence from a number of clinical studies raise the intriguing possibility that QoL might influence pathological processes such as tumor progression. If this were proven, then interventions aimed at maximizing patient QoL might also influence patient recovery.

See also: Cancer: Psychosocial Aspects; Chronic Illness, Psychosocial Coping with; Chronic Illness: Quality of Life; Health Outcomes, Assessment of

Bibliography

- Bowling A 1995 *Measuring Disease: A Review of Disease Specific Quality Of Life Measurement Scales*. Open University Press, Buckingham, UK
- Euroqol Group (Buxton M, O'Hanlon M, Pekurinen M ym) 1990 Euroqol: A new facility for the measurement of health-related quality of life. *Health Policy* **16**: 199–208
- Gill T M, Feinstein A R 1994 A critical appraisal of the quality of QoL measurements. *Journal of the American Medical Association* **272**: 619–26
- Hunt S M 1997 The problem of QoL. *Quality of Life Research* **6**: 205–12
- Joyce C R B, O'Boyle C A, McGee H M 1999 *Individual QoL: Approaches to Conceptualization and Assessment*. Harwood, The Netherlands
- Patrick D L, Erickson P 1993 *Health Status and Health Policy*. Oxford University Press, Oxford, UK
- Power M, Bullinger M, Harper A, WHOQOL Group 1999 The World Health Organization WHOQOL-100: Tests of the universality of life in 15 different cultural groups worldwide. *Health Psychology* **18**: 495–505
- Schwartz C E, Sprangers M A G 2000 *Adaptation to Changing Health: Response Shift in QoL Research*. American Psychological Association, Washington, DC
- Spilker B (ed.) 1996 *Quality-of-life and Pharmacoeconomics in Clinical Trials*, 2nd edn. Lippincott-Raven, Hagerstown, MD
- Sprangers M A G, Aaronson N K 1992 The role of health care providers and significant others in evaluating the quality of

life of patients with chronic disease: A review. *Journal of Clinical Epidemiology* **45**: 743–60

WHOQOL Group 1993 *Measuring the Quality of Life: The Development of the World Health Organization Quality of Life Instrument*. WHO, Geneva, Switzerland

C. A. O'Boyle

Quangos in Politics

Quasi-governmental (q-g) bodies exist in their thousands, particularly in the developed countries. Their degree of quasi-official standing varies greatly. A simple typology, with examples, would offer (a) direct public services by national government ('g') itself (e.g., US Post Office); (b) indirect service by a state industry ('q-g') (e.g., the UK's Royal Mail); (c) a more indirect service by a wide variety of 'agencies,' 'quangos' (the loose British term) or 'public bodies.'

This wider q-g class extends to commercial firms and other private bodies which contract, in some sense, with government to provide a regulated service, perhaps in a privileged market position (telephone, bus or train companies in many countries). Given their official legal status as companies, foundations, charities, etc., these latter deliverers of public services are more 'quasi-private' than quasi-governmental.

When the state creates or sponsors a body to act as its agent, problems of democratic accountability are raised and can only be managed politically rather than solved by any particular method of appointment or regulatory regime. These relationships do not involve full privatization (sale of public assets into the private sector) but may involve contracting public sector work to the private sector or to this third sector of q-g bodies.

1. The Quasi-governmental Sectors

In the more advanced and elaborate political systems, quasi-governmental (q-g) bodies have increasingly come to deliver or even devise public policies and services on a noncivil service basis. Managerial assertions that the q-g approach to public services offers better quality have claimed intellectual support from organization theory. Political assertions that the state should itself perform fewer tasks in general and avoid some particular types of task altogether have arisen from antistatist, procommercial or procivil society values. Separating practical issues of how to improve public policy design and delivery from ideological and normative issues of what the state should be doing with its direct budgets and civil service cadres is an essential step in understanding the q-g phenomenon.

Any body which is not part of central or local government may be intended (or may come) to act in an q-g role. Its formal standing may be as an official body ('agency' or, in the UK, 'quango' – quasi-autonomous national government organization), perhaps specifically created by law ('statutory') with its own controlling board. It may be a corporation (company, firm) whether 'for-profit' (commercial) or 'not-for-profit' (philanthropic, charitable). It may be a voluntary membership organization made up of individuals, corporations (a trade association) or other types of body such as schools or colleges. Whatever it once was or may appear to still be, any body which stands outside the formal state bureaucracy may come to behave as a quasi-official extension of state authority. This process commonly has a legal basis (the body undertakes government contracts to provide public services) or a financial one (receiving official grants for its charitable, artistic, or equivalent work). Occasionally, a private body becomes, in effect, quasi-governmental without contracts or other money passing to it but on a simply political and psychological level: the body helps the government by providing information and access to a problem area or target group and is rewarded with consultative status which draws it close to official policies and thinking.

2. Government-appointed Bodies

The rise of a quasi-governmental sector in advanced political systems displays two broad types of body within great variety of detail: 'public bodies' created by governments to do a policy job on a q-g basis, and existing private bodies (commercial, philanthropic, representative, or fraternal) which become significantly q-g in their attitude and behaviour. The former group is always more prominent and significant for public service delivery and therefore also for the problem of 'independence versus accountability' which arises from unelected bodies disposing of public functions and budgets. These bodies are distinguishable from central or local government's own bureaucratic cadres by their separate titles, budgets, governing boards and (latterly) development and 'business' plans, performance targets and other fashionable banners (even though mainstream civil services now often also display them). It is the separate, government-appointed and paid board which most clearly distinguishes these q-g bodies from the government's own direct operations because this board is a legal entity and employer, doing its work through its own staff and contracted or cooperative links with other bodies, including commercial suppliers.

Various terms have been applied since the q-g sector first attracted systematic academic attention some 30 years ago (Smith and Hague 1971, Smith 1975). The term 'the contract state' reflected the US scene with its easier acceptance of business-type contracts for public

programs, notably during and after the War on Poverty period. British assumptions (and therefore language) about public services eschewed talk of contracts, although a few had long existed, in public education, for example (Hague et al. 1975). Different theories of the essential exclusivity of state functions were encouraging or discouraging new forms of 'agency' or 'intermediary body' in different European systems, notably in the more pluralist Netherlands or Scandinavia and the more statist France or Italy, respectively.

The British tradition combined a strong state theory, stewarded by a very powerful senior civil service, within a quite rich civil society context—including, for example, the senior professions and leading charities at national level. Nominated public bodies with specialist remits and only generalized 'upwards' accountability to government ministers (and none 'downwards' to users or recipients of their services) were quite familiar devices. In this soil, the British 'quango' form multiplied (Barker 1982). The flexibility and precise control offered by a single-purpose public body was generally irresistible to both parties in government and to their senior civil servants who continually steer and supervise these bodies. By 1996, a properly functional independent analysis of British appointed public bodies found some 6,400 being funded (and therefore in some way controlled) by central government. The 5,750 executive (cf. purely advisory) bodies were judged to have spent some £60B of public money in 1994–5, an increase of 50 percent at constant prices in 16 years, following continual transfers of expenditure functions from central and local governments to new nominated bodies and the creation of new programs under these 'quango' auspices (Hall and Weir 1996).

3. The Politics of Quasi-governmental Bodies

The most ideological and most partisan of political debates about q-g bodies deliberately created by the state seems to have arisen in the UK: the 'quango issue.' Superficially (and tediously) the two main political parties have, for 25 years, accused each other of abuse of the patronage involved. A more serious criticism of all appointees is their lack of accountability. Continental Europe seems to have been spared much of this British conflict, perhaps because business, labor unions, the churches and other civil society formations accept each other and expect to share in delivering public services through special agencies or boards, under broadly consensual political oversight. The Dutch have their three 'pillars' of social service delivery (Catholic, Protestant, and secular), while in Germany it is officially polite for organized capital and organized labor to call themselves 'social partners' (Barker 1997). The USA no more follows that custom than does the UK, but the vast proliferation of

appointed and joint bodies which constitute 'America's quangos' seems not to have sparked either union-business rivalry as to patronage or specific partisan hostility. The long-term right wing rumblings against 'big government' subsumes these many 'agencies' within broader issues of overall federal or state civil service numbers and public expenditure levels. Academic critiques have often also taken a broad view of problems and pathologies (Burns 1994, Carter 1994, Kettl 1988).

4. The Rationale of Quasi-governmental Bodies

The principal practical rationale for creating a new public body to tackle either a novel or a familiar policy task is its specialist character. The new single-purpose agency or quango has been expected to succeed whereas the old, multipurpose government department would wearily add this extra burden to its already overstretched bureaucracy. If the new body's budget and staff are also outside the official civil service total budget and numbers, and if a small, 'dedicated' (specialist) office can bring sensitivity rather than standardized bureaucracy to some novel public policy challenge (such as encouraging ethnic or gender rights, offering drug or HIV health care, making grants to artists and theaters, or promoting community development) then there is no contesting this way forward. The same applies where an existing private body is to be a vehicle for such novel tasks by becoming a government contractor or grantee in return for its experienced and expert work on policy design or delivery.

A further rationale may well apply when professional expertise is crucial to a policy or program. Senior scientists and other experts are very unlikely to become civil servants, much less join the military, and can therefore be employed by the state only on their own ground and terms. The US Air Force and US Army could not get the high-quality operations researchers or mathematicians they needed except by founding the RAND Corporation (initially to analyze bombing targets in the USSR) and an applied mathematics centre on the University of Wisconsin's campus, respectively. Even in the crisis of war, the assorted, mainly British, academic intellectuals who were recruited to break the Enigma and other enemy codes at Bletchley Park ('Station X') in 1940 or the scientists and mathematicians organized into the Manhattan Project to build the US atom bomb would perform their best, or even perform at all, only in a carefully managed academic and civilian setting, as their governments realized. These civilian-military contrasts were merely one example of the lessons of organization theory which have declared the 'flat' (not hierarchical) 'open' (not sectionalized) 'organic' or flexible and small agency or company to be the best setting for creative, innovative, interactive or sensitive

tasks. As modern governments have been pressed into policy work which is either technically demanding (requiring expert staff) or socially sensitive (requiring nontraditional official styles and methods) so the appeal of specialist, flexible, and suitably qualified staff working in a small single-issue body has steadily increased. With obvious differences in funding and controls, these basic characteristics make either a specifically created new public agency ('quango') or an existing private body drawn in to q-g status the attractive option.

5. The Dilemma of Independence vs. Accountability

The essence of the q-g problem is that practical reasoning favors a democratically responsive government trying to obtain good policy or policy delivery by creating or recruiting a specialist agency ('quango') to concentrate on that job. The trickier the task, in political, technical, or social terms, the greater the appeal of this method, in many cases at least. But this very difficulty and sensitivity prompts critics to claim that democratic accountability to elected representatives (national or local) is particularly needed. Having an unelected body of government appointees running (say) the local mass transit, or even regular public grants to theaters or art galleries may be tolerable but life and death decisions on (say) drug budget allocations to hospitals or decisions affecting employment must be taken directly by elected representatives. To produce the advantages of flexibility, innovation, and effective informality and to attract the most talented and expert staffs or consultants, a q-g body does need to be independent of day-to-day control.

This dilemma of operational independence (in search of creative quality of service) versus public accountability (in support of democratic legitimacy) can only be softened, not solved. A q-g body's initial pioneering independence could be time-limited and fuller reporting and answering of questions and debate then imposed upon it. Turnover of board members could be formally encouraged or even required. In general, an 'explanatory accountability' regime is needed for the army of q-g bodies which advanced political systems now maintain (AMA 1995). Although specific cases or issues are not normally removable or reversible, board members can be made to explain their views and decisions. Rules requiring openness, transparency, and full reporting would all illuminate this explanatory accountability. Even this working culture would inhibit some valuable creative independence in some policy fields. A statutory duty to explain itself to a possibly critical public would be a major challenge for any q-g body, particularly if the courts are also open to calls for their judicial review of any such explanations of actual decisions which appear to have been based on mistakes or illogical

argument. The standard work on the British position (Skelcher 1998) explores these ideas.

6. A New Sector of Government and Society

So many tasks of policy formation and delivery; of regulation of private activity 'affected with the public interest' (to use the traditional US phrase); and of social innovation are now required in advanced democratic societies that the rise of this middle sector, standing between state and private sectors can be no surprise. Its lack of much political controversy in these systems suggests a declining popular interest in democratic principles in favor of practical, consumerist attitudes to public services.

See also: Policy Knowledge: Nonprofit Institutions; Public Administration, Politics of

Bibliography

- AMA (Association of Metropolitan Authorities) 1995 *Changing the Face of Quangos: Proposals for the Reform of Local Appointed Bodies*. AMA, London
- Barker A (ed.) 1982 *Quangos in Britain: The Making of Public Policy Networks*. Macmillan, London
- Barker A 1997 Other people's quangos: Quasi-government in seven countries. In: Flinders M (ed.) *How to Make Quangos More Democratic*. Charter 88 Publications, London
- Burns N 1994 *The Formation of American Local Governments: Private Values in Public Institutions*. Oxford University Press, Oxford
- Carter S L 1994 *The Confirmation Mess: Cleaning-Up the Federal Appointment Process*. Basic Books, New York
- Hague D C, Mackenzie W J M, Barker A (eds.) 1975 *Public Policy and Private Interests: The Institutions of Compromise*. Macmillan, London
- Hall W, Weir S 1996 The untouchables: Power and accountability in the Quango state. Democratic Audit Paper #8 The Scarman Trust, London
- Kettl D F 1988 *Government by Proxy: Managing Federal Programs*. Congressional Quarterly Press, Washington, DC
- Select Committee on Public Administration (1999) Quangos Sixth Report, 1998–99, HC 209 – 1 and 2, 9 November 1999. The Stationery Office, London, Vols 1 and 2I
- Skelcher C 1998 *The Appointed State: Quasi-Governmental Organizations and Democracy*. Open University Press, Buckingham, UK
- Smith B L R (ed.) 1975 *The New Political Economy: The Public Use of the Private Sector*. Macmillan, London
- Smith B L R, Hague D C (eds.) 1971 *The Dilemma of Accountability in Modern Government: Independence Versus Control*. Macmillan, London

A. Barker

Quantification in History

'Quantification' refers to a specific method of historical research that departs from the prevalent prac-

tice of textual analysis by actually seeking to measure changes in the past. Due to the mass of data to be handled and the complexity of the calculations involved, the application of quantitative methods usually has to be *computer-assisted*. In contrast to a mere illustration of certain tendencies by tables or graphs, quantification tends to use *analytical statistics* in order to probe relationships between certain variables. Such statistical analysis also requires a *high degree of formalization* in the research design that formulates hypotheses to be tested and tries to combine them into models. The final aim of the quantitative approach is not just to explain a particular set of developments but to make a *theoretical contribution* to some facet of historical understanding.

1. The Rise of Quantification

Within historical research, the development of quantitative methods was a product of a particular constellation during the 1960s. Already during the Enlightenment some scholars had sought to describe socioeconomic developments by compiling time-series, and by the middle of the nineteenth century national governments took on the task of determining population trends through a series of censuses. But it took the shift to the 'new social history' a century later to direct historical interests away from affairs of state to mass politics, societal changes, demographic transformations, or economic growth. At the same time the arrival of mainframe computers, followed eventually by powerful PCs, made it possible to handle large amounts of data and to use complicated statistical procedures in customized form via statistical packages such as SPSS or SAS. In order to explore these exciting new subjects with novel techniques, historians became interested in the theoretical constructs of the social sciences, borrowing eclectically from political science, sociology, or demography and somewhat less so from economics. The result of this interaction was the rapid development of a new hybrid, combining elements of historical research with social scientific analysis.

During the 1970s, this 'social science history' or 'historical social science' was successfully institutionalized. In the United States the multitude of initiatives in history and the neighboring disciplines combined in founding the Social Science History Association in 1975, and made its annual meeting a showcase of the most innovative work. In France many of the long-term community studies, produced by the *Annales School*, turned quantitative, compiling detailed time-series of population, prices, or wages and probing their inter-relationship. In Germany, the quantifiers linked with the theory-oriented project of a *Historische Sozialwissenschaft* to create a vigorous organization, called QUANTUM. Within the decade, a spate of new journals emerged to provide publication

outlets for the new work, such as the *Journal of Social History*, *Journal of Interdisciplinary History*, *Historical Methods Newsletter*, *Social Science History*, *Histoire et Mesure*, and *Historical Social Research*. New Textbooks (Dollar and Jensen 1971, Floud 1979) as well as influential anthologies (Aydelotte et al. 1972, Lorwin and Price 1972) provided a powerful platform for the propagation of quantitative research.

2. Research Contributions

The first field transformed by quantitatively-oriented work was the *new political history* that differed from its predecessor by its attention to collective phenomena (Bogue 1983). For instance, Michael Kater investigated the rise of the Nazi movement in Germany by analyzing its membership files, indicating precisely which social stratum, religious group, gender, region, etc. joined the bandwagon at what time. For democratic societies like the US, Lee Benson's or Alan Bogue's analysis of electoral behavior proved quite productive, since comparing returns over time made it possible to establish which polls were the 'critical elections' when party realignments were transformed by 'ethno-cultural' or other factors. For parliaments without party discipline, Thomas B. Alexander's roll-call analysis also offered a powerful instrument for establishing who actually voted with whom, indicating what networks dominated legislative behavior. This retro-projection of political science methods gradually provided a firmer empirical basis for an analysis of the transformation of mass politics in the last two centuries.

The second area that was deeply affected by quantitative methods and conceptual borrowing from sociology was the *new social history* (Stearns 1985). As 'history from the bottom up,' this perspective examined previously untouched serial records, such as police files, employment registers, or draft books in order to infer from the behavior of ordinary people what they might have felt and thought. The appeal of this effort was the diffusion of its topics, since the labor history of the Tillys included subjects as different as proletarian living conditions and their collective efforts at protesting through strikes. Another question, treated e.g., by Hartmut Kaelble concerned the pattern of social mobility, comparing the relative openness of societies to upward or downward movement over time or across borders. A related issue was the societal impact of education that required an analysis of the enrollment and composition of schools or universities in order to get at changes in educational opportunities. These quantitative probes produced much new and sometimes conflicting information about the lower orders, the make-up of institutions, and basic trends in societal transformations.

A third area propelled by computer-assisted source collection and statistical analysis was the field of

historical demography (Hareven and Vinovskis 1978). For time-periods before national censuses, the demographers faced the challenge of reaggregating vital statistics from the ground up by converting parish registers into reliable measures of natality, nuptiality, or mortality. Under the direction of Anthony Wrigley the Cambridge population project established the outlines of a population history for England, while the demography center in Umea undertook a national project of quantifying church records so as to explain demographic changes in Sweden. American demographers like John Demos and Tamara Hareven instead combined individual life-event data to scrutinize the development of family structures from extended via stem to nuclear families and the like. Other scholars such as Roderick Floud and Richard Steckel studied changes in the height of military recruits or slaves in order to develop more accurate measures of social deprivation. This historical demography amassed a source base for the prestatistical area and produced interesting interpretative concepts for explaining the population transition.

The final area in which quantification transformed research priorities and methods was the *new economic history* (Davis and Engerman 1987). Redirecting attention away from entrepreneurship, 'cliometrics' as it was also called, inspired efforts to establish indicators of national production and income for the prestatistical age as well as to transform information from manuscript censuses into published form. In the US, a heated debate revolved around the profitability of slavery, with Robert Fogel's and Stanley Engerman's claims producing much controversy, due to their combination of economic and moral argumentation. Much attention also focused on the key factors of economic development, such as railroads, with some scholars actually arguing that their impact had been exaggerated. In Europe discussions revolved around the pattern of industrialization instead, with some specialists like Sidney Pollard proposing a more complex regional diffusion model. On the whole, the quantitative impulse made economic history more rigorous and theoretical, but removed it at the same time further from the mainstream of historical debate.

By the 1980s these signal contributions had managed to overcome traditionalist skepticism and to establish quantification as a new and growing subfield within the historical discipline. Though they were sometimes difficult to read for laymen, high-quality monographs convinced reluctant granting agencies to fund major collective initiatives, such as the Philadelphia Social History Project, directed by Theodore Hershberg. The number of articles with graphs and tables in respected historical journals began to reach almost one-quarter (Kousser 1989). Opinion surveys of the profession suggested that most scholars expected quantification to be here to stay. Even if some critics still bemoaned the spottiness of graduate training in this area, the rapid rise of quantitative methods in

historical research had been impressive and their future prospects seemed secure.

3. *The Culturalist Challenge*

During the decade of the 1980s, however, the seemingly unstoppable progress of the quantitative history 'project' began to stall. By the middle of the decade, the number of quantitative journal articles and the frequency of graduate courses in quantitative techniques suddenly began to drop (Reynolds 1998). It seems that over the past 10 years (plus or minus five!) quantitative history has become increasingly marginalized, since a variety of other methods have risen to prominence, making so-called 'quants' seem rather quaint. Ironically, the demise of quantitative history has occurred simultaneously with the development of enhanced technology, more complete datasets, and increasingly sophisticated models and methods, which, taken together, have rendered quantitative history *potentially* more useful than ever before.

The factors responsible for this loss of attention are still being debated vigorously. Some scholars attribute the problems afflicting quantitative history to the initial overselling of the method by insufferable practitioners who promised more insights than their research was able to deliver. These critics, comprised largely of traditionalists, have always been suspicious of the proposition that history had anything to learn from the social sciences and have long bristled at heady sentiments such as that expressed by Emmanuel Le Roy Ladurie who predicted in 1967 that in the future 'the historian will be a programmer or he will be nothing' (Rabb 1983). No wonder that such skeptics gladly seized upon Bernard Bailyn's and Lawrence Stone's calls for a revival of the narrative.

Another reason was the cultural turn of many younger historians who rejected both the materialist premises and positivist vision of most practitioners of quantitative history. To these critics, quantitative representations of the past were at once too cold, too remote, too impersonal, and too abstract. Large datasets were said to dehumanize, and 'soulless' quantifiers were assumed to be ready, like Marx's bourgeoisie, to drown all human sentiment 'in the icy water of egotistical calculation' (Coclanis 1992). Moreover, in making fetishes of rigor and precision, quantitative historians purportedly attempted to exact more from the vast, vague, largely unknowable past than historians had any right to claim. Instead, the past was reconfigured as a 'foreign country,' to invoke the title of historian David Lowenthal's influential 1985 book.

The changing demographics of the historical profession, particularly in the United States, also served to reinforce the trend towards qualitative approaches.

As history opened up to women and minorities—groups, generally speaking, less inclined to pursue quantitative approaches to the past—quantitative history could be expected to decline in a relative sense. Just as women and minorities have been under-represented in the *hard* sciences over the years, they have proven less interested in the more social scientific sides of history. Though both groups have embraced social history as a vehicle for their identity politics, relatively few of their members can be found in fields such as economic history, historical demography, or quantitative political history.

During the 1980s, many historians turned instead to cultural anthropology for methodological inspiration. Rejecting the rigorous, formal and remote, data-driven 'truths' sought by quantifiers, scholars interested in experience and meaning preferred in-depth microstudies, whether of single villages, small groups, or even well-documented individuals. Gradually, 'thick description' came to trump ecological regression, crude death rates, and multiple R-squares, no matter how robust! The philosophical basis of this shift to anthropology was the rise of postmodernism, which deprecated the rationalism, universalism, and empiricism of historical social science. By the late 1990s the ascendancy of cultural studies left quantifiers off in a corner silently counting all by themselves.

4. *The Continued Importance of Quantification*

The current disregard of quantitative methods in the discipline is unfortunate, since quantification continues to be an essential technique for historical research. Some questions can best be answered through recourse to what Sir William Petty referred to in the seventeenth century as 'Number, Weight, or Measure,' and it is more than 'physics envy' that has drawn scholars to quantitative analyses of the past. Even after conceding that many early practitioners were a bit overly zealous in their claims, it seems fair to suggest that quantitative history can provide the methodological discipline, intellectual rigor, and moral purpose that subjectivist forms of history often lack. With all due respect to such influential postmodernist thinkers such as Baudrillard and Lyotard not every aspect of the past is positional, free-floating, and contextual.

The utility, indeed, the necessity of quantitative history seems more compelling in light of important new developments in the field. Technological constraints have eased considerably, for example, as miniaturization and digitalization in the electronics industry have advanced. Today's personal computers, programmable calculators, and other platforms offer tremendous power and, thus, computational possibilities to quantitative historians. Similarly new developments in statistical packages, spreadsheets, CD-ROM

technology, GIS (geographic information systems), and the like have vastly expanded the analytical reach of software that can be applied to historical questions.

The increasing availability of new and/or enhanced sources and datasets also presents unprecedented research opportunities. Materials readily retrievable from established institutions such as the ICPSR (Inter-university Consortium for Political and Social Research) in Ann Arbor, Michigan, and data available on-line from the World Bank, the IMF, and other governmental, quasigovernmental, and nongovernmental organizations and agencies are transforming the nature (and sites) of scholarly production. The new CD-ROM version of the *Historical Statistics of the United States*, datasets assembled by Yale's Cambodia Genocide Project, the University of Minnesota's IPUMS (Integrated Public Use Micro Sample) Project, drawn from US census records, and the East German central cadre files—to cite but four examples—are of similar scholarly importance, and have tremendous potential as teaching tools. Various research projects undertaken under the auspices of, or promoted by the Association for History and Computing, founded in England in 1987, have proven extremely useful as well.

At the same time, quantitative historians have in recent years developed more sophisticated methods, allowing them to analyze and interpret historical questions with greater sensitivity and precision. Economic historians, for example, have utilized both game-theoretical approaches and advanced econometric techniques to excellent effect, and anthropometric historians have employed important new procedures from auxology and biostatistics to facilitate their reconstruction of our biological past (Steckel 1995). Economic and social historians, along with historical sociologists, have also begun to employ regression analysis in more subtle and refined ways, using temporally recursive regression procedures, when appropriate, instead of standard time-series regression techniques.

Other quantitative historians have begun to 'push the envelope' in different ways, in so doing, often trespassing on the territory of scholars in other social sciences or even in the humanities. Thus, as procedures such as social network analysis (SNA), event-structure analysis, and quantitative language analysis are successfully employed in answering historical questions—quantitative analyses of major 'texts' in order to produce so-called semantic grammars, for example—non- and anti-quantifiers ignore the results at their peril. In addition, quantitative historians' recent incorporation of spatial relationships into historical analysis through spatial-effects regression models and the like has produced similarly impressive results (Griffin and van der Linden 1998).

At the beginning of the twenty-first century, the future prospects of quantification in history are therefore indistinct. Clearly, individual quantitative

historians, among the profession's most avid proponents of scientific rationalism, universalism, and empiricism, are continuing to expand the frontiers of their research and are still making important contributions. But whether or not the interest of mainstream historians will return to quantification and the subfield as such will ever again recapture the excitement and energy it possessed in the 1970s remains an open question. The relative popularity of various approaches does not just depend on the quality of their monographic work but on the larger tides of intellectual fashion. However, one unexpected development offers some cause for optimism: though quantitative *historians* hang on for their professional lives, historically-inclined political scientists, economists, and sociologists are winning kudos in their respective disciplines by bringing quantitative methods to bear on the past.

See also: Economic History; Economics, History of; Historical Demography; Historiography and Historical Thought: Current Trends; History and the Social Sciences; History: Theories and Methods; Measurement Theory: History and Philosophy; Positivism, History of; Postmodernism: Methodology; Prosopography (Collective Biography); Quantification in the History of the Social Sciences; Social History

Bibliography

- Aydelotte W O, Bogue A G, Fogel R W (eds.) 1972 *The Dimensions of Quantitative Research in History*. Princeton University Press, Princeton, NJ
- Bogue A G 1983 *Clio and the Bitch Goddess*. Sage Publications, Beverly Hills, CA
- Coclanis P A 1992 History by the numbers: why counting matters. *OAH Magazine of History* 7: 5–8
- Davis L E, Engerman S 1987 Cliometrics: the state of the science. *Historical Methods* 20: 97–106
- Dollar C M, Jensen R J 1971 *Historian's Guide to Statistics*. Holt, Rinehard and Winston, New York
- Floud R 1979 *An Introduction to Quantitative Methods for Historians*, 2nd edn. Methuen, London
- Griffin L J, van der Linden M (eds.) 1998 New methods for social history. *International Review of Social History* 43: 3–8 (Supp. 6)
- Grossbart S R 1992 Quantitative and social science methods for historians. *Historical Methods* 25: 100–20
- Hareven T K, Vinovskis M A (eds.) 1978 *Family and Population in Nineteenth Century America*. Princeton University Press, Princeton, NJ
- Haskins L, Jeffrey K 1990 *Understanding Quantitative History*. MIT Press, Cambridge, MA
- Hudson P 2000 History by numbers: An introduction to quantitative approaches. Oxford University Press, New York
- Jarausch K H, Hardy K 1991 *Quantitative Methods for Historians*. University of North Carolina Press, Chapel Hill, NC

- Kocka J 1984 Theories and quantification in history. *Social Science History* 8: 169–78
- Kousser J M 1989 The state of social science history in the late 1980s. *Historical Methods* 22: 13–20
- Lorwin V R, Price J M (eds.) 1972 *The Dimensions of the Past*. Yale University Press, New Haven, CT
- Rabb T K 1983 The development of quantification in historical research. *Journal of Interdisciplinary History* 13: 591–601
- Reynolds J F 1998 Do historians count anymore? *Historical Methods* 31: 141–8
- Stearns P N 1985 Social history and history: a progress report. *Journal of Social History* 19: 319–34
- Steckel R H 1995 Stature and the standard of living. *Journal of Economic Literature* 33: 1903–40

K. H. Jarausch and P. A. Coclanis

Quantification in the History of the Social Sciences

Quantification means, very broadly, the use of numbers to comprehend objects and events. While quantification is often identified with science, it has been equally significant for business, administration, and the professions. Quantitative methods had a crucial role in shaping the social and behavioral sciences, and remain central to them. The purpose of this article is not to supply a practical guide to the use of quantitative tools in social science, but to examine their backgrounds and assess their cultural significance from the standpoint of science studies and the history of science. In modern times, quantitative methods and the numbers they produce have assumed a powerful role in most societies as intersecting discourses of science, administration, and the public sphere. This article emphasizes the more public uses of quantification.

The broad category of quantification refers to several allied but distinct human activities. Mathematization means the characterization of objects and events in terms of mathematical relationships. While it is not the main topic of this article, it is of course relevant. The expression of scientific theories in mathematical terms has provided one important reason for relying on quantitative or numerical data. There have been other reasons of a technical, administrative, or moral kind. Among the more empirical forms of quantification, measurement grew out of such practical activities as surveying lands or buying and selling. Counting, the basic tool for assigning numbers to sets of discrete objects was deployed for such tasks as inventorying one's possessions or recording human populations. *Calculation* refers loosely to methods for combining and manipulating numerical information into more readily useable forms. Statistics, a body of

methods for analyzing quantitative information, is discussed historically in other essays in these volumes, and receives also some attention here.

1. Theories and Measures

As early as the 1660s, and more systematically from about 1830, social and economic investigators held up quantification as the proper method of science. The growth of university-based social disciplines in the twentieth century, and especially after about 1930, was attended by a widespread faith that through numbers, social research could be made genuinely scientific. A generation or more of positivistically-inclined researchers looked to systematic measurement as the proper foundation for solid, mathematical theory. Some compared its progress with the development of physics in the seventeenth century, and anticipated a breakthrough in the near future.

The most influential historical and philosophical writers on science in the postwar period took almost the opposite view. Koyré (1968), whose philosophical studies of seventeenth-century science inspired the first generation of professional historians of science in the anglophone world, seriously doubted that mere measurements could offer much to a new or unformed science. Science could never have become mathematical, he argued, through the mere processing of empirical measurements. He doubted that Galileo had learned much from rolling balls down inclined planes. The new kinematics applied not to real objects, but to abstract bodies in geometrical space, and Galileo must have derived his law of falling bodies mainly by working out logically the consequences of what he already knew. This mathematical and metaphysical work was, for Koyré, the most important achievement of the new science. Indeed, only through theory had the measures attained scientific meaning. Kuhn 1961, 1976) reaffirmed Koyré's broad argument and provided a historical perspective that reached forward to the nineteenth century. The most successful sciences of the seventeenth-century 'scientific revolution,' including mechanics, astronomy, and geometrical optics, had already been mathematical in ancient times. The more Baconian, experimental studies of the early modern period were comparatively ineffective until a 'second scientific revolution,' beginning around 1800, when new scientific objects such as electricity and heat began to be comprehended mathematically.

Scholars in science studies since about 1980 have tended to dethrone theory in favor of a more respectful attention to scientific practices. Accordingly, they take more seriously the quantification of phenomena that have not been grasped by exact laws. Just as the life of experiment has never been strictly subordinated to theory, so numbers have often been independent of higher mathematics. They are linked not just to

theoretical science, but also, and perhaps even more consequentially, to economic and administrative changes. The history of ideas alone cannot convey their full significance.

2. Quantities and Standards

Quantity is not first of all the language of science, imported as something alien into everyday affairs, but an indispensable part of the fabric of life. Yet the aspiration to science has gone a long way toward reshaping the role of numbers and measures in the world. In this—as Max Weber suggested—we find much in common between bureaucratic and scientific rationality. Mathematics has often, and plausibly, been held up as an exemplar of universalized knowledge. The rules of arithmetic, or constants like π , are the same everywhere, at least in principle. Yet the universalism of measures and numbers, particularly when these have empirical content, has been achieved only in a limited way, by overcoming great obstacles. Even now, even in the industrialized West, quantities and prices are managed in everyday activities such as shopping by techniques quite different from those taught in schools (Lave 1986). Before the rise of modern science, in the pre-industrial world, quantification had very different meanings. Numbers were pervasive, but for most purposes their use and meanings were thoroughly local. The process through which they gained general validity is one of the crucial transformations of modernity.

2.1 A Plethora of Units

As Kula (1986) has shown, measures remained immensely variable in Europe as late as the eighteenth century. This was partly by design, since small cities and lesser states often claimed the right to define their own weights and measures as a mark of autonomy and an obstacle to rule by higher authorities. A diversity of measures, however, is the natural state of things wherever the pressures of commerce and administration have been insufficient to bring about unification. At a local level, authorities took some pains to reduce the ambiguity of measurement. A bushel measure, valid for the region, would often be fixed to the town hall, and could be used to settle disputes. In the next town, however, measures might well be different, and over greater distances they varied considerably. The subtle geography of measures, moreover, was not defined exclusively by distance. Milk, oil, and wine came in different units; measures of cloth depended on the width of looms, and also on the costliness of fabric and weave. The practices of measurement, too, were complexly nuanced. A bushel measure, even if the volume was fixed, could be tall and narrow or short and wide; the wheat might or might not be compacted

by being poured in from a considerable height; and a complex set of conditions governed whether to heap the bushel, and by how much. Parties to a transaction often negotiated about measures rather than about prices. Add to this the consideration that almost nothing was decimalized, not even money, and it is evident that arithmetical calculation provided no fixed routine for managing affairs or settling disputes (Kula 1986, Frängsmyr et al. 1990). The conversions required by trade provided an important source of employment for mathematicians.

The simplification and standardization of measures was encouraged by the expansion of long-distance trade, but was achieved primarily by governments in alliance with scientists. Among Europeans, it was the British who came nearest to creating uniformity of measures in the eighteenth century. A more systematic, almost utopian, scheme, the metric system, was initiated in the 1790s under the French Revolution. In its most ambitious form this included a new calendar, with the declaration of the Republic defining the Year One, using months of 30 days divided into weeks of 10 days, and even a few decimalized clocks, beating out 100 seconds per minute. The meter was set so that 10 million of them would reach from pole to equator, in the vain hope of integrating geodetic and astronomical measures with those on a human scale (Gillispie 1997). Hence, bold expeditions were sent out into a disordered countryside to survey and measure a meridian. Although the meter was justified in part as a response to pleas in the *cahiers de doléance* for fixed measures, against the social and economic power implied by seigniorial discretion, the highly rationalized metric system of decimalized conversions and strange Greek names—of milliliters, kilometers, and centigrams—went well beyond this. Not until the July Monarchy, after 1830, did the metric system begin to succeed even in France (Alder 1995). In time, however, it has come to provide a common set of measures for most of the world, with the United States a notable, though partial, holdout.

A precise and maximally rigorous definition of units is a basic element in the infrastructure of modern life. The role of standardized measures in systems of manufacturing based on interchangeable parts is widely recognized. They are indispensable also for large-scale networks such as electric power distribution, telephony, and, by now, computers and the Internet. A collaboration of scientists and bureaucrats was required to create them. In some cases, as with the electrical units defined in the latter nineteenth century, eminent scientists such as James Clerk Maxwell, William Thomson (Kelvin), and Hermann von Helmholtz were deeply involved in the fixing of standards.

Other units, far from the domain of scientific theory, play an indispensable role in the regulation of modern life. Measures of effluent concentrations, for example, are deployed by environmental agencies to regulate

pollutants. They depend on meticulous specification not only of instruments, but also of sampling methods, training of technicians, custody of materials, and so on. Since penalties and taxes often depend on them, they are always vulnerable to challenge. In this domain, as in many others, an immense apparatus of instrumentation and control lies behind what scientists and citizens have the luxury (usually) of regarding as 'raw,' or merely 'descriptive,' data.

2.2 Producing Uniformity

Units and instruments are not by themselves sufficient to quantify effectively. A quantifiable world contains reasonably homogeneous objects, objects whose measures can be compared, or that can be grouped together and counted. Nature and unregulated human labor produce a few of these, though real uniformity occurs mainly on a scale—for example, the molecular one—that is not readily accessible to the human senses. Physics was quantified partly in terms of abstractly 'mathematical' entities, such as forces, and partly through increasingly rigorous experimental controls, involving an expanding apparatus of instruments.

In the social world, where quantification is also a tool of administration, these same regulatory processes have helped to produce new quantifiable entities. The investigation and compensation of unemployment, for example, solidified the category of 'the unemployed.' To collect unemployment insurance requires that a person meet certain standards, which are more or less closely monitored, and to which persons who lack work are encouraged to conform. They also must enroll themselves, thereby providing basic data for the statistics. The point here is not that government offices cause people to be without work, though mechanisms of compensation will certainly affect the rates. It is that official categories sharpen definitions, and so help to create countable entities. Most human and social types are variable, or at least fuzzy at the margins (Hacking 1995). Quantification is allied to processes that make them more uniform, and hence countable. A similar story can be told about ethnicity or mental illness. Economic categories such as profits, investment, and trade, are similarly dependent on the processes through which they are recorded and regulated.

3. The Authority of Calculation

Numbers and measures are not merely tools of description and analysis. They work also as guides to action, especially public action. An explicit and formal analysis, often largely quantitative in form, is by now required for a variety of public decisions involving investment or regulation, especially in the United States and Britain. Business organizations, too, rely increasingly on measures and projections in order to

make decisions. These analyses are often expressed in money terms, and their modern expansion attests to a triumph of capitalism. But public reliance on economic quantification is by no means an imitation of its commercial uses. The tools of quantitative analysis were developed for as much for government as for business purposes, and have become increasingly dependent on academic researchers. In the background is an evolving political and administrative culture, reflecting changes in the status and composition of elites.

3.1 Commensuration

It is characteristically modern to suppose that a conscientious decision should involve explicit consideration of the available data. Decision by calculation goes beyond this, in expecting a course of action to be dictated objectively by facts. The prototype of this ideal of public reason is cost-benefit analysis. In practice, cost-benefit often involves predictions in many domains, including weather patterns, health, hydrology, ecology, and commerce, but its most distinctive aspect is the attempt to convert a great diversity of considerations into a common financial metric. The difficulties of commensuration are most acute in regard to such considerations as human lives lost or saved, the disappearance of biological populations, or the degradation of a beautiful landscape. To the lay public, this appears equivalent to sacrificing a life for a sum of money, and of very doubtful morality (Sagoff 1988).

Economics provides a rationale for equations of this kind. Economists argue that people routinely accept increased risks for better wages. Neoclassical economic theories presuppose an underlying variable, utility, which individuals seek to maximize. Even if it cannot be directly measured, personal choices therefore involve some implicit commensuration of pleasures and responsibilities, acquisition and leisure, friendship and personal ambition. Cost-benefit analysis did not, however, develop out of theoretical economics, but from specific political and administrative problems of the mid-twentieth century. Engineers, not economists, initially faced the immense difficulties of incorporating life, health, recreation, and scenic beauty along with economic development as they proffered measures of the costs and benefits of water projects. Despite political opposition to some of these measures, a political logic of rationalized choice made them necessary none the less. The task was to devise methods that led to reasonably consistent answers, and where necessary to put aside what seemed their implausibility or even immorality. Beginning in the 1950s, economists and social scientists were recruited into agencies like the US Army Corps of Engineers and the Bureau of Reclamation to provide more consistent or more acceptable solutions to these thorny problems (Espeland 1998, Porter 1995).

3.2 Trust and Impersonality

Academic problems of social measurement and commensuration are, thus, embedded in social and political conditions that put a premium on scientific objectivity. In important ways, of course, science and philosophy have always been concerned with objectivity, in the sense of truth. Since the seventeenth century, the relation of fact and values has become more problematical, and for some purposes the credibility of science has been enhanced by its separation from the 'subjective' domain of values. The ambition to provide a scientific, possibly quantitative, basis for public decisions remained quixotic or utopian until at least the late eighteenth century. It remained unrealistic wherever an aristocratic political class retained power. Not everyone would have assented to Frédéric Le Play's assessment that numbers have become necessary only in democracies, where leaders no longer control the wisdom of birth and experience. But the aspiration to replace judgment with objective, quantitative rules grew up in situations where elite authority had become suspect.

The move by American state engineers to an increasingly rigorous cost-benefit methodology was a response to systematic challenge, and exemplifies this point (Porter 1995). So also do other exemplary tools of 'mechanical objectivity,' such as the accountant's bottom line and the test of statistical significance of the social scientist or medical researcher. Accountants, for example, did not abandon willingly what many saw as their professional responsibility to offer an expert judgment of the financial condition of firms. They did so, rather, in the context of regulatory pressure and public distrust. This must be understood in the context of a broad cultural and political history of the twentieth-century, and especially of the United States. Some basic features of that history would include the growth of government, frequent interagency conflict, increasingly effective opposition to administrative secrecy, and public distrust of 'bureaucracy.' The unprecedented quantitative enthusiasm of postwar social science involved, along with its optimism about science, a renunciation of the personal and the subjective that was nicely adapted to contemporary political circumstances. Allied to it was the institutionalized reliance on quantification taught in management and policy schools.

4. A Historical Perspective on Social Science Quantification

Reflective accounts about the social and behavioral sciences have often supposed that their fields have short histories, and that the pursuit of quantification has meant following the track of natural science triumphant. But the rise of a quantitative ethos in natural science can be dated no earlier than the mid-

eighteenth century. Systematic measurement until then was almost confined to astronomy, for which a host of 'mathematical instruments' had been developed in China, India, the Near East, and Europe. Many were geometrical and not numerical. Their purposes were practical as well as scientific: to survey land, navigate ships, calculate calendars and horoscopes. Barometers and thermometers helped to extend the domain of measurement into meteorology. Measurement in early modern Europe was associated as closely (or more closely) with natural history as with experimental physics. By 1750, instruments were deployed systematically to measure mountains and other natural wonders and, thus, to lend precision to the appreciation of nature. The development of an exact science of electricity in the later eighteenth century marked the triumph of experimental quantification (Heilbron 1979). Measurement had also become routine in technological endeavors, such as mineral assaying, decades before Lavoisier and his contemporaries began to insist on its centrality to chemistry.

From this perspective, it would be difficult to argue that political and economic studies lagged behind natural science in their reliance on numbers. Among the early members of the Royal Society of London, founded in 1660, were William Petty and John Graunt. Petty announced a new science of 'political arithmetic,' which in his hands involved speculative estimates of population and of wealth. Graunt is noted for his more strictly empirical compilations from the London bills of mortality. Near the same time, Jan de Wit, the Dutch Stadholder, used rudimentary probability theory to assign rates for the sale of annuities, and the astronomer Edmond Halley calculated the first mortality table. Political arithmetic, which entailed the compilation as well as the analysis of demographic and medical records, flourished in the eighteenth century. At the same time, probability theory was pursued as a mathematical theory of belief and credibility (Hacking 1975, Daston 1988). The *philosophe* and mathematician Condorcet, a leader of the French Revolution who then was hunted down by it, hoped to develop probability into the theoretical basis for elections and judicial decisions (Baker 1975).

4.1 The Nineteenth Century and the Rise of Statistics

By the early nineteenth century, as Kuhn observed, the natural-historical investigation of heat, electricity, and the physics of light had given way to sciences that were both mathematical and quantitative. Possibly the analysis of mortality records for insurance can be regarded as comparable in certain respects. But the explosive growth of quantification beginning in the 1820s, what Hacking (1990) has called an 'avalanche of printed numbers,' was not concentrated in the

experimental sciences. Rather, it involved systematic surveys of stellar positions, terrestrial magnetism, weather phenomena, tides, biogeographical distributions, and, most prominently, social statistics.

The collection and investigation of social numbers had expanded gradually over the course of the eighteenth century. Few Enlightenment states had the bureaucratic means to conduct a full census, and most were disinclined to release such sensitive information to their enemies or even their subjects. In many places, too, the people would have viewed census official as harbingers of new taxes or military conscription, and in Britain the census was blocked by a gentry class that refused to be counted. In contrast to piecemeal efforts at home, there were some systematic surveys abroad, the machinery of the information state having been turned loose first of all in colonial situations. Among European states, Sweden pioneered the regular census, in 1749; its results remained secret until 1764. The French used sampling, of a sort, and enlisted mathematicians such as Pierre Simon Laplace to estimate the population. The US census of 1790 heralded a new era of regular, public censuses, which were encouraged in Europe by French military conquest and by the pressure of wartime mobilization. The British decennial census, for example, was instituted in 1801. Only in the 1830s did bureaucratic professionals begin to take over the collection of social and economic numbers in the most advanced European states.

At the same time, an increasing pace of quantitative study by voluntary organizations made 'statistics,' pursued by 'statists,' a plausible 'science of society.' One model for it was developed by the Belgian Adolphe Quetelet, whose quantitative pursuits ranged from observational astronomy and meteorology to the bureaucratic organization of official statistics. He argued for an alliance of data gathering with mathematical probability, in the quest to uncover 'social laws.' Apart from life insurance, however, most statistical collection was in the control of reformers and administrators with little mathematical knowledge. Their enthusiasm for counting was perhaps indiscriminate, yet it was nurtured by specific anxieties about poverty, sanitation, epidemic disease, economic dislocation, and urban unrest. Social investigation, still more than natural science, was an object of contest, and there were active efforts to organize a social science to speak for working people (Yeo 1996). Karl Marx's immense statistical labors for his *Capital*, drawing mainly from official British inquiries whose results he praised as honest and reliable, testify impressively to the credibility of social numbers. Middle-class organizations had more staying power than working-class ones, and better standing to claim the mantle of 'social science.' In Britain, the Statistical Society of London (founded 1834) and the National Association for the Promotion of Social Science (1857) joined bureaucrats with professionals and even aristocratic political leaders. In Germany,

statistics became a university discipline, which, from about 1860 to 1900, combined empirical quantification with a historicized 'national economics,' in a reformist campaign for state action on behalf of workers and peasants. Some of the founders of social science in America were educated in this tradition, as was Max Weber. Emile Durkheim, too, engaged with it, as his *Suicide* of 1897 plainly attests (Porter 1986).

The nineteenth-century science of statistics embraced much of the subject matter of the modern social sciences. It overlapped, and sometimes competed, with political economy, political science, geography, anthropology, ethnology, sociology, and demography. These stood for genres, more or less distinct, though not for anything so definite as modern professional disciplines. Their quantitative methods were not identical, but neither were they clearly differentiated from each other or from measurement in biology. There was much uncertainty and debate from 1830 through the end of the century as to whether statistics was a science, with its own subject matter, or only a method. It was resolutely empirical, even to the point that sampling seemed too speculative. Far from being allied to any mathematical social science, statistics was more often at war with deduction and abstract theory. Some, like the economist William Stanley Jevons, would claim that political economy should be mathematical because it was inherently quantitative, but he could not attach the mathematical formulations he introduced to empirical numbers.

4.2 Quantification in the Modern Social and Behavioral Sciences

Toward the end of the nineteenth century, a more mathematical field of statistics began to develop. The most influential version of the new statistics was developed within a British biometric tradition by Francis Galton and Karl Pearson. At almost the same time, however, distinct social and behavioral sciences began to crystallize, in part around increasingly distinct quantitative methods. All were conceived as ways to enhance the scientific standing of their fields, and each was engaged also with issues of politics, policy, and the management of economies and populations. The typology to follow is intended to be suggestive, and certainly cannot be exhaustive.

4.2.1 Physical anthropology and eugenics. The measurement of humans was carried out for bureaucratic as well as scientific purposes throughout the nineteenth century. The increasing concern with race in anthropology was linked to a preoccupation with skulls, whose measures were used to distinguish distinct 'races' of Europeans. What Francis Galton called 'correlation' was developed to measure the

interconnections among physical human characteristics, their perpetuation from one generation to another, and their relation to measures of merit or achievement.

4.2.2 Psychometrics and psychology. Correlation provided also a solution to a problem of contemporary psychophysics, related to education. Gustav Theodor Fechner began applying the methods of error theory to the measurement of sensory discrimination in the mid-nineteenth century. Psychophysics was plagued, however, by the problem that a task—for example, of distinguishing two slightly different weights—could not be repeated indefinitely, because the subject learned over time to do better. Indeed, the left hand acquired some of the fine discrimination developed through practice with the right hand. By the end of the century, this obstacle had become a topic of investigation in its own right, as an approach to the psychology of learning. Systematic use of experimental randomization was developed for studies of this kind (Dehue 1997). They provided also a wide field of application for correlational methods. Did childhood study of dead languages really train the mind for speedier or more thorough learning of other subjects, such as science, history, or literature? The growth of mental testing in the early twentieth century opened another question of correlations. To what extent did students who proved themselves adept in some particular study excel also in others? Charles Spearman's *g*, or general intelligence, and the 'intelligence quotient' or IQ, which flourished in America, presupposed a basic unity of the mental faculties. Psychometricians and educational psychologists developed a body of statistical methods to measure this intelligence, or to decompose it, and to assess the relations of its components (Danziger 1990).

In the 1930s and 1940s, as psychology turned resolutely experimental, quantification became practically mandatory. Once again, educational psychology and parapsychology took the lead in the assimilation of novel methods of experimental design and analysis. The new paradigm of a proper experiment involved a randomized design of control and experimental populations, yielding results that would, in most cases, be subjected to analysis of variance (ANOVA). A worthwhile result was defined in part by a test of significance: the 'null hypothesis' that the treatment had no effect should be rejected at some specified level, commonly 0.05 (Gigerenzer et al. 1989).

4.2.3 Econometrics and economics. The new mathematics of marginalism was developed in the 1870s

and self-consciously reconciled to the classical tradition in the 1890s by Alfred Marshall. His work included abundant graphical illustrations, but these involved quantities that could not easily be measured. The origins of econometrics are largely distinct. From origins in the 1870s to the establishment of a society and a journal in the 1930s, the field focused on the study of business cycles (Morgan 1990). Econometricians took their numbers where they could get them, often from public sources. They were highly dependent on national accounts, prepared by governments, and indeed participated in structuring these accounts. The paradigmatic statistical tool of econometrics, adapted from biometrics mainly after 1950, was regression analysis.

4.2.4 Surveys. From the founding of statistical societies in the 1830s, social science has relied on surveys to learn about how lives and opinions. Charles Booth's monumental studies of East London beginning in the 1880s inspired a movement of surveys of towns and regions in Britain and America. At first, researchers were expected to examine everyone, but after 1900, sampling began to gain acceptance. Arthur Bowley and then Jerzy Neyman developed sampling methods that permitted generalization using probability theory. The techniques of survey sampling have linked sociology and political science to marketing and political campaigns (Bulmer et al. 1991, Desrosières 1998).

4.2.5 Modeling. By the late twentieth century, the use of models was pervasive in natural and social sciences alike. Modeling methods are far too diverse to think of summing them up, but the modeling outlook deserves comment. It is perhaps under this rubric that mathematical methods are most often integrated with the data of tallies and measurements. A quantitative model, very broadly, means working out the consequences of certain assumptions for a particular situation or problem. Often, models are used for predicting, and they can equally well be purely academic or highly pragmatic, for very specific applications. In the twentieth century they have come to pervade the sciences. Among the social disciplines, modeling is especially central to economics.

5. Conclusion

Quantification by now is virtually as important for social as for natural science. Its expansion since the 1700s reveals some key characteristics of the social and behavioral sciences: among them, the drive for the status of science, the role of applied studies in forming

academic ones, and the often problematic relationship of theoretical to empirical work. The quantification of social phenomena goes back several centuries, and is by no means a mere product of the effort to attain scientific standing. Indeed, the advance of measurement and statistical analysis in social science is not distinctively academic, but has been greatly stimulated by administrative and political demands. Yet quantification in social science has, since the 1800s, been surrounded by a certain mystique, as an indication of rigor or at least of scientific maturity. It should be regarded not as the key to all mysteries, but as a powerful set of tools and concepts, to be integrated as much as possible with theoretical understanding and with other ways of comprehending social phenomena.

See also: Econometrics, History of; Psychometrics; Scaling and Classification in Social Measurement; Scientific Revolution: History and Sociology; Statistics, History of; Statistics: The Field

Bibliography

- Alder K 1995 A revolution to measure: The political economy of the metric system in France. In: Wise N (ed.) *The Values of Precision*. Princeton University Press, Princeton, NJ pp. 39–71
- Baker K 1975 *Condorcet: From Natural Philosophy to Social Mathematics*. University of Chicago Press, Chicago
- Bulmer M, Bales K, Sklar K (eds.) 1991 *The Social Survey in Historical Perspective, 1880–1940*. Cambridge University Press, Cambridge, UK
- Daston L 1988 *Classical Probability in the Enlightenment*. Princeton University Press, Princeton, NJ
- Danziger K 1990 *Constructing the Subject: Historical Origins of Psychological Research*. Cambridge University Press, Cambridge, UK
- Dehue T 1997 Deception, efficiency, and random groups: Psychology and the gradual origination of the random group design. *Isis* **88**: 653–73
- Desrosières A 1998 *The Politics of Large Numbers*. Harvard University Press, Cambridge, MA
- Espeland W 1998 *The Struggle for Water*. University of Chicago Press, Chicago
- Frängsmyr T, Heilbron J, Rider R (eds.) 1990 *The Quantifying Spirit in the Eighteenth Century*. University of California Press, Berkeley, CA
- Gigerenzer G, Swijtink Z, Porter T, Daston L, Beatty J, Krüger L 1989 *The Empire of Chance: How Probability Changed Science and Everyday Life*. Cambridge University Press, Cambridge, UK
- Gillispie C 1997 *Pierre Simon Laplace, 1749–1827: A Life in Exact Science*. Princeton University Press, Princeton, NJ
- Hacking I 1975 *The Emergence of Probability*. Cambridge University Press, Cambridge, UK
- Hacking I 1990 *The Taming of Chance*. Cambridge University Press, Cambridge, UK
- Hacking I 1995 *Rewriting the Soul*. Princeton University Press, Princeton, NJ
- Heilbron J 1979 *Electricity in the 17th and 18th Centuries: A Study in Early Modern Physics*. University of California Press, Berkeley, CA
- Koyré A 1968 *Metaphysics and Measurement*. Harvard University Press, Cambridge, MA
- Kuhn T 1961 The function of measurement in modern physical science. In: Woolf H (ed.) *Measurement*. Bobbs-Merrill, Indianapolis, IN pp. 31–63
- Kuhn T 1976 Mathematical versus experimental traditions in the development of physical science. *Journal of Interdisciplinary History* **7**: 1–31 (reprinted in Kuhn T 1977 *The Essential Tension*. University of Chicago Press, Chicago, pp. 31–65)
- Kula W 1986 *Measures and Men* (Szreter R trans.). Princeton University Press, Princeton, NJ
- Lave J 1986 The values of quantification. In: Law J (ed.) *Power, Action, and Belief: A New Sociology of Knowledge?* Routledge, London pp. 88–111
- Morgan M 1990 *The History of Econometric Ideas*. Cambridge University Press, Cambridge, UK
- Porter T 1986 *The Rise of Statistical Thinking, 1820–1900*. Princeton University Press, Princeton, NJ
- Porter T 1995 *Trust in Numbers: The Pursuit of Objectivity in Science and Public Life*. Princeton University Press, Princeton, NJ
- Sagoff M 1988 *The Economy of the Earth: Philosophy, Law, and the Environment*. Cambridge University Press, Cambridge, UK
- Yeo E 1996 *The Contest for Social Science*. Rivers Oram Press, London

T. M. Porter

Quantifiers, in Linguistics

1. Introduction

Quantifiers are free-standing expressions whose meanings involve the notion of quantity, such as English three, several, numerous, most, every, one hundred and twenty three, all but seventeen, and so forth.

The basic semantic structure of quantification is bipartite, consisting of the quantifier itself plus the expression that it quantifies. For example, in a sentence such as John saw three cats, *three* is the quantifier, and *cats* the quantified expression.

Quantification has traditionally been of great interest to semanticists, logicians and philosophers of language, due at least in part to the perceived ‘logical’ or ‘mathematical’ nature of the meanings involved. As such, it is striking to observe how such basic and seemingly immutable meanings may, in different languages, be expressed with very different morpho-syntactic strategies, exhibiting a great degree of cross-linguistic variation.

Quantifiers may be classified internally, with respect to their morphosyntactic structure and basic semantic

properties. In addition, they may be classified externally, with regard to the syntactic and semantic relationships between the quantifiers and the quantified expressions and/or the remainder of the sentences in which they occur.

2. Internal Classifications

Internal classifications of quantification make reference to a variety of fundamental morphosyntactic and semantic properties.

2.1 Mass and Count Quantifiers

One of the most basic distinctions is between mass and count quantifiers (Pelletier 1979, Allan 1980). Mass quantifiers constitute expressions which denote an undifferentiated homogeneous mass; for example, English *much* is a mass quantifier, because it forms expressions such as *much milk*, *much cake*. In contrast, count quantifiers constitute expressions which refer to one or more countable units of characteristic size and shape; for example, English *many* is a count quantifier, because it forms expressions such as *many cats*, *many cakes*. Whereas many quantifiers belong to just one of the two types, mass or count, some quantifiers are undifferentiated with respect to the mass/count distinction, and may appear in expressions of either kind.

An interesting case is that of the so-called ‘classifier languages,’ many of which are located in the East Asian linguistic region (Craig 1986; see example (4) below). In such languages, many or all of the count quantifiers cannot occur in immediate construction with the noun that they quantify; instead, the quantifier must occur in construction with a classifier, and the quantifier-plus-classifier constituent may then occur in construction with the noun. Typically, the classifier makes reference to various characteristics—size, shape, function, and so forth—of the quantified noun. Accordingly, different nouns often require different classifiers. However, in some cases, different classifiers may occur with the same noun, sometimes resulting in subtle differences in meaning.

2.2 Existential and Universal Quantifiers

Of particular interest to logicians are the existential and universal quantifiers (Vendler 1967, Heim 1982, Gil 1995). Existential quantifiers, such as English *some* and *a(n)*, form expressions denoting at least a minimal, non-zero amount or number of the quantified expression, for example *some water*, *a boy*. Universal quantifiers, such as English *all* and *every*, form expressions referring to an exhaustive amount or number of the quantified expression, for example *all the water*, *every boy*. Existential and universal quanti-

fiers are semantically related to each other through negation by means of the classical ‘square of opposition’: thus, in English, *Not a boy came* is equivalent to *All the boys did not come*, and *Not all the boys came* is equivalent to *A boy did not come*.

Existential quantification, in many languages, is inextricably intertwined with singular number and/or indefiniteness: in some cases they are expressed identically, while in other cases they are expressed by means of the same marker occurring in different constructions, or by means of similar though not identical markers. Universal quantification, too, is expressed in a variety of different ways across languages. Some languages, such as English, do have ‘dedicated’ universal quantifiers, that is to say, words such as *all* and *every* whose primary or exclusive function is the expression of universal quantification. However, many other languages lack such words, instead making use of more complex circumlocutions if and when the expressive need arises.

Given the importance of existential and universal quantifiers in mathematical logic, it is worthy of note how relatively unimportant these quantifiers are within the grammatical systems of languages. It is indeed not all that common to find a single form, in any language, whose interpretation corresponds precisely and unproblematically to either the existential or the universal quantifier. Moreover, there is probably no language in which expressions of existential and universal quantification constitute a natural grammatical class to the exclusion of other expressions. The diversity of morphosyntactic expression of existential and universal quantifiers across the world's languages suggests that, although greatly valued by logicians, they may not be endowed with privileged status vis à vis the structure of language.

2.3 Mid-range Quantifiers

‘In between’ the existential and the universal quantifiers lie a variety of mid-range quantifiers, such as, in English, *few*, *several*, *many*, *most*, and so forth. Among the mid-range quantifiers, a particularly privileged category is that of cardinal numerals—quantifiers which refer to natural numbers, for example, *one*, *two*, *thirteen*, *one hundred* and *twenty seven* (Menninger 1969, Hurford 1975). Complex phrasal numerals often make use of syntactic patterns available elsewhere in the language; however, in addition, they sometimes exhibit unique structural features not found in other grammatical domains.

2.4 Other Internal Classifications

Quantifiers are sometimes classified as either *strong* (alternatively referred to as *definite*) or *weak* (indefinite) (Milsark 1977). Strong quantifiers, such as

English all and most, are inherently definite in their meaning, whereas weak quantifiers, such as English some and three, are not. Strong and weak quantifiers often exhibit different syntactic behaviour; for example, strong quantifiers are typically awkward or ungrammatical in existential sentences, e.g.,? There is all the water on the floor.

Some further semantic distinctions are captured by the categories of *increasing* (alternatively referred to as upward entailing) and *decreasing* (or downward entailing) (Keenan and Stavi 1986). A quantifier is increasing if a sentence containing it would be logically true in a situation in which the amount or number of quantified expressions involved is larger than that explicitly stipulated by the quantifier, for example, at least three. Conversely, a quantifier is decreasing if a sentence containing it would be logically true in a situation in which the amount or number of quantified expressions involved is smaller than that explicitly stipulated by the quantifier, for example, at most three. Other quantifiers are neither increasing nor decreasing, for example exactly three.

An interesting problem is posed by simple monomorphemic quantifiers such as three. At first blush, such quantifiers appear to be neither increasing nor decreasing: if somebody says There are three cats on the bed, the mental picture that is conjured up is, of course, one of three cats, not two, or four, or any other number. However, a moment's reflection will reveal an important asymmetry. If in fact there are exactly two cats on the bed, then the above sentence is clearly false. In contrast, if there are exactly four cats on the bed, then, in most contexts, the above sentence would obviously be misleading and inappropriate—but it would not, strictly speaking, be false. Thus, while the pragmatic appropriateness conditions of three resemble those of exactly three, the actual truth conditions of three are identical to those of at least three. Hence, three is, in fact, increasing.

3. External Classifications

External classifications of quantification make reference to the morphosyntactic and semantic relationships which obtain between the quantifiers and the quantified expressions and/or the remainder of the sentences in which they occur.

3.1 Nominal and Verbal Quantifiers

One basic semantic classification relates to the category of the quantified expression, specifically whether it is of a nominal or verbal nature (Gil 1993). For example, in the English sentence John saw three cats, the numeral three quantifies a noun; accordingly this sentence illustrates nominal quantification. In contrast, in the sentence John sang three times, the

numeral phrase three times quantifies a verb; hence this sentence instantiates verbal quantification.

Up to a point, nominal and verbal quantification exhibit parallel grammatical properties. Obvious, but still worth noting, is the fact that in both of the preceding sentences, the same numeral, three, is used. However, nominal and verbal quantification also differ in many ways; typically, verbal quantification is more highly marked than its nominal counterpart.

3.2 Continuous and Discontinuous Quantifiers

A distinct albeit related syntactic classification pertains to the position of the quantifier with respect to the quantified expression, specifically, whether or not the two form a continuous constituent (Gil 1993, Bach et al. 1995). Sentences (1)–(4) below all instantiate nominal quantification, in which a numeral quantifies an expression referring to an animal; however, each of the examples exhibits a different syntactic configuration.

- (1) HEBREW
raiti [šaloš arnavot]
 see-PAST: 1:SG three-F rabbit-PL:F
 'I saw three rabbits'
- (2) TAGALOG
Tatlo [ang nakita
 three TOP PAT.TOP: PFV-see
kong baboy]
 DIR: 1:SG-LIG pig
 'I saw three pigs'
 ('The pigs seen by me were three')
- (3) KUTENAI (Matthew Dryer, p.c.)
hu [qa'sa' wu:kati] cupqa
 1:SG three-PRV see-IND deer
 'I saw three deer' ('I three-saw deer')
- (4) JAPANESE
Watasi wa inu o
 1:SG TOP dog ACC
sanbiki mita
 three-CLF see-PFV
 'I saw three dogs' ('I saw dogs three')

In Modern Hebrew, (1), as in its English gloss, the quantifier groups with the quantified expression, *šaloš arnavot*, to the exclusion of the remainder of the sentence; in Tagalog, (2), the quantifier *tatlo* stands alone as a sister to the remainder of the sentence, which forms a constituent; in Kutenai, (3), the quantifier occurs in construction with the verb, *qa'sa' wu:kati*, to the exclusion of the remainder of the sentence, while in Japanese, (4), the quantifier *sanbiki* is a sentential adverb not forming an obvious constituent with any part of the remainder of the sentence.

Constructions (1) and (2) exemplify continuous quantification, in which the quantifier and the quantified expression form a constituent. In contrast, constructions (3) and (4) exemplify discontinuous quanti-

fication, in which the quantifier occurs apart from the quantified expression. Examples (1)–(4) above illustrate the distinction between continuous and discontinuous quantification with respect to nominal quantification. In principle, a similar distinction may also be drawn with respect to verbal quantification, though examples of discontinuous verbal quantification are more hard to come by.

3.3 Scope and Scopal Quantifiers

When two quantified expressions are present in the same construction, a variety of semantic relations may obtain between the two expressions (Jackendoff 1971, Gil 1982b). The following English sentence, containing two numerically quantified NPs, may, at least potentially, be interpreted in any or all of the following four ways.

- (5) Three boys saw two girls
- (a) ‘Each of three boys saw each of two girls’
not scope differentiated: strong symmetric.
 - (b) ‘Three boys saw two girls between them’
not scope differentiated: weak symmetric.
 - (c) ‘Three boys saw two girls each’
scope differentiated: wide scope for *three* boys
distributive key: *three boys*; distributive share: *two girls*.
 - (d) ‘Two girls were seen by three boys each’
scope differentiated: wide scope for *two girls*
distributive key: *two girls*; distributive share: *three boys*.

Interpretations (5a) and (5b) both involve a single set of three boys and a single set of two girls: each of the two NPs has independent reference. These interpretations are accordingly characterized as non-scope-differentiated, or symmetric. Interpretations (5c) and (5d) present a more complex picture. In (5c) there is a single set of three boys; however, each of the three boys is associated with a different set of two girls. Thus, while the subject NP has independent reference, the direct-object NP is referentially dependent on the subject NP. In this interpretation, then, the subject NP has scope over the direct-object NP. More precisely, the two NPs are in a relationship of distributivity, with the subject NP as distributive key, and the direct-object NP as *distributive share*. In (5d), a mirror-image situation obtains. Here, there is a single set of two girls; however, each of the two girls is associated with a different set of three boys. Accordingly, while the direct-object NP has independent reference, the subject NP is referentially dependent on the direct-object NP; the direct-object NP has scope over the subject NP. Here too then, a relation of distributivity obtains; this time, though, it is the direct object NP which is the distributive key and the subject NP which is the distributive share.

The above four interpretations are not equally readily available for speakers of English. Specifically,

the symmetric interpretations (5a) and (5b) are more easily available than the asymmetric interpretations (5c) and (5d). Moreover, between the latter two interpretations, (5c) with wide scope for the subject NP, is more readily available than (5d) with wide scope for the direct-object NP. These preferences are not accidental; rather, they are particular consequences of more general principles governing the assignment of quantifier scope across languages.

Given the general preference for symmetric interpretations, languages typically have at their disposal various lexical and morphosyntactic devices whose function is to induce the more marked asymmetric interpretations, when these are required. Consider the following English sentences:

- (6) (a) *All* the men carried three suitcases.
(b) *Every* man carried three suitcases.

Sentences (6a) and (6b) differ in the lexical choice of quantifier within the subject NP. Whereas in (6a), with *all*, a variety of scope relationships may obtain between the two NPs, and the number of suitcases may vary between three in total and three per man, in (6b), with *every*, the subject NP has scope over the direct object NP, and the suitcases number three per man. Whereas *all*, like *three*, is a simple quantifier, *every* is a *scopal* quantifier, that is to say, a quantifier endowed with an additional denotational component, one that forces a particular scope relationship to obtain between the NP in which it occurs, and some other expression. In particular, it is a distributive-key quantifier, inducing a relationship of distributivity, in which the NP containing the quantifier in question is the distributive key, or element with wide scope (Gil 1995).

A different kind of scopal quantifier is exemplified by the following pair of Maricopa sentences:

- (7) MARICOPA
- (a) *ʔipač* *ʔii xmokm*
man-PL:NOM stick 3-three-SG-DS
upaavk
3-carry-PL-REAL
‘The men carried three sticks’
 - (b) *ʔipač* *ʔii xmokxperm*
man-PL:NOM stick 3-three-SG-DIST.SHARE-DS
upaavk
3-carry-PL-REAL
(i) ‘The men carried three sticks each’
(ii) ‘The men carried the sticks in threes’

Sentences (7a) and (7b) differ only in the form of the numeral within the direct-object expression. Sentence (7a) contains a simple numeral, *xmok-* ‘three,’ and is ambiguous in roughly the same ways as its English gloss. In contrast, sentence (7b) contains the same numeral plus the suffix *-xper*, whose effect is to mark the numeral *xmok-* and the expression containing it as distributive share. The ambiguity exhibited by sentence (7b) results from different expressions being selected as the distributive key counterpart. Under interpretation (7b(i)), the subject NP *ʔipač* ‘men’ is

chosen as distributive key: the sticks accordingly number three per man. Under interpretation (7b(ii)), the verb *upaavk* 'carried' is chosen as distributive key: here the sticks accordingly number three per carrying. Cross-linguistically, distributive-share numerals are formed by a variety of formal processes, including affixation (as in the above example), periphrasis, reduplication, and suppletion (Gil 1982a).

3.4 Quantifier as a Grammatical Category

Quantifier is primarily a semantic category: an expression is a quantifier only if it bears certain meaning properties. One of the most striking characteristics of quantifiers is the absence of any distinctive shared formal morphosyntactic features, which might justify the positing of a corresponding grammatical category of quantifier. Within languages, different quantifiers often display different grammatical properties. And across languages, the 'same' quantifier may exhibit different morphosyntactic behavior as well. Thus, in terms of their formal properties, quantifiers constitute a heterogeneous collection, exhibiting a great amount of morphological and syntactic diversity.

Looking within individual languages, there is probably no language within which there is a formal grammatical category consisting exactly of all quantifiers but no other expressions. Rather, in most or all languages, different quantifiers exhibit different arrays of grammatical properties, some grouping together with expressions belonging to one category, others patterning together with expressions belonging to some other category, yet others exhibiting idiosyncratic morphological or syntactic behaviour.

Looking across languages, quantifiers with the same meaning often exhibit quite different morphological and syntactic behaviour. Examples (1)–(4) above showed how the numeral 'three,' as a nominal quantifier, may occur, in different languages, in a variety of different constructions, exhibiting continuous and discontinuous quantification of various kinds. Even if attention is restricted to the most well-known type, that of continuous NP-internal quantification, as in (1), a great amount of cross-linguistic variation may be observed. In particular, in different languages, nominal quantifiers in attributive NP internal position may resemble, to various degrees, nouns, adjectives, and/or verbs.

4. Delimitation of Quantification

The discussion in the preceding sections was based on a bipartite definition of quantifiers as 'free-standing expressions' whose meanings 'involve the notion of quantity.' Like many similar definitions, this one is useful, but not unproblematical. Specifically, in addition to prototypical quantifiers such as those dis-

cussed above, there are numerous expressions and construction types for which it is difficult to adjudicate whether or not they may appropriately be characterized as quantificational.

The first part of the definition is formal: in order for an expression to be considered as a quantifier, it must be free standing—a word or a phrase, rather than an affix or some other bound morphological unit. Accordingly, various kinds of clitics, particles, and other intermediate units smaller than words but larger than affixes may provide borderline cases straddling the boundary between what is a quantifier and what is not. The distinction between free and bound forms is what underlies the distinction between quantification and the semantically-related grammatical category of *number*. Specifically, whereas quantifiers are free standing, number markers are bound forms. However, in some instances it may be difficult to adjudicate whether a form associated with a quantificational meaning is free or bound, and hence whether it constitutes an instance of quantification or of number. In order to determine whether forms such as these are quantifiers or number markers, it may be necessary to make use of additional diagnostic criteria distinguishing quantifiers from number markers.

The second part of the definition of quantifier is the obvious semantic condition: its meaning must involve the notion of quantity. While in many instances this is quite straightforward, in other cases, quantificational meanings may shade off gradually into other kinds of meanings not generally considered to be quantificational. One class of expressions related in meaning to nominal quantifiers are those denoting size and other scalar properties. An important class of expressions related to verbal quantifiers are those marking aspectual categories, such as perfect and progressive. More generally, different quantifiers in different languages may be related to an extremely broad range of non-quantificational concepts, synchronically and/or diachronically.

See also: Lexical Semantics; Logic and Linguistics; Numeral Systems; Semantics; Word Classes and Parts of Speech

Bibliography

- Allan K 1980 Nouns and countability. *Language* 56: 541–67
- Bach E, Jelinek K E, Kratzer A, Partee B H (eds.) 1995 *Cross-Linguistic Quantification*. Kluwer Academic Publishers, Dordrecht, The Netherlands
- Craig C (ed.) 1986 Noun classes and categorization. *Typological Studies in Language* 7
- Gil D 1982a Distributive numerals. PhD. dissertation, University of California, Los Angeles, CA
- Gil D 1982b Quantifier scope, linguistic variation, and natural language semantics. *Linguistics and Philosophy* 5: 421–72
- Gil D 1993 Nominal and verbal quantification. *Sprachtypologie und Universalienforschung* 46: 275–317

- Gil D 1995 Universal quantifiers and distributivity. In: Bach E, Jeline K E, Kratzer A, Partee B H (eds.) *Quantification in Natural Languages*. Kluwer, Dordrecht, The Netherlands, pp. 321–62
- Heim I 1982 The semantics of definite and indefinite noun phrases. PhD. dissertation, University of Massachusetts, Amherst, MA
- Hurford J R 1975 *The Linguistic Study of Numerals*. Cambridge University Press, Cambridge, UK
- Jackendoff R S 1971 *Semantic Interpretation in Generative Grammar*. MIT Press, Cambridge, MA
- Keenan E L, Stavi J 1986 A semantic characterization of natural language determiners. *Linguistics and Philosophy* 9: 253–326
- Menninger K 1969 *Number Words and Number Symbols* [trans. Broneer P.]. MIT Press, Cambridge, MA
- Milsark G 1977 Toward an explanation of certain peculiarities in the existential construction in English. *Linguistic Analysis* 3: 1–30
- Pelletier F J (ed.) 1979 *Mass Terms: Some Philosophical Problems*. Synthese Language Library, Reidel, Dordrecht, The Netherlands
- Vendler Z 1967 *Linguistics and Philosophy*. Cornell University Press, Ithaca, NY

D. Gil

Quantitative Cross-national Research Methods

1. The Main Challenges of Quantitative Nation Comparisons

Quantitative cross-national comparisons have become an important feature of all the social sciences. Their popularity has been spurred, in part, by the increasing availability of good data for a growing number of countries and, in part, by the diffusion of statistical skills among researchers. There are many constraints and potential pitfalls involved in cross-national statistical analyses, and these will be the main theme of this article. But this research strategy has, regardless, proven itself of great scientific value. It has, above all, helped highlight how fundamental many international variations are, thus, providing a powerful check against parochialism and unwarranted generalizations. Quantitative macrocomparisons have also proven themselves eminently useful in testing the validity of core theoretical assumptions, such as the convergence thesis that has long dominated development and modernization theory in political science and sociology especially.

Quantitative nation comparisons will, however, inevitably face a number of tradeoffs; some more serious than others. One is that the capacity for broader generalization that quantification permits is bought at the price of less attention to the contextual reality of individual nations. The uniqueness of a

nation's culture, historical heritage, and endemic logic is lost easily when all we actually observe are a few indicators. Moreover, these indicators may mean somewhat different things in the unique context of specific countries (Ragin 1987). Scholars are frequently quite sensitive to this issue and so attempt to embed their statistical analyses within the histories and social uniqueness of the countries being studied. Also, as Ragin (1994) argues and empirically demonstrates, Boolean analysis can help overcome this dilemma by enabling us to build conjunctural models with very few cases and to analyze nonevents. But the approach is somewhat circumscribed. It needs to be guided by strong theory and background knowledge, its applicability is limited to relatively few cases, and it may be biased in favor of corroborating nonadditive, conjunctural models.

A second constraint has to do with the often limited number of observations available. This is especially true of studies of advanced (OECD) democracies where the N rarely exceeds 25. In broader worldwide comparisons, however, the N approaches 200. Many researchers attempt to supplement the paucity of nations by collecting longitudinal data on each nation, as with pooled cross-sectional and time series analyses. A 20-year time series for 20 countries yields 400 observations; a 20-year time series for 200 nations yields 4,000. As ever, longer data series for individual countries become available, the small N problem will gradually diminish.

Small samples do not necessarily pose problems of statistical inference, especially when the substantive theory is strong. They do, however, limit what statistical tools can be brought to bear. Fearon (1991) argues that the smaller the sample size, the greater is the need to make counterfactuals explicit—in other words, to tighten the theoretical formulation about the precise conditions that will (or will not) produce an outcome. Vague theory implies uncertainty about the number of contending explanations (which can grow very large) and about their mutual relationship within a causal order. The consequences of this include causal misspecification, multi-colinearity and large error terms. The last two of these can only be managed by having more observations. Without them, estimations are likely to yield results that are not robust. The less ambiguous the theory, the less the need for large N 's. Small N 's are, therefore, an especially acute problem in disciplines without a firm theoretical architecture, like sociology or political science. Western (1998a) argues that Bayesian estimation, by allowing for greater uncertainty, produces more robust results when N 's are few, theory vague, and explanations are many.

Small N 's can even hold advantages, because a limited nation sample implies that the researcher can gain more easily maximum advantage from diagnostic scrutiny of the residual plots. A set of systematic outliers, for example, can be easily identified in terms

of their nationhood, and any good comparativist should be able to pin down what variable(s) drives their deviance from the fitted regression line. Diagnostics on small samples can approach the advantages of in-depth case studies.

The constraints imposed by small samples are generally well known to researchers. There are additional problems inherent in quantitative national comparisons that are both more generic and potentially more serious, but also less recognized. Chief among these are issues related to: qualitative and limited dependent variables, the potential for selection bias, possible lack of independence between observations and, finally, the question of endogeneity of variables.

2. Qualitative and Limited Dependent Variables

Many dependent variables in cross-national research are either qualitative (assuming discrete values) or limited (they assume continuous values within some range). The choice whether to measure a variable in a qualitative or continuous way is often controversial. Bollen and Jackman (1989), for example, argue that difficulties in classifying some political regimes speak in favor of using continuous scales because 'Dichotomizing democracy blurs distinctions between borderline cases.' In contrast, Przeworski et al. (in press) prefer to treat regimes dichotomously or multinomially.

When the dependent variable is either qualitative or limited, we need to use nonlinear models. Whether we give political regimes the values of 0–1 (as do Przeworski et al. in press) or 1–100 (as does Bollen 1980), it remains that the value on the dependent variable cannot exceed its maximum when the independent variable(s) tend to infinity, and it cannot fall below its minimum when these variables tend to infinity. Linear models, at best, can provide an approximation within some range of the independent variables.

The standard model when the dependent variable is multinomial is

$$\Pr(Y = j, X = x) = F(x) \quad (1)$$

where $j = 0, 1, \dots, J-1$ and F is the cumulative distribution function. Since such models are covered by any standard textbook we need not present them here. Such models can be applied to panel data unless the number of repeated observations is large.

'Event history analysis' is one particular class of nonlinear models applied in cross-national research. In such models the dependent variable is an event, such as a revolution, regime transition, or a policy adoption. The general model is

$$\Pr(Y(t+dt) = j) = f(y(t), x(t)), dt \quad (2)$$

Most often, such models can be conveniently estimated as

$$\log S(t) = \log(1 - F(t)) \quad (3)$$

where $S(t)$ is the survival function, or the probability that an event lasts beyond time t , and $F(t)$ is the cdf.

For dichotomous dependent variables, logit and probit give very similar results. For multinomial variables, it is often assumed that the errors are independent across the values of the dependent variable, which leads to a logit specification. But this implies a strong assumption, namely the irrelevance of independent alternatives that rarely holds in practice. Multinomial probit, in turn, requires computing multiple integrals, which was until recently computationally expensive. Alternatives would be semi- and nonparametric methods.

The distributions that are commonly used in estimating survival models include the exponential, Weibul, logistic, and Poisson distributions. But, except for the Weibul and exponential, it is very difficult to distinguish statistically between such distributions.

Methods for studying qualitative dependent variables are now standard textbook fare, but what warrants emphasis is that the traditional distinction between 'qualitative' and 'quantitative' research is becoming increasingly obsolete. Phenomena such as social revolutions may be rare, but this just means they occur with a low probability. They can, nevertheless, be studied systematically using maximum likelihood methods.

3. The Problem of Selection Bias

In comparative research we are often interested in the effect of some systemic feature, institution, or policy on some outcome. Examples include the effect of labor market institutions on unemployment, the effect of political regimes on economic growth, or the effect of electoral rules on the number of political parties.

In such cases, the question is how some feature of X affects some outcomes Y in the presence of conditions Z , or $PR(Y, X, Z)$, where the hypothesis is that $y = f(x)$. The generic problem is how to isolate the effects of X and Z on Y .

If sampling is exogenous (i.e., the probability of observing any y is independent of z), the $pr(Y = yz) = pr(y)$, and standard statistical methods can be used. But if $pr(Y = yz)$ is different under different conditions Z , sampling is endogenous and the assumptions of the standard statistical model are violated.

To exemplify the problem: suppose we want to know the impact of political regimes on economic growth. We observe Chile in 1985 (Z), which was a dictatorship (X), in which case *per capita* income declined at the rate of $Y = -2.26$. To assess the effect of political regimes, we need to know what would have

been the rate of growth (Y) in Chile had it been a democracy. This cannot be answered with the available observations because Chile in 1985 was indeed a dictatorship. The comparativist in such a case would proceed quasi-experimentally, and look for a case that matches Chile in all aspects except authoritarianism. In this way one would compare an authoritarian with a democratic 'Chile.'

But what if Chile's status as a dictatorship in 1985 was due to some factors that also affected its economic performance? This could be because growth itself influences the survival of regimes; a country's level of development may be associated both with regime selection and growth; some unobservable factors (say 'enlightened leadership' or, perhaps, measurement error) may be common to both variables. Whatever the underlying selection mechanism, the basic consequence is that there will be cases without a match, and this implies that our inferences will be biased.

The problem raised by nonrandom selection is how to make inferences from what we observe to what we do not. It is a problem of identification (Manski 1995). For what do we know when we observe Chile as a dictatorship in 1985? We observe the fact that, given conditions Z , the Chilean regime (X) is a dictatorship. We also know its rate of economic growth, given that it is a dictatorship under conditions Z . We do not observe, however, its rate of growth as a democracy under conditions Z . The issue is one of counterfactual observations: the rate of growth that an observation with $Z = zi$, observed under dictatorship, would have had under democracy. We do not know what this value is and, hence, we face under-identification.

What we need to know are two distributions, $P(Y = y, x = j, Z)$ and $P(Y = y, x = k, Z)$, and their expected values, where we now think of the possible values of X more generally as $j, k = 0, 1, 2, \dots, N$. The first is the distribution of Y in the entire population if all cases were observed as dictatorships; the second is the distribution of Y in the entire population if all cases were observed as democracies. Clearly, if we want to know the impact on y of being in states x , we need to determine the difference between these two distributions, conditional on z .

It may seem strange to refer to the 'population' as including unobserved cases. But in order to compare the effects of states x on the performance of an individual case characterized by a zi , we must allow the case to be potentially observable under the full range of X . We observed this case as $X = j$, but must also imagine the possibility that it would have been Xj . We must think in terms of a 'superpopulation' consisting of a continuum of potential cases. The actual sample is then regarded as having been drawn by 'nature' from this superpopulation.

If the actual population is an exogenous sample of the potential one, any multivariate combination (xi, zi) can be drawn with a positive probability. Such exogenous sampling allows us to isolate the effect of x

on y given z , that is, to 'control' for z the effects of x on y . If the sample is exogenous, we can 'match' pairs ($xi = j, zi$) and ($xi = k, zi$). But if sampling is endogenous, there will be some cases without a 'match': some of the xi 's will not have support. Where we have nonrandom selection, quasi-experimental comparisons fail regardless of the number of observations.

The methods to correct for selection bias consist of constructing counterfactuals, that is, of filling the unobserved supports of the distribution Y for all X . Following Heckman (1979), when sampling is not random, the regressions $y = f(x, z)$ suffer from omitted variable bias. When such regressions are estimated on the basis of the observed sample, the variable that is omitted is the expected value of the error in the equation that specifies how the observations are selected into the sample

$$\Pr(x = j) = F(Z) + u \quad (4)$$

If

$$E(uj \mid x = j) = 0 \quad (5)$$

then

$$E(uj \mid x) = juE(u \mid x) \quad (6)$$

where ju is a regression coefficient of uj on u .

Finally, $E(ux = j) = j$, where the f 's are the inverse Mill ratios, or the hazard rates. Since we know that Yj is observed when $x = j$, we can write the expected values in the observed sample as

$$E(yj \mid x = j, z) = zj + juj \quad (7)$$

Note that if this equation is estimated on the basis of the observed sample, the variable j is omitted from the specification.

We can now see why controlling for the variables that enter both into selection and outcome equations may indeed worsen the selection bias. Following Heckman (1988), we distinguish first between selection on observables and on unobservables. Selection on observables occurs when the expected covariance $E(uju \mid z) \neq 0$, but once the observed variables Z are controlled for it vanishes, so that $E(uju \mid z) = 0$. Selection is on unobservables when $E(uju) \neq 0$ and $E(uju \mid z) = 0$, which means that controlling for the factors observed by the investigator does not remove the covariance between the errors in the outcome and the selection equations. Now note that the regression coefficient $ju = \text{cov}(uju) / \text{var}(uj)$. If selection is on unobservables, controlling for some variable x in the outcome equation may reduce the error variance uj without equally reducing the covariance uju . Hence, the coefficient on the omitted variable will be larger and the bias will be exacerbated.

In short, the expected values of the observed cases will be biased because they co-vary with the variable which determines which cases are observed. If selection is exclusively on observables, this bias can be corrected by traditional controlling techniques. But if it is on unobservables, such controls only worsen the bias.

Correcting for selection bias is not uncontroversial. It has been found that corrections for selection are not robust, but are highly sensitive to relatively minor changes in assumptions about distributions (Goldberger 1983). Others have found that some estimation methods fail to correct for this bias and may even exacerbate it (Stolzenberg and Relles 1990). As Heckman (1988) argues, the quandary we face is that different methods of correcting for selection bias are robust if there is no bias to begin with; if there is, there is no guarantee that the methods are robust.

The logic of the problem is similar whether we study large or small N 's. When de Toqueville concluded that revolutions do not bring about social change, the reason might be that they occur only in countries where it is difficult to change society. Even when $N = 1$, the issue of selection bias does not disappear: The French revolution in 1789 may have been caused by the same conditions that made social change so difficult. It is possible that a revolution in a country where social relations are easier to change would have provoked change. But then a revolution would not have been necessary.

Comparativists who conduct case studies cannot benefit from statistical distributions to generate the counterfactuals. Yet, as in all cases of comparison the problem remains and it should, therefore, be standard practice to ask counterfactual questions of one's case or cases.

4. The Problem of Independence

Statistical inference must assume that the observations on a variable are independent one of the other. Is country A's performance truly independent of what happens in country B? Is what happens at $t+1$ independent of events in t ? Usually not, and this implies the need for corrective procedures. In most cases, however, rigorous correction will entail that the *de facto* N (nations or years) diminishes; in some instances, statistical dependency cannot be resolved at all.

Cross-sectional analysis almost invariably assumes that nations and their properties (say budgets or institutions) are independent one of the other. This should not be assumed. We know that the Scandinavian countries have a shared history, deliberately learning from each other through centuries, thus creating similar institutions and path dependencies. The same goes for Austria and Germany, for Belgium and The Netherlands and, arguably, for all the Anglo-saxon nations. World samples have a similar problem: Japan's long hegemony in East Asia will have influenc-

ed Korean society; Confucianism has had a pervasive influence throughout the region. Similar stories are easily told for Latin America and Africa. If the World is a set of nation clusters, the real N is not 20-odd OECD countries or 150-odd World nations.

When nations form families, but are treated as if they were all unique and independent, we are likely to get biased coefficients and, very probably, unequal error variance (heteroskedasticity). Sweden alone will drive the regression line in just about any welfare state analysis, and when also Denmark and Norway are treated as discrete observations, the bias is multiplied in so far as all three in reality form part of the same political family ('Scandinavia'). Diffusion effects that operate between members of a nation-cluster can also result in heteroskedastic disturbance in the cross-section. Such can be corrected by, for example, adding a variable that captures the common underlying property that drives the disturbance (say, a dummy for being 'Scandinavia') but, again, this correction absorbs precious degrees of freedom in a small N study and, substantively, amounts to reducing the three nations to one observation.

Lack of independence in a time-series is normally taken for granted, since this year's budget or election outcome is almost inevitably related to last year's budget or the previous election. The standard assumption is a first-order (AR1) serial correlation. In comparative research virtually all time-series applications are pooled with cross-sections. But, where N 's are very small, one may as well simply compare across individual time-series estimations, as do Esping-Andersen and Sonnberger (1991). Time-series are meant to capture historical process. Yet as Isaac and Griffin (1989) argue, they easily end up being a-historical.

Pooling cross-sectional with time-series data (panel regressions) has become very widespread, especially in studies of the limited group of advanced (OECD) societies. In many cases, the panel design is chiefly cross-sectional (more nations than years); others are temporally dominated (for a discussion, see Stimson 1985). Panel models are especially problematic because they can contain simultaneous diachronic and spatial interdependence and, worse, the two may interact. The standard method for correcting contemporaneous error correlation (GLS) applies only where the t 's well exceed nations (which is rare). The consequence is that t -statistics are overestimated, errors underestimated, and the results may therefore not be robust (Beck and Katz 1995). The Beck and Katz (1995) procedure can correct for temporal and cross-sectional dependency one at a time, but if the two interact, no solution exists. See also Beck et al. (1998) for an application to maximum likelihood estimation.

Panel models can be based on two types of theoretical justification. One is that events or shocks occur over time that affect the cross-sectional variance.

There is, for example, a huge recent literature on the impact of labor market 'rigidities' on unemployment: regulations vary across nations but also across time because of deregulatory legislation (see, for example, Nickell 1997). Deregulation in a country should produce a break in its time series, and the autocorrelation element will be split into the years preceding and following the break. The second justification, not often exploited, is to interpret autocorrelation as an expression of institutional or policy path dependency. In this instant, the rho must be treated as a variable. The problem, of course, is that the rho is likely to combine theoretically relevant information as well as unknown residual autocorrelation. The researcher can accordingly not avoid including a variable that explicitly measures path dependency.

If we insist on faithful adherence to the real world, panel regressions may require so much correction against dependency that the hard-won additional degrees of freedom that come with a time-series are easily eaten up. And how many can truthfully claim that time and country dependencies do not interact? Indeed, most sensible comparativists would assume they do: if nations form part of families it should also be the case that the timing of their shocks, events, or policies is interdependent. Such intractable problems are certainly much more severe in small- N comparisons. Here, a marginal difference in measurement, the inclusion or exclusion of one country, the addition or subtraction of a year here or there, or the substitution of one variable for another, can change the entire model.

One alternative is to construct multilevel models which explicitly take into account the possibility that nations may 'cluster' (for an overview, see Goldstein 1987). Nieuwbeerta and Ultee (1999) have, for example, estimated a three level (nation, time, and individual) model of the impact of class on party choice within the context of nations' social mobility structure. A second alternative, in particular when the dependent variable is categorical, is to exploit the advantages of event history analysis. But, here the time series needs to be quite long considering that theoretically interesting events, such as revolutions, democratization, or even welfare reforms, are far between. In the event history context, analytical priority usually is given to temporal change, which brings it much closer to traditional time series analysis. But rather than having to manipulate autocorrelation, time sequencing (states and events) is actively modeled and thus gains analytic status. For an application to nation comparisons, see, for example, Western (1998b), which also can stand as an exemplar of how to minimize the interdependency problem.

There are two particular cases where the lack of independence among observations simply prohibits adequate estimation. The first, noted above, occurs when time and nation dependencies interact. The

second is when 'globalization' penetrates all nations and when many nations (such as the European Union) become subsumed under identical constraints. In this instance, existing cross-national correlations will strengthen and we may, indeed be moving towards an $N = 1$. Of course, global shocks or European Union membership do not necessarily produce similar effects on the dependent variable across nations or time. If nations' institutional filters differ, so will most likely the impact of a global shock on, say, national unemployment rates. Here we would specify interaction effects, but that would be impossible in a pure cross-section, and extremely difficult in a time series, unless we already know how the lag structure will differ according to institutional variation.

5. The Endogeneity Problem

All probabilistic statistics require conditional independence, namely that the values of the predictor variables are assigned independently of the dependent variable. The basic problem of endogeneity occurs when the explanans (X) may be influenced by the explanandum (Y) or both may be jointly influenced by an unmeasured third. The endogeneity problem is one aspect of the broader question of selection bias discussed earlier.

The endogeneity issue has been debated intensely within the economic growth literature in terms of the causal relationship between technology and growth. But it applies equally to many fields. For example, comparativists often argue that left power explains welfare state development. But are we certain that left power, itself, is not a function of strong welfare states? Or, equally likely, are both large welfare states and left power just two faces of the same coin, different manifestations of one underlying, yet undefined, phenomenon? Would Sweden have had the same welfare state even without its legendary social democratic tradition? Perhaps, if Sweden's cultural past over-determines its unique kind of social democracy and social policy. If this kind of endogeneity exists, the true X for Sweden is not left power but a full list of all that is 'Sweden.' The vector of the X 's becomes a list of all that is nationally unique.

The endogeneity problem becomes easily intractable in quantitative cross-national research because we observe variables (Y 's and X 's) that represent part of the reality of the nations we sample. Our variables are in effect a partial reflection of the society under study, and the meaning of a variable score for one nation may not be metrically equivalent to that of another—a weighted left cabinet score of 35 for Denmark and 40 for Sweden probably misrepresents the Danish–Swedish difference if, that is, the two social democracies are two different beasts.

A related, and equally problematic, issue arises in the interpretation of coefficient estimations. In the

simple cross-sectional regression, the coefficient for left power, economic development or for coordinated bargaining can only be interpreted in one way: as a constant, across-the-board effect irrespective of national context. If Denmark had 5 points more left power, its welfare state should match the Swedish (all else held constant). In fixed-effects panel regressions, the X 's are assumed to have an identical impact on Y , irrespective of country. Often we cannot accept such assumptions.

The reason that we cannot is that it is difficult to assume that variables, say a bargaining institution or party power balance, will produce homogenous, monotonically identical, effects across nations for the simple reason that they are embedded in a more complex reality (called nation) which has also given rise to its version of the dependent variable. In this case, the bias remains equally problematic whether we study few or many N 's; it will not disappear if we add more nations. Quantitative national comparisons rarely address such endogeneity problems. Studies of labor market or economic performance routinely presume that labor market regulations or bargaining centralization are truly exogenous variables, whose effects on employment is conditionally identical whether it is Germany, Norway, or the United States. The same goes for welfare state comparisons with their belief that demography and left power are fully exogenous and conditionally unitary.

There exist several, not necessarily efficient, ways of correcting for the endogeneity bias. If we have some inkling that the bias comes from variable omission, the obvious correction entails the inclusion of additional controls. An example of this kind was Jackman's (1986) argument that Norway's North Sea Oil was over-determining the results in the Lange and Garrett (1985) study. Small N studies with strong endogeneity have little capacity to extend the number of potentially necessary controls. A second approach is to limit endogeneity in X by reconceptualizing and, most likely, narrowing Y . The welfare state literature provides a prototypical example: aggregate social expenditure increasingly was replaced by measures of specific welfare state traits.

Controls, no matter how many, will however not resolve the problem under conditions of strong sampling bias. As discussed above, the best solution in such a situation is to concentrate more on the theoretical elaboration of causal relations between variables. If we can assume that our estimations are biased because Y affects the values on X , or because both are jointly attributable to a third underlying force, thinking in counterfactual terms ('would Sweden's welfare state be the same without Sweden's social democracy') will force the researcher to identify more precisely the direct or derived causal connections (Fearon 1991). If bias can be assumed to come from the assumption of monotonically homogeneous effects of the X across all nations, the researcher's attention

should concentrate on identifying more precisely the conditional mechanisms that are involved in the causal passage from an X to a Y (why and how will a 5-point rise in left power make the Danish welfare state converge with the Swedish?).

6. Conclusion

The difficulties and tradeoffs that we have synthesized above are intrinsic to all quantitative research, but are often more difficult to deal with in national comparisons because of the inevitability of small samples of countries and of the ways in which nations have developed together in interaction and will continue to do so even more in the future, given globalization. However, the difficulties should not be seen as paralyzing research but rather as stimulating attempts to deal with them. Indeed, in most practical situations today, each of these problems can be dealt with, albeit not perfectly. So, today's partial vices may be the impetus to tomorrow's virtues. Although we may never be able to completely resolve all problems, ongoing trends suggest that our capacity to conduct good quality research in this field is improving rapidly. For one, the constraints of small N 's are diminishing because of rapid progress in data quality and quantity, covering both more years and nations. Second, the field has made great strides in terms of researchers' substantial comparative knowledge of the nations and cultures under study. Hence, our ability to interpret and contextualize statistical findings has greatly strengthened. And this, in turn, is a major precondition for what we believe is the single most effective tool with which to manage the challenges of quantitative comparison: namely, to systematically think in terms of counterfactuals.

Bibliography

- Amemyia T 1985 *Advanced Econometrics*. Harvard University Press, Cambridge, MA
- Beck N, Katz J N 1995 What to do (and not to do) with times-series, cross-section data. *American Political Science Review* **89**: 634–47
- Beck N, Katz J N, Tucker R 1998 Taking time seriously: Time series-cross-section analysis with a binary dependent variable. *American Journal of Political Science* **42**: 1260–88
- Blossfeld H P, Roehwer G 1995 *Techniques of Event History Modelling: New Approaches to Causal Analysis*. L. Erlbaum, Mahwah, NJ
- Bollen K A 1980 Issues in the comparative measurement of political democracy. *American Sociological Review* **45**: 370–90
- Bollen K A, Jackman R W 1989 Democracy, stability, and dichotomies. *American Sociological Review* **54**: 612–21
- Esping-Andersen G, Sonnberger H 1991 The demographics of age in labor market management. In: Myles J, Quadagno J (eds.) *States, Labor Markets, and the Future of Old-Age Policy*. Temple University Press, Philadelphia, PA, pp. 227–49

- Fearon J D 1991 Counterfactuals and hypothesis testing in political science. *World Politics* **43**: 169–95
- Goldberger A 1983 Abnormal selection bias. In: Karlin S, Amemyia T, Goodman L A (eds.) *Studies in Econometrics, Time Series, and Multivariate Statistics*. Academic Press, New York
- Goldstein H 1987 *Multilevel Models in Educational and Social Research*. Oxford University Press, New York
- Heckman J J 1988 The microeconomic evaluation of social programs and economic institutions. In: *Chung-Hua Series of Lectures by Invited Eminent Economists no. 14*. The Institute of Economics, Academia Sinica, Taipei, China
- Hsiao C 1986 *Analysis of Panel Data*. Cambridge University Press, Cambridge, UK
- Isaac L W, Griffin L J 1989 A historicism in time-series analyses of historical process. *American Sociological Review* **54**: 873–90
- Jackman R 1986 The politics of economic growth in industrial democracies, 1974–1980: Leftist strength or North Sea oil? *Journal of Politics* **49**: 242–56
- Janoski T, Hicks A (eds.) 1994 *The Comparative Political Economy of the Welfare State*. Cambridge University Press, Cambridge, UK
- Lange P, Jarrett G 1985 The politics of growth. *Journal of Politics* **47**: 792–827
- Maddala G S 1983 *Limited-dependent and Qualitative Variables in Econometrics*. Cambridge University Press, Cambridge, UK
- Nickell S 1997 Unemployment and labor market rigidities: Europe versus North America. *Journal of Economic Perspectives* **11**(3): 55–74
- Nieuwebeerta P, Ultee W 1999 Class voting in Western industrialized countries, 1945–1990: systematizing and testing explanations. *European Journal of Political Research* **35**: 123–60
- Przeworski A, Alvarez M R, Cheibub J A, Limongi F in press *Democracy and Development: Political Regimes and Material Welfare in the World, 1950–1990*. Cambridge University Press, New York
- Ragin C 1987 *The Comparative Method*. University of California Press, Berkeley, CA
- Ragin C 1994 A qualitative comparative analysis of pension systems. In: Janoski T, Hicks A M (eds.) *The Comparative Political Economy of the Welfare State*. Cambridge University Press, Cambridge, UK, pp. 320–45
- Stimson J A 1985 Regression in time and space: A statistical essay. *American Journal of Political Science* **29**: 914–47
- Western B 1998a Causal heterogeneity in comparative research: A Bayesian hierarchical modeling approach. *American Journal of Political Science* **42**: 1233–59
- Western B 1998b *Between Class and Market: Postwar Unionization in the Capitalist Democracies*. Princeton University Press, Princeton, NJ

G. Esping-Andersen and A. Przeworski

Quasi-Experimental Designs

The theory of quasi-experimentation addresses one of the great unanswered questions in the social and

behavioral sciences: how can researchers make a valid causal inference when units cannot be randomly assigned to conditions? Like all experiments, quasi-experiments deliberately manipulate presumed causes to discover their effects. However, quasi-experiments are also defined by what they do not do: the researcher does *not* assign units to conditions randomly. Instead, quasi-experiments use a combination of design features, practical logic, and statistical analysis to show that the presumed cause is likely to be responsible for the observed effect, and other causes are not. The term nonrandomized experiment is synonymous with quasi-experiment; and the terms observational study and nonexperimental design often include quasi-experiments as a subset. This article discusses the need for quasi-experimentation, describes the kinds of designs that fall into this class of methods, reviews the intellectual and practical history of these designs, and notes important current developments.

1. The Need for Quasi-experimentation

Given the desirable properties of randomized experiments, one might question why quasi-experiments are needed. When properly implemented, randomized experiments yield unbiased estimates of treatment effects, accompanied by known probabilities of error in identifying effect size. Quasi-experimental designs do not have these properties. Yet quasi-experiments are necessary in the arsenal of science because it is not always possible to randomize. Ethical constraints may preclude withholding treatment from needy people based on chance, those who administer treatment may refuse to honor randomization, or questions about program effects may arise after a treatment was already implemented so that randomization is impossible. Consequently, the use of quasi-experimental designs is frequent and inevitable in practice.

2. Kinds of Quasi-experimental Designs

The range of quasi-experimental designs is large, including but not limited to: (a) Nonequivalent control group designs in which the outcomes of two or more treatment or comparison conditions are studied but the experimenter does not control assignment to conditions; (b) Interrupted time series designs in which many consecutive observations over time (prototypically 100) are available on an outcome, and treatment is introduced in the midst of those observations to demonstrate its impact on the outcome through a discontinuity in the time series after treatment; (c) Regression discontinuity designs in which the experimenter uses a cutoff score on a measured variable to determine eligibility for treatment, and

an effect is observed if the regression line (of the assignment variable on outcome) for the treatment group is discontinuous from that of the comparison group at the cutoff score; (d) Single-case designs in which one participant is repeatedly observed over time (usually on fewer occasions than in time series) while the scheduling and dose of treatment are manipulated to demonstrate that treatment controls outcome.

In the preceding designs, treatment is manipulated, and outcome is then observed. Two other classes of designs are sometimes included as quasi-experiments, even though the presumed cause is not manipulated (and often not even manipulable) prior to observing the outcome. In (e) case-control designs, a group with an outcome of interest is compared to a group without that outcome to see if they differ retrospectively in exposure to possible causes in the past; and in (f) correlational designs, observations on possible treatments and outcomes are observed simultaneously, often with a survey, to see if they are related. Because these designs do not ensure that cause precedes effect, as it must logically do, they usually yield more equivocal causal inferences.

3. The History of Quasi-experimental Designs

Quasi-experimental designs have an even longer history than randomized experiments. For example, around 1,850 epidemiologists used case-control methods to identify contaminated water supplies as the cause of cholera in London (Schlesselman 1982), and in 1898, Triplett used a nonequivalent control group design to show that the presence of audience and competitors improved the performance of bicyclists. In fact, nearly all experiments conducted prior to Fisher's work were quasi-experiments.

However, it was not until 1963 that the term quasi-experiment was coined by Campbell and Stanley (1963) to describe this class of designs. Campbell and his colleagues (Cook and Campbell 1979, Shadish et al. in press) extended the theory and practice of these designs in three ways. First, they described a large number of these designs, including variations of the designs described above. For example, some quasi-experimental designs are inherently longitudinal (e.g., time series, single case designs), observing participants over time, but other designs can be made longitudinal by adding more observations before or after treatment. Similarly, more than one treatment or control group can be used, and the designs can be combined, as when adding a nonequivalent control group to a time series.

Second, Campbell developed a method to evaluate the quality of causal inferences resulting from quasi-experimental designs—a validity typology that was elaborated in Cook and Campbell (1979). The typology includes four validity types and threats to validity for each type. Threats are common reasons

why researchers may be wrong about the causal inferences they draw. Statistical conclusion validity concerns inferences about whether and how much presumed cause and effect co-vary; examples of threats to statistical conclusion validity include low statistical power, violated assumptions of statistical tests, and inaccurate effect size estimates. Internal validity concerns inferences that observed co-variation is due to the presumed treatment causing the presumed outcome; examples include history (extraneous events that could also cause the effect), maturation (natural growth processes that could cause an observed change), and selection (differences between groups before treatment that may cause differences after treatment). Construct validity concerns inferences about higher-order constructs that research operations represent; threats include experimenter expectancy effects whereby participants react to what they believe the experimenter wants to observe rather than to the intended treatment, and mono-operation bias in which researchers use only one measure that reflects a construct imperfectly or incorrectly. External validity concerns inferences about generalizing a causal relationship over variations in units, treatments, observations, settings, and times; threats include interactions of the treatment with other features of the design that produce unique effects that would not otherwise be observed.

Third, Campbell's theory emphasized addressing threats to validity using design features—things that a researcher can manipulate to prevent a threat from occurring or to diagnose its presence and potential impact on study results (see Table 1). For example, suppose maturation (normal development) is an anticipated threat to validity because it could cause a pretest-post-test change like that attributed to the treatment. The inclusion of several consecutive pretests before treatment can indicate whether the rate of maturation before treatment is similar to the rate of change from during and after treatment. If it is similar, maturation is a threat. All quasi-experiments are combinations of these design features, thoughtfully chosen to diagnose or rule out threats to validity in a particular context. Conversely, Campbell was skeptical about the more difficult task of trying to adjust threats statistically after they have already occurred. The reason is that statistical adjustments require making assumptions, the validity of which are usually impossible to test, and some of which are dubious (e.g., that the selection model is known fully, or that the functional form of errors is known).

Other scholars during this time were also interested in causal inferences in quasi-experiments, such as Cochran (1965) in statistics, Heckman (1979) in economics, and Hill (1953) in epidemiology. However, Campbell's work was unique for its extensive emphasis on design rather than statistical analysis, for its theory of how to evaluate causal inferences, and for its sustained development of quasi-experimental theory

Table 1

Design elements used in constructing quasi-experiments

Assignment (Control of assignment strategies to increase group comparability)*Cutoff-based assignment.* Controlled assignment to conditions based solely on one or more fully measured covariates. This can yield an unbiased effect estimate*Other nonrandom assignment.* Various forms of ‘haphazard’ assignment that sometimes approximate randomization (e.g., alternating assignment in a two condition quasi-experiment whereby every other unit is assigned to one condition, etc.)*Matching and stratifying.* Efforts to create groups equivalent on observed covariates in ways that are stable, do not lead to regression artifacts, and are correlated with the outcome. Preference is for pretreatment measures of the outcome itself*Measurement* (Use of measures to learn whether threats to causal inference actually operate)

Post-test observations

Nonequivalent dependent variables. Measures that are not sensitive to the causal forces of the treatment, but are sensitive to all or most of the confounding causal forces that might lead to false conclusions about treatment effects (if such measures show no effect, but the outcome measures do show an effect, the causal inference is bolstered because it is less likely due to the confounds)*Multiple substantive post-tests.* Used to assess whether the treatment affects a complex pattern of theoretically predicted outcomes

Pretest observations

Single pretest. A pretreatment measure on the outcome variable, useful to help diagnose selection bias*Retrospective pretest.* Reconstructed pretests when actual pretests are not feasible—by itself, a very weak design feature, but sometimes better than nothing*Proxy pretest.* When a true pretest is not feasible, a pretest on a variable correlated with the outcome—also often weak by itself*Multiple pretest time points on the outcome.* Helps reveal pretreatment trends or regression artifacts that might complicate causal inference*Pretests on independent samples.* When a pretest is not feasible on the treated sample, one is obtained from a randomly equivalent sample*Complex predictions* such as *predicted interactions.* Successfully predicted interactions lend support to causal inference because alternative explanations become less plausible*Measurement of threats to internal validity.* Help diagnose the presence of specific threats to the inference that A caused B such as whether units actively sought out additional treatments outside the experiment*Comparison groups* (Selecting comparisons that are ‘less nonequivalent’ or that bracket the treatment group at the pretest(s))*Single nonequivalent groups.* Compared to studies without control groups, using a nonequivalent control group helps identify many plausible threats to validity*Multiple nonequivalent groups.* Serve several functions. For instance, groups are selected that are as similar as possible to the treated group but at least one outperforms it initially and at least one underperforms it, thus bracketing the treated group*Cohorts.* Comparison groups chosen from the same institution in a different cycle (e.g., sibling controls in families or last year’s students in schools)*Internal (vs. external) controls.* Plausibly chosen from within the same population (e.g., within the same school rather than from a different school)*Treatment* (Manipulations of the treatment to demonstrate that treatment variability affects outcome variability)*Removed treatments.* Showing an effect diminishes if treatment is removed*Repeated treatments.* Reintroduces treatments after they have been removed from some group—common in laboratory sciences or where treatments have short-term effects*Switching replications.* Reverses treatment and control group roles so that one group is the control while the other receives treatment, but the controls receive treatment later while the original treatment group receives no further treatment or has treatment removed*Reversed treatments.* Provides a conceptually similar treatment that reverses an effect—e.g., reducing access to a computer for some students but increasing access for others*Dosage variation.* Demonstrates that outcome responds systematically to different levels of treatment

and method over four decades. Both the theory and the methods he outlined were widely adopted in practice during the last half of the twentieth century,

and his terms like internal and external validity became so much a part of the scientific lexicon that today they are often used without reference to Campbell.

4. Contemporary Research about Quasi-experimental Design

4.1 Statistics and Quasi-experimental Design

Although work on the statistical analysis of quasi-experimental designs deserves separate treatment, several contemporary developments deserve mention here. One is the work of statisticians such as Paul Holland, Paul Rosenbaum, and Donald Rubin on statistical models for quasi-experimental designs (e.g., Rosenbaum 1995). They emphasize the need to measure what would have happened to treatment participants without treatment (the counterfactual), and focus on statistics that can improve estimates of the counterfactual without randomization. A central method uses propensity scores, a predicted probability of group membership obtained from logistic regression of actual group membership on predictors of outcome or of how participants got into treatment. Matching, stratifying, or co-varying on the propensity score can balance nonequivalent groups on those predictors, but they cannot balance groups for unobserved variables, so hidden bias may remain. Hence these statisticians have developed sensitivity analyses to measure how much hidden bias would be necessary to change an effect in important ways. Both propensity scores and sensitivity analysis are promising developments warranting wider exploration in quasi-experimental designs.

A second statistical development has been pursued mostly by economists, especially James Heckman and his colleagues, called selection bias modeling (Winship and Mare 1992). The aim is to remove hidden bias in effect estimates from quasi-experiments by modeling the selection process. In principle the statistical models are exciting, but in practice they have been less successful. A series of studies in the 1980s and 1990s found that effect estimates from selection bias models did not match results from randomized experiments. Economists responded with various adjustments to these models, and proposed tests for their appropriate application, but so far results remain discouraging. Most recently, some economists have improved results by combining selection bias models with propensity scores. Although selection bias models cannot yet be recommended for widespread adoption, this topic continues to develop rapidly and serious scholars must attend to it. For example, other economists have developed useful econometrically based sensitivity analyses. Along with the incorporation of propensity scores, this may promise a future convergence in statistical and econometric literatures.

A third development is the use of structural equation modeling (SEM) to study causal relationships in quasi-experiments, but this effort has also been only partly successful (Bollen 1989). The capacity of SEM to model latent variables can sometimes reduce problems of bias caused by unreliability of measurement, but its

capacity to generate unbiased effect estimates is hamstrung by the same lack of knowledge of selection that thwarts selection bias models.

4.2 The Empirical Program of Quasi-experimental Design

Many features of quasi-experimentation pertain to matters of empirical fact that cannot be resolved by statistical theory or logical analysis. For these features, a theory of quasi-experimental design benefits from empirical research about these facts. Shadish (2000) has presented an extended discussion of what such an empirical program of quasi-experimentation might look like, including studies of questions like the following:

(a) Can quasi-experiments yield accurate effect estimates, and if so, under what conditions?

(b) Which threats to validity actually occur in practice (e.g., pretest sensitization, experimenter expectancy effects), and if so, under what conditions?

(c) Do the design features in Table 1 improve causal inference when applied to quasi-experimental design, and if so, under what conditions?

Some of this empirical research has already been conducted (see Shadish 2000, for examples). The methodologies used to investigate such questions are eclectic, including case studies, surveys, literature reviews (quantitative and qualitative), and experiments themselves. Until very recently, however, these studies have generally not been systematically used to critique and improve quasi-experimental theory.

5. Conclusion

Three important factors have converged at the end of the twentieth century to create the conditions under which the development of better quasi-experimentation may be possible. First, over 30 years of practical experience with quasi-experimental designs have provided a database from which we can conduct empirical studies of the theory. Second, after decades of focus on randomized designs, statisticians and economists have turned their attention to improving quasi-experimental designs. Third, the computer revolution provided both theorists and practitioners with increased capacity to invent and use more sophisticated and computationally intense methods for improving quasi-experiments. Each in their own way, these three factors have taken us several steps closer to answering that great unanswered question with which this article began.

See also: Campbell, Donald Thomas (1916–96); Comparative Studies: Method and Design; Experimental Design: Overview; Experimental Design:

Randomization and Social Experiments; Experimentation in Psychology, History of; Panel Surveys: Uses and Applications

Bibliography

- Bollen K A 1989 *Structural Equations with Latent Variables*. Wiley, New York
- Campbell D T, Stanley J C C 1963/1966 *Experimental and Quasi-experimental Designs for Research*. Rand-McNally, Chicago
- Cochran W G 1965 The planning of observational studies in human populations. *Journal of the Royal Statistical Society, Series A General* **128**: 234–66
- Cook T D, Campbell D T 1979 *Quasi-experimentation: Design and Analysis Issues for Field Settings*. Rand-McNally, Chicago
- Heckman J J 1979 Sample selection bias as a specification error. *Econometrica* **47**: 153–61
- Hill A B 1953 Observation and experiment. *The New England Journal of Medicine* **248**: 995–1001
- Rosenbaum P R 1995 *Observational Studies*. Springer-Verlag, New York
- Schlesselman J J 1982 *Case-Control Studies: Design, Conduct, Analysis*. Oxford University Press, New York
- Shadish W R 2000 The empirical program of quasi-experimentation. In: Bickman L (ed.) *Validity and Social Experimentation: Donald Campbell's Legacy*. Sage Publications, Thousand Oaks, CA
- Shadish W R, Cook T D, Campbell D T in press *Experimental and Quasi-experimental Designs for Generalized Causal Inference*. Houghton-Mifflin, Boston
- Triplett N 1898 The dynamogenic factors in pacemaking and competition. *American Journal of Psychology* **9**: 507–33
- Winship C, Mare R D 1992 Models for sample selection bias. *Annual Review of Sociology* **18**: 327–50

W. R. Shadish

Queer Theory

'Queer theory' is a notoriously unstable phrase, and one much in contention. As a new theoretical movement with equally new political counterparts, it is in constant flux and development, and is characterized more by what it challenges and contests than by what it offers in the shape of a unified social theory. Drawing on the work of theorists such as Eve Sedgwick and Judith Butler, queer theory 'describes those gestures or analytical models which dramatise incoherencies in the allegedly stable relations between chromosomal sex, gender and sexual desire (Jagose 1996). In this sense, queer theory is a challenge to the 'obvious categories (man, woman, *latina*, jew, butch, *femme*), oppositions (man vs. woman, heterosexual vs. homosexual), or equations (gender = sex) upon which conventional notions of sexuality and identity rely' (Hennessy 1993). Queer theory argues instead that sexual desire and sexual practices are not reducible or

explicable solely in terms of identity categories, such as gender, race, class, or sexual orientation. It is radically anti-essentialist, in that it challenges a notion of homosexuality as intrinsic, fixed, innate, and universally present across time and space.

Queer theorists reject any mode of thought that relies on a conception of identity as unified and self-evident (e.g., I have sex with people of the opposite sex, therefore I must be heterosexual), and instead demonstrate that desires, sexual practices, and gendered identities are performances and enactments, rather than expressions of 'true' subjectivity. Heterosexuality is therefore challenged by queer theory not simply as a 'hegemonic' mode of identity, but as a false claim to unity and coherence that is constantly undermined by the incoherencies of sex and gender, incoherencies that the queer analytic hopes to expose and celebrate.

1. Intellectual Origins

In the broadest sense, queer theory emerged in what might be called the postmodern moment, when intellectual unease with unitary and cohesive frameworks of knowing reached a fever pitch. While impossible to summarize here, queer theory's allegiance to postmodern and/or poststructural modes of thought can be traced in its challenge to the notion of unitary identity (as in 'gay' or 'straight'), its refusal to understand sexuality through a singular and unified lens (homosexual desire, feminist theory, gender), a rejection of binary models (gay/straight, man/woman, biological/social, real/constructed), and a more generic critique of identity-based theories and politics that, according to poststructuralist accounts, invariably reproduce the very conditions of repression they desire to challenge. For example, the term 'gay' or 'homosexual' might be critiqued as a (fictional) category that shores up the binary opposition between 'gay' and 'straight' that is itself part of the repressive logic of identity. To claim 'gayness' is therefore not simply or solely an act of self-revelation but is also a way of corraling sexuality within the framework of a category that only appears coherent but that, when opened up, reveals its instability (e.g., Am I still gay if I sleep with a person of the opposite sex? Or if I sleep with those of the same sex but only in certain conditions and in certain ways? Or if my self-understanding is of myself as 'straight?'). Queer theory, in that sense, has developed within and through the deconstructive impulse of poststructuralism, challenging assertions of unitary identity and necessary linkages (between, say, sexual desire and gender orientation) and arguing instead for a more provisional, contingent, and fluid conception of the 'queer' in contemporary culture.

While queer theory emerges as coterminous with postmodern impulses, it also traces its intellectual origins in lesbian/gay studies and feminist theory even

as it challenges those very paradigms. While for some the term queer theory is simply a synonym for 'lesbian and gay studies' or even 'feminist theory,' for many it signifies a real shift away from what is understood to be the 'essentialist' underpinnings of those more traditionally defined fields. In that sense, queer theory positions itself as a theoretical framework that rejects the typical binarisms (man/woman, gay/straight) in favor of a more universalist and overarching critique of the whole edifice of 'normalization,' a set of rules, regulations, and normative commands that create the rubric of 'good' vs. 'bad' sexualities. Queer theory takes issue with the ways in which feminist theory has tended to analyze sexuality within the framework of gender, and contests the ways in which 'gay' assumes a lining up of identity with acts, behavior with being. Drawing on the pioneering work of postmodern thinkers such as Michel Foucault and Eve Sedgwick, as well as hard-working historians such as Jonathan Ned Katz, queer theory joins in the disarticulating of sexual identities, sexual practices, and gender that characterizes contemporary postmodern theorizing. Indeed, queer theory takes all the mainstreaming, normalizing impulses to task—be it the scientific move of disciplines or the ethnic minority model of mainstream gay organizations. In general, then, queer theory can be understood as part of the larger intellectual ferment of the late twentieth century that has challenged the very core concepts of Western humanism: identity, science, truth, and authenticity.

2. Social Context

If queer theory develops out of the intellectual movements of poststructuralism, feminist theory, and lesbian and gay studies, it also emerged in the context of recent political alliances and strategies. As the gay community faced AIDS, and governmental reluctance to address this disease was slow and insufficient, gay activist groups developed new, 'in your face' strategies, exemplified by the work of groups such as ACT-UP and Queer Nation. While not always or necessarily drawing on queer theory explicitly, these groups helped to define the terrain of the intellectual discourse by emphasizing a nonnormative, anti-assimilationist alternative to traditional identity politics and minority group formations. As Warner (1991) argues,

The insistence on 'queer'—a term defined against 'normal' and generated precisely in the context of terror—has the effect of pointing out a wide field of normalization, rather than simple intolerance, as the site of violence. Its brilliance as a naming strategy lies in combining resistance on the broad social terrain of the normal with more specific resistance on the terrains of phobia and queer-bashing, on one hand, or of pleasure on the other. 'Queer,' therefore, also suggests the difficulty in defining the population whose interests are at stake in queer politics. And as a partial replacement for 'lesbian

and gay' it attempts partially to separate questions of sexuality from those of gender.

Queer theory therefore has a strong affinity with radical political praxis and thus brings to the academy an intellectual movement informed substantively by political crises and interventions.

3. Core Challenges

Queer theory takes as its primary theoretical and analytic focus the destabilization of 'gay and lesbian' and the disruption of the category of gender, particularly in terms of how gendered analysis is often understood to be the appropriate intellectual framework for the study of sexuality. Queer theory critiques the conceptualization of gender as a self-evident and unified category, and argues that this framework leads inexorably to another kind of polarizing and oppressing logic in which categories of identity become ways to 'police' populations whose behaviors, desires, and modes of living seem somehow discordant with the identity in question.

Because queer theory understands sexuality as fluid, mobile, and plural, the concept of 'performance' has been a central one, insisting on the situational and provisional nature of identity. Thus, 'gay' or 'lesbian' or even 'woman' are not understood as self-evident and essential categories, but rather are understood to be enactments of subjectivity. In this queer reading of identity, gender is understood as process, acquisition, enactment, creation. Thus, queer theorists often take as their analytic site the spaces where these performances occur and are highlighted—cross-dressing and drag, for example. Rather than assert a universal humanism (gays are essentially the same as heterosexuals, women are 'as good as' men), queer theory embraces the vilified, denigrated, ostracized identities (much as it embraces the epithet 'queer') and sees in these hybrid spaces (the transsexual, the drag queen) possibilities for a more thorough-going critique of normative culture and its repressive apparatuses.

Therefore, for queer theorists, the problem is not primarily one of bigotry, or lack of tolerance, or even homophobia *per se*, but rather the edifice of categorization and normalization that forces even those considered 'deviant' to tie behaviors and identities to narrow lines of power, lines that always imply repression and limitation. Queer theory celebrates sexual 'otherness' outside of identity categories (even outside those putatively radical categories such as 'gay' and 'lesbian') because it believes fervently that the power of domination is to be found primarily in the production and enforcement of identity itself. Identities, even alternative ones, carry with them normative assumptions about correct and incorrect ways of living and loving. And while identities give meaning to our lives, they also curtail possibilities by promoting a

false sense of unity (this is what a homosexual is) to a human life that is invariably riven with contradictions and inconsistencies. Thus queer theorists search for and celebrate moments of sexual subjectivity that appear outside of, or resistant to, firm categories of unified identity.

4. *Queer Theory and the Academy*

Queer theory cannot, by any stretch of the imagination, be located in any singular field or characterized by any unified methodology. While it is undoubtedly true that queer theory has found a more hospitable home in the humanities (particularly in departments of comparative literature and English), increasing numbers of social scientists (particularly those working within the broad framework of cultural studies) are engaging with queer theory and adapting it to their own research agendas. However, part of the uniqueness of queer theory comes in its deployment of multiple strategies to analyze a myriad of social and cultural phenomenon. In that sense, queer theory itself becomes a model of a new sort of trans- or even antidisciplinarity in the academic world. That is why most of the interventions in queer theory emerge from specific and located microstudies, rather than from large-scale, grand theoretical ventures. In that way, queer theory evinces much similarity with the growing 'field' of cultural studies, itself marked by a traversing of typical academic and disciplinary boundaries. And, given its emphasis on performativity and on non-normativity, queer theory presents a challenge to academic discourses that seek assimilation within mainstream intellectual life. It is precisely that mainstreaming, that incorporation, that queer theory sets its sights on.

5. *Some Critical Notes*

Queer theory has been critiqued from any number of locations, but none so persuasively as feminist theorists and theorists of color. Both the feminist and race critiques of queer focus on the ways in which queer theory often assumes implicitly a white, gay maleness as the prototype of queer subjectivity. In addition, both feminists and radical theorists of color have been suspicious of any theoretical framework that vitiates the need for an aggressive identity politics (a politics based on a notion of a unified and 'real' identity—gay, woman, etc.), given that identity politics have been for many the route to social change. For many feminist theorists, queer theory is seen as naive in its claim to transcend gender as a meaningful category of analysis. For them, it is seen as having eviscerated the centrality of gender to social identity, social power, sexuality and sexual, identity by dethroning gender critique as the cornerstone of the critique of both homophobia and

sexism. Queer theory, it is argued, sees patriarchal sexuality and sexist culture as almost tangential to the analysis of the repression of sexualities, while feminists have long argued that, for example, homophobia emerges from the larger and more pervasive rubric of sexist and patriarchal ideology and practice.

The point that gender is not the only significant marker of difference is an important one, and one that deserves development and reiteration. The regimes of sexuality and gender are not identical, either historically or theoretically, but many remain skeptical of their premature separation. In much of queer writing, it appears that feminism emerges as the enemy, the agent of a repressive regime of identity categorization, rather than heterosexism and patriarchal violence. Feminism is often depicted in queer theoretical writings as an outdated and essentialist relic of a humanist past, and queer theory, in its transcendence of gender, race, and class, as the postmodern mode of the future. Martin (1994) has argued that, for Sedgwick and others, race and gender often assume a fixity, a stability, a ground, whereas sexuality (typically thematized as male) becomes the 'means of crossing' and the figure of mobility. In the process of making the female body the 'drag' on the (male) play of sexuality, 'the female body appears to become its own trap, and the operations of misogyny disappear from view.'

Queer theory's relation to the politics and theorizing of racialized identities is no less fraught with difficulties than its relation to feminism and feminist identities. The generalizing tendencies of queerness (queer as everything not normatively heterosexual) has the potential to evacuate the specificities of racialized identities in favor of a queer universalism that claims multiracial status without ever developing seriously a race-based critique of heteronormativity, and without ever considering seriously the ways in which heterosexual dominance depends on certain racial ideologies. Queer can 'de-race' the homosexual of color in much the same way 'old-time' gay studies often has, effectively erasing the specificity of 'raced' gay existence under a queer rubric in which whiteness is not problematized. Dhairyam (1994) critiques the implicit whiteness of queerness while still attempting to activate the category 'queer women of color': 'Queer theory comes increasingly to be reckoned with as critical discourse, but concomitantly writes a queer whiteness over raced queerness, it domesticates race in its elaboration of sexual difference.' Anzaldúa (1991) also accuses white academics of co-opting the term 'queer' and using it to construct 'a false unifying umbrella which all "queers" of all races, ethnicities and classes are shoved under.'

Many have argued that queer theory lives in a realm of almost utopian idealism and abstraction, where the material conditions of marginalized populations are 'performed' away by theoretical sleights of hand, as if oppression is solely a matter of sexuality, and its representation and regulation. This inattention to

material social relations (commodification, the fluctuations of international capital, shifting forms of familial life, rise in antigay activism, regressive social legislation, increasing disenfranchisement of people of color, etc.) and the academicism of much of queer writing might be problems for a lesbian/gay praxis that is both class and race conscious.

6. Future Directions

Many self-defined queer theorists have taken these criticisms to heart and are attempting to reframe queer theory to pay more attention to structural social conditions, bodily realities, and the lived vicissitudes of race, class, and gender violence. Hennessy's (1994) piece is an exemplary attempt to hold on to the insights of queer theory while insisting that we pay attention to the material social conditions (such as class) that help to determine who can or cannot have access to the brave new queer world. She demands forcefully that queer theorists pay more attention to the processes of commodification and avoid valorizing a politics of the outrageous at the expense of attending to the realities of structured social relations, relations not reducible to the discursive or cultural, although certainly not determinative of them either. Queer may hold out some possibilities for a politics and a theory that challenges the fixity and clarity of identity, that speaks to the fractured (non)self of postmodern subjectivity. The queer challenge to the notion of sexual identity as monolithic, obvious, and dichotomous is a healthy corrective to our vexing inability to see beyond the limitations of the homo/hetero opposition. In addition, the openness of the term 'queer' provides the possibility of theorizing 'beyond the hyphen,' beyond the additive models (race, class, gender, sexual orientation = oppressed identity) that have so often seemed to set up new hierarchies, or instead retreated into an empty recitation of 'difference.' In this sense, queer theory can make theoretical and political space for more substantive notions of multiplicity and the intersecting and often contradictory lines of desire, identity, and sexuality.

See also: Androgyny; Feminist Ethics; Feminist Movements; Feminist Theology; Feminist Theory; Feminist Theory: Postmodern; Feminist Theory: Radical Lesbian; Foucault, Michel (1926–84); Gay, Lesbian, and Bisexual Youth; Gay/Lesbian Movements; Gender and Feminist Studies; Gender, Class, Race, and Ethnicity, Social Construction of; Gender Ideology: Cross-cultural Aspects; Heterosexism and Homophobia; Homosexuality and Psychiatry; Lesbians: Historical Perspectives; Lesbians: Social and Economic Situation; Male Dominance; Masculinities and Femininities; Nontraditional Families and Child Development; Rationality and Feminist Thought; Regulation: Sexual Behavior; Sexual Behavior and

Maternal Functions, Neurobiology of; Sexual Behavior: Sociological Perspective; Sexual Orientation and the Law; Sexual Orientation: Biological Influences; Sexual Preference: Genetic Aspects; Sexual Risk Behaviors; Sexuality and Gender; Sexually Transmitted Diseases: Psychosocial Aspects; Transsexuality, Transvestism, and Transgender

Bibliography

- Anzaldúa G 1991 To(o) queer the writer—loca, escritora y chicana. In: Warland B (ed.) *Inversions: Writing by Dykes, Queers, and Lesbians*. Press Gang Publishers, Vancouver, BC, pp. 249–63
- Beemyn B, Eliason M (eds.) 1996 *Queer Studies: A Lesbian, Gay, Bisexual, and Transgender Anthology*. New York University Press, New York
- Bergman D 1993 Disciplining gay and lesbian studies. Andrew R. Hilén Lecture at the University of Washington, Seattle, WA
- Butler J 1990 *Gender Trouble: Feminism and the Subversion of Identity*. Routledge, New York
- Butler J 1991 Imitation and gender insubordination. In: Fuss D (ed.) *Inside/Out*. Routledge, New York, pp. 13–31
- Butler J 1993a Critically queer. *GLQ: A Journal of Lesbian and Gay Studies* 1: 17–32
- Butler J 1993b *Bodies That Matter*. Routledge, New York
- de Lauretis T 1991 Queer theory: Lesbian and gay sexualities, an introduction. *Differences* 3(2): iii–xviii
- D'Emilio J 1983 *Sexual Politics, Sexual Communities: The Making of a Homosexual Minority in the United States, 1940–1970*. University of Chicago Press, Chicago
- Dhairiyam S 1994 Racing the lesbian, dodging white critics. In: Doan L (ed.) *The Lesbian Postmodern*. Columbia University Press, New York pp. 25–46
- Doty A 1993 *Making Things Perfectly Queer: Interpreting Mass Culture*. University of Minnesota Press, Minneapolis, MN
- Duggan L 1992 Making it perfectly queer. *Socialist Review* 22(1): 11–31
- Duggan L 1994 Queering the state. *Social Text* 39: 1–14
- Editors 1990/1993 Introduction. *Radical America* 24(4): 2–13
- Fuss D (ed.) 1991 *Inside/Out: Lesbian Theories, Gay Theories*. Routledge, New York
- Harper P B, White E F, Cerullo M 1990/1993 Multi/Queer/Culture. *Radical America* 24(4): 27–37
- Hennessy R 1993 Queer theory: A review of the *Differences* special issue and Wittig's *The Straight Mind*. *Signs* 18(4): 964–73
- Hennessy R 1995 Queer visibility in commodity culture. *Cultural Critique* 29: 31–76
- Jagose A 1996 *Queer Theory: An Introduction*. New York University Press, New York
- Kader C, Piontek T 1992 Introduction. *Discourse* 15(1): 5–10
- Katz J N 1995 *The Invention of Heterosexuality*. Dutton, New York
- Kitzinger C, Wilkinson S 1994 Virgins and queers: Rehabilitating heterosexuality? *Gender & Society* 8(3): 444–63
- Martin B 1993 Lesbian identity and autobiographical difference(s). In: Abelow H, Barale M A, Halperin D M (eds.) *The Lesbian and Gay Studies Reader*. Routledge, New York pp. 274–93
- Martin B 1994 Sexualities without genders and other queer Utopias and gay and lesbian studies. *Diacritics* 24(2–3): 104–21

- Moraga C, Hollibaugh A 1983 What we're rollin around in bed with: Sexual silences in feminism. In: Snitow A, Stansell C, Thompson S (eds.) *Powers of Desire: The Politics of Sexuality*. Monthly Review Press, New York, pp. 394–405
- Morton D (ed.) 1996 *The Material Queer: A LesBiGay Cultural Studies Reader*. Westview Press, Boulder, CO
- Nunokawa J 1992 Identity crisis: Queer politics in the age of possibilities. Interview by Alisa Solomon. *Village Voice* 30 (June): 27
- Phelan S 1993 (Be)Coming out: Lesbian identity and politics. *Signs* 18(4): 773–90
- Reid-Pharr R 1990/1993 The spectacle of blackness. *Radical America* 24(4): 57–65
- Rubin G S 1993 Thinking sex: notes for a radical theory of the politics of sexuality. In: Abelove H, Barale M A, Halperin D M (eds.) *The Lesbian and Gay Studies Reader*. Routledge, New York, pp. 3–44
- Sedgwick E K 1989 Across gender, across sexuality: Willa Cather and others. *South Atlantic Quarterly* 88(1): 53–72
- Sedgwick E K 1990 *Epistemology of the Closet*. University of California Press, Berkeley, CA
- Sedgwick E K 1991 How to bring your kids up gay. *Social Text* 29: 18–27
- Sedgwick E K 1993 *Tendencies*. Duke University Press, Durham, NC
- Seidman S 1994 Queer pedagogy/queer-ing sociology. *Critical Sociology* 20(3): 167–76
- Seidman S (ed.) 1996 *Queer Theory/Sociology*. Blackwell, Cambridge, MA
- Seidman S 1997 *Difference Troubles: Queering Social Theory and Sexual Politics*. Cambridge University Press, Cambridge, UK
- Smith B 1993 Queer politics: Where's the revolution? *The Nation* 5 (July): 12–16
- Stein A 1992 Sisters and queers: The decentering of lesbian feminism. *Socialist Review* 22(1): 33–55
- Walters S 1995 *Material Girls: Making Sense of Feminist Cultural Theory*. University of California Press, Berkeley, CA
- Walters S 1996 From here to queer: Radical feminism, post-modernism, and the lesbian menace (or, Why can't a woman be more like a fag?) *Signs* 21(4) (Summer): 830–69
- Warner M 1991 Fear of a queer planet. *Social Text* 9(4): 3–17
- Warner M 1992 From queer to eternity: An army of theorists cannot fail. *Voice Literary Supplement* June: 18–19
- Warner M (ed.) 1993 *Fear of a Queer Planet: Queer Politics and Social Theory*. University of Minnesota Press, Minneapolis, MN
- Whisman V 1993 Identity crises: Who is a lesbian, anyway? In: Stein A (ed.) *Sisters, Sexperts, Queers*. Plume, New York, pp. 47–60

S. D. Walters

Questionnaires: Assessing Subjective Expectations

Questionnaires are used to elicit from individuals a broad array of objective information about their experiences and environments, as well as subjective information about their thoughts and perceptions. An important objective of social and behavioral science research is to make questionnaires more informative.

In practical terms, how should the questionnaire be administered and what should be its content? Aspects of questionnaire administration include when and where the data are collected, who does the data collection, and the mode of administration (e.g., face-to-face, telephone, or Internet). Aspects of content begin with the basic matter of determining the purposes of the questionnaire and end with specific issues of item wording, the sequencing of questions, and the formatting of responses.

This article focuses on a single topic about questionnaires—how to elicit respondents' subjective expectations of future events. For example, demographers ask women to predict their future fertility and political scientists ask citizens to predict their voting behavior in upcoming elections. Researchers use the responses to expectation questions to predict fertility, voting behavior, and other future events. They also use the responses to assess how individuals' perceptions of the future are related to the decisions they make today. Subjective expectations of future events seem to be measured with increasing frequency. Yet the social and behavioral sciences have not formed a common view on how subjective expectations should be conceptualized and measured. A particularly intriguing controversy, described in this article, concerns the relative merits of verbal and probabilistic forms of questioning.

1. Subjective Concepts and Measures

A US National Research Council (NRC) panel study on *Surveying Subjective Phenomena* has distinguished subjective from objective phenomena as follows (Turner and Martin 1984, p. 8):

Subjective phenomena are those that, in principle, can be directly known, if at all, only by persons themselves, although a person's intimate associates or a skilled observer may be able to surmise from indirect evidence what is going on 'inside.' Objective phenomena are those that can be known by evidence that is, in principle, directly accessible to an external observer.

The NRC panel acknowledged that the sharp dichotomy just drawn is not always tenable, but the essential idea clearly has content: subjective phenomena are not amenable to external verification.

The social and behavioral sciences have a common interest in understanding subjective phenomena, but the various disciplines have not been able to agree on concepts or measures. The conceptual problem is to characterize what is going on 'inside.' A reader of social and behavioral science research encounters an enormous variety of subjective concepts in use, with concepts that are central to one discipline often playing no role in others. A modest sense of this variety can be obtained by scanning the introductory chapter of Turner and Martin (1984), which mentions affect, aspirations, attitudes, disappointments, dispositions,

emotions, frustrations, fulfillment, happiness, hopes, intentions, motives, opinions, plans, preferences, quality of life, satisfaction, and values.

The measurement problem is to obtain empirical evidence on what is going on 'inside.' A reader of empirical research in sociology, social psychology, or political science finds much analysis of subjective data elicited from survey respondents through questionnaires. Yet a reader of anthropology or economics finds little reliance on questionnaire responses. Anthropologists, who tend to recoil from the tightly scripted structure of questionnaires, have preferred to elicit information through conversations. Economists, who tend to be skeptical of subjective data of any kind, have sought to infer subjective phenomena from observations of behavior and experiences.

To illuminate the diversity of perspectives on subjective phenomena, this article examines how different disciplines have used questionnaires to study subjective expectations about potential future events. How persons perceive the future is a shared concern of all social and behavioral science disciplines. Yet there is a wide spectrum of views on how expectations should be conceptualized and measured.

Section 2 examines the prevailing practice in economics of inferring expectations from realization data, and recalls early efforts to collect data on economic expectations through questionnaires. Section 3 describes qualitative attitudinal research as practiced by sociologists and social psychologists. Section 4 describes the growing body of research in economics and cognitive psychology eliciting subjective probabilities of future events. Section 5 considers the relative merits of qualitative and probabilistic modes of questioning. Section 6 poses directions for future research on subjective expectations. See Dominitz and Manski (1999) for a more detailed exposition of the subjects discussed below.

2. Inference on Expectations in Economics

Early in their careers, economics students are taught that a good economist believes only what people do, not what they say. Skeptical of subjective data, economists have sought to infer expectations of future events by combining data on past experiences with assumptions about processes of expectation formation. This is a daunting task. To infer expectations from data on past experiences, economists must speculate about what information persons possess and how they use the available information to form their expectations.

Consider, for example, the matter of income expectations. Economists often assume that persons use their own past incomes to forecast their future incomes. Perhaps so, but how do persons form expectations of future income conditional on past income? Economists often assume that persons have *rational expectations*; that is, they know the actual

stochastic process generating their income streams. Perhaps so, but what is this stochastic process? The practice has been to specify the income process up to some parameters and use available data on income realizations to estimate the parameters. Authors differ, however, in their specifications. So the literature displays competing models of expectations formation, with no clear way of distinguishing among them. See Dominitz and Manski (1997a).

If inference on expectations from data on experiences is so challenging, why do economists not collect and analyze data on expectations? The answer lies at least partly in the history of a scientific controversy that began in the 1940s and persisted until the 1960s. In the 1940s, the US Federal Reserve Board began to fund an annual Survey of Consumer Finances, conducted by the University of Michigan Survey Research Center (SRC), that elicited qualitative assessments of expected household finances. A typical question took this form:

How about a year from now—do you think you people will be making more money or less money than you are now, or what do you expect?

The usefulness of responses to such qualitative questions was controversial and the Federal Reserve Board appointed a committee to assess their value. The Federal Reserve Consultant Committee on Consumer Survey Statistics (1955) issued findings that questioned the predictive power of the SRC expectations data. The negative findings of the Committee were challenged by SRC researchers, notably George Katona, a leading proponent of research on consumer attitudes and expectations (Katona 1957). A contentious conference on expectations data followed (National Bureau of Economic Research 1960). Then Juster (1964) reported an intensive study, drawing largely negative conclusions, of the usefulness of qualitative expectations data in predicting individual behavior.

The Federal Reserve–SRC controversy only concerned the usefulness of a certain type of qualitative expectations data in predicting consumer behavior. Academic economists, however, appear to have drawn the broader conclusion that all expectations data are suspect. Economists also failed to take note when Juster (1966) proposed elicitation of probabilistic expectations of consumer durable purchases and reported favorable empirical evidence. Exploration of this idea had to wait a quarter century, as described in Sect. 4.

3. Qualitative Attitudinal Research

The Federal Reserve–SRC controversy may have soured most economists on the usefulness of qualitative expectations data, but the collection and analysis of such data have long been important activities in

other behavioral and social science disciplines. Many questions in common use have a structure similar to that of this General Social Survey (GSS) question eliciting expectations of job loss (Davis and Smith 1994):

Thinking about the next twelve months, how likely do you think it is that you will lose your job or be laid off—very likely, fairly likely, not too likely, or not at all likely?

This GSS question illustrates well a basic persistent problem that researchers face in interpreting qualitative expectations data—assessment of the inter- and intrapersonal comparability of responses. Respondents are often asked to report whether a specified event is ‘expected,’ ‘likely,’ or ‘probable’ to occur. Sometimes they are asked to report their strength of belief in the event on a numerical scale with undefined units. Do different respondents interpret these verbal phrases and numerical scales in the same way? Does a given respondent interpret them in the same way when asked about different events? Cognitive research does not give reason to think that responses should be or are comparable. Indeed, the available empirical evidence indicates that interpretations of qualitative expectations questions vary substantially between persons (see Sect. 5).

Concern about interpersonal comparability comes into even sharper focus when considering the events about which respondents are asked. Attitudinal researchers often ask respondents about vaguely defined future outcomes, as in this SRC question (Curtin 1982):

Now looking ahead—do you think that a year from now you (and your family living there) will be better off financially, or worse off, or just about the same as now?

How should a respondent interpret the phrase ‘better off financially?’ Do different respondents interpret it the same way? The data provide no way of knowing.

A second persistent problem in qualitative attitudinal research is that the coarseness of the response options limits the information contained in the responses. Consider, for example, the fertility question asked female respondents in the annual June Supplement to the Current Population Survey of the US Bureau of the Census:

Looking ahead, do you expect to have any (more) children?
Yes No Uncertain

The three response options can express little of the richness of the uncertainty that women may perceive about their future childbearing. Savage (1971, p. 795) called attention to the problem when he observed that ‘Yes, No, or Maybe is Not Enough.’ Juster (1966) hypothesized that a person facing a Yes/No expectations question responds as would a statistician asked to make a best point prediction of a future random

event. Manski (1990) formalized this hypothesis and derived an upper bound on the information about expectations a researcher can extract from responses to Yes/No questions.

4. Elicitation of Probabilistic Expectations

A growing body of research in economics and cognitive psychology elicits expectations in the form of subjective probabilities. For example, the Survey of Economic Expectations uses this question to elicit expectations of job loss (Dominitz and Manski 1997b):

What do you think is the percent chance that you will lose your job during the next 12 months?

The National Longitudinal Study of Youth: 1997 Cohort of the US Bureau of Labor Statistics elicits the fertility expectations of teenage [girls/boys] as follows:

What is the percent chance that you will become the [mom/dad] of a baby sometime between now and when you turn 20?

Probabilistic questions of similar structure are posed in other national surveys, such as the Health and Retirement Study (Juster and Suzman 1995) and the Bank of Italy’s Survey of Household Income and Wealth, as well as in small-scale surveys such as the one described in Quadrel et al. (1993).

The idea that probabilistic elicitation of expectations might improve on the traditional qualitative approaches of attitudinal research developed independently in economics and in cognitive psychology. In economics, the idea appears to have originated with Juster (1966), who recommended elicitation of purchase probabilities for consumer durable goods. Market researchers were immediately attracted to Juster’s proposal, and elicitation of purchase probabilities has since become a common practice in market research (e.g., Jamieson and Bass 1989). However, the idea that expectations might be elicited probabilistically from survey respondents did not draw serious attention within economics until the early 1990s, when economists began to re-assess the discipline’s self-imposed prohibition on the collection and analysis of subjective data.

Elicitation of subjective probabilities has a more continuous history of research in cognitive psychology. Psychologists have long conducted calibration studies assessing the accuracy of elicited expectations reports when the outcome is known. Lichtenstein et al. (1982) reviewed findings from calibration studies dating back to 1906. McClelland and Bolger (1994) updated the review with findings from 1980 through 1994. Whereas the older studies mostly examined the accuracy of experts (e.g., weather forecasters’ reported probabilities of precipitation), the more recent re-

search typically analyzes the expectations of non-experts, especially students situated in a cognitive laboratory.

5. Qualitative vs. Probabilistic Expectations

As this article is written, a consensus has yet to emerge on the relative merits in practice of qualitative and probabilistic elicitation of subjective expectations. However, there is reasonably widespread agreement that elicitation of probabilistic expectations has several *a priori* desirable features. Perhaps the most basic attraction is that probability provides a well-defined absolute numerical scale for responses. Hence there is some reason to think that responses may be inter- and intrapersonally comparable.

A second attraction is that some empirical assessment of the internal consistency and external accuracy of respondents' expectations is possible. A researcher can use the algebra of probability (Bayes Theorem, the law of total probability, etc.) to examine the internal consistency of a respondent's elicited expectations about different events. In those cases where probability has a frequentist interpretation, a researcher can compare elicited subjective probabilities with known event frequencies, as is done in calibration studies, and reach conclusions about the correspondence between subjective beliefs and frequentist realities.

A third consideration is the usefulness of elicited expectations in predicting prospective outcomes. Suppose that respondents have reasonably accurate expectations about the likelihood of future events. Then, as argued by Juster (1966), Savage (1971), and Manski (1990), numerical responses to probability questions should have more predictive power than categorical responses to qualitative expectations questions.

These *a priori* desirable features of probabilistic expectations data must be weighed against a set of unresolved issues. Perhaps the most basic concern is that there is no direct way to assess how well elicited subjective probabilities reflect respondents' thinking. Of course this concern is not specific to elicitation of probabilities. Interpretation of responses to subjective questions of any type runs up against the generic problem that a researcher cannot directly observe a respondent's thinking.

A specific concern in the elicitation of probabilities is that respondents may not think probabilistically about uncertain events. A large and diverse literature suggests that persons may think about uncertain events using less than the full structure of modern probability theory (e.g., Camerer and Weber 1992). If so, then questions eliciting precise probabilities force respondents to give point responses when only interval responses may be meaningful.

Within cognitive psychology, there has long been controversy about the way in which humans internally

represent their beliefs, and their ability and willingness to express their internal beliefs as numerical probabilities. Ferrell and McGoey (1980) posed models in which individuals may have some difficulty expressing beliefs as numerical probabilities, but nevertheless concluded that elicitation of numerical subjective probabilities is feasible. However, Zimmer (1983), who argued that humans process information using verbal rather than numerical modes of thinking, concluded that expectations should be elicited in verbal rather than numerical forms.

Wallsten et al. (1993) has reported that a majority of respondents prefer to communicate their own beliefs verbally and to receive the beliefs of others in the form of numerical probabilities. This asymmetry is intriguing but only marginally relevant to the design of expectations questions. The relevant question is not what communication mode respondents prefer to use, but rather what modes they are willing and able to use. Wallsten et al. (1993) report that virtually all of their respondents were willing to communicate their beliefs numerically, should the situation warrant it. Certainly the accumulating experience of economists using questionnaires to elicit probabilistic expectations has been that respondents are willing and able to communicate their beliefs in numerical probabilistic terms (e.g., Dominitz and Manski 1997a, 1997b, Hurd and McGarry 1995).

A related theme of cognitive research has been to learn how individuals associate numerical probabilities and verbal phrases such as 'very likely,' 'probable,' or 'rarely' (e.g., Wallsten et al. 1986). A central finding is that individuals vary considerably in the numerical probabilities that they associate with given verbal phrases. There is, however, controversy about the proper conclusion to be drawn. Should one conclude that individuals differ in their interpretation of verbal phrases, in their interpretation of numerical probabilities, or in their interpretation of both? The empirical findings are compatible with all of these conclusions.

6. Expectations Formation and Behavior

Empirical research on subjective expectations is still in its infancy. Basic questions about the conceptualization and measurement of expectations are not yet resolved. When sufficiently satisfactory answers to these questions are achieved, research will be able to move to yet more challenging and interesting matters.

Future research should seek to understand how persons form their expectations. Attitudinal researchers have given little systematic thought to this basic matter. Researchers who conceptualize expectations in probabilistic terms have not been able to agree on how persons update their subjective probabilities in light of new evidence. Economists and statistical decision analysts have commonly assumed that persons apply Bayes rule to update their expectations.

Psychologists have reported intriguing empirical evidence to the contrary (e.g., Kahneman et al. 1982), but the practical import of this evidence remains unsettled. Many social scientists hypothesize that persons form expectations by observation of the behavior and experiences of others in their environments. However, we know precious little about the nature of social learning processes.

Research on subjective expectations should aim to understand how a person's perception of the future affects behavior. Economists suppose that youth make schooling choices based on their expectations of the returns to schooling relative to labor market activity. They also hypothesize that uncertainty about future income induces individuals to increase savings, thereby decreasing consumption. Criminologists suggest that perceived risks of crime victimization may cause individuals to take precautions in their daily activities. Public health researchers speculate on how perceptions of health risks affect smoking, drug use, and unprotected sexual behavior. To shed light on these and similar questions of practical import will require much more intensive collection and analysis of expectations and behavioral data than the social and behavioral sciences have yet undertaken.

See also: Classical (Psychometric) Test Theory; Test Administration: Methodology

Bibliography

- Camerer C, Weber M 1992 Recent developments in modeling preferences: Uncertainty and ambiguity. *Journal of Risk and Uncertainty* **5**: 325–70
- Curtin R 1982 Indicators of consumer behavior: The University of Michigan Surveys of Consumers. *Public Opinion Quarterly* **46**: 340–52
- Davis J, Smith T 1994 *The General Social Surveys, 1972–1994, Cumulative File*. National Opinion Research Center, Chicago
- Dominitz J, Manski C 1997a Using expectations data to study subjective income expectations. *Journal of the American Statistical Association* **92**: 855–67
- Dominitz J, Manski C 1997b Perceptions of economic insecurity: Evidence from the survey of economic expectations. *Public Opinion Quarterly* **61**: 261–87
- Dominitz J, Manski C 1999 The several cultures of research on subjective expectations. In: Smith J, Willis R (eds.) *Wealth, Work, and Health: Essays in Honor of F. Thomas Juster*. University of Michigan Press, Ann Arbor, MI
- Federal Reserve Consultant Committee on Consumer Survey Statistics 1955 Smithies Committee report in *Reports of the Federal Reserve Consultant Committees on Economic Statistics*. Hearings of the Subcommittee on Economic Statistics of the Joint Committee on the Economic Report, 84th US Congress
- Ferrell W, McGoe P 1980 A model of calibration for subjective probabilities. *Organizational Behavior and Human Performance* **26**: 32–53
- Fishbein M, Ajzen I 1975 *Belief, Attitude, Intention, and Behavior: An Introduction to Theory and Research*. Addison-Wesley, Reading, MA
- Hurd M, McGarry K 1995 Evaluation of the subjective probabilities of survival in the health and retirement study. *Journal of Human Resources* **30**: S268–92
- Jamieson L, Bass F 1989 Adjusting stated intentions measures to predict trial purchase of new products: A comparison of models and methods. *Journal of Marketing Research* **26**: 336–45
- Juster F T 1964 *Anticipations and Purchases*. Princeton University Press, Princeton, NJ
- Juster F T 1966 Consumer buying intentions and purchase probability: An experiment in survey design. *Journal of the American Statistical Association* **61**: 658–96
- Juster F T, Suzman R 1995 An overview of the health and retirement study. *Journal of Human Resources* **30**: S7–56
- Kahneman D, Slovic P, Tversky A (eds.) 1982 *Judgment Under Uncertainty: Heuristics and Biases*. Cambridge University Press, New York
- Katona G 1957 Federal Reserve Board Committee Reports on Consumer Expectations and Savings Statistics. *Review of Economics and Statistics* **39**: 40–6
- Lichtenstein S, Fischhoff B, Phillips L 1982 Calibration of probabilities: The state of the art to 1980. In: Kahneman D, Slovic P, Tversky A (eds.) *Judgment Under Uncertainty: Heuristics and Biases*. Cambridge University Press, New York
- Manski C 1990 The use of intentions data to predict behavior: A best case analysis. *Journal of the American Statistical Association* **85**: 934–40
- McClelland A, Bolger F 1994 The calibration of subjective probabilities: Theories and models 1980–94. In: Wright G, Ayton P (eds.) *Subjective Probability*. Wiley, New York
- National Bureau of Economic Research 1960 *The Quality and Economic Significance of Anticipations Data*. Princeton University Press, Princeton, NJ
- Quadrel M, Fischhoff B, Davis W 1993 Adolescent (in)vulnerability. *American Psychologist* **48**: 102–16
- Savage L 1971 Elicitation of personal probabilities and expectations. *Journal of the American Statistical Association* **66**: 783–801
- Schuman H, Johnson M 1976 Attitudes and behavior. *Annual Review of Sociology* **2**: 161–207
- Turner C, Martin E (eds.) 1984 *Surveying Subjective Phenomena*. Russell Sage Foundation, New York
- Wallsten T, Budescu D, Rapoport A, Zwick R, Forsyth B 1986 Measuring the vague meanings of probability terms. *Journal of Experimental Psychology: General* **115**: 348–65
- Wallsten T, Budescu D, Zwick R, Kemp S 1993 Preferences and reasons for communicating probabilistic information in verbal or numerical terms. *Bulletin of the Psychonomic Society* **31**: 135–8
- Zimmer A 1983 Verbal vs. numerical processing of subjective probabilities. In: Scholz R (ed.) *Decision Making Under Uncertainty*. North-Holland, Amsterdam

C. F. Manski

Questionnaires: Cognitive Approaches

Respondents' answers to standardized questions presented in self-administered questionnaires or face-to-face and telephone interviews are a major source of social science data. These answers, however, are

profoundly influenced by characteristics of the question asked, including the wording of the question, the format of the response alternatives, and the context in which it is presented. The cognitive and communicative processes underlying these influences are increasingly well understood and this article reviews their implications for question wording and questionnaire construction.

1. Respondents' Tasks

Answering a question requires that respondents (a) interpret the question to understand what is meant and (b) retrieve relevant information from memory to form an answer. In most cases, they cannot provide their answer in their own words but (c) need to map it onto a set of response alternatives provided by the researcher. Finally, (d) respondents may wish to edit their answer before they communicate it to the interviewer for reasons of social desirability and self-presentation. Respondents' performance at each of these steps is context dependent and profoundly influenced by characteristics of the questionnaire (Sudman et al. 1996, Chap. 3), including its administration mode (self-administered, or face-to-face, or telephone interview; Schwarz et al. 1991).

1.1 Question Comprehension

The key issue at the question comprehension stage is whether respondents' understanding of the question matches the meaning the researcher had in mind. Not surprisingly, methodology textbooks urge researchers to write simple questions and to avoid unfamiliar or ambiguous terms (see Sudman and Bradburn 1983, for good practical advice). However, understanding the literal meaning of a question is not sufficient to provide an informative answer. When asked, 'What have you done today?' respondents are likely to understand the words—yet, they still need to determine what kind of activities the researcher is interested in. Should they report, for example, that they took a shower, or not? Providing an informative answer requires inferences about the questioner's intention to determine the pragmatic meaning of the question.

To infer the pragmatic meaning, respondents draw on contextual information, including the content of adjacent questions and the nature of the response alternatives. Their use of this information is licensed by the tacit assumptions that govern the conduct of conversation in daily life, which respondents apply to the research interview (Clark and Schober 1992, Schwarz 1996). These pragmatic inferences have important implications for questionnaire construction.

1.1.1 Question context. Respondents interpret questions in the context of the overall interview and a

term like 'drugs' acquires different meanings when presented in survey pertaining to respondents' medical history rather than to crime in the neighborhood. Moreover, they assume that adjacent questions are meaningfully related to one another, unless otherwise indicated. They may infer, for example, that an ambiguous 'educational contribution' refers either to fellowships that students receive, or tuition they have to pay, depending on the content of adjacent questions. Finally, they hesitate to reiterate information they have already provided, consistent with conversational norms that discourage redundancy. For example, questions pertaining either to happiness with one's life or to life-satisfaction typically elicit highly similar answers, unless both questions are presented together. In the latter case, respondents assume the researcher wouldn't ask for the same thing twice and differentiate between both questions, resulting in less similar answers. In self-administered questionnaires, question interpretation is influenced by preceding as well as following questions, whereas only preceding questions can exert an influence in face-to-face and telephone interviews (see Clark and Schober 1992, Sudman et al. 1996, Schwarz 1996 for extensive reviews).

1.1.2 Response alternatives. Most questions are presented with a fixed set of response alternatives, which facilitates administration and data analysis. Far from being 'neutral' measurement devices, however, response alternatives influence question interpretation (Schwarz and Hippler 1991). First, consider the differences between open and closed response formats. When asked in an open response format 'What have you done today?' respondents are likely to omit activities that the researcher is obviously aware of (e.g., 'I gave a survey interview') or may take for granted anyway (e.g., 'I took a shower'). Yet, if these activities were included in a closed list of response alternatives, most respondents would endorse them. At the same time, such a list would reduce reports of activities that are not represented on the list, even if an 'other' option were provided, which respondents rarely use. Both of these question form effects reflect that response alternatives can clarify the intended meaning of a question by specifying what the researcher is interested in. When using a closed response format, it is therefore important to ensure that the list of response alternatives covers the full range of behaviors or opinions.

Similarly, suppose respondents are asked how frequently they felt 'really irritated' recently. To infer if 'really irritated' refers to major or to minor annoyances, respondents may draw on an apparently formal feature of the question, namely the numeric values of the frequency scale. When the scale presents low (high) frequency alternatives, respondents infer that the

researcher is interested in rare (frequent) events. Hence, they report on major annoyances (which are relatively rare) when given a low frequency scale, but on minor annoyances when given a high frequency scale. As a result, the same question elicits reports of substantively different experiences, depending on the frequency scale used (Schwarz 1996, Chap. 5, Schwarz and Hippler 1991).

In sum, apparently formal differences in the response alternatives can profoundly influence respondents' interpretation of the question. Far from reflecting superficial responding, such effects indicate that respondents do their best to make sense of the question asked by drawing on the full range of information available to them—including information the researcher never intended to provide. To safeguard against unintended question interpretations, researchers have developed a number of different pretesting procedures, reviewed in Sect. 3.

1.2 *Recalling or Computing a Judgment*

Once respondents determined what the researcher is interested in, they need to recall relevant information from memory. In some cases, respondents may have direct access to a previously formed relevant judgment that they can offer as an answer. In most cases, however, they will not find an appropriate answer readily stored in memory and will need to compute a judgment on the spot. The different judgmental processes pertaining to behavioral questions and opinion questions are discussed in Sect. 2 and *Attitude Measurement*, respectively.

1.3 *Formatting the Response*

Unless the question is asked in an open response format, respondents need to format their answer to fit the response alternatives provided by the researcher. Respondents observe these question constraints and avoid answers that are not explicitly offered, as already seen in Sect. 1.1.2. For example, many more respondents answer 'don't know,' or choose a middle alternative, when these options are explicitly offered than when they are not (see Schwarz and Hippler 1991, for a review). Under which conditions the addition or omission of these response options results in different substantive conclusions, and whether some respondents are more likely to be affected by the response alternatives than others, is the topic of some debate (see Krosnick and Fabrigar in press).

Response formatting processes are particularly relevant when rating scales are used (Sudman et al. 1996, Chap. 4). Specifically, respondents use the most extreme stimuli to anchor the endpoints of the scale. Hence, a given stimulus will be rated as less extreme if presented in the context of a more extreme one, than if

presented in the context of a less extreme one. In addition, if the number of stimuli to be rated is large, respondents attempt to use all categories of the rating scale about equally often. Hence, two similar stimuli may receive different ratings when only a few stimuli are presented, but may receive the same rating when a large number of stimuli has to be located along the scale. Accordingly, ratings of the same object cannot be directly compared when they were collected in different contexts, rendering comparisons over time or between studies difficult. The psychometric properties of rating scales have been the topic of extensive investigation (see Krosnick and Fabrigar in press). Seven point scales seem to be best in terms of reliability, percentage of undecided respondents, and respondents' ability to discriminate between scale points. Moreover, scales that provide verbal labels for each scale value seem more reliable than scales with verbal endpoints only.

Finally, respondents' choice of a given response alternative is often influenced by the order in which the alternatives are presented. Under the auditory presentation conditions of telephone interviews, a given alternative is more likely to be endorsed when it is presented last rather than first (referred to as a recency effect). Under the visual presentation conditions of a self-administered questionnaire, a given alternative is more likely to be endorsed when it is presented first rather than last (referred to as a primacy effect; see Sudman et al. 1996, Chap. 6, for a discussion of the underlying processes).

1.4 *Editing the Response*

Finally, respondents may want to edit their response before they communicate it, reflecting considerations of social desirability and self-presentation (see DeMaio 1984 for a review). Not surprisingly, the impact of self-presentation concerns is more pronounced in face-to-face interviews than in self-administered questionnaires. It is important to emphasize, however, that influences of social desirability are limited to potentially threatening questions and are typically modest in size. These influences can be reduced through techniques that ensure the confidentiality of the answer (see Sudman and Bradburn 1983 for detailed advice).

2. *Asking Questions about Behavior*

Most questions about respondents' behavior are frequency questions, pertaining, for example, to how often the respondent has bought something or has missed a day at work during some specified period of time. Researchers would ideally like respondents to identify the behavior of interest; to scan the reference period; to retrieve all instances that match the target

behavior; and to count these instances to determine the overall frequency of the behavior. This, however, is the route that respondents are least likely to take.

Except for rare and very important behaviors, respondents are unlikely to have detailed representations of numerous individual instances of a behavior stored in memory. Instead, the details of various instances of closely related behaviors blend into one global representation, rendering it difficult to distinguish and retrieve individual episodes. Accordingly, a 'recall-and-count' model does not capture how respondents answer questions about frequent behaviors or experiences. Rather, their answers are likely to be based on some fragmented recall and the application of inference rules to compute a frequency estimate. Accordingly, the best we can hope for under most conditions is a reasonable estimate, unless the behavior is rare and of considerable importance to respondents.

Conway (1990) provides a useful introduction to psychological research into autobiographical memory (i.e., individuals' memory for their experiences and behaviors) and Bradburn et al. (1987), Sudman et al. (1996, Chaps. 7–9), Tourangeau et al. (2000), and the contributions in Schwarz and Sudman (1994) review the implications for questionnaire construction. Section 2.1 summarizes what researchers can do to facilitate recall and Sect. 2.2 addresses respondents' estimation strategies.

2.1 Facilitating Recall

2.1.1 Recall cues and reference periods. If researchers are interested in obtaining reports that are based on recalled episodes, they may simplify respondents' task by providing appropriate recall cues and by restricting the recall task to a short and recent reference period. There are, however, important drawbacks to both of these strategies. Although the quality of recall will generally improve as the retrieval cues become more specific, respondents are likely to restrict their memory search to the particular cues presented to them, reflecting that the cues constrain the meaning of the question. As a result, respondents will omit instances that do not match the specific cues, resulting in under-reports if the list is not exhaustive. Moreover, using a short reference period may result in many 'zero' answers from respondents who rarely engage in the behavior, thus limiting later analyses to respondents with a high behavioral frequency. Finally, reference periods of different length may result in different question interpretations. Long reference periods ('How often have you been angry last year?') suggest that the researcher is interested in relatively rare events, whereas short reference periods ('How often have you been angry last week?') indicate an interest in frequent

events. Hence, respondents may report on substantively different behaviors, e.g., rare and intense vs. frequent but minor episodes of anger.

Not surprisingly, different recall cues are differentially effective. The date of an event is the poorest cue, whereas cues pertaining to what happened, where it happened, and who was involved have been found to be very effective. In addition, recall will improve when respondents are given sufficient time to search memory. Recalling specific events may take up to several seconds and repeated attempts to recall may result in the retrieval of additional material, even after a considerable number of previous trials. Unfortunately, respondents are unlikely to have sufficient time to engage in repeated retrieval attempts in most research situations, and may often not be motivated to do so even if they had the time. This is particularly crucial in the context of survey research, where the available time per question is usually less than one minute.

Moreover, the direction in which respondents search memory may influence the quality of recall. Specifically, better recall is achieved when respondents begin with the most recent occurrence of a behavior and work backward in time than when they begin at the beginning of the reference period. This presumably occurs because memory for recent occurrences is richer and the recalled instances may serve as cues for recalling previous ones. Given free choice, however, respondents tend to prefer the less efficient strategy of forward recall.

Yet, even under optimal conditions, respondents will frequently be unable to recall an event or some of its critical details, even if they believed they would 'certainly' remember it at the time it occurred. In general, the available evidence suggests that respondents are likely to under-report behaviors and events, which has led many researchers to assume that higher reports of mundane behaviors are likely to be more valid. Accordingly, a 'the-more-the-better' rule frequently is substituted for external validity checks.

2.1.2 Dating recalled instances. After recalling or reconstructing a specific instance of the behavior under study, respondents have to determine if this instance occurred during the reference period. This requires that they understand the extension of the reference period and that they can accurately date the instance with regard to that period. Reference periods that are defined in terms of several weeks or months are highly susceptible to misinterpretations. For example, the term 'during the last 12 months' can be construed as a reference to the last calendar year, as including or excluding the current month, and so on. Similarly, anchoring the reference period with a specific date ('Since March 1, how often ...?') is not very helpful because respondents will usually not be able to relate an abstract date to meaningful memories.

The most efficient way to anchor a reference period is the use of salient personal or public events, often referred to as landmarks. In addition to improving respondents' understanding of the reference period, the use of landmarks facilitates the dating of recalled instances. Given that the calendar date of an event will usually not be among its encoded features, respondents were found to relate recalled events to other, more outstanding events in order to reconstruct the exact time and day. Accordingly, using public events, important personal memories or outstanding dates (such as New Year's Eve) as landmarks was found to reduce dating biases (see Sudman et al. 1996, Tourangeau et al. 2000 for reviews).

Without a chance to relate a recalled event to a well-dated landmark, time dating is likely to reflect both forward and backward telescoping, that is, distant events are assumed to have happened more recently than they did, whereas recent events are assumed to be more distant than they are (Sudman et al. 1996, Chap. 8).

2.2 Estimation Strategies

In most cases, respondents need to rely on estimation strategies to arrive at an answer (Sudman et al. 1996, Chap. 9). Which strategy they use is often a function of the research instrument and hence to some extent under researchers' control.

2.2.1 Decomposition strategies. Many recall problems become easier when the recall task is decomposed into several subtasks. To estimate how often she has been eating out during the last three months, for example, a respondent may determine that she eats out about every weekend and had dinner at a restaurant this Wednesday, but apparently not the week before. Based on such partial recall she may arrive at an estimate of '18 times during the last 3 months.' Such estimates are likely to be accurate if the respondent's inference rule is adequate, and if exceptions to the usual behavior are rare.

In the absence of these fortunate conditions, however, decomposition strategies are likely to result in overestimates because people usually overestimate the occurrence of low frequency events and underestimate the occurrence of high frequency events. Hence, asking for estimates of a global, and hence frequent, category (e.g., 'eating out') is likely to elicit an underestimate, whereas asking for estimates of a narrow, and hence rare, category (e.g., 'eating at a Mexican restaurant') is likely to elicit an overestimate. The observation that decomposition usually results in higher estimates does therefore not necessarily reflect better recall.

2.2.2 Subjective theories. A particularly important inference strategy is based on subjective theories of

stability and change (Ross 1989). In answering retrospective questions, respondents often use their current behavior or opinion as a bench-mark and invoke an implicit theory of self to assess whether their past behavior or opinion was similar to, or different from, their present behavior or opinion. The resulting estimates are correct to the extent that the subjective theory is correct. This, however, is often not the case.

In many domains, individuals assume a rather high degree of stability, resulting in retrospective reports that are too close to the current state of affairs, thus underestimating the degree of change that has occurred over time. For example, retrospective estimates of income or of tobacco, marijuana, and alcohol consumption were found to be heavily influenced by respondents' income or consumption habits at the time of interview. On the other hand, when respondents have reason to believe in change, they will detect change, even though none has occurred. Assuming, for example, that therapy is helpful, respondents may infer that their problem behavior prior to therapy was more frequent than was actually the case. Relying on such theory-driven retrospective reports, researchers may infer therapeutic success where none has occurred (see Ross 1989 for examples).

2.2.3 Response alternatives. In many studies, respondents are asked to report their behavioral frequencies by checking the appropriate alternative on a frequency scale of the type shown in Table 1. Assuming that the researcher constructed a meaningful scale, respondents commonly believe that the values in the middle range of the scale reflect the 'average' or 'usual' behavioral frequency, whereas the extremes of the scale correspond to the extremes of the distribution. Given this assumption, respondents can use the range of the response alternatives as a frame of reference in estimating their own behavioral frequency. This strategy results in higher estimates along scales that present high rather than low frequency response alternatives, as shown in Table 1. As may be expected, the impact of frequency scale is most pronounced for frequent and mundane behaviors, which are poorly represented in memory (see Schwarz 1996 for a review).

In addition to influencing respondents' behavioral reports, response alternatives may also affect subsequent comparative judgments. For example, a frequency of '2½ h a day' constitutes a high response on the low frequency scale, but a low response on the high frequency scale shown in Table 1. Checking this alternative, a respondent may infer that their own TV consumption is above average in the former case, but below average in the latter. Hence, respondents in this study reported lower satisfaction with the variety of things they do in their leisure time when the low frequency scale suggested that they watch more TV than most other people.

Table 1
Reported daily TV consumption as a function of response alternatives

	Low frequency alternatives		High frequency alternatives
Up to $\frac{1}{2}$ h	7.4 percent	Up to $2\frac{1}{2}$ h	62.5 percent
$\frac{1}{2}$ h to 1h	17.7 percent	$2\frac{1}{2}$ h to 3h	23.4 percent
1h to $1\frac{1}{2}$ h	26.5 percent	3h to $3\frac{1}{2}$ h	7.8 percent
$1\frac{1}{2}$ h to 2h	14.7 percent	$3\frac{1}{2}$ h to 4h	4.7 percent
2h to $2\frac{1}{2}$ h	17.7 percent	4h to $4\frac{1}{2}$ h	1.6 percent
More than $2\frac{1}{2}$ h	16.2 percent	More than $4\frac{1}{2}$ h	0.0 percent

Note $N = 132$. Adapted from Schwarz et al. (1985). Response scales: Effects of category range on reported behavior and comparative judgments: Reprinted by permission from *Public Opinion Quarterly* 49: 388–95

To avoid these systematic influences of response alternatives, it is advisable to ask frequency questions in an open response format, such as, 'How many hours a day do you watch TV? ... hours per day.' Note that such an open format needs to specify the relevant units of measurement, for example, 'hours per day' to avoid answers like 'a few.'

As another alternative, researchers are often tempted to use vague quantifiers, such as 'sometimes,' 'frequently,' and so on. This, however, is the worst possible choice (see Pepper 1981 for a review). Most importantly, the same expression denotes different frequencies in different content domains, for example, 'frequently' suffering from headaches reflects higher absolute frequencies than 'frequently' suffering from heart attacks. Moreover, different respondents use the same term to denote different objective frequencies of the same behavior. For example, suffering from headaches 'occasionally' denotes a higher frequency for respondents with a medical history of migraine than for respondents without that migraine history. Accordingly, the use of vague quantifiers reflects the objective frequency relative to respondents' subjective standard, rendering vague quantifiers inadequate for the assessment of objective frequencies, despite the popularity of their use.

3. Pretesting

Given the complexity of the question answering process, extensive pretesting is of paramount importance. Traditionally, questions were often assumed to 'work' if respondents could provide an answer. Recent psychological research, however, facilitated the development of more sensitive pretest procedures (Sudman et al. 1996, Chap. 2, Schwarz and Sudman 1996). These procedures include the extensive use of probes and think-aloud protocols (summarily referred to as cognitive interviewing), detailed codings of interview transcripts, and the use of expert systems that alert researchers to likely problems. Drawing on these procedures, researchers can increase the odds that respondents understand the question as intended

and use cognitive strategies that are likely to produce a meaningful answer.

See also: Content Validity; Questionnaires: Assessing Subjective Expectations; Sample Surveys, History of; Sample Surveys: Methods; Social Survey, History of

Bibliography

- Bradburn N M, Rips L J, Shevell S K 1987 Answering autobiographical questions: The impact of memory and inference on surveys. *Science* **236**: 157–61
- Clark H H, Schober M F 1992 Asking questions and influencing answers. In: Tanur J M (ed.) *Questions about Questions*. Sage, New York, pp. 15–48
- Conway M A 1990 *Autobiographical Memory: An Introduction*. Open University Press, Milton Keynes, UK
- DeMaio T J 1984 Social desirability and survey measurement: A review. In: Turner C F, Martin E (eds.) *Surveying Subjective Phenomena*, Vol. 2. Russell Sage Foundation, New York, pp. 257–281
- Krosnick J A, Fabrigar L in press *Handbook of Attitude Questionnaires*. Oxford University Press, Oxford, UK
- Pepper S C 1981 Problems in the quantification of frequency expressions. In: Fiske D W (ed.) *Problems with Language Imprecision (New Directions for Methodology of Social and Behavioral Science)*, Vol. 9. Jossey-Bass, San Francisco
- Ross M 1989 The relation of implicit theories to the construction of personal histories. *Psychological Review* **96**: 341–57
- Schuman H, Presser S 1981 *Questions and Answers in Attitude Surveys*. Academic Press, New York
- Schwarz N 1996 *Cognition and Communication: Judgmental Biases, Research Methods, and the Logic of Conversation*. Erlbaum Associates, Mahwah, NJ
- Schwarz N, Hippler H J 1991 Response alternatives: The impact of their choice and ordering. In: Biemer P, Groves R, Mathiowetz N, Sudman S (eds.) *Measurement Error in Surveys*. Wiley, Chichester, UK, pp. 41–56
- Schwarz N, Strack F, Hippler H J, Bishop G 1991 The impact of administration mode on response effects in survey measurement. *Applied Cognitive Psychology* **5**: 193–212
- Schwarz N, Sudman S (eds.) 1994 *Autobiographical Memory and the Validity of Retrospective Reports*. Springer Verlag, New York
- Schwarz N, Sudman S (eds.) 1996 *Answering Questions: Methodology for Determining Cognitive and Communicative Processes in Survey Research*. Jossey-Bass, San Francisco

- Sudman S, Bradburn N M 1983 *Asking Questions*. Jossey-Bass, San Francisco
- Sudman S, Bradburn N M, Schwarz N 1996 *Thinking about Answers: The Application of Cognitive Processes to Survey Methodology*. Jossey-Bass, San Francisco
- Tanur J M (ed.) 1992 *Questions about Questions*. Sage, New York
- Tourangeau R, Rips L J, Rasinski K 2000 *The Psychology of Survey Responses*. Cambridge University Press, Cambridge, UK

N. Schwarz

Quetelet, Adolphe (1796–1874)

1. Life and Career

Born in Ghent (Belgium) to a French father and Flemish mother, the young Quetelet started his career as a teacher, and besides his interest in mathematics, he also had artistic talents.

Quetelet was the first student to obtain a doctorate in 1819 at the newly founded State University of Ghent. In his dissertation he presented a new geometric function called the focal curve. In 1819 he was appointed mathematics professor at the Atheneum in Brussels and in 1820 he was elected to the *Académie royale des sciences et des belles-lettres*. The academy became the central place from which Quetelet directed all his other activities. In 1823 his scientific orientation took a decisive turn: he proposed the foundation of a new observatory to the Dutch government, and in the same year he was sent to Paris for further preparation of this plan. At the Paris observatory, contacts were established with Laplace, Poisson, and Fourier, and Quetelet's interests in probability theory and astronomy stem from this short period in France. The construction of the Brussels observatory was completed only in 1832, but in the meantime Quetelet had also become involved in the planning of the 1830 population census.

Unfortunately this census was never completed as a result of the Belgium's secession from The Netherlands in 1830. But, elected as secretary for life at the Academy in 1834, Quetelet remained in control not only of the activities in meteorology, astronomy, and physical geography located at the observatory, but also of the activities in demography and statistics at the *Bureau de Statistique* and, from 1841 on, at the *Commission centrale de Statistique*. The procedures to be used in the new census were tried out first in Brussels in 1842, and the nationwide census took place in October 1846. Both the method of census taking and the subsequent analyses with detailed breakdowns by age and marital status set new standards for similar undertakings in most other European countries. In the

meantime Quetelet had gained an international reputation, first by the publication of his *Physique Sociale* in 1835, in which he brought together his earlier statistical work on demographic subjects, anthropometry, and criminology, and then by his relentless efforts to establish statistics as a new internationally oriented and empirically grounded scientific discipline. The latter was accomplished mainly through organizing a steady succession of international statistics conferences and by inspiring the foundation of statistical societies in other countries.

In 1855, however, Quetelet suffered a major stroke, and from then on his own scientific contribution was seriously reduced. But, until his death in 1874, he continued to inspire statistical applications in a wide variety of fields, and above all, to promote international comparability of statistical information. In Desrosières' words (1993, p. 95), 'Quetelet was the orchestra conductor of 19th Century statistics.'

2. Quetelet's Contribution to Statistics and Demography

Quetelet's statistical work was profoundly influenced by early probability theory—particularly the binomial distribution of events with equal odds—and by the use that Laplace had made of the Gaussian bell curve in astronomy. From Laplace he had gained the insight that the binomial distribution could also represent measurement error in astronomical observations, and Quetelet was convinced that this would also hold for measurements in the physical, social, and 'moral' domains. In the 1820s, measurements and frequency distributions pertaining to these domains and based on large numbers were still very rare. Quetelet had to use the frequency distribution of the chest circumferences of over 5,000 Scottish recruits to verify his hypothesis that these would conform to the binomial distribution (Hacking 1990, pp. 108–11). In a series of heights of French recruits, however, he detected a deviation from the expected distribution and therefore suspected the presence of systematic errors associated with attempts to avoid conscription. This illustrates how Quetelet used theoretical distributions to assess measurement reliability.

His use of the bell curve led him also into far deeper water: he proposed the notion of *l'homme moyen* or the average man, of which all members of a given population would be imperfect copies. For Quetelet, and for many after him, the emergence of a bell curve is a strong indication that there must be a link between the individuals producing this particular frequency distribution. Individuals would be drawn as balls from the same urn, and populations would correspond to different urns. The average men, each reconstructed from a different population, would therefore facilitate comparisons between populations, which themselves would be sufficiently homogeneous. Quetelet's 'aver-

age man' has often been misunderstood and misused. First, Quetelet was never preoccupied exclusively by averages. Whenever possible he presented complete distributions, and one of his contributions to demography is precisely his systematic presentation of *age-specific distributions* of vital events or of other occurrences (e.g., crime). This is the beginning of the statistical study of the life cycle. Second, Quetelet would be quite shocked to see his average man being associated with national prejudice, e.g., the stingy Scot, the thrifty Frenchman, ..., for the simple reason that these would all be *unmeasured* attributes. In fact, he perceived his average man as an antidote against statements based on prejudice, anecdote, and impression. It is for this reason that Quetelet can be considered as one of the founding fathers of modern empirical sociology: adequate statistical measurement presupposes operationalization with satisfactory reliability and validity.

The most fundamental critique of the average man came from A. Lexis who discovered that the variances of observed distributions are generally too large to uphold the thesis of balls drawn from the same urn. The hypothesis of a homogeneous population ceased to be tenable, and this would have major consequences for the advancement of statistics and for theories in the biological and social sciences (cf. Desrosières 1993, 111ff). In present day multivariate analyses we now routinely calculate 'little average men' in terms of subgroup means, odds ratios for different subpopulations at risk, and for different combinations of categories of co-variables. Furthermore, these numerous 'little average men' are compared and the results are tested for the presence or absence of significant differences.

The use of the bell curve was taken one step further by Quetelet's younger colleague and professor of mathematics at his alma mater in Ghent. Pierre-François Verhulst (1804–49) imagined that the growth rate r of a population would evolve according to a normal distribution. The size of a population, or of any other stock, would then follow a logistic curve, i.e. the elongated S-curve, with 50 percent of the total growth corresponding to the maximum of r , and then leveling off toward a saturation plateau. The logistic curve proved to have many applications: later demographic transitions produced population evolutions that very closely resemble the logistic curve, and diffusion processes modeled along the principle of contagion (of rumors, knowledge, disease, technology, etc.) or of competition commonly led to growth curves in accordance with Verhulst's logistic.

It is rather ironic that Verhulst's view of this particular form of population change did not feed back into Quetelet's own demographic work. Quetelet continued to think within the framework of a stationary population. Like several observers before him (e.g., Vauban, Süssmilch), Quetelet had been impressed by the observation that both numbers and distribu-

tions of vital events (deaths, births, marriages, ages at marriage, and age differences between spouses) showed a remarkable stability over time. His early work with Smits on the demography of the low countries had convinced him even more. Only major disturbances, such as a revolution, were capable of producing a temporary distortion (cf. his *causes constantes* et *causes accidentelles* in the *Physique sociale*). Quetelet kept thinking in terms of a homeostatic model, in the same way as Malthus had before him.

In his construction of the Belgian life table of 1841–50, centered around the census of 1846 and in his 'population tables' (population by age, sex, and marital status simultaneously) of 1850, Quetelet explicitly discusses the properties of a stationary population and shows that the actual age composition ought to be the same as the l_x or ${}_nL_x$ functions of the life table (l_x = number of survivors at each exact age x ; ${}_nL_x$ = number of person-years lived in the age interval x to $x+n$). Furthermore, Quetelet goes on to show that the hypothesis of constant mortality could be relaxed (1835 p. 310, 1850 p. 16): 'the necessary condition for deducing a population table (i.e., an age structure) from a life table is that the deaths by age annually preserve the same ratios [in French: *rapports*] between them' (present author's translation and clarifications).

Quetelet was clearly on the way to show that there exists a neutral mortality decline which does not affect the shape of the age distribution (for the proof see A. J. Coale 1972 pp. 33–6). What Quetelet never developed—probably because there was not yet a need for it in pre-transition Belgium—was the model of a stable population with a growth rate different from zero. Yet, his demographic work shows that he can be considered as an early founding father of the branch of contemporary demography called 'indirect estimation.' In this branch extensive use is made of population models (stationary, stable, quasistable), of model life tables, and of the mathematical properties of particular model distributions for detecting and correcting errors in recorded data, and especially for estimating basic demographic parameters of mortality, fertility, and nuptiality on the basis of fragmentary information. Modern historical demography and the demography of developing countries progressed rapidly thanks to these techniques and ways of thinking.

3. Quetelet's Contribution to Sociology

In contrast to Auguste Comte, Quetelet never developed a general plan for a new social science, and by 1900 many of his ideas had already become outdated (cf. J. Lottin 1912, pp. 391–515). He is, however, highly typical of the beginnings of sociology: there is this new entity, called 'society,' that can be studied and ana-

lyzed from the outside with objective methods. This entity follows its own ‘laws’ (i.e., patterning, regularities) and these can be detected via statistical methods. At first, Quetelet’s sociology had a deterministic ring, and his distinction between *causes constantes* and *causes accidentelles* led straight into the debate pitting collective determinism against individual freedom of action. Quetelet is clearly more impressed by the *causes constantes*, but in various places in his work he also proposed a solution to the conundrum. In an early letter to Villermé of 1832, for instance, Quetelet writes the following: ‘As a member of society [in French: *corps social*], the individual experiences at each moment the social requirements [*la nécessité des causes*] and pays them regular tribute; but as an individual he masters these influences [*causes*] to some extent, modifies their effect and can seek to reach a superior state’ (quoted in Desrosières 1993, pp. 103–4; present author’s translation). In other words, through socialization the actor becomes an integral part of society, but individual and institutional agency (for Quetelet: the advancement of the sciences) play decisive roles in social change.

For many of his contemporaries the deterministic ring of the *Physique Sociale* of 1835 continued to be a source of steady criticism, particularly when Quetelet went on with his assertion that the *causes accidentelles* merely compensate each other. Later on, however, Emile Durkheim viewed the social influence on individuals and the feedback stemming from individual and collective action in very much the same way as Quetelet. Durkheim’s equally holistic view of society, as in his *Formes élémentaires de la vie religieuse*, was directly influenced by Quetelet’s work.

In actual practice, however, Quetelet remains essentially a master of ‘comparative statics’ rather than of ‘social dynamics.’ He paid little attention to the effects of early industrialization (definitely a *cause constante*) or to the impact of the economic, social, and health crises of the 1840s. This is very surprising for a chief statistician and demographer. The social reform movements, the rise in real wages, and the decline in adult mortality in Belgium since the 1860s also remain without comment. His successors in sociology were much more alert to analyzing the dynamics of the major societal transformations they were witnessing.

4. Contributions to Psychology

Quetelet’s 1835 work on the development of human faculties has also had a major impact on the field of developmental psychology (Baltes 1983, Lindenberger and Baltes 1999). His work is the first effort to consider, based on empirical evidence, the nature of intellectual development from childhood into old age.

Treating human-ontogenetic development as a life-long process is a fairly recent advent in psychology.

Researchers in this field have been drawn to Quetelet’s early efforts in this regard. They have also emphasized that Quetelet in his discussion of methodological issues was able to prefigure some of the issues that are of high significance for modern researchers. Among these are the role of period effects and selective survival.

See also: Censuses: Demographic Issues; Censuses: History and Methods; Demography, History of; Distributions, Statistical: Approximations; Distributions, Statistical: Special and Continuous; Distributions, Statistical: Special and Discrete; Elicitation of Probabilities and Probability Distributions; Individual/Society: History of the Concept; Life Table; Measurement Theory: History and Philosophy; Measurement Theory: Probabilistic; Population Size Estimation and Capture–Recapture Methods; Simultaneous Equation Estimates (Exact and Approximate), Distribution of; Statistical Methods, History of: Pre-1900; Statistical Systems: Censuses of Population; Statistics, History of

Bibliography

- Académie Royale de Belgique 1997 *Actualité et universalité de la pensée scientifique d’Adolphe Quetelet*. Classe des sciences, actes du colloque 24–25.10.96, Brussels, Belgium
- Baltes P B 1983 Life-span developmental psychology: Observations on history and theory revisited. In: Lerner R M (ed.) *Developmental Psychology: Historical and Philosophical Perspectives*. Erlbaum, Hillsdale, NJ, pp. 79–111
- Coale A J 1972 *The Growth and Structure of Human Populations—A Mathematical Investigation*. Princeton University Press, Princeton, NJ
- Desrosières A 1993 *La politique des grands nombres—Histoire de la raison statistique*. Editions la Découverte, Paris
- Hacking I 1990 *The Taming of Chance*. Cambridge University Press, Cambridge, UK
- Lindenberger U, Baltes P B 1999 Die Entwicklungspsychologie der Lebensspanne (Lifespan-Psychologie): Johann Nicolaus Tetens (1736–1807) zu Ehren. *Zeitschrift für Psychologie* **207**: 299–323
- Lottin J 1912 *Quetelet, statisticien et sociologue*. Institut Supérieur de Philosophie, Louvain, France
- Mailly E 1875 *Essai sur la vie et les ouvrages de L-A-J. Quetelet*. *Annuaire de l’Académie Royale de Belgique* **41**: 5–191
- Porter T M 1986 *The Rise of Statistical Thinking*. Princeton University Press, Princeton, NJ
- Quetelet A 1832 Sur la possibilité de mesurer l’influence des causes qui modifient les éléments sociaux—Lettre à M. De Villermé. *Correspondance Mathématique et Publique* **7**: 321–46
- Quetelet A 1833 *Recherches sur le penchant au crimes aux différents âges*. M. Hayez, Brussels, Belgium
- Quetelet A 1835 *Sur l’homme et le développement de ses facultés, ou essai de physique sociale*. Editions Bachelier, Paris [2nd edn. 1869 eds C. Muquart, Brussels]
- Quetelet A 1847 De l’influence du libre arbitre de l’homme sur les faits sociaux, et particulièrement sur le nombre des mariages. *Bulletin de la Commission Centrale de Statistique Brussels* **2**: 135–55

- Quetelet A 1846 *Lettres à S.A.R. le Duc Regnant de Saxe Cobourg et Gotha, sur la théorie des probabilités, appliquée aux sciences morales et politiques*. M. Hayez, Académie Royale, Brussels, Belgium
- Quetelet A 1848 *Du système social et des lois qui le régissent*. Guillaumin et Cie, Paris
- Quetelet A 1849 Nouvelles tables de mortalité pour la Belgique. *Bulletin de la Commission Centrale de Statistique, Brussels* 4: 1–22
- Quetelet A 1850 Nouvelles tables de population pour la Belgique. *Bulletin de la Commission Centrale de Statistique, Brussels* 4: 1–24
- Stigler S M 1986 *The History of Statistics—The Measurement of Uncertainty Before 1900*. Belknap Press of Harvard University Press, Cambridge, MA

R. Lesthaeghe

Queues

The intuitive notion that everybody has of the so familiar ‘queues’ or waiting lines closely parallels its mathematical meaning. Queueing systems are simplified mathematical models to explain congestion. This is a field of impressive activity and growth, which has mainly been concerned with the building of the models and the study of its probability properties. Relatively little effort has been put into their statistical analysis. In this article, the main concepts underlying simple queueing systems are discussed; some Bayesian inferential tools are also given.

1. Queueing Systems

Broadly speaking, a queueing system occurs any time ‘customers’ demand ‘service’ from some facility; usually both the arrival of the customers and the service times are assumed to be random. If all of the ‘servers’ are busy when new customers arrive, these will generally wait in line for the next available server. In some examples, the terms ‘customer,’ ‘service’ and ‘server’ have the usual, everyday meaning. For instance, queueing systems occur when customers check out in a grocery store, or require the service of a teller in a bank; here, the queueing term ‘customer’ refers to the actual customers at the store or the bank, the required service is to be checked out or to resolve the bank inquiry they have, and if the server (the check-out person or the teller) is busy when a customer require his/her service, then the customer has to wait in line, that is, queue up for service. In general, the terms ‘customers,’ ‘service,’ and ‘servers’ are very broad and can refer to almost anything; for instance, telephone calls (‘customers’) arriving at a switchboard (the service time being the time each call needs to be completed), machine failures (the ‘customer’ is the

machine, and the server the repair person), demands of CPU time (server is the CPU, and the service the completion of the task), etc. Sometimes, a ‘waiting line’ (real or conceptual) never occurs. For instance, buses following their route can be considered a queueing system in which the customers are the buses, and the service is the completion of the route; here, customers are immediately ‘served’ and no waiting is necessary (technically, infinite servers, or as many servers as the number of buses, are assumed). In many interesting applications, customers require more than one service from different servers, and upon completion of service in one node, they have to proceed to another node. If the new server is busy, the customer has to queue up again for service. This is a typical situation modeled by the so called networks of queues, which are not treated in this article. An example occurs when patients enter an emergency department in a hospital: they might have to ‘queue up’ for primary attention, x-rays, lab service, surgery, and so on.

Queues (and queueing networks) have become very useful tools for evaluating (at least approximately) the performance of complex interacting stochastic systems in which congestion occurs. Obvious applications occur in airport terminals, health-care centers, and check-out lines at different counters. They have also been traditionally applied to biology (migration, population models), electrical engineering (flow models) and chemistry (polymerization, clustering) among others. The most prevalent applications of queueing systems and networks occur in the fields of manufacturing, computer networking, and telecommunications and broadcasting. But there are also many applications in the social sciences:

(a) Several types of airport queues: aircrafts waiting to take off or to land, queues of passengers at terminals (they have to go through ticket counters, check-in counters, security and passport controls, baggage carousels, customs, etc.).

(b) Police, fire, and emergency medical systems can all be regarded as queueing systems. Here the main concern is with responding to the calls for assistance as quickly as possible, important questions being to decide the number of units and which unit or units to dispatch to a specific call.

(c) Health-care systems provide many examples of analyses well suited for queueing systems, as modeling the movement of coronary patients in a hospital, determining the staffing of a department, modeling waiting queues for transplants, etc.

(d) In justice scenarios, cases queue up for trial (and the length of the trial depends on the actions of both the ‘customer’ (that is, the lawyers), and the judge.

(e) Analyzing and evaluating the delays in the mail, studying queues at voting polls to optimize the number (and type) of voting machines, and so on.

(f) Latent phase of diseases (such as Aids) in which the times of infection are arrivals to the queueing

system, the incubation is the ‘service,’ and the onset (or diagnosis) of the disease is the departure from the queueing system. Questions of interest are the number of people in the system and the time that ‘customers’ stay in the system (that is, from infection to onset or diagnosis).

Since the economical and social impact of congestion can be considerable, the importance of analyzing queueing systems has been clear ever since Erlang’s pioneering work on telecommunications at the beginning of the twentieth century. For mathematicians, queues offer an added bonus, namely that the mathematics involved can be complex and intriguing; they also have an enormous potential for varying the conditions and the assumptions of the system. It does not come as a surprise that queues developed quickly into the huge area of study that it is today. Thousands of papers, and many journals, meetings, and books are entirely devoted to the subject. It is impossible to list here even the key references, and only a small selection will be given, from which other references can be obtained. Among the numerous books, it is worth mentioning the general books of Cox and Smith (1961), Takács (1962), Cooper (1981), Gross and Harris (1985), Medhi (1991), and Nelson (1995); Kleinrock (1975, 1976) and Newell (1982) are good sources of applications; Cooper (1990) is a nice survey with an extensive bibliography. Very useful reviews of many different aspects of queueing theory (including history, statistical analysis, simulation methods, applications, etc.) can be found in Dshalalow (1995, 1997).

Simple queueing systems are defined by specifying the following:

(a) *The arrival pattern*, which among other things can be deterministic or stochastic, individual or in batches (like travelers arriving at an airport passport control), etc. Also customers can abandon the line before being served, or can interact in some other ways with the rest of the customers.

(b) *The service mechanism*, which describes the number of servers, how many customers at the time the server can handle (usually one, but bulk servers such as buses, elevators, etc. can serve more than one), and the distribution of service times (which can be independent among servers, time-dependent, dependent on queue length, etc.).

(c) *Queue discipline*, usually FIFO (first-in-first-out), but others are possible, such as last-in-first-out (LIFO) (as the subroutines in a computer program), priority customers (as express check-out lines, giving priority to shorter service times), customers switching lines, etc. Besides these, other aspects can also be relevant, such as the capacity of the waiting room, or the size of the population of customers.

In the early 1950s, Kendall introduced a useful shorthand notation for simple queueing systems. It consists on a series of symbols separated by slashes, as $A/B/c$. ‘A’ refers to the distribution of the inter-arrival times between customers; thus *Er* stands for Erlang, *G*

for General, *M* for Exponential (Markovian), etc. ‘B’ characterizes the service distribution for each server, with the same symbols as before, and ‘c’ refers to the number of servers, who are usually assumed to be equally efficient. More symbols are added when needed, to impose restrictions on the system capacity, and/or on the size of the population that feeds the system, and/or to specify the queue discipline if other than FIFO. One of the simplest and most studied queues is the $M/M/c$ queue, especially the $M/M/1$ and $M/M/\infty$; the latter are particularly useful for modeling self-service facilities or as approximations to systems with a very large number of servers (such as emergency services).

2. Equilibrium: Measures of Performance

From the probabilistic point of view, properties of queues are usually derived from the properties of stochastic processes associated with them. Let $N(t)$ denote the number of customers in the system at time $t \geq 0$. The *state probabilities* are defined as $P_j(t|\theta) = \Pr(N(t) = j|\theta)$, for $j = 0, 1, 2, \dots$, where θ denotes all parameters governing the arrival and service processes. The derivation of $P_j(t|\theta)$ is crucial in the probabilistic analysis of queueing systems. For many simple systems, the stochastic process $\{N(t), t \geq 0\}$ is a birth–death process (a ‘birth’ occurring with every arrival to the system, and a ‘death’ with every departure), greatly simplifying their study. (A brief introduction to birth–death, and other, stochastic processes with their connections to queues can be found, for example, in Ross (1993), as well as in the general references for queues given in Sect. 1). More complex systems sometimes resort to studying other *embedded* stochastic (hopefully Markovian) processes by contemplating the system as it would be seen by an arriving customer or by a departing customer.

In all but the simplest queues, determination of these probabilities is, unfortunately, extremely difficult. Often, however, it is possible to determine their large time limit, the so-called *equilibrium* or *steady-state* distribution

$$P_j(\theta) = \lim_{t \rightarrow \infty} P_j(t|\theta)$$

without actually having to compute $P_j(t|\theta)$. The equilibrium distribution does not depend on the initial conditions of the system and it is stationary (does not depend on t once the statistical equilibrium is reached). The ergodic conditions give the restrictions on θ under which the system will eventually reach the equilibrium (with probability one). Thus, for instance, in an $M/M/1$ queue, the ergodic condition establishes that the arrival rate (λ) should be smaller than the service rate (μ) (that is, the average number of customers arriving in a unit time is smaller than the average

number that the server can handle, a most intuitive condition).

For the most part, queueing theory deals with computations of the steady-state probabilities $P_j(\theta)$, and their use in computing other (steady-state) measures of performance of the queue, such as the expected number of customers in the system and/or in the waiting line, the distribution of the time spent queueing or in the system or their expected values, or the mean (or distribution) of the time the servers are busy (or idle). These steady-state distributions are also often very difficult to compute, and many require solving systems of differential equations, inversion of transforms (generating probability functions, Laplace transforms,...), etc. When only the expected values are required, an extremely useful formula for systems in equilibrium is *Little's law* which establishes that the expected number of customers in the system (L) is equal to the average time that a customer spends in the system (W) times the rate (λ) at which customers enter the system.

3. Statistical Analysis: A Bayesian Approach

Most of the vast effort in queueing theory has been devoted to the probabilistic development of queueing models and to the study of its mathematical properties; that is, the parameters governing the models are, for the most part, assumed given. Statistical analyses, in which uncertainty is introduced, are comparatively very scarce. A good review is Bhat et al. (1996). It could be argued that the general results of the (well developed) area of inference for stochastic processes (see for instance, Basawa and Prakasa Rao 1980, Lehoczky 1990) could be applied to queues. These general results do not take advantage of the special structure of queueing systems, however, nor do they address its specific questions.

Inference in queueing systems is not easy: From a frequentist perspective, development of the needed sampling distributions can be very involved and often the analysis is restricted to asymptotic results. Also, estimates can be different, for example, if the observational process differ, if the initial queue length is or is not included in the analysis. The statistical analysis is simpler if approached from the Bayesian perspective, and we take this approach below (see *Bayesian Statistics*). More arguments in favor of Bayesian analyses of queueing systems are given in Armero and Bayarri (1999), where a comprehensive list of references with Bayesian analyses of queues is also given. As an illustration of the main issues, the analysis of the much studied $M/M/1$ queue is outlined here.

According to the assumptions of the $M/M/1$ queue, the interarrival times, Y , are i.i.d. exponential with arrival rate λ , that is

$$f(y|\lambda) = \lambda e^{-\lambda y}, \quad y > 0. \quad (1)$$

Similarly, the service times X are also i.i.d., $X \sim \text{Ex}(\mu)$, independently of Y , where μ is the service rate. That is

$$f(x|\mu) = \mu e^{-\mu x}, \quad x > 0. \quad (2)$$

Steady-state distributions exist if and only if the ergodic condition

$$\rho = \lambda/\mu < 1 \quad (3)$$

holds, where ρ is the called traffic intensity and plays an important role in the study of $M/M/1$ queues.

A queue can be observed in a variety of ways and different quantities can be recorded. The most traditional experiments (providing complete information) in queues observe the system over a continuous period of time $(0, T]$. T can either be a value fixed in advance, or be determined by a suitable stopping rule. For instance, the system can be observed until the busy time reaches some preassigned fixed value, until a fixed number of customers have departed from the system, until a fixed number of transitions have been recorded, and so on (references can be found in Armero and Bayarri, 1997). Different quantities can be recorded (instants of arrivals, of departures, of commencing of service, number of arrivals and/or departures, interarrival times, service times, number of transitions, busy times, etc). If the queue is assumed to be in equilibrium, the observation of the number of customers at time 0 can also be added, providing more informative but rather cumbersome likelihood functions.

Suppose that an experiment providing n_a interarrival times, $Y_1, \dots, Y_{n_a}$ and n_s service completions, $X_1, \dots, X_{n_s}$ has been performed. Then, from (1) and (2), a likelihood function for (λ, μ) based on the observed $\mathbf{x} = (x_1, x_2, \dots, x_{n_s})$ and $\mathbf{y} = (y_1, y_2, \dots, y_{n_a})$ is

$$\ell(\lambda, \mu) \propto f(\mathbf{x}, \mathbf{y} | \lambda, \mu) \propto \lambda^{n_a} e^{-\lambda t_a} \mu^{n_s} e^{-\mu t_s} \quad (4)$$

where $t_a = \sum_{i=1}^{n_a} y_i$, $t_s = \sum_{j=1}^{n_s} x_j$ are the totals of observed interarrival times and service times, respectively. Notice that this likelihood is valid whether n_a and/or n_s are fixed or random (as long as the likelihood function is proportional to (4)). Also, the observation of the arrival and the service processes does not need to be simultaneous and the X 's and Y 's does not need to correspond to successive customers. (Of course, many of the observational procedures mentioned above result in likelihood functions proportional to (4).) Notice also that (4) applies whether or not the system is in equilibrium and whether or not equilibrium can eventually be reached.

Maximum likelihood estimators are immediate from (4), and given by

$$\hat{\lambda} = \frac{n_a}{t_a}, \quad \hat{\mu} = \frac{n_s}{t_s}, \quad \hat{\rho} = \frac{\hat{\lambda}}{\hat{\mu}} \quad (5)$$

However, standard errors of confidence intervals do require careful consideration of the specific stopping rule used, and different experimental designs will in general produce different sampling distributions for these estimators, and hence different standard errors and confidence intervals.

Bayesian analyses are insensitive to (noninformative) stopping rules, and all that is required from the data is the likelihood function (4). A prior distribution has also to be assessed quantifying whatever is known about the system before the data is collected. Usually, there is plenty of information *a priori* about the queue, especially if it is assumed to be in equilibrium (which implies, at the very least, that the system has been working for an extended period of time). A prior distribution incorporating all this information should then be assessed. However, to keep the parallelism with the likelihood analysis and also to avoid the extra work of quantifying subjective beliefs, we use a Bayesian analysis usually called ‘objective’ because the prior distribution used is of the ‘non-informative’ or ‘objective’ type. Since both λ and μ are scale parameters, the invariant measures for each will be used, and they will be assumed independent *a priori*, that is

$$\pi(\lambda, \mu) \propto \frac{1}{\lambda} \frac{1}{\mu}. \quad (6)$$

Note that no assumptions about the stationarity of the queue are made in (6). In fact, they are not needed to produce posterior distributions, although they could be incorporated if desired (see (9) below). If it is wished to incorporate prior information about the values of λ and μ , a very simple prior distribution to use is the conjugate $\pi(\lambda, \mu) \propto Ga(\lambda|a, b) Ga(\mu|c, d)$, a product of two independent gamma distributions, with the hyperparameters a, b, c, d chosen to reflect the prior information. (Again, this prior makes no assumptions about stationarity). In fact, the prior (6) can be seen as a limiting case of this conjugate prior when all hyperparameters tend to 0 (‘vague’ prior information).

Combining (4) with (6) with Bayes theorem, we get the joint posterior distribution

$$\begin{aligned} \pi(\lambda, \mu | \mathbf{x}, \mathbf{y}) &\propto \pi(\lambda, \mu) f(\mathbf{x}, \mathbf{y} | \lambda, \mu) \\ &\propto Ga(\lambda | n_a, t_a) Ga(\mu | n_s, t_s) \end{aligned} \quad (7)$$

so that λ and μ are independent *a posteriori* with scaled chi-squared distributions $2t_a\lambda \sim \chi^2_{(2n_a)}$, $2t_s\mu \sim \chi^2_{(2n_s)}$, and since they are independent, $(t_a/t_s)\rho \sim F_{2n_a, 2n_s}$, an *F* distribution. These distributions will be *exactly* valid for all experiments producing the likelihood function (4), irrespective of the stopping rule. Usual estimates of λ , μ and ρ are the means (or the modes) of their (marginal) posterior distributions, and ‘standard errors’ associated to these estimates are most naturally given by the standard deviations of those distributions. Intervals with a given probability content can similarly

be given. All these will numerically reproduce their non-Bayesian counterparts when the design has n_a and n_s fixed in advance.

An important issue when studying queues is how likely it is that steady-state can be reached. From the marginal distribution of ρ , this and other questions of direct interest can be answered. In particular, the probability that the ergodic condition holds is computed from the *F* distribution above as

$$\begin{aligned} \pi_e &= Pr(\rho < 1 | \mathbf{x}, \mathbf{y}) = \int_0^1 \pi(\rho | \mathbf{x}, \mathbf{y}) d\rho \\ &= \Psi_{(2n_a, 2n_s)}(t_a/t_s) \end{aligned} \quad (8)$$

where $\Psi_{(2n_a, 2n_s)}$ is the distribution function of an $F_{2n_a, 2n_s}$ distribution.

Most often, queues are assumed to be in equilibrium. Under this assumption, the restriction $\rho < 1$ imposed by the ergodic conditions should be incorporated into the analysis. This is easily done, and in fact amounts to truncating the unrestricted posterior distribution for ρ giving

$$\pi(\rho | \mathbf{x}, \mathbf{y}, \rho < 1) = \frac{\pi(\rho | \mathbf{x}, \mathbf{y})}{\pi_e} \quad \text{for } \rho < 1, \quad (9)$$

and $\pi(\rho | \mathbf{x}, \mathbf{y}, \rho < 1) = 0$ for $\rho \geq 1$. From (9), estimates, standard errors, credible intervals, etc. that take into account the restriction can be produced. (In this simple case, the probability π_e in (8), the proportionality constant in (9) and the moments of the truncated posterior can be given close form expressions in terms of Gauss’ hypergeometric functions.) In Armero and Bayarri (1999), an explicit expression for a factor that can be used for adjusting the usual MLE $\hat{\rho}$ is given, as well as (adjusted) ‘standard errors.’ It can there be seen that incorporating the restriction $\rho < 1$ has the effect of ‘shrinking’ the usual MLE, the shrinkage being larger the larger $\hat{\rho}$ and the smaller the sample sizes, a most intuitive result.

When statistically analyzing queueing systems, inference about the parameters is rarely (if ever) of direct interest. Usually the main goal is prediction of observable quantities that describe the utilization and degree of congestion of the system. From the Bayesian perspective, it is also very easy to produce predictive distributions incorporating the uncertainty in the estimation of the unknown parameters. The predictive distribution of the steady-state number, N , of customers in the system is computed as

$$\begin{aligned} p(N = n | \mathbf{x}, \mathbf{y}, \rho < 1) \\ = \int_0^1 Pr(N = n | \rho) \pi(\rho | \mathbf{x}, \mathbf{y}, \rho < 1) d\rho \end{aligned} \quad (10)$$

where $Pr(N = n | \rho)$ is given by the Geometric $Ge(1$

– p) distribution and $\pi(\rho | \mathbf{x}, \mathbf{y}, \rho < 1)$ is given in (9). In this case, again, (10) can be given a close form expression (see Armero and Bayarri 1994, for details).

Similar arguments apply for prediction of the number of the (steady-state) number of customers in the waiting line, time spent by a customer in the system (or waiting in line), length of busy (idle) periods, etc. All these predictive distributions are a most natural tool for deciding future interventions in the queue (see Armero and Bayarri 1997).

4. Conclusion

Queues systems are useful to stochastically model systems in which subjects interact and congestion occurs. They can describe a wide variety of such situations, in virtually all areas of application. In probabilistic terms, queues have been extensively studied under the assumption that the governing parameters are known. Inferences about these parameters can be very challenging when approached from a frequentist perspective, and the common approach of simply replacing the unknown parameters by some estimates when predicting measures of performance of the queues can produce deceiving results. Bayesian methods appear to be ideally suited to handle inferential aims in queuing systems, as incorporating restrictions in the parameter space (imposed by the ergodic conditions), predicting measures of performance, designing the system, etc. Moreover, its independence of the stopping rule makes the inferential task much simpler.

Bibliography

- Armero C, Bayarri M J 1994 Bayesian prediction in $M/M/1$ queues. *Queueing Systems* **15**: 401–17
- Armero C, Bayarri M J 1997 Bayesian questions and answers in queues. In: Bernardo J M, Berger J O, Dawid A P, Smith A F M (eds.) *Bayesian Statistics 5*. Oxford University Press, Oxford, UK, pp. 3–23
- Armero C, Bayarri M J 1999 Dealing with uncertainties in queues and networks of queues: a Bayesian approach. In: Ghosh S (ed.) *Multivariate Analysis, Design of Experiments and Survey Sampling*. Marcel Dekker, New York, pp. 579–608
- Basawa I V, Prakasa Rao B L S 1980 *Statistical Inference for Stochastic Processes*. Academic Press, New York
- Bhat U N, Miller G K, Rao S S 1996 Statistical analysis of queueing systems. In: Dshalalow J H (ed.) *Frontiers in Queueing: Models and Applications in Science and Engineering*. CRC Press, Boca Raton, FL, pp. 351–94
- Cooper R B 1981 *Introduction to Queueing Theory*. 2nd edn. North-Holland, New York
- Cooper R B 1990 Queueing Theory. In: Heyman D P, Sobel M J (eds.) *Hand-books in OR and MS, Vol 2*. Elsevier, Amsterdam, pp. 469–518
- Cox D R, Smith W L 1961 *Queues*. Wiley, London
- Dshalalow J H (ed.) 1995 *Advances in Queueing: Theory, Methods, and Open Problems*. CRC Press, Boca Raton, FL
- Dshalalow J H (ed.) 1997 *Frontiers in Queueing. Models and Applications in Science and Engineering*. CRC Press, Boca Raton, FL
- Gross D, Harris C M 1985 *Fundamentals of Queueing Theory*. 2nd edn. Wiley, New York
- Kleinrock L 1975 *Queueing Systems, Vol. 1: Theory*. Wiley, New York
- Kleinrock L 1976 *Queueing Systems, Vol. 2: Computer Applications*. Wiley, New York
- Lehoczky J 1990 Statistical Methods. In: Heyman D P, Sobel M J (eds.) *Handbooks in OR and MS, Vol 2*. Elsevier, Amsterdam, pp. 469–518
- Medhi J 1991 *Stochastic Models in Queueing Theory*. Academic Press, Boston
- Nelson R 1995 *Probability, Stochastic Processes and Queueing Theory*. Springer Verlag, New York
- Newell G F 1982 *Applications of Queueing Theory*. 2nd edn. Chapman and Hall, London
- Ross S M 1993 *Introduction to Probability Models*. 5th edn. Academic Press, New York
- Takács L 1962 *Introduction to the Theory of Queues*. Oxford University Press, New York

C. Armero and M. J. Bayarri

Copyright © 2001 Elsevier Science Ltd.

All rights reserved.

R

Race and Gender Intersections

In the 1980s and 1990s there has been a proliferation of scholarship by and about women of color that challenges heretofore accepted theories of society. Yet, this writing is even more marginalized than thinking that centers on white women. While a great deal of feminist work has been written to challenge existing sociological theory, there has been only limited impact on new theory formation. Currently, in most social theory texts, gender generally is segregated into a section on feminist theory, or ignored completely. Moreover, few articles in the *American Journal of Sociology* and the *American Sociological Review* contain discussions of feminist theory.

Thus, while 'feminist theory' remains marginalized, or ignored, the theorizing by and about women of color remains marginalized within the margins. Lest we think that this is simply a byproduct of white male hegemony, this marginalization is also present in feminist theorizing. To be sure, most edited volumes now include articles by and about women of color, the question remains to what extent have these writings influenced mainstream feminist theory?

Theorizing by and about women of color has had a bifurcated influence on the development of feminist theory. As Hooks (2000) notes, there is a 'high' and 'low' theory, and the former still continues to exclude much of the experientially based theorizing of women of color. Caraway (1991) notes that 'in academic practice, some white feminists have exhibited a tendency to ghettoize the creative and expressive writings of Black women as "empirical" supplements to the "formal" theory texts of white feminists.' For example, in a brief overview of some of the main feminist theory texts, women of color generally are segregated into a section on difference, multiculturalism, or women of color. Or, the text is peppered with reprints of 1980s writings by prominent women of color such as Bell Hooks, Pat Hill Collins, Adrienne Rich, Patricia Williams, Audre Lorde, Gloria Anzaldua, and Angela Davis. But often more recent scholarship on women of color is absent. In an overview of the literature, Glenn (1999) concludes that, '... despite increased recognition of the interconnectedness of gender and race, race remained undertheorized even in the writings of women of color. In the absence of a "theory" of race comparable to a "theory" of gender, building a com-

prehensive theory has proven elusive.' Following the work of Liu (1991), Higginbotham (1992), Kaminsky (1994), and Stoler (1996), Glenn calls for a synthesis through social constructionism, arguing that race and gender are 'relational concepts whose construction involves both representation and social structural processes in which power is a constitutive element' (1999).

1. Intersectionality Theory

Theorizing about women of color has had a major impact on postmodernist feminist theory. Postmodernism is committed to the deconstruction of absolute truths and views knowledge as socially constructed. Women of color have developed 'Intersectionality Theory' that subscribes to many postmodern tenets, and is committed to what McLaren (1994) calls resistance postmodernism. As Cordova (1998) states, 'Along with the critiques of modernism comes the notion that social "facts" are social constructions; social reality is mediated by one's positionality; that identity is tied to location (subjectivity); and that meaning is unstable and connected to the advanced capitalist construction of consumerism.'

'Intersectionality Theory' objects to the notion that all truths are equal and to the notion that individual narratives should not become a basis for shared resistance to oppressive power relations. So, while postmodernism emphasizes deconstructionism, 'Intersectionality Theory' insists upon an equal attention to constructionism. Equally important is that 'Intersectionality Theory' is written in a language more readily accessible to the general public than is that of French Postmodernism. It is often written in the language of experience.

Many women of color have contributed to the development of 'Intersectionality Theory,' which is interdisciplinary (Hooks 1984, Anzaldua 1987, 1990, Collins 1990, 1998, Crenshaw 1989, 1991, 1997, de la Torre and Pesquera 1993, Glenn 1985, Lorde 1984, Williams 1991, Zinn and Dill 1993, Brewer 1990, King 1988, Espiritu 1997). Collins (1990) argued for a more complex conceptualization of gender relations; one that included an understanding of the 'matrix of domination,' that is the intersection of gender, class, race, sexual preference, and other relevant systems of oppression. For Collins and other feminists of color,

these intersections are not simply separate cumulative variables by which we can assess one's experience of oppression. Rather it is the way these positions interlock in interactive or unique ways at their intersections that provides the experience of oppression. She writes: 'replacing additive models of oppression with interlocking ones creates possibilities for new paradigms. The significance of seeing race, class, and gender as interlocking systems of oppression is that such an approach fosters a paradigmatic shift of thinking inclusively about other oppressions, such as age, sexual orientation, religion and ethnicity.'

Hooks (1989) offers a similar analysis with her notion of a 'politic of domination,' arguing that, all of these axes are imbued with 'notions of superior and inferior ... component.' 'Intersectionality Theory' simultaneously challenges multiple levels of oppression, while viewing race and gender as something more than the sum of their parts (King 1988). This theoretical approach arose in parallel with mainstream feminist theory as evidenced by the statement of the Combahee River Collective (1983), which was formed in the mid-1970s by a group of Black lesbian feminists. They wrote: 'The most general statement of our politics at the present time would be that we are actively committed to struggling against racial, sexual, heterosexual, and class oppression, and see as our particular task the development of integrated analysis and practice based upon the fact that the major systems of oppression are interlocking. The synthesis of these oppressions creates the conditions of our lives. As Black women we see Black feminism as the logical political movement to combat the manifold and simultaneous oppressions that all women of color face.'

'Intersectionality Theory' is more than the notion that the whole does not equal the sum of its parts. It also encompasses some creative, novel, methodological approaches to the study of social phenomena. For example, Sandoval (1998) writes about the extent to which Anzaldúa's (1987) book influences feminists of color: 'This "borderlands" feminism, many argue, calls up a syncretic form of consciousness made up of transversions and crossings; its recognition makes possible another kind of critical apparatus and political operation in which *mestiza* feminism comes to function as a working chiasmus (a mobile crossing) between races, genders, sexes, culture, languages, and nations. Thus conceived, *La conciencia de la mestiza* makes visible the operation of another meta-form of consciousness that insists upon polymodal forms of poetics, ethics, identities, and politics not only for Chicanas/os but for any constituency resisting the old and new hierarchies of the coming millennium.'

In this view, the study of any social problem entails a holistic approach, one that includes analyses of race, class, gender, sexuality, language, culture, and nation.

2. Intersectionality Theory and Mainstream Sociology

'Intersectionality Theory' encompasses a methodological approach critical to the study of social phenomena, and it has been employed by many women of color social scientists in such areas as the study of family, social movements, race and ethnicity, social stratification, legal institutions and crime, organizations, work and welfare, and politics. The following discussion provides examples of ways in which analysis of the intersectionality of race and gender enrich our theoretical understanding in several areas of study.

Let us begin with family studies. Wilson's (1987) work on the American Black underclass, and its implications for the Black family, has gained national attention. By contrast, the research conducted by Black women scholars, many of whom contextualize his findings, has received less attention. Dickerson (1995) examines predominant approaches to the study of African-American families that often judge them by their similarity to white families. She finds these approaches less useful than a form of 'centrism,' an Afrocentric paradigm that 'centers' groups in terms of their own cultural meanings, attitudes, and values, and gathers data from the *standpoint* of the group being studied. This approach is very much in keeping with 'Intersectionality Theory.'

Using this approach, Sudarkasa (1993) and Jewell (1988) both argue against the claim that 'female-headed households are the root cause of the deplorable conditions in which many Blacks find themselves in urban ghettos' (Sudarkasa 1993). Rather, they argue that the most critical factors for family survival and success are economic resources and proximity to kin. Harriet McAdoo's research indicates that kin systems are critical for the survival of black single mothers and the elderly. Wilson (1987) seems to drop the importance of kin and translates the need for economic resources into a need for a marriageable pool of Black males. Black male unemployment, he argues, is a critical intervention in the fight to uplift the underclass. While this is certainly of central importance, Black women's employment is equally important. After all, many upper-class white women are choosing to bear children out of wedlock, but there has not been the corresponding outcry, because they are economically capable of sustaining themselves and their offspring. But it is most difficult for Black women to support their children. Black women are disadvantaged by both racial and gender differences in earnings, which persist even after levels of educational attainment are held constant (Herring 1999).

A perspective that emphasizes the intersectionality of race, class, and gender are absent in most mainstream theories of racial/ethnic families, and this omission leaves such theories inadequate.

This pattern of omission is also visible in social movement theory. While feminists such as Verta

Taylor, Myra Marx Ferree, Pam Oliver, Nancy Whittier, and several others have made significant contributions to social movement theory, scholarly contributions by and/or about women of color are largely absent from major theoretical texts in this area, ignoring the contributions of the few women of color who are writing in this area, such as Patricia Zavella, Belinda Robnett, and Bernice McNair Barnett. Zavella's (1987) work on union organizing has shown that recruitment strategies must specifically address issues of race/class/gender as they affect the reality of potential recruits. Both Barnett (1993) and Robnett (1996, 1997) have shown the importance of Black women's contributions to the civil rights movement. Robnett (1997) argues that the intersectionality of race/class/gender shaped the organization of social movement organizations and leadership in social movements. Robnett (1997) shows that the experiences and contributions of women in the civil rights movement are varied by race and class. African-American women were critical bridge leaders in the civil rights movement, forging ties between movement organizations and the grass roots participants. Without their participation, the movement would not have succeeded. Her intersectionality approach lends insight into the decline of that movement, as the clash between white politicians, Black male elites, and rural Black women led to the disillusionment and disengagement of grass roots leadership. Scholars such as Dill (1988), Gilkes (1988), and Sacks (1988), have written about the redefinition of what we define as 'political,' suggesting that the boundaries are much too narrow, which has important implications for social movement theory, including what counts as a political identity or political activity.

Indeed, the theorizing by and about women of color has been excluded in a number of areas including work, the state, and the sociology of knowledge. For example, while feminist scholars have analyzed social reproduction, that is, childbearing and childcare, its racialization has been noted less. Glenn (1999) points out that while scholars theorize about how gender leads women to be assigned to child care in the context of a sex division of labor, they often ignore the fact that women of color are the ones who end up in the situation of leaving their own children at home or with kin, while they take care of other women's children to make the money to support their own children. Class, race, and gender affect the life chances of women of color, making them the group often concentrated in low paid childcare because of a lack of better employment options.

Glenn notes a similar omission in studies of citizenship. Whereas many feminists such as Pateman (1988), Young (1989), and Okin (1979) have challenged conceptualizations of the 'universal citizen,' short shrift has been paid to race/gender. Glenn (1999) asserts, 'The problem with looking at citizenship as only gendered or only raced is the familiar one:

women of color fall through the cracks.'

Collins's (1999) critique of feminist analyses of science is equally instructive. She suggests that, 'feminist analyses of gender and scientific knowledge might benefit from closer association with emerging scholarship on intersectionality.' Collins convincingly illustrates the ways in which 'selecting a specific social location, social practice, group history, or topic, and subjecting it to an intersectionality analysis,' challenges the notion that Western science is an objective, apolitical, ahistorical endeavor.

In short, the social location (in the margin in the margin) of the scholarship by and about women of color, impedes the development of our understanding of social phenomena.

See also: Altruism and Prosocial Behavior, Sociology of; Civil Rights Movement, The; Conflict Sociology; Critical Race Theory; Feminist Theory and Women of Color; Gender and Feminist Studies; Gender and Feminist Studies in Sociology; Gender, Class, Race, and Ethnicity, Social Construction of; Gender History; Multicultural Feminism: Cultural Concerns; Multiculturalism: Sociological Aspects; Poverty and Gender in Affluent Nations; Race: History of the Concept; Race Identity; Racial Relations; Racism, History of; Racism, Sociology of

Bibliography

- Anzaldúa G 1987 *Borderlands/La Frontera: The New Mestiza*. Aunt Lute Press, San Francisco
- Anzaldúa G (ed.) 1990 *Making Face, Making Soul: Haciendo Caras: Creative and Critical Perspectives by Women of Color*. Aunt Lute Press, San Francisco
- Barnett B 1993 Invisible southern black women leaders in the civil rights movement: The triple constraints of gender, race, and class. *Gender and Society* 7(2): 162-82
- Brewer R 1990 In: Albrecht L, Brewer R M (eds.) *Bridges of Power: Women's Multicultural Alliances*. New Society Publishers, Philadelphia, PA
- Brewer R 1993 Theorizing race, class and gender: The new scholarship of Black feminist intellectuals and Black women's labor. In: James S M, Busia A P A (eds.) *Theorizing Black Feminisms: The Visionary Pragmatism of Black Women*. Routledge, New York, pp. 13-30
- Caraway N 1991 *Segregated Sisterhood: Racism and the Politics of American Feminism*. University of Tennessee Press, Knoxville, TN
- Collins P H 1990 *Black Feminist Thought: Knowledge, Consciousness, and Empowerment*. Unwin Hyman, Boston
- Collins P H 1998 *Fighting Words: Black Women and the Search for Justice*. University of Minnesota Press, Minneapolis, MN
- Collins P H 1999 Moving beyond gender: Intersectionality and scientific knowledge. In: Ferree M M, Lorber J, Hess B (eds.) *Revisioning Gender*. Sage, Thousand Oaks, CA, pp. 261-84
- Combahee River Collective 1983 Combahee river collective statement. In: Smith B (ed.) *Home Girls: A Black Feminist Anthology*. Kitchen Table—Women of Color Press, New York, pp. 272-82

- Cordova T 1998 Power and knowledge: Colonialism in the academy. In: Trujillo C (ed.) *Living Chicana Theory*. Third Woman Press, Berkeley, CA, pp. 17–45
- Crenshaw K 1989 Demarginalizing the intersection of race and sex: A black feminist critique of antidiscrimination doctrine, feminist theory, and antiracist politics. *University of Chicago Legal Forum* 7: 139–67
- Crenshaw K 1991 Mapping the margins: Intersectionality, identity politics, and violence against women of color. *Stanford Law Review* 43: 6
- Crenshaw K 1997 Intersectionality and identity politics: Learning from violence against women of color. In: Shanley M, Narayan U (eds.) *Reconstructing Political Theory: Feminists Perspectives*. Pennsylvania State University Press, University Park, PA
- De la Torre A, Pesquera B (eds.) 1993 *Building with Our Hands: New Directions in Chicana Studies*. University of California Press, Berkeley, CA
- Dickerson B 1995 (ed.) *African American Single Mothers: Understanding their Lives and Families*. Sage, Thousand Oaks, CA
- Dill B T 1988 Making your job good yourself: Domestic service and the construction of personal dignity. In: Bookman A, Morgen S (eds.) *Women and the Politics of Empowerment*. Temple University Press, Philadelphia, PA, pp. 33–52
- Espiritu Y L 1997 *Asian American Women and Men*. Sage, Thousand Oaks, CA
- Gilkes C T 1988 Building in many places: Multiple commitments and ideologies in black women's community work. In: Bookman A, Morgen S (eds.) *Women and the Politics of Empowerment*. Temple University Press, Philadelphia, PA, pp. 53–76
- Glenn E N 1985 Racial ethnic women's labor: The intersection of race, gender, and class oppression. *Review of Radical Political Economics* 17: 86–108
- Glenn E N 1999 The social construction and institutionalization of gender and race: An integrative framework. In: Ferree M M, Lorber J, Hess B B (eds.) *Revisioning Gender*. Sage, Thousand Oaks, CA, pp. 3–41
- Herring C 1999 African Americans in contemporary America: Progress and retrenchment. In: Dworkin A G, Dworkin R J (eds.) *The Minority Report*. Harcourt Brace, Orlando, FL, pp. 181–208
- Higginbotham E 1992 African American women's history and the metalanguage of race. *Signs* 17: 251–74
- Hooks B 1984 *Feminist Theory from Margin to Center*. South End Press, Boston
- Hooks B 1989 *Talking Back: Thinking Feminist, Thinking Black*. South End Press, Boston
- Hooks B 2000 Theory as liberatory practice (from *Teaching to Transgress* (1994)). In: Kolmar W, Bartkowski F (eds.) *Feminist Theory: A Reader*. Mayfield Publishing, Mountain View, CA, pp. 28–33
- Jewell K S 1988 *From Mammy to Miss America and Beyond: Cultural Images and the Shaping of US Social Policy*. Routledge, New York
- Kaminsky A 1994 Gender, race, raza. *Feminist Studies* 20: 3–32
- King D K 1988 Multiple jeopardy, multiple consciousness: The context of a black feminist ideology. *Signs* 14(1): 42–72
- Liu T 1991 Teaching differences among women from a historical perspective: Rethinking race and gender as social categories. *Women's Studies International Forum* 14(4): 265–76
- Lorde A 1984 *Sister Outsider: Essays and Speeches*. Crossings Press, Trumansburg, NY
- McLaren P 1994 *Critical Pedagogy and Predatory Culture: Oppositional Politics in a Postmodern Era*. Routledge, New York
- Okin S 1979 *Women in Western Political Thought*. Princeton University Press, Princeton, NJ
- Pateman C 1988 *The Sexual Contract*. Polity, Cambridge, MA
- Robnett B 1996 African American women in the civil rights movement: Gender, leadership, and micromobilization. *American Journal of Sociology* 101(6): 1661–93
- Robnett B 1997 *How Long? How Long? African American Women in the Struggle for Civil Rights*. Oxford University Press, New York
- Sacks K 1988 *Caring by the Hour*. University of Illinois Press, Chicago
- Sandoval C 1998 Mestizaje as method: Feminists-of-color challenge the canon. In: Trujillo C (ed.) *Living Chicana Theory*. Third Woman Press, Berkeley, CA, pp. 352–70
- Stoler A 1996 Carnal knowledge and imperial power: Gender, race and morality in colonial Asia. In: Scott J W (ed.) *Feminism and History*. Oxford University Press, New York, pp. 209–66
- Sudarkasa N 1993 Female-headed African American households: Some neglected dimensions. In: McAdoo H P (ed.) *Family Ethnicity: Strength in Diversity*. Sage, Newbury Park, CA, pp. 81–9
- Williams P 1991 *The Alchemy of Race and Rights: Diary of a Law Professor*. Harvard University Press, Cambridge, MA
- Wilson W J 1987 *The Truly Disadvantaged: The Inner City, the Underclass, and Public Policy*. University of Chicago Press, Chicago
- Young I 1989 Polity and group difference: A critique of the ideal of universal citizenship. *Ethics* 99: 250–74
- Zavella P 1987 *Women's Work and Chicana Families: Cannery Workers of the Santa Clara Valley*. Cornell University Press, Ithaca, NY
- Zinn M B, Dill B T (eds.) 1993 *Women of Color in United States Society*. Temple University Press, Philadelphia, PA

B. Robnett

Race and the Law

1. Definition of the Field

The study of race and the law is most comprehensively developed and best defined as a genre of scholarship in the US. There the field is primarily, but far from exclusively, occupied by scholars trained in law and working in law schools. Significant contributions to the field have also been made in Canada and Europe, where the field is more evenly divided among lawyers and social scientists trained in anthropology, sociology, and cultural studies.

1.1 Mainstream vs. Critical Scholarship

Scholarship addressing race and the law is split between mainstream approaches and critical ap-

proaches. Mainstream approaches generally conceive of racial discrimination as a minor and isolated exception to the otherwise liberal and enlightened societies of the industrialized West. They understand racial injustice as largely unconnected to other institutions in the societies in which it operates and are therefore optimistic that racism can be eliminated 'surgically,' without requiring significant or widespread changes in those societies. Critical scholarship by contrast understands racism as deeply implicated in the institutions and traditions of societies and is therefore accordingly both more radical in its prescriptions and more circumspect in its predictions of success (Bell 1989, Crenshaw 1995, Williams 1991).

Because the definition of race is malleable, fluctuating, and contested, the study of race and the law encompasses not only 'race' as it is colloquially understood but also issues of ethnicity, national origin, caste, and religious and cultural difference. The field could therefore be understood to include the study of ethnic and nationalist divisions in Asian countries, caste hierarchy in Hindu societies, tribal conflict in central Africa, and religious conflict worldwide. It implicates the study of nationalism and pan-nationalist movements, communities of diaspora and indigenous peoples. Moreover, the field is deeply concerned with the intersection of race with issues of gender, sexual orientation, and class (Crenshaw 1995, A Harris 1990, Thomas 1995, Williams 1991).

Generally speaking, the study of race and the law encompasses two major projects. The first is a critique that demonstrates how legal institutions and formal laws produce and perpetuate racism and racial hierarchy. The second is the analysis of legal reform that might eliminate or minimize the effects of racial hierarchy. This article will address each in turn.

1.2 The Critique of Law as a Cause of Racism

Legal scholarship addresses the various ways in which law has produced and continues to produce racial hierarchy. Law contributes to racism and racial hierarchy in at least three distinct ways. First, law codifies racial categories, which are then employed by both public and private actors to erect and maintain racial hierarchies. Second, law institutionalizes racial hierarchy by creating formal rules requiring or encouraging racial hierarchy: by segregating the races, by excluding the members of certain races from participation in social, economic, and political activities and institutions, or by legally codifying the subordinate status of one or more races. These first two types of legal action are commonly known as *de jure* racial discrimination because they explicitly employ or create racial categories by force of law. Third, law allows, encourages, and requires the maintenance of institutional practices and social customs that perpetuate racial subordination by entrenching the

racial hierarchies of the past, even when no formal legal rule requires or promotes racial subordination on its face. This latter type of legal action is known as *de facto* racial discrimination because it in fact produces and enforces racial subordination, although it does not do so explicitly.

2. Racial Categorization

Many scholars have studied the development and use of racial classificatory schemes (Davis 1991, Haney-Lopez 1994, Omi and Wynant 1994). Virtually all scholars agree that, historically, schemes of racial classification were not based on sound scientific theories and therefore do not reflect any 'real' biological differences between the races. Given the dubious scientific basis for any racial classification of human beings, it is not an exaggeration to assert that law not only enforced racial hierarchy, it also (along with other social institutions and practices) in fact produced a specific type of racial difference.

2.1 The Ontological Status of Racial Categories

There is disagreement about the ontological status of racial categories and the centrality of law in producing racial categories. Some scholars believe that racial categories have a substance that predates or transcends legally enforced and encoded racial categorization. While few scholars would argue that races are well understood in terms of biological or genetic difference, many argue for or assume that cultural differences distinguish the races (Gotanda 1995, Peller 1995). The belief that racial categories describe an essence or positive (in the sense of 'positive science' not necessarily in the sense of 'good') attribute of human beings—whether biological or cultural—is known as racial essentialism.

Other scholars argue that social practices, including legal codification, produced racial differences where there were no such differences before. These scholars contend race is not an attribute of individuals or groups but instead the effect of a social discourse that ascribes difference to individuals and groups (Appiah 1996, Ford 1999). They insist that without the discourse of race, there would be no races. The belief that race is not an attribute or essence of individuals or groups but is instead the effect of the social discourses that describe it and the social practices that make it relevant is known as anti-essentialism or social constructivism.

2.2 The Legal Codification of Racial Categories

Legal historians have examined the development and transformation of racial categories and definitions. A

vast literature details the variety of racial classificatory schemes employed in the USA, Europe, and South Africa. The contemporary literature is overwhelmingly critical of most racial categorization, both because of its opportunistic use of scientific data (or its use of questionable science to justify racial categories) and because of the oppressive use to which racial categories have been put.

There is, however, some dissension on the question of remedial or so-called benign racial classification. A widespread understanding of constitutional law and a pervasive popular attitude holds that the legal codification of *any* racial classification is socially deleterious and morally indefensible (Thernstrom and Thernstrom 1999). This view asserts that the law should be *race blind*. Many scholars who may not insist on strict color blindness nevertheless have serious misgivings about the use of racial classifications, even for remedial purposes (Carter 1992). Others believe that racial classification is not *per se* undesirable and argue that remedial racial classification is the necessary antidote to centuries of oppressive racial classification. They assert that the law should be affirmatively *race conscious* (Crenshaw 1995, Gotanda 1995, Matsuda and Lawrence 1997, Peller 1995).

2.3 The Legal Enforcement of Racial Hierarchy

Along with the study of racial categorization, a core concern of the study of race and the law is the legal enforcement of racial hierarchy. Of course the two are closely related: racial categories developed hand in glove with policies of racial exclusion, segregation, disenfranchisement, and involuntary servitude. The study of racial hierarchy has traditionally distinguished between formally enforced racial distinctions, hierarchies, and exclusions—*de jure* racial practices—and the formal requirement and enforcement of practices that have the effect of establishing or perpetuating racial hierarchy—*de facto* racial practices.

2.3.1 De jure racial subordination. Legal historians have detailed the legal mechanisms that historically have enforced and perpetuated racial hierarchy. Historians of the USA study the institution of racial hierarchy through chattel slavery (Tushnet 1981) and the post-emancipation development of 'Jim Crow' laws, which required racial segregation in public institutions and accommodations such as school rest rooms, drinking fountains, eating facilities, railroad coaches, and theater or stadium seating (Woodward 1989). Scholars of South Africa study the emergence, transformation, and eventual dismantling of racial *apartheid*, a policy that codified a belief in black inferiority as explicit policy and required segregation of the races in virtually every aspect of political and social life (Goldberg 1994, McClintock and Nixon

1991). Comparative scholars analyze the difference and similarities of South African apartheid and American Jim Crow laws (Ford 1999, Goldberg 1994).

Much of the historical study of European racial policy takes place as part of the study of European colonialism (Appiah 1993, Bhabha 1994, Gilroy 1991). Explicit colonial racial policy and the treatment of non-whites in Europe are understood to implicate the ideology and political economy of European colonial rule.

2.3.2 De facto racial subordination. Formal racial exclusion, segregation, and hierarchical classification are rare in the post-industrial mass democracies of today. They have been rejected by mass opinion, eliminated by legislation, and struck down by courts applying constitutional law. Nevertheless, racial hierarchy persists as a significant feature of every society with a history of racial differentiation. Accordingly, the study of race and the law focuses a significant amount of attention on informal or implicit practices that effectively perpetuate racial hierarchy.

The focus on *de facto* racial subordination almost certainly began as a response to racial practices in the post civil war USA. Federal constitutional guarantees of emancipation from slavery, equal protection and the right to vote prohibited certain explicit racial barriers (see *Civil Rights; Fundamental Rights and Constitutional Guarantees*). But many states, especially those of the former confederacy, found mechanisms to effectively re-establish these racial barriers without explicit reference to race. For instance, although blacks had a formal right to vote, many states effectively eliminated the black franchise through poll taxes and literacy requirements that were informally waived or ignored for whites but strictly enforced against blacks. And despite the formal emancipation of black slaves, many blacks were effectively forced into perpetual indentured servitude through the notorious sharecropping system of Southern agriculture.

Informal and *de facto* forms of racial subordination emerged again as the response to more comprehensive civil rights reforms of the Civil Rights Act of 1964. Significantly, the Civil Rights Act prohibited racial discrimination in employment and accommodation in private institutions such as hotels, restaurants, and theaters, and in the sale and rental of housing (see *Civil Rights; Fundamental Rights and Constitutional Guarantees*). Many of the private institutions covered by the Act sought effectively to reproduce racial exclusion or discrimination through less explicit means. Lawyers, activists, and scholars, armed with a knowledge of post civil war history, were able to identify and critique these practices. The attack on *de facto* racial subordination and discrimination became a staple of civil rights litigation as scholars and lawyers worked to expose the variety of practices that effectively per-

petuated racial subordination. Antiracist scholarship and legal practice has attacked a host of formally colorblind practices as effectively racially discriminatory, including electoral gerrymandering (Guinier 1995), standardized tests (Matsuda and Lawrence 1997), neighborhood schools in a context of racially segregated neighborhoods (Ford 1995), restrictive land use regulations (Ford 1995, C Harris 1995), informal racial steering in real estate sales (Ford 1995), and informal racial discrimination in hiring and promotions (Crenshaw 1995, Williams 1991).

3. Antiracist Legal Reform

Scholarship that examines, analyzes, and advocates the development of antiracist legal reform encompasses a vast range of proposals and is based on a number of methodologies and ideological commitments. Many approaches to antiracist reform build on civil rights reforms of the 1960s in the USA, advocating the expansion of such reforms and/or the adaptation of such reforms in other national contexts (see *Civil Rights*).

A significant split has emerged between mainstream and critical approaches. Critical approaches question the efficacy and the underlying ideologies of mainstream civil rights and liberal antiracism (see *Critical Race Theory*). Beginning from an analysis that sees the problems of racism as deeply imbedded in the structure of social practices and institutions—including law itself—critical approaches are skeptical of reform proposals that attempt ‘surgically’ to correct or ameliorate racial bias while leaving most institutions and social practices more or less unchanged (Bell 1989, Crenshaw 1995). Nevertheless, many critical approaches often build on mainstream reforms, arguing for their radical expansion (Matsuda and Lawrence 1997, Williams 1991). Others suggest that mainstream reforms do more harm than good and should be abandoned altogether in favor of more comprehensive changes (Tushnet 1982).

3.1 Integrationism vs. Separatism

Crudely put, antiracist legal reform is characterized by a split between two strategies and/or normative commitments, one which favors the integration of the races and one which favors the development of distinct, autonomous and separate communities (see Ford 1995, Peller 1995).

Historically, racial hierarchy has been legally enforced through spatial quarantine: American Jim Crow laws enforced the physical separation of the races in public accommodations and state institutions; racial zoning and restrictive covenants have divided the metropolitan landscape into racial neighborhoods; South African apartheid enforced racial segregation on a comprehensive regional scale and imposed

racially specific curfews in white-only jurisdictions. Spatial segregation not only had potent symbolic consequences but also facilitated the inequitable distribution of resources: white-only institutions were consistently better situated, better funded, and socially and politically privileged compared to those assigned to racial minorities.

Given this history, a significant approach to remediation has been to encourage the integration of the races in public and private institutions and residential environments. Many American civil rights reformers believed that antiracist legal reform should embody a positive preference for integration *qua* integration. Others believed that integration was desirable only because it would undo or stymie the inequitable distribution of resources in a society. These two rationales for integration produced a powerful consensus that has dominated American civil rights discourse since the 1960s and which has been influential in many other national contexts.

At the same time, however, a competing view has held that the development of distinctive and empowered minority communities is essential to racial justice and equality (Peller 1995). This view focuses on the need for political solidarity, social support, and cultural autonomy. Racial ‘separatism’ has grown in popularity as many racial minorities have become disillusioned with the limitations of mainstream civil rights and the difficulties and compromises entailed in integration.

3.2 Anti-discrimination

A dominant conceptual and practical approach to antiracist legal reform has been that of anti-discrimination. The anti-discrimination concept presupposes that racial hierarchy is the result of irrational prejudice or bigotry based on race that results in racial discrimination. Therefore, legal reform prohibits intentional racial discrimination and practices and some policies that disparately impact members of racial minority groups. The anti-discrimination approach to antiracism focuses on specific wrongful actions by identifiable decision makers: if no wrongful act can be identified, this approach recognizes no legally cognizable or remediable harm (see Freeman 1995; *Discrimination; Equality and Inequality: Legal Aspects*).

3.3 Anti-subordination

An alternative approach to antiracist legal reform is that of anti-subordination. The anti-subordination approach views racism as a set of practices that have the effect of systematically reproducing racial hierarchy, even if no identifiable decision maker can be said to be racially biased. Anti-subordination approaches therefore find it uncontroversial that practices that adversely effect racial minorities should

come under legal scrutiny, even in the absence of evidence of intent or bias. More importantly anti-subordination approaches look for patterns and systems in which several distinct practices combine to produce regular, predictable, and systematic racial subordination even when each individual practice, viewed in isolation, might seem unproblematic (Crenshaw 1995).

3.4 Identity Politics and Multiculturalism

One significant strand of antiracist politics seeks to advance a project of racial recognition or multiculturalism, emphasizing the distinctiveness of racial minority communities in terms of culture. This strand seeks legal reform prohibiting discrimination against the cultural practices of racial minority groups as well discrimination on the basis of race *qua* status (Gotanda 1995, Peller 1995, Kymlicka 2000, Taylor 1994).

This approach can be understood as a logical extension of the anti-subordination approach to antiracist legal reform: discrimination against cultural practices effectively harms members of racial minority groups and can be seen as a component of systematic racial subordination.

One necessary consequence of racial multiculturalism is that it requires a definition or widely shared understanding of the content of racial identity or a racial culture. Such understandings or definitions are often controversial, both in society at large and within the given racial group itself. Moreover, the idea that racial identity or culture exists in a static or stable form is itself controversial. Some scholars would reject the very idea of a static or widely shared understanding of racial identity or culture as mistakenly essentialist (see Sect. 2.1.1).

3.5 Anti-essentialist and Cosmopolitan Conceptions of Racial Identity

An emerging strand of antiracist scholarship first articulated within postcolonial African, Caribbean, and Indian communities approaches racial identity as fluid and supple, yet also the source of rich political, social, and cultural affiliations. These approaches define racial identity not in terms of a biological or cultural essence, but instead in terms of identification (Appiah 1996, Bhabha 1994, Gilroy 1991). According to one strand of this scholarship, members of racial minority groups share common 'passions' based on shared or analogous experiences of racial subordination (Gilroy 1991). This approach dispenses altogether with any remnant of biological or inheritance notions of racial identity: members of what might traditionally have been several different 'races' can and do share a cosmopolitan racial identity which can transcend national boundaries, cultural traditions, skin color, and phenotype. This scholarship adopts an explicitly anti-subordination approach to antiracism,

but rejects the multiculturalist focus on cultural traditions within distinctive racial or ethnic communities as traditionally defined. Instead it embraces a cosmopolitan or diasporic culture that joins racial minorities of various ethnicities and cultural backgrounds.

See also: Affirmative Action: Comparative Policies and Controversies; Affirmative Action: Empirical Work on its Effectiveness; Affirmative Action Programs (India): Cultural Concerns; Affirmative Action Programs (United States): Cultural Concerns; Affirmative Action, Sociology of; Civil Liberties and Human Rights; Conflict Sociology; Critical Race Theory; Discrimination; Gender, Class, Race, and Ethnicity, Social Construction of; Race: Genetic Aspects; Race: History of the Concept; Race Relations in the United States, Politics of; Racism, History of; Racism, Sociology of; Segregation Indices

Bibliography

- Appiah K A 1993 *In My Father's House, Africa in the Philosophy of Culture*. Oxford University Press, New York
- Appiah K A 1996 Race, culture, identity. In: Appiah K A, Gutman A (eds.) *Color Conscious*. Princeton University Press, Princeton, NJ
- Bell D 1987 *And We Are Not Saved: The Illusive Quest for Racial Justice*. Basic Books, New York
- Bell D 1992 *Race, Racism and American Law*. Little Brown, Boston
- Bhabha H K 1994 *The Location of Culture*. Routledge, New York
- Carter S L 1991 *Reflections of an Affirmative Action Baby*. Basic Books, New York
- Crenshaw K 1995 Race, reform and retrenchment: Transformation and legitimation in anti-discrimination law. In: Crenshaw K, Gotanda N, Peller G, Thomas K (eds.) 1995 *Critical Race Theory: The Key Writings that Formed the Movement*. New Press, New York, pp. 103–26
- Crenshaw K, Gotanda N, Peller G, Thomas K (eds.) 1995 *Critical Race Theory: The Key Writings that Formed the Movement*. New Press, New York
- Davis F J 1991 *Who is Black? One Nation's Definition*. Penn St. Press, PA
- Delgado R, Stephanic J (eds.) 1999 *Critical Race Theory: The Cutting Edge*, 2nd edn. Temple University Press, PA
- Denton N A, Massey D S 1993 *American Apartheid: Segregation and the Making of the Underclass*. Harvard University Press, Cambridge, MA
- Ford R T 1995 The boundaries of race. In: Crenshaw K, Gotanda N, Peller G, Thomas K (eds.) 1995 *Critical Race Theory: The Key Writings that Formed the Movement*. New Press, New York, pp. 449–64
- Ford R T 1999 Law's territory (a history of jurisdiction). *Michigan Law Review* 97: 843–930
- Freeman A D 1995 Legitimizing racial discrimination through anti-discrimination law: A critical review of supreme court doctrine. In: Crenshaw K, Gotanda N, Peller G, Thomas K (eds.) *Critical Race Theory: The Key Writings that Formed the Movement*. New Press, New York, pp. 29–45
- Gates H L Jr. (ed.) 1986 *Race, Writing and Difference*. University of Chicago Press, Chicago

- Gilroy P 1991 *There Ain't No Black in the Union Jack: The Cultural Politics of Race and Nation*. University of Chicago Press, Chicago
- Goldberg D 1994 *Racist Culture*. Blackwell, London
- Gotanda N 1995 Our constitution is colorblind. In: Crenshaw K, Gotanda N, Peller G, Thomas K (eds.) *Critical Race Theory: The Key Writings that Formed the Movement*. New Press, New York, pp. 257–75
- Guinier L 1994 *The Tyranny of the Majority: Fundamental Fairness in Representative Democracy*. Free Press, New York
- Haney-Lopez I F 1994 *White by Law*. New York University Press, New York
- Harris A P 1990 Race and essentialism in feminist legal theory. *Stanford Law Review* 42: 581–616
- Harris C 1995 Whiteness as property. In: Crenshaw K, Gotanda N, Peller G, Thomas K (eds.) *Critical Race Theory: The Key Writings that Formed the Movement*. The New Press, New York, pp. 276–91
- Kennedy R 1998 *Race, Crime and the Law*. Vintage Books, New York
- Kymlicka W 2001 *Politics in the Vernacular: Nationalism, Multiculturalism, and Citizenship*. Oxford University Press, Oxford, UK
- Matsuda M J, Lawrence C R III 1997 *We Won't Go Back: Making the Case for Affirmative Action*. Houghton Mifflin, Boston
- McClintock A, Nixon R 1991 No names apart: The separation of word and history in Derrida's 'Le dernier mot du racisme.' In: Gates H (ed.) 1986. *Race, Writing and Difference*. University of Chicago Press, Chicago, pp. 339–53
- Omi M, Wynant H 1994 *Racial Formation in the United States: From the 1960s to the 1990s*, 2nd edn. Routledge, New York
- Peller G 1995 Race consciousness. In: Crenshaw K, Gotanda N, Peller G, Thomas K (eds.) *Critical Race Theory: The Key Writings that Formed the Movement*. The New Press, New York, pp. 127–58
- Taylor C 1994 The politics of recognition. In: Gutman A, Taylor C (eds.) *Multiculturalism: Examining the Politics of Recognition*. Princeton University Press, Princeton
- Thernstrom S, Thernstrom A 1999 *America in Black and White: One Nation, Indivisible*. Touchstone Books, New York
- Thomas K 1995 Rouge et Noir reread. In: Crenshaw K, Gotanda N, Peller G, Thomas K (eds.) 1995 *Critical Race Theory: The Key Writings that Formed the Movement*. The New Press, New York, pp. 465–94
- Tushnet M V 1981 *The American Law of Slavery, 1810–1860: Considerations of Humanity and Interest*. Princeton University Press, Princeton, NJ
- Tushnet M 1984 An essay on rights. *Texas Law Review* 62: 1363–403
- Williams P J 1991 *The Alchemy of Race and Rights*. Harvard University Press, Cambridge, MA
- Woodward C V 1989 *The Strange Career of Jim Crow*, 3rd edn. Oxford University Press, New York

R. T. Ford

Race and Urban Planning

In general, the term 'race and urban planning' refers to those aspects of urban planning which affect or are affected by the racial variety of the human condition,

as this is expressed in multiracial societies. 'Urban planning,' also known as urban and regional planning, or as town and country planning, depending on the country of practice, helps to develop and enhance various aspects of life in human settlements. These include the location and arrangement of housing, community services, and commercial facilities, or the broader management of land use. The word 'race' typically refers to one of a few broad categories of humanity, such as Caucasian, Asian, or African. Although urban planning should ideally be race-neutral, in fact the racial characteristics of various populations have often affected their receipt of services, their spatial location or lack of spatial mobility, and their interaction with or participation in planning decisions. The lessons learned about how multiracial contexts affect urban planning can often be applied to contexts diversified by some other dimension, such as ethnicity, tribe, religion, or nationality.

1. Background of the Issue

The linkage between urban planning and race arose in several modern multiracial societies because of the obvious ways in which decision makers were using urban planning as a tool for racial oppression. The context in these early cases was one of at least one dominant and one subordinate racial group. For example, in the USA during the early part of the twentieth century, some of the first tools which professional urban planners used to regulate land use helped enforce the segregation of African American households from white households (Silver 1997). In societies characterized by colonialism, professional designers of new cities controlled by European powers made sure that the spatial arrangement of urban settlements sent at least two clear signals. The first, symbolized by grand approaches to isolated government buildings, was that the European colonial government was to be respected and feared. The second was that those sections of the city which housed Europeans were favored with higher standards of municipal services and housing than other sections set aside for 'natives' (Hall 1996).

As multiracial urban societies evolved during the twentieth century, many governments abandoned the purposeful isolation of supposedly subordinate racial groups through such means. Other manifestations of the problem arose, however. In the USA during the 1930s through the 1960s, some localities used programs allied with urban planning and redevelopment to isolate racial minorities, either through constructing racially segregated public housing or through selectively moving people out of and bulldozing minority neighborhoods by using federal 'urban renewal' funds. These activities brought mixed reactions: in only a few cases during the World War II era did civil rights organizations protest the blatantly segregationist public housing decisions, for example, which were some of

the first manifestations of racism in modern public policy. Outrage over the abuses of the federal 'urban renewal' program, on the other hand, brought spirited protests. The public outcry over 'urban renewal,' which became nicknamed 'Negro removal,' led to significant reforms in federal policy and in local urban planning activities after the 1960s. Yet racial segregation was so entrenched in US society that it proved difficult to eliminate its many manifestations simply by changing federal and local policy making. Similarly, as colonialism withered and died, former colonies arose to govern and plan their own cities and to eliminate the most pervasive abuses of colonial urban planning. Yet in urban places they found themselves dealing with significant social divisions nonetheless, because of caste, class, ethnicity, and any number of other dimensions. Once set in place, city structure proved hard to change.

Although in most cases the purposeful use of urban planning tools to enforce racial subordination, isolation, and segregation has waned as time has passed, the problem is by no means irrelevant to the present. As we will see in the sections below, subtle exclusion still characterizes seemingly innocuous urban planning decisions in modern countries. One society that has continued, until very recently, to use overt techniques of racial oppression through urban planning is South Africa. Other examples of cities with continuing conflict over related issues include Jerusalem, Israel, and Belfast, Northern Ireland. Yet the nature of the issues has changed markedly as broader societal conditions have changed. In Johannesburg, South Africa, for example, the end of apartheid has brought the end of legally enforced spatial segregation by race, but not the end of racial segregation and uneven development. In such a case, it may be harder to address present conditions and problems with meaningful solutions than it was to operate under the pernicious but forthright apartheid.

2. Areas of Scholarly Inquiry

Contemporary scholars of urban planning have focused on several general areas of connection between race and urban planning. Some scholarly works are connected with other more traditional social science endeavors in that they emphasize the empirical analysis of such phenomena as racial segregation. Other scholars, sometimes using qualitative or holistic methodologies such as history or case studies, document the existence of problems of racial estrangement or oppression in the context of private or public decisions concerning planning and development. Particularly valuable are those discussions which deal with the distribution of power and decision making, when urban planning is being practiced in a context of racial or other inequality.

In addition, several authors focus on specific dimensions of the problems. Certain findings which relate to

land use, zoning, housing, and neighborhoods, as well as those related to services such as transportation, or to general urban conditions such as the environment, are now briefly summarized.

2.1 General Connections Between Race and Urban Planning

Several general works about urban history or urban planning history have addressed the use of planning for racial purposes through case studies or 'stories' about injustice or oppression. Such works have found that poor race relations have strongly influenced local development and planning policy as well as federal policies. A typical story in this tradition, for example, might trace the deleterious effects of racial conflict upon local efforts to improve one city, perhaps covering all or part of a period extending from the early twentieth century until the 1990s. Topics covered for a US scholar might include decisions about public housing and urban renewal, effects of civil rebellion or racial disorder upon inner-city redevelopment, housing discrimination, or failed attempts to stabilize neighborhoods in the context of racial change. In more recent years, issues covered might include how racial conflict bars attempts to bridge the gap between cities and suburbs in metropolitan areas, or weakens coalitions which might otherwise unify in order to revitalize neighborhoods or central business districts (Thomas 1997, Catlin 1993).

Other books in this general category document the continued existence of racial segregation, or suggest policy actions designed to lessen the effects of such segregation. More connected with the disciplinary areas of the social sciences than with urban planning, these works have nonetheless continued to have a major influence on thinking about the problems of racial separation and isolation. In some cases, recent works documenting conditions of racial and income segregation and of metropolitan fragmentation have helped heighten the dialogue about how to undertake fundamental society reforms, many of which could profoundly affect the work of urban planners. Examples include possibilities of merging racially segregated municipalities or heightening their cooperation, sharing tax bases, dispersing low-income housing, or in other ways halting or minimizing the effects of society's tendency to disperse and separate in racially exclusionary ways (Rusk 1995, Orfield 1998).

Within this same category of 'general' scholarship concerning urban planning and race also belong those writings connecting issues of race and empowerment, particularly as this concerns participation in making decisions that affect the communities in which racial minorities live. In addition to the extensive literature on urban renewal which deals with this issue, a few additional works have addressed directly the problem of how racial minority communities have tried to wrestle control of their destinies from the hands of city

officials. In some cases in the USA, it turns out, such minority communities have had outstanding success in doing so; one case study documents this evolution in Birmingham, Alabama, for example (Connerly and Wilson 1997). In other cases, it was federal programs from the 1960s and 1970s such as the Office of Economic Opportunity or Model Cities that led to increases in power and influence, since these programs required extensive citizen participation by low-income populations in the areas they served.

Sometimes planners in strongly segregated societies have made valiant attempts to overcome historic misuse of planning activities for purposes of subjugation, but have discovered that the issues are not so easily resolved. As apartheid has ended in South Africa, for example, attempts have been made to develop a more egalitarian approach to municipal decision making concerning planning. As a few observers have begun to document, however, the lingering effects of unequal power and unequal resources continue. The 'deracialization' of urban space in Cape Town led to laudable efforts to promote racial mixture in previously segregated neighborhoods. In addition, local planners sponsored public forums which tried to develop an inclusionary model for guiding decisions which would affect the future layout of residential and commercial areas. Yet the 'forum' approach to egalitarian decision making wrongfully assumed that all stakeholders would be included, would participate, and would have equal voice. In a society beset by profound disadvantages for people of color, such assumptions have sometimes proved premature (Watson 1998).

Where the history of racial segregation and oppression is less debilitating than in South Africa, planners in multiracial and multicultural contexts are beginning to make creative strides. Urban planners in several Canadian cities, for example, are being forced to examine the cultural and political biases inherent in a wide range of their activities. They are becoming more attentive to the need to invite participation in decision making by a wide range of racial, ethnic, and national groups, and to listen to these communities when they protest that various regulations hamper their free cultural expression. As a simple example, Chinese communities expect that their 'shopping centers' will meet different functions and have a different scale than would the mainstream population, which has set regulations according to European standards. As another example, many regulations concerning religious facilities assume that the building to be sited will be a church. An immigrant-rich context in which Middle Eastern Muslims are increasing in numbers has required creative flexibility concerning the need for progressive congregational growth. Small Muslim communities often start with meetings in houses which cannot therefore be strictly limited to residential use. These groups may then move to rented stores or other buildings, and then construct mosques

with unique architectural features not allowed by regulations designed only for Christian churches (Qadeer 1997).

2.2 Land Use and Zoning

Relevant research in this area documents the fact that land use and zoning decision making in cities, towns, and suburban areas routinely excludes people on the basis of socioeconomic status or race. While in decades past exclusion specifically focused on race, more modern and subtle versions keep people out through informal means, or through socioeconomic exclusion. An example would be zoning ordinances which exclude low-income or multifamily housing from well-to-do suburbs. The general term for such discriminatory action is exclusionary zoning. In metropolitan areas in which low-income racial minorities are isolated and segregated, such action may have much the same effect as would zoning exclusions based on racial characteristics, which are now, in most places, illegal.

A related phenomenon is expulsive zoning. This refers to the practice of using minority neighborhoods as dumping grounds for Locally Unwanted Land Use (known in planning circles by the acronym LULU, or as a NIMBY, which stands for 'Not In My Back Yard'). Examples of such uses are landfills, transfer stations, and inappropriately large concentrations of group homes for people with chemical dependencies. Localities often accomplish unfair concentrations of LULUs by granting inappropriate variances, conditional use permits, or spot zonings in residentially zoned low-income minority neighborhoods. These actions effectively 'expel' residential use and residents from an area and lower the property values of the remaining residences (Rabin 1989).

In many cities, residential minority neighborhoods (often stable living environments for most of their history) are also externally controlled by the fact that, years ago, they were zoned for commercial or industrial use. When in order to maintain stability they attempt to seek a less intense, residential classification, which is known as downzoning, such neighborhoods may be rebuffed because of their history, minority status, or lack of power within the political arena. Zoning also controls aspects of the accessibility and availability of childcare, elder care, and other services needed to maintain neighborhood livability.

In the USA, one drawback to effective action is that the federal Supreme Court, in a long series of cases, has made it clear that socioeconomic exclusion through zoning will not be substantively challenged from the courts. Yet seemingly innocuous opposition to subsidized or low- or moderate-income housing in towns or suburban communities may indeed have enormous racial implications. The failure to grant standing (the ability to be heard as a substantive party in court cases) to several low-income plaintiffs who

wished to press for approval of affordable housing in suburban communities has effectively shut off an important route to access to suburban housing for inner-city minorities. Extensive surveys of zoning ordinances show that many local communities discriminate, through seemingly innocuous means, against 'nontraditional' families, such as single female-headed families or extended families which choose to live together to reduce expenses. Such means may include strict zoning definitions of family, barring nonrelated family groupings or extended families from residential areas zoned for single-family housing, or placing limitations on home-based occupations, meaning that low-income residents cannot legally support themselves or supplement their incomes by opening up home businesses (Ritzdorf 1997).

2.3 Housing, Neighborhoods, Mortgage Lending

Yet another area of concern focuses on the effects of continued racial segregation in residential markets, and of discrimination among lending institutions, both of which can have a marked effect on the nature, stability, and viability of urban neighborhoods. To some extent it has been necessary to document the problem, since it has not always been accepted that discrimination among lenders existed, or that segregation in housing was anything but good business practice. The problem of segregation itself is largely the purview of sociologists and geographers, and three particularly influential urban scholars in the contemporary USA are Douglas Massey, Nancy Denton, and William Julius Wilson. They have clearly documented the fact that racial segregation has continued to be a problem in the last half of the twentieth century, and has had profound effects in reducing opportunity, shaping city growth, and crippling several generations of black youth (Wilson 1996, Massey and Denton 1993). Several good works have also documented continuing problems over equitable treatment for racial minorities in obtaining money for buying, rehabilitating, or insuring their homes and their other property (Squires 1997, Yinger 1995).

These issues are particularly important to planners because, with racially skewed housing markets, it becomes very difficult for planners or other local decision makers to trust private housing markets to supplement public investments. Yet no country has sufficient public investment money to reshape urban areas without significant support by the private sector. Neither can planners undertake lasting reforms via neighborhood revitalization if the population is in constant turmoil because of racial change. This is particularly a problem in the USA, which is characterized by extensive decentralization of middle-class populations to racially homogeneous suburbs. When neighborhoods in multiracial societies become identified exclusively with the subordinate class, it becomes

very difficult to market them as viable places to live for a wide variety of people. Only in a few cases have US scholars been able to discover means of stabilizing such communities, such as using fair housing initiatives—that is, initiatives designed to promote residential decisions based on the principles of free access for racial minorities—and purposeful efforts to convince home buyers of one race to move to residential neighborhoods populated by another race. Cleveland, Ohio, is one US metropolitan area that has attempted to promote a proactive agenda of this nature. Observers have found, however, that the private housing market will resist policies designed to remedy racial imbalances unless these are supported by home buyers, renters, politicians, realtors, financial institutions, and insurers. Only in a few metropolitan areas have prointegrative efforts by several local organizations and governments resulted in neighborhoods characterized by long-term racial diversity, however (Keating 1994).

The other side of coin is that in some national communities, diversity of neighborhoods is a sign of healthy diversity of cultural choice, rather than of a history of exclusionary discrimination. Again, the example of Canadian cities offers some insight into the implications of such variety for the activities which urban planners undertake (Qadeer 1997). In such cases, the challenges are to remove the last vestiges of policies based on the unconscious assumption that dominant cultural values rule the world, and to begin to allow for a multitude of life-styles and cultural expressions.

A related concept concerns community development, which refers to the purposeful improvement of residential communities using, in part, the participatory involvement of community residents. Given the previous discussions on segregation and on empowerment, it is fitting to note that in some cases isolated minority communities have been able to use tools of urban development and of urban planning to improve their own conditions. Several observers have offered inspiring case studies of successful attempts by African American communities to overcome the handicaps of race and class in order to create viable living spaces (Leavitt and Saegert 1989, McDougall 1993).

2.4 Transportation and Environment

Several other aspects of urban life can also relate to issues of race. Two are briefly mentioned here: transportation and the environment. Examples related to transportation, would be the negative effects of highway construction policies upon the viability of minority neighborhoods, or the problem of lack of accessibility to means of transportation by racial minorities (Bullard and Johnson 1997). This problem is also evident in several historical case studies of African American community life, since federal highway legislation functioned much as redevelopment

legislation did in the physical clearance of African American neighborhoods.

A major race-related issue of concern to planners is the disproportionately negative impact that environmental pollutants have upon racial minorities. Reference was made earlier to expulsionary zoning (see Sect. 2.2), indicating some of the problems that can arise. Examples of such problem areas include the siting of regular and hazardous waste landfills in minority neighborhoods, the prevalence of lead poisoning in minority housing, and the consumption by indigenous hunters and fishers of fish and wildlife contaminated by harmful substances. Social scientists have helped to document the extent of the problem, and have joined with civil rights and other groups to examine the racial implications of such cases (Bullard 1994). One of the claims among activists in the USA is that environmental pollution disproportionately affects people of color, and negatively impacts their health and well-being. Another key concern is that the mainstream environmental movement has largely ignored these racial issues. Recent environmental justice initiatives, however, have forced the environmental movement to become more inclusive. Environmental racism is an important phenomenon for urban planners because of their role in helping to determine land uses, including the location of landfills and heavy industry, and the reclamation of environmentally contaminated land for urban redevelopment projects.

3. Responses to Racial Inequality and Injustice

Urban planners and urban planning scholars have reacted in several ways to the continued problems related to the connections between racial injustice and urban planning. Scott Bollens has offered a typology of responses that is useful as a framework for discussion (Bollens 1988).

The first possible response by an urban planner is to side with the status quo in the name of objectivity, even in the face of blatant inequities. The urban planner or urban planning policy is supposedly 'neutral,' but in fact reinforces the dominant race's position. Such a 'neutral' planner may refuse to acknowledge or address issues of segregation, estrangement, or polarization. This approach is sometimes associated with so-called 'technical' planners, who prefer to focus on the technical, non-social aspects of their work without dealing with social problems. The shield, under such circumstances, is that urban planning is a professional exercise not involved with politics or with issues of social justice. This particular attitude is not uncommon among professionals, and may indeed be relatively harmless. It is not so harmless when exemplified by planners in racially conflicted situations, particularly when those planners are members of the dominant race or ethnic group, or support that dominant group to the exclusion of their own.

The claim that the planner is 'color-blind,' in such cases, may be only a poor excuse for collaboration with systematic injustice.

A second set of responses might be called 'partisan.' Under such a situation, racial inequality and injustice are obvious, but the planner rejects the concerns of the powerless group and sides with the powerful one. The city's government leaders view neighborhoods and policies through a lens colored by race, and the interests of the dominant group—to exclude, isolate, and perhaps even oppress the subordinate group—become the norm for professional urban planners. As the political leaders set the policies, the urban planners carry them out, even in the face of blatant favoritism or discrimination. Bollens offers several examples from South Africa, but also notes the same syndrome in Jerusalem, where the systematic and purposeful isolation of Arab neighborhoods has gained the force of righteousness, even though it is, in many cases, patently unjust. As Bollens notes, somehow many local urban planners have been able to function without any noticeable discomfort about the obvious partisanship of their actions.

A third category attempts to reduce inequalities between groups, sometimes through remedial action for the oppressed group. Many US urban planners consider the equity planning movement to be one of their most important responses to the racial crisis in America's cities. According to its advocates, equity planning offers a reasoned and practical response to the crisis of local policy caused by racial segregation and poverty in America's cities. That is, urban planners (and by implication other urban professionals) are not bound by the confines of their municipal jobs to an endless cycle of repression or inattention toward the disadvantaged. Instead, urban planners can and do arise above the limitations of government bureaucracy, and pursue redevelopment and other local policies to channel direct benefits to people hindered by race or socioeconomic status. Several key authors explicitly address the important role of race in helping to create the unjust conditions to which equity planners must respond (Krumholz and Forester 1990).

Finally, Bollens refers to 'a resolver strategy,' which he indicates extends beyond equity planning. A resolver approach attempts to address the root causes of power imbalances, competition among racial and ethnic groups, and disempowerment. It then aims to confront the status quo and profoundly transform the system. Although this approach is in some ways simply a more impassioned form of equity planning, it also has broader implications in countries such as South Africa. There, for example, 'resolvers' have come forward who have begun to carry out development planning that differs markedly from traditional urban planning activities, which favored white areas and white concepts even after apartheid. Development planning, in contrast, more fully integrates spatial

planning with social and economic planning than did traditional urban planning; it restructures policy directions by cutting across traditional departments and governments; and, perhaps most importantly, it includes 'a participatory process aimed at empowering the poor and marginalized.' It also happens to involve more members of the victim class as professional planners than was previously the case.

That the urban planning profession maintains an ongoing dialogue about this and other approaches to resolving racial injustice is a healthy sign. Although many societies remain severely fragmented and alienated by race, the situation is open to proactive efforts to improve existing conditions. As larger societies become more comfortable with multi-racialism and with multiculturalism, urban planning as a field can and will adjust.

See also: Community Economic Development; Income Distribution: Demographic Aspects; Neighborhood: General; Neighborhood Revitalization and Community Development; Poverty: Measurement and Analysis; Poverty, Sociology of; Racism, Sociology of; Social Inequality in History (Stratification and Classes); Social Stratification; Urban Geography; Urban Growth Models; Urban Poverty in Neighborhoods; Urban Sociology

Bibliography

- Bollens S 1988 Urban planning amidst ethnic conflict: Jerusalem and Johannesburg. *Urban Studies* 35: 729–50
- Bullard R (ed.) 1994 *Unequal Protection: Environmental Justice and Communities of Color*. Sierra Club, San Francisco, CA
- Bullard R, Johnson G (eds.) 1997 *Just Transportation: Dismantling Race and Class Barriers to Mobility*. New Society Press, Branford, CT
- Catlin R 1993 *Racial Politics and Urban Planning: Gary, Indiana 1980–1989*. University Press of Kentucky, Lexington, KY
- Connerly C, Wilson B 1997 The roots and origins of African American planning in Birmingham, Alabama. In: Thomas J M, Ritzdorf M (eds.) *Urban Planning and the African American Community: In the Shadows*. Sage, Thousand Oaks, CA
- Hall P 1996 *Cities of Tomorrow: An Intellectual History of Urban Planning and Design in the Twentieth Century*, 2nd edn. Blackwell, Oxford, UK
- Keating W D 1994 *The Suburban Racial Dilemma: Housing and Neighborhoods*. Temple University Press, Philadelphia, PA
- Krumholz N, Forester J 1990 *Making Equity Planning Work: Leadership in the Public Sector*. Temple University Press, Philadelphia, PA
- Leavitt J, Saegert S 1989 *From Abandonment to Hope: Community Households in Harlem*. Columbia University Press, New York
- Massey D, Denton N 1993 *American Apartheid*. Harvard University Press, Cambridge, MA
- McDougall H 1993 *Black Baltimore: A New Theory of Community*. Temple University Press, Philadelphia, PA
- Orfield M 1998 *Metropolitics: A Regional Agenda for Community and Stability*. Brookings Institution Press, Washington, DC

- Qadeer M A 1997 Pluralistic planning for multicultural cities. *Journal of the American Planning Association* 63: 481–94
- Rabin Y 1989 Expulsive zoning: the inequitable legacy of Euclid. In: Harr C, Kayden J (eds.) *Zoning and the American Dream: Promises Still to Keep*. American Planning Association Press, Washington, DC, pp. 101–20
- Ritzdorf M 1977 Locked out of paradise: Contemporary exclusionary zoning, the Supreme Court, and African Americans, 1970 to the present. In: Thomas J M, Ritzdorf M (eds.) *Urban Planning and the African American Community: In the Shadows*. Sage, Thousand Oaks, CA, pp. 43–57
- Rusk D 1995 *Cities Without Suburbs*. Woodrow Wilson Press, Washington, DC
- Silver C 1997 The racial origins of zoning in American cities. In: Thomas J M, Ritzdorf M (eds.) *Urban Planning and the African American Community: In the Shadows*. Sage, Thousand Oaks, CA, pp. 23–42
- Squires G D (ed.) 1997 *Insurance Redlining: Disinvestment, Reinvestment, and the Evolving Role of Financial Institutions*. Urban Institute Press, Washington, DC
- Thomas J M 1997 *Redevelopment and Race: Planning a Finer City in Postwar Detroit*. Johns Hopkins University Press, Baltimore, MD
- Thomas J M, Ritzdorf M (eds.) 1997 *Urban Planning and the African American Community: In the Shadows*. Sage, Thousand Oaks, CA
- Watson V 1998 Planning under political transition—lessons from Cape Town's metropolitan planning forum. *International Planning Studies* 3: 335–50
- Wilson W J 1996 *When Work Disappears*. Knopf, New York
- Yinger J 1995 *Closed Doors, Opportunities Lost: The Continuing Costs of Housing Discrimination*. Russell Sage Foundation, New York

J. M. Thomas

Race: Genetic Aspects

Natural species are genetically polymorphic, in two related, but distinct ways. Different individuals may carry different alleles of the same gene (polymorphism within populations), and different populations may be characterized by different alleles, or by different frequencies of the same alleles (polymorphism among populations). Differences among populations depend on various evolutionary factors, but populations that exchange migrants tend to resemble each other, whereas mutually isolated groups and communities tend to diverge genetically.

The relative importance of polymorphism within and among populations varies across species. In many plant and animal species, the largest share of genetic diversity occurs among populations or groups thereof. When that is the case, most individuals of unknown origin can be assigned to one group on the basis of their genetic characteristics, with a high degree of statistical confidence. If distinct groups occupy different geographical areas, it is customary to define them as subspecies or races. The question is how

faithfully this concept also describes the levels and patterns of genetic diversity observed among humans.

One of the latest definitions of race is as follows: 'A large population of individuals who have a significant fraction of their genes in common, and can be distinguished from another race by their common gene pool' (Vogel and Motulsky 1986). But the concept of race is older than genetics. Long before Mendel the term referred to people sharing the same 'blood,' that is, to groups of genealogically-related individuals, especially those of noble families (Furetière 1690, cited in Cohen 1991). In the eighteenth and nineteenth centuries, the term race came to designate constant human types, and came to be used not only to describe, but also to explain, human diversity. People who look different belong to different races, and they are different because they belong to different races. That view is still widespread; it seems based on an empirical observation, and on an assumption. The observation is that biologically related people look alike, and the assumption is that our species is naturally subdivided in clusters of individuals who look alike because they are genealogically related, or races.

Is that assumption based on solid scientific grounds? Or are human races arbitrary groups of populations, which do not share a common evolutionary history more than other populations classified into different races? Only in the second half of the twentieth century has it become possible to address this question using genetic data. For the purposes of this article, the hypothesis being tested is whether the observed patterns of genetic diversity suggest that humankind is subdivided in nearly discrete, and internally homogeneous, groups. The results of those tests, and their evolutionary implications, will be summarized in this article.

1. Listing Human Races

Prior to measuring genetic differences within and among human groups, one has to define these groups. The many lists of races which have appeared in the course of history may be regarded as successive

attempts to identify the set of physical traits which leads to the best discrimination of groups, that is, traits for which the among-group component of variation is maximum. Ideally, once such 'best' physical traits are identified, most individuals and populations should be classified unambiguously into the appropriate group, except perhaps for a minority of them, traditionally regarded as derived from interracial crosses.

Table 1 summarizes some lists of human races proposed in the last three centuries. Although this selection is necessarily incomplete (the interested reader is referred to Cohen 1991 and Molnar 1998) a problem, and a trend, are evident. The problem is, any group defined on the basis of morphological traits harbors some degree of internal variation, and different criteria, say skin color and skull shape, lead to different classifications. This is clearly shown by the fact that, although the existence of human races has been challenged only recently (Livingstone 1964), little agreement has evidently been reached on how many races exist, and which. The trend is a consequence of the problem. As more and more groups came in contact with the European scientists who were developing these classifications, the number of proposed races increased steadily.

There is an exception to that trend. In 1963, Coon proposed that there are only five subspecies among humans. In that context, subspecies was synonymous to race, and it was defined as a regional population 'which meets two tests: (1) it occupies a distinct geographical territory; (2) it differs from other subspecies of the same species in measurable characteristics to a considerable degree' (1962, p. 19). Coon's definition overlaps with the one given by Vogel and Motulsky (1986), and both can be tested using genetic data.

2. Apportioning Human Diversity

Because there is no agreement on the best system of human racial classification, the groups or putative races being compared in the studies here reviewed were seldom the same. In general, when the groups con-

Table 1
Some lists of proposed human races

Author	No. of races	Races proposed
Linnaeus (1735)	6	Europaeus, Asiaticus, Afer, Americanus, Ferus, Monstruosus
Buffon (1749)	6	Laplander, Tartar, South Asiatic, European, Ethiopian, American
Blumenbach (1795)	5	Caucasian, Mongolian, Ethiopian, American, Malay
Cuvier (1828)	3	Caucasoid, Negroid, Mongoloid
Deniker (1900)	29	
Weinert (1935)	17	
Von Eickstedt (1937)	38	
Biasutti (1959)	53	
Coon (1962)	5	Congoid, Capoid, Caucasoid, Mongoloid, Australoid

Source: Coon 1962, Cohen 1991, and Molnar 1998

Table 2

Percentages of the overall genetic variance at three levels of population subdivision, as estimated from protein (P), RFLP (R), microsatellite (M), or single-nucleotide (S) polymorphisms; G are multilocus genotypes, that is, sets of loci jointly analyzed which, (for mitochondrial DNA and the Y chromosome) are also jointly inherited, as a single locus. Genome location refers to the fact that the loci of interest are autosomal (A), mitochondrial (M), or Y-linked (Y)

Authors	N Loci	Type	Genome location	Within populations	Between populations	
					within groups	Between groups
Lewontin (1972)	17	P	A	85.4	8.3	6.3
Latter (1980)	18	P	A	83.8–87.0	5.5–6.6	7.5–10.4
Ryman et al. (1983)	25	P	A	86.0	2.8	11.2
Barbujani et al. (1997)	79	R	A	84.5	3.9	11.7
Barbujani et al. (1997)	30	M	A	84.5	5.5	10.0
Barbujani et al. (1997)	16	G	A	83.6	8.4	8.0
Barbujani et al. (1997)	109	M + R	A	84.4	4.7	10.8
Seielstad et al. (1998)	n.a.	S	M	81.4	6.1	12.5
Seielstad et al. (1998)	n.a.	S	Y	35.5	11.8	52.7
Santos et al. (1999)	30	S	Y	59.0	25.0	16.0
Poloni et al. (1997)	9	R	Y	70.1	6.9	23.0
Barbujani and Excoffier (1999)	100	R	M	65.5	7.1	27.5
Barbujani and Excoffier (1999)	5	G	M	77.4	7.3	15.3

Note: Latter 1980 gave three sets of values, based on three different statistical approaches. n.a.: not available

sidered are many, the variance among them is inflated. At the same time, on the other hand, as the number of groups increases, the boundaries between them tend to be ill-defined, which may lead to overestimating variation within groups.

The statistical methods used to apportion human diversity vary too, when one considers different studies. Nevertheless, the conceptual framework is the same, and is as follows. Several groups are defined on the basis of independent evidence. Each individual is compared to all other individuals studied, and a number representing the amount of genetic differentiation between them is calculated. Three components of genetic variance are then estimated from those numbers, namely: (a) between individuals of the same population; (b) among populations of the same group; (c) among major groups (see Lewontin 1972, Latter 1980, Barbujani et al. 1997). In this way, the overall genetic diversity is broken down in three hierarchical components.

2.1 Genetic Diversity at the Protein Level

In 1972, Lewontin first quantified the amount of genetic variation within and among major human groups. He chose a seven-race system, comprising Caucasians, Black Africans, Mongoloids, Amerinds, Oceanians (the five groups of Blumenbach's classification; see Table 1), plus South Asian and Australian aborigines. Analysis of 17 independent genetic loci, for which information on all available populations was collected, gave somewhat different results, but the average proportion of the overall variance that fell within populations was 85.4 percent (with a maximum

of 99.7 percent for the *Xm* locus and a minimum of 63.6 percent for the *Duffy* locus). Additional differences between populations of the same race represented, on average, 8.3 percent of the total, and differences between races accounted for the remaining 6.3 percent of the total. At the protein level, therefore, variation among seven races seems to represent less than one tenth of the overall genetic diversity of our species.

Lewontin's conclusion that the human racial classification has virtually no genetic or taxonomic significance was variously criticised, not always on scientific grounds. One serious objection was that only a fraction of the total DNA is actually translated into proteins. As a consequence, protein polymorphism may not fully reflect the underlying DNA diversity. Although successive studies of proteins confirmed that more than four fifths of the overall genetic diversity of our species is accounted for by differences among members of the same community (Table 2), skepticism remained widespread.

2.2 Genetic Diversity at the DNA Level

The first global study of DNA diversity, based on 109 genetic loci in 16 populations, appeared only 25 years later (Barbujani et al. 1997). The samples were chosen so as to form well-separated groups over the map of the world, whereas geographically intermediate populations were not considered. Therefore, if sampling affected the genetic variances estimated, it did so by increasing the weight of the between-continent components. In addition, this made the samples of each continent fall within one of the racial groups of

Blumenbach's classification, without ambiguity. The loci chosen were 30 short-tandem repeats (STR or microsatellite loci, whose alleles differ in length, that is, in the number of nucleotides they comprise), and 79 biallelic DNA polymorphisms (restriction-fragment length polymorphisms, or RFLP). For 16 RFLPs, compound genotypes referring to subsamples of individuals were available, and variance components were also estimated from them. By jointly analyzing many loci in this way, one may hope to identify significant differences better than by separately considering each gene.

The within-population variance component was significantly greater than 0 at all loci, representing between 54.4 percent and 94.6 percent of the total. Approximately one-third of the loci studied showed significant variance also between populations or between continents. Very similar results were obtained in the analysis of single RFLP polymorphisms, of the multilocus RFLP genotypes, and of the STR loci (Table 2).

Three recent papers provided additional data (Table 2) about genes inherited only from one parent. Every couple possesses four potentially transmissible copies of most chromosomes, two maternal and two paternal. But there is only one (paternal) copy of the Y chromosome; and there are several, presumably identical, copies of the mitochondrial DNA, which are only transmitted by the mother. Population genetics theory predicts that genes on these chromosomes should diverge four times as fast as the biparentally-transmitted (or autosomal) genes considered in the previous studies, all other demographic and evolutionary factors being equal. There is no simple model quantifying the effect of such a fast divergence on components of variance, but the differences among populations and continents (jointly estimated around 15 percent for autosomic genes) are expected to represent a substantially larger fraction of the genetic variance estimated from Y-chromosome and mitochondrial data.

Contrary to those expectations, mitochondrial variation appeared distributed much like autosomic variation. Such a lower-than-expected population diversity raises evolutionary problems that will not be discussed here, related with possible effects of natural selection on mitochondrial DNA, or with a high prehistoric female mobility. On the other hand, two analyses of the Y chromosome did show an increased differentiation among continents and populations, but confirmed that the largest component of human diversity is within populations (Poloni et al. 1997, Santos et al. 1999). However, in a third study of the Y chromosome, Seielstad et al. (1998) estimated even higher among-continent variances, exceeding 50 percent of the total. Note that the values estimated by Santos et al. (1999), which are closer to the other figures listed in Table 2, derive from the analysis of a different set of the same genetic markers, single-

nucleotide polymorphisms, that is, variable sites in the Y-chromosome sequence. The fact that such large differences have only been found in the male-transmitted portion of the genome, and only for one set of data, is not easy to interpret.

In synthesis, it seems that the within-population component of the overall genetic diversity of our species is close to 85 percent, regardless of the genetic markers and of the populations studied. Despite the limited polymorphism of the protein markers considered, and despite the choice of different sets of populations, Lewontin's results essentially are confirmed at the DNA level, for more than 300 genes analyzed so far. Greater among-continent diversities are estimated, as predicted by theory, on the basis of uniparentally-transmitted genes, and the allelic variants of the Y chromosome are more restricted geographically than the alleles of autosomal and mitochondrial loci. Among the plausible explanations of such a high Y-chromosome diversity are, again, selection differences in the males' and females' pre-historical migration rates, and a high variability of the males' reproductive success (discussed in Hammer et al. 1997). The conclusions of the other studies seem rather coherent, and they can be summarized as follows: if we arbitrarily set to 100 the average genetic difference between two very distant members of our species, the expected difference between members of the same community (who are not relatives, of course) seems to be approximately 85, and not 10 or 40, as previously illustrated definitions of race appear to imply.

3. *Are We All Equal, Then?*

The fact that genetic differences among major human groups are relatively small does not mean that all humans are genetically identical. In fact, the among-population and among-group components of variance are significantly greater than zero, not only for the Y chromosome, but also at several autosomic loci (Barbujani et al. 1997). That fraction of diversity, albeit small in relative terms, contains all the information that population geneticists use to reconstruct human evolutionary past. From its study, we know that both allele frequencies and DNA sequences vary nonrandomly in space, often forming gradients over entire continents (reviewed by Cavalli-Sforza et al. 1994). We also know that zones of sharp genetic change, or genetic boundaries, exist, and often overlap with various kinds of migrational barriers, both physical (mountain ranges, arms of sea) and cultural (especially language boundaries) (Sokal et al. 1990). However, very seldom do genetic boundaries form complete rings around single populations, or groups thereof. Over much of the planet, continuous variation is the rule, with the genetic characteristics of the various regions changing more or less smoothly in the

geographical space. Locating on the map the place at which one racial group ends and another one begins proved, so far, an impossible task. But nearly all populations studied are somewhat different from their neighbors.

4. Comparisons with Other Species

The subspecies concept is widely used in zoology and botany, and the suspicion may be raised that moral and political, rather than scientific, considerations may influence our willingness to use the same concept for humans. Although only a few authors have addressed this problem so far, their results suggest that humans are actually a comparatively uniform species. To cite just one example, the fruit fly *Drosophila pseudoobscura* shows a 10-fold greater nucleotide diversity than our species (Li and Sadler 1991).

Templeton (1999) compared Barbuji et al.'s (1997) estimates of genetic variances with those referring to 12 large-bodied, highly mobile mammals. Two such species (East African Buffalo and Kenyan Waterbuck) show smaller between-population diversity than humans. However, these are not cosmopolitan species like ours, but dwell in restricted geographical zones. All other species compared, including Elephant, Gazelle, and gray Wolf, show higher, or much higher, differentiation among populations and groups, with values of those variance components reaching 80 percent of the total.

On a smaller taxonomic scale, Ruvolo et al. (1994) compared mitochondrial diversity in the five genera of hominoids, that is, humans and apes. Variation appeared minimal within humans and pygmy Chimpanzees; the latter, once again, occupy a much smaller territory than humans. On the other hand, common Chimpanzee, Orangutan, and especially Gorilla, show much higher genetic differentiation. In particular, the most mitochondrially diverse humans studied appear to be less different than random pairs of lowland gorillas from West Africa.

5. Evolution of Human Genetic Diversity

Understanding human diversity also means to understand how that diversity came about. The low genetic variation among humans, as compared to other species, points to two, not incompatible, evolutionary processes, namely a high dispersal ability leading to limited isolation among populations, and/or a recent origin, too recent indeed for extensive genetic variation to have accumulated ever since.

Various evolutionary scenarios have been proposed to explain the available fossil, archaeological, and genetic evidence on human origins. In Coon's (1962) view, some 700,000 years ago, the population of our ancestor, *Homo erectus*, was subdivided geographi-

cally in five groups. Coon's five subspecies of *Homo sapiens* would differ because they are descended from those groups, with very little interbreeding, if any. This is no longer considered a valid scientific hypothesis, but the idea that the differences among contemporary humans started accumulating before the appearance of anatomically modern humans is still part of the so-called multiregional theory of human evolution (Wolpoff and Caspari 1997). This theory differs from Coon's in that the major human lineages are here regarded as only partly isolated, with extensive genetic exchange even between distant populations.

The basis of the multiregional theory of human evolution is the observation that certain traits persist in some regions of the earth from *H. erectus* times through *H. sapiens* times, and are unique to those regions. However, the statistical significance of that observation has been called into question (Lahr 1996), and it also seems that implausibly high levels of gene flow would be required, for hundreds of millennia, to maintain the unity of our species under a multiregional model (Rouhani 1989). The alternative hypothesis, the so-called out-of-Africa model (Stringer and Andrews 1988), regards the contemporary population as derived from ancestors who lived in Africa until recently. When those people expanded into Eurasia and Australia, perhaps 100,000 years ago, they replaced without admixture previous local populations of *Homo erectus* and (in Europe) *Homo sapiens neanderthalensis*.

The multiregional and the out-of-Africa models have several implications. If the main human groups separated before the appearance of anatomically modern *Homo sapiens*, deep differences are to be expected among them, and quite distinct evolutionary lineages should exist, in agreement with Vogel and Motulsky's (1986) definition of race. On the contrary, if all continental populations derived recently from the same ancestors, even distant human groups should share many alleles, and the differences among groups should be relatively small. The studies reviewed in this article seem to clearly support the latter model, although Templeton (1998) argued that high levels of migration could lead to low variances between continents, even under a multiregional model. Be that as it may, most investigators, including Templeton, agree that the available genetic evidence does not suggest the existence of distinct clusters of humans.

Support to this view also comes from studies which, although not explicitly addressing the issue of racial subdivision, are in evident agreement with a recent African origin of humankind. First, genetic diversity is greater in Africa than in any other continent (Armour et al. 1996, Hammer et al. 1997, Jorde et al. 1997), an observation that a multiregional model can hardly explain. Second, genetic diversity within populations decreases as one moves away from Africa, with near populations showing subsets of African genes, and more distant populations showing smaller subsets of

those genes (Tishkoff et al. 1996). Third, Neandertal man is considered a direct ancestor of the Europeans, and only of them, under the multiregional model; but the only Neandertalian DNA sequences obtained so far fall out of the range of current human variation, and bear no special resemblance to the European sequences (Krings et al. 1997). All these findings, and other paleontological data, emphasize the crucial role of Africa in human evolution (Foley 1998), and suggest that, during their dispersal in Eurasia, Australia, and the Americas, human groups spent too little time in isolation to significantly diverge genetically and form separate races.

6. Future Prospects

Several problems remain open. The first one has to do with the exact amount of genetic diversity that can be attributed to the various levels of population subdivision. Figures around 10 percent of the total diversity occurring among continents have been estimated by analyzing samples that were well separated in space, so that no ambiguity existed as for the continent each of them belonged to. However, in nature, populations are not so discretely distributed. May that sampling structure have led to overestimating the among-group component of genetic variance? And may the choice of the DNA markers studied (most of them falling in noncoding regions of the DNA) have affected the variance estimates too, perhaps in the opposite direction?

A second question is whether the apparently continuous distribution of genetic variation also implies that recognizable geographical groups do not exist, as has been argued by Templeton (1998). In fact, as previously mentioned, genetic variances among groups, although quantitatively small, are significantly greater than zero at many loci; this may have a meaning.

Some DNA alleles occur only in one specific population or continent. Individuals of unknown origin who carry those alleles can be assigned with high statistical confidence to the continent they belong to. However, it is not clear yet, and this is a third question, whether also broad human groups can be defined based on the presence of those alleles (rather than on the basis of most loci of the genome, as implied by traditional definitions of race); and it is unclear whether such groups, if any, could be called races. Most studies reviewed in this article do not support the idea that *the majority* of the members of our species can be put into a specific continental or racial group, on the basis of their genotype. This is a statistically tractable problem, and investigation on it is currently in progress. Good candidate genes for discriminating among groups could be the Y-chromosome single-nucleotide polymorphisms which, as mentioned, vary substantially across continents (Seielstad

et al. 1998). However, population groups defined on the basis of Y-chromosome diversity only could hardly be called races. A statement such as 'Human races exist, but only among males' would sound puzzling.

The genes determining morphological traits such as skull and eye shape, hair texture, or skin color, have not been identified yet, and some have argued that, once they will be known, a clearer identification of human groups will be possible. There are three reasons to doubt it, though. First, as shown in Table 1, centuries of studies of the phenotypes determined by those genes have not led to a consensus on the existing human races; whether and how a better knowledge of the underlying genes may clarify this picture is not obvious. Second, groups sharing traits typically used for racial classification, such as Melanesians and Africans, have been shown to be very distant genetically, both at the protein and at the DNA level (Cavalli-Sforza et al. 1994). Third, morphological traits are determined by the interaction between several genes and environmental factors. Theory shows that this polygenic and multifactorial inheritance results in a continuous distributions of traits, that is, the distributions which are less likely to show the sharp discontinuities necessary to discriminate among groups. Probably, morphological variation reflects differential selection by climate (Owens and King 1999), or the effects of sexual selection, both sometimes leading genetically distant groups to resemble each other.

Grouping individuals in rough categories based on morphology may still be useful for some practical purposes. In countries with extensive recent immigration, different groups may have different disease risks. Despite the low among-continent genetic variances discussed so far, associating a different risk to different, sometimes ill defined, groups (e.g., 'Hispanics') may simplify the clinicians' work. Similar problems also exist in forensic science, for DNA-based personal identification; and similar solutions, based on the definition of somewhat arbitrary racial or ethnic groups, have been put forward to reduce the risk of a wrong decision against a suspect (Nichols and Balding 1991). Despite these examples, genetic studies have shown that humans do not neatly fall within few well-distinct racial groups. Any population differs from its neighbors for several genetic features and, with appropriate sample sizes, these differences may achieve statistical significance. But if the existence of at least one significant genetic difference were the criterion for defining races, every village would be occupied by a different race.

In conclusion, newly discovered polymorphisms, especially those of the Y chromosome, and larger sampling efforts, may modify our ideas on the levels and patterns of human genetic diversity. But at present, clear discontinuities among major groups have not been detected, and there is no clear evidence of largely independent evolution of distinct human

lineages. In addition, nobody has yet identified sets of genes whose variation is correlated, so that a classification based on one of them will not be contradicted by the classification based on any other of them. Judging from the available genetic information, therefore, Vogel and Motulsky's (1986) definition of race does not seem to plausibly describe the structure of the human population. New definitions may be proposed, and, if so, they will have to stand the comparison with the data. At present, the burden of the proof seems on those who maintain that our species is subdivided into races that can be objectively recognized by studying their genes.

See also: Ethical Issues in the 'New' Genetics; Eugenics as the Basis of Population Policy; Gender, Class, Race, and Ethnicity, Social Construction of; Human Evolutionary Genetics; Human Genome Diversity and its Value; Population Policy: International; Y-chromosomes and Evolution

Bibliography

- Armour J A L, Anttinen T, May C A, Vega E E, Sajantila A, Kidd J R, Kidd K K, Bertranpetit J, Pääbo S, Jeffreys A 1996 Minisatellite diversity supports a recent African origin for modern humans. *Nature Genetics* **13**: 154–60
- Barbujani G, Excoffier L 1999 The history and geography of human genetic diversity. In: Stearns S (ed.) *Evolution in Health and Disease*. Oxford University Press, Oxford, UK, pp. 27–40
- Barbujani G, Magagni A, Minch E, Cavalli-Sforza L L 1997 An apportionment of human DNA diversity. *Proceedings of the National Academy of Sciences USA* **94**: 4516–9
- Cavalli-Sforza L L, Menozzi P, Piazza A 1994 *The History and Geography of Human Genes*. Princeton University Press, Princeton, NJ
- Cohen C 1991 Les races humaines en histoire des sciences. In: Hublin J J, Tillier A M (eds.) *Aux Origines d'Homo sapiens*. Presses Universitaires de France, Paris, pp. 9–47
- Coon C 1962 *The Origin of Races*. Knopf, New York
- Foley R 1998 The context of human genetic evolution. *Genome Research* **8**: 339–47
- Hammer M F, Spurdle A B, Karafet T, Bonner M R, Wood E T, Novelletto A, Malaspina P, Mitchell R J, Horai S, Jenkins T, Zegura S L 1997 The geographic distribution of human Y chromosome variation. *Genetics* **145**: 787–805
- Krings M, Stone A, Schmitz R W, Krainitzki H, Stoneking M, Pääbo S 1997 Neandertal DNA sequences and the origin of modern humans. *Cell* **90**: 19–30
- Jorde L B, Rogers A R, Bamshad M, Watkins W S, Krakowiak P, Sung S, Kere J, Harpending H 1997 Microsatellite diversity and the demographic history of modern humans. *Proceedings of the National Academy of Sciences USA* **94**: 3100–3
- Lahr M M 1996 *The Evolution of Human Diversity*. Cambridge University Press, Cambridge, UK
- Latter B D H 1980 Genetic differences within and between populations of the major human subgroups. *American Naturalist* **116**: 220–37
- Lewontin C 1972 The apportionment of human diversity. *Evolutionary Biology* **6**: 381–98
- Li W H, Sadler A 1991 Low nucleotide diversity in man. *Genetics* **129**: 513–23

- Livingstone F B 1964 On the nonexistence of human races. In: Montagu A (ed.) *The Concept of Race*. Free Press, Glencoe, NY, pp. 46–60
- Molnar S 1998 *Human Variation. Races, Types, and Ethnic Groups*. Prentice Hall, Upper Saddle River, NJ, p. 19
- Nichols R A, Balding D J 1991 Effects of population structure on DNA fingerprint analysis in forensic science. *Heredity* **66**: 297–302
- Owens K, King M C 1999 Genomic views of human history. *Science* **286**: 451–3
- Poloni E S, Semino O, Passarino G, Santachiara-Benerecetti A S, Dupanloup I, Langaney A, Excoffier L 1997 Human genetic affinities for Y-chromosome P49a,f/TaqI haplotypes show strong correspondence with linguistics. *American Journal of Human Genetics* **61**: 1015–35
- Rouhani S 1989 Molecular genetics and the pattern of human evolution: Plausible and implausible models. In: Mellars P, Stringer C B (eds.) *The Human Revolution: Behavioural and Biological perspectives on the Origins of Modern Humans*. Princeton University Press, Princeton, NJ, pp. 47–61
- Ruvolo M, Pan D, Zehr S, Goldberg T, Disotell T R, von Dornum M 1994 Gene trees and hominoid phylogeny. *Proceedings of the National Academy of Sciences USA* **91**: 8900–4
- Ryman N, Chakraborty R, Nei M 1983 Differences in the relative distribution of human gene diversity between electrophoretic and red and white cell antigen loci. *Human Heredity* **33**: 93–102
- Santos F R, Pandya A, Tyler-Smith C, Pena S D J, Schanfield M, Leonard W R, Osipova L, Crawford M H, Mitchell R J 1999 The Central Siberian origin for native American Y chromosomes. *American Journal of Human Genetics* **64**: 619–28
- Seielstad M T, Minch E, Cavalli-Sforza L L 1998 Genetic evidence for a higher female migration rate in humans. *Nature Genetics* **20**: 278–80
- Sokal R R, Oden N L, Legendre P, Fortin M J, Kim J Y, Thomson B A, Vaudor A, Harding R M, Barbujani G 1990 Genetics and language in European populations. *American Naturalist* **135**: 157–75
- Stringer C B, Andrews P 1988 Genetic and fossil evidence for the origin of modern humans. *Science* **239**: 1263–8
- Templeton A R 1998 Human races: A genetic and evolutionary perspective. *American Anthropologist* **100**: 632–50
- Tishkoff S A, Dietzsch E, Speed W, Pakstis A J, Kidd J R, Cheung K, Bonné-Tamir B, Santachiara-Benerecetti A S, Moral P, Krings M, Pääbo S, Watson E, Risch N, Jenkins T, Kidd K K 1996 Global patterns of linkage disequilibrium at the CD4 locus and modern human origins. *Science* **271**: 1380–7
- Vogel F, Motulsky A G 1986 *Human Genetics: Problems and Approaches*, 2nd edn. Springer-Verlag, Berlin, p. 534
- Wolpoff M H, Caspari R 1997 *Race and Human Evolution*. Simon & Schuster, New York

G. Barbujani

Race: History of the Concept

All societies resort to the use of genealogy. The genealogical myth is the first form of historical thought and metaphysical enquiry. The concept of race dom-

inated modes of thought when research into the origins of man developed in scientific terms. From this point of view, the classification of all living beings by the Swedish scientist Karl Linné (1707–78) was a significant event. Man, or *homo sapiens*, was classified according to species, gender, and class, as were other living beings. From this point onwards, interpretation of the history of humanity was considered in terms of race until the 1930s.

Since the 1930s, most biologists have criticized the notion of race, arguing that, even if phenotypic variations are genetically determined, there is no general classification of human beings according to the category of race. 'The concept of race has lost all its value' (Jacob 1981).

Biologists and social scientists employ the term in different ways. Biologists inquire whether humans are divided into racial categories and whether belonging to a particular race has effects on behavior. Social scientists are less interested in whether the idea of race is well founded or not and instead focus on the social consequences of the fact that man believes that racial groups exist. The main themes of their research are as follows: social relationships which are thought of in terms of race; the consequences of the belief that races exist on social relations and even on political organization as a whole. Although today race no longer has any meaning for biologists, the belief that there is a white race and a black race has had and still continues to have social consequences. Sociologists argue, according to Thomas' theorem, that if men refer to a situation or phenomenon as real, its consequences become real.

1. The Notion of Race Before the Science of the Races

The term race predates the emergence of racial theories claiming to be scientific. From the sixteenth century onwards, the British interpreted their history in these terms. They claimed that their political genius was due to their Saxon blood ties and the superiority of the Germanic race. So, according to them, William the Conqueror was seen to have imposed the Norman yoke upon free Anglo-Saxon peoples. If the parliament had asserted its rights against the Stuart Monarchy during the Glorious Revolution in 1688, then this was because the British were Germanic, in the way that Tacite had described them: free, asserting their liberty by limiting sovereign power (Banton 1977).

Thought in terms of race took up old representations and themes, in the name of science. The stigmatization of blacks already was current during the Antiquity period. According to a sixth-century statement attributed to Jesus: 'the children of Israel are among those people on the same level as Negroes....' The 'Purity of Blood-ties Statutes,' introduced in Spain after the *Reconquista* and the expulsion of

Jews and Muslims in 1492, established social categories based on 'blood' by distinguishing between pure-blooded Old Christians and the New Christians: Jews and Muslims who had converted to Catholicism, yet whose blood still remained 'impure.' A biological defect, unerasable even by baptism, was therefore passed on from generation to generation. This system of stigmatization did not, however, make any claims to scientific truth, which would become the case from the end of the eighteenth century onwards.

Racist thought developed more in English and German-speaking countries in particular. However, although thought in terms of race had less effect in France, the quarrel between two races, or in other words, the myth of two French peoples developed from the sixteenth century onwards. The noble classes claimed that the fact that they were of Germanic blood justified the freedom and privilege they enjoyed, the masses, being made up of descendants of the Gauls who had been defeated by the German invaders. Even Montesquieu, subscribing to this explanation, referred to the 'former Germans' as *our fathers*, and praised their admirable simplicity and traditions of liberty and independence.

2. The Century of Racial Ideas

Genealogical myth took on a new meaning during the modern period which saw the development of scientific ambition in fields such as physical anthropology and biology (Guillaumin 1972). Feelings and attitudes which had always been expressed towards certain stigmatized groups in the past were now reinterpreted and stamped with the scientifically legitimate trademark. Classic racist thought became fully established from the end of the eighteenth century onwards, as was the classification of living beings, which culminated in erasing the philosophical or religious distinctions between man, gifted with reason, responsibility, and thus liberty and other living beings.

This approach was revived with the development of social Darwinism. For the Darwin theorists, the decisive influence was biological and biology was defined by race. They believed that hostility between different races was natural and that the mixing of races led to the decline of the species concerned. They were obsessed by the need to preserve racial purity, fearing the dangers of intermarriage and its consequences, namely decline (Gobineau 1853). In order to define races, it became common to take measurements of the cranium and the angle of the nose, etc. It was scientifically claimed that blond *dolichocephalic* races were superior to brown-haired *brachycephalic* peoples.

Race gradually appeared to be the factor that explained all historical phenomena. According to Robert Knox (1791–1862), 'Human character, in-

dividual and national, is traceable solely to the nature of that race to which the individual or nation belongs' (Banton 1977). In France, Gustave Le Bon (1841–1931) wrote in 1894 that 'different races are unable to feel, think and act in the same manner, and subsequently are also unable to understand one another.' The revolutions of 1848, which can be seen as political events, *par excellence*, were explained by Gobineau in terms of race. The Aryan myth was used by the British in order to legitimise the colonisation of India. Walter Scott explained the conflict between the Scottish and the English by their different 'blood.' The concept of race was also used as an argument for all nationalist demands. These interpretations conformed to the scientism of the times and satisfied the human need for one sole explanation of the diversity and complexity encountered in the social world.

The word racism was first used at the end of the 1930s to refer to the doctrine which claimed that race determined culture. However, the theory predated the emergence of the term itself. This theory can be summarized by the two following claims: first, that human races exist; they are biologically different and therefore will always be unequal, and second, that there is a necessary link between biological characteristics and social behavior. In other words, the biological determines the social. There is an analytical distinction between race and racism. It is possible to argue that different races exist, without recourse to the biological determinism and essentializing ideas which characterize racist thought. Central to racist thought is the notion that races exist but above all that history and human behavior are determined by race. Some thinkers who believed that races did exist still criticized racist ideas. On the other hand, those with anti-Semitic views did not necessarily refer to the 'Jewish race' in justification of their hostility towards Jews. There are also those who do not believe in the existence of races but still conceive of social relations in a deterministic manner by explaining relations between people in terms of their culture, their class, or their gender. Nevertheless, it is equally true that thinking in terms of race has, more often than not, led to racist ideas.

The interpretation of historical events in terms of race has often been linked to the need to justify European domination over the rest of the world (Park and Burgess 1924). This is the case of Marxist thinkers in particular. Cox, for example, claims that race relations first emerged in 1493 when the borders showing the limits of Spanish and Portuguese control in the Americas were first marked out (Cox 1948). Yet the negative representations of black-skinned peoples in relation to whites by far predated the age of European expansionism. The development of racist thought cannot be linked to one sole cause but must be seen as the result of a number of converging factors. What is remarkable is the capacity of deterministic racist thought to renew and reinvent itself throughout the different periods of history.

3. Development of the Social Science Critique

Although the first American sociologists justified slavery, social scientists have, at different stages, criticized the concept of racial determinism. Max Weber clearly denounced racial explanations of historical events. 'That there exists today, only one precise and concrete factor, relevant to Sociology, which clearly and permanently reduces any given sociological category down to innate or hereditary characteristics, whilst claiming that another category cannot ever not possess those same qualities, I categorically deny, and I will continue to deny this until it is proven right before my eyes' (Weber 1974, p. 120).

The work of the Chicago school sociologists (such as Charles S. Johnson, John Dollard, Allison Davies, E. Franklin, and even Gunnar Myrdal whose work was based on the Chicago school's research) marked another stage in the social science critique. Through study of the 'assimilation' of various populations in American society, they replaced the notion of race with that of *relations* between races or *race relations*. Race, in the biological sense of the term, was replaced by a social definition: race became a social category which contributed to defining an individual's position in society, according to his origins which were inferred by certain biological characteristics. However, it is clear that this category is only acknowledged if society accepts that 'race' is meaningful. In this way, it was henceforth taken as given by all sociologists that it is racism that creates race, not vice versa.

The rediscovery of the weight of ethnicity by American sociologists in the 1960s led to a redefinition of different groups in terms of 'ethnic groups': an historical definition thus replaced a biological definition. However, these sociologists did not all unanimously agree that Afro-Americans or the native-Americans constituted a group like any other. It was argued that the term 'ethnic group' was perhaps no more than just a euphemism or hypocrisy as it tended to obscure the stigmatization, segregation, and discrimination which Afro-Americans and native-Americans had suffered as a result of their race. Wasn't 'ethnic' anything else than 'racial'? Shouldn't race relations be distinguished from ethnic relations, as individuals could abandon their ethnic group but could not escape their racial group?

The sociologists of *race relations* continue, however, to use the concept of race as though races do actually exist, and the sociologists of ethnicity as though ethnic groups are static, fixed entities. When, from the 1960s onwards, the term *ethnic group* took over from *race*, it continued to imply a certain hereditary link. Furthermore, both terms are closely linked to the American experience, a society made up of immigrants (ethnic groups), autochthonous peoples, and those brought over as slaves (racial groups). Should we not take into account the consequences of the critique of the sociology of race relations and stop claiming that

the term race is meaningless, only to then use it continuously, even in the titles of scholarly journals? Is it not possible to reach a new stage by adopting the concept of *historical collectivity*? If, as Weber had already observed, those who we refer to as Afro-Americans in the United States make up a particular group, this is not due to the color of their skin, but rather because they consider themselves and are considered by others to be the descendants of Africans who were subjected to a certain historical experience, which makes them a particular *historical collectivity* (Schnapper 1998).

Societies use natural differences and relations as a way to organize relationships between people. This is the case for gender and relationships between parents and children. In the same way, human beings use the physical differences, thought of in terms of race, in order to establish and identify social roles. Sociological thought has progressively criticized the notion of race. The relations between the 'races' are now analyzed as part of the processes of inclusion and exclusion of the different social groups in a more general analysis of social structures and political legitimacy.

See also: Critical Race Theory; Genetics and Mate Choice; Race and the Law; Race Identity

Bibliography

- Banton M P 1977 *The Idea of Race*. Tavistock, London
 Cox O C 1948 *Caste, Class & Race. A Study in social Dynamics*. Doubleday, New York
 Gobineau J A 1853 *Essai sur l'inégalité des races*. Gallimard, La Pléiade, Paris
 Guillaumin C 1972 *L'idéologie raciste. Genèse et langage actuel*. Mouton, Paris and La Haye
 Jacob A 1981 La science face au racisme. In: Olender M (eds.) *Le racisme, Mythes et sciences*. Complexe, Brussels
 Park R E, Burgess E W 1924 *Introduction to the Science of Sociology*. University of Chicago Press, Chicago
 Schnapper D 1998 *La Relation à l'autre. Au coeur de la pensée sociologique*. Gallimard, Paris
 Weber M 1974 Max Weber et les théories bioraciales du XXème siècle. *Cahiers Internationaux de Sociologie* 1: 115–26

D. Schnapper

Race Identity

The concept of 'racial identity' has generated many and different bodies of literature that bear directly upon its multiple deployments in the social and behavioral sciences. Given that every individual is universally and simultaneously also part of a larger group, from family to community to language group

to (for most of the last three centuries) nation–state, the salient question for the social sciences is: under what conditions do individuals come to think of themselves as 'racial'? The corollary, but by no means symmetrical question is: under what conditions do larger social units, from nation–states to communities, come to see, classify, and identify the individual as 'racial'?

1. Racial Identity and Individual Identity

There is a vast literature on the psychology of personal identity that provides a particular angle for approach to the topic. Identity formation as a developmental process through adolescence is one significant track, and a subset of this work is early-stage racial identity formation. This research can be traced back more than a half-century to the studies of the preferences of preschool children for white or black dolls (Clark and Clark 1947). A parallel track of scholarship emphasizes the power of 'group identity' to shape behavior through the imposition of cultural and national identities (Burman and Harrell Bond 1979, Shibutani and Kwan 1969). From political mobilization to choice of a marital partner, each of these renderings of racial identity has substantially penetrated the disciplines of sociology, political science, anthropology, and social psychology. In the USA, the courts became the site of the connection between the vastly different worlds of research on racial identity formation and social and political institutions. The Supreme Court's 1954 decision on school desegregation was based partly upon research claiming that lowered self esteem of the subordinated group was 'inherently a function of racial segregation' (Brown vs. Board of Education), and explicitly referred to the doll studies noted above. This generated a research trajectory involving studies on individual responses to racial identification that would extend through the remainder of the twentieth century (Montalvo 1999).

While the concept of individual identity has no clearly locatable beginning, it has been traced, in the West, to 'the discovery of the individual' (Morris 1972, Doi 1985). Individuals typically identified with their families, then perhaps with the local village or community (Linton 1955). As communities grew and members became mobile, region, language group, religion, and the nation-state would each claim some measure of allegiance via identity. Multiplicity of identities for a single individual can also be understood in terms of the simultaneity of those identities. That is, a single individual can sense and affirm a complex set of identities, being simultaneously a grandfather and being Jewish, being French and being working-class, being an alcoholic and a chess player. No single identity need claim any essentialist primacy nor be privileged as 'fundamental' across time, space, and culture. The social context frames which identity will

be called up and acted upon, or be the reference point of further action for the individual.

The mobilization of constituencies around a perceived collective interest is the essence of political and social movements, and in this sense all political mobilization is 'identity politics.' In a work that would set the frame for later research on racial identity and political consciousness, E. P. Thompson's (1963) monumental social history of the 'making of the English working class' detailed how class identity is forged from a particular set of economic and social circumstances. In a different and parallel kind of scholarship, other researchers have sought to explain the emergence of a racial identity. For example, the US civil rights movement (1956–72) produced a number of scholarly analyses that demonstrated how different sets of circumstances explain the emergence and sustenance of racial identity (Hampton and Fayer 1990, Bloom 1987, Morris 1984). However, the most direct descendant of Thompson's (1963) demonstration of the historical specificity of political mobilization around identity, relevant to race, is David Roediger's *The Wages of Whiteness: Race and the Making of the American Working Class* (1991). Roediger mines the archives of the US labor movement of the nineteenth century to provide rich historical material showing how the racial identity of whites in general, and the Irish in particular, were formed in relationship to the enslavement of blacks.

2. *Politics and Identity Politics*

Thus racial identity was generated by massive historical, economic, and political forces that far transcend the level of individual choice (Banton 1977, Miles 1989). Indeed, there is a substantial body of research indicating that in the nineteenth century European and US academics expended much energy trying to demonstrate the intellectual and moral superiority of whites (Gould 1981, Bernal 1987). The biological and medical sciences of the late nineteenth century assumed a Darwinian model of racial hierarchy of adaptation to civilization (Frazer 1900). Thus, in order to counter the assertion of white superiority, from Europe to South Africa to the North American continent to Australia, groups that wished to mobilize to change their position in a racial hierarchy often found that racial identity was a useful and powerful basis for achieving group solidarity (Kelley 1997, Fredrickson 1995, Crenshaw et al. 1995). Obviously, these groups would encounter competing group identities, such as tribal, regional, ethnic, linguistic, and class. This list would permit the white colonialists the advantage of playing one off against the other, and the strategic use of ethnic identity to trump 'race' could be deployed. For an elaboration of the distinction between race and ethnicity, see Fenton (1999).

If we take even a casual excursion through the last few centuries of racial classification, there is overwhelming evidence on the side of those who have argued that race is arbitrary, shifting, and often biologically and sociologically inconsistent and contradictory. A consortium of leading scientists across the disciplines from biology to physical anthropology issued a 'Revised UNESCO Statement on Race' in 1995—a definitive declaration that summarizes 11 central issues, and concludes that in terms of 'scientific' discourse, there is no such thing as a 'race' that has any scientific utility: '... the same scientific groups that developed the biological concept over the last century have now concluded that its use for characterizing human populations is so flawed that it is no longer a scientifically valid concept. In fact, the statement makes clear that the biological concept of race as applied to humans has no legitimate place in biological science' (Katz 1995, pp. 4, 5).

The rule that one drop of black blood makes one black is the easy mark along a full continuum of contradictory and incoherent taxonomies. 'Passing' and 'miscegenation laws' and slaveowner/slave offspring do more than simply dot the landscape of this topic (Harris 1995, Haney-López 1996). This continuum extends well into the present period where we find more and more people asserting a mixed-race identity. Since the classification of race is arbitrary and often whimsical (e.g., one drop of blood), accepting the idea that race is something identifiable with fixed borders that could be crossed and thus 'mixed' is just another indication of the power of social forces shaping racial identity. At the biochemical level of blood types and hematology, at the neurological level of neurotransmission patterns, at the level of cell function—at all these levels, we are all 'mixed' by any taxonomy or measure of allele frequencies in large population groups (Molnar 1992).

Yet a number of scholars of race have noted the increasing use of the category of 'mixed race' as the self-identity of a younger generation, mainly of relatively privileged and middle-class persons. Paul Spickard (1989) and Werner Sollers (1997) have each produced important studies of 'mixed race' identities that cross boundaries of Europe and the Americas. Maria Root (1996) has culled a collection of essays that deal with identity construction along ethnic and racial lines, and further explores these issues well beyond the black and white focus that has riveted much of this literature. The difference between 'individual choice' of a racial identity as opposed to having a racial identity 'imposed' by the state and/or the obdurate empirical reality of a group consensus (based primarily upon phenotype-stereotype) is profound, yet variable over time and place. There are concrete economic, political, social, and sometimes psychological benefits to choosing certain identities. That is why scores of thousands of light-skinned 'blacks' under slavery in Brazil, Jamaica, and the USA

chose to 'pass' as white (Small 1994). For parallel reasons, thousands of persons in the USA have attempted to reclassify themselves as Native American or Latino or African American to benefit from affirmative action laws.

For an extensive theoretical treatment of the way in which racial taxonomies shift over time, see Omi and Winant (1986). In addition, for specific cases of this phenomenon, see Harris (1995), Haney-López (1996) and Takaki (1998). From the other side, namely, individuals who come to collectively redefine their own situation, and thereby reveal how 'race' is sometimes in flux due as much to agency, interpretation, and circumstance, see Blumer (1958), Lal (1990), and Perry (1998). For an 'interest-driven' version of the collective will to create, sustain, or maintain a particular racial identity, see Lipsitz (1998) and Wellman (1993).

3. White Identity

In recent years, the concept of white racial identity has come under increasing focus as a basis for a research agenda. Blumer's (1958) classic article reframed race as a structural relationship between competing group interests. This position was adopted and developed by an early entry into this field, David Wellman's *Portraits of White Racism* (1993). Roediger (1991, 1994), Frankenberg (1993), Allen (1994), Jacobson (1998), Lipsitz (1998), and Perry (1998) have each examined the ways in which white privilege is embedded in a faceless and taken-for-granted normalness of white identity. Twine and Warren (2000) have assembled an anthology in which they bring to greater consciousness the ways in which researchers in racially stratified societies ignore at their peril the subject world of race as an integral feature of methodology, largely through the racial identity of subjects and researchers.

4. Which Identity Gets Triggered?

Given that there are multiple and simultaneous identities, the task for the social analyst is to try to understand the social circumstances in which a particular set of identities gets 'triggered' or brought to the fore. That is, rather than arguing for the primacy or 'essentialist' character of an identity, the analytic task is to understand why identity A, B, or C emerges as primary. War between nation-states is an especially propitious time for a 'national identity' to be forged, or come to the fore. Thus, it may be heuristically useful to think of a series of latent multiple identities that get turned on or off based upon social and political circumstances—and determine 'which button gets pushed.' In wartime, the national identity button gets pushed frequently, and lights up. In courtship, the

gender button is pushed, and so forth. In these terms, the empirical question of when and why the 'racial identity' is triggered is no less compelling when posed as a theoretical question.

We can expect the next period to generate many more investigations of racial identities deployed as one among multiple identities—moving away from static conceptions of fixed boundaries of racial identity, with richer explorations of its contingent and sometimes fluid character. An example of this kind of work at the vanguard would be Gloria Anzaldúa's (1987) *Borderlands/La Frontera: The New Mestiza* and George Lipsitz' (1991) *Collective Memory and American Popular Culture*. These authors identify 'border identities,' which are emergent identities crafted and recrafted continually by movements across different borders. As the new millennium witnesses the increasing thousands and even millions of border crossings (and recrossings) with the movement of capital and labor, studies of 'borderland identities' around issues of race and ethnicity will come to play a significant role in this field of inquiry.

See also: Civil Rights Movement, The; Identity: Social; Integration: Social; Racial Relations; Residential Segregation: Sociological Aspects

Bibliography

- Allen T 1994 *The Invention of the White Race*. Verso, New York
- Anzaldúa G 1987 *Borderlands/La Frontera: The New Mestiza*. Spinsters/Aunte Lute Press, San Francisco
- Banton M 1977 *The Idea of Race*. Tavistock, London
- Bernal M 1987 *Black Athena: The Afroasiatic Roots of Classical Civilization*. Rutgers University Press, New Brunswick, NJ
- Bloom J M 1987 *Class, Race, and the Civil Rights Movement*. Indiana University Press, Bloomington, IN
- Blumer H 1958 Race prejudice as a sense of group position. *Pacific Sociological Review* 1(Spring): 3–6
- Burman S B, Harrell-Bond B E 1979 *The Imposition of Law*. Academic Press, New York
- Clark K B, Clark M P 1947 Racial identification and preference in Negro children. In: Newcomb T M, Hartley E (eds.) *Readings in Social Psychology*. Henry Holt, New York, 602–11
- Crenshaw K, Gotanda N, Peller G, Thomas K 1995 *Critical Race Theory*. The Free Press, New York. pp. 609–611
- Doi T 1985 *The Anatomy of Self: The Individual versus Society*. Kodansha International, Tokyo
- Fenton S 1999 *Ethnicity: Racism, Class and Culture*. Rowman & Littlefield, London
- Frankenberg R 1993 *White Women, Race Matters: The Social Construction of Whiteness*. University of Minnesota Press, Minneapolis, MN
- Frazer J 1900 *The Golden Bough: A Study in Magic and Religion*, 2nd edn. Macmillan, New York
- Fredrickson G M 1995 *Black Liberation: A Comparative History of Black Ideologies in the United States and South Africa*. Oxford University Press, New York
- Gould S J 1981 *The Mismeasure of Man*. W. W. Norton, New York

- Hall S 1991 Ethnicity: Identity and difference. *Radical America* 23(4): 9–19
- Hampton H, Fayer S (eds.) 1990 *Voices of Freedom: An Oral History of the Civil Rights Movement from the 1950s Through the 1980s*. Bantam Books, New York
- Haney-López I 1996 *White By Law: The Legal Construction of Race*. New York University Press, New York
- Harris C I 1995 Whiteness as property. In: Crenshaw K, Gotanda N, Peller G, Thomas K 1995 (eds.) *Critical Race Theory*. The Free Press, New York, pp. 276–91
- Ifekwunigwe J O 1999 *Scattered Belongings. Cultural Paradoxes of 'Race,' Nation and Gender*. Routledge, New York
- Jacobson M F 1998 *Whiteness of a Different Color: European Immigrants and the Alchemy of Race*. Harvard University Press, Cambridge, MA
- Jordan W D 1968 *White Over Black, White Over Black: American Attitudes Toward the Negro, 1550–1912*. University of North Carolina Press, Chapel Hill, NC
- Katz S H 1995 Is race a legitimate concept for science? *The AAPA Revised Statement on Race: A Brief Analysis and Commentary*. University of Pennsylvania, February
- Kelley R D G 1997 Identity politics and class struggle. *New Politics* VI(2): 84–97
- Lal B B 1990 *The Romance of Culture in an Urban Civilization: Robert E. Park on Race and Ethnic Relations in Cities*. Routledge, New York
- Linton R 1955 *The Tree of Culture*. Alfred A. Knopf, New York
- Lipsitz G 1991 *Time Passages: Collective Memory and American Popular Culture*. University of Minnesota Press, Minneapolis, MN
- Lipsitz G 1998 *The Possessive Investment in Whiteness: How White People Profit from Identity Politics*. Temple University Press, Philadelphia, PA
- Miles R 1989 *Racism*. Routledge, London
- Molnar S 1992 *Human Variation: Races, Types, and Ethnic Groups*, 3rd edn. Prentice Hall, Englewood Cliffs, NJ
- Montalvo F F 1999 The critical incident interview and ethno-racial identity. *Journal of Multicultural Social Work* 7: 1–43
- Morris A D 1984 *The Origins of The Civil Rights Movement: Black Communities Organizing for Change*. Free Press, Collier Macmillan, New York
- Morris C 1972 *The Discovery of the Individual, 1050–1200*. Harper & Row, New York
- Omi M, Winant H 1986 *Racial Formation in the United States: From the 1960s to the 1980s*. Routledge & Kegan Paul, New York
- Perry P 1998 Beginning to see the white. Unpublished PhD thesis, University of California, Berkeley
- Roediger D R 1991 *The Wages of Whiteness, Race and the Making of the American Working Class*. Verso, New York
- Roediger D R 1994 *Towards the Abolition of Whiteness: Essays on Race, Politics, and Working Class*. Verso, New York
- Root M P P 1996 *The Multiracial Experience: Racial Borders as the New Frontier*. Sage, Thousand Oaks, CA
- Shibutani T, Kwan K W 1969 *Ethnic Stratification: A Comparative Approach*. Macmillan, London
- Small S 1994 *Racialised Barriers: The Black Experience in England and the United States in the 1980s*. Routledge, New York
- Smedley A 1993 *Race in North America: Origin and Evolution of a Worldview*. Westview Press, Boulder, CO
- Sollers W 1997 *Neither Black Nor White, Yet Both: Thematic Explorations of Interracial Literature*. Oxford University Press, New York

- Spickard P 1989 *Mixed Blood: Intermarriage and Ethnic Identity in Twentieth-century America*. University of Wisconsin Press, Madison, WI
- Takaki R T 1998 *Strangers From a Different Shore: A History of Asian Americans*. Little Brown, Boston
- Thompson E P 1963 *The Making of the English Working Class*. Vintage Books, New York
- Twine F W, Warren J (eds.) 2000 *Racing Research, Researching Race*. New York University Press, New York
- Wellman D T 1993 *Portraits of White Racism*, 2nd edn. Cambridge University Press, New York

T. Duster

Race Relations in the United States, Politics of

The problem of race has been very badly handled in political science. Except for Marxists who see racism and capitalism as intertwined, postmodern skeptics, and those like V. O. Key Jr., who studied the South, democratic theorists have not systematically addressed the racial factor in American politics (McClain and Garcia 1993, Dawson and Wilson 1991). Yet although the problems of racism, racial discrimination, and inequality have scarcely been resolved, many scholars of this current post-civil rights period have continued to ignore the profound and dynamic relationship between race and American politics. A large part of their failure to address the problem of race in America stems from the pervasiveness of a 'gradualist' liberal bias that views the race problem as ultimately resolving itself through integration and assimilation of racial minorities as well as through the disappearance of White racism. From Myrdal (1944) to Thernstrom and Thernstrom (1997), widely disseminated and important treatments of race relations in the USA remain rooted in this gradualist bias. Some scholars, however, recognize that race and ethnicity remain central features of a rational and modernizing society, and have begun to explore the political foundations of racial conflict and racism. The latest research is expanding the question of race to include a focus on groups other than the politics of American Blacks and Whites.

1. Black Politics

The dominant theories put forth by leading American democratic theorists following World War II obviously excluded African-Americans in broadly characterizing the American political process as one at its core, where no one group is excluded and all, although hardly equals, can be heard. Even after such theories were later revised to portray the inclusion of Blacks as

an inevitability, because as Robert Dahl explained, 'democratic cultures have considerable capacity for correcting their own failures' (1989, p. 180), what politics would look like after their inclusion was poorly foreshadowed and misunderstood by such scholars. Democratic theorists like Dahl (1961) and David Truman (1971) generally assumed that the political incorporation of Blacks would 'normalize' American politics, with all the various groups and subgroupings able to engage in pressure politics. Browning et al. (1984) in an important book, would argue that eventually minorities were integrated and empowered in the American pluralist system (Eisenger 1982, Sigelman and Welch 1982, Sonenshein 1993). These same authors and others, would later critically re-evaluate their claim of 'empowerment' for American minorities (Hero 1992, Browning et al. 1991). But others scholars, including Adolph Reed Jr. (1988) would seriously question whether the pluralist model could ever be refashioned to fit American Blacks, since to do so would be to ignore the 'racial element' in American politics (see also Holden 1973, Jones 1972, McLemore 1972, Morris 1975, Pinderhughes 1987).

Given its exclusion in the discipline, research on the politics of Blacks would constitute a separate inquiry, apart from the mainstream (McClain and Garcia 1993). Research on Black politics would not only establish the continuing significance of race in American politics (Carmichael and Hamilton 1967, Jones 1972), but also the fact that politics as practiced, encompasses many more activities than recognized by mainstream scholars. Social movement scholars established how protest is not unlike electoral politics (McAdam 1982, Morris 1984) and the continuity between the two in the postcivil rights era (Holden 1973, Smith 1981, Tate 1993, Walton 1973). Jesse Jackson's 1984 and 1988 presidential bids clearly reflected the continuity (Marable 1985, 1990). Religion and social context, important to the emergence of the civil right movement, remain central to Black electoral politics (Cohen and Dawson 1993, Harris 1994). The racial identities of Blacks, a factor in the movement, continues to play a critical role in their political participation and attitudes today (Dawson 1994, Miller et al. 1981, Tate 1993, Shingles 1981, Verba and Nie 1972).

Nevertheless, William J. Wilson's (1981) controversial thesis that race is receding in importance relative to class in the Black community, lay the seeds for a new controversy in Black politics over the degree to which race matters. Carol M. Swain's (1993) award-winning book would question the relevance of race in the political representation of Blacks; others would contest it (Cannon 1999, Whitby 1997) or offer qualifications of it (Lublin 1996). While this debate continues, the study of the politics of Blacks which was generally the only way political scientists addressed the question of race has actually reshaped the contours

and direction of mainstream work in political science. New attention is now paid to the nonelectoral antecedents of politics (e.g., Verba et al. 1995) and on the relationship between the institution and political behavior (e.g., Rosenstone and Hansen 1993) as others insisted was critical in the study of Black politics (e.g., Barker et al. 1998, Walton 1985).

2. The Debate over Racial Prejudice

Although the bulk of political science has ignored race, there exists nevertheless a voluminous literature devoted to the empirical study of how racial attitudes influence the policy attitudes of Americans. Nearly all of this empirical work has focused on Whites, although some scholars have addressed the racial and policy attitudes of Blacks (Hochschild 1995, Sigelman and Welch 1991). Broadly, two lines of argument have emerged in the study of White racial attitudes. Those on one side of the debate argue that racial prejudice remains the fundamental factor behind Whites' opposition to racial policies such as busing school children to achieve racially balanced public schools, and affirmative action. White racism has not disappeared, as Whites' responses to the general principle of racial equality items imply. For some scholars, education has merely given more White Americans the ability to appear more racially tolerant than they really are (Jackman and Muha 1984). Rather, racial intolerance along with fundamental American values about the work ethic and individualism blended together, explains the negative attitudes that Whites have toward government efforts to eradicate racial inequality (Kinder 1986, Kinder and Sanders 1996, Kinder and Sears 1981, McConahay 1986, McConahay and Hough 1976, Sears 1988, Sears et al. 2000). Using their defense of traditional values as a cover for their resentment and fears of Blacks, Whites strongly object to policies that would change the racial status quo.

Those on the other side of the debate have sought to account for White opposition to racial policies differently, choosing to focus on factors other than racial prejudice, and not only because they provide a more flattering portrayal of White Americans. For those seeking alternative explanations, ideology and moralistic values and beliefs are posited to be the major reasons underlying White opposition to racial implementation policies. Politically conservative Whites object to such policies because of their dislike of 'big government.' Kuklinski and Parent 1981, Margolis and Hague 1981, Sniderman et al. 1984, 1991). In addition, these scholars point to the 'natural contradiction' between the White majority's commitment to individual rights and freedoms, its preference for a limited government, and support for new government programs that would favor Blacks (Carmines and Merriman 1993, Sniderman and Hagen 1985). Pro-

grams that are designed to promote racially equal results at once contradict the very values, including color-blindness, that most Whites now espouse (Sniderman and Piazza 1993).

Along these lines, other scholars have sought to link White racial attitudes to their beliefs about the causes of racial inequality and racial stratification. Several scholars have shown that Whites will not support programs that assist Blacks if they believe the system to be fair and that Blacks themselves are not working hard enough to get ahead (Kluegel and Smith 1986, 1986, Sniderman and Hagen 1985). While civil rights organizations have asserted that Blacks are denied equal opportunities to Whites to get ahead, most White Americans express full faith in the abundance of opportunity in American society. Their faith contrasts greatly with that of Blacks (Kluegel and Bobo 1993, Kluegel and Smith 1986, pp. 67–68, Sigelman and Welch 1991). In the 1990 General Social Survey, 72 percent of Blacks, but only 37 percent of Whites, thought that racial discrimination was a main factor in the unequal outcomes for Blacks in jobs, income, and housing. Given that most Whites believe that Blacks currently have equal opportunities to Whites and that the economic system works well, most oppose programs that demand equality of outcomes.

As one of the more radical theoretical shifts, the meaning of racial prejudice has changed (see Sniderman et al. 1993, for a similar discussion). Once considered to be an archaic, irrational need located deep in a person's heart to scapegoat others, research has consistently favored another view, that racial bias is ubiquitous, forming naturally in a person's mind through social group interactions (Nisbett and Ross 1980). The new definition of prejudice as an instinctive and widespread phenomenon can be interpreted in two ways. On the one hand, the very ubiquitousness of racial stereotypes undermines their relevance in the study of White racial attitudes. Because stereotypes are so widespread, Whites' opposition to racial policies must more critically depend on other factors, such as politics, ideologies, and the framing of the issue. On the other hand, the ubiquitous nature of racial stereotypes exposes its foundation in the social stratification system or a racial hierarchy and in history. Scholars working in sociology and social psychology argue that racism, new or old, persists because it justifies and extends one group's domination over another (Sidanius and Pratto 1993, Sidanius et al. 1996).

3. The New Institutionalism in Racial Politics

There is a new and different approach to the study of race and politics that I classify under the heading of the 'new institutionalism in racial politics.' The core thesis is that politics is as much an underlying cause of America's racial cleavages as it gives expression to

such cleavages. Although consistent with a growing branch of political science that argues that institutions structure political behavior, not only political outcomes (March and Olsen 1984), to argue that politics is structurally implicated in America's racial divisions still represents an important break from the intellectual tradition in political science.

American federalism was a pragmatic response to the contentious issue of slavery, but greatly prolonged the institution of slavery and the subjugation of American Blacks. States would use their constitutional powers to deny citizenship to Mexicans and Asian immigrants, as well as deny Blacks their new right to vote under the Fourteenth Amendment. An entire subfield exists that examines the ways that electoral arrangements and practices negate the Black vote (e.g., Davidson and Grofman 1994, Grofman et al., 1992, Kousser, 1999).

New research on racial politics and voting has continued to identify factors other than the racial attitudes of Whites in their rejection of Black candidates and their defection from the Democratic party (see Giles and Buckner 1993, Giles and Evans 1986, Giles and Hertz 1994, Glaser 1994, Huckfeldt and Kohfeldt 1989, Reeves 1997). Racial prejudice is not deemed unimportant or irrelevant in these new studies, but the racial behavior of Whites is conceptualized as more situational-based than attitudinally-based. Candidates, political parties, and the media can make race the focal point of the election generally, but not always, to the detriment of the Black candidate (Lublin and Tate 1995, Metz and Tate 1995, Reeves 1997).

Finally, while sociologists and social psychologists contend that racial categories are socially constructed, some political scientists (Smith 1997) have begun to show how in the USA racial identities have been politically constructed through US laws covering immigration, state laws that defined race, the US census, and affirmative action guidelines along with legal and political discourse. Comparative race studies have emerged, underscoring the role that state policies play in defining race (Gilroy 1991, Marx 1998).

4. Beyond Black and White

Finally, most of the research on racial prejudice and racial politics has been conceptualized and tested on Whites and Blacks. Researchers have begun to recognize that even race relations defined as narrowly as Black–White relations is critically related to status and politics of other American minorities as well, notably America's fastest-growing groups, Asian-Americans and Latinos (Bedolla 2000, Cain et al. 1991, De la Garza et al. 1992, DeSipio 1996, Espiritu 1992, Garcia 1997, Hardy-Fanta 1993, Hero 1992, Jennings 1994, Jennings and Rivera 1984, Saito 1998, Uhlaner et al. 1989, Wrinkle et al. 1996). Whether these groups suffer discrimination on the basis of their nationality, and

the degree to which national origin affects their politics is often the unstated subtext of in these body in work (but see Kim, 2000). In response to recent highly publicized conflicts between Blacks, Latinos, and Koreans in cities such as Los Angeles and New York, much of this work addresses interminority conflict (Abelmann and Lie 1995, Kim, in press, McClain 1993, McClain and Karnig 1990) Finally, new work has explored not just interracial or interethnic politics, but intraminority group politics. The limits of treating African American and Latino politics in 'gender-and-sex neutral' terms, for example, have been exposed in Cathy Cohen's (1999) study of politics of AIDS in the Black community and Carol Hardy-Fanta's (1993) work on community organizing among Latinos. Moving beyond the Black-White dichotomy is critical for reaching a better understanding of not only the politics of the neglected 'other' minorities, but for Black politics and US politics more broadly.

While research on race exists and grows, it still remains a neglected focus in political science. The 1990s scholarship I have presented clearly establishes that race does not lie on a progressive-regressive continuum as the gradualist assumption posits. This idea that we are headed in the direction of a color-blind society remains extremely popular to scientists and nonscientists alike. Nevertheless, ample social science research in social and cognitive psychology shows that people routinely discriminate, that racial stereotypes persist (Nisbett and Ross 1980). Race therefore remains a dynamic feature of American politics, whose complexities we have only just begun to understand and appreciate. At best, one can talk about minimizing its significance. It is vital that we understand how politics is implicated in the persistence of racial divisions if we are to have a true democracy organized in a way that is truly inclusive, working for the common good of us all, not just for those at the top of the racial order. To do this, political scientists must pay greater and careful attention to the problem of race than they have in the past.

See also: Civil Rights Movement, The; Ethnic and Racial Social Movements; Ethnic Conflict, Geography of; Ethnic Conflicts; Ethnic Conflicts and Ancient Hatreds: Cultural Concerns; Political Protest and Civil Disobedience; Race Identity; Racial Relations; Racism, History of; Racism, Sociology of; Residential Segregation: Geographic Aspects; Residential Segregation: Sociological Aspects; Urban Politics: United States; Urban Poverty in Neighborhoods

Bibliography

Abelmann N, Lie J 1955 *Blue Dreams: Korean Americans and the Los Angeles Riots*. Harvard University Press, Cambridge, MA

- Barker L J, Jones M H, Tate K 1998 *African Americans and the American Political System*, 4th edn. Prentice Hall, Englewood Cliffs, NJ
- Bedolla L G 2000 They and we: Identity, gender, and politics among Latino youth in Los Angeles. *Social Science Quarterly* **81**: 106-22
- Bobo L 1983 Whites' opposition to busing: Symbolic racism or realistic group conflict? *Journal of Personality and Social Psychology* **45**: 1196-210
- Bobo L 1986 Group conflict, prejudice, and the paradox of contemporary racial attitudes. In: Katz P A, Taylor D A (eds.) *Eliminating Racism: Means and Controversies*. Plenum Press, New York
- Browning R P, Marshall D R, Tabb D H 1984 *Protest is Not Enough*. University of California Press, Los Angeles and Berkeley
- Browning R P, Marshall D R, Tabb D H (eds.) 1991 *Racial Politics in American Cities*, 2nd edn. Longman, New York
- Cain B E, Kiewiet D R, Uhlaner C J 1991 The acquisition of partisanship by Latinos and Asian Americans. *American Journal of Political Science* **35**: 390-442
- Cannon D T 1999 *Race, Redistricting, and Representation*. University of Chicago Press, Chicago
- Carmichael S, Hamilton C V 1967 *Black Power*. Vintage Press, New York
- Carmines E G, Merriman W R Jr. 1993 The changing American dilemma: Liberal values and racial policies. In: Sniderman P M, Tetlock P E, Carmines G E (eds.) *Prejudice, Politics, and the American Dilemma*. Stanford University Press, Stanford, CA, pp. 237-55
- Cohen C J 1999 *The Boundaries of Blackness: AIDS and the Breakdown of Black Politics*. University of Chicago Press, Chicago
- Cohen C J, Dawson M C 1993 Neighborhood poverty and African-American politics. *American Political Science Review* **87**: 286-302
- Dahl R 1961 *Who Governs? Democracy and Power in an American City*. Yale University Press, New Haven, CT
- Dahl R 1989 *Democracy and Its Critics*. Yale University Press, New Haven, CT
- Davidson C, Grofman B (eds.) 1994 *Quiet Revolution in the South: The Impact of the Voting Rights Act of 1965-1990*. Princeton University Press, Princeton, NJ
- Dawson M C 1994 *Behind the Mule: Race and Class in African-American Politics*. Princeton University Press, Princeton, NJ
- Dawson M C, Wilson E J III 1991 Paradigms and paradoxes: Political science and African American politics. In: Crotty W (ed.) *Theory and Practice of Political Science*. Northwestern University Press, Evanston, IL
- De la Garza R O, DeSipio L, Garcia F C, Garcia J A, Falcon A 1992 *Latino Voices: Mexican, Puerto Rican, and Cuban Perspectives on American Politics*. Westview Press, Boulder, CO
- DeSipio L 1996 *Counting on the Latino Vote*. University of Virginia Press, Charlottesville, VA
- Espiritu Y L 1992 *Asian American Panethnicity: Bridging Institutions and Identities*. Temple University Press, Philadelphia
- Eisenger P 1982 Black empowerment in municipal jobs: the impact of Black political power. *American Political Science Review* **76**: 380-92
- Fanta-Hardy C 1993 *Latina Politics, Latino Politics*. Temple University Press, Philadelphia
- Garcia F C 1997 *Pursuing Power: Latinos and the Political System*. University of Notre Dame Press, Notre Dame, IN

- Giles M W 1977 Percent black and racial hostility: An old assumption reexamined. *Social Science Quarterly* 58: 412–17
- Giles M W, Buckner M 1993 David Duke and black threat: An old hypothesis revisited. *Journal of Politics* 55: 702–13
- Giles M W, Evans A 1986 The power approach to intergroup hostility. *Journal of Conflict Resolution* 30: 469–85
- Giles M W, Hertz K 1994 Racial threat and partisan identification. *American Political Science Review* 88: 317–26
- Gilroy P 1991 *There Ain't No Black in the Union Jack*. University of Chicago Press, Chicago
- Glaser J M 1994 Back to the black belt: Racial environment and white racial attitudes in the South. *Journal of Politics* 56: 21–41
- Grofman B, Handley L, Niemi R G 1992 *Minority Representation and the Quest for Voting Equality*. Cambridge University Press, New York
- Harris F 1994 Something from within: Religion as a mobilizer of African-American political activism. *Journal of Politics* 56: 42–68
- Hero R 1992 *Latinos and the Political System: Two-tiered Pluralism*. Temple University Press, Philadelphia
- Hochschild J 1995 *Facing Up to the American Dream: Race, Class, and the Soul of the Nation*. Princeton University Press, Princeton, NJ
- Holden M 1973 *The Politics of the Black Nation*. Chandler Publishing, New York
- Huckfeldt R, Kohfeld C W 1989 *Race and the Decline of Class in American Politics*. University of Illinois Press, Urbana and Chicago
- Jackman M, Muha M J 1984 Education and inter-group attitudes: Moral enlightenment, superficial democratic commitment, or ideological refinement. *American Sociological Review* 49: 93–102
- Jennings J (ed.) 1994 *Blacks, Latinos, and Asians in Urban America: Status and Prospects for Politics and Activism*. Praeger, Westport, CT
- Jennings J, Rivera M (eds.) 1984 *Puerto Rican Politics in Urban America*. Greenwood Press, Westport, CT
- Jones M H 1972 A frame of reference for Black politics. In: Henderson L J (ed.) *Black Political Life in the United States*. Chandler Publishing Company, San Francisco
- Kim C J In Press. 1997 *Bitter Fruit: The Politics of Black-Korean Conflict in New York City*. Yale University Press, New Haven, CT
- Kim C J 1999 The racial triangulation of Asian-Americans. *Politics and Society* 27: 105–38
- Kinder D R 1986 The continuing American dilemma: White resistance to racial change 40 years after Myrdal. *Journal of Social Issues* 42: 151–71
- Kinder D R, Sanders L M 1996 *Divided by Color: Racial Politics and Democratic Ideals*. University of Chicago Press, Chicago
- Kinder D R, Sears D O 1981 Prejudice and politics: Symbolic racism versus racial threats to the good life. *Journal of Personality and Social Psychology* 40: 414–31
- Kousser J M 1999 Colorblind injustice: Minority voting rights and the undoing of the second reconstruction. University of North Carolina Press, Chapel, NC
- Kluegel J R, Bobo L 1993 Dimensions of whites' beliefs about the black-white socioeconomic gap. In: Sniderman P M, Tetlock P E, Carmines E G (eds.) *Prejudice, Politics, and the American Dilemma*. Stanford University Press, Stanford, California, pp. 127–47
- Kluegel J R, Smith E R 1986 *Beliefs about Inequality*. Aldine de Gruyter, Hawthorne, NY
- Kuklinski J H, Parent T W 1981 Race and big government: Contamination in measuring racial attitudes. *Political Methodology* 7: 131–59
- Lublin D 1997 *The Paradox of Representation: Racial Gerrymandering and Minority Interests in Congress*. Princeton University Press, Princeton, NJ
- Lublin D, Tate K 1995 Racial group competition in US mayoral elections. In: Peterson P E (ed.) *Classifying by Race*. Princeton University Press, Princeton, NJ
- Marable M 1985 *Black American Politics: From the Washington Marches to Jesse Jackson*. Thetford Press, London
- Marable M 1990 *Race, Reform, and Rebellion*, 2nd edn. University Press of Mississippi, Jackson, MS
- March J G, Olsen J P 1984 The new institutionalism: Organizational factors in political life. *American Political Science Review* 78: 734–49
- Margolis M, Hague K 1981 Applied tolerance or fear of government? An alternative interpretation of Jackman's finding. *American Journal of Political Science* 25: 241–55
- Marx A W 1998 *Making Race and Nation: A Comparison of South Africa, the United States, and Brazil*. Cambridge University Press, New York
- McAdam D 1982 *Political Process and Development of Black Insurgency, 1930–1970*. University of Chicago Press, Chicago
- McClain P D 1993 The changing dynamics of urban politics—Black and Hispanic municipal employment: Is there competition? *Journal of Politics* 55: 399–414
- McClain P D, Garcia J A 1993 Expanding disciplinary boundaries: Black, Latino, and racial minority group politics in political science. In: Finifter A W (ed.) *Political Science: The State of the Discipline II*. The American Political Science Association, Washington, DC, pp. 247–79
- McClain P D, Karnig A K 1990 Black and Hispanic socioeconomic and political competition. *American Political Science Review* 84: 535–45
- McConahay J B 1986 Modern racism, ambivalence, and the modern racism scale. In: Dovidio J, Gaertner S L (eds.) *Prejudice, Discrimination, and Racism*. Academic Press, Orlando, FL, pp. 91–125
- McConahay J B, Hough J C 1976 Symbolic racism. *Journal of Social Issues* 32: 23–45
- McLemore L B 1972 Toward a theory of Black politics—The Black and ethnic models revisited. *Journal of Black Studies* 2: 323–31
- Metz D H, Tate K 1995 The color of the campaign. In: Peterson P (ed.) *Classifying by Race*. Princeton University Press, Princeton, NJ
- Miller A, Gurin P, Malunchuk O 1981 Group consciousness and political participation. *American Journal of Political Science* 25: 494–511
- Morris A 1984 *The Origins of the Civil Rights Movement*. Free Press, New York
- Morris M D 1975 *The Politics of Black America*. Harper and Row, New York
- Myrdal G 1944 *An American Dilemma: The Negro Problem and Modern Democracy*. Harper and Brothers, New York
- Nisbett R, Ross L 1980 *Human Inference: Strategies and Shortcomings of Social Judgment*. Prentice-Hall, Inc, Englewood Cliffs, NJ
- Pinderhughes D 1987 *Race and Ethnicity in Chicago Politics*. University of Illinois Press, Urbana and Chicago
- Reed A Jr. 1988 The Black urban regime: Structural origins and constraints. In: Smith M P (ed.) *Power, Community, and the City*. Transaction Books, New Brunswick, NJ, Vol. 1

- Reeves K 1997 *Voting Hopes or Fears?* Oxford University Press, New York
- Rosenstone S J, Hansen J M 1993 *Mobilization, Participation, and Democracy in America*. Macmillan Publishing Co., New York
- Saito L T 1998 *Race and Politics: Asian Americans, Latinos, and Whites in a Los Angeles Suburb*. University of Illinois Press, Urbana and Chicago
- Sears D O 1988 Symbolic Racism. In: Katz P A, Taylor D A (eds.) *Eliminating Racism: Profiles in Controversies*. Plenum Press, New York
- Sears D O, Kinder D R 1985 Whites' opposition to busing: On conceptualizing and operationalizing conflict. *Journal of Personality and Social Psychology* **48**: 1141–7
- Sears D O, Sidanius J, Bobo L 2000 *Racialized Politics: The Debate About Racism in America*. University of Chicago Press, Chicago
- Shingles R 1981 Black consciousness and political participation. *American Political Science Review* **75**: 76–91
- Sidanius J, Pratto F 1993 The inevitability of oppression and the dynamics of social dominance. In: Sniderman P M, Tetlock P E, Carmines E G (eds.) *Prejudice, Politics, and the American Dilemma*. Stanford University Press, Stanford, CA, pp. 173–211
- Sidanius J 1996 Racism, conservatism, affirmative action and intellectual sophistication: a matter of principled conservatism or group dominance? *Journal of Personality and Social Psychology* **70**: 476–90
- Sigelman L, Welch S 1991 *Black Americans' Views of Racial Inequality: The Dream Deferred*. Cambridge University Press, New York
- Smith R C 1981 Black power and the transformation from protest to politics. *Political Science Quarterly* **96**: 431–43
- Smith, R M 1997 *Civic Ideals: Conflicting Visions of Citizenship in US History*. Yale University Press, New Haven, CT
- Sniderman P M, Hagen M G 1985 *Race and Inequality: A Study in American Values*. Chatham House, Chatham, NJ
- Sniderman P M, Piazza T 1993 *The Scar of Race*. Harvard University Press, Cambridge, MA
- Sniderman P M, Tetlock P E 1986 Symbolic racism: Problems of motive attribution in political analysis. *Journal of Social Issues* **42**: 129–50
- Sniderman P M, Brody R A, Kuklinski J 1984 Policy reasoning and political values: The problem of racial equality. *American Journal of Political Science* **28**: 75–94
- Sniderman P M, Piazza T, Tetlock P E, Kendrick A 1991 The new racism. *American Journal of Political Science* **35**: 423–47
- Sniderman P M, Tetlock P E, Carmines E G 1993 Prejudice and politics: An introduction. In: Sniderman P M, Tetlock P E, Carmines E G (eds.) *Prejudice, Politics, and the American Dilemma*. Stanford University Press, Stanford, CA, pp. 1–31
- Sonenshein R 1993 *Politics in Black and White: Race and Power in Los Angeles*. Princeton University Press, Princeton, NJ
- Swain C M 1993 *Black Faces, Black Interest: The Representation of African Americans in Congress*. Harvard University Press, Cambridge, MA
- Tate K 1993 *From Protest to Politics: The New Black voters in American Elections*. Harvard University Press and the Russell Sage Foundation, Cambridge, MA
- Thernstrom S, Thernstrom A 1997 *America in Black and White*. Simon and Schuster, New York
- Truman D B 1971 *The Governmental Process: Political Interests and Public Opinion*, 2nd edn. Knopf, New York
- Uhlman C J, Cain B, Kiewiet D R 1989 Political participation of ethnic minorities in the 1980s. *Political Behavior* **11**: 195–231
- Verba S, Nie N H 1972 *Participation in America: Political Democracy and Social Equality*. Harper and Row, New York
- Verba S, Schlozman K L, Brady H E 1995 *Voice and Equality*. Harvard University Press, Cambridge, MA
- Walton H 1973 *The Study and Analysis of Black Politics*. Scarecrow Press, Metuchen, NJ
- Walton H 1985 *Invisible Politics*. State University of New York Press, Albany, NY
- Whitby K J 1997 *The Color of Representation: Congressional Behavior and Black Interests*. University of Michigan Press, Ann Arbor, MI
- Wrinkle R D, Stewart J Jr, Polinard L, Meier K J, Arvizu J R 1996 Ethnicity and nonelectoral political participation. *Hispanic Journal of Behavioral Sciences* **18**: 142–53

K. Tate

Racial Relations

The term *racial relations* denotes the relationships between groups socially defined as 'races.' In common usage this term is ambiguous and is taken by many popular analysts to mean that racial relations are not primarily matters of power and oppression, but rather mostly involve relationships between groups with relatively equal resources. However, in many cases racial relations are more accurately termed *racist relations*. For example, in areas of the globe where there was significant European colonization—eventually most areas of the globe—systems of racist oppression were created by European colonists who enriched themselves substantially at the expense of indigenous peoples' land, resources, and labor. This often brutally executed enrichment became part of the economic foundation for Europe's and North America's long-term economic prosperity (Du Bois 1965 [1946]).

1. The Idea of 'Race'

Prior to the late eighteenth century the word 'race' was used for descendants of a common ancestor, emphasizing kinship rather than skin color or other physical characteristics. By the late eighteenth century the word had evolved into its modern sense of a category of human beings with distinctive physical characteristics transmitted by descent. In the 1770s Immanuel Kant was the first prominent scholar to use the term 'races of mankind' (in German) in the sense of physically and biologically distinct categories. Soon thereafter, the German anatomist Johann Blumenbach laid out a ladder of races that became influential (Hannaford 1996, pp. 205–7). At the top were the 'Caucasians,' a term Blumenbach coined for Europeans. Groups lower on the ladder were, in order, the

Mongolians (Asians), the Ethiopians (Africans), the Americans (Native Americans), and the Malays (Polynesians).

In the late eighteenth and early nineteenth centuries Kant, Blumenbach, and other leading European intellectuals increasingly proclaimed the virtues and privileges of whiteness. Educated Europeans and European Americans crystallized the distinction between the superior 'white race' and 'inferior races' and disseminated these views to the general public. This social construction of races (see *Conflict Sociology*) developed as Europeans expanded their colonial adventures overseas. Colonizers developed racist rationalizations for their destruction of indigenous societies, enslavement of Africans, and other colonial pursuits. This congealed in an *ideological racism* that legitimated oppression and exploitation, especially as the racist ideas became integrated into the general ideology dominant in Western societies (Balibar and Wallerstein 1991).

2. Racial and Ethnic Groups

Taking into account this history of racial relations and racial oppression, most social scientists define a *racial group* as a social group that people inside or outside that group have decided is important to single out as inferior or superior, typically on the basis of physical characteristics subjectively selected. For marking off racial groups certain characteristics like skin color are chosen, and others like hair color are rejected, solely for social reasons. Similarly, an *ethnic group* is a group that is set apart by insiders or outsiders primarily on the basis of cultural or national-origin characteristics subjectively selected (Feagin and Feagin 1999).

Racial and ethnic groups are not fixed groups or essences that endure forever but are temporary inventions shaped in sociopolitical struggles at historical times and places. They have significance only in group interaction. While the two types of groups can be distinguished on theoretical grounds, in the everyday operation of societies, racial and ethnic markers, identities, and oppressions often intertwine.

3. Conceptual Approaches to Racial Relations

Most conceptual approaches to intergroup relations, including racial relations, can be roughly classified as (a) order-assimilation theories or (b) power-conflict theories. Order-assimilation theories tend to accent the gradual incorporation of immigrant groups into a dominant culture and stability over time in intergroup relations. In contrast, power-conflict theories give central attention to the outcomes of racial contacts that result in oppression, labor exploitation, and genocide.

3.1 Assimilation Approaches

In Europe and the United States much theorizing about racial-ethnic relations has placed an emphasis on assimilation, on the more or less orderly adaptation of a migrating group such as Italian immigrants to the ways of an established host group, such as an Anglo-Protestant core culture in the United States. Related theories accent individual or group competition seen as resolving itself in the direction of eventual accommodation (see, for example, Banton 1983). In this order-assimilation theorizing little attention is paid to outcomes of intergroup contact taking the form of long-term oppression. One early theorist, Robert E. Park (1950), argued that intergroup relations move through stages of a *race relations cycle*—contacts, competition, accommodation, and eventual assimilation—that are more or less progressive and irreversible. Scholars in this tradition argue that there is a long-term trend toward orderly adaptation of immigrant groups in Western societies and that racial-ethnic conflict will diminish with the eventual assimilation of these incoming groups.

Giving primary attention to immigrants to the United States, Milton Gordon (1964) views the typical trend of adaptation to be in the direction of immigrants giving up much of their heritage for the pre-existing Anglo-Protestant core culture. This adaptation takes place in seven different arenas of assimilation: cultural assimilation, structural (primary group) assimilation, marital assimilation, attitude-receptional assimilation (breakdown of prejudices), behavior-receptional assimilation (decline of discrimination), identificational assimilation, and civic assimilation. Classical assimilation theorists have had in mind as their main examples of racial-ethnic adaptation the European groups migrating more or less voluntarily to the United States or European nations. Until relatively recently, theorists who accent assimilation gave less attention to the adaptation and assimilation of non-European groups to the core culture. When they did, they often viewed non-Europeans as assimilating like other groups. One prominent European writer on US racial relations, Gunnar Myrdal (1964 [1944], Vol. 2, p. 929), argued that as a practical matter it is to the advantage of black Americans 'as individuals and as a group to become assimilated into American culture, to acquire the traits held in esteem by the dominant white Americans.' Even today, many assimilation analysts expect the importance of racial-ethnic stratification to decline as the powerful universalistic forces of industrialization and urbanization wipe out the vestiges of earlier particularistic value systems in various societies.

3.2 Power-conflict Approaches

Generally speaking, assimilation theorists are most concerned with cultural (such as language) or primary-

group network (such as intermarriage) features of the adaptation of migrating groups to a host society, while power-conflict theorists are most concerned with racial-ethnic oppression, exploitation, and subordination. An early US advocate of this latter perspective, Oliver C. Cox (1948), emphasized the role of the capitalist class in creating racial conflict by exploiting the labor of groups like African Americans and by propagating racist ideologies to divide the working class in Western societies. Racial (white-on-black) stratification grew out of the expansion of profit-oriented capitalism.

Some scholars have adopted the concept of *internal colonialism* to describe what happens to many non-European groups coming into Western societies. An emphasis on great power inequalities between whites and people of color is at the heart of this model. Thus, people of color, such as Native Americans and African Americans, were brought into the emerging political-economic system of North America by force. Subsequent adaptation was characterized by dominant group coercion and oppression, as well as by subordinate group resistance. Racial stratification has a material base, a firm underpinning rooted in a specific political-economic history. Internal-colonialism theorists are not primarily concerned with white immigrant groups but rather with the persistence of systemic oppression and the control processes that maintain white dominance (Blauner 1972).

Power-conflict analysts look at the immigration of non-European peoples as involving much more than assimilation. John Rex (1973) has analyzed the immigration of such groups from the former European colonies to the metropolitan cities of the colonizing countries. In the latter countries the colonial immigrants get the most inferior and marginal jobs and suffer the highest unemployment—thereby becoming structurally distinct from the native-born, white working class and, thus, an outcast group of alien workers.

The concept of *institutional racism* accents the importance of patterns of racism built into a society's major institutions and patterns. Racism is much more than whites' racist attitudes, for it encompasses a complex array of inegalitarian relationships developed over generations and imbedded in all institutions. In racist systems most whites become stakeholders in a hierarchical structure of opportunities, wealth, and privileges linked to a long history of racial exploitation and oppression. Joe R. Feagin (2000) has shown that the concept of *systemic racism* encompasses not only the construction of racist ideologies but also the maintenance of white privilege, wealth, and power. People do not experience racial relations in the abstract but in concrete and recurring social situations. Individuals, whether they are the perpetrators or recipients of discrimination, are caught in a complex web of alienating racist-relations. A racist system categorizes and divides human beings from each other and

severely impedes the development of a common consciousness. At the macrolevel, large-scale institutions routinely perpetuate racial subordination and inequalities. These institutions are created and recreated by the routine discriminatory practices of dominant group individuals at the microlevel of everyday interaction. Individual acts of discrimination regularly activate whites' group power position and authority, a point underscored by the European scholar (Philomena Essed 1991).

In much theorizing about contemporary racial conflict, there is an accent on class or on socioeconomic status. Some analysts argue that systemic racism is of declining importance in Western societies like the United States and is more or less consigned to the fringes, while class divisions are now increasing (Wilson 1978). Yet the evidence (see below) suggests that such a conclusion is premature. Critics (for example, Omi and Winant 1994) of this view of the declining significance of racism have shown that, for the United States at least, race cannot be reduced to class or socioeconomic status, but remains an autonomous field of intergroup conflict, political organization, and ideological meaning.

4. *Continuing Racial Oppression: The US Case*

Today, people of African descent are still the globe's largest racially oppressed group; they are now resident in a number of countries as a result of the overseas slave trade carried out by Europeans. One of the largest groups in this African diaspora is in the United States, where there remains much evidence of continuing white-on-black oppression.

Recent field studies and opinion surveys show that among white Americans racist attitudes remain widespread. In one survey three-quarters of white respondents openly agreed with one of more of the eight antiblack stereotypes presented to them, while just over half agreed with two or more (Anti-Defamation League 1993). Discrimination against African Americans and other Americans of color also remains commonplace. One Los Angeles survey (Bobo and Suh 1995) found that within the past year six in ten black workers had encountered workplace discrimination; that proportion increased with workers' education. This study also found that a majority of highly educated Asian and Latino American workers faced workplace discrimination. More generally, researchers in the United States have found that, after decades of affirmative action programs attempting to address discrimination in employment, at least 95 percent of the top positions in major economic, political, and educational organizations are still held by white men. In the United States housing patterns are still highly segregated, largely as a result of active discrimination by white landlords, homeowners, and real estate

agents. In a national survey using 3,800 test audits in 25 metropolitan areas, black renters and homebuyers were found to face discrimination about half the time in seeking housing (Turner et al. 1991).

In the United States, as in other Western nations, whites also direct much racist hostility and discrimination at other groups of color. A number of research studies have shown that many Asian Americans are moving into the mainstream of the US economy and suburban neighborhoods, yet even as they do so they still encounter much racial discrimination. White-collar workers, including administrators and professionals in historically white institutions, report ingrained patterns of anti-Asian discrimination (Cheng 1997). Similarly, Latino Americans also report much continuing discrimination in employment and housing. Political discrimination also targets Latino and Asian Americans. During the 1990s California voters, especially white voters, supported several ballot propositions designed to abolish government services for Latino and Asian immigrants and eliminate affirmative action and bilingual programs. In addition, hostility to Spanish speaking Americans persists in widespread mocking of the Spanish language and Latino cultures.

5. *Racism in Europe: The Case of France*

The United States is not the only Western nation pervaded by racist relations. The reality of racial oppression can be seen in the outbreak of hate crimes across the European continent since the 1980s, as well as in routine discrimination in housing and employment in many European countries. For example, in France immigrants from Algeria and other African countries often face racist practices much like those in the United States. France imposed colonial rule on Algeria in the nineteenth century; later, beginning in the 1960s, large numbers of Algerians migrated to France. Defining immigrant groups as undesirable, whites there publicly accented the lack of assimilation by African immigrants into French culture as the major issue, playing down or ignoring the way in which such realities as housing segregation result from institutionalized, race-linked discrimination against Africans. In France racist-right political parties have grown in influence in recent years, and the national government has responded periodically with repressive attacks on African immigrants (Batur-VanderLippe 1999).

For numerous European countries, overseas colonialism involved the use of violence to exploit the resources of the colonies and oppress their peoples—practices generally legitimated with racist ideologies. More recently, the descendants of the formerly colonized face racist oppression in the colonizers' home countries. In a number of European countries the legacy of colonialism can be seen graphically in the

often violent expressions of racialized hatred for the identity and culture of African, Asian, and West Indian immigrants.

6. *Racism in Former Empires: The South African Case*

In many ways South Africa is the symbol of racist relations as they developed between European and non-European peoples in the ever spreading European empires. In the South African case, two colonizing powers, England and Holland, invaded the lands of indigenous African peoples, and their descendants became the economically and politically dominant group. In the first decade of the twentieth century a new country, called the Union of South Africa, was formed, combining residents of British and Dutch origin indigenais as the ruling group over a number of African groups. Black Africans were gradually forced into a system of segregation (apartheid) where the relations between the white and black groups, as well as between these groups and middle groups like the 'Asians' and the 'coloreds' (those of mixed racial backgrounds), were determined by an extensive system of segregation laws. Over time the system of apartheid came to encompass every aspect of the relations between whites and non-whites, including relations in the institutional areas of employment, education, and housing (Batur-VanderLippe 1999, Feagin and Feagin 1999).

As in the case of African Americans, this intensive system of racial oppression generated resistance in the form of major organizations like the African National Congress (ANC). In the 1990s the black African majority finally took back political control. The ANC, working with other anti-apartheid groups, eventually forced an end to apartheid and brought new hope for the oppressed majority. Racial relations dramatically changed. A formerly total-racist society, the Union of South Africa moved from white to black political control and began to change the patterns of its social and economic apartheid with relatively little bloodshed. Still this change has so far marked only a beginning, for the political changes have not yet ended economic domination by whites, and racist practices are still found in housing and employment.

Late twentieth century events in South Africa have linked black peoples around the globe in a common struggle against the descendants of European colonists, who long maintained economic or political control in a number of nations where people of color are the numerical majority. Since the early 1900s black American efforts for liberation from racism (such as the recurring civil rights movements) have provided inspiration for the struggle in South Africa and other nations (for example, Brazil), and in a similar fashion the recent political successes by black Africans in South Africa have encouraged African Americans in

their struggle against oppression in the United States. Black Americans have long drawn on the history, spirituality, and symbolism of Africa as part of their individual and group strategies for coping with the white-dominated society. This is yet another example of the way in which racial relations in one nation are set within a larger global context.

7. *The Future of the Global Racial Order*

In many countries across the globe challenges to white domination and racism are arising from population and sociopolitical changes. Europeans and European Americans are a decreasing fifth of the world's population and a decreasing proportion of US population. In the United States, at the beginning of the twenty-first century, whites are a statistical minority in many of the largest cities and in the states of New Mexico and Hawaii. By the early 2000s whites will be a minority in the states of California and Texas, with other major states to follow in the next few decades. Demographic projections indicate that shortly after the year 2050 white Americans will become a minority of the US population.

As whites become the minority in the United States and in many international organizations where they once dominated, the traditional patterns of racial relations will be under even greater pressure to change. For the US case, by the 2050s, the majority of the labor force will no longer be white, and the population and labor force will be older. This will likely give people of color ever more political and economic clout. Tensions will likely increase between a retired population that is mostly white and workers who are mostly not white. In several Western nations, white politicians opposed to legal immigration by non-Europeans and to remedial programs for racial discrimination will not likely be elected when the majority of their political constituencies become citizens of color. As constituencies change, juries and justice systems, educational systems, and government policies will also be likely to change.

Across the globe the relations of colonialism and racism are slowly but dramatically changing. The deep-lying contradictions of the global racial order were set into motion by the colonial adventures of numerous European nations. This colonialism generated sociopolitical structures that have long embedded the practices, ideologies, and institutions of racist subordination in numerous nations. Today, international relations, global markets, the global mass media, and multinational corporations are all racialized, with white European perspectives and executives often in control. Half a millennium of European-dominated national and international institutions is slowly coming to an end. As the twenty-first century develops, racial relations in Europe, the United States, and the rest of the world are likely to

change. Many once oppressed groups and societies will move farther out from under the dominance of Europeans and European Americans. Over the last few decades strong antiracism movements have arisen in South Africa, Brazil, and the United States. The possibility of a global democratic order rid of the oppressive relations of racism remains only a dream, but the South African revolution shows that it is a powerful dream for many people.

See also: Affirmative Action Programs (India): Cultural Concerns; Affirmative Action Programs (United States): Cultural Concerns; Affirmative Action, Sociology of; Colonization and Colonialism, History of; Conflict Sociology; Discrimination: Racial; Environmental Justice; Ethnic Conflicts; Ethnic Groups/Ethnicity: Historical Aspects; Ethnicity, Sociology of; Hate Crimes; Policing; Race and Gender Intersections; Race and the Law; Race and Urban Planning; Race: History of the Concept; Race Identity; Race Relations in the United States, Politics of; Racism, History of; Racism, Sociology of; Residential Segregation: Sociological Aspects; Slavery as Social Institution; Slaves/Slavery, History of

Bibliography

- Anti-Defamation League 1993 *Highlights from an Anti-Defamation League Survey on Racial Attitudes in America*. Anti-Defamation League, New York
- Balibar E, Wallerstein I 1991 *Race, Nation, Class: Ambiguous Identities*. Verso, London
- Banton M P 1983 *Racial and Ethnic Competition*. Cambridge University Press, Cambridge, UK
- Batur-VanderLippe P 1999 Colonialism and post-colonialism: the global expansion of racism. In: Feagin J R, Feagin C B (eds.) *Racial and Ethnic Relations*. Prentice-Hall, Upper Saddle River, NJ
- Blauner R 1972 *Racial Oppression in America*. Harper & Row, New York
- Bobo L, Suh S A 1995 Surveying Racial Discrimination: Analyses from a Multiethnic Labor Market. Research Report. Department of Sociology, University of California, Los Angeles
- Cheng C 1997 Are Asian American employees a model minority or just a minority? *Journal of Applied Behavioral Science* 33: 277–90
- Cox O C 1948 *Caste, Class, and Race*. Doubleday, Garden City, NY
- Du Bois W E B 1965 [1946] *The World and Africa*. International Publishers, New York
- Essed P 1991 *Understanding Everyday Racism*. Sage, Newbury Park, CA
- Feagin J R 2000 *Racist America: Old Roots, Current Realities, and Future Reparations*. Routledge, New York
- Feagin J R, Feagin C B 1999 *Racial and Ethnic Relations*, 6th edn., Prentice-Hall, Upper Saddle River, NJ
- Gordon M M 1964 *Assimilation in American Life*. Oxford University Press, New York
- Hannaford I 1996 *Race: The History of an Idea in the West*. Johns Hopkins University Press, Baltimore, MD

- Myrdal G 1964 [1944] *An American Dilemma*. McGraw-Hill, New York
- Omi M, Winant H 1994 *Racial Formation in the United States: From the 1960s to the 1990s*, 2nd edn. Routledge, New York
- Park R E 1950 *Race and Culture*. Free Press, Glencoe, IL
- Rex J 1973 *Race, Colonialism and the City*. Routledge and Kegan Paul, London
- Turner M A, Struyk R J, Yinger J 1991 *Housing Discrimination Study: Synthesis*. US Dept. of Housing and Urban Development, office of Policy Development and Research, Washington, DC
- Wilson W J 1978 *The Declining Significance of Race: Blacks and Changing American Institutions*. University of Chicago Press, Chicago, IL

J. R. Feagin

Racism, History of

In common usage the term racism is often employed in a loose and unreflective way to describe the hostile or negative feelings of one ethnic group or historical collectivity toward another and the actions resulting from such attitudes. But historians and social scientists normally require a more precise and limited definition, although they may disagree on what it should be. It is useful, therefore, to review the historiographic discourse on the meaning of racism before attempting a brief survey of what contemporary historians might accept as its principal manifestations.

1. Historiography of Racism

The term racism was coined in the 1930s mainly in reference to the theories on which the Nazis based their persecution of the Jews. The first historian to focus directly on the subject was Barzun. His book *Race: A Study in Superstition* (originally published in 1937 and reprinted in 1965) set two important precedents for most future historians of racism: it presumed that claims of the innate inferiority of one 'race' to another were false or at least unproven, and its main concern was with the history of ideas rather than with the social and political applications of prejudiced beliefs and attitudes. Barzun's concern with the intellectual origins of Nazi antisemitism was shared by the many scholars who wrote on the subject after World War II in response to what the Holocaust had revealed about the horrifying consequences of racist ideas. A notable example was Poliakov's *Le Myth Aryen*, translated into English as *The Aryan Myth* (1974). But disagreements developed on whether biological racism was continuous with earlier antisemitic attitudes that were, ostensibly at least, based on religion or whether it was a radical new departure, a *sine qua non* for something like the Holocaust.

In the immediate postwar decades, the term racism was also applied with increasing frequency to the relationships between whites or Europeans and people of color, especially Africans or those of African descent, in colonial, postcolonial, and former slave societies. The rising concern with black-white relations in the US, South Africa, and (to a somewhat lesser extent) in Great Britain, beginning in the 1950s and early 1960s, gave rise to an enormous literature on the origins and development of white supremacy as an ideology. Historians of ideas did most of the early work, but gradually the emphasis shifted somewhat from the racist ideas themselves to the patterns or racial discrimination and domination for which they served as a rationale. Many historians of black-white relations used the term casually and without reflecting on its exact meaning. At its most imprecise, it could mean anything that whites did or thought that worked to the disadvantage of blacks.

There have been relatively few attempts to write general histories of racism, which encompass both the antisemitic and color-coded varieties. For the most part, the two historiographies have not addressed each other, and the definitions employed have sometimes been incompatible. Some historians of antisemitic racism have found the term applicable only if the aim was the elimination of the stigmatized group. Some historians of white supremacy, on the contrary, would limit the term to ideologies associated with patterns of domination or hierarchical differentiation. The first attempt at a comprehensive history of American racism, Gossett's *Race: The History of an Idea in America* (1963) traced consciousness of race to the ancient world. Its treatment of specifically American manifestations cast its net quite wide to include representation and treatment of American Indians and immigrants from southern and eastern Europe, including Jews, as well as blacks. But most subsequent work on the history of American racism has been group specific and has concentrated most heavily on attitudes toward African Americans. Mosse's general history of European racism, on the other hand, pays most of its attention to the growth of antisemitism between the Enlightenment and the Holocaust (Mosse 1978). But there appear to be only two significant attempts to cover Western attitudes toward race comprehensively: Hannaford's *Race: The History of an Idea in the West* (1996) and Geiss' *Geschichte des Rassismus* (1988). Hannaford's study, as its title indicates, is strictly an intellectual history and considers race as a concept more than racism as an ideology. It argues strenuously that no clear concept of race existed before the seventeenth century, thus raising the issue of whether anything that existed before the invention of race in the modern sense can be legitimately labeled racism. Geiss, on the contrary, sees racism as anticipated in most respects by the ethnocentrism or xenophobia that developed in the ancient world, as reflected, for example, in the Old Testament.

2. Toward a Definition of Racism for Historical Purposes

Somewhere between the view that race is a peculiar modern idea without much historical precedent and the notion that it is simply an extension of the ancient phenomena of ethnocentrism and xenophobia may lie a working definition that is neither too broad for historical specificity or too narrow to cover more than the limited span of Western history during which a racism based on scientific theories of human variation was widely accepted. If racism is defined as an ideology rather than as a theory, links can be established between belief and practice that the history of ideas may obscure. But ideologies have content, and it is necessary to distinguish racist ideologies from other belief systems that emphasize human differences and can be used as rationalizations of inequality. The classic sociological distinction between racism and ethnocentrism is helpful, but not perhaps in the usual sense, in which the key variable is whether differences are described in cultural or physical terms. It is actually quite difficult in specific historical cases to say whether appearance or 'culture' is the source of the salient differences, because culture can be reified and essentialized to the point where it has the same deterministic effect as skin color. But it would stretch the concept of racism much too far to make it cover the pride and loyalty that may result from ethnic identity. Such group-centeredness may engender prejudice and discrimination against those outside the group, but two additional elements would seem to be required before the categorization of racism is justified. One is a belief that the differences between the ethnic groups involved are permanent and ineradicable. If conversion or assimilation is genuinely on offer, we have religious or cultural intolerance but not racism. The second is the social side of the ideology—its linkage to patterns of domination or exclusion. To attempt a short formulation, we might say that racism exists when one ethnic group or historical collectivity dominates, excludes, or seeks to eliminate another on the basis of differences that it believes are hereditary and unalterable.

3. The Emergence of Racism

Historians who can accept this middle-of-the-road definition can agree with the scholars who have examined conceptions of human diversity in classical antiquity and the early Middle Ages without finding clear evidence of color prejudice or racism. Neither the classical conception that one becomes civilized through being able to engage in politics nor the Christian belief that everyone regardless of color or ancestry is a potential convert with a soul to be saved could readily sustain a racist view of the world. The period between the fourteenth and the seventeenth

centuries is much more problematic. The literal demonization of the Jews in the late middle ages, the belief that they were in league with the devil and plotting the destruction of Christianity, could lead unsophisticated folk to believe that they were outside the bounds of humanity, even though church authorities persisted in advocating their conversion. Folk antisemitism resulted in massacres and expulsions of Jews in virtually every country in Western Europe. A degree of elite or official sanction for such attitudes came in sixteenth century Spain when Jews who had converted to Christianity and their descendants became the victims of a pattern of discrimination and exclusion. When purity of blood, or *limpieza de sangre*, became a qualification for certain offices and honors it signified that what people could accomplish or achieve no longer depended on what they did but on who they were. Precedents for the notion that blood will tell could be found in the inheritability of royal or noble status but the designation of an entire ethnic group as incapable of a reliable conversion to Christianity constituted, at the very least, a transitional stage between religious bigotry and biological racism.

The period of the Renaissance and Reformation was also the time when Europeans were coming into increasing contact with people of darker pigmentation in Africa, Asia, and the Americas and were making judgments about them as they traded with them, fought with them over territory, or enslaved them. A belief that all these non-Europeans were children of God with souls that missionaries could save did not prevent a brutal expropriation of their land and labor, but it did inhibit the articulation of a racial justification for how they were being treated. When Africans were carried by force to the New World to labor on plantations, slave traders and slave owners sometimes invoked an obscure passage in the book of Genesis to explain the color of their human property. Ham had committed a sin against his father Noah that resulted in a divine curse on his descendants to be 'servants unto servants.' When Africans were enslaved, first by Arabs, and then by Europeans, it became convenient to believe that the punishment of God had included a blackening of the skin. But the passage had other possible meanings, and no orthodox religious authorities ever endorsed the interpretation of the curse popular among slaveholders, because they feared that it would impede conversion of blacks to Christianity. The standard and official justification for enslaving Africans was that they were heathens and that enslavement made heaven accessible to them.

An implicit racism became evident, however, when converted slaves were kept in bondage, not because they were actual heathens but because they had heathen ancestry. As with the doctrine of purity of blood in Spain, descent rather than performance became the basis for determining the qualifications for membership in a community that was still theoretically based on a shared Christian faith. Beginning in the late

seventeenth century laws were passed in English North America forbidding marriage between whites and blacks and discriminating against the mixed offspring of informal liaisons. Without clearly saying so, such laws implied that blacks were unalterably alien and inferior.

4. *The Rise of Modern Racism*

A modern naturalistic form of racism could not be articulated fully until the Enlightenment. A secular or scientific theory of race required a new way of looking at human diversity, one that moved the subject away from the Bible, with its insistence on the essential unity and homogeneity of the human race and its collective elevation above the animal kingdom. Eighteenth-century ethnologists began to think of human beings as part of the natural world and subdivided them into three to five races, usually considered as varieties or subspecies. Most eighteenth-century ethnological theorists believed that whiteness was the original color of humanity and that blackness or brownness resulted from a process of degeneration caused by climate and conditions of life. Although the white or Caucasian race was generally considered more beautiful and civilized than the Mongolian and especially the Negro, the belief that human beings constituted a single species with a common origin remained dominant. In the late eighteenth and early nineteenth centuries, however, an increasing number of ethnologists asserted that the races were separately created and constituted distinct species. As debates over the morality of slavery and the slave trade erupted in Western Europe and the US in the wake of democratic revolutions and the rise of humanitarian movements, some of the most militant defenders of black bondage adopted these polygenetic arguments as a justification for enslaving Africans and holding them in permanent servitude. But the fact that polygenesis seemed to conflict with scripture limited the usage of this extreme version of scientific racism. Many defenders of slavery in the American South, where evangelical religion enforced biblical literalism, reverted to the Curse on Ham or made the degeneracy hypothesis of Enlightenment ethnology into an argument for the permanent or irreversible divergence of the races from their common origin.

The nineteenth century was an age of emancipation, nationalism, and imperialism—all of which contributed to the growth and intensification of ideological racism in Europe and in the US. Although the emancipation of blacks from slavery and Jews from the ghettos, which seemed a logical consequence of the democratic revolutions of the late eighteenth century, received most of its support from religious or secular believers in an essential human equality, the consequences of these reforms was to intensify rather than diminish racism. Race relations became less paternalistic and more competitive. Poor or working class

whites in the US, for example, feared competition with emancipated blacks for work, land, and social status. Middle class (especially lower middle class) Germans viewed the increasing involvement of Jews in finance, commerce, the professions, journalism, and the arts during the late nineteenth century as a potential threat to their own economic security, social status, and traditional values. Of course, it was the legacy of inherited stereotypes about 'the Other,' in addition to the immediate economic and social circumstances, that gave edge and intensity to these feelings of insecurity and hostility. The idiom of Darwinism, with its emphasis on 'the struggle for existence' and concern for 'the survival of the fittest' was conducive to the development of a new and more credible scientific racism in an era that increasingly viewed race relations as an arena for conflict rather than as a stable hierarchy.

The growth of nationalism, especially romantic cultural nationalism, encouraged the growth of a more mystical variant of racist thought. When it could be assumed that one had to be of German ancestry to participate in the German *Volkgeist*, Jews could be discriminated against or excluded just as effectively as if their deficiencies were clearly defined as genetic or biological. For most of the nineteenth century the question of whether German Jews could or should be converted to Christianity was an open one. Beginning in the late 1870s and early 1880s, however, the coiners of the term 'antisemitism' made explicit what some cultural nationalists had previously implied—that to be Jewish was not simply to adhere to a set of religious beliefs or cultural practices but meant belonging to a race that was the antithesis of the race to which true Germans belonged. The latter was variously designated as Teutonic, Aryan, Nordic, or simply Germanic. In Great Britain and the US romantic nationalism took the form of a veneration for Anglo-Saxon ancestors and legacies that could turn racist, as it did for a time in the US, when confronted with large-scale immigration from eastern and southern Europe. In France, an ethnocentric cultural nationalism fed the fury against the Jews that surfaced in the Dreyfus affair at the turn of century. It did not triumph against the universalistic civic nationalism inherited from the revolutionary past, but it did make antisemitic racism a strain of right-wing French thought that would resurface during the 1930s and especially under the German occupation.

The climax of Western imperialism in the late nineteenth century 'scramble for Africa' and parts of Asia and the Pacific represented an assertion of the competitive ethnic nationalism that existed among European nations (and which, as a result of the Spanish-American War came to include the US). A belief that different races, subraces, or racial mixtures inhabited each nation led some, especially the Germans and the English, to suppose that the quality of each nation's racial stock was being tested for its

fitness. But imperialism also required the massive subjugation and colonial domination of non-European populations. An ideology proclaiming that whites were superior to 'lesser breeds' and responsible for ruling over them and tutoring them in the rudiments of civilization was a prime rationale for this new burst of European expansionism, even if it was not a cause of it. The relation of racism to imperialism is more problematic than one might suppose, however, because the most consistent and extreme racists were often anti-imperialists. They believed that nothing useful could be done with people whose inferiority was so profound and permanent that they were incapable of being civilized. For them the taming or domestication of such savages was not worth the trouble. Within colonies that attracted substantial numbers of European settlers, such as French Algeria or some of the British colonies of South and East Africa, a more extreme and explicit racism than the one usually professed by government officials or missionaries could be found in the discourse of white colonists about their 'native' servants or farm laborers.

5. *The Climax: Racism in the Twentieth Century*

The climax of the history of racism came in the twentieth century in the rise and fall of what might be called overtly racist regimes. In the American South, the passage of racial segregation laws and restrictions on black voting rights reduced African Americans to lower caste status, despite the Constitutional Amendments that had made them equal citizens. Extreme racist propaganda, which represented black males as ravening beasts lusting after white women, served to rationalize the practice of lynching. These extra-legal executions increasingly were reserved for blacks accused of offenses against the color line, and they became more brutal and sadistic as time went on; by the early twentieth century victims were likely to be tortured to death rather than simply being killed. A key feature of the racist regime maintained by state law in the South was a fear of sexual contamination through rape or intermarriage, which led to efforts to prevent the conjugal union of whites with those with any known or discernable African ancestry. The effort to guarantee 'race purity' in the American South anticipated aspects of the official Nazi persecution of Jews in the 1930s. The Nuremberg laws of 1935 prohibited intermarriage or sexual relations between Jews and Gentiles, and the propaganda surrounding the legislation emphasized the sexual threat that predatory Jewish males presented to German womanhood and the purity of German blood. Racist ideology was eventually of course carried to a more extreme point in Nazi Germany than in the American South of the Jim Crow era. Individual blacks had been hanged or burned to death by the lynch mobs to serve as examples to insure that the mass of southern African

Americans would scrupulously respect the color line. But it took Hitler and the Nazis to attempt the extermination of an entire ethnic group on the basis of a racist ideology.

Hitler, it has been said, gave racism a bad name. The moral revulsion of people throughout the world against what the Nazis did, reinforced by scientific studies undermining racist genetics (or eugenics), served to discredit the scientific racism that had been respectable and influential in the US and Europe before World War II. But explicit racism also came under devastating attack from the new nations resulting from the decolonization of Africa and Asia and their representatives in the United Nations. The Civil Rights movement in the US, which succeeded in outlawing legalized racial segregation and discrimination in the 1960s, was the beneficiary of revulsion against the Holocaust as a logical outcome of racism. But it also drew crucial support from the growing sense that national interests were threatened when blacks in the US were mistreated and abused. In the competition with the Soviet Union for 'the hearts and minds' of independent Africans and Asians, Jim Crow and the ideology that sustained it became a national embarrassment with possible strategic consequences.

The one overtly racist regime that survived World War II and the Cold War was the South African, which did not in fact come to fruition until the advent of Apartheid in 1948. The laws passed banning all marriage and sexual relations between different 'population groups' and requiring separate residential areas for people of mixed race ('Coloureds'), as well as for Africans, signified the same obsession with 'race purity' that characterized the other racist regimes. However, the climate of world opinion in the wake of the Holocaust induced apologists for apartheid to avoid straightforward biological racism and rest their case for 'separate development' mainly on cultural rather than physical differences. The extent to which Afrikaner nationalism was inspired by nineteenth-century European cultural nationalism also contributed to this avoidance of a pseudo-scientific rationale. No better example can be found of how a 'cultural essentialism' based on nationality can do the work of a racism based squarely on skin color or other physical characteristics. The South African government also tried to accommodate itself to the age of decolonization. It offered a dubious independence to the overcrowded 'homelands,' from which African migrants went forth to work for limited periods in the mines and factories of the nine-tenths of the country reserved for a white minority that constituted less than a sixth of the total population.

6. *The End of Racism?*

The defeat of Nazi Germany, the desegregation of the American South in the 1960s, and the establishment of

majority rule in South Africa suggest that regimes based on biological racism or its cultural essentialist equivalent are a thing of the past. But racism does not require the full and explicit support of the state and the law. Nor does it require an ideology centered on the concept of biological inequality. Discrimination by institutions and individuals against those perceived as racially different can long persist and even flourish under the illusion of nonracism, as historians of Brazil have discovered recently. The use of allegedly deep-seated cultural differences as a justification for hostility and discrimination against newcomers from the Third World in several European countries has led to allegations of a new 'cultural racism.' Similarly, those sympathetic to the plight of poor African Americans and Latinos in the US have described as 'racist' the view of some whites that many denizens of the ghettos and barrios can be written off as incurably infected by cultural pathologies. From the historian's perspective such recent examples of cultural determinism are not in fact unprecedented. They rather represent a reversion to the way that the differences between ethnoracial groups could be made to seem indelible and unbridgeable before the articulation of a scientific or naturalistic conception of race in the eighteenth century.

See also: Anti-Semitism; Discrimination; Discrimination, Economics of; Discrimination: Racial; Ethnic Groups/Ethnicity: Historical Aspects; Ethnic Identity, Psychology of; Ethnicity: Anthropological Aspects; Holocaust, The; Prejudice in Society; Race and the Law; Race: History of the Concept; Race Identity; Race Relations in the United States, Politics of; Racial Relations; Racism, Sociology of

Bibliography

- Banton M 1998 *Racial Theories*. Cambridge University Press, Cambridge, UK
- Barzun J 1965/1937 *Race: A Study in Superstition*. Harper and Row, New York
- Fredrickson G M 1981 *White Supremacy: A Comparative Study in American and South African History*. Oxford University Press, New York
- Geiss I 1988 *Geschichte des Rassismus*. Suhrkamp, Frankfurt, Germany
- Gossett T F 1963 *Race: The History of an Idea in America*. Southern Methodist University Press, Dallas, TX
- Hannaford I 1996 *Race: The History of an Idea in the West*. Johns Hopkins University Press, Baltimore, MD
- Langmuir G I 1990 *Toward a Definition of Antisemitism*. University of California Press, Berkeley, CA
- Lauren P L 1988 *Power and Prejudice: The Politics and Diplomacy of Racial Discrimination*. Westview Press, Boulder, CO
- Mosse G L 1978 *Toward the Final Solution: A History of European Racism*. H Fertig, New York
- Pieterse J N 1992 *White on Black: Images of Africa and Blacks in Western Popular Culture*. Yale University Press, New Haven, CT
- Poliakov L 1974 *The Aryan Myth: A History of Racist and Nationalist Ideas in Europe*. Chatto for Essex University Press, London
- Smedley A 1993 *Race in North America: Origin and Evolution of a World View*. Westview Press, Boulder, CO
- Taguieff P 1997 *Le Racisme*. Flammarion, Paris

G. M. Fredrickson

Racism, Sociology of

Few concepts in social science have generated more heat and less light than 'race' and its derivative 'racism,' the former because of its multiplicity of meanings, and the latter because of its dual career as a concept and as an ideological bludgeon.

1. Definitions

'Race' has been used with four distinct meanings:

- (a) As a synonym for a biological species, as in 'the human race.'
- (b) As a synonym for a biological subspecies as in 'the Caucasoid race.'
- (c) As a synonym for a nation or ethnic group, as in 'the French race.'
- (d) As a term to designate a group of diverse origins thought to be a biological subspecies in a given society, as in 'the Jewish race' in Nazi Germany. To reduce the confusion and distinguish this usage from the second, the term 'social race' has sometimes been used, but generally the qualifier is omitted.

The first usage is largely metaphorical and, thus, unproblematic. However, the other three, are hopelessly confusing, either because they have no clear empirical referent (as in the second), or because they overlap empirically with one another (as in the second, third, and fourth), or both. The result is a conceptual muddle that has plagued the social science literature since the turn of the twentieth century.

'Racism,' i.e., the ideology or attitudes associated with one or more of the listed definitions of 'race,' is in an even more parlous state of conceptual confusion. The following list includes some but not all of the current usages:

- (a) As an ideology or set of attitudes accompanying the second meaning of 'race,' that is, a belief in the relative inferiority or superiority in moral, intellectual, physical, and other attributes of biological subspecies of *Homo sapiens*.
- (b) As an ideology or set of attitudes corresponding to the third meaning of 'race,' that is, as a synonym for ethnocentrism, as in the belief that one's own ethnic group is superior to others.

(c) As an ideology or set of attitudes linked to the fourth meaning of 'race,' namely 'social race.'

(d) As a set of practices or institutions, divorced from individual beliefs or attitudes, that, by design or otherwise, have a differential impact on various racial or ethnic groups. This is usually labeled 'institutional racism,' and is a synonym for discrimination, witting or unwitting. Thus, an ostensibly neutral test on which some groups consistently perform better than others is said to be 'institutionally racist.'

(e) As a general term of opprobrium hurled at any person, policy, or practice one disapproves of. Examples would be the 'Zionism is racism' resolution passed by the United Nations, or accusations leveled at opponents of race-based affirmative action in the USA. In the latter case, and indeed in much contemporary usage in the USA, race-blindness is said to be racist, while race discrimination is defined as liberal, thus turning the concept on its head.

2. Evolution in the Use of the Term 'Racism'

Clearly, racism as a social concept has outlived whatever use it may have had until the 1960s, before its usage became broadened beyond recognition. At mid-twentieth century, a clear distinction was maintained in most social literature between attitudes and behavior, between prejudice and discrimination, and between race and ethnicity. The concept of racism was largely restricted to prejudicial 'attitudes' applied to 'social races.' Discrimination was treated as varying in partial independence of prejudice, and discrimination was seen as a possible (but not necessary) consequence of racist attitudes. Prejudicial attitudes applied to ethnic groups were termed 'ethnocentrism,' as distinct from racism. Driven by an ideological agenda of stigmatizing any beliefs or practices thought to sustain any kind of group differential in society (e.g., group differences in performance in tests, in admission to universities, in crime or incarceration rates, in school drop-out rates, in mortality or morbidity rates, in representation in various occupations, and so on), social scientists since the 1960s have gradually blurred or eliminated this double distinction between prejudice and discrimination, and between race and ethnicity. The result has been to universalize the concept of racism. Everybody can now be called a racist, including those who steadfastly refuse to make racial distinctions.

To blur the distinction between prejudice and discrimination, as the concept of 'institutional racism' does, achieves at least two aims. First, it dissociates 'racism' from individuals, and makes it an inherent property of social institutions, or indeed of entire societies. Second, it dissociates racism from consciousness and intentionality. Both individuals and institutions can be labeled 'racist' irrespective of beliefs,

attitudes, or intentions. A classical liberal, for instance, who believes in philosophy of individual rights and equality of opportunity, without any consideration of race or ethnicity, can easily be labeled a racist if, for instance, she opposes 'race norming' of tests in employment or education. Thus, we find opponents on the two sides of a clear debate on whether to take race into account in social policy hurl the epithet of 'racist' at each other. 'Racism' is now void of meaning and reduced to pure invective.

The confounding of race and ethnicity has also served a variety of political agenda. The analytical use of the distinction between race and ethnicity is that racial (e.g., physical) categorization of groups leads to much more rigid, enduring, invidious, and stigmatizing social cleavages than ethnic (e.g., cultural) criteria of group membership. For example, the categorically different position of blacks in the USA compared to other minority groups, such as Hispanics or Asian Americans, can only be understood because they are overwhelmingly defined by the double stigma of racial slavery.

African-Americans were essentially treated as a stigmatized pariah group in American society. However, redefining American blacks, as merely one ethnic minority in a 'multicultural' society may, it is hoped, destigmatize their uniquely invidious position, but it does nothing to explain the stubborn resilience of their disabilities and predicaments in American society. The confounding of race and ethnicity in this cited example cuts both ways ideologically: it may on the one hand promote pride in 'black culture,' but it also underpins the racist argument of innate black inferiority. If blacks are but one ethnic group among many, how come they fail where others (notably Asian Americans) succeed?

3. Recent Developments

The unintended consequences of this broadening of the meaning of racism are three-fold. First, it robbed social science of a useful tool that permitted complex analysis of complicated social phenomena. Examples of such analyses are the relationship between prejudice and discrimination (one can be for instance, an unprejudiced discriminator or a prejudiced non-discriminator, depending on a social context which rewards or punishes discrimination); the difference between profoundly racist societies such as the Ante Bellum US and apartheid South Africa, and less racist ones such as Mexico or Brazil, where ethnic and class distinction often overshadow racial ones; and, the social consequence of stressing physical phenotypes such as skin color, as distinguished from cultural characteristics, such as language or religion, for the dynamics of group conflicts. Broadening the meaning of racism not only transformed a concept into an

epithet; it simultaneously impoverished social analysis. It also made the teaching of race relations in race-polarized societies a minefield of political correctness liberally sprinkled with euphemisms and impervious to rational discourse.

Second, the broadening usage of racism and the confounding of race and ethnicity have noticeably weakened the effectiveness of remedial policies by facilitating and legitimating a shift from universalistic, race-blind policies of equal opportunity and individual rights, to one of groups rights, proportional representation, quota systems, and collective grievances and guilt. Policies such as race-based affirmative action in the USA are not only ineffective but even retrogressive because they are based on bad social science. The inherent contradiction involved in race-based affirmative action is that the remedy of 'positive discrimination' uses as a selection criterion the very stigma the policy seeks to eliminate. Surely, the way to reduce the social significance of race is not to classify the population into official race groups, to use these categories in official statistics and reports, and to request individuals to identify themselves by race for the purpose of school attendance, employment, promotion, college admission, faculty recruitment, etc.

4. Racism and Sociobiology

Third, the extension of the concept of race and its confounding with ethnicity are congruent with the sociobiological approach to the study of race and ethnicity that has been slowly gaining ground in the social sciences over the last 20 years of the twentieth century. The motor of natural selection is differential reproduction. Organisms are selected to behave in ways that are reproductively successful. Reproductive success can be direct or indirect. An organism can have offspring, or it can help other related organisms (that share a proportion of their genes) to reproduce. In short, nepotism—the favoring of kin—is a common strategy for maximizing inclusive fitness. This clearly operates in humans as in other social species, and is particularly evident in nuclear and extended families. However, members of ethnic and racial groups also define themselves as linked by common biological descent, that is, as large extended families. Racism and ethnocentrism can thus be seen as extended forms of nepotism. Members of such groups are typically more closely related to one another on average, than to members of other groups.

Racial and ethnic groups, in short, share the fundamental property of real, or at least putative, common descent, although they only constitute diluted forms of nepotism. The distinction between racial and ethnic groups is in the criteria used to determine membership: either physical traits in the case of races, or cultural characteristics in the case of ethnics. Preference for other individuals who resemble one

physically is common where certain visible traits, such as skin color, enable one to discriminate easily between groups. However, in the case of neighboring groups that closely resemble one another (such as Swedes and Norwegians, or Zulu and Swazi), cultural traits, notably language and dialect, are more useful than biological features in establishing group membership.

The choice of racial or cultural criteria of group membership is thus opportunistically dictated by which does the best job of cheaply, quickly, easily, and reliably discriminating between co-existing groups. Sometimes both racial and cultural traits are used, but purely physical traits are often unreliable because neighboring groups frequently interbreed. Ethnocentrism, therefore, is far more general a phenomenon than racism, even though racism is generally present where it 'works,' e.g., in situations of recent long-distance migration rapidly putting in close contact large numbers of phenotypically distinct populations. Military conquest, colonialism, slavery, and labor migration are examples of such situations where racism thrives. Conversely, after several generations of contact and interbreeding, racism tends to wane, unless forcibly maintained as in the USA and South Africa, by a discriminatory political and legal system.

Seen in evolutionary perspective, racism becomes a natural propensity to favor individuals to whom one assesses to be more closely biologically related to one, and thus an expression of the nepotism that drives natural selection through inclusive fitness. This general preference for 'look-alikes' operates both at the individual level (e.g., in assortative mating between members of the same group), and between groups where group differences in the distribution of certain phenotypes are pronounced enough to make reliable and easy discrimination possible.

According to this evolutionary perspective, racism is both primordial, that is rooted in the biology of nepotism, and circumstantial or instrumental, varying with the degree to which co-existing groups look alike or different. It is neither inevitable nor unchangeable, but it is likely to occur whenever physical phenotypes, especially highly viable ones, such as skin pigmentation, sharply differentiate between groups. Extensive interbreeding, however, can quickly undermine it. Racism is also responsive in both directions, to state-imposed policies that either reward, or punish it. A state determined to maintain a racial caste society can do so for centuries, as US and South African history amply demonstrate. Conversely, the state can create an environment in which racial tolerance can thrive. The persistence or waning of racism is also a function of the degree of overlap between class and race. Racism is more likely to endure when large socioeconomic differences exist between racially defined groups, and to subside where those differences are reduced.

In conclusion, the double side of racism is that while, in all likelihood, it is indeed 'in our genes' as an

expression of the biology of nepotism, it takes definite environmental conditions to become expressed. In this respect, it is no different from all our genetic propensities. The phenotype is always the product of interaction between the genotype and the environment. But that environment is not only changeable; it can also be manipulated for malign or benign ends. However, the main lesson of multiracial societies is that any state policy based on race, whether purportedly malign or benign, consolidates and perpetuates racial distinctions.

See also: Colonialism: Political Aspects; Colonization and Colonialism, History of; Discrimination: Racial; Ethnic Conflicts; Ethnicity, Sociology of; Eugenics, History of; Gender, Class, Race, and Ethnicity, Social Construction of; Genotype and Phenotype; Hate Crimes; Poverty, Sociology of; Prejudice in Society; Race and Gender Intersections; Race and the Law; Race Identity; Racial Relations; Racism, History of; Social Stratification; Urban Poverty in Neighborhoods; Xenophobia

Bibliography

- Allport G W 1954 *The Nature of Prejudice*. Addison-Wesley, Cambridge, MA
 Banton, M 1977 *The Idea of Race*. Tavistock, London
 Francis E K 1976 *Interethnic Relations*. Elsevier, New York
 Hutchinson J, Smith A D (eds.) 1996 *Ethnicity*. Oxford University Press, Oxford, UK
 Rex J, Mason D (eds.) 1986 *Theories of Race and Ethnic Relations*. Cambridge University Press, Cambridge, UK
 Schermerhorn R A 1970 *Comparative Ethnic Relations*. Random House, New York
 Sowell T 1994 *Race and Culture*. Basic Books, New York
 van den Berghe P L 1981 *The Ethnic Phenomenon*. Elsevier, New York

P. L. van den Berghe

Radicalism

The word 'radicalism' is derived from the Latin 'radix' meaning 'root.' Radicalism is any stance, practical, intellectual, or both, that goes to the root of existing practices, beliefs, or values. By transposition the term has shifted from meaning things related to the root or foundation of an institution or set of arrangements, to describing the view that existing arrangements should be transformed on the basis of foundations, roots, which are either to be rediscovered and re-asserted, or to be transplanted from the ideal to the actual. It is normally associated with the left, with liberalism (see *Liberalism*) and socialism (see *Socialism*), but since

the term is relative, any fundamental criticism of or assault on existing practices can reasonably be termed radical.

1. Radicalism is a Relationship Term not a Content Term

Radicalism is, like conservatism (see *Conservatism: Theory and Contemporary Political Ideology*), a relationship term not a content term. It describes not any particular set of aspirations and aversions, but a distance between what exists and what is desired, and the way in which proposed changes are justified. Its particular character is therefore dependent on the historical circumstances in which it exists. Views put forward by radicals in Communist China might be shared by conservatives in Connecticut; the radicalism of California might be the traditionalism of Catalunya. Hannah Arendt's observation that the 'most radical revolutionary will become a conservative on the day after the revolution' may thus describe not only a change of heart following the acquisition of power, but a change in the relationship between a substantive policy which remains consistent, but which moves from opposition to implementation.

2. Some General Features of Radicalism as a Form Can Be Identified

Despite the contingency of radicalism however, some general features of its form can be identified. Radicalism normally involves the detection or description of principles that are considered to lie at the root of both existing and desired arrangements or ways of thinking. This concentration on a single and logical hierarchy of cause and effect or structure and foundation, is accompanied by a goal and a program that similarly concentrates on a single policy or set of policies the pursuit of which is intended or expected to transform a broad swathe of social life. The transformative keys are as various as are the forms of radicalism: democracy (see *Democratic Theory*), the market, religious faith.

3. Attempts to Describe Radicalism as a Feature of Personality or Social Situation

Various attempts have been made to account for radicalism in terms of the psychological disposition of radicals. This generally has been done by those who do not welcome whatever project radicals may currently be pursuing. These attempts are vitiated by the manner in which a change in the political location of an actor or group can alter the descriptive appropriateness of the term. Stalinists can move from radicalism to conservatism simply by successfully seizing and ex-

exercising power. Their critical stance towards established authority comes to an end once they are themselves the embodiment of that authority.

4. *Radicalism, Conservatism, Pragmatism, and Traditionalism*

Since it goes to the roots, radicalism tends to ignore the complexity of the branches. It has, therefore, been contrasted with pragmatism, traditionalism, and conservatism. This contrast can be misleading. Pragmatism, traditionalism, and conservatism can be terms almost as contentless and historically contingent as radicalism. On the other hand, each of them operates as a matter of both historical fact and logical necessity with principles of selection, and none of them accepts unconditionally all aspects of existing arrangements and practices. Even the most stubborn conservative will have criteria for preferring Renoir to rats. Conversely, radicalism is a part of the tradition of many if not all societies, and the title of the English socialist and Christian Tawney's collection of essays, *The Radical Tradition* (1964) was only superficially paradoxical. It is, in other words, a conservative fallacy that tradition is homogeneous. Radicalism is one of its historically contingent components in most if not all societies. It is conversely a radical fallacy that radicalism seeks a transformation of all aspects of a society's life. A radical program seldom lacks foundations in some existing practices, institutions, or values, however peripheral or subordinated those may be at the time the program is advanced. And if radicals believe their own particular principles are deeply embedded in their society's traditions, they can recommend a return to roots rather than a departure from them. Franklin D. Roosevelt's remark, therefore, that the radical is someone with both feet firmly planted in the clouds, misses the point. The radical, rather, is someone who stands on their own patch of solid ground and wants the small cloud which they can see from that real but partial vantage point to encompass the whole sky.

5. *Radicalism and Reason*

The term 'radical' has been used most frequently to describe those who, whilst their criticism may be fundamental, do not reject the democratic and liberal methods of reason and persuasion in favour of insurrection or coup. Radicalism normally implies an appeal to reason, and therefore a dependence on persuasion, reasoned assent, and free individual choice. But appeals to other roots, instinct, faith, national myth are possible and have been made. Radicalism, as a matter of historical nomenclature, is more likely to be used to describe the ambitious advocacy of democracy or markets than of totali-

tarianism (see *Totalitarianism*) or authoritarianism. But it can describe Fascists who appeal to racial myth or religious extremists who demand that social life be determined, and law applied, on the basis of their interpretation of Christianity, Islam, or Hinduism.

6. *Radicalism and Rulers*

Were radicalism simply an abstract disposition, then it might be found to characterize equally rulers and their critics. But because of its association with challenges to existing arrangements, if it does characterize rulers, it is presented frequently by them as if they were Davids facing the Goliaths of custom or entrenched interest. Radicalism as a characteristic of government has usually laid great stress on the allegedly tremendous, and unfairly insuperable, obstacles in its path. Radicalism out of power has the real adversary of incumbent elites. Radicalism in power must to an extent invent enemies to maintain its radical credentials. It is not an accident that those most eager to use the powers of government to shape economy, society, or religion to their own favored roots are most likely to detect enemies within.

7. *Approbatory and Critical Uses of the Term*

'Radical' and 'radicalism' have been used with approbatory, pejorative, critical, or neutral tone according to the different historical locations in which they have been employed. Whereas in the 1960s and 1970s 'radical' was often used by conservatives to dismiss the proposals of their opponents, by the end of the twentieth century the term could be employed by left and right alike with entirely positive purpose. In the United Kingdom, Prime Minister Tony Blair, after the Labour Party's election victory of 1997, used radical to mean progressive, optimistic, forward looking, and subsumed socialism and liberalism under the title. The leader of the Conservative Party, William Hague, was equally able to recommend what he termed radical measures, for the term can mean thorough, imaginative, unhidebound.

8. *Historical Instances of Radicalism*

As an abstract term, therefore, radicalism lacks content, though it may have a distinguishable form, and can gain substance only in context. Its substantial meanings have, therefore, been historical and particular, proper names rather than taxonomic categories. 'Radical' as an adjective qualifying a noun has been more common than 'radicalism' as a noun by itself: radical liberalism in eighteenth century Europe; radical philosophy, the radical right, and radical feminism in the twentieth century.

Because any set of arrangements can be subjected to radical appraisal, the terms 'radical' and 'radicalism' have been applied to an extremely wide range of ideas and movements, from Lollards in fourteenth century Europe, Diggers and Levellers during the English revolution of the seventeenth century, to neo-Nazis in twenty-first century Europe and North American. The earliest substantial use was to describe democrats and liberals in the eighteenth century, those inspired by human rights, and by the examples of the American and French Revolutions. Radical liberalism and radical democracy were movements of active opinion that challenged the old order in Europe with demands based on individual social, civil, and political rights. In the nineteenth century, radicalism concentrated on the extension of political rights. The various fascist movements of the mid-twentieth century have been described as the radical right, in that they were in part an accentuation of aims and aversions already present in conservative and right wing politics. In the 1960s and 1970s, the New Left in Europe and North American was an instance of radicalism with a broad range of targets, but a narrower range of solutions from within which each segment chose its favorite: the abolition of capitalism, the overthrow of gerontocracy, the transformation of consciousness with drugs, or the decentralization and de-industrialization of western society. A different sort of radicalism, though one which also stressed individual liberation, was found within the New Right of the 1980s and 1990s. Here, the character of the radical demand was made more complex by a double track radicalism, on the one hand calling for a root and branch destruction of what were seen as restraints on the market and hence on individual utility maximization and self-fulfilment through choice about the use of property, but on the other calling for a radical imposition of cultural and moral controls, if necessary by government. By the end of the twentieth century radicalism was an aspect of environmental or green theory (see *Green Parties*), and of feminism. Radicalism had, too, religious avatars in the New Christian Right in the US, Islamism in Asia and the Middle East (see *Islamic Fundamentalism*), and political Hinduism in India. But in each case, within the variety of the movement, the breadth of targets was given coherence by the unity of ascribed causes for identified malaises.

9. Conclusion

Not only does the term radicalism lack any firm historical content, it possesses an equal normative flexibility, and is used neutrally and descriptively, and as a term of both approval and condemnation. For those who believe that language is the only reality we have, it provides a congenial illustration.

A radical may be a person with a vision, but it is usually a vision derived from experience, an extension

of the radical's own values to a whole society. There is a conservative foundation for radical politics, which is why, once radicals achieve power, they can appear to shift from change to conservation. Radicals are not for change as such, but only, like conservatives, for particular sets of arrangements, for the cultivation of which change may appear necessary, and for the achievement of which they believe they have identified a single cause or cluster of causes.

See also: Communism; Communist Parties; National Socialism and Fascism; Parties/Movements: Extreme Right; Revolutions, History of; Revolutions, Sociology of; Revolutions, Theories of; Social Movements, Sociology of

Bibliography

- Belchem J 1996 *Popular Radicalism in Nineteenth-century Britain*. Macmillan, Basingstoke, UK
- Biagini E F 1996 *Citizenship and Community: Liberals, Radicals and Collective Identities in the British Isles, 1865–1931*. Cambridge University Press, Cambridge, UK
- Blinkhorn M (ed.) 1990 *Fascists and Conservatives: The Radical Right and the Establishment in Twentieth-century Europe*. Unwin Hyman, London
- Inglis F 1982 *Radical Earnestness: English Social Theory 1880–1980*. Robertson, Oxford, UK
- Kaplan J, Weinberg L 1999 *The Emergence of a Euro-American Radical Right*. Rutgers University Press, New Brunswick, NJ
- Kramnick I 1990 *Republicanism and Bourgeois Radicalism: Political Ideology in Late Eighteenth Century England and America*. Cornell University Press, Ithica, NY
- Tawney R H 1964 *The Radical Tradition: Twelve Essays on Politics, Education and Literature*. George Allen & Unwin, London
- Wilks S 1993 *Talking About Tomorrow: A New Radical Politics*. Pluto/New Times, London

R. Barker

Radio as Medium

Radio, the transmission of sound via electromagnetic waves, is the most widely used electronic means of communication on the globe. It is second only to spoken language and nonverbal interaction as a medium of human communication. Despite the broad reach of television and the growing role of cable and Internet, radio remains an important arena of cultural expression, commerce, politics, and education in advanced capitalist countries. In many developing countries, including most of Africa, much of Asia, and the Middle East, and parts of Latin America, Eastern Europe, and Oceania, radio continues to be the primary means of mass communication. Due to its mobility, low cost, and flexibility, radio is one of the fastest growing media of communication worldwide.

Although radio has a variety of scientific and commercial applications as a technology of radio wave transmission (e.g., radio astronomy, radio telegraphy, radio telephony), as a social medium radio has been configured principally as a means of point-to-mass communication or broadcasting. Two broad models of radio broadcasting emerged during the first half of the twentieth century: commercial broadcasting and public monopoly broadcasting. While both models defined the radio wave spectrum as a public resource or good, they differed on whether the spectrum should be exploited by competing private entities for profit or developed by a single public entity as a nonprofit service. During the last quarter of the twentieth century both commercial broadcasters and community-based nonprofit broadcasters challenged the public monopoly model and launched new and hybrid broadcasting forms. This article examines the cultural, technological, economic, and political factors that shaped radio as a social medium of broadcasting.

1. *Technology and Cultural Form*

Radio is comprised of a dynamic sound technology wedded to a flexible yet resilient set of cultural practices. This combination of technological and cultural factors both enables and constrains radio as a medium. Radio production, for example, is shaped by the capacities of the technology (real-time, sound-only) as well as the forms of expression that have been adapted to the medium over time (program genres, formats, and presentation styles). Radio reception is likewise constrained. For example, a Walkman radio differs from a console radio in both its technological configuration and in the cultural assumptions that shape its use (mobile-individualized listening vs. situated-shared listening). This interaction of technology and culture shapes not only the form and content of the radio medium, but radio's relationship to time, space, community, and identity.

In this sense, a defining feature of radio as a technology and cultural form is its conflicted relationship to modernity. On the one hand, radio broadcasting represents a heightened sense of movement towards modernity; that is, movement toward a condition in which human social relations are increasingly abstract, individualized, and future-directed. To the extent that radio is understood to speed up processes of social progress and development (through, for example, the diffusion of innovations) the medium evokes a sense of 'hypermodernity.' On the other hand, radio provides a means of resisting or allaying the impact of increased mobility and rationalization in social life. To the extent that radio is used as a tool for building intimacy and community in an individuated and changeable modern world, broadcasting practices are 'antimodern.'

Radio broadcasting both reflects and constitutes uniquely modern forms of social organization. It works to integrate and coordinate an increasingly mobile, dispersed, and complex society, at the same time that it promotes the privatization of social life. The simultaneous access of vast numbers of isolated listeners to a single radio content creates a virtual common space that tends to become coterminous with the nation (due, in part, to government regulation and commercial market formation). In this way, radio, like print and cinema, acts as a medium for the 'community in anonymity which is the hallmark of modern nations' (Anderson 1991 p. 36).

At the same time, however, radio voices have a materiality that reproduces the immediate, tactile quality of primary orality in a number of ways. Roland Barthes's concept of the 'grain' of the voice is helpful in characterizing the somatic presence of radio voices. Barthes (1990) describes the grain as 'the body in the voice as it sings;' that is, the listener's sense of the physical presence of the performer in the sound of his or her voice (p. 299). Walter Ong (1995) argues that radio talk has the same ability as oral speech to envelope and immerse listeners, uniting the audience 'with themselves and with the speaker' (p. 74). Indeed, broadcasters actively work to recreate the sociability and interactivity of oral settings by developing intimate, chatty forms of radio talk. In this way, radio gives listeners a sense of continuity with older forms of intimacy and community at the same time that it enables distinctly modern forms of social life.

2. *History and Political Economy*

Radio has historically been deployed by a range of political and economic institutions largely within the bounds of the nation-state. During the last two decades of the twentieth century radio changed from a predominantly public service medium to an eminently commercial one. To understand the meaning of this global transformation it is necessary to investigate the social forces that guided and shaped radio broadcasting beginning in the 1920s. Among the most important influences on the medium have been national and transnational communication corporations, the state (nationalist and expansionist) and community groups.

2.1 *Beginnings: The 1920s and 1930s*

Radio expanded from a means of point-to-point communication to a medium of point-to-mass communication, or broadcasting, during the 1920s and 1930s. The rise of broadcasting was both shaped by and reflective of larger political and economic changes in the first quarter of the twentieth century, including the rise of transnational communication corporations following World War I (e.g., Philips, RCA, Tele-

funken), the growth of product marketing, increased trade in cultural goods (e.g., film and recorded music), and the escalation of both nationalism and international expansion. Technological developments including voice broadcasting (replacing Morse code) and high-power radio transmissions were also necessary ingredients for broadcasting.

Approaching the medium as a strategic communication technology, scarce resource, and tool for national integration, central governments began asserting control over the radio spectrum in the 1910s and 1920s. In part, this followed the pattern established by state intervention in telegraph and telephone development. As a US Commerce Department publication lamented, 'When the social possibilities of broadcasting became apparent many countries determined upon maintaining it as a public utility, particularly those where government ownership of communication services had previously been the rule' (Batson 1932, p. 2). Even in the USA, however, the central government claimed radio as a public resource and used regulatory powers to promote a national broadcasting system with citizenship and public service requirements for broadcasters. These regulations, codified in the 1927 Radio Act and again in the Communications Act of 1934, granted concessions or licenses for commercial broadcasting and actively encouraged the development of commercial networks on a national scale.

Radio was shaped by the climate of intense nationalism and international expansion in which it was born. In industrialized countries radio was viewed as a means of consolidating the nation as both market and polity, strengthening ties with colonies and dependencies and opening foreign markets. For example, the British Broadcasting Corporation (BBC) began its Empire Service in 1932, while the Japanese government established broadcasting stations in occupied territories in both Korea and China. US media corporations such as the Radio Corporation of America, General Electric, and Westinghouse financed broadcasting stations in foreign markets in order to promote the sale of radio receivers (e.g., Latin America). In less industrialized countries and colonial territories broadcasting rarely reached beyond a few major cities and was often limited to the interests of colonial governments (e.g., Hong Kong, Kenya).

In post-revolutionary societies such as Mexico and Soviet Russia, as well as fascist countries like Germany and Italy, the state turned to radio as a tool for building political solidarity and cultural identity through official broadcasting projects and the distribution of radio sets for both public and in-home listening. In Nazi Germany inexpensive radio sets known as the 'People's Wireless' or 'Goebbels' Lips' allowed listeners to tune in to Adolf Hitler's speeches but blocked the reception of foreign stations (Euro-media Research Group 1997, p. 79). In almost all independent countries radio developed within a na-

tionalistic framework: Broadcasters were required to be citizens transmitting from national soil, the state was given special access to the medium on the grounds of national interest and security, and content was frequently regulated to promote and 'improve' the national culture (e.g., language and music requirements).

Although radio broadcasting took many different forms, two broad types of radio organization—commercial broadcasting and public monopoly broadcasting—emerged during this period. In the commercial model, radio programs were designed to attract large audiences that could be sold to consumer product advertisers through program sponsorship or 'spot' advertisements. This model predominated only in the USA and Latin America, although Canada, Australia, and New Zealand had significant commercial radio sectors and small pockets of commercial radio existed elsewhere. In France and Luxembourg, for example, private broadcasters transmitted commercial programs in direct competition with the BBC (e.g., Radio Luxembourg, Radio-Normandie). In the vast majority of countries and occupied territories radio was a public monopoly (or modified monopoly) financed by government subsidy or receiver license fees and subject to varying degrees of government control. These two general models yielded different geographies of broadcasting: Whereas public monopolies tended to produce a small number of stations serving the national community (with limited regional or local service), commercial systems tended to generate a large number of stations serving the local community. With the rise of national networks, however, local commercial affiliates developed a strong national orientation.

Only countries in North America and Western Europe approached universal radio access by the end of the 1930s when, for example, about 75 percent of the British population had radios. One of the most influential broadcasts of the era, Franklin Roosevelt's first 'Fireside Chat' of March 18, 1933, was heard by an estimated audience of 60 million (out of a total population of about 120 million). Urban and industrializing areas of Latin America, Oceania, and Asia, including the Soviet Union, had small but growing numbers of radio listeners. Rural parts of these regions, along with most of Africa and the Middle East, had negligible radio listenership.

2.2 *Take-Offs and Transitions: The 1940s and 1950s*

World War II had a major impact on radio organization, content, and listenership. Timely broadcasts of news and information greatly increased the utility of the medium, promoted frequent listening and raised demand for radio sets both during and after the war. Radio listening grew even in countries where most people could not afford radio sets (e.g., Egypt) 'because merchants or owners of coffee shops would

have a set to attract customers' (Dabbous 1994, p. 65). The role of the central state in broadcasting also increased significantly on both the domestic and international fronts. At home, the state took an interest in the security and efficiency of the radio system. Official reports and government-sponsored programs proliferated on both public monopoly and commercial radio stations. In some cases (e.g., Vichy France) the state used the wartime emergency to prohibit private broadcasting and institute a public monopoly system.

International broadcasting was exploited by all of the belligerents during the war, largely in response to Italy and Germany who invested heavily in short-wave broadcasting. The BBC, for example, became the medium for Charles de Gaulle's appeals to the French Resistance, beginning with his celebrated broadcast of June 18, 1940. The US government developed international broadcasting projects aimed at Asia, Europe, and Latin America. In the case of Latin America, the US Coordinator of Inter-American Affairs developed a wartime broadcasting offensive that worked to solidify US hegemony in the region while strengthening the cultural influence of US advertisers and a small number of Latin American broadcasters. In 1948 the Coordinator's radio unit was transferred to the US State Department where it became the Voice of America (VOA) broadcasting service. By the end of the war, international broadcasting became the provision of the central state.

Radio broadcasting became one of the most important ideological and political tools of the Cold War. Along with international broadcasting, the USA, Soviet Union, and their allies sponsored clandestine radio stations in countries around the world (e.g., Guatemala, Turkey, Yugoslavia). The mobility and low cost of radio, however, also made it a powerful tool for grassroots revolutionary movements in the postwar period. Communist insurgents in China in the late 1940s and revolutionaries in Cuba in the late 1950s used clandestine broadcasting stations to maneuver against their enemies and mobilize popular support. Fidel Castro's Radio Rebelde, for example, became a model of effective clandestine broadcasting for decades to come.

Most of the countries that fought for independence at the end of World War II (e.g., China, India, Indonesia) followed a path similar to earlier postrevolutionary societies: They established state-directed broadcasting systems that controlled language and content and used the medium for nationalist ends. New governments struggled to improve broadcasting coverage and increase listenership quickly and efficiently. Consequently, broadcasting services were often highly centralized and limited to one or two national services (e.g., the Central People's Broadcasting Station in Beijing, China).

The history of radio in the postwar period was also shaped by the rise of a second broadcasting medium:

television. Although television fundamentally transformed radio broadcasting in the USA during this period, television's impact was much more gradual worldwide. Radio remained a public monopoly in most countries and a strong competitor for the national audience. In parts of Europe (e.g., Italy, Norway, Sweden) television failed to reach a majority of the population until the early 1960s and in Latin America the high cost of television guaranteed that radio would continue as the primary mass medium. Indeed, this period can be characterized as a golden age of radio in countries with developed broadcasting systems. In Latin America, for example, popular music (e.g., mambos, rancheras, sambas) provided the primary material for broadcasting while the *radionovela* (drama serial) reached the height of its popularity. In both commercial and public monopoly systems this was an important period of genre formation, with the live production of musical variety programs, comedies, soap operas, novelas, news, and game shows filling the airwaves.

In the USA radio changed radically beginning in the early 1950s. In order to develop national television networks, broadcasting corporations like the National Broadcasting Company stripped capital, talent, and program genres from radio and ultimately dismantled their national radio networks. Without national network programming, radio stations were forced to reorient their activities to the local and regional level and cultivate new markets (e.g., teenagers, ethnic minorities, and out-of-home listeners). Radio content became increasingly diversified and localized as radio shifted to a repertoire of recorded music supervised by disk jockeys (DJs) who advertised products and promoted the radio station. By the late 1950s stations were catering to narrow market segments with increasingly standardized program formats (e.g., 'beautiful music', and 'top-40').

Overall the early postwar period was marked by two divergent trends in radio broadcasting. While richer Western countries moved away from radio and toward television as the primary broadcast medium, many newly independent, postcolonial countries looked to radio as the primary tool for national communication and unification projects. Although television came to symbolize development, it remained impractical for the vast majority of countries due to high costs and infrastructure limitations. Beginning in the postwar decades, then, a significant divide emerged and grew between countries that saw high levels of television penetration beginning in the second half of the 1950s and those in which radio remained the primary mass medium through the late 1970s and beyond.

2.3 Change and Competition: The 1960s and 1970s

Public monopoly broadcasting remained dominant during this period despite technological changes and

increasing competition from television. In the countries that emerged from the last major wave of decolonization (e.g., Algeria, Nigeria, Uganda) public monopoly broadcasting with strong government influence was the norm. In their study, *Broadcasting in the Third World*, Elihu Katz and George Wedell (1977) noted the tendency towards government control of radio broadcasting in many developing countries. In the case of Nigeria, for example, they observed that: 'The demands of a government needing to assert its authority over a country of 50 million people of diverse ethnic, linguistic, and cultural characteristics, between 70 and 80 percent of whom were illiterate, made closer control of broadcasting inevitable' (p. 83).

At the same time, however, public monopoly broadcasters faced growing criticism for controlling and limiting public information. In India, for example, a 1964 government report noted that 'Suspicion of official information has deepened in India because of an incorrect, even improper use of media for personalized publicity and an undue accent on [government] achievements ...' (Ninan 1998, p. 4). In France, during the political crisis of 1968, listeners turned to commercial radio stations outside of France (the peripheral stations) to hear balanced coverage of events that French radio covered with a progovernment bias. In many countries protests against the public monopoly system were registered in the form of unlicensed, pirate stations operating within the national territory. While some countries managed to stave off reforms, significant liberalization began during this period. For example, Italy legalized private broadcasting in 1976 and Sweden permitted noncommercial 'neighborhood radio' as early as 1978.

During the same period commercial broadcasting (both television and radio) came under attack as a wasteland of low-quality programming that failed to meet basic public needs for information, education, and culture. In the USA, legislators responded with the Public Broadcasting Act of 1967 which created a public corporation that helped to finance educational television as well as National Public Radio, a public service national radio network. In Latin America, where radio became accessible to the majority of the population during this period, most governments either initiated or greatly expanded the use of radio in national education projects. Overall, this period witnessed a sizable growth in educational and nonprofit broadcasting by diverse civic groups, including religious organizations, universities, and other nonprofit groups.

Although radio listening increased worldwide during this period, the divide between rich and poor countries continued. Radio's growth in the developed world was driven, in part, by innovations such as frequency modulation broadcasting (FM) and transistor radios. While FM radio made high-fidelity broadcasting possible, inexpensive transistor sets increased radio's portability outside of the home and

facilitated the expansion of car radio. In the USA, for example, both innovations helped to push the number of radio sets to 1,813 per 1,000 population by 1975. In the developing world, however, FM expanded much more slowly due to the costs of conversion from amplitude modulation (AM) receivers. In addition, although inexpensive transistor sets made radio much more accessible to communities without electricity, the price of batteries remained prohibitively high for many listeners. Thus radio penetration remained relatively low in poorer regions of the world at the end of the decade. Asia and the Middle East, for example, had 62 sets per 1,000 population while Africa had 70 sets per 1,000 population in 1977 (Katz and Wedell 1977, p. 62).

2.4 *Liberalization, Deregulation, and Consolidation: The 1980s and 1990s*

The move to reform public monopoly broadcasting that began in the 1970s grew enormously during the 1980s. Liberalization was a response to a variety of pressures: social pressure to increase political and cultural pluralism over the air, political pressure to reduce public spending (especially during a time of economic crisis), and ideological pressure in favor of free market capitalism. This process was particularly evident in Western Europe, where local noncommercial stations were legalized in the early 1980s, followed by commercial stations in the mid- to late-1980s (e.g., Belgium, Denmark, FRG). In many cases, the radio system was opened to allow competition at the local level while the public monopoly on national radio broadcasting remained in place (e.g., the UK).

After the break-up of the Soviet Union and the Eastern Bloc in the late 1980s, most of the world's radio systems faced liberalization of some kind. Much of Eastern Europe became an unregulated market open to competition from North American and European companies (although, in fact, development remained limited due to severe economic dislocation in the region). In Africa most countries opened their radio systems to commercial broadcasting, although commercial development was also highly circumscribed by economic conditions. South Africa, for example, began to liberalize its broadcasting system in the mid-1990s with the privatization of some government-owned stations, licensing of new commercial stations, and the establishment of community broadcasting. Only in the Arab world did monopoly systems remain largely unscathed.

What began as liberalization in the 1980s became a move toward market regulation in the 1990s. As commercial broadcasters grew and faced local competition, they moved toward networking, syndication, and consolidation at a national level. In Sweden, for example, although the Radio Act of 1996 prohibited multiple ownership 'all but a few of the more than 90

commercial stations [were] closely integrated into four more or less national networks' (Euromedia Research Group 1997, p. 218). A similar process took place in France where a political movement in opposition to the public broadcasting monopoly was ultimately exploited by commercial interests. Whereas the 1980s had been dominated by small, activist stations, the 1990s witnessed the 'oligopolization of the private radio sector in France' at the hands of professional broadcasting corporations (Kuhn 1995, p. 107).

The move towards market regulation also produced oligopoly in the US radio industry. Beginning in the early 1980s, deregulation removed federal content regulations and reduced the limits on cross-ownership of media. With the Telecommunications Act of 1996 all national limits on station ownership were removed, opening the door to a series of mergers and consolidations. By the year 2000 radio broadcasting had moved from monopolistic competition to oligopoly: Many major cities had one station owner with 40 percent of the market and some had two owners with 70 percent of the market. According to Charles Fairchild (1999) the 1996 deregulation aimed, in part, to make radio competitive with other consolidated media industries, thus preparing the radio system for exploitation by integrated media conglomerates. At the same time, developments such as Internet radio and digital audio broadcasting (DAB) provided ideal media platforms for the convergence of recorded music, text, and image through a real-time broadcasting-style format.

Radio reached almost every person on the globe by the end of the twentieth century. Radio listening continued to grow in the world's expanding urban areas, primarily in the form of in-car and portable radio listening. Radio listening also grew in rural areas due to electrification projects as well as the development of low-tech, wind-up radio receivers that finally made radio affordable in areas without electricity. Radio's popularity also increased during this period as broadcasters exploited the potential interactivity and intimacy of the medium. Listener-participation talk shows reinvigorated the medium and promoted public discussion of personal topics (especially personal relationships) as well as political issues in countries as different as China, Mexico, and the USA.

In the growing commercial sector, radio stations delivered increasingly standardized program formats to targeted market segments. Format-specific radio chains developed a 'brand value' or 'lifestyle statement' that could be 'applied profitably to other aspects of their listeners' lives,' including other forms of media consumption (Hendy 2000, p. 47). Convergence and conglomeration in radio and related media industries (recorded music, television, Internet), as well as the growth of satellite delivery systems, set the stage for transnational deployment of increasingly 'branded' radio services. Although linguistic diversity and na-

tional regulatory policies continued to pose obstacles, regional radio integration was feasible in a few linguistically and economically unified regions (e.g., the Southern Cone of Latin America and the European Economic Community).

In contrast, the last decade of the twentieth century was also a boom period for decentralized community-based radio broadcasting. Community broadcasting, which emphasized community control and participation, was nurtured by democratization movements around the globe. The World Association of Community Radio Broadcasters (AMARC) provided a venue for the organization and promotion of community radio and lobbied for 'social enterprise' broadcasting as a means of encouraging economically viable community stations. Just such a model of community radio was initiated in South Africa in the second half of the 1990s, where nonprofit community broadcasters were allowed to support themselves through advertising. These community stations, some of which served populations of over 50,000, provided local service where none had existed before and covered events such as the local Truth and Reconciliation Hearings. By 2000, despite the vocal protests of the commercial broadcasting industry, even the United States Federal Communication Commission had approved noncommercial low-power FM broadcasting at the community level.

3. The Future of Radio Broadcasting

Radio's future is likely to unfold along two quite different paths in the early twenty-first century. First, radio's flexibility as a real-time, sound-only medium will ensure its survival as a social medium of broadcasting. Radio will continue to accompany people while they drive cars, work, and do other activities that require visual attention and mobility. Radio's flexibility and low cost will continue to drive the proliferation of small, decentralized stations at the local level. In less developed regions of the world this form of traditional radio broadcasting will continue to be a viable medium for development by commercial, governmental, and nonprofit groups.

Second, radio will tend to merge with other digital media. Both industrial conglomeration and technological convergence will work to collapse the distinction between radio, television, and Internet. This will occur primarily in advanced capitalist countries where radio will be integrated into multimedia formats that can be delivered via satellite, cable, and mobile radio telephony to privatized consumers. At the same time, a real-time 'radio' style of mass intimacy and anonymous interactivity will continue to shape electronic media and their contradictory relationship to modern life.

See also: Broadcasting: General; Broadcasting: Regulation; International Communication: History; International Communication: Regulation; Media Effects; Media Imperialism; Media, Uses of; Talk Show in Media; Television: History; Television: Industry

Strauss N (ed.) 1993 *Radiotext(e). Semiotext(e)*. Columbia University, New York
Williams R 1974 *Television: Technology and Cultural Form*. Schocken Books, New York

J. E. Hayes

Bibliography

- Anderson B R O'G 1991 *Imagined Communities*. Verso, London
Barthes R 1990 The grain of the voice. In: Firth S, Goodwin A (eds.) *On Record: Rock, Pop, and the Written Word*. Pantheon Books, New York, pp. 293–300
Batson L D 1932 *Radio Markets of the World*. Government Printing Office, Washington, DC
Dabbous S 1994 Egypt. In: Kamalipour Y R, Mowlana H (eds.) *Mass Media in the Middle East*. Greenwood Press, Westport, CT, pp. 64–8
Douglas S J 1999 *Listening In: Radio and the American Imagination: from Amos 'n Andy and Edward R. Murrow to Wolfman Jack and Howard Stern*. Times Books, New York
Euromedia Research Group (ed.) B. S. Ostergaard 1992 *The Media in Western Europe*. Sage, London
Fairchild C 1999 Deteritorizing radio: deregulation and the continuing triumph of the corporatist perspective in the USA. *Media Culture and Society* 21: 549–61
Hayes J E 2000 Radio. In: Balderston D, González M (eds.) *Encyclopedia of Contemporary Latin American and Caribbean Cultures*. Routledge, London, Vol. 1, pp. 1234–8
Hayes J E 2000 *Radio Nation: Communication, Popular Culture and Nationalism in Mexico, 1920–1950*. University of Arizona Press, Tucson, AZ
Hendy D 2000 *Radio in the Global Age*. Blackwell Publishers, Malden, MA
Hilmes M 1996 *Radio Voices: American Broadcasting, 1922–1952*. University of Minnesota Press, Minneapolis, MN
Hong L 1994 Radio Revival in China. *Beijing Review* 37: 15–18
Kamalipour Y R, Mowlana H (eds.) 1994 *Mass Media in the Middle East*. Greenwood Press, Westport, CT
Katz E, Wedell G 1977 *Broadcasting in the Third World*. Harvard University Press, Cambridge, MA
Kuhn R 1995 Radio. In: *The Media in France*. Routledge, London, pp. 77–108
Lent J A (ed.) 1978 *Broadcasting in Asia and the Pacific*. Temple University Press, Philadelphia, PA
McDaniel D O 1994 *Broadcasting in the Malay World*. Ablex, Norwood, NJ
Miller T 1997 Radio. In: Cunningham S, Turner G (eds.) *The Media in Australia*. Allen & Unwin, St. Leonards, NSW, pp. 47–69
Ninan S 1998 History of Indian broadcasting reform. In: Price M E, Verhulst S (eds.) *Broadcasting Reform in India*. Oxford University Press, Delhi, India, pp. 1–21
Ong W 1995 *Orality and Literacy*. Routledge, London
Ritchie S E 1999 The impact of community radio broadcasting on rural development in South Africa. In: *Radio Africa '99: Proceedings of the Third Regional Workshop on Radio Communication in Africa*. University of Botswana, Gaborone, Botswana, pp. 134–7
Scannell P, Cardiff D 1991 *A Social History of British Broadcasting*. Blackwell, Oxford

Random Assignment: Implementation in Complex Field Settings

Random assignment of units to treatment and control groups allows researchers to make the strongest possible causal connection between the treatment and observed outcomes. This article considers the way in which random assignment is implemented in educational and social field settings.

1. Widespread Acceptance of Random Assignment

The last three decades of the twentieth century have seen widespread acceptance of random assignment as an important aspect of high-quality research studies designed to test the effectiveness of interventions. In the 1970s and 1980s, proponents of randomized designs for studying program impacts began to assemble a literature attesting to their feasibility and frequent use (Boruch 1974, Boruch et al. 1978, Boruch and Wothke 1985). By the 1990s, randomized experiments in the social sciences had become so numerous that Greenberg and Shroder (1997) prepared a 500-page compendium of recent experiments in various content areas. The US Congress now routinely specifies that random assignment be used in studies of the effectiveness of federally funded programs, and recently called for a random assignment study of Head Start, America's flagship social program for disadvantaged children (Advisory Committee on Head Start Research and Evaluation 1999).

Randomized experiments have limitations as research vehicles (see Cook and Campbell 1979 for a discussion of the strengths and weaknesses of randomized experiments), and they are not universally accepted as the best way to learn about program effectiveness. One sector of the evaluation community espouses the virtues of qualitative research in place of randomized experiments and other quantitative approaches (e.g., Guba and Lincoln 1982), while another sector is more interested in issues of generalizability than internal validity (e.g., Cronbach 1982). Cook (1999) has written about the anti-random assignment culture that exists in American schools of education. Still, there is no denying that American society relies heavily on randomized experiments as one source of information about the effectiveness of social programs (St. Pierre and Puma 2000).

2. The Purpose of Random Assignment

As long as researchers take the time to explain the strengths and weaknesses of random assignment studies, those who commonly question experiments are easily able to understand the rationale underlying random assignment. Perhaps most helpful is the argument that random assignment is a special purpose tool, and while it is good only for one purpose, for making causal inferences, it is far better for that purpose than any alternative. As a special purpose tool, random assignment has many limitations. For example, it does not assure group comparability over time (attrition from the evaluation sample may result in nonrandom analytic groups), it cannot ensure that subjects remain in the intervention being studied (with or without random assignment, interventions may incur high dropout rates), and it does not help with issues of generalizability (a one-site random assignment study is no more generalizable than a one-site quasi-experiment). Cook and Campbell (1979) stated it clearly: 'Though random assignment is germane to the research goal of assessing whether the treatment may have caused any observed effects, it is conceptually irrelevant to all other research goals' (p. 343).

3. Common Objections to Random Assignment

Most of the common objections to random assignment in studies of program effectiveness do not dispute its utility as a tool for understanding causal connections. Instead, objections take other forms. Some are made in an attempt to avoid participation in a research study. Take the oft-made claim that 'random assignment is unethical' or that 'random assignment means denying services to some subjects.' These are political statements, either made without understanding what random assignment means, or made in an attempt to retain control over which subjects are to be served by a given social program.

When there is excess demand for a social intervention, program operators are in the position of denying services to needy families every day as they make decisions about which families are the neediest or the most deserving. As long as a randomized experimental study of such an in-demand program does not reduce the total number of subjects served, then there is no denial of services—the same aggregate number of families are served with or without an experiment. What an experiment does change is the mechanism for selecting which families receive the services; random assignment is used to decide who participates in the program instead of the judgment of program staff. Thus, when program operators say that random assignment is unethical or involves denying services, they usually are saying that they don't want to give up the power to determine which families are served.

But what about denying services to families? Is this always unethical? Suppose a funding agency wants to learn about the effectiveness of one of its social programs and in the absence of evidence of effectiveness, the funding agency plans to withdraw its support. Suppose further that there is no waiting list and hence that a randomized experiment would make it impossible to maintain aggregate service levels since some families would be assigned to a control group. In this case, the ethical problem of doing an experiment which reduces the number of families serviced by a social program has to be weighed against the possibly greater ethical problem of not doing an experiment and risking reduced funding for the entire program.

An underlying, and typically unvoiced, objection to random assignment is that program staff or funders assume the effectiveness of social programs instead of worrying about whether scarce resources are being spent in the most appropriate manner. This is understandable, since scientific skepticism seems to counteract the ability of program operators to do their best work. But again, this is a political issue—that it is appropriate to assume the effectiveness of a social program instead of subjecting it to a rigorous test.

Another objection to random assignment is that assignment to a control group 'feels bad' to program staff and to the families they serve. The argument that randomization is like a lottery doesn't work in many social program settings. Why? Because a lottery usually has only one winner, and the losers can take solace in the fact that they are in the great majority. On the other hand, those assigned to the control group in a randomized study do not view their assignment the same way they view a lottery loss, since the odds of winning the lottery are miniscule, whereas the odds of being randomly assigned to the desired social program are usually 50/50.

Some other arguments are that randomized experiments are expensive (Levitan 1992) or that they take a long time. The response to this sort of complaint is, 'Compared to what?' A randomized experiment is expensive and time-consuming when compared with the alternative of doing no research at all. But this is rarely a realistic alternative. Many quasi-experimental alternatives are just as expensive (any design in which data are collected from a treatment group and a nonequivalent comparison group) and time-consuming (the common pre-post study) as randomized studies. Some alternative approaches are quick and inexpensive, for example, relying on data collected for other purposes. However, the fact is that none of these alternatives are capable of doing the one thing that randomized studies excel at—providing strong evidence about causal connections. The problem with using alternatives to random assignment to study program effectiveness is that even if they are less expensive, even if they require less time, even if they are easier to implement, they rarely are worth the time and money spent on them. Why? Because the results

are not helpful for their intended purpose—for understanding program effectiveness. To paraphrase the common aphorism about education, ‘If you think randomized studies are expensive, try ignorance.’

4. Conditions Making Random Assignment Easier

There are several circumstances that reduce the difficulty of implementing random assignment studies. Suppose an intervention program is in short supply and there is a list of subjects waiting to participate. This often happens with Head Start (the largest federal preschool program in the US) projects in large metropolitan areas. If being placed on a waiting list means that a family is likely to miss a substantial portion of the Head Start year, then randomly assigning families from the waiting list to be in the program or in a control group might well be preferable to being on the waiting list with certainty. The absence of a waiting list does not mean that subjects are uninterested in a program. It often means that program staff do not want to spend time and energy on recruiting until they have openings. Hence, it is possible in short-supply situations for researchers to generate a waiting list by helping to recruit subjects and informing them about the planned study.

Suppose a treatment can be delayed with no ill effect on the participants. This is likely the situation in a school that wishes to test the short-term effectiveness of a one-semester school health curriculum. In this case, it would be possible to randomly assign students to take the school health course in the fall semester or in the spring semester.

Suppose a program’s proponents are skeptical and are searching for the best solution to a social problem instead of trying to promote a specific intervention. Or, suppose a demand for evidence of program effectiveness cannot be ignored, for example, a congressional mandate. In these cases, political objections to random assignment are eased because program implementers are supportive of or at least are willing to cooperate with the research.

Suppose that a social program has many local projects, for example, the Head Start program has about 1,600 local projects and the Even Start program (a family literacy program) has about 800 projects. An experimental evaluation of these programs could be done by selecting a relatively large number of projects and randomly assigning only a few subjects within each project. This approach minimizes the unpleasant effects of random assignment for any given project, making it more palatable to local program staff.

Assigning higher-level units is another way to ease the pain of random assignment. Suppose that there is an interest in testing three different solutions to a given social problem. One way to construct this test is to

randomly assign individuals to one of three interventions implementing the three different solutions. A different approach which avoids many of the common objections to random assignment is to randomize groups of individuals. For example, a federal agency could hold a grant competition to test three different approaches to preschool education. School districts would not be allowed to receive a grant unless they indicated their willingness to implement any of the three approaches. School districts that receive grants would then be randomly assigned to implement one of the three preschool approaches.

5. Working Towards Correct Random Assignment

Random assignment is more likely to be successfully implemented if done by researchers and research organizations with prior experience in conducting experimental studies. Certain research organizations have built the capacity to and reputation for conducting random assignment studies. Greenberg et al. (1999) report that almost half of the social experiments started in the US since 1983 have been done by three large research organizations.

In the same vein, random assignment is more likely to be successful if it is under the control of researchers rather than program implementers (Conner 1977). The latter are likely to make exceptions or may misunderstand random assignment rules, even if those rules are carefully prepared by researchers. The research team does not have to be physically present in order to control the random assignment. One approach is to have program staff be responsible for recruiting subjects, explaining the experimental alternatives, and transmitting lists of study subjects to the research team via fax, email, or other means. The research team then does the random assignment and sends back listings of research subjects, sorted into treatment and control groups. A related approach is for the research team to prepare a random assignment computer program for use by program staff who recruit subjects and enter basic data on a micro-computer which does the random assignment on the spot.

The way in which random assignment is implemented depends on the way in which applicants are recruited for an intervention. Sometimes a large pool of applicants are recruited. In these cases, information on the entire pool can be sent to the research team at the same time, simplifying the random assignment. The process is more complicated when applicants enter a program on a rolling basis. In these cases, a random assignment algorithm has to be developed so that applicants are told their treatment/control status soon after they apply. However, simple assignment rules such as assigning the first applicant to the

treatment, the second to control, the third to treatment, etc. are easily manipulated by program staff who often have control over the flow of applicants. More complex assignment systems are less subject to local control. For example, random assignment of families to an intervention could be done in blocks of four, where the order of treatment and control assignment is random within each block.

6. Preserving Randomization

Implementing random assignment is difficult. Once achieved, how can it be preserved throughout the life of an evaluation? One issue has to do with when randomization takes place. This involves tradeoffs between generalizability and internal validity. Families interested in participating in the Even Start family literacy program could be randomized as soon as they are first recruited for the program, after they have had the evaluation explained to them and have indicated their willingness to accept any of the alternatives, or after they have completed a one-month 'tryout period' during which they can try Even Start and see whether they truly are interested in the program. Early randomization draws on the largest possible pool of families (increased generalizability) but risks losing many families (decreased internal validity) who find out that they are not assigned to the desired group or who leave the program after a tryout period. Late randomization winnows the pool of families (decreased generalizability) to those who are most likely to participate in the study (increased internal validity). Gueron and Pauly (1991) and Rieken et al. (1974) give additional examples.

Whether assigned to an intervention or a control group, subjects may refuse to cooperate with research measurements once an experiment has been started. Such attrition has the potential to destroy the comparability of the intervention and control groups. Clearly, it is best to use all means possible to avoid attrition of subjects. In some circumstances (e.g., studies of welfare programs), subjects can be sanctioned if they do not participate in research; in other cases, incentives for participation can be offered. Attrition is all but unavoidable in most social experiments since families move, are ill or otherwise unavailable for measurement activities, or simply refuse to participate in the research. Methods for determining whether attrition is related to the treatment and ways of dealing with this problem are discussed by Cook and Campbell (1979).

Are there ways to reduce the likelihood that subjects will refuse to go along with the randomization, or refuse to participate in the study? One approach is to plan the study so that subjects who agree to participate in the randomization have a better chance of getting the desired intervention than they would have if the

study had not existed. This works best when there is a waiting list and when the intervention has a large number of openings relative to the number of subjects needed for the study. A second approach is to offer a competing intervention to members of the control group. This could work in several different ways. One competing intervention could be a financial or in-kind incentive for participating in the study. Another competing intervention could be a reduced or modified version of the treatment being studied. Yet a third alternative could be to offer control group subjects the opportunity to enroll in the desired intervention at a later point in time.

Bibliography

- Advisory Committee on Head Start Research and Evaluation 1999 *Evaluating Head Start: A Recommended Framework for Studying the Impact of the Head Start Program*. US Department of Health and Human Services, Washington, DC
- Boruch R 1974 Bibliography: Illustrative randomized field experiments for program planning and evaluation. *Evaluation* 2: 83-7
- Boruch R, McSweeney A J, Soderstrom E J 1978 Randomized field experiments for program planning, development and evaluation. *Evaluation Quarterly* 2: 655-95
- Boruch R, Wothke W 1985 Randomization and field experimentation. *New Directions for Program Evaluation* 28
- Conner R F 1977 Selecting a control group: An analysis of the randomization process in twelve social reform programs. *Evaluation Quarterly* 1: 195-244
- Cronbach L J 1982 *Designing Evaluations of Educational and Social Programs*, 1st edn. Jossey-Bass, San Francisco
- Cook T D 1999 *Considering the major arguments against random assignment: An analysis of the intellectual culture surrounding evaluation in American schools of education*. Paper presented at the Harvard Faculty Seminar on Experiments in Evaluation, Cambridge, MA
- Cook T D, Campbell D T 1979 *Quasi-experimentation: Design & Analysis Issues for Field Settings*. Rand McNally College Pub. Co., Chicago
- Greenberg D, Shroder M 1997 *The Digest of Social Experiments*, 2nd edn. The Urban Institute Press, Washington, DC
- Greenberg D, Shroder M, Onstott M 1999 The social experiment market. *Journal of Economic Perspectives* 13: 157-72
- Guba E G, Lincoln Y S 1982 *Effective Evaluation*, 1st edn. Jossey-Bass, San Francisco
- Gueron J M, Pauly E 1991 *From Welfare to Work*. Russell Sage Foundation, New York
- Leviton S A 1992 *Evaluation of Federal Social Programs: An Uncertain Impact*. George Washington University, Center for Social Policy Studies, Washington, DC
- Rieken H W, Boruch R F, Campbell D T, Coplan W, Glennan T K, Pratt J, Rees A, Williams W 1974 *Social Experimentation: A Method for Planning and Evaluating Social Innovations*. Academic Press, New York
- St. Pierre R G, Puma M J 2000 Toward the dream of the experimenting society. In: Bickman L (ed.) *Advances in Social Research Methods: The Heritage of Donald Campbell*. Sage Publications, New York

R. G. St. Pierre

Random Numbers

Random number generators (RNGs) are available from many computer software libraries. Their purpose is to produce sequences of numbers that appear as if they were generated randomly from a specified probability distribution. These *pseudorandom numbers*, often called *random numbers* for short (with a slight abuse of language), are crucial ingredients for a whole range of computer usages, such as statistical experiments, simulation of stochastic systems, numerical analysis with Monte Carlo methods, probabilistic algorithms, computer games, cryptography, and gambling machines, to name just a few. These RNGs are small computer programs implementing carefully crafted algorithms, whose design is (or should be) based on solid mathematical analysis. Usually, there is a basic uniform RNG whose aim is to produce numbers that imitate independent random variables from the uniform distribution over the interval $[0,1]$ (i.i.d. $U[0,1]$), and the random variables from other distributions (normal, chi-square, exponential, Poisson, etc.) are simulated by applying appropriate transformations to the uniform random numbers.

1. Random Number Generators

The words *random number* actually refer to a poorly defined concept. Formally speaking, one can argue that no number is more random than any other. If a fair dice is thrown eight times, for example, the sequences of outcomes 22222222 and 26142363 have exactly the same chance of occurring, on what basis could the second one be declared more random than the first?

A more precisely defined concept is that of independent random variables. This is a pure mathematical abstraction, which is nevertheless convenient for modeling real-life situations. The outcomes of successive throws of a fair dice, for instance, can be modeled as a sequence of independent random variables uniformly distributed over the set of integers from 1 to 6, i.e., where each of these integers has the probability $1/6$ for each outcome, independently of the other outcomes. This is an abstract concept because no absolutely perfect dice exists in the real world.

An RNG is a computer program whose aim is to imitate or simulate the typical behavior of a sequence of independent random variables. Such a program is usually purely deterministic: if started at different times or on different computers with the same initial state of its internal data, i.e., with the same *seed*, the sequence of numbers produced by the RNG will be exactly the same.

When constructing an RNG to simulate a dice, for example, the aim is that the statistical properties of the

sequence of outcomes be representative of what should be expected from an idealized dice. In particular, each number should appear with a frequency near $1/6$ in the long run, each pair of successive numbers should appear with a frequency near $1/36$, each triplet should appear with a frequency near $1/216$, and so on. This would guarantee not only that the virtual dice is unbiased, but also that the successive outcomes are uncorrelated.

However, practical RNGs use a finite amount of memory. The number of states that they can take is finite, and so is the total number of possible streams of successive output values they can produce, from all possible initial states.

Mathematically, an RNG can be defined (see L'Ecuyer 1994) as a structure (S, μ, f, U, g) , where S is a finite set of states; μ is a probability distribution on S used to select the initial state (or seed) s_0 ; the mapping $f: S \rightarrow S$ is the transition function; U is a finite set of output symbols; and $g: S \rightarrow U$ is the output function.

The state evolves according to the recurrence $s_n = f(s_{n-1})$, for $n \geq 1$. The output at step n is $u_n = g(s_n) \in U$. These u_n are the so-called random numbers produced by the RNG. Because S is finite, the generator will eventually return to a state already visited (i.e., $s_{i+j} = s_i$ for some $i \geq 0$ and $j > 0$). Then, $s_{n+j} = s_n$ and $u_{n+j} = u_n$ for all $n \geq i$. The smallest $j > 0$ for which this happens is called the period length ρ . It cannot exceed the cardinality of S . In particular, if b bits are used to represent the state, then $\rho \leq 2^b$. Good RNGs are designed so that their period length is close to that upper bound.

Choosing s_0 from a uniform initial distribution μ can be crudely approximated by using external randomness such as picking balls from a box or throwing a real dice. The RNG can then be viewed as an extensor of randomness, stretching a short random seed into a long stream of random-looking numbers.

2. Randomness from Physical Devices

Why use a deterministic computer algorithm instead of a truly random mechanism for generating random numbers? Simply calling up a system's function from a program is certainly more convenient than throwing dice or picking balls from a box and entering the corresponding numbers on a computer's keyboard, especially when thousands (often millions or even billions) of random numbers are needed for a computer experiment.

Attempts have been made at constructing RNGs from physical devices such as noise diodes, gamma-ray counters, and so on, but these remain largely impractical and unreliable, because they are cumbersome, and because it is generally not true that the successive numbers they produce are independent and uniformly distributed. If one insists on getting away

from a purely deterministic algorithm, a sensible compromise is to combine the output of a well-designed RNG with some physical noise, e.g., via bitwise exclusive-or.

3. Quality Criteria

Assume, for the following, that the generator's purpose is to simulate i.i.d. $U[0,1]$ random variables. The goal is that one cannot easily distinguish between the output sequence of the RNG and a sequence of i.i.d. $U[0,1]$ random variables. That is, the RNG should behave according to the null hypothesis $\mathbb{H}_0$: 'The u_n are i.i.d. $U[0,1]$,' which means that for each t , the vector $(u_0, \dots, u_{t-1})$ is uniformly distributed over the t -dimensional unit cube $[0,1]^t$. Clearly, $\mathbb{H}_0$ cannot be exactly true, because these vectors always take their values only from the finite set

$$\Psi_t = \{(u_0, \dots, u_{t-1}) : s_0 \in S\}.$$

If s_0 is random, Ψ_t can be viewed as the sample space from which the output vectors are taken. The idea, then, is to require that Ψ_t be very evenly distributed over the unit cube, so that $\mathbb{H}_0$ be *approximately true for practical purposes*. A huge state set S and a very large period length ρ are necessary requirements, but they are not sufficient; good uniformity of the Ψ_t is also needed. When several t -dimensional vectors are produced by an RNG by taking non-overlapping blocks of t output values, this can be viewed in a way as picking points at random from Ψ_t , without replacement.

It remains to agree on a computable figure of merit that measures the evenness of the distribution, and to construct specific RNGs having a good figure of merit in all dimensions t up to some preset number t_1 (because computing the figure of merit for all t up to infinity is infeasible in general). Several measures of discrepancy between the empirical distribution of a point set Ψ_t and the uniform distribution over $[0,1]^t$ were proposed and studied in the last few decades of the twentieth century (Niederreiter 1992, Hellekalek and Larcher 1998). These measures can be used to define figures of merit for RNGs. A very important criterion in choosing such a measure is the ability to compute it efficiently, and this depends on the mathematical structure of Ψ_t . For this reason, it is customary to use different figures of merit (i.e., different measures of discrepancy) for analyzing different classes of RNGs. Some would argue that Ψ_t should look like a typical set of random points over the unit cube instead of being too evenly distributed. That it should have a chaotic structure, not a regular one. However, chaotic structures are hard to analyze mathematically. It is probably safer to select RNG classes for which the structure of Ψ_t can be analyzed and understood, even if this implies more regularity,

rather than selecting an RNG with a chaotic but poorly understood structure.

For a concrete illustration, consider the two linear recurrences

$$\begin{aligned} x_{1,n} &= (a_{1,1}x_{1,n-1} + \dots + a_{1,k}x_{1,n-k}) \bmod m_1 \\ x_{2,n} &= (a_{2,1}x_{2,n-1} + \dots + a_{2,k}x_{2,n-k}) \bmod m_2 \end{aligned}$$

where k , the m_j 's, and the $a_{i,j}$'s are fixed integers, ' $\bmod m_j$ ' means the remainder of the integer division by m_j , and the output function

$$u_n = (x_{1,n}/m_1 - x_{2,n}/m_2) \bmod 1$$

This RNG is a *combined linear multiple recursive generator* (L'Ecuyer 1999a). Its state at step n is the $2k$ -dimensional vector

$$(x_{1,n}, \dots, x_{1,n-k+1}, x_{2,n}, \dots, x_{2,n-k+1}),$$

whose first k components are in $\{0, 1, \dots, m_1-1\}$ and last k components are in $\{0, 1, \dots, m_2-1\}$. There are thus $(m_1m_2)^k$ different states and it can be proved that Ψ_k is the set of all k -dimensional vectors with coordinates in $\{0, 1/m, \dots, (m-1)/m\}$, where $m = m_1m_2$. This implies, if m is large, that for each $t \leq k$, the t -dimensional cube $[0,1]^t$ is evenly covered by Ψ_t . For $t > k$, it turns out that in this case, Ψ_t has a regular lattice structure (Knuth 1998, L'Ecuyer 1998), so that all its points are contained in a series of equidistant parallel hyperplanes, say at a distance d_t apart. To obtain a good coverage of the unit cube, d_t must be as small as possible, and this value can be used as a selection criterion.

L'Ecuyer (1999a) recommends a concrete implementation of this RNG with the parameter values: $k = 3$, $m_1 = 2^{32} - 209$, $(a_{11}, a_{12}, a_{13}) = (0, 1403580, -810728)$, $m_2 = 2^{32} - 22853$, and $(a_{21}, a_{22}, a_{23}) = (527612, 0, -1370589)$. This RNG has two main cycles of length $\rho \approx 2^{191}$ each (the total number of states is approximately 2^{192}). Its lattice structure has been analyzed in up to $t = 48$ dimensions and was found to be excellent, in the sense that d_t is not far from the smallest possible value that can be achieved for an arbitrary lattice with $(m_1m_2)^k$ points per unit of volume. These parameters were found by an extensive computer search.

For a second example, consider the *linear congruential generator* (LCG) defined by the recurrence $x_n = ax_{n-1} \bmod m$ and output function $u_n = x_n/m$. This type of RNG is widely used and is still the default generator in several (perhaps the majority of) software systems. With the popular choice of $m = 2^{31} - 1$ and $a = 16807$ for the parameters, the period length is $2^{31} - 2$, which is now deemed much too small for serious applications (Law and Kelton 2000, L'Ecuyer 1998). This RNG also has a lattice structure similar to the previous one, except that the distances d_t are much larger, and this can easily show up in simulation

results (Hellekalek and Larcher 1998, L'Ecuyer and Simard 2000). Small LCGs of this type should be discarded.

Another way of assessing the uniformity is in terms of the leading bits of the successive output values, as follows. Suppose that S (and Ψ_t) has cardinality 2^e . There are tl possibilities for the l most significant bits of t successive output values $u_n, \dots, u_{n+t-1}$. If each of these possibilities occurs exactly 2^{e-tl} times in Ψ_t , for all l and t such that $tl \leq e$, the RNG is called *maximally equidistributed* (ME). Explicit implementations of ME or nearly ME generators can be constructed by using linear recurrences modulo 2 (L'Ecuyer 1999b, Tezuka 1995, Matsumoto and Nishimura 1998).

Besides having a long period and good uniformity of Ψ_t , RNGs must be efficient (run fast and use little memory), repeatable (the ability of repeating exactly the same sequence of numbers is a major advantage of RNGs over physical devices, e.g., for program verification and variance reduction in simulation; Law and Kelton 2000, and portable (work the same way in different software/hardware environments). The availability of efficient methods for jumping ahead in the sequence by a large number of steps is also an important asset of certain RNGs. It permits the sequence to be partitioned into long, disjoint substreams and to create an arbitrary number of virtual generators (one per substream) from a single backbone RNG (Law and Kelton 2000).

Cryptologists use different quality criteria for RNGs (Knuth 1998, Lagarias 1993). Their main concern is the unpredictability of the forthcoming numbers. Their analysis of RNGs is usually based on asymptotic analysis, in the framework of computational complexity theory. This analysis is for families of generators, not for particular instances. They study what happens when an RNG of size k (i.e., whose state is represented over k bits of memory) is selected randomly from the family, for $k \rightarrow \infty$. They seek RNGs for which the work needed to guess the next bit of output significantly better than by flipping a fair coin increases faster, asymptotically, than any polynomial function of k . This means that an intelligent guess quickly becomes impractical when k is increased. So far, generator families conjectured to have these properties have been constructed, but proving even the existence of families with these properties remain an open problem.

4. Statistical Testing

RNGs should be constructed based on a sound mathematical analysis of their structural properties. Once they are constructed and implemented, they are usually submitted to empirical statistical tests that try to detect statistical deficiencies by looking for empirical evidence against the hypothesis $\mathbb{H}_0$ defined previously. A test is defined by a test statistic T , function of a fixed set of u_n 's, whose distribution under

$\mathbb{H}_0$ is known. The number of different tests that can be defined in infinite and these different tests detect different problems with the RNGs. There is no universal test or battery of tests that can guarantee, when passed, that a given generator is fully reliable for all kinds of simulations. Passing a lot of tests may improve one's confidence in the RNG, but never proves that the RNG is foolproof. In fact, no RNG can pass all statistical tests. One could say that a bad RNG is one that fails simple tests, and a good RNG is one that fails only complicated tests that are very hard to find and run. In the end, there is no definitive answer to the question 'What are the good tests to apply?'

Ideally, T should mimic the random variable of practical interest in a way that a bad structural interference between the RNG and the problem will show up in the test. But this is rarely practical. This cannot be done, for example, for testing RNGs for general-purpose software packages. For a sensitive application, if one cannot test the RNG specifically for the problem at hand, it is a good idea to try RNGs from totally different classes and compare the results.

Specific statistical tests for RNGs are described in Knuth (1998), Hellekalek and Larcher (1998), Marsaglia (1985), and other references given there. Experience with empirical testing tells us that RNGs with very long periods, good structure of their set Ψ_t , and based on recurrences that are not too simplistic, pass most reasonable tests, whereas RNGs with short periods or bad structures are usually easy to crack by statistical tests.

5. Non-uniform Random Variates

Random variates from distributions other than the uniform over $[0, 1]$ are generated by applying further transformations to the output values u_n of the uniform RNG. For example, $y_n = a + (b-a)u_n$ produces a random variate uniformly distributed over the interval $[a, b]$. To generate a random variate y_n that takes the values -1 and 1 with probability 0.2 each, and 0 with probability 0.6 , one can define $y_n = -1$ if $u_n \leq 0.2$, $y_n = 1$ if $u_n > 0.8$, and $y_n = 0$ otherwise.

Things are not always so easy, however. For certain distributions, compromises must be made between simplicity of the algorithm, quality of the approximation, robustness with respect to the distribution parameter, and efficiency (generation speed, memory requirements, and setup time). Generally speaking, simplicity should be favored over small speed gains.

Conceptually, the simplest method for generating a random variate y_n from distribution F is *inversion*: define $y_n = F^{-1}(u_n) \triangleq \min\{y | F(y) \geq u_n\}$. Then, if we view u_n as a uniform random variable, $P[y_n \leq y] = P[F^{-1}(u_n) \leq y] = P[u_n \leq F(y)] = F(y)$, so y_n has distribution F . Computing this y_n requires a formula or a good numerical approximation for F^{-1} . As an example, if y_n is to have the geometric distribution with

parameter p , i.e., $F(y) = 1 - (1-p)^y$ for $y = 1, 2, \dots$, the inversion method gives $y_n = F^{-1}(u_n) = 1 + \ln(1 - u_n)/\ln(1-p)$. The normal, student, and chi-square are examples of distributions for which there are no close form expression for F^{-1} , but good numerical approximations are available (Bratley et al. 1987).

In most simulation applications, inversion should be the method of choice, because it is a monotone transformation of u_n into y_n , which makes it compatible with major variance reductions techniques (Bratley et al. 1987). But there are also situations where speed is the real issue and where monotonicity is no real concern. Then, it might be more appropriate to use fast non-inversion methods such as, for example, the *alias method* for discrete distributions or its continuous *acceptance-complement* version, the *acceptance/rejection* method and its several variants, *composition* and *convolution* methods, and several other specialized and fancy techniques often designed for specific distributions (Bratley et al. 1987, Devroye 1986, Gentle 1998).

Recently, there has been an effort in developing so-called automatic or black box algorithms for generating variates from an arbitrary (known) density, based on acceptance/rejection with a transformed density (Hörmann and Leydold 2000). Another important class of general non-parametric methods is those that sample directly from a smoothed version of the empirical distribution of a given data set (Law and Kelton 2000, Hörmann and Leydold 2000). These methods shortcut the fitting a specific type of distribution to the data.

6. A Consumer's Guide

No uniform RNG can be guaranteed against all possible defects, but the following ones have fairly good theoretical support, have been extensively tested, and are easy to use: the Mersenne twister of Matsumoto and Nishimura (1988), the combined MRGs of L'Ecuyer (1999a), the combined LCGs of L'Ecuyer and Andres (1997), and the combined Tausworthe generators of L'Ecuyer (1999b). Further discussion and up-to-date developments on RNGs can be found in Knuth (1998), L'Ecuyer (1998) and from the following web pages:

<http://www.iro.umontreal.ca/~lecuyer>
<http://random.mat.sbg.ac.at>
<http://cgm.cs.mcgill.ca/~luc/>
http://www.webnz.com/robert/rng_links.htm
 For non-uniform random variate generation, software and related information can be found at:
<http://cgm.cs.mcgill.ca/~luc/>
<http://statistik.wu-wien.ac.at/src/unuran/>

See also: Monte Carlo Methods and Bayesian Computation: Overview; Random Assignment: Implementation in Complex Field Settings

Bibliography

- Bratley P, Fox B L, Schrage L E 1987 *A Guide to Simulation*, 2nd edn. Springer, New York
- Devroye L 1986 *Non-uniform Random Variate Generation*. Springer, New York
- Gentle J E 1998 *Random Number Generation and Monte Carlo Methods*. Springer, New York
- Hellekalek P, Larcher G (eds.) 1998 *Random and Quasi-random Point Sets, Vol. 138, Lecture Notes in Statistics*. Springer, New York
- Hörmann W, Leydold J 2000 Automatic random variate generation for simulation input. In: Joines J A, Barton R R, Kang K, Fishwick P A (eds.) *Proceedings of the 2000 Winter Simulation Conference*. IEEE Press, Piscataway, NJ, pp. 675–82
- Knuth D E 1998 *The Art of Computer Programming, Vol. 2: Seminumerical Algorithms*, 3rd edn. Addison-Wesley, Reading, MA
- Lagarías J C 1993 Pseudorandom numbers. *Statistical Science* **8**(1): 31–9
- Law A M, Kelton W D 2000 *Simulation Modeling and Analysis*, 3rd edn. McGraw-Hill, New York
- L'Ecuyer P 1994 Uniform random number generation. *Annals of Operations Research* **53**: 77–120
- L'Ecuyer P 1998 Random number generation. In: Banks J (ed.) *Handbook of Simulation*. Wiley, New York, pp. 93–137
- L'Ecuyer P 1999a Good parameters and implementations for combined multiple recursive random number generators. *Operations Research* **47**(1): 159–64
- L'Ecuyer P 1999b Tables of maximally equidistributed combined LFSR generators. *Mathematics of Computation* **68**(225): 261–9
- L'Ecuyer P, Andres T H 1997 A random number generator based on the combination of four LCGs. *Mathematics and Computers in Simulation* **44**: 99–107
- L'Ecuyer P, Simard R 2000 On the performance of birthday spacings tests for certain families of random number generators. *Mathematics and Computers in Simulation* **55**(1–3): 131–7
- Marsaglia G 1985 A current view of random number generators. In: *Computer Science and Statistics, Sixteenth Symposium on the Interface*. North-Holland, Amsterdam, pp. 3–10
- Matsumoto M, Nishimura T 1998 Mersenne twister: A 623-dimensionally equidistributed uniform pseudo-random number generator. *ACM Transactions on Modeling and Computer Simulation* **8**(1): 3–30
- Niederreiter H 1992 *Random Number Generation and Quasi-Monte Carlo Methods, Vol. 63, SIAM CBMS-NSF Regional Conference Series in Applied Mathematics*. SIAM, Philadelphia, PA
- Tezuka S 1995 *Uniform Random Numbers: Theory and Practice*. Kluwer, Norwell, MA

P. L'Ecuyer

Ranke, Leopold von (1795–1886)

Leopold Ranke was born in Wiehe an der Unstrut in Thuringia on the December 20, 1795 as the son of the lawyer Gottlob Israel Ranke. His family had long served as evangelical pastors in the area of Saxony

between the rivers Saale and Unstrut. From 1809 to 1814 he attended the renowned Pforta school at Naumburg in Saxony, and from the spring of 1814 to the summer of 1818 he studied theology and classical literature in Leipzig. He completed his doctorate in 1817 with a (lost) dissertation on Thukydides. In 1818 Ranke started teaching history at the Fridericianum Gymnasium in Frankfurt an der Oder. In 1825 he was appointed associate professor of history at the University of Berlin, and in 1834 he was promoted to a full professor. In 1841 he was appointed official historian to the Prussian state by Friedrich Wilhelm IV, and in 1854 to the renewed parliament. Ranke was ennobled in 1865, becoming Leopold von Ranke, and in 1867 he was named chancellor of the order *pour le mérite*. He rejected countless offers to teach at other universities, at Munich among others, and remained in Berlin, where he died on the May 23, 1886. From 1880 onwards he was almost blind, and was forced to dictate his later works to assistants.

Ranke's career as a historian was influenced less by the historical science of the time, but rather by his interest in classical literature. His earliest existing studies are on the subject of translation and interpretation of classical Greek texts. Recent research has shown that the German literature of his time, Goethe, Schiller, and Friedrich Schlegel, also had an important influence on his conception of historical writing. It is also known that Ranke admired the literary form of contemporary British and French historiography. Ranke's early years were during the era of French hegemony over Germany and Europe, and the time of the rising German nationalist movement from 1806 onwards. These experiences of contemporary history left lasting impressions on Ranke's thinking. His political sympathies lay with Prussian state conservatism. On behalf of the Prussian government, but with independent control, Ranke published the *Historisch-politische Zeitschrift* from 1832 to 1836, through which the government aimed to exert a moderate conservative influence on the public. Ranke had to give up the journal after a short time because of lack of success; however, he did use the opportunity to publish a number of historical-political essays, in which he developed important methodological principles, and also began to sketch out his later, larger works. The most important of these articles are for example *Über die Zeit Ferdinands I. und Maximilians II.* (On the Time of Ferdinand I. and Maximilian II.), *Frankreich und Deutschland* (France and Germany), *Vom Einfluß der Theorie* (On the Influence of Theory) (all 1832), *Rom 1815–1823* (Rome from 1815 to 1823), *Die großen Mächte* (The Great Powers) (both 1833), and *Das politische Gespräch* (The Political Conversation) (1836). Ranke placed importance on his good relations with the Prussian court, but was also held in high esteem by the Bavarian King Maximilian II, who had attended his lectures as a student at the University of Berlin, and for whom Ranke held a series of free

lectures in Berchtesgaden in the September and October 1854 which were taken down in shorthand. The king had wished for 'a short and condensed overview of the motivating ideas of the various centuries of history from the Christian era on,' and Ranke developed basic characteristics of his historical perception in eight lectures which were then published in 1888 under the title *Über die Epochen der neueren Geschichte* (On the Epochs of Recent History).

Ranke's works total 63 volumes, including nine volumes of a world history which he dictated late in his life but was unable to complete. His first work *Geschichte der romanischen und germanischen Völker von 1494–1514* (History of the Latin and Teutonic Nations from 1494 to 1514) (1824) was a great success. It already showed some of the basic characteristics of his conception of Europe, and was of historiographical importance particularly because Ranke made an exemplary critical analysis of his sources in a separate volume, *Zur Kritik neuerer Geschichtsschreiber* (On the Critical Methods of Recent Historians). In this work he raised the method of textual criticism used in the late eighteenth century, particularly in classical philology to the standard method of scientific historical writing. In 1827 he visited the archives of Vienna, Venice, Rome, and Florence and various German libraries with a stipend of the Prussian government and discovered the since famous Venetian relations, legation reports from the sixteenth and seventeenth centuries, which became an important basic source for the series of his great historical works beginning in 1832. The experiences of the July Revolution in France in 1830 led him to interpret the modern history of the European states more strongly in perspective of the relation between the present and the past. From 1832 to 1836 he published the three volumes of *Die römischen Päpste, ihre Kirche und ihr Staat im 16 und 17. Jahrhundert* (Ecclesiastical and Political History of the Popes during the 16th and 17th Centuries), and from 1839 to 1847 the six volumes of *Die deutsche Geschichte im Zeitalter der Reformation* (History of the Reformation in Germany). These two works were developed in a single process and complement one another. Further main works followed in 1847–1848, *Preußische Geschichte* (History of Prussia), 1852–1861, *Französische Geschichte* (History of France), and 1859–1868, *Englische Geschichte* (History of England), as well as the incomplete world history of 1881–1888. After the foundation of the German Reich in 1870 and 1871, Ranke partly revised his *Preußische Geschichte* by emphasizing the 'German' role of Prussia in the old Reich more strongly in a second edition. Numerous studies concentrating for the most part on Prussian and Austrian history complete his work. It must be noted that the focal point of these works, with editions and investigations on the Prussian state chancellor von Hardenberg and on Friedrich Wilhelm IV, was in the area of contemporary history.

From 1830 onwards Ranke's thought centers on the history of the European states of the sixteenth and seventeenth centuries, based on his experiences of the time. However, depending on his estimation of the problems and their continuations, he did look back into the late and even the early middle ages, and on the other hand pursued the configurations of the European states not only up to 1789–1795, but also into the beginning of the nineteenth century. We can see from the focal point of the sixteenth and seventeenth centuries, the age of the secularized Renaissance papacy, the reformation and religious wars, the modern process of building states and the crystallization of the European powers system and its functioning between repeated hegemonic claims and equilibrium, that Ranke viewed the European historical process as essentially grounded in the tension between state and religion. The state and the church are buttresses of the system, at the same time however also buttresses of change, as they tend to broaden and reinforce their power so widely that they cause opposition and opposite tendencies, just as do the acting personalities themselves. In Ranke's historical perception however, each state and church, as well as their representatives, have a specific 'task,' with which they contribute to the overall human task of cultural development. The task of historians is to recognise this 'task' and the 'leading ideas,' and to portray at which points the protagonists deviate from them, causing turbulences and even historical catastrophes, and where they return to realizing these leading ideas and thereby to a harmonious equilibrium of the forces. Multiplicity is essential for the workings of this self-regulating system of states and religion, or churches. 'States are "Thoughts of God"' (*Politisches Gespräch*, ed. U. Muhlack, p. 95), and historical research pursues their development over the centuries. Ranke's great historical writings begin at the point when the medieval universalism of the church and the 'Sacrum Imperium' of the Germans started to break down and the specific European culture developed, in which the pretention of religious or state universality repeatedly appeared, but was crushed by the resistance of the state and national individualities. Multiplicity and unity are connected in this history of the great European nations and their summing together in the European powers system in such a way as to enable human energies to be realized more productively than ever before.

The recognition of this historical process has for Ranke, a value in itself. Ranke formulated this thought, which is a plea against liberal progressive thinking, repeatedly; most markedly in the sentence from the lectures '*Über die Epochen der neueren Geschichte*': 'Each age is immediate to God, and its value is not based on that which it brings forth, but in its own existence, in its own self, which however does not rule out that something else emerged from it' (von Ranke et al. 1971, p. 60). Ranke hereby opposes

every functionalization of historical events and situations for contemporary interests, yet makes it clear that the present cannot be understood without the historical process and that contemporary action may not ignore the way things have come to pass. For example one can learn from the past that is not possible to re-establish the absolute monarchy or the old importance of the church in the 'Age of Revolution.' Therefore Ranke did not limit himself to a purely contemplative ideal of historical insight beyond all practical applications. In his inaugural lecture in Berlin in 1836, *Über die Verwandtschaft und den Unterschied der Historie und der Politik* (On the Relation and the Difference of History and Politics), he declared that an insight into the past as free from prejudice as possible, was a condition for the possibility of reasonable politics in the present. Historical insight explains the conditions under which contemporaries act. However, these must be drawn into the leading ideas of their predecessors, and may not criticize them from the standpoint of their contemporary interests. Ranke also described this concept of historical objectivity in various works, the most well known of which being the formulation: 'We have given history the task of judging the past, of teaching the world for the use of future years: the present attempt does not forbid itself such high tasks: it merely aims to show how it really was' (SW 33, VII).

Ranke founded his ideal of objectivity for the most part on historical theology, not on logical insights. The basis of historical insight was for him a certain knowledge of God: 'History is religion' ... 'At least there is the closest connection between them' (letter to Georg Waitz, December 30, 1841, Briefwerk 313). As Ranke saw God as manifested in all historical appearances, and especially in their inner relations to one another, historical insight represented for him the specific modern and up to date search for the recognition of God. The religious conviction that all historical action is carried by providence meant on one hand a limitation of the possibilities of insight, because the historian can at best guess at God's will. On the other hand this conviction lends Ranke's historical perception a harmonistical character, which has earned sharp criticism in the past decades of historicism-critical historical science. His ideal of objectivism was rejected by his Berlin colleague and methodological opponent Johann Gustav Droysen but also by his pupil Heinrich von Sybel and by Heinrich von Treitschke, the protagonist of the politically active, Prussia-oriented liberal-conservative historiography in late nineteenth century Germany.

In the field of German history of the late nineteenth and early twentieth century, however, Ranke was very highly respected. His rather too uncritical successors unfortunately produced a perception of history which did not take sufficient account of the specific forces of the modern age. As mentioned previously, Ranke himself had always related history to the present and

had included the French Revolution as well as the re-establishment of the European system of powers at the Congress of Vienna in his continuity-oriented thinking. He saw in the French Revolution in particular, a dynamic element which gave the stagnating European powers system new energies and ability to integrate, so that European state system could regenerate and reinforce itself through increased nationalization. There was no longer room for industrialization, the modern competitive economy and capitalism, liberalism, even socialism and communism in this historical perception.

Leopold Ranke is counted along with Johann Gustav Droysen (1808–1884), Barthold Georg Neibuhr (1776–1831), and Theodor Mommsen (1817–1903) as one of the most important founders of modern history in Germany and as the most influential international representative of classical German historicism. His own research work and his teaching at the University of Berlin made an essential contribution to the circulation and final establishment of modern source of criticism. He pushed ahead the autonomization of history as a science, by distancing himself both from the rhetorical tradition of history as an ‘art,’ and also from the contemporary historical philosophy, particularly that of Georg Wilhelm Friedrich Hegel. With his systematic work from sources he emphasized the importance of empiricism, and simultaneously made it repeatedly clear in theoretical comments and in his historical practice than the individual event or fact can only be interpreted in its surrounding context. His art of portrayal was essential to his success and to his, at times, world recognition in the developing modern international field of history, with which style he has gone down in history as one of the founders of modern historiography. Ranke expressly aimed to make history into a ‘pleasure.’ His mastery of specifically narrative forms contributed to bridging the gap between the empirical and specialized historian and the bourgeois readership of the nineteenth century. As mentioned previously, one consequences of this esthetization could also be a harmonizing view of events and situations. However, Ranke’s narrative mastership did allow him to lead from the apparently coincidental individual event on to the contemplation of the most general insights, in a way that is very plausible for the historical-political consciousness of the nineteenth century. His stories deal with the responsibility of individuals for their fate, without suggesting that they could control this fate. That is the meaning of the sentence, that the historian is to deal with humanity, ‘as it is, explainable or unexplainable: the life of the individual, the generations, the peoples, at times the hand of God above them’ (SW 33/34, VIII).

See also: Historicism; Historiography and Historical Thought; Modern History (Since the Eighteenth

Century); History: Overview; Objectivity, History of; Religion, History of

Bibliography

- Berding H 1971 *Leopold von Ranke*. In: Wehler H-U (ed.) *Deutsche Historiker*, Göttingen, vol. 1, pp. 7–24
- Fulda D 1996 *Wissenschaft aus Kunst. Die Entstehung der modernen deutschen Geschichtsschreibung 1760–1860*. Berlin, pp. 296–410
- Hardtwig W 1991 Geschichtsreligion—Wissenschaft als Arbeit—Objektivität. Der Historismus in neuer Sicht. *Historische Zeitschrift* **252**: 1–32
- Hardtwig W 1997 Historismus als ästhetische Geschichtsschreibung. *Geschichte und Gesellschaft* **23**: 99–114
- Hinrichs C 1754 *Ranke und die Geschichtstheologie der Goethezeit*. Göttingen/Frankfurt Main/Berlin
- Iggers G 1971 *Deutsche Geschichtswissenschaft*. Munich, pp. 86–119
- Iggers G, von Moltke K (eds.) 1973 *The Theory and Practice of History, Leopold Ranke*. Indianapolis, New York
- Krieger L 1977 *Ranke. The Meaning of History*. Chicago, London
- Metz K-H 1979 *Grundformen historiographischen Denkens. Wissenschaftsgeschichte als Methodologie*. Dargestellt an Ranke, Treitschke und Lamprecht, Munich
- Muhlack U 1988 Leopold von Ranke. In: Hammerstein N (ed.) *Deutsche Geschichtswissenschaft um 1900*. Stuttgart, pp. 11–36
- Vierhaus R 1957 *Ranke und die soziale Welt*. Münster
- von Ranke L 1897–1890 *Sämtliche Werke*, 54 vols. [Ed. Dove A] Leipzig, Germany
- von Ranke L 1926 *Werke Gesamtausgabe der Deutschen Akademie* [ed. Joachimsen P] Munich
- von Ranke L 1949 *Das Briefwerk* [ed. Fuchs W P] Hamburg
- von Ranke L 1953 *Die römischen Päpste in den letzten vier Jahrhunderten*, with an introduction by V. Friedrich. Baethgen, Stuttgart
- von Ranke L 1957 *12 Bücher preußischer Geschichte* [ed. Andreas W] Hamburg, 4 vols.
- von Ranke L 1963 *Die großen Mächte, politisches Gespräch*. Göttingen
- von Ranke L 1964–75 *Aus Werk und Nachlaß* [eds. Fuchs W P and Schieder T]. Munich/Vienna, 4 vols.
- von Ranke L 1988 *Und die moderne Geschichtswissenschaft*. Stuttgart
- von Ranke L 1989 *Über die Epochen der neueren Geschichte*. Darmstadt
- von Ranke L 1995 *Die großen Mächte, politisches Gespräch*. Frankfurt Main/Leipzig

W. Hardtwig

Ranking Models, Mathematics of

Suppose that k items are ranked from best (1) to worst (k) by a nondeterministic process. For example, the observed ranking could represent an arbitrarily selected voter’s preferences for the $k = 5$ candidates in an election, or a judge’s ranking of $k = 4$ salad dressings by their perceived tartness. A probability model on rankings—or, more simply, a *ranking model*—specifies

the probability, $P(\pi)$, of observing each of the $k!$ rankings π . This article discusses many of the ranking models that have been proposed in the statistical and psychological literature, and interrelates their basic properties. The focus will be on three widely used methods for constructing ranking models, based on: (a) order statistics, (b) paired comparisons, and (c) distances between permutations.

Sect. 1 describes Thurstone order statistics models. In Sect. 2, several thematic properties of ranking models are introduced, and then these general properties are discussed in the particular context of Thurstone order statistics models. Sects. 3 and 4 consider ranking models induced by paired comparisons and distance-based ranking models, respectively. For each model, its underlying motivation is studied, as well as its relationship to the basic properties of Sect. 2.

Much of this material is from the survey article by Critchlow et al. (1991). Other useful general references include the comprehensive monograph by Marden (1995) and the volume edited by Fligner and Verducci (1993).

Some basic notation is as follows. Once the k items are labeled arbitrarily as item 1 to item k , a *ranking* of the k items corresponds to a permutation function π from $\{1, \dots, k\}$ onto $\{1, \dots, k\}$, where $\pi(i)$ is the rank assigned to item i , $i = 1, \dots, k$, and where $(\pi(1), \dots, \pi(k))$ is the associated vector of ranks. Alternatively, any ranking π can be described by the corresponding *ordering* of the k items, denoted by the bracketed vector $\langle \pi^{-1}(1), \dots, \pi^{-1}(k) \rangle$, where $\pi^{-1}(r)$ is the item assigned rank r , $r = 1, \dots, k$. Often, the simpler notation $\langle i_1, \dots, i_k \rangle$ will be used. For example, $\langle 3, 1, 2 \rangle$ denotes the ordering in which item 3 is ranked as best, item 1 is ranked as second best, and item 2 is ranked as worst.

1. Order Statistics Models

In his research on psychophysical phenomena, Thurstone (1927) hypothesizes that the actual perception of each item is a random variable. The relative ordering of these random variables then determines the judge's ordering of the items. Thurstone's original method considers only paired comparisons of the items, but Daniels (1950) extends the method to full orderings of the k items.

Formally, let X_i be a continuous random variable associated with item i , for $i = 1, \dots, k$. A random ranking π is defined by setting $\pi(i)$ equal to the rank, from smallest to largest, of X_i among $\{X_1, \dots, X_k\}$. In other words, the ranking π with ordering $\langle i_1, \dots, i_k \rangle$ is assigned the probability

$$P(\pi) = P\{X_{i_1} < \dots < X_{i_k}\} \quad (1)$$

If the random variables X_i are allowed to have arbitrary dependencies, then *any* probability distribution on rankings can be obtained from this con-

struction. To restrict the model, it is often assumed that the X_i are independent, and that their cumulative distribution functions $F_i(x) = F(x - \mu_i)$ constitute a location family, where F is fixed and where the μ_i ($i = 1, \dots, k$) are shift parameters. The resulting ranking model is referred to as a *Thurstone order statistics model* (see, e.g., Yellott 1977).

Two well-studied cases of the Thurstone model are the Thurstone (1927)–Mosteller (1951)–Daniels (1950) model (*TMD model*) where F is the standard normal distribution, and the Luce (1959)–Plackett (1975) model where $F(x) = 1 - \exp(-\exp x)$ is the Gumbel distribution. Brook and Upton (1974) use the TMD model to analyze voters' rankings of candidates for public office, while Henery (1981) uses the Luce–Plackett model to study outcomes of horse races. As discussed in the next section, the Luce–Plackett model affords a simple expression for the ranking probabilities given in Eqn. (1). Also, this model coincides with the first order model in Plackett's (1975) system of logistic models. Further development of order statistics models can be found, for example, in Yellott (1977), Henery (1983), Stern (1990), and Bockenholt (1993).

2. Properties of Ranking Models

This section introduces the study of ranking models with respect to certain thematic properties. In particular, the properties of reversibility, complete consensus, and L-decomposability—which are shared by many diverse types of ranking models—are motivated and defined. Each of these properties is then discussed in the context of Thurstone order statistics models. In later sections, the same thematic properties will be reconsidered in relation to additional classes of ranking models.

2.1 Reversibility

Let $\mathcal{P}$ denote a *class* of ranking models $\{P(\pi): P \in \mathcal{P}\}$, usually indexed by a set of parameters. Assume that a particular strategy of ranking items from best (rank 1) to worst (rank k) corresponds to a ranking model $P \in \mathcal{P}$. Suppose one now reverses the meaning of the ranks, so that rank 1 is given to the worst item and rank k is given to the best item. Is the resulting ranking model $P^*(\pi)$ still in the same class $\mathcal{P}$ —in other words, does P^* correspond to another choice of the parameters? If so, then the class $\mathcal{P}$ of ranking models is said to be *reversible* (Luce 1959 pp. 69–70, Marley 1968, 1982).

For fixed F , consider the class $\mathcal{P}_F$ of Thurstone order statistics models, with shift parameters $\mu_1, \dots, \mu_k$. Without loss of generality, assume that F has median zero. If the distribution F is symmetric, then the class $\mathcal{P}_F$ is reversible. Indeed, if particular parameter values

$\mu_1, \dots, \mu_k$ correspond to ranking from best to worst, then the values $-\mu_1, \dots, -\mu_k$ correspond to ranking from worst to best. Consequently, the class $\mathcal{P}_F$ of TMD ranking models (F standard normal) is reversible. On the other hand, Pendergrass and Bradley (1960) show that the class $\mathcal{P}_F$ of Luce–Plackett ranking models—with F the asymmetric Gumbel distribution—is not reversible (for recent work, see Yellott 1997).

2.2 Complete Consensus

An important problem is to identify a single ranking that best summarizes a probability distribution or sample of rankings (see, e.g., Marley 1993). By definition, a ranking model $P(\pi)$ is *unimodal*, with modal ranking π_0 , if π_0 uniquely maximizes P . Any ranking model satisfies this simple condition for unimodality, provided there is not a tie for the most likely ranking. However, for many ranking models, the items may be ordered in a much stronger sense than that implied by the mere existence of a modal ranking π_0 . For any two items i and j , suppose that $\pi_0(i) < \pi_0(j)$, and that $P[\pi(i) < \pi(j)]$, the probability that item i is preferred to item j , exceeds 0.5. If this probability continues to exceed 0.5 given any fixed assignment of ranks to the other $k-2$ items, then item i is *strongly preferred* to item j . If every pair of items can be related by strong preference, then the ranking model $P(\pi)$ is said to have *complete consensus* with respect to the modal ranking π_0 . In brief, the condition is that if item a is ranked ahead of item b in π_0 , then

$$P\{\langle i_1, \dots, a, \dots, b, \dots, i_{k-2} \rangle\} > P\{\langle i_1, \dots, b, \dots, a, \dots, i_{k-2} \rangle\}$$

for any fixed assignment of ranks to the other $k-2$ items $i_1, \dots, i_{k-2}$. The notion of complete consensus is discussed by Henery (1981), Fligner and Verducci (1988), and Critchlow et al. (1991).

One can construct distributions F for which the resulting Thurstone order statistics model does not even have a unique modal ranking (see Savage 1956). However, if the μ_i are distinct and if the density F' has monotone likelihood ratio, that is, if $F'(x-\mu_i)/F'(x-\mu_j)$ is decreasing in x for $\mu_i < \mu_j$, then the Thurstone order statistics model has the complete consensus property (Savage 1956, 1957), and the ordering of the items in the modal ranking coincides with the ordering of their associated parameters μ_i from least to greatest. Thus, for example, the TMD and Luce–Plackett models have complete consensus.

2.3 L-decomposability

The property of L-decomposability is a generalization of Luce's (1959) ranking postulate, defined in Eqn. (2)

below. If L-decomposability holds for a ranking model, then the ranking process can be viewed as a sequence of 'independent' stages: at stage 1 the most preferred item is chosen, at stage 2 the most preferred of the remaining items is chosen, and so forth until a full ranking is produced. The probability of the resulting ranking is assumed to be the product of the choice probabilities across the various stages.

Formally, for any subset B of the item labels $\{1, \dots, k\}$, and for each $i \in B$, let the *choice probability* $P_B(i)$ be the probability that item i is most preferred among those in B . Note that by definition, the choice probability $P_B(i)$ can depend only on i and B , and not on the relative ordering of any items not in B . Luce's ranking postulate states that for every ranking π with associated ordering $\langle i_1, \dots, i_k \rangle$

$$P(\pi) = P_{\{i_1, \dots, i_k\}}(i_1) \cdot P_{\{i_2, \dots, i_k\}}(i_2) \cdots P_{\{i_{k-1}, i_k\}}(i_{k-1}). \quad (2)$$

Any ranking model $P(\pi)$ that can be expressed in the form (2) for *some* collection of choice probabilities $P_B(i)$ is said to be *L-decomposable* (Critchlow et al. 1991, Fishburn 1994).

In formulating the ranking postulate, Luce had a *particular* family of choice probabilities $P_B(i)$ in mind—discussed in Eqn. (4) below—and from these he obtained a particular ranking model $P(\pi)$. On the other hand, the notion of L-decomposability is a slight generalization: given the ranking model $P(\pi)$, does there exist some collection of choice probabilities—not necessarily Luce's—that give rise to $P(\pi)$ via Eqn. (2)?

Critchlow et al. (1991) prove that a necessary and sufficient condition for L-decomposability is that for each rank $r = 3, \dots, k$ and for each $i_1, \dots, i_r$, the conditional probability

$$P[\pi^{-1}(r) = i_r | \pi^{-1}(1) = i_1, \dots, \pi^{-1}(r-1) = i_{r-1}] \quad (3)$$

is invariant under permuting the $i_1, \dots, i_{r-1}$. If a ranking model is L-decomposable, then it is easy to generate random rankings from the model, simply by generating the ranking stagewise, and using the choice probabilities at each stage.

The Luce–Plackett model is an important example of a Thurstone order statistics model that satisfies L-decomposability, and indeed this property is closely linked with Luce's (1959) original formulation of the model. Luce's approach is to construct choice probabilities $P_B(i)$ on the basis of his Choice Axiom, which implies that for each item i there is a value p_i —the 'merit' of item i —such that

$$P_B(i) = p_i / \sum_{j \in B} p_j \quad (4)$$

for any subset B of $\{1, \dots, k\}$. Without loss of generality, the p_i can be scaled to sum to 1, so that p_i is interpreted as the probability that item i is most preferred among the full set of items. In conjunction

with the ranking postulate in Eqn. (2), the choice probabilities in Eqn. (4) yield the ranking model

$$P\{i_1, \dots, i_k\} = \prod_{r=1}^{k-1} \left(p_{i_r} / \sum_{j \in B_r} p_j \right) \quad (5)$$

where $B_r = \{i_r, \dots, i_k\}$ is the set of items remaining at stage r . Moreover, it can be proved that this is indeed an alternative description of the Luce–Plackett model, since Eqn. (5) coincides with the Thurstone order statistics model with Gumbel F and with $\mu_i = -\log p_i$. See, for example, Yellott (1977), who attributes this result to Holman and Marley see also *Luce's Choice Axiom*).

3. Ranking Models Induced by Paired Comparisons

Babington Smith (1950) uses a conditioning argument for inducing a probability model on rankings from a set of arbitrary paired comparison probabilities. For each pair of items $i < j$, let p_{ij} be the probability that item i is preferred to item j ($i \neq j$) in a paired comparison of these two items. Imagine a tournament in which all of the $C(k, 2) \equiv k(k-1)/2$ possible paired comparisons are made independently. If the results of this tournament contain no circular triads ($h \neq i \neq j \neq h$), then the tournament corresponds to a unique ranking π of the items; otherwise the entire tournament is repeated until a unique ranking is obtained. The probability of any resulting π is thus given by

$$P(\pi) = C \prod_{\{(i, j): \pi(i) < \pi(j)\}} p_{ij} \quad (6)$$

where p_{ij} for $i > j$ is defined by $p_{ij} \equiv 1 - p_{ji}$, and C is a constant chosen to make the probabilities sum to 1.

With regard to the properties of Sect. 2, the class of all *Babington Smith models* in Eqn. (6) is reversible, since if the parameter values p_{ij} (for $i < j$) correspond to ranking from best to worst, then the values $p_{ij}^* = 1 - p_{ij}$ correspond to ranking from worst to best. For a Babington Smith model to have complete consensus, Critchlow et al. (1991) prove that sufficient conditions are that no $p_{ij} = 0.5$ and that the $\{p_{ij}\}$ are strongly stochastically transitive (i.e., $p_{ij} \geq 0.5$ and $p_{jm} \geq 0.5$ imply $p_{im} \geq \max\{p_{ij}, p_{jm}\}$; see David 1988)). Finally, to verify L-decomposability, it is straightforward to check that Eqn. (3) is invariant under permuting the $i_1, \dots, i_{r-1}$, for all models of the form Eqn. (6).

Although the Babington Smith model is indexed by $k(k-1)/2$ parameters, constraints on the $\{p_{ij}\}$ proposed by Mallows (1957) lead to three important

subclasses of the general Babington Smith model, having significantly fewer parameters. These subclasses are called the Mallows–Bradley–Terry (MBT) model, the rho-based model, and the ϕ -model, respectively.

To describe the MBT model, suppose the paired comparison probabilities have the Bradley–Terry (Bradley and Terry 1952) form

$$p_{ij} = v_i / (v_i + v_j)$$

for some nonnegative parameters $v_1, \dots, v_k$. Mallows suggests substituting the Bradley–Terry paired comparison probabilities into Eqn. (6), which leads to the *MBT ranking model*: for any ranking π with associated ordering $< i_1, \dots, i_k >$,

$$P(\pi) = C(\mathbf{v}) \prod_{r=1}^{k-1} (V_{i_r})^{k-r} \quad (7)$$

where $\mathbf{v} = (v_1, \dots, v_k)$ and the constant $C(\mathbf{v})$ is chosen to make the probabilities sum to 1. Since the Bradley–Terry paired comparison probabilities are strongly stochastically transitive, any MBT model has complete consensus, provided the v_i are distinct. Moreover, the ordering of the v_i from largest to smallest corresponds to the ordering of the items from best to worst in the modal ranking.

A further simplification of the MBT model has just one continuous parameter $\theta \in (0, 1)$ and a vector of discrete parameters $(\pi_0(1), \dots, \pi_0(k))$, where $\pi_0(i)$ is the rank assigned to item i by the modal ranking π_0 . Assuming that $v_i = \theta^{2\pi_0(i)}$ in Eqn. (7) gives Mallows's *rho-based model*

$$P(\pi | \theta, \pi_0) = C(\theta) \theta^{R^2(\pi, \pi_0)} \quad (8)$$

where

$$R^2(\pi, \pi_0) = \sum_{i=1}^k [\pi(i) - \pi_0(i)]^2$$

is the Spearman's rho-distance between the rankings π and π_0 , and is related to Spearman's (1904) correlation coefficient. The rho-based model has the interpretation that the probability $P(\pi)$ decreases geometrically, as the R^2 distance from π to π_0 increases. θ is a dispersion parameter: as $\theta \neq 1$, the rho-based model approaches the uniform distribution on all $k!$ possible rankings, whereas when $\theta \neq 0$, the model becomes increasingly concentrated about the modal ranking π_0 .

The final simplification of the general Babington Smith model is the well-known ϕ -model (Mallows 1957, Mann 1945). Unlike the rho-based model, the ϕ -model is not an MBT model. To describe the ϕ -model, suppose that in Eqn. (6) the corresponding paired

comparison probabilities have the simple form $p_{ij} = (\phi + 1)^{-1} \geq 0.5$ for $\pi_0(i) < \pi_0(j)$, so that the probability that item i is preferred to item j only depends upon their relative order in the modal ranking. Then the resulting ranking model is

$$P(\pi | \phi, \pi_0) = C(\phi) \phi^{T(\pi, \pi_0)} \quad (9)$$

where

$$T(\pi, \pi_0) = \sum_{i < j} I\{\pi(i) - \pi(j)[\pi_0(i) - \pi_0(j)] < 0\}$$

is the Kendall's tau-distance between the rankings π and π_0 , and $I(\cdot)$ is the indicator function: $I(A) = 1$ if the event A occurs, and $= 0$ otherwise. Thus, the ϕ -model and its parameters have interpretations analogous to the rho-based model, except the tau-distance is used instead of Spearman's rho-distance. The ϕ -model has been developed, applied, and extended subsequently by many authors, for example Feigin and Cohen (1978), Schulman (1979), Smith and Evans (1991), and Huh et al. (1995), to name just a few.

4. Distance-based Ranking Models

By letting $\lambda = -\log \phi$ in (8), the ϕ -model can be written in the form of an exponential family as

$$P(\pi | \lambda, \pi_0) = C(\lambda) \exp\{-\lambda T(\pi, \pi_0)\}$$

where the constant $C(\lambda) = [\sum_{\pi \in S_k} \exp\{-\lambda T(\pi, \pi_0)\}]^{-1}$. Diaconis (1988 p. 104) uses this representation to motivate a general class of metric-based ranking models

$$P(\pi | \lambda, \pi_0) = C(\lambda) \exp\{-\lambda d(\pi, \pi_0)\}$$

where d is an arbitrary metric on the group S_k of all permutations.

Diaconis (1988 pp. 112–9) considers the following metrics on S_k to be among the most widely used in applied scientific and statistical problems:

$$T(\pi, \nu) = \sum_{i < j} I\{\pi(i) - \pi(j)[\nu(i) - \nu(j)] < 0\}$$

is the Kendall's tau metric:

$$R(\pi, \nu) = \left(\sum_i [\pi(i) - \nu(i)]^2 \right)^{1/2}$$

is the Spearman's rho metric:

$$F(\pi, \nu) = \sum_i |\pi(i) - \nu(i)|$$

is the Spearman's footrule metric:

$$H(\pi, \nu) = \sum_i I\{\pi(i) \neq \nu(i)\}$$

is the Hamming metric:

$C(\pi, \nu)$ = minimal number of transpositions needed to transform π into ν

is Cayley's metric and

$U(\pi, \nu) = k -$ (the maximal number of items ranked in the same relative order by π and ν)

is Ulam's metric. All of these satisfy the usual metric axioms of positivity, symmetry, and the triangle inequality, and thus give rise to metric-based ranking models. Additionally, the class of metric-based models may be enlarged to include Mallows's rho-based model, Eqn. (8), by dropping the triangle inequality requirement for d . The resulting function d is called a distance rather than a true metric, and the resulting model is a distance-based ranking model (Fligner and Verducci 1986). Not much is lost in this generalization, although the triangle inequality property has a desirable interpretation with respect to the behavior of likelihood ratio tests for π_0 (Critchlow 1985 pp. 101–2).

Properties of distance-based ranking models are related to analogous properties of the distance d . Let $\mathcal{P}_d$ denote the class of all ranking models associated with a fixed distance d . Then $\mathcal{P}_d$ is reversible if the distance d satisfies

$$d(\pi^*, \nu^*) = d(\pi, \nu) \quad \text{for all } \pi, \nu \in S_k \quad (10)$$

where $\pi^*(i) \equiv k + 1 - \pi(i)$ and $\nu^*(i) \equiv k + 1 - \nu(i)$ are the permutations that rank the items in exactly the reverse order as π and ν , respectively. All of the distances T , R , R^2 , F , H , C , and U satisfy Eqn. (10), and hence give rise to reversible classes.

A distance-based model has complete consensus if and only if d possesses the transposition property (Critchlow 1985 pp. 101–2): $d(\pi, \nu) \leq d(\pi', \nu)$ whenever π , π' , and ν are permutations such that π and π' assign identical ranks to all items except for i and j , and ν agrees with π on the relative ordering of i and j . Since the distances T , R , R^2 , and F satisfy this transposition property, the models based on these distances exhibit complete consensus.

Finally, Critchlow et al. (1991) prove that a distance-based ranking model is L-decomposable if and only if the distance d is additively decomposable, in the sense that for each $r = 2, \dots, k$, there exist functions f_r and g_r such that

$$d(\pi, e) = f_r[\pi^{-1}(1), \dots, \pi^{-1}(r-1)] + g_r[\pi^{-1}(r), \dots, \pi^{-1}(k)]$$

where e is the identity permutation $e(i) \equiv i$. It follows that the models based on T , R^2 , F , and H are L-decomposable.

4.1 Extensions

Some important multiparameter extensions of the distance-based and ϕ -models include the ϕ -component models and multistage models of Fligner and Verducci (1986, 1988), as well as the orthogonal contrast models (Chung and Marden 1991, 1993, Marden 1992).

Critchlow (1985) develops distance-based models for partially ranked data, such as rankings of only the q best items out of k , where $q < k - 1$. The set of all partial rankings of the best q items is identified with the quotient space S_k/S_{k-q} of the permutation group, consisting of all right cosets of the subgroup $S_{k-q} = \{\pi \in S_k \mid \pi(i) = i, i = 1, \dots, q\}$. Any metric d on S_k induces a Hausdorff metric d^* on the quotient space S_k/S_{k-q} , which serves as the metric for these models. Critchlow gives the Hausdorff extensions of T , R , F , H , C , and U , and applies the resulting models to partially ranked data sets.

See also: Consensus Panels: Methodology; Non-parametric Statistics: Rank-based Methods; Order Statistics; Ordered Relational Structures; Scaling and Classification in Social Measurement; Unfolding and Vector Models

Bibliography

- Babington Smith B 1950 Discussion on Professor Ross's paper. *Journal of the Royal Statistical Society B* **12**: 153–62
- Bradley RA, Terry MA 1952 Rank analysis of incomplete block designs. I. *Biometrika* **39**(3): 324–45
- Bockenholt U 1993 Applications of Thurstonian models to ranking data. In: Fligner MA, Verducci JS (eds.) *Probability Models and Statistical Analyses for Ranking Data*. Springer-Verlag, New York, pp. 157–72
- Brook D, Upton G J G 1974 Biases in local government elections due to position on the ballot paper. *Applied Statistics* **23**: 414–9
- Chung L, Marden J I 1991 Use of nonnull models for rank statistics in bivariate, two-sample, and analysis-of-variance problems. *Journal of the American Statistical Association* **86**: 188–200
- Chung L, Marden J I 1993 Extensions of Mallows' ϕ model. In: Fligner MA, Verducci JS (eds.) *Probability Models and Statistical Analyses for Ranking Data*. Springer-Verlag, New York, pp. 108–39
- Critchlow DE 1985 *Metric Methods for Analyzing Partially Ranked Data*. *Lecture Notes in Statistics*, Vol. 34. Springer-Verlag, New York
- Critchlow DE, Fligner MA, Verducci JS 1991 Probability models on rankings. *Journal of Mathematical Psychology* **35**: 294–318
- Daniels HE 1950 Rank correlation and population models. *Biometrika* **33**: 129–35

- David H A 1988 *The Method of Paired Comparisons*, 2nd edn. Oxford University Press, New York
- Diaconis P 1988 *Group Representations in Probability and Statistics*. Institute of Mathematical Statistics, Hayward
- Feigin P, Cohen A 1978 On a model for concordance between judges. *Journal of the Royal Statistical Society B* **40**: 203–13
- Fishburn PC 1994 On 'choice' probabilities derived from ranking distributions. *Journal of Mathematical Psychology* **38**: 274–85
- Fligner MA, Verducci JS 1986 Distance-based ranking models. *Journal of the Royal Statistical Society B* **48**: 859–69
- Fligner MA, Verducci JS 1988 Multi-stage ranking models. *Journal of the American Statistical Association* **83**: 892–901
- Fligner MA, Verducci JS (eds.) 1993 *Probability Models and Statistical Analyses for Ranking Data*. *Lecture Notes in Statistics*, Volume 80. Springer-Verlag, New York
- Henery R J 1981 Permutation probabilities as models for horse races. *Journal of the Royal Statistical Society B* **43**: 86–91
- Henery R J 1983 Permutation probabilities for gamma random variables. *Journal of Applied Probability* **20**: 822–34
- Huh M-H, Critchlow DE, Verducci JS, Kiecolt-Glaser J, Glaser R, Malarkey WB 1995 A symmetric analysis of paired rankings with application to temporal patterns of hormonal concentration. *Biometrics* **51**: 1361–71
- Luce RD 1959 *Individual Choice Behavior*. Wiley, New York
- Mallows CL 1957 Non-null ranking models. I. *Biometrika* **44**: 114–30
- Mann HB 1945 Non-parametric tests against trends. *Econometrica* **13**: 245–59
- Marden J I 1995 *Analyzing and Modeling Rank Data*. Chapman and Hall, London
- Marley A A J 1968 Some probabilistic models of simple choice and ranking. *Journal of Mathematical Psychology* **5**: 311–32
- Marley A A J 1982 Random utility models with all choice probabilities expressible as 'functions' of the binary choice probabilities. *Mathematical Social Sciences* **3**: 39–56
- Marley A A J 1993 Aggregation theorems and the combination of probabilistic rank orders. In: Fligner MA, Verducci JS (eds.) *Probability Models and Statistical Analyses for Ranking Data*. Springer-Verlag, New York, pp. 216–40
- Mosteller F 1951 Remarks on the method of paired comparisons: I. The least squares solution assuming equal standard deviations and equal correlations. *Psychometrika* **16**: 3–9
- Pendergrass RN, Bradley RA 1960 Ranking in triple comparisons. In: Olkin I, Ghurye SG, Hoeffding W, Madow WG, Mann HB (eds.) *Contributions to Probability and Statistics: Essays in Honor of Harold Hotelling*. Stanford University Press, Palo Alto, CA, pp. 331–51
- Plackett RL 1975 The analysis of permutations. *Applied Statistics* **24**: 193–202
- Savage IR 1956 Contributions to the theory of rank order statistics—the two-sample case. *Annals of Mathematical Statistics* **27**: 590–615
- Savage IR 1957 Contributions to the theory of rank order statistics—the 'trend' case. *Annals of Mathematical Statistics* **28**: 968–77
- Schulman RR 1979 Ordinal data: An alternative distribution. *Psychometrika* **44**: 3–20
- Smith LA, Evans DA 1991 Estimating a population distribution of sequences of k items from cross-sectional data. *Applied Statistics* **40**: 31–42
- Spearman C 1904 The proof and measurement of association between two things. *American Journal of Psychology* **15**: 72–101
- Stern HS 1990 Models for distributions on permutations. *Journal of the American Statistical Association* **85**: 558–64

- Thurstone L L 1927 A law of comparative judgement. *Psychological Reviews* 34: 273–86
- Yellott J 1977 The relationship between Luce's choice axiom, Thurstone's theory of comparative judgment, and the double exponential distribution. *Journal of Mathematical Psychology* 15: 109–44
- Yellott J 1997 Preference models and irreversibility. In: Marley A A J (ed.) *Choice, Decision, and Measurement: Essays in Honor of R. D. Luce*. Erlbaum, Hillsdale, NJ, pp. 133–53

D. E. Critchlow

Rape and Sexual Coercion

The following summarizes the biological, and cultural approaches to the understanding of rape in human societies and evaluates these approaches by reference to studies demonstrating the incidence and social correlates of sexual coercion in the United States and cross-culturally.

1. The Biological View

The biological explanation for rape and sexual coercion is grounded in the argument that sexual aggression is an evolved adaptation in human males. The argument builds on Charles Darwin's theory of natural and sexual selection. Natural selection is the doctrine that in the struggle for existence evolutionary progress is achieved by the inheritance of advantageous characteristics which prosper at the expense of less advantageous ones. Sexual selection is a specific case of natural selection applied to the evolution of sex differences. Darwin claimed that the greater size and strength of human males was due 'in chief part to inheritance from his half-human male ancestors.' According to Darwin such characteristics 'would have been preserved or even augmented during the long ages of man's savagery, by the success of the strongest and boldest men, both in the general struggle for life, and in their contest for wives.' He measured success by the number of children some men left as compared with their 'less favoured brethren' (Darwin 1936, pp. 872–3).

In the nineteenth and early twentieth centuries, the Darwinian doctrine of sexual selection was applied to treatises on human sexual behavior produced by the famous sexologists R. F. von Krafft-Ebing and Havelock Ellis. In his magnum opus, *Psychopathia Sexualis*, first published in German in 1886, Krafft-Ebing argued that 'gratification of the sexual instinct [is] the primary motive in man as well as in beast' (quoted by Sanday 1996 p. 125). In the third volume of his *Studies in the Psychology of Sex*, published in Philadelphia between 1897 and 1910, Ellis comes very

close to making rape the *sine qua non* of sexual arousal. Ellis conceives of human sexual behavior as a game of combat. Playing the role of hunted animal the female conceals her sexual passion by adopting a demeanor of modesty in order that the male may be more ardent and forceful. Ellis claims that as the hunt becomes more sexually charged, 'an element of real violence, of undisguised cruelty' is introduced. Accepting the Darwinian theory of natural selection with its emphasis on competition and brute strength, Ellis claimed that a woman who resisted 'the assaults of the male' aided natural selection 'by putting to the test man's most important quality, force' (quoted by Sanday 1996, pp. 127–8) (see *Male Dominance* for more on the social meaning of sex differences in Western thought).

At the end of the twentieth century, the Darwinian approach was again applied to the subject of the 'biological bases of sexual coercion' by Randy Thornhill and Craig T. Palmer in their book *A Natural History of Rape* (2000). Like Krafft-Ebing and Ellis they make a distinction between sexual desire in human males and females. Citing the work of anthropologist Donald Symons (1979, pp. 264–7), these authors claim that rape is due to certain adaptations making males more easily aroused, with a greater sex drive, a reduced ability to abstain from sexual activity, in need of greater sexual variety, and less discriminating in choice of sexual partner (Thornhill and Palmer 2000, p. 62). However, while Symons suggests that rape is a by-product of such adaptations, Thornhill and Palmer suggest that rape is itself an adaptation, which they call 'the human rape-adaptation' (Thornhill and Palmer 2000, p. 62). By this they mean that rape has a direct biological basis rather than being a behavioral consequence of other evolved behaviors.

To support their argument for the evolved nature of such an adaptation, Thornhill and Palmer (2000, p. 62) say it is necessary to identify 'mechanisms involved in rape that were designed by selection in the past specifically for reproduction by means of rape'. Because such mechanisms exist in the morphology of certain insects, they seem to suggest, they must also exist in humans. To make their case, the authors draw parallels with scorpion flies, which they point out have a physical adaptation for rape. Male scorpion flies have an organ, 'a pair of clamp-like structures, one on either side of the penis,' that serves to keep unwilling females in a mating position (2000, p. 63).

Admitting that there is no 'conspicuous morphology that might be a rape adaptation' in human males, Thornhill and Palmer (2000, pp. 64–5) look 'to the male psyche for candidates for rape adaptations.' They propose (2000, pp. 65–6) a number of 'psychological mechanisms' as 'adaptations' that can be viewed as 'analogous' to the rape adaptations observed in 'male insects'.

In a review of the argument put forward by Thornhill and Palmer, Frans B. M. de Waal, a prominent biologist, points out that they have not demonstrated

that rape in humans is a product of natural selection. According to Waal, 'for natural selection to favor rape, rapists would have to differ genetically from nonrapists and need to sow their seed more successfully, so to speak, causing more pregnancies than nonrapists, or at least more than they would without raping' (Waal 2000, p. 24). 'Not a shred of evidence for these two requirements is present' Waal concludes (2000, p. 24). Waal also points out that the psychological mechanisms Thornhill and Palmer propose do not demonstrate natural selection but have to do 'with judgment of people and situations, a multi-purpose capacity also present in women' (2000, p. 24). Waal's emphasis on judgement leads us to the cultural argument.

2. The Cultural Approach

Against a strictly biological explanation of human behavior anthropologist Clifford Geertz (1973, p. 46) argues that without culture patterns (i.e. 'organized systems of significant symbols') human behavior would be shapeless and without direction, 'a mere chaos of pointless acts and exploding emotions.' Applied to sexual behavior, the cultural approach, accepted by many anthropologists and sociologists, suggests that sexual behavior is part of a broader cultural system which includes notions of masculinity and femininity, ideas about appropriate sexual relations, cues for sexual arousal, and messages about the consequences of sexual expression. As anthropologist Malinowski said long ago, '[s]ex in its widest meaning... is rather a sociological and cultural force than a mere bodily relation of two individuals' (Malinowski 1929, p. xxiii).

Peggy Reeves Sanday (1981) puts Malinowski's claim to the test in a study of the sociocultural context of rape cross-culturally. Using a standard sample of band and tribal societies, Sanday found that 47 percent of the societies for which there was adequate information showed a low frequency of rape, compared with 18 percent which were unambiguously 'rape prone.' Commenting on the pattern of correlations revealed by her study, Sanday (1981, p. 5) concludes that rape by males is part of 'a cultural configuration' which includes interpersonal violence among men as well as between men and women and male social dominance. Rape-prone societies are more likely to be characterized by male segregation in separate houses, an ideology of male toughness, emphasis on competition, low respect for women as citizens, and the absence of women in the public domain of economic and political affairs. In the more rape-free societies, there is an ethos of cooperation and consensus in human affairs. The social separation of the sexes is less marked and both sexes are more likely to hold exalted positions in public decision-making and to be integrated and equal in the affairs of everyday life.

Anthropologist Carole Vance (1984, pp. 7–8) also cites the diversity in human sexual practices to suggest that biology does not 'directly or simply determine the configuration or meaning of sexuality'. If this were the case, she argues, we would encounter uniformity cross-culturally rather than the startling diversity that actually exists in which 'activities condemned in one society are encouraged in another, and ideas about what is attractive or erotic, or sexually satisfying or even sexually possible vary a great deal.' Vance concludes that 'the body and its actions [must be] understood according to prevailing codes of meaning.' She suggests that an important question for research must be 'What is the nature of the relationship between the arbitrariness of social construction and the immediacy of... bodily sensations and functions?' (Vance 1984, pp. 7–9; see also Caplan 1987, pp. 1–27 on the cultural construction of sexuality).

In a study of the sociocultural and historical contexts of rape in the United States, Sanday (1996, p. 26) coins the term 'sexual culture' to describe the system of meanings, standards, and expected behaviors by which male and female sexuality is judged and understood. She argues that historically theories like those of Darwin, Krafft-Ebing, and Ellis functioned as cultural templates for popular sexual stereotypes by means of which males and females judged appropriate sexual behavior and acted accordingly. Because these stereotypes condoned sexual aggression by making it 'natural,' such as seen in the idea that 'boys will be boys,' US rape laws applied primarily to the use of force by a stranger. This changed after feminist activism in the 1970s resulted in more stringent legal codes. The same activism initiated studies on the incidence and social correlates of rape in America showing that acquaintance rape was much more common than stranger rape (Sanday 1996, pp. 161–207).

3. Rape Statistics in the United States

US statistics come from numerous studies utilizing national, community, or college samples. In general, these studies define rape as nonconsensual sexual intercourse irrespective of whether the parties know one another. Nonconsent is defined using the modern legal definition of rape. In most states, rape is defined as nonconsensual sex either due to the use of force or taking advantage of a person's inability to consent because of intimidation or because of alcohol or drugs. Although most studies focus on heterosexual rape, some authors note that rape also affects same-sex couples.

The first study designed to measure the percentage of those claiming they had been raped in a specified sample population was conducted in 1957 by sociologist Eugene Kanin on a college campus. This survey revealed that 55 percent of the 291 college

women interviewed said they had experienced offensive episodes 'at some level of erotic intimacy.' Twenty-one percent said they were offended 'by forceful attempts at intercourse' and 6.2 percent by 'aggressively forceful attempts at sex intercourse in the course of which menacing threats or coercive infliction of physical pain were employed.' Of the latter group 48 percent told no one and none reported to anyone in authority (Kirkpatrick and Kanin 1957, pp. 53, 56).

Another well-known path breaking study was conducted by Diane Russell in 1978 in San Francisco. Russell interviewed 930 women ranging in age from 18 to 80 of diverse social classes and racial/ethnic groups. Twenty-four percent of the women interviewed reported at least one completed rape, and 31 percent reported at least one attempted rape. The majority experienced rape by acquaintances and a much smaller percentage by strangers (Russell 1984, pp. 34–8).

Two large-scale national studies yielded similar results. In 1985 Mary Koss joined with *Ms.* magazine to survey a national sample of 6,159 students on 32 college campuses. This study found that many more women and men were having sex in the 1980s compared to the figures reported by Alfred Kinsey in his landmark studies of male and female sexual behavior published in the 1940s and 1950s. One-quarter of the men interviewed in the Koss–*Ms.* study reported involvement in some form of sexual aggression, ranging from unwanted touching to rape. A high percentage of the males did not name their use of force as rape. Eighty-eight percent said it was definitely not rape. Forty-seven percent said they would do the same thing again (summarized by Sanday 1996, pp. 191–2; see also Koss 1988, 1992, Warshaw 1988).

One in four of the women surveyed by Koss said yes to questions asking about behaviors that in most states would be legally defined as rape or attempted rape (for this definition see above). The power of popular stereotypes for defining acceptable sexual behavior and the adversarial nature of these stereotypes is reflected in the fact that most of the women reporting incidents that would legally qualify as rape did not call it rape. The term 'date rape' evolved from this study because many of the women surveyed said that the behavior often occurred on dates. Although most did not call it rape, Koss reported that 'the great majority of rape victims conceptualized their experience in highly negative terms and felt victimized whether or not they realized that legal standards for rape had been met' (Koss 1992, pp. 122–6).

By the end of the 1990s a number of scientifically designed research studies revealed similar statistics (for a summary of studies see Koss 1993 and Sanday 1996, pp. 184–206). The findings reported in these studies are corroborated by a national study of sex in America conducted in 1992 employing a representative probability sample of 3,432 Americans (see Michael et al. 1994). The findings fall within the range of many other studies: '22 percent of women were forced to do

something sexually at some time,' while 'just 2 percent of men were forced.' The authors of this study concluded that 'as many as one in five women' experience some form of sexual coercion, including rape, in the United States (Michael et al. 1994, pp. 33, 221–2).

4. *The Social Correlates of Rape in US Studies*

A number of studies conducted late in the twentieth century indicate that the adversarial view of male–female sexual interaction described by the early sexologists is frequently held by sexually coercive males. In a study of a large southeastern university, Boeringer (1996, pp. 137–9) reports that 56 percent of the males he interviewed admitted to obtaining sex by verbal harassment (i.e., 'threatening to end a relationship unless the victim consents to sex, falsely professing love, or telling the victim lies to render her more sexually receptive'). One-quarter of the males reported using drugs or alcohol to obtain sex and 9 percent reported at least one use of force or threatened force to obtain sex. Boeringer concludes that such verbally coercive tactics suggest an 'adversarial view of sexuality in which one should use deceit and guile to win favors from a woman' (p. 140).

Sexually aggressive men, from convicted rapists to college males answering questions on social surveys, share a remarkably similar set of attitudes. Most believe that sexual aggression is normal, that sexual relationships involve game playing, that men should dominate women, that women are responsible for rape, and that relations between the sexes are adversarial and manipulative on both sides (Koss and Leonard 1984, pp. 221, 223). Reanalyzing Koss' data to pinpoint attitudes held by the self-admitted sexually aggressive men of her study (i.e., those who admitted to forcing a woman to have sex in Koss's 1985 questionnaire), Sanday (1996, pp. 196–7) found that these men often expressed adversarial-like beliefs. For example, many stated that they believe that a woman's No means Yes and that women say no to intercourse because they don't want to seem loose but really hope the man will force her. Compared with men who do not admit to forcing a woman, more of these men also think that being roughed up by a man is sexually stimulating to women and that women have an unconscious wish to be raped. Few of the women in Koss' sample hold such attitudes.

To conclude, there is considerable evidence suggesting that whatever the biological basis of male sexual aggression might be, family background and cultural context make a difference. The incidence of rape is far lower among men raised in stable, peaceful environments be it a family context in a complex society like the United States or the intimate communities of small-scale traditional societies in which the ethos is one of mutual respect between the sexes.

To say that some men will rape in all societies and to use this fact to make generalizations about the sexual predilections of all men may bring about the very sexual culture one is trying to avoid by creating a lore that makes sexual aggression a 'natural' part of masculinity. More important for the education of young males is the conclusion based on studies in the US and cross-culturally that most men in most societies do not rape, and those who do face social rejection or incarceration.

5. Prevention

Examples of prevention strategies based on the biological approach are offered by Thornhill and Palmer (2000, pp. 169–88). Believing that rape occurs 'in all the environments in which humans societies [sic] have been known to exist' (p. 171), they are less interested in changing the environment by challenging popular stereotypes than with focusing on the 'exact nature of the psychological mechanisms that guide male sexual behavior' (p. 172). They offer a number of suggestions including certain types of educational programs. Citing the view of rape proposed by Camille Paglia, claim Thornhill and Palmer (2000, p. 183) that since men 'have evolved sexual preferences for young and healthy women and are attracted to women who signal potential availability by means of dress and behavior,' women should be informed of the risk factors and use the information to lower their risk of rape by paying attention, for example, to dress and appearance. They also suggest more direct social interventions such as separating vulnerable females from males.

The cultural approach argues for a change in rape laws and sexual stereotypes so that nonconsensual sex is treated unambiguously as a crime. Increasingly, this approach defines rape not just in terms of the use of force or drugs, but more broadly as failure to obtain verbal consent. The operative words of the new American sexual ideology are 'affirmative verbal consent,' 'freely given agreement,' and 'sexual communication.' Such modern adaptations would be acceptable to Darwin's view of natural selection because they promote sexual health and well being in an era of sexually transmitted diseases. One point left out of discussions of the biological bases of rape is the fact that as much as Darwin stressed competition, he also stressed cooperation or acting on behalf of what he called 'the general good or welfare of the community' (quoted by Sanday 1996, p. 288). Called 'altruism' by students of human evolution, some suggest that the primary selective pressure in human evolution was not for violent men but for those who were able to cooperate with women and other males in organized food gathering. Some anthropologists believe that this ability to cooperate was responsible for the evolution of human culture (Tanner 1981). In light of the ever increasing incidence of sexually transmitted diseases

together with the evolution of nuclear weapons, one can also note that today there is a fine line between aggressive behavior that results in reproductive success and behavior that leads human populations instead to the brink of extinction.

See also: Domestic Violence: Sociological Perspectives; Family and Gender; Family Bargaining; Female Genital Mutilation; Feminist Legal Theory; Feminist Theory; Feminist Theory: Psychoanalytic; Feminist Theory: Radical Lesbian; Gender and Feminist Studies; Gender and Reproductive Health; Gender and the Law; Gender Ideology: Cross-cultural Aspects; Heterosexism and Homophobia; Male Dominance; Masculinities and Femininities; Pornography; Prostitution; Rape and Sexual Coercion; Regulation: Sexual Behavior; Reproductive Rights in Affluent Nations; Reproductive Rights in Developing Nations; Sex Offenders, Clinical Psychology of; Sexual Attitudes and Behavior; Sexual Behavior: Sociological Perspective; Sexual Harassment: Legal Perspectives; Sexual Harassment: Social and Psychological Issues; Sexual Perversions (Paraphilias); Sexual Risk Behaviors; Sexuality and Gender; Teen Sexuality; Treatment of the Repetitive Criminal Sex Offender: United States; Women's Health

Bibliography

- Boeringer S B 1996 Influences of fraternity membership, athletics, and male living arrangements on sexual aggression. *Violence Against Women* 2(2): 134–47
- Caplan P (ed.) 1987 *The Cultural Construction of Sexuality*. Tavistock, London
- Darwin C 1936 *The Origin of Species by Means of Natural Selection or the Preservation of Favoured Races in the Struggle for Life and the Descent of Man and Selection in Relation to Sex*. The Modern Library, New York
- Geertz C 1973 *The Interpretation of Cultures: Selected Essays*. Basic Books, New York
- Kirkpatrick C, Kanin E J 1957 Male sex aggression on a university campus. *American Sociological Review* 22: 52–8
- Koss M P 1988 Hidden rape: Sexual aggression and victimization in a national sample of students in higher education. In: Burgess A W (ed.) *Rape and Sexual Assault II*. Garland Publishing, New York, pp. 3–25
- Koss M P 1992 Defending date rape. *Journal of Interpersonal Violence* 7(1): 122–6
- Koss M P 1993 Rape: Scope, impact, interventions, and public policy responses. *American Psychologist* 48(10): 1062–9
- Koss M P, Leonard K E 1984 Sexually aggressive men: Empirical findings and theoretical implications. In: Malamuth N M, Donnerstein E (eds.) *Pornography and Sexual Aggression*. Academic Press, New York, pp. 213–32
- Malinowski B 1929 *The Sexual Life of Savages in North-Western Melanesia: An Ethnographic Account of Courtship, Marriage, and Family Life among the Natives of the Trobriand Islands, British New Guinea*. G. Routledge & Sons, London
- Michael, R T, Gagnon J H, Laumann E O, Kolata G 1994 *Sex in America: A Definitive Survey*. Little, Brown and Company, Boston

- Russell D E H 1984 *Sexual Exploitation: Rape, Child Sexual Abuse, and Workplace Harassment*. Sage, Beverly Hills, CA
- Sanday P R 1981 The socio-cultural context of rape: A cross-cultural study. *Journal of Social Issues* 37(4): 5–27
- Sanday P R 1996 *A Woman Scorned: Acquaintance Rape on Trial*. University of California Press, Berkeley, CA
- Symons D 1979 *The Evolution of Human Sexuality*. Oxford University Press, New York
- Tanner N M 1981 *On Becoming Human*. Cambridge University Press, Cambridge, UK
- Thornhill R, Palmer C T 2000 *A Natural History of Rape: Biological Bases of Sexual Coercion*. MIT Press, Cambridge, MA
- Vance C S 1984 Pleasure and danger. In: Vance C (ed.) *Pleasure and Danger: Exploring Female Sexuality*. Routledge and Kegan Paul, Boston, pp. 1–28
- Warshaw R 1988 *I Never Called It Rape: The Ms Report on Recognizing, Fighting and Surviving Date and Acquaintance Rape*. Harper & Row, New York
- de Waal F B M 2000 Survival of the rapist. *NY Times Book Review* April 2: 24–25

P. R. Sanday

Rational Choice and Organization Theory

Rational Choice Theory (RCT) and Organizational Theory (OT) are two different, yet—in the works of some neoclassical economists, rational choice sociologists, and economic organizational theorists—closely related theories. Rational choice theory is used by economic organizational theorists as a method of aggregation. Specifically, it is used when it is necessary to link the change in actions of individuals with change in characteristics of organizations. However, organizational analysts generally take a broader view in which organizational context, organizational structure, and individual actions interact and move together to change organizational functioning. Such a perspective does not rest on the assumption that aggregated individual choices explain organizational behavior. In some cases, organizational theory explains how individual actions articulate with structure. In other cases, organizational theory does not deal with individual actions. Rather, it deals with articulating relationships between and among organizational and contextual characteristics.

The purpose of this analysis is to compare RCT with OT. Rational choice theory is based on the premise of self-interested utility maximizing actions/choices of individuals. Organizational theory is based on the premise of organizational rationality, the efficient functioning of organizations through means/ends relationships. Organizational economic theory, which consists of transactions cost economics and agency theory, forms an indirect link between the

RCT and OT. First, differences by defining the origin and premises of RCT and premises of Organizational Theory are elaborated. This is followed by demonstrating the link between the two, as developed in the premises of transaction cost analysis and agency theory. Although transaction theory holds to the premises of organizational efficiency, it finds the premise of self-interest utility maximization problematic. On the other hand, while agency theory holds the premise of self-interested, utility maximizing individual choice of RCT, and finds the guiding premise of organizational theory, the efficient functioning of firms, to be problematic. Thus, even the economic organizational theories, which purport to link RCT and OT are able to do so only if they accept either the RCT premises or the OT premises. As one set of premises are assumed, the other becomes problematic and becomes the questions to be examined in the analysis. No organizational theory has successfully integrated RCT and OT.

1. Rational Choice Theory: Origin and Premises

Rational Choice Theory originates in the British utilitarian tradition, via neo-classical economics. It is the core of classical economic theory. Although often qualified beyond recognition to meet conditional constraints of organizational analysts, it has been adopted in sociology (social choice) and in political science (public choice, game theory). In organizational theory and analysis, rational choice is most often used for the development of economic models of organizations. Transaction cost economics (TCE) and agency theory (AT) are the dominant economic organizational perspectives.

The premises at the basis of RCT as derived from neoclassical economic theory are: (a) self-maximizing individual utility driven actors, unconstrained by others and social norms (Coleman 1990, p. 503); (b) actors possess complete information and knowledge of their tastes, existing resources, the resources of others and market characteristics of the products and services they seek; (c) actors calculate and recalculate the changing conditions and act rationally; (d) however, preferences do not change, they are stable, and they are given often implied from the very behavior they are to predict; (e) actors are not emotional, irrational, impulsive, habitual in their choices, they are rational; (f) interaction is defined as an exchange between buyer and seller achieved when utility and cost converge in a market transaction of supply and demand equilibrium.

Rational choice theory (RCT) is an axiomatic system that provides a basis for making individual decisions/choices. The axioms of RCT define rational choice on the basis of choices between pairs of alternatives, which meet the axiomatic requirements of: consistency, transitivity, independence, continuity,

and monotonicity. Consistency demands that it be possible for all the decision maker's options to be ranked. The assumption of consistency is also called the assumption of connectedness. Preferences must be either equal or unequal. Unequal preferences can be rank ordered for comparison across the decision maker's entire set of preferences. Transitivity is the assumption that if A is preferred to B, and B is preferred to C, then consistency requires that A is preferred to C (Green and Shapiro 1994, pp. 14–15). The distance between preferences or magnitude of preferences does not have to be known to the subject or the analyst. Independence assumes that all preferences are independent of all other preferences. The continuity assumption is that preferences hold across time and space. Rational choice theorists assume monotonicity—all decision rules and preferences are stable across individuals and time. Allowing for assumptions of interpersonal variation produces enormous tractability problems (Storm 1990, p. 126). Another assumption of RCT is that individuals 'maximize the expected value of their payoff measured on some utility scale' and that, usually, 'numerical probabilities can be attached to different eventualities' (Luce and Raiffa 1957, p. 50). Eventuality is calculated as the value of the outcome 'weighted by the probability of achieving it' (Elster 1986, Harsanyi 1986).

2. Organizational Theory: Origin and Premises

Rational choice theory and organizational theory are brought together by the term rational. However, the term has different meanings (Zey-Ferrell 1981) for RCT and OT. RCT is based on the idea that an individual seeks to maximize his/her utilities as they conduct organization business. The central tenant of rational choice is that individuals at all levels of organizations act selfishly based on their own preferences to maximize their utility. This is rational individual behavior. OT defines rational as the way in which economic organizations are designed—the most efficient method of organizing. This is rational organizational behavior.

Rationality in OT originates in Weber's ([1922] 1968) concept of 'rational bourgeois capitalism,' which connotes the systematic organization of ideas, people, and resources in the interests of instrumental efficiency and effectiveness. Weber wrote that rationality and its accompanying bureaucratic organizational form, based on formalization of the rule of law and differentiation of functions, pervaded not only the modern economy, but public administration, the legal system, and religions. The rationalization of bureaucratic structures led to increased efficiency. In Weber's model, organizational rationality and rationalization are essentially the same. This is not a (micro-level) individual process as it is in RCT. It is an organizational process or specification of mean/ends

relationships based on systematic calculation and recalculation to achieve the most efficient organizational functioning. Subsequent analyses of organizations have assumed this rationality is the basis of organizations and developed theories around this premise.

3. Divergences of Rational Choice and Organizational Theories

In addition to definitional divergence of the term rational, the most apparent divergence is the difference in the unit of analysis. RCT attempts to explain individual choice behavior. RCT is based on the utility maximizing behavior of individual choice, while OT is based on the assumption of efficiency seeking organizational design. OT explains behavior of organizations. The utility maximizing behavior of individuals and the assumption of efficiency-seeking organizational design are similar, but are at different levels of analysis. If the unit is the individual, the utility being maximized is individual owner or employee self-interest. If the unit is the organization, the utility being maximized is profitability. Profitability of the firm does not always translate into income or other self-interests of firm members. Likewise, rational efficient organizational functioning does not always maximize the self-interest of organizational members.

Economic organizational theorist define organizations as firms doing business in markets. Neoclassical economists begin with market assumptions and work backward to firms and individuals assuming that actors operating in firms act in ways consistent with the theory of the market. However, assumptions that hold for RCT do not hold for organizational actors. Let us take just one assumption, that of assuming that all actors have full information about costs. Because each of them knows the cost and adjusts the quantity of their purchase, it is reasonable to assume that the market will settle on an equilibrium. In organizations these assumptions are false. Many types of decisions are not afforded to each actor. The choices available to each actor are not the same. Some actors' choices impose constraints on the choices available to others. Some actors do not know that a particular choice exists. For example, information about pending mergers and acquisitions. During mergers and acquisitions, parts of corporations maintain secrecy, not only from other firms, but from other units of their own firm and individual within their own units (Zey 1993, 1999). For example, securities firms are required by law to construct walls of secrecy, 'Chinese Walls,' between M&A units and traders to prevent insider trading. Information is also kept secret to prevent oppositional reactions to present and new allocations of organizational resources. Information is not equally distributed among or within organizational units for deliberate power-based reasons. Finally, the norms of

Table 1

Rational Choice Theory (RCT) and Organizational Theory (OT): Ideal type contrasts

	Rational choice theory	Organizational theory
Intellectual Origins	Neumann and Morgensterns Milton Freideman	Max Weber and Herbert Simon Oliver Williamson Michael Jensen
Unit of Analysis	Individual	Organization
Content of Analysis	Preferences	Characteristics of organization and relationships to context
Human Nature	Self-interested	Self- and other interested
Rationality	Individual, self-interested Utility maximizer	Means/ends efficiency of functioning
Utility Maximization	Self-interests	Profits in firms
Human Preferences	Fixed hierarchy	Variable over time and place
Basis for Interaction	Assumed trust	Trust is problematic
Constraints	Economic actions are constrained by tastes and scarcity of resources	Organization actions are constrained by scarcity but also by social structure, and historical context
Method	Predictive and explanation Not descriptive. Formal, mathematical model building no empirical data (clean hands)	Many methodologies comparative, historical based on data from empirical observations (dirty hands)

organizational rationality, efficiency dictates that various units within an organization should not receive the same information, but only what is necessary for their organizational entity. For all units to receive all information would be inefficient.

4. *The Link Between Rational Choice Theory and Organizational Theory*

Although rational choice theory and organizational theory are based on different premises and units of analysis, one type of theory, organizational economic theory, has attempted to link the two sets of premises and units of analysis. There are two dominant perspectives in organizational economic theory transaction cost economics and agency theory.

Economic organizational theory, both agency theory (AT) and transaction cost economics (TCE), begins with the relationships between individuals and organizations that are often conceptualized as events or transactions in which goods and services are exchanged (transaction cost analysis) and as contracts, which are agreements to conduct specific transactions (agency analysis). One is immediately aware that the unit of analysis is aggregated individuals. Transactions and contracts are aggregated either cross-sectionally

or over time. In contrast, OT deals with the properties and characteristics of organizations, not individuals.

Coase (1937) from whom Williamson elaborates his TCE, explains the origin of firms as alternatives to inefficient markets, while organizational theorists observe that people create organizations to generate income and power, to create larger empires with which they are identified, for themselves or groups to which they belong. Mergers to create new organizations often take place to take advantage of new technologies, to take advantage of tax changes, or to reposition the organization in a growing market. These changes may have little to do with rational economic organization or inefficient market functioning.

5. *Convergence and Divergence of Economic Theories of the Firm and Organizational Theory*

Transaction cost economics (TCE) and agency theory (AT), generally labeled economic organizational theories, have attempted to link RCT and OT and their different units of analysis and methods of theorizing through a set of assumptions that aggregate utility-maximizing individual behavior so that their behavior promotes organizational efficiency (profitability). However, survey research has demonstrated that not everyone has the same (or even individually

consistent) utilities. Thus, the aggregation, specifically non-additive aggregation functions, of individual choices and action into organizational actions is not clearcut, because the complexity of the aggregation process often violates RCT assumptions.

Furthermore, the two economic organizational theories, AT and TCE, do not operate under the same premises. AT is based on the utility-maximizing individual of RCT (Jensen 1983, 1993, Jensen and Meckling 1976), while transaction cost economics assumes that humans are boundedly rational at best. AT attempts to explain why organizations are more or less efficient while TCE assumes that organizations are an efficient alternative to inefficient markets (Williamson 1975, 1985, 1996). The utility-maximizing behavior of organizations is efficiency-seeking organizational design. If the unit of analysis is the organization, and the utility being maximized is profit, then efficiency maximization is implied when income seeking is not compromised. Organizational theorists question these assumptions because organizations, even firms, are never constraint free and individuals within organizations do not operate under the same constraints.

Agency theory assumes that individuals are utility maximizers but does not assume that organizations are structured in the way they are because it is the most efficient method of organizing. This is demonstrated, not only in the assumption that managers and workers act as opportunistic agents—that is, not in the interest of the owners—but also by the preoccupations of agency theorists with governance and methods of controlling nonowners to maximize profits (returns to shareholders) to the owners (Jensen and Ruback 1983). Agency theorists and economic organizational theorists in general do not integrate into their analyses historical changes, which provide opportunities and constraints on individual and organizational actions as organizational theorists do. Jensen's presidential address to the American Finance Association demonstrates that he had not until recently discovered Chandler's work (1962, 1977, 1990) and that organizational forms which are most efficient change over time with each historical context in which they develop.

Transaction cost economics assumes that organizations are structured in the way they are because it is the most efficient method of organizing, but TCE does not assume that individuals within organizations are utility maximizers. For example, in Williamson's work, *Markets and Hierarchies* (1975), he assumes that organizations more efficiently eliminate (by management fiat) opportunistic ('self-interested with guile') individual behavior than market mechanisms. TCE theorists analyze uncertainty and asset specificity of transactions. Individual utility maximization is problematic because of the obvious tendency of rational actors to act opportunistically under conditions of uncertain and asset specificity. TCE, like AT ignores historical changes and processes which constrain and provide opportunities for organizations.

Both contextual and organizational changes are ignored. Organizational change, especially transformational change cannot be explained by economic theories of the firm (see Zey and Swenson 1998, 1999). This is problematic because of the major transformations of firms at the turn into the twenty-first century.

In both Williamson's transactions cost economics (1975) and in Jensen's agency theory (Jensen and Meckling 1976), the idea of self-interest in contractual arrangements between and within organizations is problematic. Ultimately, questions of reconciliation of differential self-interest in the absence of trust, is impossible. Trust, rules of fairness, norms of reciprocity, and conflict, are non-existent and irrelevant in RCT and economic theories of the firm. They are essential to organizational formation and perpetuation.

Neither TCE or AT holds both the basic premise of RCT, self-interested, utility maximizing individuals, and the basic premise of OT, that firms are designed in the most efficient form of organizational design. Agency theory holds the former, while TCE holds the latter. Each finds the others basic premises problematic and grist for the mill of organizational analysis. Thus, economic organizational theory has not yet integrated rational choice theory and organizational theory. This author does not think it is possible to do so because of conflicting premises and different units of analysis.

See also: Bounded and Costly Rationality; Decision and Choice: Economic Psychology; Decision Making: Nonrational Theories; Firm Behavior; Game Theory; Hierarchies and Markets; Institutional Economic Thought; Market Structure and Performance; Rational Choice Theory: Cultural Concerns; Transaction Costs and Property Rights; Utilitarian Social Thought, History of; Utilitarianism: Contemporary Applications; Weber, Max (1864–1920)

Bibliography

- Chandler A D Jr. 1962 *Strategy and Structure: Chapters in the History of the Industrial Enterprise*. MIT Press, Cambridge, MA
- Chandler A D Jr. 1977 *The Visible Hand: The Managerial Revolution in American Business*. Belknap Press, Harvard University Press, Cambridge, MA
- Chandler A D Jr. 1990 *Scale and Scope: The Dynamics of Industrial Capitalism*. Belknap Press, Harvard University Press, Cambridge, MA
- Coase R 1937 The nature of the firm. *Economica* 4: 386–405
- Coleman J S 1990 *Foundations of Social Theory*. Belknap Press, Harvard University Press, Cambridge, MA
- Elster J (ed.) 1986 *Rational Choice*. New York University Press, New York
- Green D P, Shapiro I 1994 *Pathologies of Rational Choice Theory: A Critique of Applications in Political Science*. Yale University Press, New Haven, CT

- Harsanyi J C 1986 Advances in understanding rational behavior. In: Elster J (ed.) *Rational Choice*. New York University Press, New York
- Jensen M 1983 Organizational theory and methodology. *Accounting Review* 55(2): 319–39
- Jensen M 1993 Presidential address: The modern industrial revolution, exit, and the failure of internal control systems. *Journal of Finance* 48(3): 225–64
- Jensen M, Meckling W 1976 Theory of the firm: Managerial behavior, agency costs, and ownership structure. *Journal of Financial Economics* 3: 305–60
- Jensen M, Ruback R S 1983 The market for corporate control. *Journal of Financial Economics* 11: 5–50
- Luce D R, Raiffa H 1957 *Games and Decisions: Introduction and Critical Survey*. A Study of the Behavioral Models Project Bureau of Applied Social Research, Columbia University. Wiley, New York
- Storm G S 1990 *The Logic of Lawmaking: A Spatial Theory Approach*. Johns Hopkins University Press, Baltimore, MD
- Weber M [1922] 1968 *Economy and Society*. Roth G, Wittich C (eds.). Bedminster Press, New York
- Williamson O E 1975 *Markets and Hierarchies: Analysis and Antitrust Implications: A study in the Economics of Internal Organization*. The Free Press, New York
- Williamson O E 1985 *The Economic Institutions of Capitalism: Firms, Market, Relational Contracting*. Free Press, New York
- Williamson O E 1996 *The Mechanisms of Governance*. Oxford University Press, New York
- Zey-Ferrell M 1981 Criticisms of the dominant perspectives in organizations. *The Sociological Quarterly* 22: 181–205
- Zey M 1993 *Banking on Fraud: Drexel, Junk Bonds, and Buyouts*. Aldine de Gruyter, Hawthorne, NY
- Zey M 1998 *Rational Choice Theory and Organizational Theory: A Critique*. Sage, Thousand Oaks, CA
- Zey M 1999 The subsidiarization of the securities industry and the organization of securities fraud networks to return profits in the 1980s. *Work and Occupations* 26: 50–76
- Zey M, Swenson T 1998 Corporate tax laws, corporate restructuring, and subsidiarization of corporate form, 1981–95. *The Sociology Quarterly* 39: 555–82
- Zey M, Swenson T 1999 The transformation of the dominant corporate form from multidivisional to multisubsidiary: The role of the 1986 Tax Reform Act. *The Sociological Quarterly* 40: 241–67

M. Zey

Rational Choice Explanation: Philosophical Aspects

Rational choice explanation is the progeny of the Scottish Enlightenment and neoclassical economics. It is essentially reasoning from economic assumptions of what actors value to reach conclusions about both individual actions and aggregate or collective-level outcomes. It is applied to virtually any significant behavior or social outcome, from political participation, crime, and drug addiction to institutional structures, legislative and interest group politics, law,

norms, failures of collective action, and revolution. The most extensively developed parts of such theory are in the theory of elections and voting, the analysis of law under the rubric of law and economics, the study of group behavior, and game theory. Rational choice explanations gain their power from the attempt to reduce particular motivations to some narrower set of motivations, often simply to interests, which can be moderately well defined and measured across a large variety of contexts and actions.

The most common criticism of rational choice explanation is that it is false to the facts of actual motivations of people. This is a by-inspection claim. Because very much of actual behavior surely is motivated by interests, the dispute must actually be whether rational choice explanation is likely to help understand certain activities but not others. This is a question that cannot be answered in the abstract. Sometimes the criticism of rational choice trades on the fact that we commonly resolve many apparent choices by acting from habit or routine. As A. N. Whitehead quipped, civilization advances by the reduction of more and more choices to habit. A similar criticism of utilitarianism is that, if we spend as much time in calculating costs and benefits of our choices as would seem to be necessary to get our choices right, we will have no time left for life. J. S. Mill (1969) responded, rightly, that a sailor need not work out the positions of the stars anew each trip to sea but can simply rely on an astrolabe.

To illustrate the nature and power of rational choice explanation, as well as its plausible limits, I will address the aggregate-level phenomena of collective or group action, majority rule, and democracy. I will then address game theory and the value theory of rational choice. Finally, I will address the application of rational choice explanation to norms and social order.

1. Fallacies of Composition

A common criticism of rational choice explanation is that it cannot handle aggregate-level phenomena because it is fundamentally grounded in individual-level interests, choices, and actions. On the contrary, the chief successes of such theory have been to relate aggregate-level phenomena to individual-level actions. Several of the most important findings have been the—typically negative—demonstration that various aggregate-level claims are fallacies of composition from individual-level actions, choices, or values. It is a fallacy of composition to suppose that a collection of items has itself the characteristics of those items. This can be true, but it must be shown and not merely assumed. Consider three of these demonstrations.

The logic of collective action suggests that, although individuals may share a general interest in supplying themselves some collective benefit, they individually may have no interest in contributing effort or resources to the project. My contribution will buy a small part of

our good, of which an even smaller part will go to me, so that my cost of contribution will be larger than my benefit from it. All members of a group may therefore be individually worse off because they failed to take some action (Olson 1965, Hardin 1982). This is a problem that is pervasive in societies, ranging from very small societies with limited resources to very large and complex industrial societies.

The Arrow theorem (Arrow 1963, Barry and Hardin 1982) concludes that a collective utility or welfare function cannot be aggregated from individual utility functions if both are ordinal (see discussion of ordinal utility theory below). In essence, the theorem says that there cannot be a collective preference that has the simple features of an individual preference function. Among the implications of this theorem is that every standard choice procedure is potentially incoherent. Majority rule, for example, can produce cyclic preferences such that candidate A would defeat B by a majority, B would defeat C, and C would defeat A. Hence, simple majority rule is potentially indeterminate. It often works well enough—or seems to work—in practice, however, because there might not be cyclic preferences or because there are devices that block such preferences from being effective and even from being revealed. For example, plurality elections or the use of run-offs between the top two candidates in a multicandidate election can hide any cyclicities because there cannot be cycles over only two candidates.

The economic theory of democracy explains why candidates might not be substantially differentiated in certain common contexts and why voters often have little interest in voting. It also makes sense of the otherwise surprising level of ignorance of typical voters. In this latter case, however, there may be no fallacy of composition, because individual-level ignorance often translates into collective ignorance. Consider the problem of voting, as analyzed by Anthony Downs (1957). The chance that my vote would make a difference in an election decided by a large electorate is very nearly nil. In a 1974 statewide contest for senator in the very small state of New Hampshire, with somewhat more voters than Iceland, the vote was so close that errors in counting made it practicably undecidable who had won. No-one's single vote mattered. Suppose that the trouble to me of voting is such that I would sooner spend several dollars to vote by telephone to avoid going to the polls. The value to me of electing one candidate rather than another would have to be in the many thousands of dollars to offset my costs of voting, because the benefit would have to be discounted by the very low likelihood that my vote would make a difference. Yet many people do vote. Much behavior seems not to fit the rational choice theory and must be explained in other ways.

If it would not be worth my very slight effort to vote, Joseph Schumpeter (1950) concludes, then it would be

even less worth the much more substantial effort it would take for me to be well informed about the candidates and issues at stake. Evidence from many national contexts suggests, in fact, that citizens are abysmally ignorant when they vote. Here rational choice explanations seem to fit well. Editorial campaigns against voter ignorance are often misframed as being about the irrationality of citizens, who are acting all too rationally.

2. *Game Theory*

Rational choice explanations are commonly based in game theory, neoclassical economics, or commonsense arguments such as those of Smith and Hume. Game theory is perhaps the framework that best characterizes the nature of most of the problems in rational choice theory, although many of the most important insights in such theory predate game theory, which originated roughly during World War II (Von Neumann and Morgenstern 1953). Although it is called a theory, game theory is little more than a descriptive framework on which many theories about particular or quite general phenomena are founded. That framework takes many forms, the most transparent of which is the matrix form in which each of the players or participants faces a set of possible strategy choices. Outcomes of a play of a game are jointly determined by the strategy choices of all the players.

This characterization of our individual choices—of strategies, not of outcomes—makes game theory virtually antideontological. Judging action types as right or wrong makes no sense in interactive contexts, which are the principal contexts for most of moral theory. An action theory that takes what are essentially individual choices of outcomes as central is almost irrelevant to choice in social contexts in which, typically, important outcomes necessarily are not under the control of individual deciders (Hardin 1988, pp. 68–70). An ethic of rules makes nonsense of choices of strategies, which are chosen because of their effect on outcomes, as these are influenced by the strategy choices of others. Game theory can readily be fitted to consequentialism and especially utilitarianism. Utilitarianism essentially generalizes the principle of rational choice to choosing the welfare of all instead of only the welfare of the individual chooser.

To fit utilitarianism to game theory is to revamp its central concern with outcomes as valued by their benefits and costs. In interactive social choice, we typically cannot say that a particular action is right or wrong in producing a good overall outcome. We must, rather, judge strategy choices and we must judge them contingently. A strategy choice is right only to the extent that, in interaction with the choices of others, it produces a good overall outcome. If the others with whom I am interacting are sometimes utilitarian, sometimes self-seeking, and sometimes rule-guided in their choices, I may have difficulty determining my

best strategy choice for producing a utilitarian outcome. This is not different from the difficulty that self-interested players have in determining the best strategy choices for themselves. Game theory is indeterminate in many contexts, and therefore both pragmatic and utilitarian choices in a real-world setting are apt to be indeterminate as well. Merely seeing the nature of interactive choice in descriptive game theory changes our understanding of the possibilities of rational choice theory and of utilitarianism.

3. Value Theory: Interests, Consumptions, and Welfare

Rational choice accounts are commonly grounded in interests. Sometimes, interests are the whole story of someone's motivations in a particular context. But typically, my interest in having more resources, such as money, is itself grounded in my desire to consume various things. Resources, capabilities, and opportunities matter to us only because they enable us to consume or experience various things, and consuming or experiencing these things is what constitutes our welfare. Resources are worthless otherwise. Focusing only on interests is therefore a shorthand for what the resources can enable us to do. Moreover, we could relate our consumptions to the overall welfare that they bring us. It would be a mistake to suppose interests translate smoothly into welfare or consumptions. Indeed, consumptions typically trade off against each other and against interests. If I invest much of my resources in a holiday, I may not have enough left for other things, such as a new refrigerator. Rational choice explanations involve all three of these categories—interests, consumptions, and welfare.

It would also be a mistake to suppose that my welfare is wholly self-interested. One of the things I might wish to consume or experience might be your happiness. One may say this is against my interests because it means using resources that I could use for my own happiness in some other, perhaps more direct, way. But this is a confusion in at least two ways.

First, the only point in having resources is to be able to consume in order to contribute to one's welfare, as is evident to everyone but a Scrooge for the resource of money. True, there is a conflict between using my resources for one consumption and saving them for another, but this is properly seen not as a conflict between consumptions and interests but as a conflict between different consumptions, or between different paths to welfare.

Second, there is no reason to suppose that, say, enjoying an opera, a good meal, or a lovely journey is a more direct or more successful route to enhancing my welfare with some bit of my resources than providing for my son's or lover's, or friend's happiness with that same bit of resources. More generally, my welfare is enhanced best by a mixture of consumptions

and experiences, including contributions to the welfare of others, especially certain others. With those certain others, my welfare will depend on our sharing intentions to do various things for and with each other.

If we could suppose that economic interests as measured in resources were the primary issue for people, we could fairly easily explain actions as though people were all alike except for their circumstances. Rational choice explanations typically reduce the focus to interests, rather than other demands, such as demands for moral positions (ranging from deontological to utilitarian). In part they do so for methodological reasons. Analysis is more productive if we keep it clean and focused, and simplification is required for virtually all social analyses. But in part, the focus on interests is often basically right, especially for aggregate or collective issues. Other concerns matter to collective groups and societies, but they matter the way they do in life more generally. They often do not conflict with interests although, of course, occasionally they do and they may even trump interests in many contexts. For example, voters may vote against their personal interests in support of welfare programs and policies favoring ill-represented ethnic groups. But those who voted against the British Raj in India, French control of Algeria, or the US involvement in Vietnam may generally have voted with their interests as well as with moral concern. Still, many social and individual phenomena cannot readily be explained as the product of interest-seeking by individuals.

4. Value Theory: Cardinal, Ordinal, Ramsey–Neumann Cardinal

Apart from the structure or content of the value theory discussed above, there has been extensive debate over the form of such theory. Jeremy Bentham supposed that pleasures and pains during a day of someone's life could be defined in such a way that we could add the value of pleasures and subtract the value of pains for an overall measure of benefit to the person for that day. Today, we would speak of such a welfare or utility measure as cardinal in the sense that each experience one has could be assigned a utility number that could be added straightforwardly to other utilities.

Bentham's terminology was confused in that he spoke of the utility one gets from consuming some object and also of the utility inherent in the object. The latter notion underlies the labor theory of value that corrupted much of economic thinking throughout the nineteenth century and often still corrupts it today, in part under the doleful influence—on this matter—of Karl Marx. Bentham contributed to two opposing streams of value theory: that of the labor theory of value and that of the subjective theory that holds value to be strictly what a person gets out of an experience or consumption. The value to me of consuming some-

thing such as, say, beluga caviar might be far below the value to me of consuming other things that take equal expenditures of my resources. The value to you, even with the same overall level of resources, might be radically higher. Hence, you might consume a lot of beluga caviar and I might consume none after an initial taste. The subjective theory obviously has the disadvantage that we cannot always generalize from one person to others, although often we can generalize enough not to need to do in-depth studies of the tastes and preferences of everyone in order to explain overall behavior.

Bentham supposed further that the utility for one person could be added to that for another. This supposition would allow us to add up the utilities to all under one set of circumstances and to compare the total to that under another set of circumstances. Supposing we could do the measures, we could then say that, for example, one form of government is superior to another.

The so-called ordinal revolution in demand theory in economics overthrew all of the cardinal assumptions and set in their place a utility theory that is entirely subjective and that cannot be given interpersonal comparisons. The remarkable fact for economics is that such a limited theory clears up many prior problems and makes coherent sense of demand and supply relationships and, hence, prices in market exchange. Unfortunately, this revolution came only in the 1930s when moral theorists had ceased also to be economic theorists, as Smith, Hume, Mill, and Henry Sidgwick had been. It has therefore not had as much influence in general social theory as in economics, although it clearly should be brought to bear on the former as much as on the latter. A clear implication of interpersonally noncomparable value theory is that many traditional questions of interaction between people cannot be readily settled, so that moral and social theory must often be open on many issues.

In what might count as a minor sideline, Frank Ramsey (1931) and John von Neumann (Von Neumann and Morgenstern 1953) independently proposed a way to cardinalize preferences that are merely ordinal. Suppose you face a list of mutually exclusive outcomes, and suppose you can rank order any two outcomes. Assign a value of 0 to the least preferred, Z, and a value of 100 to the most preferred, A, outcomes in your list. For some outcome, *K*, between these two in your ranking, you can rank it against any lottery over A and Z. For example, suppose you would be indifferent to getting *K* and getting the combination of a *p* percent chance of getting A and a 100–*p* percent chance of getting Z ($0 < p < 100$). On the scale of 0 to 100, *K* has a cardinal value to you of *p*.

The Ramsey–Von Neumann cardinalization of utility (RN) clearly applies to personal decisions over outcomes. But we cannot use it to construct a Benthamite cardinal utility that is additive across persons, because there is no sense in which your 0 and

my 0 are the same or your scale and mine, because each of them is arbitrary. If I assign a value of *p* and you assign a value of *q* to outcome *K*, we cannot meaningfully say that *K* has a joint value of *p* + *q*. Indeed, we can perform a linear transformation on your utility scale without affecting the results that follow from it. For example, we could divide the value of each outcome by 10 and add 37 to it, so that your scale would then run from 37 to 47, and this transformed scale would get all the results that your original scale from 0 to 100 would get. In addition, we can assign utility values to outcomes that fall outside the original ranking from A to Z, so that some outcomes might have negative utility and others might have extremely large positive values. This is suggested by the fact that the original normalization was performed over the arbitrary set of outcomes from A to Z. Hence, there is no meaning to any claim that your RN utility can be added to mine to yield an overall joint utility. The RN theory is therefore primarily of interest in individual decision making under risk, and is seldom of interest for more general social theory. It is irrelevant to Benthamite hopes for a cardinally additive utilitarianism.

5. Norms

There is a growing literature on the rational choice explication of norms. Edna Ullmann-Margalit (1977) gave rational reconstructions of norms for dealing with strategic contexts that are modeled by prisoner's dilemma, coordination, and partiality or unequal coordination games. In this categorization the descriptive apparatus of game theory is used to establish and characterize types of norms according to the strategic problems they resolve. The first of these norms essentially handle collective action problems because, without the force of the norms, many individuals would free ride, and any effort to provide a collective good would fail. The second handle such coordinations as the driving convention. That convention surely requires no norm to keep drivers in compliance with it, because the threat of grievous harm from violating the convention is sufficient. But many conventions are less regulated by such clear incentives and some of these might be more forceful if backed by norms. Norms in Ullmann-Margalit's third category handle unequal coordinations such as the division of labor by gender roles or social order in any hierarchical society in which obedience to norms might keep even those at the bottom of the hierarchy loyal and orderly. Again, these conventions are grounded in coordination incentives and they might be fully self-enforcing, but norms that moralize the inequalities might make those in the less well off positions less sullen in acquiescing in the arrangements.

More generally we may say that there are two broad categories of norms: norms of exclusion and universa-

listic norms (Hardin 1995, Chaps. 4 and 5). Norms of exclusion are specific to particular groups, which can enforce the norms by sanctioning violators within the group; the sanctions can go so far as to shun and even exclude members. The sanctioning is grounded in the interests of individual sanctioners, so that the norms are commonly interest based, although group members might claim they are inherently moral or right. A norm of exclusion can work to help enforce cooperativeness in a group or community, or to enforce a coordination that might serve the interests of some but not all. For example, a minority culture might enforce upon its members the use of a minority language or a particular dress code. Using that language or following that dress code might reduce the possibilities for some members of the community to gain employment outside the community. The community might protect its hegemony over deviant members by shunning those who violate the language or dress code.

Universalistic norms can be dyadic, as in promising or truth telling, or large-number in the sense of applying to actions toward the whole society, as in not littering. Dyadic norms are commonly enforced through the sanctions inherent in the iterated prisoner's dilemma. If you lie to me or break your promise to me, I can simply avoid further interaction with you, to your likely cost. If you break a norm that affects everyone, such as a norm against littering or in favor of voting, no particular individual is likely to have an interest in sanctioning you. Hence, such norms are likely to be relatively weak in their force and in our adherence to them. Oddly, therefore, it is the often odious norms of exclusion, such as racist and ethnic norms, that have strongest force, while universalistic norms are relatively weak unless they are played out essentially in dyadic interactions. Those who insist that, morally, norms should commonly trump interests, should be bothered that it is interests that trump any plausible moral content of many norms, and that norms not backed by interests are likely to be weak. Large-number universalistic norms may be backed primarily by moral or psychologically inculcated commitments. If so, then it is of interest whether they are, in fact, very effective in motivating people.

6. Social Order

There are three grand schools of social theory. One, associated with the French sociologist Émile Durkheim and many others, holds that society is held together by shared values. The second, associated with Karl Marx and Plato's *Thrasymachus*, holds that conflict is pervasive and that social order must therefore be maintained through the imposition of power to coerce when necessary. The third, associated with Adam Smith, holds that society is organized by and for exchange, which involves a combination of coordination and conflict. Each of these schools has

many strong contemporary advocates. A fourth plausible view is that much of social life is simply organized by coordination (Hardin 1999, Chap. 1). We need not share particular values in order to share interests in avoiding conflicts with each other. The ideal type of this last form of social order is coordination on traffic rules, especially on the rule to drive right (or to drive left). We can be hostile to each other in many ways and still find it sensible to avoid needless conflict over many things, so that we would willingly be coordinated. A shared value theory is also a coordination theory, but its advocates ground it in value commitments and are generally hostile to the inclusion of interests among those value commitments. The conflict and exchange schools are commonly grounded in interests, although the socially destructive conflict that sparked much of early liberal thought was over commitment to divergent religious views.

The pre-eminent coordination theorists among classical political philosophers were Thomas Hobbes and David Hume. They were both protogame theorists in their grasp of the fundamental nature of various social interactions. While Hobbes is commonly thought to be a major conflict theorist, his actual resolution of the problem of social order is to arrange for coordination on an all-powerful sovereign, whose task is essentially to coordinate us on order so that we may construct our own separate lives and engage with one another in productive pursuits (Hobbes 1968). Hume's (1978, Book 3) articulation of a remarkable array of coordination problems qualifies him to be seen as the greatest coordination theorist, although his complex insights were largely neglected for two centuries, in part because he lacked the technical vocabulary of game theory and did not himself bother to work out the general principles of it. He did, however, work out a general moral and political theory that was substantially grounded in strategic understanding (Hardin 1988, Chap. 2).

In complex modern societies, coordination may be the central, dominant mode of social order. It does not require consensus on values—pluralists can coordinate on an order for the society in which they seek their diverse values. It does perhaps fit especially well with modern individualist liberalism that allows individuals to live by their own lights. Indeed, coordination is a simple version of Mill's principle that we should be free to pursue our own ends without interference so long as our doing so does not interfere with others (Mill 1977). When there is such interference, of course, other principles may be required to regulate our interactions.

7. Concluding Remarks

Rational choice theory is a diverse set of approaches to the study of society that are based in assumptions of individual rationality. Indeterminacies in such theory often mirror indeterminacies in social relations and

individual understandings of these. Indeed, many rational choice explanations have demonstrated the indeterminacy of social choice. Among the most important contributions of such theorizing is the descriptive clarity it often adds to our understanding of issues that have been poorly characterized before. Rational choice explanation is inherently about states of affairs, although it can include in a state of affairs concerns for anything that motivates people. It is therefore easy to see how it fits with utilitarianism.

Rational choice value theories, when introduced into the study of law, have led to the burgeoning movement of law and economics, which has produced the most systematic analysis of the broad sweep of law and legal rules that legal scholarship and practice have hitherto seen. The quick march of that movement—roughly beginning with the Coase theorem (Coase 1988)—through virtually all areas of law has provoked counterattacks, but none of these is itself grounded in a value theory systematic enough to permit more than piecemeal objections. The Coase theorem says, with audacious generality and simplicity, that if there are no transaction costs to impede economic cooperation between owners of various resources, then ownership will not affect production, although it presumably will affect how the profits from the production are divided. That theorem, along with earlier work of Ronald Coase and many others, has led to the massive study of transaction costs.

We have a rich understanding of the problems of rational choice value theories just because those theories have been a major focus in the development and articulation of the entire discipline of economics. The theories are a product of centuries of determined and often brilliant debate with dozens of major contributions. Alternative value theories have had far too few advocates and critics to yield much understanding at all. Indeed, value claims commonly seem to be *ad hoc* and not systematically generalizable. Such considerations are no argument for the superiority or rightness of any value theory. But they may be a tonic for those who wonder how the only well-articulated class of value theories is so easy to criticize: It has enough content to be subject to extensive and varied criticism.

See also: Decision Theory: Bayesian; Economics and Ethics; Economics, Philosophy of; Game Theory: Noncooperative Games; Intentionality and Rationality: A Continental-European Perspective; Intentionality and Rationality: An Analytic Perspective; Irrationality: Philosophical Aspects; Practical Reasoning: Philosophical Aspects; Prisoner's Dilemma, One-shot and Iterated

Bibliography

Arrow K J 1963 *Social Choice and Individual Values*, 2nd edn. Wiley, New York

- Barry B, Hardin R (eds.) 1982 *Rational Man and Irrational Society?* Sage, Beverly Hills, CA
- Coase R H 1988 The problem of social cost. In: Coase R H *The Firm, the Market and the Law*. University of Chicago Press, Chicago, pp. 95–156
- Downs A 1957 *An Economic Theory of Democracy*. Harper, New York
- Hardin R 1982 *Collective Action*. Johns Hopkins University Press, Baltimore, MD
- Hardin R 1988 *Morality within the Limits of Reason*. University of Chicago Press, Chicago
- Hardin R 1995 *One for All: The Logic of Group Conflict*. Princeton University Press, Princeton, NJ
- Hardin R 1999 *Liberalism, Constitutionalism, and Democracy*. Oxford University Press, Oxford, UK
- Hobbes T 1968 *Leviathan*. Penguin, London
- Hume D 1978 *A Treatise of Human Nature*, 2nd edn. Oxford University Press, Oxford, UK
- Mill J S 1977 On Liberty. In: Robson J M (ed.) *Collected Works of John Stuart Mill*. University of Toronto Press, Toronto, Vol. 18, pp. 213–310
- Mill J S 1969 Utilitarianism. In: Robson J M (ed.) *Collected Works of John Stuart Mill*. University of Toronto Press, Toronto, Vol. 10, pp. 203–59
- Olson M Jr 1965 *The Logic of Collective Action*. Harvard University Press, Cambridge, MA
- Ramsey F P 1931 Truth and probability. In: Braithwaite R B (ed.) *The Foundations of Mathematics and Other Logical Essays by F. P. Ramsey*. Harcourt Brace, New York
- Schumpeter J A 1950 *Capitalism, Socialism and Democracy*, 3rd edn. Harper, New York
- Ullmann-Margalit E 1977 *The Emergence of Norms*. Clarendon, Oxford, UK
- Von Neumann J, Morgenstern O 1953 *Theory of Games and Economic Behavior*, 3rd edn. Princeton University Press, Princeton, NJ

R. Hardin

Rational Choice in Politics

‘Rational choice’ (henceforth RC) is one of the principal approaches to political science as practiced in the USA. The approach has generated important insights relevant to a wide range of political phenomena, but it has also generated controversy as it moved into each of the various subfields. Much of the controversy arises from misunderstandings, as well as a failure to appreciate that RC is not a monolithic enterprise.

1. The Rational Choice Approach

The central premise uniting the various schools of RC is that behavior is purposive. Political behavior is not the product of psychological drives, childhood socialization, organizational norms, or other a-rational influences stressed by the various schools of behaviorists. Such influences certainly shape indi-

vidual preferences, but rational choice further posits that individuals pursue their goals as rationally as their knowledge, their resources, and the context permit. For example, much congressional scholarship of the 1950s and 1960s treated members of Congress as acting primarily in response to role expectations and norms established and enforced by other members. Rational choice scholars in the 1970s suggested that a more fundamental starting point would be the presumption that, first and foremost, members of Congress wanted to be re-elected—and acted accordingly. That premise would not describe all members all the time of course, but it would fit more members more of the time than any other assumption, and thus was the most appropriate starting point.

1.1 *Origins of RC*

Although RC shares with economics the maintained hypothesis of purposive behavior, RC is not a direct import from economics, as implied by critics who level charges of ‘economic imperialism.’ Some schools of RC owe much to economics, but others have independent roots. Many of the modeling tools first utilized in RC research were borrowed from operations research and other branches of applied mathematics. In particular, political scientists saw potential in game theory earlier than did economists, who continued to use mechanistic maximizing models for some years after RC practitioners in political science had gravitated toward strategic models. The non-cooperative game theoretical revolution brought the two groups closer together, but present convergence should not mask the eclectic origins of the approach. Consider that the membership of the Committee on Nonmarket Decision Making (the forerunner of the Public Choice Society) included William Riker, William Mitchell, and Vincent Ostrom (political science), Kenneth Arrow, James Buchanan, and John Harsanyi, (economics), Herbert Simon (then public administration), Gordon Tullock (law), John Rawls (philosophy), and James Coleman (sociology).

1.2 *Schools of RC*

There are at least four recognizable schools of RC, although as the approach has spread into the mainstream of political science the distinctions have eroded. The most influential school within political science is the Rochester school, founded by Riker. A product of one of the country’s most traditional political science programs (Harvard); Riker had both dissatisfaction with the state of political science in the middle of the twentieth century and a clear vision of the direction in which it should go. His program at the University of Rochester began producing graduates in the late 1960s and within a decade they had established thriving groups at Carnegie-Mellon, Washington University,

and Caltech. Within another decade Rochester graduates and their students had accepted positions at the country’s top political science departments.

In contrast, the Virginia School is based on economics. Led by Buchanan and Tullock this school is most closely identified with the Public Choice Society in particular, and the term ‘public choice’ in general. From the beginning the Virginians emphasized empirical work and especially policy relevant work. Virginians tend to be less abstract in their methods and more focused on outcomes than their Rochester counterparts. And more than Rochester the Virginia School is associated with distinct normative point-of-view, advocating markets as a better means of allocating resources than governments.

The Chicago School is associated most closely with George Stigler and Gary Becker. Reflecting their roots in economics, Chicagoans generally assume that actors’ goals are egoistic and material. This view of humans as selfish wealth maximizers strikes many political scientists as too narrow and has colored non-rational choice political scientists’ views of the larger RC enterprise. In addition, the Chicago School discounts the importance of institutions, believing that rational actors easily find ways to get around institutional constraints. Virginians and Rochesterians, in contrast, take institutions seriously. Virginians focus heavily on constitutional choice and Rochesterians are at the forefront of the ‘new institutionalism.’ Thus Americanists and comparativists working in rich institutional contexts have found Chicago’s emphasis on freewheeling interest groups less useful than international relations scholars who study an anarchic arena populated by weak institutions.

Finally, the Indiana School is smaller and more specialized. Founded by Vincent and Elinor Ostrom, the Indiana School takes as its special focus collective action problems in their various manifestations. Like Rochesterians, Indianans are more of a homegrown school of thought that attaches considerable importance to institutions. They are also more comparative in their orientation, tracking the operation of general principles across space and time.

2. *Evolution of the RC Approach in Political Science*

The RC approach first became widespread in the subfield of American politics. Probably this reflects the interests of several of its founders in democratic politics, as well as the fact that the American subfield was the most methodologically advanced and thus the most receptive to the sophisticated modeling tools that RC practitioners brought to the table. The main source of criticism within the subfield was theoretically and conceptually based. In the 1960s the subfield was dominated by behaviorists, who had tried to construct a scientific political science by borrowing concepts and

theories from sociology and psychology. From their studies of presidential elections, for example, behaviorists had concluded that the most important determinant of vote choice was party identification, then viewed as largely apolitical—a product of childhood socialization. The RC emphasis on rational, goal-oriented behavior seemed highly inconsistent with much of the behavioral literature.

The disruptions of the late 1960s undercut that judgment. As demonstrations raged outside, as an insurgent third candidate—George Wallace—contended for the Presidency, and as long-time Democratic identifiers defected from their party, it became increasingly untenable to assert that people voted on the basis of immutable, apolitical party identifications. RC scholars joined others in refocusing attention on the importance of issues and government performance and re-examined party identification, finding a political component within it.

By the 1980s RC had moved into the subfield of international relations. This was a natural development for the portion of the subfield focused on international economic relations, but natural as well for the international security side of the subfield, for whom coalitions, strategy, rational deterrence, and similar concepts had long been part of their stock in trade. Indeed, much of the early work on game theory and other operations research methods was done at the Rand Corporation with funding from the Department of Defense.

Most recently, RC has been moving into the subfield of comparative politics where its reception has perhaps been more controversial than in the other subfields. Some possible reasons for this contentious debate are offered.

2.1 Criticisms of RC

A common criticism of RC is that it is a hegemonic research program, not content to coexist with other programs but intent on replacing them. There is some truth to this charge, but it does not apply uniquely to RC—traditionalists leveled the same charge at behaviorists. Emphasizing the volitional element of politics does not require that the non-rational and emotional be ignored. But applying an approach to every possible research question helps scholars to determine its limits.

For example, it seems clear that RC is most directly applicable to the behavior of the elite, whose individual decisions make a difference, than to the behavior of the masses, whose individual decisions are submerged in those of millions of others. Attempts to behave rationally are most likely where the stakes are relatively high and the number of players relatively low—making a careful decision is not worth the effort if the consequences are trivial or your decision makes no difference to the outcome, or both. Thus, RC fails

to provide a compelling explanation of why millions of ordinary citizens vote in a presidential elections, but that does not imply that the two presidential candidates do not make every effort to behave rationally!

Others quarrel with the RC approach because it is assertively social scientific. Formalization and generalization are social science ideals, and RC is closely associated with both, although nothing requires that all RC research be expressed in terms of mathematical symbols. Understandably, those who prize detailed narratives of particular cases feel uncomfortable with RC, as do those who believe that a country's unique history or culture determines the operation of its institutions and political processes. In both cases the critics' quarrel is as much with social science as with RC.

Of course, even some card-carrying social scientists have qualms about the RC approach. Not everyone views human behavior as purposive. Moreover, RC is more self-consciously theoretical than other approaches. Practitioners engage in model-building, making simplifying assumptions and drawing logical consequences using the classic deductive method. Scholars whose research commitments are inductive and empirical sometimes feel that RC models omit much of the rich detail they have painstakingly compiled.

Finally, some part of the criticism of RC is undoubtedly ideological. Scholars have political commitments, and it is natural, although regrettable, that these commitments sometimes influence their work. As noted, rational choice is associated with a free-market viewpoint. There is nothing inherent in RC that logically requires such a point-of-view. Indeed, there are rational choice scholars who are Democrats, Liberal Democrats even! But the perception exists, and scholars who position themselves more on the social democratic side of the spectrum may criticize the approach because of its perceived ideological associations.

The acrimonious nature of the debates that have accompanied the movement of RC into comparative politics probably reflects the fact that many comparative politics scholars are on the same side of several, if not all, of the preceding divisions. Some comparativists believe in thick description, not parsimonious formal models, and many of these same scholars believe that unique cultural and historical conditions are the key elements of political systems, not timeless underlying principles common to all systems. Some comparativists emphasize the non-rational nature of cultural and historical practices, not rational calculations about means to ends. And finally, some comparativists, especially those who specialize in the developing areas, are personally farther to the left on the political spectrum than Americanists steeped in the liberal tradition and international relations scholars comfortable with discussions of realism and the national interest.

2.2 The Evolution of RC

In the years to come the controversies surrounding the RC enterprise are likely to diminish. The principal reason is the convergence generated by attempts to explain real world behavior. As noted, RC forced Americanists to reconsider the attempt to explain politics primarily in sociological or psychological terms. But just as clearly, the attempt to explain mass behavior such as political participation indicates that recourse to sociology and psychology will be necessary. Similarly, RC researchers recognize that there are phenomena underlying labels such as 'cultures' and 'norms,' and provocative current research seeks to explain such phenomena as 'equilibria' or 'focal points' of human interaction that evolve over time through processes of trial and error, learning, selection, or what-not, and persist if their consequences continue to benefit the community. Trends suggest movement toward a richer RC approach, but also toward political sociology and political psychology with a firmer grounding in first principles of human behavior.

See also: Rational Choice and Organization Theory; Rational Choice Explanation: Philosophical Aspects; Rational Choice Theory: Cultural Concerns; Rational Choice Theory in Sociology

Bibliography

- Amadae S M, Bueno de Mesquita B 1999 The Rochester school: The origins of positive political theory. *Annual Review of Political Science* 2: 269–95
- Fiorina M P 1995 Rational choice, empirical contributions, and the scientific enterprise. *Critical Review* 9: 85–94
- Hardin R 1987 Rational choice theories. In: Ball T (eds.) *Idioms of Inquiry*. State University of New York Press, Albany, NY, pp. 67–91.
- Mitchell W C 1988 Virginia, Rochester, and Bloomington: Twenty-five years of public choice and political science. *Public Choice* 56: 101–19
- Mueller D C 1997 *Perspectives on Public Choice: A Handbook*. Cambridge University Press, Cambridge, UK
- Stigler G (ed.) 1988 *Chicago Studies in Political Economy*. University of Chicago Press, Chicago

M. P. Fiorina

Rational Choice Theory: Cultural Concerns

The notions of rationality, rational choice, and rational beliefs have a central place in many of the social sciences. Although originally deployed mainly in economics, they are increasingly being incorporated into sociology and political science. (They have not

gained the same importance in social anthropology.) Some scholars argue that by virtue of its simplicity, universality and explanatory power the theory of rational choice is, or should be, the unifying framework for all the social sciences. As an explanatory theory it harbors formal and empirical problems that suggest a need for alternative approaches. At the same time, whatever its explanatory flaws, the notion of rationality retains a normative privilege, because we want to be rational (Føllesdal 1982). Rational choice theory is primarily prescriptive: it tells agents what to do to achieve their ends as well as possible. Social scientists may then try to explain their behavior by assuming that they do what the theory tells them to do.

1. The Development of Rational Choice Theory

The modern notion of rationality did not emerge until this century. Earlier references to rationality were to *ratio*, or 'reason,' understood as the faculty of calm and deliberative choice, as opposed to emotion. In the words of David Hume (1960, p. 415), 'Reason is, and ought only to be, the slave of the passions.' Some writers opposed reason both to interest and emotion: 'Nothing is easier for passion than to overcome reason; its greatest triumph is to conquer interest' (De La Bruyère 1990, IV.77). In this tradition, reason implies not only a dispassionate but also a disinterested and impartial attitude.

1.1 The Economic Meaning of Rationality

The modern concept of rationality as, essentially, means–end efficiency owes much to the marginalist revolution in economics in the last third of the nineteenth century. By conceptualizing the consumer's choice and the producer's choice as involving trade-offs among different consumption goods or different factors of production, the marginalists made it possible to think more systematically about the optimal choice of means for a given end. A profit-maximizing employer, for instance, would react to a wage increase by using more nonlabor inputs and less labor.

1.2 Max Weber

In *Economy and Society*, Weber (1968, p. 24) offered an influential definition of instrumentally rational (*zweckrational*) action: it is 'determined by expectations as to the behavior of objects in the environment and of other human beings; these expectations are used as "conditions" or "means" for the attainment of the actor's own rationally pursued and calculated ends.' Although Weber's professed aim was to offer a purely subjective notion of rationality as coherence among action, ends, and beliefs, some of his analyses

point to a more objective definition in terms of successful adaptation to the environment (Elster 2000). The closing pages of the *Protestant Ethic* (Weber 1958) suggest that he was influenced by the fact that market competition generates objectively successful behavior by eliminating those who fail.

1.3 The Emergence of Game Theory

Weber (1968) writes that 'in attempting to explain the campaign of 1866, it is indispensable in the case of [the two main generals] to attempt to construct imaginatively how each, given fully adequate knowledge both of his own situation and of that of his opponent, would have acted.' Weber does not seem to be aware of the need to 'construct imaginatively' how each general would try to 'construct imaginatively' what the other would be doing, including their further constructive imaginings, and so on. The modern theory of interdependent decisions, or (in somewhat misleading language) 'game theory,' emerged only after Weber's death. Neumann and Morgenstern (1944) and Nash (1950) showed how one can avoid the infinite regress lurking in the phrase 'and so on' by the notion of 'equilibrium', defined as a set of individual choices each of which is optimal, given the others.

1.4 The Emergence of Decision Theory

In much of rational choice theory, the intuitive notions of desire and belief are replaced by those of preference (or utility) and subjective probability. In a pioneering work, Ramsey (1931) showed how one can elicit utilities and subjective probabilities by asking subjects to respond to various hypothetical bets. Later, Savage (1954) offered an alternative demonstration. The key conclusion of these and later works in the same tradition is that rational actors maximize expected utility. The agent is assumed to face a well-defined set of feasible courses of action, each of which can with definite probabilities result in various outcomes, each of which has a definite utility for the agent. A rational agent will then choose the action that maximizes the average (or expected) utility of the outcomes associated with it.

2. The Structure of Rational Choice

In defining rationality, one can emphasize rational choice or rational action. In the latter approach, one can define rational action as action that results from rational choice or more objectively as success or adaptation to the environment. Even if social agents do not make rational choices, they may be constrained by their environment to act rationally. As noted above, this outcome may be produced by competition. In a population of firms, we may observe profit-maximizing behavior even if none of the decision-makers have

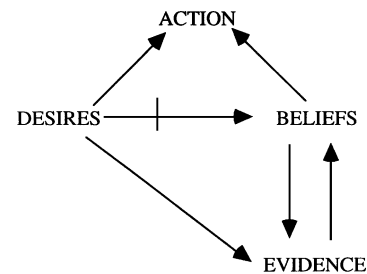


Figure 1
Rational action

profit-maximizing intentions, simply because firms that do not maximize profit are eliminated (Alchian 1950). As noted by Nelson and Winter (1982), the argument assumes that the environment is sufficiently stable for the process of elimination to have time to do its work. With a rapidly changing environment, the selection process is tracking a moving target and may not bring about optimal behavior.

2.1 A Basic Model

A widely used model of rational choice (and rational action) is captured in Fig. 1. In this diagram, the arrows represent three optimality relations. To be rational, an action has to be the best means to realize the agent's desire, given his or her beliefs. The beliefs must be well-grounded in the evidence available to the agent. The information must reflect an optimal investment in the acquisition of new information. (Because investing in information-gathering is an action, we may think of it as a shadow action that accompanies the primary action.) A rational choice is the choice of a rational action because it is the rational thing to do, excluding cases in which the agent makes the rational choice by fluke.

2.2 The Model is Subjective Through and Through

A choice and the subsequent action are rational only by virtue of the optimality relations among elements of the agent's mental set. A rational action may fail, if the agent has false (but well-grounded) beliefs. The relation between beliefs and evidence must not be confused with the relation between beliefs and the world. In inferring beliefs from the evidence a rational agent would use the procedures which in the long run yield the largest number of true beliefs, weighed by their subjective importance to the agent. Yet on any given occasion this procedure may lead to the formation of a false belief.

In the model, the desire is the 'unmoved mover.' The agent's desires allow us to assess the optimality of actions, beliefs, and information-gathering, but they are not themselves subject to rationality or optimality assessments. In some cases, this may seem counter-

intuitive. In the case of the agent's time preferences (a subset of the desires), we might want to say that an agent who discounts the future very heavily is irrational. The life of somebody who pays very little attention to the future consequences of present behavior is likely to be nasty, brutish, and short. Yet for the purposes of explaining the agent's behavior, this fact is irrelevant. Similarly, it may be tempting to say that a belief which (to us) is obviously false is irrational. Yet if it is part of the belief that the costs of testing it are very high, the agent may be caught in a belief trap (Mackie 1996) in which it is rational to hold on to the false belief. In applying rational choice theory, one has to adopt the point of view of the agent and recognize that the eye cannot see beyond its horizon.

2.3 *Belief-Desire Interactions*

As shown in Fig. 1 the model explicitly excludes a direct influence of desires on beliefs, as in wishful thinking. The converse possibility is that desires might be shaped by the beliefs, as when the belief that something is inaccessible stifles the desire for it ('sour grapes') or enhances it ('the grass is greener'). The general principle that desires cannot be assessed as more or less rational precludes us from qualifying desires shaped in these ways as irrational, however undesirable they might be on other grounds.

Rationality is, however, compatible with an indirect influence of desires on beliefs. As indicated in Fig. 1, the optimal amount of investment in evidence-gathering is shaped not only by beliefs (about the expected costs and benefits of gathering information), but also by desires. As beliefs are rationally shaped by evidence, it follows that beliefs may be rationally sensitive to desires. Consider two doctors who have to diagnose two identical patients with severe injuries. One is a GP who wants to save the patient's life and who knows that spending too much time on the correct diagnosis might undermine that goal. The other is a Nazi doctor in a concentration camp, who is unconcerned about the patient's survival and thus of the opportunity costs of information-gathering. Because of the different desires, the latter is more likely to form a correct belief.

2.4 *Some Common Misunderstandings*

Rational choice theory is often criticized, sometimes with good arguments, and sometimes with bad. Although some of the bad arguments may apply to bad versions of the theory, critics ought to address the best versions. The most common misunderstanding is that the theory assumes agents to have selfish motivations. Rationality is consistent with selfishness, but also with altruism, as when someone is trying to select the charity where a donation can do the most good.

Rational choice theory is also sometimes confused with the principle of methodological individualism.

True, the theory presupposes that principle. Talk about beliefs and desires of supra-individual entities, such as a class or a nation, is in general meaningless. The converse does not hold, however. Some of the alternatives to rational choice theory also presuppose methodological individualism.

Furthermore, it is sometimes asserted that the theory is atomistic and that it ignores social interactions. Almost the exact opposite is true. Game theory is superbly equipped to handle three important interdependencies: (a) the welfare of each depends on the decisions of all; (b) the welfare of each depends on the welfare of all; and (c) the decision of each depends on the decisions of all. It is also sometimes asserted that the theory assumes human beings to be like powerful computers, which can instantly work out the most complex ramifications of all conceivable options. In principle, rational choice theory can incorporate cognitive constraints on a par with physical or financial constraints. In practice, the critics often have a valid point.

Finally, it is sometimes asserted that the theory is culturally biased, reflecting (and perhaps describing) modern, Western societies or their subcultures. However, the minimal model set out above is transhistorically and transculturally valid. This statement does not imply that people always and everywhere act rationally, nor that they have the same desires and beliefs. It means that the normative ideal of rationality embodied in the model is one that is explicitly or implicitly shared by all human beings. People are instrumentally rational because they adopt the principle of least effort. 'Do not cross the river to fetch water,' says a Norwegian proverb. Also, as people know that acting on false beliefs undermines the pursuit of their ends, they want to use cognitive procedures that reduce the risk of getting it wrong when getting it right matters. That they often fail to adopt the right procedures does not undermine the normative ideal.

3. *Some Problems*

As an explanatory approach, rational choice theory faces two objections. The first is formal: under certain circumstances the notions of rationality or optimality are not uniquely defined or not defined at all. This is the problem of indeterminacy. The other is empirical: sometimes the social agents do not behave as the theory predicts they will. This is the problem of irrationality.

3.1 *Indeterminacy*

The problems of non-unicity or non-existence of an optimal solution arises at each of the three levels in Fig. 1

At the level of action, indeterminacy can arise if the agent's preference ordering is incomplete, i.e., for a

pair of options **a** and **b**, the agent neither prefers **a** to **b**, nor **b** to **a**, nor is indifferent between them. This case can arise if the options are very different from each other and from what the agent has experienced in the past, as in choosing a career or a spouse.

At the level of beliefs, indeterminacy—meaning that the agent has no reliable grounds for assigning probabilities to the possible outcomes—can take the form either of raw uncertainty or of strategic uncertainty. The former is illustrated by natural disasters or by the process of technical innovation. The latter is illustrated by games in which there are several equilibria none of which can be unambiguously singled out as ‘superior’ or ‘natural’ (in the sense of Schelling 1960). An example involving both forms of uncertainty involves investment by firms in research and development (Dasgupta and Stiglitz 1980).

At the level of information-gathering, indeterminacy arises through an infinite regress. To decide how much information to gather, one must first form beliefs about the expected costs and benefits of information. The formation of these beliefs also requires the acquisition of information, and so on (Winter 1964). Although the regress can be avoided in repeated and stereotyped situations such as many medical diagnoses, it is unavoidable when, as in warfare, there is both an urgent need to act and no precedent to rely on.

3.2 Irrationality

In this case, too, irrationality arises at all three levels. The problems are of two kinds. ‘Cold’ problems are caused by cognitive malfunctioning without any motivational bias. ‘Hot’ problems arise through such bias.

At the level of action, cold problems arise in many ways. Choices are sensitive to how questions are framed—whether the glass is described as half empty or as half full (Tversky and Kahneman 1981). Many people are prone to the ‘sunk cost fallacy,’ i.e., to pursue a project simply because they have already invested much in it (Arkes and Blumer 1985). Hot problems can arise when people act under the influence of emotion, e.g., by wanting to take action immediately instead of taking time to maximize the chance of success. Under the influence of addictive cravings one may fail to carry out a decision to quit. (See Elster 1999 for both issues.)

At the level of belief formation, cold problems can arise because people fail to understand the nature of statistical inference. During World War II Londoners were persuaded that the Germans systematically concentrated their bombing in certain parts of their city, because the bombs fell in clusters. This invalid inference reflected a lack of understanding of the statistical principle that random processes usually generate clustering (Feller 1968, p. 161). Flight instructors who noticed that praise for an unusually

smooth landing is typically followed by a poorer landing on the next try while criticism after a rough landing is followed by an improved performance, concluded that rewards are less effective than punishments. This invalid inference reflected a lack of understanding of the statistical principle of regression to the mean (Tversky and Kahneman 1974). Hot problems arise when ‘the wish is father to the thought,’ as in self-deception and wishful thinking. Coffee drinkers are less inclined to believe evidence of negative effects of caffeine (Kunda 1987). Those who work in risky occupations tend to believe the risks are less than they actually are (Akerlof and Dickens 1982).

Less is known about cold errors in information gathering. One might conjecture that people may invest too much in information because they underestimate opportunity costs compared to out-of-pocket expenses (Thaler 1980). A common hot mistake also involves overinvestment, when people have a strong desire to have sufficient reasons for their decisions even when cost–benefit analysis suggests that flipping a coin might be superior (Elster 1989a). Motivational biases can also, however, lead to underinvestment: ‘Marry in haste, repent at leisure.’

4. Alternatives

Problems such as those discussed above have stimulated the search for alternative theories. Some theories address the issue of indeterminacy as well as ‘cold’ deviations from rationality, others the ‘hot’ deviations.

4.1 Bounded Rationality and Behavioral Economics

Some alternatives try to identify the cognitive mechanisms that produce deviations from what rational choice theory predicts.

The research program on ‘bounded rationality’ stems from Herbert Simon (1954). Nelson and Winter (1982) is an outstanding application. The core assumption is that agents ‘satisfice,’ by choosing an option that is ‘good enough,’ rather than maximize. Agents do not look for better alternatives if the status quo is satisfactory. If it falls below the agent’s aspiration level, a search for new options is triggered until a new, satisfactory outcome is found. The theory has a good match to common sense (‘Never change a winning team’) and captures the heavy dependence of many choices on search procedures. By eschewing the maximizing assumption, it avoids problems of indeterminacy.

The research program on ‘behavioral economics’ or ‘quasi-rational economics’ (Thaler 1991) stems from the work of Kahneman and Tversky (Kahneman et al. 1982, Kahneman and Tversky 2000). In addition to identifying many of the ‘cold’ anomalies discussed above, these authors tried to account for them by

means of 'prospect theory.' It deviates from rational choice theory by assuming that agents evaluate gains and losses by how much they deviate from a reference point rather than in absolute terms, and that they are more risk-averse for losses than for gains. These two assumptions explain, for instance, that people who refuse to use credit cards when they are said to carry a surcharge accept them when the charge is redescribed as a cash discount.

4.2 Emotions and Social Norms

Other theories try to identify the motivational mechanisms that generate anomalies.

Frijda (1986, pp. 118–21) argues that emotions cause deviations from rationality in many ways: (a) they affect 'probability and credibility estimates' concerning events outside one's control; (b) they 'cause some measure of belief in the efficacy of actions one would not believe in under other conditions'; (c) they induce fantasy behavior, as when a widow for many years after the death of her husband sets the table for two each day; (d) they induce various forms of 'painful fantasies that, however painful, are yet sought by the subject himself.'—Othello's jealousy is cited as an example; and (e) they cause irrational behavior, such as 'anger at some deed that cannot be undone by the angry aggression, nor its recurrence prevented.' In addition, as mentioned earlier, they induce suboptimal investment in information. Loewenstein (1996) generalizes these ideas, by appealing to a broad range of 'visceral factors' that interfere with rational choice and rational cognition. In addition to emotion, these include hunger, thirst, pain, and the euphoria or dysphoria created by addictive drugs or abstinence from them.

Rational choice theory is consequentialist, in that it enjoins people to act in certain ways because of the expected outcomes. Social norms typically prescribe acts for their own sake, not for their consequences. In their simplest form, they take the form of unconditional imperatives: 'always wear black at funerals.' They may be contrasted with conditional, outcome-oriented imperatives: 'always wear black in strong sunshine' (as do people in Mediterranean countries to maintain circulation of air between the clothes and the body). Social norms regulate the use of money, how hard people work at their jobs, the occasions for and modalities of revenge, sexual behavior, and many other matters (Elster 1989b). They are sustained largely by the emotions of contempt and shame.

5. Conclusion

Many rational choice theorists will respond either that some of the alleged problems are empirically insignificant or that the alternative explanations can be

restated within the rational choice framework. These are ongoing debates with many unresolved questions. Overall, however, it is increasingly hard to deny that rational choice theory has serious limitations. Empirical research increasingly shows that human beings are severely limited in their cognitive functioning, and, when under the influence of emotion and other strong feelings, liable to act in ways that undermine their ends. Paraphrasing Veyne (1976, p. 676), beliefs born of passion serve passion badly. At the same time, there is no single rival or contender that could replace rational choice theory. For one thing, most people *are* rational much of the time. For another, the various alternatives to the theory have not been integrated with one another, and are unlikely to be so. There is no other theory that explains as much of human behavior, although the jury is still out on how much it explains.

See also: Bounded and Costly Rationality; Bounded Rationality; Intentionality and Rationality: A Continental-European Perspective; Rational Choice and Organization Theory; Rational Choice Explanation: Philosophical Aspects; Rational Choice Theory in Sociology; Rational Theory of Cognition in Psychology; Rationality and Feminist Thought; Rationality in Society

Bibliography

- Akerlof G A, Dickens W T 1982 The economic consequences of cognitive dissonance. *American Economic Review* **72**: 307–19
- Alchian H 1950 Uncertainty, evolution, and economic theory. *Journal of Political Economy* **58**: 211–21
- Arkes H R, Blumer C 1985 The psychology of sunk costs. *Organization Behavior and Human Decision Processes* **35**: 124–40
- Dasgupta P, Stiglitz J 1980 Uncertainty, industrial structure, and the speed of R&D. *Bell Journal of Economics* **11**: 1–28
- De la Bruyère J 1990 *Les Caractères*. Garnier-Flammarion, Paris
- Elster J 1989a *Solomonic Judgements*. Cambridge University Press, Cambridge, UK
- Elster J 1989b *The Cement of Society*. Cambridge University Press, Cambridge, UK
- Elster J 1999 *Strong Feelings: Emotion, Addiction and Human Behavior*. MIT Press, Cambridge, MA
- Elster J 2000 Rationality, economy and society. In: Turner S (ed.) *The Cambridge Companion to Weber*. Cambridge University Press, New York
- Feller W 1968 *An Introduction to Probability Theory and its Applications*, 3rd. edn. Wiley, New York, Vol. 1
- Frijda N H 1986 *The Emotions*. Cambridge University Press, Cambridge, UK
- Føllesdal D 1982 The status of rationality assumptions in interpretation and in the explanation of action. *Dialectica* **36**: 301–16
- Hume D 1960 *A Treatise of Human Nature*. Oxford University Press, Oxford, UK

- Kahneman D, Slovic P, Tversky A (eds.) 1982 *Judgment Under Uncertainty*. Cambridge University Press, Cambridge, UK
- Kahneman D, Tversky A (eds.) 2000 *Choices, Values, and Frames*. Cambridge University Press, Cambridge, UK
- Kunda Z 1987 Motivated inference: Self-serving generation and evaluation of causal theories. *Journal of Personality and Social Psychology* **53**: 636–47
- Loewenstein G 1996 Out of control: Visceral influences on behavior. *Organizational Behavior and Human Decision Processes* **65**: 272–92
- Mackie G 1996 Ending footbinding and infibulation: A convention account. *American Sociological Review* **61**: 999–1017
- Nash J 1950 Equilibrium points in *N*-person games. *Proceedings of the National Academy of Sciences* **36**: 48–9
- Nelson R R, Winter G 1982 *An Evolutionary Theory of Economic Change*. Belknap Press of Harvard University Press, Cambridge, MA
- Neumann J, Morgenstern O 1944 *Theory of Games and Economic Behavior*. Princeton University Press, Princeton, NJ
- Ramsey F P 1931 Truth and probability. In: Ramsey F P (ed.) *The Foundations of Mathematics*. Routledge and Kegan Paul, London
- Savage L J 1954 *The Foundations of Statistics*. Wiley, New York
- Schelling T C 1960 *The Strategy of Conflict*. Harvard University Press, Cambridge, MA
- Simon H 1954 A behavioral theory of rational choice. *Quarterly Journal of Economics* **69**: 99–118
- Thaler R H 1980 Towards a positive theory of consumer choice. *Journal of Economic Behavior and Organization* **1**: 39–60
- Thaler R 1991 *Quasi-rational Economics*. Russell Sage, New York
- Tversky A, Kahneman D 1974 Judgment under uncertainty: Heuristics and biases. *Science* **185**: 1124–31
- Tversky A, Kahneman D 1981 The framing of decisions and the psychology of choice. *Science* **211**: 453–8
- Veyne P 1976 *Le pain et le cirque*. Seuil, Paris
- Weber M 1958 *The Protestant Ethic and the Spirit of Capitalism*. Scribner, New York
- Weber M 1968 *Economy and Society*. Bedminster, New York
- Winter S 1964 Economic 'natural selection' and the theory of the firm. *Yale Economic Essays* **4**: 225–72

J. Elster

Rational Choice Theory in Sociology

Outside the confines of economics and closely related disciplines such as the decision sciences, where it is adhered to almost universally, rational choice theory (RCT) has not gained wide acceptance. Due to its relative success in economics and the limited achievements of other theoretical approaches adopted elsewhere, particularly in sociology and political science, rational choice theory gained significant support in the late twentieth century.

Use of RCT in sociology ranges from formulating mathematically deductive models assuming fully informed and flawlessly calculating actors, to a much broader approach, whence actors are seen as 'doing

their best' (i.e., subjectively optimizing), given certain preferences/desires/affects/utilities, with only a limited understanding of their environment and partial reasoning or calculative ability. It is this flexibility which recommends RCT to many of its proponents, ranging from those for whom it comprises the hallmark of any acceptable social science to those for whom it is pragmatically the best framework currently available.

RCT may be used in two complementary ways. First, in a positive manner by proposing explanations of individual and collective actions (decisions) and consequently responding to the question—how (in context) *do* actors behave? Second, in a normative manner, responding to the question of how they (again in context) *ought*, given their preferences, to behave. While strong in economics, the normative approach is not widely embraced in sociology.

Rigorous applications of RCT are almost invariably based upon an assumption of the maximization of (often subjective (expected) utility). Actors are thus conceived as (a) capable, at any instant, of taking one of a series of feasible actions, (b) each of which leads to an estimable set of outcomes, with (c) a known (subjective/objective) probability. Further, each outcome carries affect for the actor. The actor can then form a preference ordering over the actions in terms of affects/utilities. Actions (decisions) under certainty, risk and uncertainty are thus special cases of the general model. Various axiomatic systems are available, though the one attributed to Savage is closest to sociological concerns. It is now recognized that actors do not always abide by the precepts of expected utility theory. Prospect theory and developments thereof, which introduce a wide range of 'biases,' are widely accepted as falling under the broad ambit of RCT, as are models based upon limited rationality (like satisficing).

Three related types of RCT modeling frameworks can be distinguished. First, independent decision-making (actions), where the actor's environment may be treated as fixed (i.e., parametric); second, where the environment involves other actors and where consequently decisions or actions are interactive (i.e., strategic), where one shot and (finitely and infinitely) repeated game theory and their associated Nash equilibria are appropriate. Third, evolutionary game theory, where strategies are selected by imitation of either the relatively frequent or relatively successful strategies of others. While the first two frameworks, analytically speaking, usually commence with the full assumption of well informed and rationally calculating individuals, the latter requires no such initial assumptions. It does, though, usually assume that evolutionary process is driven towards an evolutionary stable equilibrium (or some refinement thereof).

The assumptions of expected utility maximization (more generally RCT) in complex environments (parametric or strategic) prove demanding on both the

reasoning power and knowledge (information) of the actors concerned, and are sometimes deemed by critics to be entirely unrealistic in a *positive* context. RCT is, however, strongly wedded to the idea of equilibrium, and in off-equilibrium adjustments all manner of trial and error may be allowed for. Equilibrium, in this view, is not necessarily arrived at (learned) because actors are rational, but rather because actors appear rational (i.e., discover rationality) once an equilibrium has been achieved. Evolutionary game-theoretic models derive the same conclusion by myopic processes of selection (emulation) eventuating in a evolutionary stable equilibrium.

All manner of RCT models may be said to rely initially upon three assumptions, each of which may be regarded as controversial in some sociological circles. They are:

Individualism. Only individual persons ultimately take actions.

Optimality. Individual actions are chosen optimally (at least at equilibrium).

Self-regard (or egoism). Individuals' actions are ultimately concerned with their own welfare (and nobody else's).

Though any one or more of these assumptions may be adjusted in a perspicacious manner (see below), rational choice theory may nevertheless be said to retain its paradigmatic privilege in the sense that it (with its standard assumptions) provides an analytical point of departure.

Despite its explicit individualism, system level or collective/macro relationships between variables are a perfectly acceptable currency for RCT as long as they are construed ultimately as 'shorthand' for the conjunction of relationships that ultimately run through individuals. Indeed, a theory about the connection between two system-level variables must ultimately imply a mechanism involving individual actions. The theoretical question is precisely this: How has the system level relationship been generated by the actions of individuals? The assumption of individualism in RCT implies no more than this about the explanation of system or collective level empirical generalizations.

No contradiction exists between RCT and a view which pictures individuals' actions as 'shaped' by their membership in a group (more generally a collectivity). If, as some 'structuralists' opine, individuals are often so circumscribed by internalized standards and external constraints upon their actions that they barely have any choice at all, then this may call into question optimality and self-regard, but it does not undermine the assumption of individualism in the sense that it is still the motor power behind individual actions (albeit 'programmed ones') that drive the social world.

Things happen in the social world because individuals do and do not (i.e., forbear) to do things, and they are the only things that do and do not do things. All statements that attribute 'doing' to other entities can, in principle if not in practice, be translated,

without loss, into statements about individuals doing things.

Individuals choose optimally if no other action exists whose consequences are preferred to the chosen action. An actor's preference-driven action will in turn depend upon that person's (a) beliefs or reasoning about the consequences of the available actions, and (b) affect (desire for/against) the consequences (and perhaps for the action itself). It is thus possible by derivation to speak of optimal beliefs and affects. It is this broad assumption of optimality that gives RCT its paradigmatic privilege.

The major alternative assumptions to this viewpoint are:

(a) That individuals frequently do not act optimally in terms of their preferences.

(b) That although individuals may act optimally in terms of their preferences, senses exist in which their preferences are themselves not formed optimally (non-optimal beliefs and affects).

(c) That individuals do not act upon choices but are driven by forces entirely beyond their control (structuralism).

Cases are well documented in which 'weakness of will' may be said to prevent people from acting in the way they would 'really prefer.' Preferences may also be imprecise or unclear, actions entirely experimental or based upon wishful thinking, and more than one optimal choice may be available. All of this is acknowledged by RCT, although in practice most of the problems reduce to examples of nonoptimal preference formation.

Some take the view (particularly many economists) that preferences are not only stable (compared to constraints) but also are formed in ways that entirely resist the attentions of RCT. Their genesis is either uninteresting (changed constraints accounting for changed action) or beyond rationality. Sociologists reject the first view almost unanimously, but find some sympathy for the second. Not infrequently it is also claimed that the idea of optimality is so weak, when taken by itself, as to constitute an affront to serious understanding. It is observed that, if we are to be permitted any sort of utility arguments, then actions may (certainly *ex-post facto*) always be made to look as if they are optimal.

The accusation is that, under such circumstances, the assumption of optimality can be protected entirely against potential refutation. In this context, sociologists think in terms of a series of derivative optimality questions:

(a) Are preferences formed optimally with respect to beliefs and affects?

(b) Are beliefs formed optimally with respect to the information available?

(c) Is available information (collected/processed) optimal with respect to affects?

(d) Are affects optimal with respect to individual autonomy?

An action is objectively optimal when conditions (a) to (d) hold true. Such actions may take place under certainty, risk, or uncertainty, according to standard assumptions about the nature of beliefs. These assumptions are collectively very strong, and many social theorists maintain that it is only rarely, if ever, that actions will conform to the ideal pattern.

The issue at stake is whether we may assume that affective dispositions in some sense flow from a balanced or autonomous personality. No doubt forms of behaviour/actions exist that flow from affective dispositions that are 'irrational' or pathological. When this occurs, it usually implies that individuals resist all occasions to consider the opportunity costs of maintaining the disposition (e.g., constantly striving for what is beyond their grasp or manifestly damages them).

It is points (a), (b) and (c) that have attracted the most attention. It is often proposed that preferences cannot be formed optimally for a variety of reasons, prominent among which are the following:

Satisficing. Beliefs are not performed optimally with respect to information available.

Search. No adequate theory of optimal search exists; therefore, once again, beliefs are not optimal with respect to information, nor information with respect to affects.

Decision biases. Preferences are not optimal with respect to beliefs and affects.

Exclusions. Not all possible actions are considered.

Limited deliberation. Not all consequences of all actions are thought through.

Received styles of reasoning. These prevent optimal belief formation.

Ritual/routine/habit/copying. These systematically lead to exclusions and do not evolve optimally in an evolutionary sense.

In each case a conceptual understanding of the apparent limitations of RCT is established in terms of the departure from full optimality. This is one significant sense in which RCT can make a claim for paradigmatic privilege: It still serves, even in default, as a benchmark. Furthermore, many of the above factors are modelled in terms of limited information environments.

The standard version of RCT usually starts with the assumption that individuals act (optimally) in order ultimately to satisfy their self-regarding preferences (utilities). That is, optimal actions are taken in a way that is indifferent to the welfare/utility/consumption of others, considered either individually or collectively. Although nothing intrinsic to the theory mandates this assumption, it is frequently regarded as a natural analytical starting point, to be discarded only cautiously if and when the facts speak otherwise.

Furthermore, if other-regarding preferences are invoked eventually, not infrequently it is maintained that they must in some way be explained or accounted for, whereas self-regarding sentiments do not invite

explanation, being in some way natural or self-explanatory. Thus many theorists adopt the practice of assuming that, if no evidence exists to the contrary, then postulate self-regarding utilities. If, however, independent evidence exists for other-regarding sentiments, it would be a mistake to ignore it and go on to ask for an explanation of the source. However, many RCT models interpret other regarding sentiments as instrumental to self-regard (enlightened self-interest). Such models are usually excluded before other regard is introduced into the utility function (affective orientation).

Evolutionary biologists (and evolutionary game theorists) continue to debate the issue as to whether altruistic motives can evolve at an evolving stable equilibrium. Sociologists are required increasingly to relate their own theories to those of colleagues working in these adjacent disciplines. This is particularly so when they adopt the view whereby it is the *strategic interaction* between two or more actors which constitutes much social interaction. These developments are bringing sociological theory increasingly into contact with game theory and evolutionary game theory. Although games of incomplete information currently play a limited role in sociology, we can expect this to change in the near future. Perfect Bayesian equilibria will thus come centre stage.

The paradigm of sociological investigation which is emerging from an RCT perspective takes the following form:

(a) Initially, assume the actors who are responsible for generating a collective outcome (which may not be intended but only a consequence of what is intentionally brought about) to be rationally self-interested, fully informed and endowed with calculative ability given their (exogenous) objective resources, and (exogenous) preferences/utilities.

(b) Model the structure of the actors interdependencies (social interaction) in the sense that the outcome depends jointly upon what each does. Thus each actor has to take account of what others do (will do, are likely to do, and, if appropriate, have done in the past). Such modeling will usually define a feasible set of actions strategies for each actor.

(c) In the light of (a) and (b), determine the course of action each actor will (rationally) pursue. If these are not unique, then specify their ranges of actions (i.e., multiple equilibria). If the interactions are repeated, adopt the appropriate repeated game theoretical framework.

(d) Thus predict the outcome(s) and test these against descriptions of the outcomes one wishes to explain. Of course, if the model possesses no unique outcome, then all one can require is that the event in question is among the set of possible outcomes predicted by the model (an envelope rather than point prediction). Specific cultural/historical occurrences may select one equilibrium (i.e., focal points) and suggests path-dependence on a particular outcome.

(e) If this simple model fails, construct further models which might involve one or more of the following possibilities:

(i) Relax assumptions of full information, probably following recognized procedure in game theory.

(ii) Relax the assumption of self-interest, introducing richer utility/preference functions (e.g., altruism, malice, indifference, envy, etc.).

(iii) Relax the assumption of objectively calculated resources introducing subjective beliefs about resources.

(iv) Relax the assumption of a set of feasible actions, calculated objectively, introducing subjective beliefs about what is feasible.

(v) Re-compute the structure of interdependencies.

When either (i), (ii), (iii), or (iv) are seen to hold, this will inevitably prompt a further question as to why it should be so. Answers to these questions might also be couched in terms of some of the RCT precepts outlined above (e.g., strategic limitations of information). It is only if this very general framework fails to provide the answers we seek that we should then reach for an alternative theoretical framework.

See also: Action, Theories of Social; Altruism and Self-interest; Bounded Rationality; Coleman, James Samuel (1926–95); Game Theory; Interests, Sociological Analysis of; Macrosociology–Microsociology; Methodological Individualism in Sociology; Rational Choice Explanation: Philosophical Aspects; Rational Choice in Politics; Rational Choice Theory: Cultural Concerns; Sociology: Overview; Theory: Sociological; Traditions in Sociology

Bibliography

- Abell P (ed.) 1991 *Rational Choice Theory*. Elgar, Aldershot, UK
 Abell P 1996 Sociological theory and rational choice theory. In: Turner B S (ed.) *Social Theory*. Blackwell, Oxford, UK
 Boudon R 1989 Subjective rationality and the exploration of social behaviour. *Rationality and Society* 12: 173–96
 Coleman J S, Fararo T J 1992 *Rational Choice Theory: Advocacy and Critique*. Sage, Newbury Park, CA
 Moser P K (ed.) 1990 *Rationality in Action: Contemporary Approaches*. Cambridge University Press, Cambridge, UK

P. Abell

Rational Theory of Cognition in Psychology

1. People Appear Irrational in Cognitive Experiments

Michael Watkins has said that a cognitive theory 'is a bit like someone else's toothbrush—it is fine for that

individual's use, but for the rest of us ... well, we would just rather not, thank you' (Watkins 1984, p. 86). To say whether people are behaving rationally requires a definition of what it means to behave rationally, and like a toothbrush everyone has their own. For the purposes of this article, rational behavior will be defined as follows: to behave rationally in some context is to display behavior that corresponds to a normative standard of behavior for that context. The choice of the normative standard determines what constitutes rational behavior. For the economist, behaving rationally involves maximizing utility; for the logician it is following the deductive rules of logic; and for the (Bayesian) statistician it is acting according to Bayes's rule. One need not look far to find that people do not reason rationally with respect to a variety of normative standards. There are entire literatures on how and why people violate the rules of deductive logic and statistical inference. In the Wason (1968) card task, central to one such literature, people are given a rule of the form 'if p, then q,' such as 'if there is a vowel printed on one side of a card, then an even number is printed on the other.' The participants are next presented with a set of cards, such as A, K, 2, 7. Their task is to choose only those cards that need to be flipped to check whether the rule holds. In this example, only the A-card (p) and the 7-card (~q) need to be checked. An odd number on the opposite side of the A-card would clearly violate the rule as would a vowel on the other side of the 7-card (~q). The rule says nothing about what is on the opposite side of a consonant, so flipping the K-card (~p) does not test the rule. Also, flipping the 2-card (q) cannot disconfirm the rule, because the rule does not restrict what is on the opposite side of an even-numbered card. In general people perform terribly on this task, or at least their performance does not appear to be rational with respect to the rules of deduction. Oaksford and Chater (1998) surveyed 13 studies, covering a variety of Wason tasks. The proportions of people who flip the p(A), ~p(K), q(2), ~q(7) cards were 89, 16, 62, and 25 percent respectively. That is, people rarely flip the ~q card, which they should flip, and frequently flip the q card, which they need not do.

People are similarly deficient in drawing inferences from statistical information. Consider the following example that requires people to interpret the results of a positive colon cancer test (Gigerenzer 1998). Suppose that 0.3 percent of people have colon cancer, there is a 50 percent chance that a colon cancer test will detect a cancer (a hit), and a 3 percent chance that it will indicate that there is cancer when there is none (a false positive). What is the probability that a person with a positive test has cancer? When Gigerenzer posed this problem to physicians their median estimate was 47 percent. The normative answer is more like 4.8 percent (see Table 1). This answer is obtained by applying Bayes's rule to the statistics given. Bayes's rule shows the correct way to make statistical inferences. Bayes's

Table 1

Bayes's rule applied to Gigerenzer's (1998) cancer example

This shows the probability of some hypothesis, H_j being true given some data, D . Bayes's rule takes the following form:

$$p(H_j|D) = \frac{p(H_j) p(D|H_j)}{\sum_i p(H_i) p(D|H_i)}$$

where $p(H_j)$ is the belief that the hypothesis H_j is true prior to observing the data, D , and $p(D|H_j)$ is the probability of observing the data, given the hypothesis is true. The denominator is the probability of observing the data. It assumes that there are i different hypotheses, one of which is true. We can apply Bayes's rule to Gigerenzer's (1998) cancer example. The two hypotheses are that the patient has cancer, or that he or she does not ($\sim$ cancer). The base rate, or prior, probability of having colon cancer is 0.3 percent. The hit rate of the test is 50 percent, and the false alarm rate is 3 percent. Applying Bayes's rule to this example yields an estimate of having cancer to be 4.8 percent, a much more reassuring estimate than given by the doctors.

$$p(\text{cancer} | + \text{test}) = \frac{p(\text{cancer}) * p(+ \text{test} | \text{cancer})}{p(\text{cancer}) * p(+ \text{test} | \text{cancer}) + p(\sim \text{cancer}) * p(+ \text{test} | \sim \text{cancer})} = \frac{.003 * .5}{.003 * .5 + .997 * .03} = .0477$$

rule implies that given the rarity of cancer, a positive test is more likely to follow a false positive than the actual detection of cancer. As with the Wason (1968) task, subjects' performance on statistical tasks like this clearly shows that they are behaving irrationally. The standard explanation for this is that people tend to ignore base rates; in this example they fail to take into account that cancer is uncommon. That is, only 0.3 percent of people have cancer.

2. People Employ Ecologically Valid Strategies in the Laboratory

Performance on Wason (1968) card tasks, Gigerenzer's (1998) statistical tasks, and other experiments demonstrate that people are, in fact, irrational, when measured against accepted standard inferential rules. There has been a number of proposals to describe what people are doing (e.g., Tversky and Kahneman 1974). The idea behind these approaches is that people are applying heuristics, rules of thumb that tend to be successful in their daily lives, to these laboratory situations where they do not apply. Anderson (1990) has taken this idea one step further, arguing that people are not just bringing real-world heuristics into the laboratory, but rather are employing optimal solutions to certain problems faced in their natural environments. However, behavior that is rational in natural environments may not necessarily be rational in the peculiar environments that experimental psychologists concoct. Some of the roots of this idea are discussed next.

3. David Marr's Levels of Explanation

Anderson's motivations closely follow the arguments that David Marr laid out in his influential book *Vision*.

Marr (1982) argues that fully understanding an information processing system requires considering the system from multiple levels. Marr was interested in the visual system, which he argues evolved to 'tell(s) us about shape and space and spatial arrangement' (Marr 1982, p. 36). The distinction that Marr makes among the levels of explanation can be more readily understood by referring to a far simpler information processing system than vision—a clock. What does a clock do? It indexes time. How does it do it? It achieves this goal by measuring the passage of time by incrementing counters at fixed intervals. Marr says that descriptions like this are at the computational level.

The next level down, called the representation and algorithm level, describes how the goals should be achieved. In the case of a clock it would map the cycles of an oscillator into seconds, minutes, and hours. Descriptions at this level of analysis require a specification of the representations and the algorithms that operate on them to achieve the goals specified at the computational level. Many combinations of representation and algorithms can achieve these goals; a 12 or 24 hour clock can index time. The choice of representation does constrain the choice of algorithm, such as how seconds are rolled over into minutes, and minutes into hours (e.g., what happens at 12:59). Further, not all representation and algorithms are equivalent; some computations may be simpler with one representation than another. Calculating the duration of a trip is simpler when the train leaves at 10:00 and arrives at 14:00 than when it leaves at 10:00 a.m. and arrives at 2:00 p.m.

The lowest level of description in Marr's hierarchy, the hardware implementation level, is concerned with describing the physical entities that carry out the computations of the representation and algorithm level. Here, the current time could be represented by the configuration of hands on a clock face, the coordination of seconds and minutes handled by brass

wheels, the oscillator could be realized as a pendulum. Marr's point is that if you stumbled onto a morass of gears, you would be better able to make sense of these gears if you knew that they were part of a clock as opposed to a cash register, or a sewing machine. Similarly, one is going to be better able to understand people as information processing systems, if one understands what those systems are trying to accomplish. As Marr puts it, 'trying to understand perception by studying only neurons is like trying to understand bird flight by studying only feathers: it just cannot be done. In order to understand bird flight, we have to understand aerodynamics; only then do the structures of feathers and the different shapes of birds' wings make sense' (Marr 1982, p. 27).

4. *John Anderson's Principle of Rationality*

Marr (1982) demonstrated the utility of approaching the problem of vision from multiple levels, and particularly from the computational level. Based on Marr's success with vision, Anderson (1990) hypothesized that the approach might work well for higher-level cognitive processes. Anderson argued that most theorizing in cognitive psychology concerns representations and the processes that act on them, corresponding to Marr's representation and algorithm level. Cognitive neuroscience and neuroscience focus on the hardware implementation level. Relatively little theorizing, he points out, occurs at Marr's computational level. Anderson (1990) explored whether it would be useful to think about higher level cognition, such as categorization (Anderson and Matessa 1992) and memory (Anderson and Milson 1989, Anderson and Schooler 2000) at the rational level, analogous to Marr's computational level. His principle of rationality is that 'the cognitive system operates at all times to optimize the adaptation of the behavior of the organism' (Anderson 1990, p. 28). Anderson renamed Marr's computational level the rational level, because the computational level sounds like it should be describing the algorithm and representation level. His choice of the term rational also serves as an allusion to economic theorizing that often takes place at Marr's computational level. Economists focus more on the decisions agents make, rather than on the processes involved in coming to those decisions. In the economic realm, it is easy to appreciate that it is rational for firms to maximize profits (or, as in the case of Ben and Jerry's, the famous American premium ice-cream maker, to maximize a combination of profits and social good). For the cognitive system, the currency is less clear. Thus, the critical step in what Anderson calls a 'rational analysis' is to figure out what quantity the cognitive system might be optimizing, and to make predictions based on this about how people will behave in particular experimental tasks.

5. *Oaksford and Chater's Rational Analysis of the Wason Card Task*

There would seem to be a weak case for the rationality of the cognitive system in light of people's irrational behavior on the Wason (1968) card task and Gigerenzer's (1998) cancer problem. However, Oaksford and Chater's (1996) rational analysis of the Wason card task shows that typical performance on the Wason, while violating traditional notions of rationality, is indeed quite rational when seen in a broader context of how people seek information in the world.

Consider the following situations that parents may face. In one situation, hearing an utterance from the baby he was caring for, a young father might ask 'was that just babbling?' or if the baby says 'baba,' then the baby wants a bottle. If the mother found her husband giving the baby a bottle, and if she wanted to know whether 'baba' was serving as the baby's signal it wanted a bottle, would it be rational for the mother to ask whether the baby had said 'baba' before the feeding? Now imagine 16 years have passed, and the mother is teaching her son to drive. She wonders whether he knows that when the oil light is on, he should stop the car. Would it be rational for the mother to ask whether the oil light is on? It would seem quite natural for the young mother to ask about what the baby had said, and for the middle-aged mother to remain silent. Asking in the first case would be like flipping the 2-card, irrational with respect to deductive inference. Asking cannot disconfirm the hypothesis that the baby knows how to use the word, since logically many circumstances (e.g., fussing) can suggest an infant is hungry. In contrast, asking about the oil light would be like flipping the 7-card, rational with respect to the rules of logical inference. If the mother found that the oil light was on, this would be a violation of the rule. Our intuitions are at odds with logic.

Perhaps our intuitions are correct that the young mother is rational to ask about what the baby said, and the older mother is rational in remaining silent. The essential difference between the two situations is the amount of information the answer to the question is likely to yield. Though asking whether the baby said 'baba' could not provide a definitive answer, knowing what the baby said tells something about whether the baby knows the word. In contrast, asking about the oil light, though potentially highly informative in the unlikely event that the oil light is on, will tell nothing in the more likely event that the oil light is off.

Oaksford and Chater's (1996) analysis of the Wason task formalizes these intuitions. They assume that people apply experimental, information-seeking strategies to deductive tasks like the Wason. Though the details of their mathematical analysis are beyond the scope of this article, the flavor of it can be given here. As is true for nearly all rational analyses, they assume

that people are behaving as if they are following Bayes's rule. They suggest that people are not treating the task as a test of logic, but rather are attempting to gauge the causal relation between two events. More specifically, they assume what people are really trying to do in deductive tasks is decide between two hypotheses: when p occurs (e.g., baby says 'baba'), q must follow (e.g., baby wants a bottle), or the alternative hypothesis that event p is independent of event q . Sometimes the baby says 'baba,' and sometimes the baby is hungry, and it is only by chance that the baby says 'baba' when it is hungry. In the case of the Wason task described earlier, the competing hypotheses are that an even number depends on a vowel or the alternative that evens and vowels are independent of each other. The question, then, is which card will provide the most evidence in terms of discriminating between these two rival hypotheses. Since people do not know in advance how an experiment (i.e., flipping a card) is going to turn out, they make their decisions based on how much information they expect to gain.

A critical difference between the parenting examples and the Wason (1968) task is that people have experience with how children learn words, but relatively little experience with numbered and lettered cards. Lacking any relevant experience about the cards, Oaksford and Chater (1996) assume that people treat them as if they are typical of causal relations they have seen in the past. When their model of information-seeking is combined with the assumption that causally related events are relatively rare, it predicts the observed subject preferences for flipping cards, namely p (e.g., A-card) is chosen more than q (e.g., 2-card), q more than $\sim p$ (e.g., K-card), and $\sim p$ more than $\sim q$ (e.g., 7-card). So in the Wason task we have a situation where the behavior is irrational with respect to the laws of deduction, but can be understood to be rational in the context of how people seek information.

6. *People are Rational When Ecologically Valid Strategies are Appropriate*

Gigerenzer's (1998) colon cancer experiment demonstrated that people are irrational with respect to proper Bayesian inference. This result seems to be at odds with Anderson's rational hypothesis, as Bayesian analyses underlie nearly all rational analyses. Since the problem was laid out perfectly for Bayesian inference, we would expect people to perform well. Gigerenzer argues that people perform poorly, because the problem format is all wrong. People did not evolve, he argues, to process probabilities, but rather to draw inferences based on their direct experience. His prediction is that people should do well on statistical tasks when the data are presented more naturally in terms of raw frequencies. Consider again the colon cancer problem, but this time in raw frequencies. Thirty out

of every 10,000 people have colon cancer. Of these 30, 15 will test positive. Of the remaining 9,970 people without cancer, 300 will still test positive. What proportion of people who test positive will actually have cancer? Now it is clear that only 15 of the 315 (or 4.45 percent) of those who test positive will have cancer. When the problem was presented this way to another group of doctors, 67 percent answered correctly, as compared to 4 percent when the data were presented in terms of probabilities. When the problem presentation is consistent with how we experience events in the world, people's behavior corresponds to the prescriptions of a normative standard, a Bayesian inference. As with the Wason task, people appear to be behaving irrationally, when the experimental task and conditions conflict with the natural environment.

7. *In Practice, Rational Analyses are Bounded*

The discussions of the Wason (1968) task and Gigerenzer's (1998) task were couched strictly at the rational (or computational) level. In practice, a rational analysis cannot focus strictly on the rational level, but must also consider the algorithm and representation level. The reason for this is that sometimes the rational solution requires calculations that would be physically impossible for any system to perform. In such circumstances a mapping needs to be made from the results of the rational level into algorithms and representations that approximate the computations called for by the rational analysis. It is at the algorithm and representation level that compromises have to be made. In particular, assumptions have to be made about processing limitations. This ties in with Simon's (1956) notion of bounded rationality. That is, people are rational within the constraints of their ability to process information. For example, Simon (1990) pointed out that ignoring processing limitations suggests that knowledge of the rules of chess should lead to perfect play. If a person or machine had an infinite amount of time to contemplate a chess move, and the potential countermove of the opponent to that move, and the countermove to all the opponent's potential countermove, *ad infinitum*, then the person could select an initial move that would inevitably lead them to check mate. The problem is that more board positions would have to be considered than there are molecules in the universe (Simon 1990). Instead, people and chess programs use heuristics to mimic the behavior of this rational strategy. Similarly, because of processing limitations, most rational analyses are approximations to the optimal solution.

There are some games, however, for which it is possible to know the game completely. For example, many adults know the moves and countermove in tic-tac-toe to guarantee that they will at least not lose. Like tic-tac-toe, the Wason (1968) task is atypical. Thus the number of potential hypotheses and experi-

ments raised by Oaksford and Chater's (1996) analysis of the task is relatively small compared to the number raised by problems people often face in their daily and work lives. For example, consider again the problem of interpreting what the baby means when it says 'baba.' The potential number of hypotheses for what 'baba' means certainly exceeds the number of configurations of a chessboard. If Oaksford and Chater's information-seeking strategy were applied to the problem of vocabulary acquisition, where large numbers of hypotheses and experiments are potentially relevant, they would need to employ heuristics and build processing bounds into their system as well.

8. *Whether People are Rational Depends on Your Perspective*

This article demonstrated that there is no definitive answer to the question of whether people behave rationally. For example, the Wason (1968) task clearly demonstrates that people behave irrationally with respect to logic, while Oaksford and Chater's (1996) analysis shows that the behavior is rational when the context is broadened to encompass a person's "normal" life conditions' (Brunswick 1943, p. 259). In short, people can be seen to behave rationally with respect to the environment, but appear to be operating irrationally with respect to a particular task, especially when the task that subjects are performing differs from the one intended by the experimenter.

See also: Functionalism, History of

Bibliography

- Anderson J R, Matessa M 1992 Explorations of an incremental, Bayesian algorithm for categorization. *Machine Learning* **9**: 275–308
- Anderson J R, Milson R 1989 Human memory: An adaptive perspective. *Psychological Review* **96**: 703–19
- Anderson J R, Schooler L J 2000 The adaptive nature of memory. In: Tulving E, Craik F I M (eds.) *Handbook of Memory*. Oxford University Press, New York
- Brunswick E 1943 Organismic achievement and environmental probability. *Psychological Review* **50**: 255–72
- Gigerenzer G 1998 Ecological intelligence: An adaptation for frequencies. In: Cummins D D, Allen C (eds.) *The Evolution of Mind*. Academic Press, San Diego, CA
- Gigerenzer G, Todd P M, the ABC Research Group 1999 *Simple Heuristics that Make Us Smart*. Oxford University Press, New York
- Manktelow K 1999 *Reasoning and Thinking*. Psychology Press, Hove, UK
- Marr D 1982 *Vision*. Freeman, San Francisco
- Oaksford M, Chater N 1996 Rational explanation of the selection task. *Psychological Review* **103**: 381–91
- Oaksford M, Chater N 1998a *Rationality in an Uncertain World*. Psychology Press, Hove, UK
- Oaksford M, Chater N (eds.) 1998b *Rational Models of Cognition*. Oxford University Press, New York
- Simon H A 1956 Rational choice and the structure of the environment. *Psychological Review* **63**: 129–38

- Simon H A 1990 Invariants of human behavior. In: Rozenweig M R, Porter L W (eds.) *Annual Review of Psychology*. Annual Reviews, Palo Alto, CA
- Tversky A, Kahneman D 1974 Judgement under uncertainty: Heuristics and biases. *Science* **185**: 1124–31
- Wason P C 1968 Reasoning about a rule. *Quarterly Journal of Experimental Psychology* **20**: 273–81
- Watkins M J 1984 Models as toothbrushes. *The Behavioral and Brain Science* **7**: 86

L. Schooler

Rationalism

When philosophy consisted of set piece battles between grand schools of thought (such as realism, scepticism, or monism), 'rationalism' referred to the belief that human beings disposed of a faculty called 'reason' which gave them access to the structure of reality. Man, wrote Aristotle (384–322 BC), is a rational animal, but Aristotle recognized that reason had to be trained. In most human beings, the understanding was distorted by passion. Aristotle's teacher Plato (427–347 BC) had argued in *The Republic* and other dialogues that the world we experience was a confused copy of a world of forms or ideas which could be discovered by philosophical inquiry. His parable of the cave (*Republic*, Book VII) is the founding image of rationalism. In its classical Greek version, rationalism assumed that we might understand the structure of the universe by the power of reason, an assumption connected with the fact that rationalism began its philosophical career as a generalization of the procedures of mathematics. In his dialogue the *Meno*, Plato had argued that knowledge is accessible to rational inquiry independently of experience. Later, the Stoics argued that moral knowledge of the laws of nature is available to any rational creature who looks into himself.

Medieval philosophy revived Greek ideas and cross-fertilized them with Christian doctrine. St. Thomas Aquinas (1225–1274 AD) combined reason and revelation by taking his account of nature from Aristotle, and his account of higher things (or 'supernature') from Christian revelation. This synthesis, however, soon began to fall apart, and in modern times rationalism has commonly been taken to refer to rejection of Christian revelation because of rational criticism.

1. *Modern Rationalism*

This was not, however, the view taken by the great rationalist philosophers of the seventeenth century, for whom science was the model of understanding. Descartes invoked God's veracity as a foundation of his system, while Spinoza was a monist who argued

that there was only one substance in the world, which he called 'God or nature'—*Deus sive natura*. Reason and logic alone could generate certain knowledge. Yet even at this point the idea that nature was to be understood as a result of observation was emerging in thinkers such as Francis Bacon who sought 'to put nature to the torture' of experiment. The reason in which seventeenth century thinkers put their faith was less and less the human faculty supposedly giving us access to ends and values than what later came to be called 'instrumental reason' which revealed causes and linked ends and means. Confusingly, reason and rationalism have commonly been used to cover both ideas.

Philosophers had seldom entirely lost sight of the fact that understanding of the world depended upon observation and the seventeenth century growth of science brought this idea to the fore. Its germ lay in the medieval doctrine that we can only fully understand what we can make ourselves—a class of things thought to include not only artefacts but also society and language. After Bacon, the observational basis of knowledge was developed in the *Essay Concerning Human Understanding* by John Locke (1632–1704). Locke was sceptical of syllogistic reasoning, and denied in particular the Cartesian doctrine that reason was innate in man. Locke described himself as an 'underlaborer' in a generation of giants, by which he particularly referred to Isaac Newton the great physicist. Although the founder of empiricism as a coherent alternative to rationalism, Locke's philosophy did not fully free itself from rationalism, particularly in his account of 'powers of the mind.' Empiricism was refined in the work of Bishop Berkeley (1685–1753) and particularly David Hume (1711–76), a line of philosophical development that for some purposes culminated in the synthesis of Immanuel Kant (1724–1804), who argued that we can understand the phenomenal world only in terms of the set of categories (substance, causality, etc.) built into the human mind.

Thomas Hobbes (1588–1679) had used instrumental reason to generate the principles on which the state might be based. In *Leviathan* (1651), he attempted to demonstrate that civil order could be based only on obedience to a sovereign power. The Sovereign guaranteed laws of nature which are rationally derivable as axioms of prudence from the human propensity to avoid death. Hobbes's work was philosophical in that it analyzed the concept of a Commonwealth, but like most political philosophy, it contained some programmatic elements for practical life—positions clearly revealed in his *Behemoth* (written in 1668). In the next century, many thinkers of the French Enlightenment turned reason into an instrument for criticizing society. Writers such as Condillac and Helvetius adopted the empiricism of Locke, but by their emphasis on ideas, and their hostility to Catholicism and the *ancient regime*, they turned it into a form of rationalism. Reason led these *philosophers* to

reject Christian revelation because of its conflict with logic (in the case of the Trinity) or science (in the case of miracles for example). Reason was also invoked against monarchy in favor of a revival of the classical idea of citizenship. These powerful movements of thought exploded in 1789 into the French Revolution, in which reason as the generator of natural rights was pitted not only against experience in epistemology, and against revelation in theology, but against tradition in politics. Reason was now a revolutionary principle seeking to reform everything in social and political life.

2. Rationalism in History

In the nineteenth century, rationalism invaded history by schematizing, as the doctrine of progress, the evolution of mankind from stone age life to the advanced condition of modern Europe. In rationalist terms, the progress of mankind led from superstition and prejudice towards forms of mutual understanding. Each such theory generally located the present as the hour before the dawn of a new era. Auguste Comte the positivist (1798–1857) thought mankind had moved from a theological stage through a metaphysical period and was now moving into a positivist era. The most famous of these adventures of ideas was that of Karl Marx (1818–88), who incorporated his account of human progress into a lossely connected narrative which was passed off as history. The drive to new rationalist accounts of society was inspired by technological optimism. Given that human beings had learned how to control nature, ought they not also to learn how to control society in the interests of human happiness? Rationalism thus became the vehicle of active programs of social reform, while those who took a more pessimistic view of human possibility were stigmatized as reactionary, or even irrational.

Such activist movements often collided with each other, and liberal and socialist programs for the inclusive improvement of the human condition competed for support with exclusivist movements such as those of the Fascists and the Nazis. In terms of the progress of human beings towards universal betterment, the oppression of totalitarian movements seemed not merely irrational, but actually irrationalist. Such irrationalism was often taken to be a form of nihilism, symbolised by the Francoist general whose cry was *Viva la muerte* ('Long live death'). Rationalism thus developed a contrast with irrationalism, understood as any general rejection of the liberal program of rights and the liberal search for universal happiness.

Rational understanding seeks to grasp the universal structure of whatever it studies, and is therefore more at home with theory and universal principles—with science and philosophy—than with history. Science and philosophy are thus pre-eminently rational in a sense in which history, as explanation in terms of

event and circumstance, is not. Further, philosophy assembles in a coherent manner whatever is known about an activity. If knowledge is power, then theoretical knowledge ought to give us power over the world. The most powerful rationalist impulse in recent centuries has been to create an understanding of human life which will allow us to build what is conceived to be a better world.

The problem is that theory and practice seem not to be related in this way. Plato's *Republic* is a masterpiece of political philosophy, but Plato's attempt to improve the politics of Syracuse was a failure. Nevertheless, the belief that successful practice could be encapsulated in a doctrine of a more or less philosophical kind has powerfully influenced the character of modern Western societies. The fusion of a philosophical method with practical ambitions is one of the central meanings of the elusive idea of ideology. Marx along with Friedrich Engels (1820–95), for example, directed their theory towards a new class of political actor, which they called the 'proletariat,' which was instructed in how to take over political power. Niccolo Machiavelli (1469–1527) was a pioneering exponent of this kind of rationalism, and Locke's *Two Treatises of Government* and Jeremy Bentham's (1748–1832) doctrine of utilitarianism might also be understood in this way as handbooks of political skill designed for new actors on the political stage.

3. Critics of Rationalism

The central principle of rationalism in politics, as analyzed by Michael Oakeshott (1901–90), is the sovereignty of technique. The skill of politics (which Lenin (1870–1924) and others thought remarkably easy once mystifications had been shed) can be learned from a book containing basic principles. Politics in this sense is thus no longer the conduct of public business by a small set of rulers, but a universal activity, not basically affected by the particularities of any given political culture. This is why for example Bentham's ignorance of Latin American politics was thought an actual qualification for the project of writing constitutions for that area.

The most common rationalist image of politics is that of a building program, in which abstract foundations logically generate a good society. Such projects are often based on abstractions such as democracy, justice, rights, or community. Rationalist programs focus on only one such desirable foundation, for the danger (as Plato had already recognized) is that two principles might come into conflict. The aim is to solve the problem of conflict, and both rulers and ruled need instruction from theoreticians (called 'education') in the philosophical foundations of the proposed good society.

Contemporary normative political philosophy is rationalist in its concern to analyze some basic norm of current liberal societies, commonly with an eye to

improving practice. In the analysis of a just society advanced by John Rawls in *A Theory of Justice*, the future members of the society guarantee the rationality of the rules of justice because the rules have been agreed behind the famous 'veil of ignorance' in order to avoid the contamination of any particular knowledge. The equality of the Rawlesian citizens is thus matched by their vacuity.

The practical aspects of this form of rationalism are driven by instrumental reason in the sense that thinking takes off from whatever is construed as a problem (usually, indeed, a 'social problem') and the aim is to achieve a solution. Rationalism is sometimes criticized as assuming that all human beings are basically good, and only made bad by bad institutions, but the more fundamental assumption is that human beings are plastic entities who can be made into whatever type of social being is desired through the power of ideas. Politics thus becomes not so much social as attitudinal engineering; it requires removing the wrong ideas (racism, sexism, etc.) from society and restocking it with the right ideas. The general project of improving the world by 'modernizing' it is one of the most powerful of current rationalist projects.

In the second half of the twentieth century, rationalism as the search for the right foundations of society was challenged by people who said that the search was futile. There were no rational foundations, merely the changing responses of different sets of people. The derivation of this form of rationalism was from Heidegger (1889–1976) and Wittgenstein (1889–1951) but it was popularized in the United States by disciples of John Dewey such as Richard Rorty in *Philosophy and the Mirror of Nature* (1980). Radical critics charged that the whole rationalist project was designed to establish the hegemony of some specific group—capitalists, white males, commissars, and the like. Ethical opinions such as 'political correctness' and political projects such as multiculturalism were juxtaposed against rationalism. These revived forms of scepticism look like what rationalists would see as an irrationalist challenge to Western life, but, as is common in forms of scepticism, there are discernible traces of rationalist dogmatism in the egalitarian morality and communitarian politics taken for granted in the practices of the challengers.

See also: Aristotle (384–322 BC); Comte, Auguste (1798–1857); Hobbes, Thomas (1588–1679); Justice: Philosophical Aspects; Justice: Political; Locke, John (1632–1704); Marx, Karl (1818–89); Progress: History of the Concept; Rational Choice Explanation: Philosophical Aspects; Rationality in Society; Social Justice

Bibliography

Berlin I 1998 *The Proper Study of Mankind*. Farrar, Straus and Giroux, New York

- Gellner F 1992 *Reason and Culture*. Basil Blackwell, Oxford, UK
- Oakeshott M 1991 *Rationalism in Politics and Other Essays*. Liberty Press, Indianapolis, IN
- Popper K R 1966 *The Open Society and its Enemies*, 5th edn., Routledge, London
- Rorty R 1980 *Philosophy and the Mirror of Nature*. Basil Blackwell, Oxford, UK

K. Minogue

Rationality and Feminist Thought

One of the most profound reorganizations of the twentieth century has been a thorough re-evaluation of concepts of rationality as they have evolved in Western culture for the past three or so millennia (Gigerenzer et al. 1999, Labouvie-Vief 1994). This reorganization has been of profound interest to women, who have actively participated in the many lively debates about the nature of rationality. Models of rationality are, in large part, aimed at specifying the nature of humans and their desirable and less desirable characteristics, and, since these characteristics often were thought to find their most prototypical expression in men rather than in women, the issue of gender occupies a central place in this profound metatheoretical shift.

In response to traditional dualistic notions of rationality and the implied devaluation of women's rationality, some women have pointed to the fact that women traditionally were prevented from cultivating rationality (Hollingsworth 1926). Others have suggested that the traditional masculinist concept of rationality does not apply to women who utilize uniquely 'feminine' ways of knowing (Belenky et al. 1986, Gilligan 1992). Neither of these alternatives, however, transcends the traditional dualistic concept of rationality. Thus, a further suggestion has been that we move away from dualistic concepts of rationality altogether. Accordingly, concepts of reason are being transformed and new models are emerging that acknowledge that rational and nonrational processes mutually enhance each other and contribute in the process of knowing (Labouvie-Vief 1994).

1. Rationality and Women's Mind in History

Concepts of rationality have been central to definitions of human nature. Often couched explicitly in terms of the mind and its relationship to emotions, social context, and culture, rationality was defined dualistically—as that which transcends these categories, i.e., as conscious thought, abstract thinking, and universal laws of truth (Labouvie-Vief 1994). Its relation to gender often was more implicit, but the rational being was a 'man of reason' (Lloyd 1993).

This association of reason with masculinity—and of femininity as a lack of reason—widely pervaded Western intellectual history and is visible not only in philosophy and science, but also in art and literature, myth and religion, and ultimately, the social structures and social regulation processes that have supported these representations.

1.1 The Gendered Mind in Philosophy

The dualistic theory of rationality and the mind became institutionalized when Greek philosophers began to speculate about human nature. These efforts culminated in the work of Plato, who proposed a two-layer view of human nature and reality to supplant more ancient views (Labouvie-Vief 1994). For more ancient views of the self as a collection of motor actions, bodily processes, or mythic-divine injunctions, Plato substituted a language of a self no longer identified primarily with its bodily processes and concrete actions, but of a mental agent different from its bodily manifestations. The new language was that of a self who was the author of those actions, of a psychological causal agent who was at their center and who was responsible for them.

Plato's dualistic, two-layered conception of the individual contrasted a layer of mind, abstract thought, soul, spirit, or ideals to a layer of body, matter, and enslavement to concrete sensory and organismic happenings. These layers were arranged hierarchically, with mind-spirit forming the superior pole, body, and senses the inferior one. That hierarchical worldview also became associated with gender. Thus, in the *Symposium*, Plato argued that the spiritual love between men who create ideas is more valuable than the physical and material love with women, since women's creativity is merely one of bodily procreation rather than the creation of universal ideas in which men participate (Labouvie-Vief 1994).

1.2 Women's Mind in Western Science

The association of the masculine with mind and spirit, and of the feminine with body and matter, widely came to characterize thinking about and representations of women. It has characterized not only the view of most philosophers but has also remained basic to other theoretical accounts of human nature. For example, Aristotle (*De Generatione Animalium*) thought that women's contribution to fertilization consisted of a passive and material principle rather than an active and spiritual masculine one (Lerner 1986). Freud proposed that women's lack of phallic sexuality made their sexual and psychological orientations more passive and masochistic (Freud 1925). Others taught that the anatomy of women's brains made them less capable of conscious, rational thought and of extraordinary creativity or achievement (Bem

1993, Labouvie-Vief 1994). Even in the twentieth century, the notion became popularized by de Beauvoir (1952), who argued that woman is closer to nature because physiology makes her so: Her mode of creating is by submitting to a natural bodily function, by remaining enslaved to her biological fate, while the man creates by 'acts that transcend his animal nature.' By creating culture, the man is thought to be involved in a new, freer form of creation, one that transcends what is given organismically and shapes the future rather than passively suffering it.

This gendered dualistic notion of mind has been influential in shaping theoretical notions of development. Traditionally, development has been construed as an onward rush, a heroic and energetic striving for mental ideals and control, a triumph of conscious, abstract forms of thinking, a victory of rationality over emotions and the body. Since these features of the mind also are identified with masculinity, femininity was represented as an inferior and less developed state of being—a form, as it were, of mental and physical castration. Freud (1925) was most explicit in equating development of mind and identity with a progression from the domain of the castrated mother to the potent and genital father, but most major theories of development of the past followed this same pattern of the devaluation of the emotive, imaginative, and sensory aspects of the self (Labouvie-Vief 1994).

1.3 Women's Mind in Myth, Art, and Religion

The notion of the masculine as identified with mind, spirit, and transcendence is also pervasive in mythology and religion, literature, and art (Labouvie-Vief 1994, Ortner and Whitehead 1981, Rosaldo and Lamphere 1974). In all of those, one polarity of human functioning tends to be idealized and even inflated as 'masculine': it is based on an imagery of rise, light, and sun, and it symbolizes such psychological qualities as leadership, rational self-control, self-assertiveness, individuality, and willfulness. The other polarity, in turn, is degraded and devalued: its imagery is based on falling, being conquered and surrendering, and engulfment in dark spaces, and it symbolizes suffering, powerlessness, and entrapment in unconscious processes. In mythology and religion, such imagery often occurs as the theme of the overthrow of original female divine figures by the male godheads or heroes who came to be the original creators, creating the world not organically but conceptually, by pronouncing a word (Lerner 1986). A well known example of this process of the replacement of the feminine principle is offered by the great Greek tragedies by Aeschylus and Euripides centering on the myths of Agamemnon and Orestes. These myths address the question of which of two murders was a more heinous act. Was it Klytemnestra's killing her husband Agamemnon to revenge his sacrificing their daughter Iphi-

genia to the gods so as to secure a good wind for his war ships. Or was it their son Orestes's killing of his mother to avenge the murder of his father? The answer was that Orestes could be forgiven because he represented a new masculine order that superseded an earlier feminine order—a new paradigm in which fatherhood was said to have replaced motherhood as a principle of creation (Labouvie-Vief 1994).

2. Explanations of Dualistic Theories of Mind

How universal is the equation of masculinity with mind and rationality, and of femininity with an 'older' emotional and organic order of being? Much writing has suggested that such patriarchal ideologies in actuality came to replace more women-focused ones of previous times. Originally proposed by Bachofen in 1861 (Duby and Perrot 1992/93), this 'matriarchy' thesis holds that there must have been, in prehistoric times and before the rise of the major patriarchal societies, the reign of a maternal principle. However, the historical accuracy of this thesis has recently been doubted. Instead, a number of anthropological, sociological, and psychological explanations have been proposed.

2.1 Anthropological Evidence

Most historical and anthropological evidence (Duby and Perrot 1992/94, Ortner and Whitehead 1981, Rosaldo and Lamphere 1974) does not support the view that earlier and/or less civilized societies granted women more power universally. On the contrary, strong patriarchal systems in which women are devalued along a mind-nature dimension exist worldwide in preliterate societies. Often such societies also have myths of early matriarchies that eventually were superseded by the rule of man. However, such matriarchy myths can be part of an ideology that devalues women (Rosaldo and Lamphere 1974). Around the world, such myths often tell of the past bad behavior and evil of women that eventually was ended only when men overturned the power of women and took control. In this way, matriarchy myths can be used to justify the male order. They form part of a cultural code distinguishing men from women in terms of morality and authority, they incorporate the values that grant men more power in social life. Thus, even though they talk about a time before the current social order, they fix that order as legitimate and invariant.

2.2 Sociological Mechanisms

Although there is considerable variation in women's behavior and activities across cultures and historical times, there are nearly universal differences in what is believed to constitute the nature of women (Brown

and Gilligan 1992, Labouvie-Vief 1994, Rosaldo and Lamphere 1974)—and, correspondingly, the nature of men (Gilmore 1990). In many cultures, women are held to be nurturers, to be gentle and concerned with caring for others rather than self-expression, they are considered passive and unaggressive and to readily subordinate themselves to those in power. They are felt to be closer to children—more emotional and expressive, less capable of rational decision-making and abstract thought. They are also believed to be poorly suited for positions of leadership and power (Bem 1993, Gutmann 1987, Ortner and Whitehead 1981, Rosaldo and Lamphere 1974). Sometimes these beliefs are attributed to biological bases of gender, but many think they root in the social roles and realities of women's lives.

In most societies women are subordinated socially or, at least, they do not hold publicly recognized power and authority surpassing that of men (Ortner and Whitehead 1981, Rosaldo and Lamphere 1974). In some societies, individual women may have achieved considerable power and recognition, but in general all societies tend to exclude women from certain crucial economic, political, and intellectual activities. Even in societies where women exert considerable power by doing much productive work and controlling resources, their activities are not associated with the same social status as those of men. Their roles as wives and mothers are associated with fewer powers and prerogatives than are the roles of men (Bernard 1987, Duby and Perrot 1992/93, Ortner and Whitehead 1981, Rosaldo and Lamphere 1974). In turn, this social devaluation is often attributed to some major biological deficiencies (Bem 1993, Labouvie-Vief 1994). These same 'lacks' also were thought to make women eminently suitable to take on positions subservient to men.

2.3 Psychological Mechanisms

The near universality of the devaluation of women's mind also has suggested that the construal of women as mentally inferior and nonrational may be rooted in psychological processes through which gender is defined. Psychoanalyst Stoller (Herdt and Stoller 1990) and sociologist Chodorow (1978) suggested that these processes render the gender development of boys particularly vulnerable. This vulnerability results from the mother's role in early development. For both boys and girls, the mother defines an early core of identity, which in the case of the boy's development becomes problematic. He, unlike the girl, needs to change his core identity from 'feminine' to 'masculine', a change that leaves a life-long vulnerability of self. As a result, men may defensively devalue women as shown in cultural practices (usually, related to initiation rites) aimed at defining the unique nature of the 'masculine' as a way of creation unique to men and 'better' than the 'feminine.' Complementary to the definition of

manhood, womanhood in many cultures is defined by practices aimed at creating a sexual and social object to the action of men—whether it is direct genital mutilation or isolation from participation in public life through withholding education and high status work (de Beauvoir 1952, Labouvie-Vief 1994, Ortner and Whitehead 1981).

However, such mechanisms may be themselves relative to the family structure that operates in a society strongly divided along gender lines and devaluing women (Bem 1993, Labouvie-Vief 1994). But with family structure and gender roles changing to create more shared and equitable activities among men and women, these domains are becoming less clearly related to gender, *per se*.

3. Rationality and Women's Minds

Traditional dualistic theories of rationality, with their devaluation of femininity and inflation of masculinity, have influenced deeply women's process of growing up. Whereas masculine development is associated with increases in abstract thought, in individuality, independence, and self-confidence, women often have experienced their development as renunciation, surrender, and victimization—a loss of voice, a blunting of mind, a process of silencing (Bem 1993, Brown and Gilligan 1992, Labouvie-Vief 1994). Such struggles of women with their own sense of creativity is expressed very poignantly in the lives of many writers of the twentieth century, such as Sylvia Plath or Virginia Woolf.

As educational and employment opportunities have opened up for women and as women have begun to participate in defining themselves, research has upheld few objective differences in intellectual status between men and women. Although women tend to score somewhat higher on verbal and men on mathematical abilities, these differences are overall small (Linn and Hyde 1989). Moreover, since Hollingsworth's (1926) pioneering work showing that gender differences in intellectual can be explained usually in terms of differences in educational opportunities, research has documented that over time, as disparities of opportunity are declining, gender differences in intellectual functioning are becoming smaller, too (Linn and Hyde 1989).

Yet despite evidence that female intellectual endowment does not lag behind that of men, women on average continue to lag in achievement behind men. Moreover, this gender lag increases in the course of development. Indeed, in early childhood, girls surpass boys on most measures of school achievement. However, as they grow older, girls often are found to retreat from achievement-related challenges; standing out for women becomes a source of conflict and shame, and they attempt to hide their excellence (Brown and Gilligan 1992, Labouvie-Vief 1994). Thus, women's feelings of shame and discounting of their achievement

persist well into the present day. However, in many cultures women's transition into later life is often characterized by a degree of freeing from the cultural constraints of 'womanhood.' No longer defined as sexual objects, women often move from a world of confinement to one of daring openness, of a release from their earlier interiority. Thus, from young adulthood to midlife, women have been found to become less traditionally 'feminine' but more dominant, independent, objective, and intellectually oriented. In many societies this figure of the older woman emerges as the woman of power who holds authority over the family and who can control healing, magic, witchcraft, and other powerful (and often dangerous) resources (Gutmann 1987, Rosaldo and Lamphere 1974).

4. *Transforming Concepts of Mind and Gender*

This review has shown that traditional concepts of rationality have implied a strong gender bias, one that has corresponded with the stereotyping of women as nonrational and more closely bound to nature and their bodies. Such a feminist critique ultimately does more than restore to women the claim to be rational, or to claim for them a unique form of knowing that is different from the rationality of men. Rather, the intricate association of traditional concepts of gender and rationality reveals that the very claim that women lack rationality is problematic, since it was on a narrow concept of rationality in the first place. Thus, the association of mind and gender indicates the need to reorganize the very meaning of 'rationality' in theories of mind and knowledge. A transformation of the concept of rationality is not just a feminist enterprise. Criticisms of traditional views of rationality have been a mainstay of the intellectual history of this century, which systematically has eroded beliefs in dualistic concepts of rationality (Baynes et al. 1987). The mind and knowledge are no longer believed to stand secure above nature and human foible. Instead, we are witnessing a growing belief that the dimensions of the mind and self that were abandoned in the classical view must in fact be included in a viable theory of mind and knowledge. Emotion and intuition, time and process, variability and disorder—all the elements previously devalued as feminine have taken a forefront place in philosophical discussions, and are being incorporated into new theories of the mind. Thus, reason also has an immanent pole and must be explained through the emotional propensities, cultural scenarios, linguistic practices, and narrative and symbolic structures of meaning.

Freud's (1925) original proposal that rational processes are built upon nonrational processes was one early voice warning that rationality ultimately must include a balance between our biological nature and our strivings for transcendent realities. It was he who pointed out the paradox that 'rationality,' when one-

sided and unbalanced, in fact is a sign of the defensive break through of 'irrational' processes. However, Freud's very concept of 'irrational' processes still implied remnants of a dualistic and oppositional view, since it relegated these processes to a problematic status. Instead of such dualism, notions of balance and exchange have been carried forward into modern theories of the mind. In those theories, a tremendous part of knowledge is, in fact, based on processes that are profoundly adaptive, yet nonrational. Such theories follow Darwin's insight that biology provides the basis for the behavioral patterns—such as emotions—that are of the most urgent adaptive significance. More generally, we are realizing that an important contributor to knowledge are other non-rational decision systems or 'heuristics' that allow us to deal with information processing in situations where time and resources are limited (Gigerenzer et al. 1999, Labouvie-Vief 1994).

Sometimes the revaluation of the nonrational is taken as a dethroning and destructive critique of rationality. Instead, calling for a transformation of the concept of rationality implies that it merely constitutes one important pole or function of the mind. Thus body, nature, emotions, and concrete existence vs. mind, culture, reason, and necessary, deductive law no longer are considered in hierarchical terms. Instead, the mind becomes a cooperative interplay of these two sets of processes, rational and nonrational. Such dual processing theories of the mind are, in fact, proliferating and becoming the standard in theories of the mind (Chaiken and Torpe 1999, Gigerenzer et al. 1999, Labouvie-Vief 1994).

Such a reorganization of theories of the mind inevitably also involves a reexamination of the role of gender in the order of things. Indeed, such a reevaluation is visible in many strata of contemporary thought and life, visible not only philosophy and scientific theory, but also religion, literature, and art. These efforts allow us to acknowledge that rationality and non-rationality are processes of a mind that is, at last—and at least theoretically—part of nature and independent of gender.

See also: Androgyny; Child Care and Child Development; Deconstruction: Cultural Concerns; Feminist Economics; Feminist Epistemology; Feminist Ethics; Feminist Legal Theory; Feminist Theology; Feminist Theory; Feminist Theory and Women of Color; Feminist Theory: Ecofeminist and Cultural Feminist; Feminist Theory: Liberal; Feminist Theory: Marxist and Socialist; Feminist Theory: Postmodern; Feminist Theory: Psychoanalytic; Feminist Theory: Radical Lesbian; Foucault, Michel (1926–84); Gay, Lesbian, and Bisexual Youth; Gay/Lesbian Movements; Gender and School Learning; Mathematics and Science; Gender and the Law; Gender, Feminism, and Sexuality in Archaeological Studies; Gender-related

Development; Heterosexism and Homophobia; Homosexuality and Psychiatry; Irrationality: Philosophical Aspects; Lesbians: Historical Perspectives; Lesbians: Social and Economic Situation; Love and Intimacy, Psychology of; Mind–Body Dualism; Post-modernism: Philosophical Aspects; Queer Theory; Rationalism; Rationality in Society; Regulation: Sexual Behavior; Sex-role Development and Education; Sexual Behavior: Sociological Perspective; Sexual Orientation and the Law; Sexuality and Gender; Theory: Sociological; Transsexuality, Transvestism, and Transgender

Bibliography

- Baynes K, Bohman J, McCarthy T T (eds.) 1987 *Philosophy: End or Transformation*. MIT Press, Cambridge, MA
- de Beauvoir S 1952 *The Second Sex*. Knopf, New York
- Belenky M F, Clinchy B M, Goldberger N R, Tarule J M 1986 *Women's Ways of Knowing*. Basic Books, New York
- Bem S L 1993 *The Lenses of Gender*. Yale University Press, New Haven, CT
- Bernard J S 1987 *The Female World from a Global Perspective*. Indiana University Press, Bloomington, IN
- Bettelheim B 1962 *Symbolic Wounds: Puberty Rites and the Envious Male* (rev. edn.). Collier Books, New York
- Brown L M, Gilligan C 1992 *Meeting at the Crossroads: Women's Psychology and Girls' Development*. Harvard University Press, Cambridge, MA
- Chaiken S, Trope Y (eds.) 1999 *Dual-process Theories in Social Psychology*. Guilford Press, New York
- Chodorow N 1978 *The Reproduction of Mothering*. University of California Press, Berkeley, CA
- Duby G, Perrot M (eds.) 1992/94 *A History of Women in the West*, Vols. 1–4. Harvard University Press, Cambridge, MA
- Linn M C, Hyde J S 1989 Gender, Mathematics, and Science. *Educational Researcher* 18(8): 17–27
- Freud S 1925 Some psychological consequences of the anatomical differences between the sexes. In: Rieff P (ed.) 1963 *Sexuality and the Psychology of Love*. Macmillan, New York pp. 183–93
- Gigerenzer G, Todd P M, ABC Research Group 1999 *Simple Heuristics that Make Us Smart*. Oxford University Press, New York
- Gilmore D D 1990 *Manhood in the Making: Cultural Concepts of Masculinity*. Yale University Press, New Haven, CT
- Gutmann D 1987 *Reclaimed Powers: Toward a New Psychology of Men and Women in Later Life*. Basic Books, New York
- Herd G H, Stoller R J 1990 *Intimate Communications: Erotics and the Study of Culture*. Columbia University Press, New York
- Labouvie-Vief G 1994 *Psyche and Eros: Mind and Gender in the Life Course*. Cambridge University Press, New York
- Lerner G 1986 *The Creation of Patriarchy*. Oxford University Press, New York
- Lloyd G 1993 *The Man of Reason: 'Male' and 'Female' in Western Philosophy* 2nd edn. University of Minnesota Press, Minneapolis, MN
- Ortner S B, Whitehead H (eds.) 1981 *Sexual Meanings: The Cultural Construction of Gender and Sexuality*. Cambridge University Press, New York

- Rosaldo M Z, Lamphere L 1974 *Women Culture and Society*. Stanford University Press, Stanford, CA
- Schott R M (ed.) 1997 *Feminist Interpretations of Immanuel Kant*. Pennsylvania State University Press, University Park, PA

G. Labouvie-Vief

Rationality in Society

Human behavior is rational when appropriate to its goals. The brain provides the biological foundation for human rationality; it accepts sensory stimuli from both the natural and social environments and stores knowledge, beliefs, and values that guide the thought processes leading to action. This article focuses upon the social environment of human rationality.

Rationality has been an object of scholarly analysis since Greek times, but largely as part of ethics, in its concern with goals, sound reasoning and logic. Empirical study of human rational behavior is mostly quite recent, although we find such significant precursors as Machiavelli. Even Adam Smith provides no explicit theory of rationality in his *Wealth of Nations*, and such theories begin to appear only in the nineteenth century (e.g., Cournot 1838), initially in economics and, much later, in sociology (Parsons 1937, see also *Action, Theories of Social*). In the twentieth century, utility theory took shape in economics, while more realistic but less formal theories of rationality developed in the other social sciences.

1. Theories of Rationality

The social context molds behavior in three main ways. First, it largely determines what actions are available. Second, choice must try to anticipate the reactions of others to contemplated actions. Third, people's society is the major source and shaper of the knowledge, skills, and goals they use in choosing actions.

Hence, although a large part of today's understanding of rationality derives from psychological experiments with individuals, most of the contents of individuals' knowledge and beliefs, and their cognitive processes derive from social sources. Moreover, experimental stimuli themselves often are stated in language that has rich social content. Thus, even 'individual' psychology is the study of highly socialized beings (Simon 1992). The entrepreneur and consumer of economics, the voter and legislator of political science, the husband and delinquent, and soldier of sociology and anthropology are deeply social.

1.1 The Rationality of Experts

Cognitive psychology has acquired substantial knowledge about the decision-making of experts and novices

in such professional domains as medical diagnosis, chess-playing, and scientific discovery, and has successfully simulated these behaviors on computers (Anderson 1999, Newell and Simon 1962, Shrager and Langley 1990; see also *Expert Memory, Psychology of; Scientific Reasoning and Discovery, Cognitive Psychology of; Chess Expertise, Cognitive Psychology of; Medical Expertise, Cognitive Psychology of*). Within the professional domain, an expert employs two main procedures: (a) recognizing familiar situations and patterns that give access to knowledge in memory about handling these situations; and (b) searching selectively for satisfactory problem solutions, guided by learned rules of thumb, or heuristics. The same procedures are used to find better problem definitions and representations, instruments and tools for observing or reasoning, experimental designs, and search strategies. They even account for the problem solving that is called 'intuitive,' 'insightful,' or 'creative.'

1.2 Rationality as Utility Maximization

Utility maximization, the best developed formal theory of rationality, which forms the core of neoclassical economics, does not refer to the social context of action (see also *Decision Theory: Classical*). It postulates a utility function, which measures the degree to which an individual's (aggregate) goals are achieved as a result of their actions. The rational actor chooses the action, from among those given, which maximizes utility. If the actor's goals are food and sleep, then rationality calls for choosing the attainable combination of food and sleep that yields the greatest utility. The core theory does not specify the content of the utility function (Simon 1983).

Without a concrete meaning for utility, the theory has few refutable consequences; its relevance to real situations depends mainly on adding auxiliary assumptions. This is not wholly true, for the theory does assume constancy of the utility function over sequences of choices. Many experiments have shown that this implication is testable—and frequently false (Tversky and Kahneman 1974). Human actors often make sequences of choices that are incompatible with any consistent utility function.

In order to predict more than consistency in any particular situation, additional strong assumptions must be added. Economic applications usually add the vastly oversimplified assumption that utility is measured by total income or wealth, which, in turn, may be expressed in terms of goods possessed. Additional assumptions introduce markets, in which goods can be exchanged (for each other or for money), and firms, which consume goods to produce other goods.

Without auxiliary assumptions, the theory is also silent about the range of available alternatives and the processes used to find the optimal action. It says nothing about how to select auxiliary assumptions, or

what empirical support they should require. As a result, assumptions are often introduced in order to preserve or enhance mathematical tractability.

To produce an empirically sound theory, a much richer set of empirically based auxiliary assumptions must be provided, to take account of the actual limits on human thinking, and the influences of its social context (Green and Shapiro 1994).

1.3 Forms of Rationality

An action may be regarded as rational because of the way it was chosen (procedural rationality), or because of its outcomes (substantive rationality) (see also *Decision Research: Behavioral; Bounded Rationality; Decision and Choice: Economic Psychology; Heuristics for Decision and Choice*).

Attributions of rationality may or may not respect the limits of actors' abilities to adapt actions to goals. Utility-maximizing behavior, as discussed previously, selects an action that is among those objectively available, attains a real-world optimum, and is of a comprehensive function of wants and needs (utility function).

As an alternative, bounded rational behavior selects some action among those that memory, observation, and thought bring to attention. It is aimed at attaining outcomes that satisfy (attain aspirations) for the wants and needs that have been evoked (Conlisk 1996). The expert behavior described in Sect. 1.1 is an example of bounded rationality.

There is a close affinity between optimizing and substantive rationality, and an affinity between bounded and procedural rationality. The utility-maximizing optimum is independent of the choice process; the outcome is all that counts. In bounded rationality, the choice depends upon the process leading up to it. The social sciences, other than neoclassical economics, usually explicitly incorporate human cognitive limitations and social influences into their theories, using some version (seldom highly formalized) of bounded rationality. Bounded and procedural rationality require generating courses of action, assessing their possible outcomes under different futures and estimating how much satisfaction they will attain. They postulate procedures that respect human limits on knowledge of present and future, on abilities to calculate the implications of knowledge, and on abilities to evoke relevant goals (Simon 1983).

1.4 Habits, Emotion, and Motivation

Although dictionaries often define rational action narrowly, as the action chosen by reasoning, numerous actions that are rational (i.e., appropriate to goals) are taken without conscious deliberation or any deliberation at all. Rather, these rely on heredity responses,

or prior knowledge and belief (e.g., see also *Perception without Awareness*, *Psychology of Emotion in Cognition*). Evolution has seen to it that eyes blink when something comes flying at them. Similarly, but as a result of learning and experience, when a car suddenly stops, the driver in the car behind applies brakes without deliberation. These goal-oriented, but unreflective actions make up much of the activity of everyday life, and greatly expand the capacity for (boundedly) rational response. Rationality is inextricably interwoven with emotion and motivation. Action is directed toward goals by such motivators as hunger, sexual drive, anger, fear, pride, and others. These are represented by the aspirations of bounded rationality. Evoked into consciousness, or activating 'automatic' responses, they provide the goals for rational choice. Additionally, cognitive processes cannot handle large quantities of information simultaneously. Deliberative and conscious thoughts have to pass through the narrow straits of short-term memory, which hold only a few symbols (approximately six), and can attend to only one thing at a time (or perhaps two or three, by alternation of attention over intervals of 1–2 seconds). Recent research using functional molecular resonance imaging to record neuronal activity has shown that even simple acts (reading a short sentence) employ a sequence of neural processes (see also *Short-term Memory*, *Cognitive Psychology of Working Memory*, *Psychology of*).

Because human beings must satisfy many drives, rationality fixes attention on particular goals, usually until these are satisfied, but shifts attention to other goals when they become more pressing. Interruption is triggered both by sensory processes, such as peripheral vision, and by emotions and inner drives, such as fear and hunger. So rationality, is not just a cognitive function, but depends on motivation and emotion to control attention.

2. Rationality in Social Context

The social context of action that plays such an enormous role in determining choices will now be examined more closely.

2.1 The Sources of Knowledge

Most of the knowledge that enters into either habitual responses or deliberate choice is received from other persons, beginning at birth. Direct experience of the nonhuman world contributes only a small fraction of what people know and believe. Language, a major tool of thought, is a social artifact (Simon 1992).

Of course, what people know is also strongly influenced by the natural environment, some of it sensed directly. Nonetheless, much of one's knowledge of this environment is acquired by social processes,

not direct sensation; sensory stimuli are interpreted by social information. People believe that cholesterol is bad for them, but their belief derives from doctors, the media, and friends, not their own laboratory experiments. The same rains fall on humankind around the world, wetting anyone it strikes. It enters into each person's beliefs and knowledge in many ways, and social processes have created a wide range of different expedients for mitigating its discomforts and using its waters. The range of available actions and the criteria and procedures for choosing among them are largely circumscribed by the actions, criteria, and procedures discovered and propagated by each society.

2.2 The Social Sciences are Historical Sciences

Because of the social origin of knowledge, the social and behavioral sciences are, essentially, historical, for the behavior of each person changes with the enveloping society's states of belief and knowledge. Functionalism in anthropology and sociology has pointed out the common functions shared by all societies, but also the variety of ways in which these functions are performed in particular societies, depending on their values, beliefs, and knowledge.

Of course, dependence on history is not peculiar to the social sciences. Historical laws are found throughout natural science: in modern astrophysics, the Big Bang and its sequelae; in geology and oceanography, mountain formation and continental drift; evolutionary biology is wholly historical.

Today, the social sciences differ greatly in the extent to which their theories attend to history. Historical change in basic laws has played only a small role in utility-maximizing economic theory. The central theoretical structure in economics today, as in Adam Smith's day, is an idealized set of markets with vestigial business firms (each characterized by a production function). Perhaps these roughly match Adam Smith's world, but they are a far cry from the economy of large international business corporations throughout today's world; and this model does not explain the growth (or even the possibility) of these large organizations.

On recognizing its historical dependence, a science must reexamine the meaning of natural law. It must seek to find invariant underlying relations that account for continuing change as well as equilibrium in dynamic systems. In the social sciences, temporal structural change is still absent from most formal theory.

2.3 Social Determinants of Motives

Social modification of motives and actions is usually dealt with in contemporary social science by introducing evolutionary mechanisms (Nelson and Winter 1982). In most current versions of Darwinism, the sole

effective motive is to increase fitness (essentially rate of population growth); altruism will not compete. This has often been interpreted as implying that only selfish human motives survive, and that, consequently, maximization of utility means maximization of income or wealth, profit of the business firm, or power (the latter in Public Choice Theory in political science and sociology). In this view, only the special genetics of ants and bees, and some other social insects, allow nonpropagating offspring to exhibit altruism to close kin.

However, it has been shown (Simon 1997) that, in the presence of bounded rationality, altruism may increase the fitness of altruists, and, hence, may survive and grow within a society. If bounded rationality admits a large capability for social learning, persons willing and able to learn (docile), will have a large fitness advantage over others. By natural selection the society will help individuals acquire knowledge and beliefs that usually enhance their individual fitness, but are sometimes advantageous to the society's fitness at the individual's expense (e.g., volunteering for military service). As long as social influence, on balance, increases individual fitness, altruism can also thrive, contributing to the fitness and growth of the society (see also *Altruism and Self-interest*).

2.4 Altruism and Group Identification

Perhaps the most important human expression of altruism is identification with, and loyalty to, groups ranging from the family, through ethnic or religious groups, to whole societies or nations. Group identification is a major determinant both of goals (defining the 'we') and of knowledge and beliefs, for both are formed in the groups that people associate with in work and leisure. Identification is the primary basis for the intense within-group collaboration and between-group conflict that is an enduring characteristic of human interaction. It is also one of the bases for the ability of large organizations to coordinate human efforts efficiently (see also *Identity: Social*).

2.5 Organizations

In a modern society, formal work organizations—companies, government agencies, schools—provide environments for a large fraction of human waking hours (Simon 1997). This has presumably come about because organizations can, through appropriate specialization and coordination of activities, enormously increase individual productivity. Effective specialization requires dividing up activities in such a way that each specialty calls on only a small part of the knowledge and skill that is required for the whole, and each component is mainly concerned with the initial inputs and final outputs of the others, and need not be involved with their internal processes.

At the same time, it is rarely possible to obtain complete independence among components, so that some provision, often involving the exercise of authority, must be made for securing coordination where it is required (the axles that one department makes must fit the wheels made by another). The theory of organizations, in business and public administration, is concerned with the architectural arrangements that secure a good balance of specialization and coordination, and the means of using authority, organizational identification and rewards to secure acceptance of the structure by those who inhabit it. Because of the difficulty of measuring the precise contributions of individuals' work efforts to an organization's product and success, economic incentives cannot account for the high level of support for organization goals that is often observed (for example, in the behavior of employees who do not share in profits or bonuses). Much of this support can be attributed to organizational identification, and it is one of the novelties of the modern world (with perhaps some precedents in armies, even of the distant past) that strong loyalties can be evoked by formal organizations as well as by more traditional social groups. Human attachment to organizational goals has been a major force in converting the market economy of Adam Smith's time into the very different organizations-cum-markets society of the present (Simon 1997).

Recent developments in the formal theory of complex organizations also suggest that a delicate balance between specialization and coordination is essential if large systems are to innovate and continue to evolve successfully. Excessive need for coordination prevents one unit from evolving without damaging the efficiency of others; excessive independence of components limits the sophistication of the products of the whole organization.

2.6 Social Structure

Although more nebulous than formal organizations, the framework of laws and customs of a whole society also provides an important object of identification for members, and in particular a stable cognitive framework within which a creature of bounded rationality can operate with some probability of predicting correctly the consequences of action. An important component of the structure of most societies is a system of markets and social practices that maintain them. Every modern industrial society is a composite system of organizations-cum-markets, where the markets permit a considerable degree of decentralization and independence of action of the individual organizations. The division of labor in the social sciences between organization theorists, on the one hand and economists on the other has left something of a lacuna in our understanding of the interactions between these two prominent kinds of social structure.

2.7 Mutual Expectations: Game Theory

What concept of rationality suits situations where several or many actors have competitive but highly interdependent goals? Cournot raised the problem in 1838 of two or more firms competing in an imperfectly competitive market, and threatening market instability by trying to outguess each other. Although von Neumann and Morgenstern (1944), in their *Theory of Games and Evolution*, proposed a solution to the problem, they defined rationality in a rather implausible way that did not guarantee unique choices or equilibria (Aumann and Sergiu 1992). Although their work has had a large impact, especially in economics and international relations theory, its most important result has been to show the near-impossibility of framing a univocal 'logical' definition of rationality in conditions of mutual uncertainty and interdependence. The ambiguities can sometimes be resolved by retreating from utility maximizing to bounded rationality, and by determining empirically which of their many options humans in fact employ in these situations (Axelrod 1980).

Political revolutions provide excellent demonstrations of the importance of stability of mutual expectations to the operation of a society. Revolutions dissolve expectations about what reactions actions will evoke from opposing groups. This 'gaseous' dissolution of social structure is ultimately reversed by the ascendance to power of new groups with internal identifications strong enough to remove the uncertainties (see also *Revolutions, Sociology of*).

3. Conclusion

Today, considerable effort is going into reexamining the concept of rationality, questioning the assumptions of utility maximization theory and seeking foundations for an empirically grounded theory of bounded rationality that gives proper attention to the social environment of choice. There is considerable new experimental work, particularly experiments with markets and games where 'outguessing' is important, but much more empirical study, in the field as well as the laboratory, will be needed to arrive at a theory that takes full account of the limits on human cognitive capabilities, and the social context which informs human reasoning.

Bibliography

- Anderson J R 1999 *Cognitive Psychology and its Implications*, 5th edn. WH Freeman, New York
 Aumann R J, Sergiu H (eds.) 1992 *Handbook of Game Theory*. North-Holland, Amsterdam
 Axelrod R 1980 *The Evolution of Cooperation*. Basic Books, New York
 Conlisk J 1996 Why bounded rationality? *Journal of Economic Literature* 34: 669–700

- Cournot A 1838 *Researches into the Mathematical Principles of the Theory of Wealth*. Reprinted 1960, Augustus M. Kelley, New York
 Green D P, Shapiro I 1994 *Pathologies of Rational Choice Theory*. Yale University Press, New Haven, CT
 Nelson R R, Winter S G 1982 *Evolutionary Theory of Economic Change*. Harvard University Press, Cambridge, MA
 Newell A, Simon H A 1962 *Human Problem Solving*. Prentice-Hall, Englewood Cliffs, NJ
 Parsons T 1937 *The Structure of Social Action*. McGraw-Hill, New York
 Rapoport A, Chammah A M 1965 *Prisoner's Dilemma*. University of Michigan Press, Ann Arbor, MI
 Shrager J, Langley P 1990 *Computer Models of Scientific Discovery*. Morgan Kaufman, San Mateo, CA
 Simon H A 1947 *Administrative Behavior*, 4th edn. The Free Press, New York
 Simon H A 1983 *Reason in Human Affairs*. Stanford University Press, Stanford, CA
 Simon H A 1992 What is an 'explanation' of behavior? *Psychological Science* 3: 150–161
 Simon H A 1997 *Models of Bounded Rationality Vol 3: Empirically Grounded Economic Reason*. MIT Press, Cambridge, MA
 Tversky A, Kahnemann D 1974 Judgment under uncertainty: heuristics and biases. *Science* 185: 1124–31
 von Neumann J, Morgenstern O 1944 *The Theory of Games and Economic Behavior*. Princeton University Press, Princeton, NJ

H. A. Simon

Reactance, Psychology of

The theory of psychological reactance (Brehm 1966) pertains to the reactive human tendency to protect or defend threatened freedoms. Unlike most conceptions about motivation, reactance theory does not speak of the human as arriving at a particular goal, achieving clarity in ambiguous stimulus fields, reducing conflict, or preserving self-worth. On the contrary, the reactance process is one of remaining open, of preserving one's freedom of movement, in the face of pressures to firm up or narrow the direction of movement.

1. Freedom: The Starting Point

The theory can best be introduced by considering single behavior realms in which freedom for the individual is conceivable. For instance, while shopping a person expects the freedom to choose from a vast array of breakfast cereals or brands of coffee. In more important matters, the person has a choice among numerous kinds of automobiles. In commerce with others, the person can potentially choose between being overtly friendly vs. taking a more distanced stance toward others. Even in such affective realms as emotions there is the potential—as in a setting that would normally set off sadness or grief—to react appropriately or to begin laughing.

In short, the theory addresses the freedoms associated with segments of human existence. These usually entail concrete behavioral scenarios in which a decision is eventually possible, or in which the person senses that several different solutions or directions can be chosen or tried out. In the context of the theory one does not speak of 'the free person' vs. 'the unfree person.' Freedom here is defined by the concrete endeavor of interest, such as finding oneself in a coffee bar and having to choose between espresso and normal coffee.

By no means does the theory assure that people automatically feel free across all such settings. On the contrary, Brehm (1966, 1968, 1993) postulates that, in each behavioral setting, the person will have a strong (or less strong, ranging down to zero) sense of freedom as a function of one of two factors. (a) Prior experience in the setting will be a guide as to expectations of freedom, as in, for instance, the case of a child who has always been entrusted with the freedom to select any of the breakfast cereals in the grocery store. This case may be contrasted with the child who has always been instructed to pick only the one brand of cereal that contains no sugar, preservatives, or coloring. (b) The other factor underlying the person's sense of being free, or less free, in a particular setting has to do with culturally-defined expectations. For example, pedestrians may (or may not) expect the freedom to cross against the red light, depending on local customs.

In practice, this all-important starting point for the theory—the person's expecting to be free or not—has been handled in an experimental manner. In some studies research subjects are told that they can expect to select freely from among several available consumer alternatives. For other participants, the expectation is created that they will receive one or more of the consumer items on a random basis, or at the whim of the person who controls the situation. People in this latter case thus find themselves expecting to gain something, such as a CD, pen, or article of clothing, but there is no person-freedom associated with their gaining that end.

With this illustration the reader can see that the theory is highly specific in its treatment of freedom, choice, and the free person. Setting-by-setting, individuals expect to be able to move about and choose freely, or not. One child expects to choose freely among cereals; another does not. Students in one culture might expect to choose which textbook to read for an exam; students in another culture may have no such expectation.

2. Reactance Arousal

Given that a person expects freedom of movement in a setting, infringements on that freedom will instigate a motivational force of reactance. The state of reactance, going into gear as a function of the strength of

threat to freedom and the importance of the freedom, brings the person to resist and counter the shrinking of the pertinent freedom. This is manifested in a direct behavioral fashion and also in terms of affective reactions.

Take the instance of the child who expects freedom of choice among breakfast cereals. If the child is standing before the array of breakfast products, pondering what to choose, reactance will come about by another person's picking a box of cornflakes from the top shelf and dropping it into the child's grocery basket. This unequivocal social influence will set up reactance in the child, resulting in the tendency to remove the cornflakes from the basket and to select a different variety. Further, the motivational state will show up in terms of affective changes; as an outcome of the social interference with freedom, the child will become relatively unattracted to the cornflakes and increasingly attracted to other salient, interesting brands.

3. Frustration

Reactance is not about attaining the goal that is top-ranked in the person's need or preference hierarchy. If the child had initially ranked Swiss muesli as No. 1, and cornflakes as No. 4, then the adult's placing the cornflakes in the child's grocery basket would have been directly frustrating. Perhaps obviously, the child would have reacted negatively to the frustration of a simple need to select the top-ranked item. However, the deciding quality of reactance theory is in its addressing freedom *per se*, and not the issue of the person's attaining the most-preferred goal object. The implications of this statement are far reaching; as long as the child's freedom to choose any of the considered alternatives is interfered with, reactance will be engendered. Quite aside from whether the muesli is No. 1 and the cornflakes No. 4, or vice versa, the adult's constraining the choice will produce reactance and the consequent behavioral/affective effects.

The point regarding frustration is similar when reactance is brought about in areas such as freedom of values, attitudes, or emotions. There is a sense in which it is frustrating to be pressured to take a certain opinion position, particularly one at variance with one's usual opinions. When a member of the American Civil Liberties Union is confronted with others who authoritatively argue in favor of restricting certain rights, the felt pressure to succumb to such pressure toward change is in a sense frustrating. But for reactance theory, the actual direction of such social pressure is not a terribly critical factor. If the member of the ACLU feels free to hold any of several positions along the human rights continuum, then this person will also experience reactance when a forceful communicator insists that the only correct opinion is a totally liberal one. Reactance, in the case of pressure

to change an opinion, value, or emotion, is manifested in terms of 'boomerang' effects, i.e. switching the opinion to a direction opposing the communication.

4. Areas of Application

Given the breadth of the theory, it seems reasonable that it should be applied to such diverse realms as choices among products, among people (e.g., choosing a partner), and to numerous social realms such as whether or not to help another person and whether or not to abide by another's advice or efforts to influence. The research evidence addressing these topics is largely experimental, and follows the formula given by the theory. First the investigator must know whether or not the person to be studied expects freedom of choice in the given context, and second, the strength of the threat to that freedom is varied. Within these general methodological guidelines, research participants have been shown to resist persuasive attempts, refuse to help, and frequently to upgrade their attraction to potential choice objects or people whose availability has been threatened. Reviews of these many effects can be found in Brehm (1966, 1968, 1993), Brehm and Brehm (1981), Wicklund (1974, 1997), and Wortman and Brehm (1975).

5. The Longer-range Dynamics

These dynamics can best be depicted in three phases, with the assistance of an expanded theory by Wortman and Brehm (1975). The first two phases, as described above, are (a) the instigation of an expectation of acting freely in a domain, and (b) infringements on that freedom, producing reactance. The issue raised by Wortman and Brehm—implying a third phase—has to do with the effect of repeated restrictions on freedom. A person may be accustomed to choosing freely between espresso and normal coffee in a certain coffee bar. When the management announces one day that normal coffee is not available, the reactance-produced response is one of increased attraction to normal coffee and a corresponding downgrading of espresso. But what transpires upon repeated, continued instances of this restriction? Will the person continue to express reactance on the second, third, fourth, and fifth days of the freedom restriction?

In alternative theoretical language, such repeated restrictions are called helplessness inducing (Seligman 1975), and in the Wortman and Brehm model, the idea is that repeated violation of the freedom will eventually bring the person to no longer expect freedom of action within that domain, as in the particular coffee shop. The expectation of free choice is altered into the expectation of espresso, and nothing more. This adaptation to the repeated restriction is thus the third phase of the longer-term reactance process, and as

such, describes the ebb and flow of reactance in numerous situations over the course of a child's development, as well as the ebb and flow of reactance in diverse areas as a result of cultural change.

6. Individual Differences

Brehm (1993) differentiates sharply between the chronic need to expand areas of freedom and control, and the defense of a specific area of freedom or control. The former, normally associated with such concepts as control need or competence striving, addresses a human who is, across time and place, chronically interested in enlarging the scope of freedom, thus in maximizing personal control in a general sense. Such a chronic personal striving can be operationalized in terms of certain individual differences, such as Rotter's Internal/external control scale and, to be sure, researchers have introduced these very individual differences into a variety of reactance paradigms. The research suggests that those with a high chronic need for control appear to evidence stronger responses in reactance-creating situations (see Brehm and Brehm 1981).

On the other hand, there is little theoretical reason to expect that general human tendencies will be pertinent to the reactance process as described in detail in the numerous experiments. This is because the question of whether a person will evidence the counter-behaviors, the boomerang effect attitude changes, and other freedom-restoring maneuvers, is a highly concrete issue. The parameters of each behavioral scenario, combined with the individual's expectation of freedom within those confines, determines the momentary predilection to protect and regain freedom.

7. Future Directions

A salient direction to take in future development of the theory involves the social circumstances which give rise to expectations of freedom. While the original theory refers to such expectations as stemming from (a) direct, personal experience and (b) societal conventions regarding specific classes of actions, there is a further possibility: a person who is raised in a community in which a variety of points of view are acknowledged will stand to gain a sense of freedom to express those many points of view. This should apply no matter whether they are expressed in discussions of opinions and values, or even in solutions to problems. In contrast, a person coming from an environment in which there is 'one correct opinion,' or solution, will be less likely to incorporate the idea that multiple opinions or solutions are possible or desirable (Wicklund 1999). Accordingly, we may be able to speak of a person's more general sense of freedom of opinions and values, as long as the person's socializ-

ation experiences are in a setting that fosters such multiple values systems, or even multiple personality traits. The direction for corresponding research lies both in developmental psychology and in the detailed descriptions of cultural settings allowed by cultural anthropology.

See also: Control Behavior: Psychological Perspectives; Control: Social; Freedom/Liberty: Impact on the Social Sciences

Bibliography

- Brehm J W 1966 *A Theory of Psychological Reactance*. Academic Press, New York
- Brehm J W 1968 Attitude change from threat to attitudinal freedom. In: Greenwald A G, Brock T C, Ostrom T M (eds.) *Psychological Foundations of Attitudes*. Academic Press, New York
- Brehm J W 1993 Control, its loss, and psychological reactance. In: Weary G, Gleicher F, Marsh K L (eds.) *Control Motivation and Social Cognition*. Springer, New York, pp. 3–30
- Brehm S S, Brehm J W 1981 *Psychological Reactance: A Theory of Freedom and Control*. Academic Press, New York
- Seligman M E P 1975 *Helplessness: On Depression, Development, and Death*. Freeman, San Francisco
- Wicklund R A 1974 *Freedom and Reactance*. Erlbaum, Hillsdale, NJ
- Wicklund R A 1997 The theory of psychological reactance: a reliably revolutionary perspective. *Contemporary Psychology* 42: 679–81
- Wicklund R A 1999 Multiple perspectives in person perception. *Theory and Psychology* 9: 667–78
- Wortman C B, Brehm J W 1975 Responses to uncontrollable outcomes: An integration of reactance theory and the learned helplessness model. In: Berkowitz L (ed.) *Advances in Experimental Social Psychology*. Academic Press, New York, Vol. 8, pp. 277–336

R. A. Wicklund

Readability, Applied Psychology of

Almost since the beginning of research into the process of text comprehension, researchers and educators have attempted to find accurate, meaningful, and simple methods to assess readability. Early researchers such as Quantz, Cattell, Huey, and Dearborn—most of their research was performed around the turn of the twentieth century—focused on differences among readers' abilities, offering what might now be considered 'process models,' albeit simple ones. The factors they studied, however, were more perceptual in nature, such as the effects of saccadic eye movements and eye-voice spans in reading aloud. Despite some important insights, their work waned in influence during the period from 1910 until 1970, or so. Not

coincidentally, this decline in the interest of process models mirrored the rise in influence of behaviorism in American psychology. Behaviorists tended to focus on observable, easily measured factors. Reading researchers directed their attention to properties of the texts, rather than how a text might induce changes in the processing of the text by the reader.

During the middle portion of the twentieth century, the first readability formulas were developed. Consistent with the underlying philosophy of behaviorism, these formulas considered readability to be a property of the stimulus—that is, the text itself. Most of the commonly used readability formulas, such as the Gunning Fog index and the Flesch Readability scales, were developed during this time.

The so-called cognitive revolution of the 1960s renewed interest in natural language. In the last 25 years, especially, the psychological process of language comprehension has received considerable attention. A number of wide-reaching theories of text comprehension have been developed (Gernsbacher 1990, Just and Carpenter 1980, Kintsch 1988, Kintsch and van Dijk 1978). These models do not assume the text to be a static entity; instead, they emphasize the process of reading, or more specifically, the interaction between the text and the reader. Ironically, contemporary researchers have essentially returned to the same philosophical positions that drove researchers such as Huey and Dearborn a century before.

1. Early Readability Research

The challenge facing educational psychologists in the 1930s and 1940s was rather pragmatic, and is still one of the goals of contemporary researchers and educators: to evaluate and compare readability of texts used in a variety of school settings. In theory, this would allow teachers to select age- (or ability-) appropriate materials for use in the classroom. In addition, educators were looking for a way to assess the readability of various books students might read on their own.

Understandably, these formulas tended to follow from behaviorally oriented research: identify the objective (although generally superficial), easily measured properties in the stimuli—the texts. Early work tended to focus primarily on such simple ideas as vocabulary difficulty, word frequency, and sentence length (e.g., Gray and Leary 1935). It is important to note that the question of *why* these variables were related to readability was considered largely irrelevant. An understanding that sentence length was important because it provided an estimate of the demands placed on working memory was decades away.

Within a few years, readability formulas were developed that could be applied virtually to any text (Dale and Chall 1948a, 1948b, Flesch 1948, Klare 1963, 1974/5, 1984, Vogel and Washburne 1928).

Though each of these differed in some respect, in truth they were remarkably similar. Most of these scales used some combination of word length (presumably related to word frequency or word difficulty), number of sentences per paragraph (a crude measure of complexity), sentence length, and/or number of syllables per word. They soon became widely adopted in academic, educational, and other applied settings. Standardized testing procedures also made wide use of these formulas. Some governmental agencies even specified readability levels for things such as insurance policies; in order to be 'approved,' for example, an insurance policy must have a readability grade level of 9 or below.

Despite the sophistication of contemporary reading research, most of the readability formulas in wide use today continue to use the same key variables. For example, recent versions of Microsoft Word provide an assessment of readability using the Flesch Reading Ease scale, as well as a normed score using the Flesch-Kincaid Grade Level scale (which scales readability relative to US public school grade levels). Cassady and Greene's program *Grammarians* also provides an assessment of readability using the Gunning Fog index. (The Gunning Fog index uses the number of 'long words': those of three or more syllables.) These formulas are somewhat successful, at least to a point, and provide a reference from which texts can be compared.

Perhaps such widespread use of these simple formulas is not warranted. After all, these formulas analyze only a few variables, and cases where the formulas fail are abundant. However, the limitations of the formulas (the fact that they analyzed only a few components) were also their most attractive feature—they were easy to use. Thus, these criticisms, even when recognized by the researchers, have not attenuated the popularity of readability formulas.

Do revisions intended to improve the readability of a text actually produce more easily understood texts? One of the most thorough investigations in rewriting and readability was performed by Britton and Gulgoz (1991). They used a text of roughly a thousand words long, detailing the success and failure of the air bombing 'Operation Rolling Thunder' offensive in the Vietnam war (Greene 1985). They compared the original text (which had a grade level equivalent of about 13.5 years, as measured by the Flesch-Kincaid Index) to one of three revisions: a readability revision, which used shorter sentences and more frequently used words, improving the readability level to 12 years; a principle revision, which modified the text according to the comprehension model of Kintsch and van Dijk (1978) and had a readability level the same as the original text (13.5 years); and a heuristic revision, performed by an expert in text revision (Britton himself), designed to make the text as understandable as possible. The reading level of the latter version was lowered to grade 12. The only revision which actually

improved comprehension was this final revision. Lowering the readability score itself produced no significant improvement—simply using shorter sentences and more frequent words was not sufficient to make a text more understandable.

Improving readability has one other benefit—texts written at a level consistent with a reader's ability are monitored more accurately while reading. Simply put, comprehension monitoring refers to the accuracy with which individuals can assess their comprehension. After reading a passage, individuals are asked to predict later performance on a test. The correlation between predicted and actual performance provides a measure of comprehension monitoring accuracy.

Weaver and Bryant (1995) examined the effect of readability on comprehension monitoring. They had college students read texts that were either easy (reading level below grade 8), standard (reading level around grade 12), or difficult (reading level around grade 16). Comprehension performance was related to readability: recognition performance increased as reading material became easier (more readable). Comprehension monitoring, in contrast, showed an inverted U-shaped function. The greatest metamemory accuracy occurred not for the easiest materials, but for those materials best matching the readers' ability (standard readability, approximately 12th grade).

2. Summary

The theoretical case against readability formulas has been argued many times (e.g., from the present perspective by Kintsch and Vipond 1979). Their merits are clear, and center on their ease of use. But the drawbacks are equally clear. At best these formulas provide an oversimplification, too often a serious distortion. Nevertheless, they continue to be used, and will be as long as comprehension research does not provide viable alternatives.

Great progress into the reading process has been made in the past 30 years. But why have these advances not led to the development of viable alternatives to readability formulas?

The answer to this question has three parts. First of all, most of the research discussed above has dealt with the construction of internal representations, often called textbases. A great deal is now known about how to construct comprehensible, recallable, and summarizable texts for given populations. However, while this knowledge is quite adequate for laboratory research on discourse processing, simple, easy-to-use procedures for practical applications have not yet been developed.

The second part of the answer involves a fundamental change in the nature of readability. Readability has typically been considered as a property of the text alone. Clearly, true readability is more

complicated than that and is probably best considered as an interaction between the text itself and the readers' *ability* and *knowledge*. Unfortunately, a simple readability formula incorporating readers' abilities has yet to be developed.

The final part of the answer is purely pragmatic. When using a text in an experiment—often a relatively brief text—it does not matter whether the procedures for text analysis are time-consuming, require expertise, and so on. For the applied settings in which most readability formulas are used, though, ease of use matters a great deal. What is needed is a large-scale effort directed toward the development of practical methods of text analysis that allow non-experts to do with relatively little effort what only experts can do today with considerable effort. While such an enterprise would be nontrivial, it is certainly feasible. For the moment, however, readability measures have little connection to the cognitive science research on text comprehension.

In light of this fact, Bruce et al. (1981; see also Weaver and Kintsch 1991) have proposed that readability formulas be used only when the following criteria are met:

(a) The material will be freely read, and reading time will not be determined by 'external factors' such as experimental control or (in their example) captions that are presented along with television programs (for the hearing impaired).

(b) The text is 'honestly written': that is, not directly written in such a way to satisfy the criteria of readability formulas.

(c) The higher-level features of a text, such as organization and intention, correlate with the sentence and word features. As current formulas make little use of the top-level information, such material is otherwise ignored.

(d) The purpose of the reading is similar to that of the readers in the validation studies. If the two groups are reading with different strategies, then the formula may not be valid across the situations.

(e) The formulas are applied across a large number of texts, and individual texts are not singled out. Bruce et al. (1981) emphasize that readability scores are composite averages and may not apply to specific texts.

(f) The population of readers is the same in the 'real-world' setting as it was in the validation studies. If a text is to be used as a remedial text, it should not be validated on a group of average readers.

See also: Comprehension, Cognitive Psychology of; Eye Movements in Reading; Inferences in Discourse, Psychology of; Knowledge Activation in Text Comprehension and Problem Solving, Psychology of; Learning from Text; Literary Texts: Comprehension and Memory; Macrostructure in Discourse Comprehension, Psychology of; Memory for Text; Narrative Comprehension, Psychology of; Reading,

Neural Basis of; Reading Skills; Sentence Comprehension, Psychology of; Speed Reading; Text Comprehension: Models in Psychology; Textbooks

Bibliography

- Britton B K, Gulgoz S 1991 Using Kintsch's computational model to improve instructional text: Effects of repairing inference calls on recall and cognitive structures. *Journal of Educational Psychology* **83**: 329–45
- Bruce B, Rubin A, Starr K 1981 *Why Readability Formulas Fail*. Center for the Study of Reading, Champaign, IL
- Dale E, Chall J S 1948a A formula for predicting readability. *Educational Research Bulletin* **27**(January): 11–12
- Dale E, Chall J S 1948b A formula for predicting readability. *Educational Research Bulletin* **27**(February): 37–54
- Flesch R F 1948 A new readability yardstick. *Journal of Applied Psychology* **32**: 221–33
- Gernsbacher M A 1990 *Language Comprehension as Structure Building*. Erlbaum, Hillsdale, NJ
- Gray W S, Leary B E 1935 *What Makes a Book Readable?* University of Chicago Press, Chicago
- Greene M P 1985 *US Air Power: Key to Deterrence*. Maxwell Air Force Base
- Just M A, Carpenter P A 1980 A theory of reading: From eye fixations to comprehension. *Psychological Review* **87**: 329–54
- Kintsch W 1988 The role of knowledge in discourse comprehension: A constructive-integration model. *Psychological Review* **95**: 163–82
- Kintsch W, van Dijk T A 1978 Toward a model of text comprehension and production. *Psychological Review* **85**: 363–94
- Kintsch W, Vipond D 1979 Reading comprehension and readability in educational practice and psychological theory. In: Nilsson L G (ed.) *Perspectives on Memory Research*. L. Erlbaum Associates, Hillsdale, NJ, pp. 329–65
- Klare G R 1963 *The Measurement of Readability*. Iowa State University Press, Ames, IA
- Klare G R 1974/5 Assessing readability. *Reading Research Quarterly* **10**: 62–102
- Klare G R 1976 A second look at the validity of readability formulas. *Journal of Reading Behavior* **8**: 129–52
- Klare G R 1984 Readability. In: Pearson P D, Barr R, Kamil M L, Mosenthal P (eds.) *Handbook of Reading Research. Vol. I*. Longman, New York, pp. 681–744
- Vogel M, Washburne C 1928 An objective method of determining grade placement of children's reading material. *Elementary School Journal* **28**
- Weaver C A III, Bryant D S 1995 Monitoring of comprehension: The role of text difficulty in metamemory for narrative and expository text. *Memory & Cognition* **23**: 12–22
- Weaver C A, III, Kintsch W 1991 Expository text. In: Barr R, Kamil M, Mosenthal P, Pearson P D (eds.) *The Handbook of Reading Research*. Longman, White Plains, NY, Vol. 2, pp. 230–45

C. A. Weaver, III and A. Renken

Reading, Neural Basis of

Psychologists and brain researchers who investigate reading usually focus on the basic cognitive processes

and brain regions that are necessary to recognize and pronounce a printed word. More complex aspects of reading, such as the ability to understand the grammatical roles played by each word in a written or spoken sentence, typically are assumed to rely upon general language functions that are not specific to reading (see *Syntactic Aspects of Language, Neural Basis of*). This chapter will focus on three components of skilled reading, and the brain regions that support them: a) the representation of visual word forms (orthography), b) the use of orthographic information to access information about word sounds (phonology), and c) the use of orthographic information to access information about meaning (semantics).

Claims about the neural bases of component processes associated with reading are based upon three primary methods. Studies of subjects with brain damage (neuropsychological studies) are used to understand how different aspects of reading performance are impaired and preserved following brain injury or degeneration (see *Dyslexia (Acquired) and Aphasia*). Studies of brain blood-flow changes (functional neuroimaging studies) use either PET (Positron emission tomography) or fMRI (functional magnetic resonance imaging) to localize and understand the functional contributions of brain regions associated with reading. ERP (see *Event-related/ Evoked Potentials*) and MEG (see *Magnetoencephalography*) studies measure brain electromagnetic changes, and provide information about the temporal sequence of brain processing during reading, as well as some information about the location of brain regions associated with reading.

1. An Overview of the Brain Regions Involved in Reading

An overview of the brain regions that support reading can be obtained by using neuroimaging to compare the pattern of brain activation associated with reading aloud a single word, versus merely looking at a small cross or dot. Some of the areas identified in this task comparison support relatively general sensory and motor functions. For instance, the primary motor cortex, which is located in the precentral gyrus (Fig. 1), is usually active during any task that incorporates movement. The identification of the motor cortex in the reading versus fixation comparison reflects the fact that a task that involves mouth movements is being compared to a task that does not. One approach for identifying regions that are more specifically involved in reading is to compare reading aloud to a control task that has similar sensory and motor components, but which minimizes orthographic, phonological, or semantic processing. For example, reading aloud has been compared to a task in which subjects say a simple utterance in response to a random string of letters. This type of contrast subtracts out basic visual,

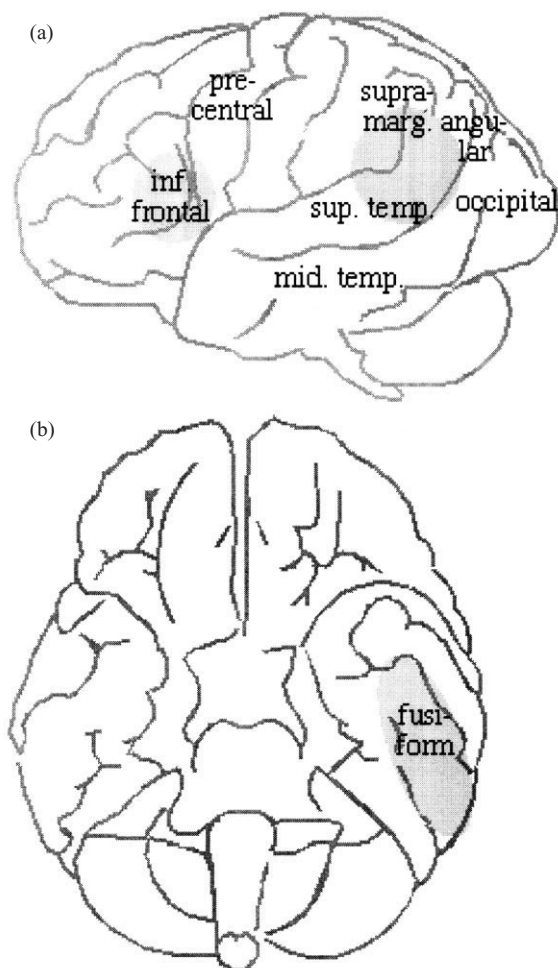


Figure 1

Brain areas associated with reading. Brain regions involved in higher-level aspects of reading can be identified through comparisons of reading to simpler tasks that contain similar sensory and motor components. Three broad areas identified through such task comparisons have attracted a significant amount of attention. A left inferior frontal area (1a) is centered on the inferior frontal gyrus. A left temporoparietal area (1a) is centered on the junction of the posterior superior temporal cortex and the supramarginal and angular gyri of the inferior parietal lobe. A left basal temporal area is centered within the fusiform gyrus, which lies between temporal and occipital cortex. This area is best viewed from the underside of the brain (Fig. 1b)

auditory, and motor areas, and reveals a more restricted set of regions that support higher-order aspects of reading (Fiez and Petersen 1998).

The exact number and location of the brain regions that support higher-order aspects of reading is a matter

of active investigation, but three broad areas have attracted the most attention (Fiez and Petersen 1998; Price 1998). As illustrated in Fig. 1, these areas are centered in the left inferior frontal cortex, the left temporoparietal cortex, and the left basal temporal cortex (primarily the fusiform gyrus). Most of the remainder of this article will focus on the potential relationships between these brain areas and the basic components of skilled reading outlined in the introduction. It is important to note, as will become clearer in the discussion of the basal temporal area, that a more precise mapping of the brain regions that contribute to reading is still needed. It is likely that each of the broad areas focused on in this article contains multiple functionally distinct subregions. Furthermore, it is likely that other areas that are not discussed, such as the left middle temporal gyrus (Fig. 1), make some contribution to reading.

2. *The Representation of Orthography*

The organization of the visual system has been a topic of longstanding interest in the neuroscience and cognitive neuroscience research communities. One question that is highly relevant for studies of reading is whether different regions of the visual object recognition system (located in the ventral occipital and temporal cortex) are specialized for processing different types of stimuli, such as faces or text. In the domain of reading, the search for a region that selectively processes text has generally involved two types of comparisons. The first has been comparisons of words and nameable pictures. This allows regions that are specifically involved in orthographic processing to be teased apart from regions that are more generally involved in the naming of meaningful stimuli. The second approach has been comparisons between different types of orthographic stimuli. Most often, words and pseudowords (pronounceable nonwords, such as 'floop') have been compared to nonword stimuli that are visually similar to words (e.g., stimuli composed of artificial letters, or 'false fonts,' or random consonant strings). This allows regions that are specifically involved in the representation of words to be dissociated from regions that are involved in lower-level processes, such as letter recognition and the detection of line segments.

Several regions that may mediate orthographically specific processing have been proposed. The most promising 'word form area' is in the left middle fusiform gyrus, near the border between the temporal and occipital cortex. Activation in the fusiform gyrus is found reliably when subjects read words aloud (Fig. 1), and one reason to suspect that it may be involved in orthographic processing is the fact that it lies near areas associated with simple visual processing and visual object recognition. More direct evidence comes from studies that have found responses in the fusiform

to words and pseudowords, but not visually similar nonword stimuli, such as random letter strings (Fiez and Petersen 1998). Evidence that this area is influenced by the degree to which an orthographic stimulus functions as a whole word (lexical) unit comes from studies in Japanese readers. The Japanese writing system contains two types of characters: Kana characters represent syllables, while Kanji characters represent entire words. Significant activation in the fusiform is found when subjects read words aloud written with Kanji versus Kana characters, and when subjects are given a task that requires them to access the orthographic form of a Kanji character, such as writing the Kanji equivalent of a word written in Kana (Nakamura et al. 2000, Sakurai et al. 2000).

Data from neuropsychology also point toward the important role of occipitotemporal areas, such as the fusiform gyrus, in orthographic processing. Subjects with brain damage in this area can exhibit the syndrome of pure alexia (Patterson and Lambon-Ralph 1999). In pure alexia, the ability rapidly to access or represent the orthographic form of an entire word, or the ability to use orthographic information to access phonological and semantic information, is lost. As a result, subjects use a laborious letter-by-letter strategy to identify and read aloud a word. Subjects with pure alexia generally are able to perform simple visual perceptual tasks, recognize and name visual objects, and write words normally. Japanese subjects can also exhibit alexia following damage to the fusiform, but long-lasting impairments are typically limited to the reading and writing of Kanji characters (Iwata 1984).

Results from ERP and MEG studies indicate that there are differences in the electromagnetic current evoked by words and wordlike stimuli, as compared to nonword stimuli such as letter strings (Cohen et al. 2000; Tarkiainen et al. 2000). The temporal information gained from these studies indicates that orthographically specific processing at the word level begins 100–200 milliseconds (ms) after the onset of a stimulus. Attempts to localize the source of the waveform differences have generally implicated the left basal temporal cortex. More specific localization has come from a study in which recordings were done using an electrode placed in the fusiform gyrus of patients prior to brain surgery. This study found neurons that selectively responded to orthographic stimuli (Nobre et al. 1994).

While there is a significant amount of accumulated evidence across multiple methodologies that points towards a role for the fusiform gyrus in orthographic processing, the precise function of this area remains a point of debate. One issue is whether domain-specific regions—'face areas' or 'word form areas'—exist in the visual system. Since reading is a skill acquired very recently on the evolutionary time-scale, it is unlikely that any brain region would be innately specified to represent word forms. Rather, any specialization is

most likely to occur as a result of extensive experience with a relatively unique class of visual stimuli. An alternative point of view is that the analysis of orthography places heavy demands on particular types of visual analysis, such as the ability to resolve input with high-frequency spatial information, or to represent a set of distinct objects that are perceptually very similar. Portions of the fusiform may be particularly suited for these types of analysis (either innately or through experience), and thus may be used to a greater extent for processing orthographic versus non-orthographic stimuli.

A second issue is the level of information that may be represented in the fusiform. Recent evidence indicates that letter-by-letter reading is typically accompanied by visual-perceptual difficulties below the level of the whole word. Deficits in lower-level processing could impair the access to or generation of higher-level orthographic representations (Behrmann et al. 1998). Conversely, imaging evidence reviewed more extensively below indicates that portions of the fusiform are active during linguistic tasks that are not thought to require access to orthographic information. Activation associated with orthographic processing may thus reflect more abstract lexical (whole word) or semantic information that can be accessed through either printed words or speech.

3. *The Transformation Between Orthography and Phonology*

In order to read a word aloud, it is essential that information about the printed form of the word is transformed into information about the corresponding sound, or pronunciation, of the word. Two general brain regions have been identified which appear to play a major role in phonological (word sound) representation and processing. One region is located in the left inferior frontal lobe, and the other is in the temporoparietal cortex.

Much of the neuropsychological evidence that these regions are important for reading stems from studies of patients with damage to either or both of these regions. These patients often exhibit the syndrome of phonological dyslexia, which is characterized by great difficulty reading aloud pseudowords (Fiez and Petersen 1998; Patterson and Lambon-Ralph 1999). The ability to read aloud pseudowords is considered to be a marker of phonological processing, because the pronunciation of pseudowords cannot be specified by stored lexical (whole word) or semantic knowledge. Instead, pronunciation is thought to rely upon some type of sublexical procedure for deriving a phonological representation. Subjects with phonological dyslexia can read real words relatively well. This is thought to reflect their ability to take advantage of stored lexical and semantic information using brain regions that are not damaged. Interestingly, when

direct electrical stimulation is applied to the temporoparietal region during a reading task, subjects appear to exhibit a temporary phonological dyslexia: their ability to pronounce nonwords is temporarily disrupted, but their reading of real words is intact (Simos et al. 2000; see also *Electrical Stimulation of the Brain*).

Conclusions based upon neuropsychological data are supported by converging evidence from neuroimaging. Both the left frontal and temporoparietal regions are reliably active during normal word reading (Fig. 1), and during a wide variety of tasks that encourage phonological processing (Demonet et al. 1996). ERP and MEG evidence indicates that the activation of these areas occurs at least 300ms after the onset of a stimulus, which is after the temporal differences observed in the basal temporal cortex (Simos et al. 2000). The importance of these regions in reading is further demonstrated by comparisons of normal adults to adults who experienced difficulties learning to read as children (developmental dyslexics). Many investigators believe that subjects with developmental dyslexia have a core deficit in phonological processing that persists into adulthood. When subjects with developmental dyslexia read words aloud, neuroimaging and MEG studies have revealed abnormally low levels of activity in the temporoparietal area (Shaywitz et al. 1998, Rumsey et al. 1999, Simos et al. 2000). The relative reductions in temporoparietal activation can be paralleled by above-normal levels of activity in the left frontal region. The left frontal increases may reflect use of a compensatory strategy for reading that normally plays a secondary role in pronunciation for skilled readers (except under more challenging conditions, such as when a stimulus is a nonword).

Several broad schemes that might distinguish the contributions of the frontal and temporoparietal regions to reading and phonological processing exist. One potential distinction is whether the phonological information and processes supported by a particular brain region are related to the acoustics of speech sounds (input phonology), or the articulation of speech sounds (output phonology). The proximity of the temporoparietal region to basic auditory areas, coupled with the fact that lesions to this area produce deficits in speech comprehension, indicates that it may support input phonology. Conversely, the proximity of the left frontal region to general motor-related areas, coupled with the fact that lesions to this area produce deficits in speech production, indicates that it may support output phonology.

A second potential distinction comes from broad differences between the roles of the frontal and posterior brain regions in cognition. Frontal regions are recruited most often by tasks that require conscious and effortful processing, or controlled processing. When a task is well learned and relatively automatic, activation in the frontal cortex is generally

minimal. The frontal cortex may thus represent an area used for controlled phonological processing, whereas the temporoparietal region may represent an area used for automatic phonological processing. This could explain why subjects with developmental dyslexia, who do not develop the normal level of reading skills and automaticity, show less activation in the temporoparietal area and more in more frontal areas. It is also possible that different subregions within the inferior frontal cortex are responsible for different aspects of speech production (e.g., output phonology, grammatical analysis, etc.).

A third potential distinction comes from theoretical models of reading, which have outlined different types of mechanisms that could support orthographic to phonological transformation. Two broad classes of models have been proposed, dual-route (Coltheart et al. 1993) and connectionist models of reading (Plaut et al. 1996). In both types of models, access to the pronunciation of a word is mediated by two different 'routes,' though the routes function in different ways. In the dual-route models, a lexical route functions to associate orthographic information with stored representational information about the pronunciation of the whole word. A sublexical route functions to break a word or pseudoword into smaller parts, such as individual letters or letter clusters (graphemes), that can be associated with corresponding sounds (phonemes); the sounds can then be put together to yield a pronunciation for the entire item.

In connectionist models, direct mappings between orthography and phonology are used to capture statistical regularities between orthography and phonology at both the lexical (or whole word) and the sublexical (e.g., the 'b' in 'back' and 'bear' has the same sound) levels. The pronunciation of a written stimulus is represented as a distributed pattern of activity across a set of interconnected phonological units. Indirect mappings between orthography and phonology via semantics can also be used to support the pronunciation of a word. That is, you could represent the meaning of the word 'dog,' and also represent the fact that this meaning is associated with the pronunciation/dog/. As in the direct mappings between orthography and phonology, these mappings are represented as a distributed pattern of activity, but unlike the direct mappings the mappings should be relatively insensitive to sublexical information (since there is little relationship between parts of words and word meanings, at least for simple 'mono-morphemic' words like 'dog' and 'table').

Simple assumptions have been used to suggest potential mappings between theoretical components in models of reading and specific brain regions (Fiez and Petersen 1998, Price 1998). For instance, it has been suggested that the left frontal area contributes to a sublexical phonological process. This mapping is based upon the fact that in models of reading the pronunciation of pseudowords is primarily based

upon output from a sublexical phonological procedure, and in the brain greater activation is seen in the left frontal area when subjects read pseudowords as compared to words. Even this potential correspondence, however, falls short of identifying whether the left frontal region functions according to dual-route or connectionist principles (or some alternative process), and it fails to acknowledge puzzling differences between words and pseudowords in other brain regions. In order to reach more complete levels of description, it will be essential to develop more rigorous methods for relating computational modeling and neuroimaging results.

4. *The Transformation Between Orthography and Lexical Semantics*

In addition to lexical procedures that facilitate direct associations between an orthographic and a phonological form of a word, it is also possible that other types of stored representational information can be used to support pronunciation. This could include lexical information about the word itself, the meaning of the word, or the concept denoted by the word. Once such lexical-semantic information has been accessed, it could be used to access information about the pronunciation of the word or concept, as suggested in connectionist models of reading. Importantly, access to most lexical-semantic information is expected to be modality and task independent. Either a written word or a picture, for instance, can be used to represent a word name or a semantic concept, and individuals can produce a spoken word in response to a written word, picture, or their own internal thoughts.

Of the three general areas that are commonly active during word reading (Fig. 1), the basal temporal region is the best candidate for an area that contributes to lexical-semantic processing. This is the same general region that has been implicated in orthographic representation, and the potential overlap between orthographic and lexical-semantic processing is discussed at the end of this section. Assertions that the basal temporal region contributes to lexical-semantic processing are based upon evidence that the region responds to a variety of stimuli and task conditions that emphasize the retrieval of word names or concepts. For instance, neuroimaging studies have found overlapping activation during picture naming and word naming tasks (Moore and Price 1999). Outside the domain of reading, basal temporal activation is commonly associated with tasks that emphasize semantic processing or word retrieval, such as deciding whether a pair of words belongs to the same semantic category (Price 1998; see also *Semantic Knowledge: Neural Basis of*).

Neuropsychological studies also point towards the role of the inferior temporal area in lexical-semantic processing. With respect to reading, the most critical

information comes from studies of patients with acquired surface dyslexia. In the purest forms of surface dyslexia, subjects tend to mispronounce inconsistent words, which are words with unusual spelling-to-sound correspondences. For instance, the word 'pint' is not pronounced like 'mint,' 'lint,' or 'hint.' Theoretical models of reading have postulated that the pronunciation of inconsistent words, especially those that occur infrequently in print, is heavily dependent upon access to stored lexical-semantic information. Most reported subjects with surface dyslexia have large regions of damage or degeneration affecting the left middle and inferior temporal lobe, and the deficits of these subjects typically extend beyond the domain of reading (Patterson and Lambon-Ralph 1999). For instance, they often have difficulty accurately naming pictures of objects. Surface dyslexics are generally able to read aloud most words with consistent spelling-to-sound correspondences, and pseudowords. This preserved ability may reflect the use of sublexical orthographic to phonological procedures that are supported by intact brain regions, such as the left frontal area.

An issue that is not well understood is whether the reading-related area discussed in Sect. 2 is the same as the lexical-semantic area discussed in this section. One reason to believe that they may be the same is the fact that in many neuroimaging studies of reading, only a single region of activation is identified in the fusiform gyrus. If they do overlap, future research will need to reconcile the orthographic and lexical-semantic points of view to provide a unified interpretation of the role of the inferior temporal area in reading. However, there is some evidence that there may be more than one functional region within the basal temporal area. A neuroimaging study of word versus picture naming, for example, identified multiple regions within the fusiform that had different patterns of activation (Moore and Price 1999). This is consistent with converging data from animal and human studies of visual object recognition. Such studies indicate that modality and domain specific visual processing, such as that involved in orthographic representation of lexical information, may be located in the posterior fusiform. In contrast, the anterior portions of the fusiform may be associated with increasingly conceptual and modality-independent processing. This could include the conceptual representation of objects and ideas that are denoted by pictures, and by written and spoken words (Ricci et al. 1999).

5. The 'Tuning' of Brain Regions that Support Reading

Neuropsychological and neuroimaging studies of Chinese, Japanese, Italian, German, and English-speaking subjects have reached similar conclusions

about the basic cognitive processes and brain regions involved in reading. This consistency reinforces the conclusion that the network of brain areas summarized in Fig. 1 is somehow specialized to support reading and other language processes. However, it should not be taken as evidence that the neural network that underlies reading is inflexible. Quite the contrary, there is growing evidence that a number of factors can have a profound impact on how the network is tuned to most effectively support reading.

Some of the tuning of the reading system may reflect differences in experience or training. For instance, in Italian the mapping between how words look and how they sound is much more consistent than it is in English. Italian-speaking subjects take advantage of this consistency to read aloud words and pseudowords more quickly than English-speaking subjects. Furthermore, when they read words aloud, they produce more brain activation in the left temporoparietal region than do English-speaking subjects; the converse is true in the left inferior temporal region (Paulesu et al. 2000).

Other differences may reflect the use of compensatory or alternative strategies. For instance, as noted above, subjects with developmental dyslexia show a hyperactivation of the left frontal region, and a hypoactivation of the left temporoparietal region. More subtle differences in the activation of brain areas associated with phonology are even apparent in readers at different ends of the normal distribution of reading skill (Rumsey et al. 1999). In subjects with acquired brain damage, neuroimaging studies indicate that right hemisphere brain regions, typically similar in location to sites of damage in the left hemisphere, may begin to play a more substantial role in language tasks (Buckner et al. 1996).

See also: Lexical Processes (Word Knowledge): Psychological and Neural Aspects

Bibliography

- Behrmann M, Plaut D C, Nelson J 1998 A literature review and new data supporting an interactive account of postlexical effects in letter-by-letter reading. *Cognitive Neuropsychology* **15**: 7–52
- Buckner R L, Corbetta M, Schatz J, Raichle M E, Petersen S E 1996 Preserved speech abilities and compensation following prefrontal damage. *Proceedings of the National Academy of Science USA* **93**: 1249–53
- Cohen L, Dehaene S, Naccache L, Lehericy S, Dehaene-Lamberg G, Henaff M A, Michel F 2000 The visual word form area. Spatial and temporal characterization of an initial stage of reading in normal subjects and posterior split-brain patients. *Brain* **123**: 291–307
- Coltheart M, Curtis B, Atkins P, Haller M 1993 Models of reading aloud: Dual-route and parallel-distributed-processing approaches. *Psychological Review* **100**: 589–608
- Demonet J F, Fiez J A, Paulesu E, Petersen S E, Zatorre R 1996 PET studies of phonological processing: A critical reply to Poeppel. *Brain and Language* **55**: 352–79

- Fiez J A, Petersen S E 1998 Neuroimaging studies of word reading. *Proceedings of the National Academy of Science USA* **95**: 914–21
- Iwata M 1984 Kanji versus Kana: Neuropsychological correlates of the Japanese writing system. *Trends in Neuroscience* **7**: 290–3
- Moore C J, Price C J 1999 Three distinct ventral occipito-temporal regions for reading and object naming. *Neuroimage* **10**: 181–92
- Nakamura K, Honda M, Okada T 2000 Participation of the left posterior inferior temporal cortex in writing and mental recall of kanji orthography. A functional MRI study. *Brain* **123**: 954–67
- Nobre A C, Allison T, McCarthy G 1994 Word recognition in the human inferior temporal lobe. *Nature* **372**: 260–3
- Patterson K, Lambon-Ralph M A 1999 Selective disorders of reading? *Current Opinion in Neurobiology* **9**: 235–9
- Paulesu E, McCrory E, Fazio F, Menoncello L, Brunswick N, Cappa S F, Cotelli M, Cossu G, Corte F, Lorusso M, Pesenti S, Gallagher A, Perani D, Price C, Frith C D, Frith U 2000 A cultural effect on brain function. *Nature Neuroscience* **3**: 91–6
- Plaut D C, McClelland J F, Seidenberg M S, Patterson K 1996 Understanding normal and impaired word reading: Computational principles in quasi-regular domains. *Psychological Review* **103**: 56–115
- Price C J 1998 The functional anatomy of word comprehension and production. *Trends in Cognitive Sciences* **2**: 281–8
- Ricci P T, Zelkowitz B J, Nebes R D, Meltzer C C, Mintun M A, Becker J T 1999 Functional neuroanatomy of semantic memory: Recognition of semantic associates. *Neuroimage* **9**: 88–96
- Rumsey J M, Horwitz B, Donohue B C, Nace K L, Maisog J M, Andreason P 1999 A functional lesion in developmental dyslexia: Left angular gyral blood flow predicts severity. *Brain and Language* **70**: 187–204
- Sakurai Y, Momose T, Iwata M, Sudo Y, Ohtomo K, Kanazawa I 2000 Different cortical activity in reading of Kanji words, Kana words and Kana nonwords. *Cognitive Brain Research* **9**: 111–15
- Shaywitz S E, Shaywitz B A, Pugh K R, Fulbright R K, Constable R T, Mencl W E, Shankweiler D P, Liberman A M, Skudlarski P, Fletcher J M, Katz L, Marchione K E, Lacadie C, Gatenby C, Gore J C 1998 Functional disruption in the organization of the brain for reading in dyslexia. *Proceedings of the National Academy of Sciences USA* **95**: 2636–41
- Simos P G, Papanicolaou A C, Breier J I, Fletcher J M, Wheless J W, Maggio W W, Gormley W, Constantinou J E C, Kramer L 2000 Insights into brain function and neural plasticity using magnetic source imaging. *Journal of Clinical Neurophysiology* **17**: 143–62
- Tarkiainen A, Helenius P, Hansen P C, Cornelissen P L, Salmelin R 2000 Dynamics of letter string perception in the human occipitotemporal cortex. *Brain* **122**: 2119–213

J. A. Fiez

Reading Nonalphabetic Scripts

1. The Japanese Language

Linguistically, Japanese does not relate to Chinese, Korean, or English, though it is similar to Korean in syntax, but not in speech sounds or vocabulary.

However, Japanese has borrowed many words and characters, mainly from Chinese. Almost all indigenous people living in Japan speak Japanese, and ethnic Japanese living outside of Japan also do so. There are about 130 million Japanese speakers, and the language ranks sixth among the languages of the world after Chinese, English, Russian, Hindi, and Spanish.

The Japanese language employs five vowels and 16 consonants. Compared to other languages such as English, the Japanese sound system contains few sounds. The Japanese sound system uses a unit called a 'mora,' which is a short beat, the time taken to pronounce a short syllable. The mora is the main unit in the Japanese language, and is also the unit of Japanese phonetic script—hiragana (or hirakana) and katakana. Each Japanese syllabary sign is based on a mora. There are 110 different moras, or syllables, in Japanese. It is one of the smallest inventories of any major languages. For example, there are 400 syllables in Chinese and several thousand in English. The small inventory of Japanese sounds gives rise to many homophones, but it is easy to represent the language by a simple syllabary.

The Japanese vocabulary consists of four word types: Japanese origin (47.5 percent), Sino-Japanese (36.7 percent; long ago adopted from China), foreign loan words (9.8 percent; mostly European adopted in modern times), and hybrids (6 percent; Sino-Japanese or foreign word stems with Japanese endings).

2. The Japanese Writing System

The Japanese writing system consists of five kinds of script—kanji, hiragana, katakana, roman alphabet, and arabic numerals. Among these, the first three script types are employed most frequently in writing Japanese. Japanese text is therefore written using these five kinds of script, and it is also true that both the traditional vertical (tategaki) from top to bottom, and the horizontal (yokogaki), from left to right, styles are used depending on the kind of being written. For example, newspapers are printed in tategaki style, whereas academic journals are printed in yokogaki style.

3. Kanji Script and Reading

As a result of waves of Chinese and Korean emigration to Japan during the fourth and seventh centuries AD, Chinese-based culture, mainly in the form of Confucian classics, medicine, and Buddhism, came to Japan. A writing system with Chinese characters was, therefore, the only script until two forms of syllabary, hiragana and katakana were created from kanji during the ninth century AD. 'Kan-on' kanji readings were borrowings from Confucian scholarship. Chinese 'hanzi,' associated with the Han Dynasty period, became Japanese 'kanji,' which remains the backbone of the Japanese writing system. As Japanese kanji

borrowed Chinese hanzi in distinct historical periods, different pronunciations were adopted for these kanji, and makes reading complicated in modern writing systems. This feature of having multiple readings for many kanji characters in Japanese is different from those of Chinese and Korean, which have only a single reading.

The Japanese language uses not only Chinese and native readings of the logographic characters, but also several varieties of each type of reading, which makes the oral reading of characters complicated. For example, Mandarin Chinese read the simple character for 'ten' 十 as one sound 'shi,' while the Korean reading is 'sip' (no tone or soundless). Both languages use only one reading, but Japanese speakers read the character as 'to,' 'toh,' 'so,' 'jyu,' 'jutt-,' all without tone (Taylor and Taylor 1995). The first three sounds are examples of Kun readings (Japanese native readings), while the others are 'On' readings (Chinese readings). The Kun reading is quite different from the Chinese because the two languages are unrelated. Besides the Kun reading, most kanji have one or more On readings which are not identical with Chinese sounds. Depending upon when and where the Chinese characters were borrowed, four different types of On reading (go-on, kan-on, toh-soh-on, and the habitual one) were agreed. Most Japanese kanji are given only one of the four types of On reading.

The number of kanji is enormous. Kanji dictionaries contain between 12,000 and 50,000 entries. In 1981, the Japanese Ministry of Education attempted to restrict the number of kanji as 'joyo kanji' to be used in the educational system. This contains 1945 kanji, but because almost all kanji possess both On and Kun readings, Japanese people therefore learn a total of 4084 kanji readings during their first 9 years of education. Generally, each kanji consists of a combination of radicals. There are 250 kinds of radical, each consisting of between one and 17 components (strokes). Most radicals consist of two to nine strokes (Ueda et al. 1993). As the number of kanji is so huge, this reduces any difficulty in understanding homophones. However, there is a clear asymmetry between reading and writing ability, and it is true that some adult Japanese fail to read even joyo kanji.

Cognitive psychologists once suggested that the processing route of reading depended on the type of orthography—that is, kanji has a single route from orthography to semantics, bypassing the contribution of phonological information, while the kana travel a phonological route to access semantics. But this is too simple an explanation. Recent reviews of kanji studies suggest that it is reasonable to conclude that, depending upon the task settings, familiarity, frequency, and complexity of kanji, one of the most efficient processing routes of phonology-based or grapheme-based routes (or both routes) is employed to access semantics (Kess and Miyamoto 1999). However, in the case of high-frequency and high-familiarity kanji,

Table 1
Samples of radicals

Radicals	Examples	
<i>Hen</i>	波(wave)	河(river)
<i>Tsukuri</i>	清(clear water)	晴(clear weather)
<i>Kanmuri</i>	草(grass)	花(flower)
<i>Ashi</i>	熊(bear)	馬(horse)
<i>Kamae</i>	国(state)	園(to enclose)
<i>Tare</i>	庭(garden)	店(shop)
<i>Nyo</i>	返(to return)	逃(to escape)

semantic access is achieved only through the grapheme-based route, and this is not specific to Japanese. There are examples in alphabetic systems where the cognitive route is also direct from grapheme to semantics.

As shown in Table 1, complex kanji consist of semantic and phonetic radicals, *hen* (left hand), *tsukuri* (right hand), *kanmuri* (top), *ashi* (bottom), *kamae* (enclosure), *tare* (top-left), and *nyo* (bottom-left). Each acts differentially in reading. That is, generally, semantic radicals give clues to meanings, whereas phonetic radicals give clues to pronunciation. Research findings also suggest that the right hand radical is more informative than the left.

Many words in Japanese are represented not only by individual kanji but also by a compound which is a polysyllabic composite of between two and four kanji, usually carrying an On reading. More than 50 percent of the words in most Japanese dictionaries are compound kanji. Research findings also suggest that the reading processes of kanji compound words constitute another grouping. In addition to the frequency of the compound word itself, the first kanji of a compound is important, because the first kanji defines the range of possible collocations in the cohort of potential kanji compounds.

There is strong evidence that the individual kanji are processed as a recognition unit and kanji compounds themselves are also recognition units. A case has been that one agraphic patient was able to read the single kanji 手 'hand' and 紙 'paper' but failed to read their coupling as 手紙 'letter,' while another patient had less difficulty in reading a compound word, but could not read the single kanji which constituted the compound word (Hatta et al. 1998). This double dissociation suggests that single and compound kanji words are stored with different addresses in the mental lexicon.

4. Kana Scripts and Reading

In addition to kanji, there are two syllabary types of script in Japanese—hiragana and katakana (see Table 2). The kana unit represents a sound unit, which corresponds roughly to a syllable. Both scripts contain

Table 2

Hiragana, katakana, and kanji descriptions of the words 'firefly' and 'dreams'

	Firefly	Dream
Hiragana	ほたる	ゆめ
Katakana	ホタル	ユメ
Kanji	螢	夢

46 basic symbols and 25 additional symbols with diacritic marks. Hiragana and katakana share the same syllabic reference points, therefore the same syllable can be written by either system. Hiragana is cursive, as shown in Table 2, and used mainly for some content words, morphological endings, function words, and the rest of the grammatical scaffoldings of sentences. In contrast, katakana is angular in shape and is used for writing foreign loan words.

Kana developed gradually out of Chinese characters in the following order—kanji, kanji as phonetic signs, simplified kanji shapes, and kana. During this development, mainly female authors used hiragana to write letters, poems, diaries, and stories; this is because hiragana is a graceful, cursive form of its original kanji. For a long time, it has been the convention that a cursively simplified hiragana is used for grammatical morphemes, and the angular simplified katakana is used for foreign loan words and onomatopoeia. However, an explanation for the use of both types of script because of differing roles in syntax cannot be the only reason why the Japanese have kept three kinds of script usage irrespective of high cost. Japanese people might have found it useful or convenient to express semantic emotional information by employing different types of script.

Conventionally, there is a connection between orthography type and the words used. This has a strong effect on Japanese word readings. Japanese people possess a strong, emotional semantic image for each type of script—kanji is seen as masculine, hiragana feminine, and katakana exotic. Japanese people sometimes employ this emotional image of a script when writing, irrespective of the conventional usage of scripts and words. For example, reading experiments suggested that when Japanese people see the kanji word for chair 椅子, an image of an old-fashioned, strongly-built chair is generated; when they see the katakana word for chair イス, they imagine a modern and elegant chair; and in the case of the hiragana word for chair いす, they may imagine a simple wooden chair (Ukita et al. 1996).

5. Summary of Research Findings in Cognitive and Neuropsychological Studies

The results of earlier cognitive psychological experiments suggest that orthographic configurations can be used to access directly phonological form, meaning, or

both, depending on the processing task requirement. Although the role and the importance of these factors may differ for each orthography in respect to the processing strategy favored and the processing times required, Japanese seems not to be an exceptional writing system, but possesses idiosyncratic requirements of the world's writing systems.

However, neuropsychological studies have suggested that kanji and kana are processed differentially and stress the importance of orthography difference in reading (Iwata 1984, Yamadori 1975). In alexia with agraphia due to angular gyrus lesions, kana reading is impaired more severely than kanji, whereas left posterior–inferior temporal lesions impair kanji reading (Iwata 1984, Kawamura et al. 1987). These findings support the hypothesis that there are two pathways of reading in Japanese: one is the ventral pathway from the visual area in the occipital cortex via the posterior part of the inferior and middle temporal gyli of the left hemisphere, which is used in semantic reading of kanji; the other is the dorsal route from the visual area via the angular gyrus to Wernicke's area, and is used for the phonological processing of kana reading.

Regarding the visual half field, several studies have revealed that the right hemisphere contributes more than the left in kanji recognition. For example, a tendency toward a right hemisphere advantage was reported in the case of single kanji (Hatta 1977), and in the case of two-kanji non-word pairs (Sasanuma et al. 1977). On the other hand, a left hemisphere advantage was reported in the case of two-kanji compound words (Hatta 1978). By employing a very recent new research method (MEG), Koyama et al. (1998) revealed that not only the left but also the right hemisphere is activated in reading kanji words, and support the view of the right hemisphere engagement in kanji reading. These studies suggest that hemisphere function in kanji reading is not the same as in the reading of English and kana words.

The discrepancy between cognitive psychological and neuropsychological research findings in Japanese reading leaves us to conclude that there is a language non-specific cognitive processing in reading. Further studies concerning non-alphabetical scripts are necessary to answer this question.

See also: Dyslexia (Acquired) and Agraphia; Dyslexia, Developmental; Letter and Character Recognition, Cognitive Psychology of; Reading, Neural Basis of; Reading Skills; Writing, Evolution of; Writing Systems

Bibliography

- Hatta T 1977 Recognition of Japanese kanji in the left and right visual fields. *Neuropsychology* **15**: 685–88
Hatta T 1978 Recognition of Japanese kanji and Hiragana in the left and right visual fields. *Japanese Psychological Research* **20**: 51–9

- Hatta T, Kawakami A, Tamaoka K 1998 Writing errors in Japanese kanji. *Reading and Writing* **10**: 457–70
- Iwata M 1984 Kanji versus kana: Neuropsychological correlates of the Japanese writing system. *Trends in Neuroscience* **7**: 290–93
- Kawamura M, Hirayama K, Hasegawa K, Takahashi N, Yamaura A 1987 Alexia with agraphia of kanji (Japanese morphograms). *Journal of Neurology, Neurosurgery and Psychiatry* **50**: 1125–29
- Kess J F, Miyamoto T 1999 *The Japanese Mental Lexicon*. Benjamins, Amsterdam
- Koyama S, Kakigi R, Hoshiyama M, Kitamura Y 1998 Reading of Japanese kanji (morphograms) and kana (syllabograms): A magneto-encephalographic study. *Neuropsychology* **36**: 83–98
- Sasanuma S, Itoh K, Mori K, Kobayashi Y 1977 Tachistoscopic recognition of kana and kanji words. *Neuropsychology* **15**: 547–53
- Taylor I, Taylor M M 1995 *Writing and Literacy in Chinese, Korean, and Japanese*. Benjamins, Amsterdam
- Ueda M, Okada M, Iijima T, Sakaeda T, Iida D 1993 *Shin Daijiten*. Kodansha, Tokyo
- Ukita J, Sugishima M, Minagawa M, Inoue M, Kashu K 1996 *Psychological Research on Orthographic Forms in Japanese*. Psychology Monographs 25, Japanese Psychological Association
- Yamadori A 1975 Ideogram reading in alexia. *Brain* **98**: 231–8

T. Hatta

Reading Skills

The question of reading skill includes definitional and substantive components. The definitional question is answered as follows: reading skill is an individual's standing on some reading assessment. Skilled readers are those who score above some standard on this assessment; readers of low skill are those who score below some standard. The substantive question is this: What are the processes of reading that produce variation in assessed reading skill? This question is the focus here: given that two individuals differ in some global assessment of their reading, what differences in reading processes are candidates to explain this difference?

1. Normal and Special Variation in Reading Skill

Is a single analysis of reading skill sufficient to characterize both 'garden variety' variation and severe reading difficulties? Severe difficulties in reading in the absence of general intellectual problems is the standard definition of *dyslexia*, or specific reading disability. The discrepancy between achievement in reading and achievement in other domains is what sets dyslexia apart. However, there are reasons to blur the distinction between specific and nonspecific reading problems. Individuals in the two categories may differ

only in their achievements in some other nonreading area. The processes that go wrong in a specific disability may not be much different from those that go wrong for an individual who also has a problem in some other area (Stanovich and Siegel 1994). For both groups, difficulties in reading must be understood in terms of the processes of reading. What are the processes that can go wrong?

2. A Process Analysis

Reading processes depend on the language of the reader and the writing system that encodes that language. The units of the writing system are converted into mental representations that include the units of the language system. Specifically important are (a) the identification of words and (b) the engagement of language and general cognitive mechanisms that assemble these words into messages.

It is visual word identification that is the process most distinctive to reading. Beginning with a visual input—a string of letters—perceptual processes produce the activation of the *grapheme units* (individual and multiple letters) that constitute words. In traditional models of human cognition, the words are represented in a *lexicon*, the reader's mental representation of word forms and meanings. Successful word reading occurs when there is a match between the input letter string and a word representation. As part of this process, *phonological units*, including individual phonemes associated with individual letters, are also activated. The joint contribution of graphemic and phonological units brings about the identification of a word. It is common to refer to the phonological contribution to this process as *phonological mediation*. It is also common to represent two pathways, one from graphemic units to meaning directly, and one from graphemic units to phonological units, and then to meaning (the mediation pathway). The issues of mediation and one-or-two routes are central to alternative theoretical models of the reading process.

The alternative models do not assume a representation system that stores words. In one class of nonrepresentational models, words emerge from patterns of parallel and distributed activation (Plaut et al. 1996). In resonance models, word identification results from the stabilization of dynamic patterns that are continuously modified by interactions among inputs and various dynamic states resulting from prior experience (Van Orden and Goldinger 1994). An interesting feature of this model is that patterns of graphic-phonological activation stabilize more rapidly than do patterns of graphic-semantic activation. In effect, a word form becomes identified primarily through the convergence of orthography and phonology. Meaning is slower to exert an influence on the identification process.

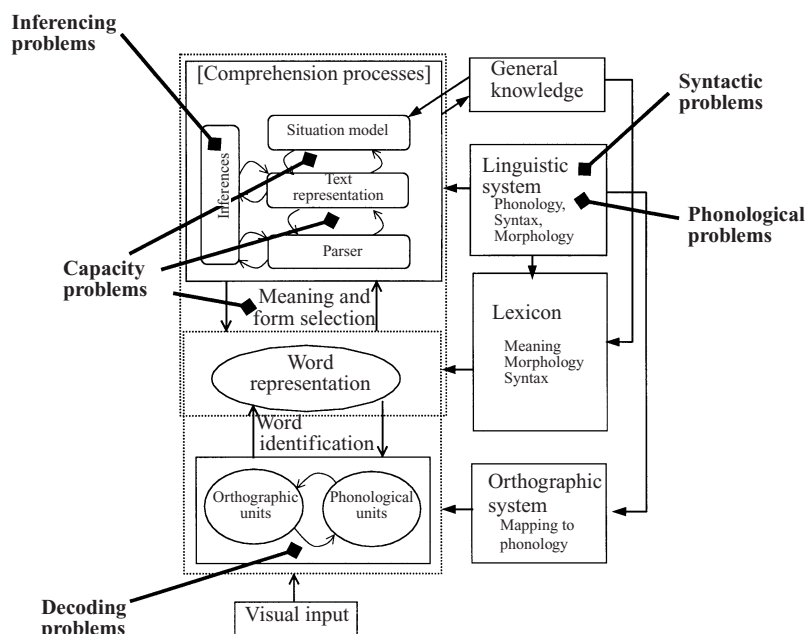


Figure 1

A schematic view of reading processes with candidate sources of reading problems. Adapted from Perfetti (1999)

In considering reading skill, it is possible to ignore differences among theoretical models to a certain extent. However, the theoretical models of reading processes actually make some commitments about the sources of reading problems. For example, a dual route model allows two very different sources of word reading difficulties: Either the direct (print-to-meaning) route or the indirect (print-to-phonology-to-meaning) route can be impaired (Coltheart et al. 1993). This provides a model for both developmental and acquired dyslexia. In acquired dyslexia, surface dyslexics are assumed to have selective damage to the direct route; phonological dyslexics are assumed to have selective damage to the phonological route. For developmental dyslexia, children may have a phonological deficit or an 'orthographic' (direct route) deficit. However, single mechanism models with learning procedures can give an alternative account of developmental dyslexia: only one type, phonological dyslexia, is the result of a processing defect; surface or orthographic dyslexia becomes a delay in the acquisition of word-specific knowledge (Harm and Seidenberg 1999). Such an example illustrates that understanding individual differences in reading is part of the same problem as theoretical understanding of reading processes.

Figure 1 schematizes an organization of the cognitive architecture for reading. Reading begins with (a) a visual input that, with immediate use of phonology, leads to word identification that (b) yields semantic information connected to the word as con-

strained by the current context. A word immediately is (c) integrated syntactically with an ongoing sentence parse and (d) semantically with an ongoing message interpretation (proposition). As sentences are read (e), an integrated text representation is developed, consisting of interrelated propositions. To establish a reasonably specific understanding of a text, (f) inferences establish a coherent referential model of what is being read.

Potential reading difficulties are seen in each component. At the word level, a reader could have trouble with processing letter strings or in selecting the required meaning of a word. A reader might have defective phonological processes, limiting word identification and memory for words. On the comprehension end, a reader could have syntactic difficulties or fail to generate inferences or to monitor comprehension. Because these processes can demand mental resources from a limited pool, a very general constraint is working memory limitations. Thus reading difficulty can arise from a general limit in functional working memory capacity or from the failure of some processes to escape this limit. The following sections consider some of these possibilities.

2.1 Lexical Orthographic-Phonological Processes

Readers who fail to read words accurately fail to comprehend. Thus, word-level problems are potentially the most important in creating reading failures,

because they lead both to word reading problems themselves and to derivative comprehension problems. For example, verbal efficiency theory (Perfetti 1985) assumes that readers who lack efficient word identification procedures are at risk for comprehension failure.

A word identification problem is a failure of phonological and orthographic information to cohere onto a word identity (Fig. 1). Knowledge of letters and knowledge of phonemes that connect with letters in specific letter environments are necessary. Less skilled readers tend to have weaker knowledge of word spellings (orthography) and of word sounds (phonology), and especially weaker knowledge of orthography–phonology connections (decoding). Various measures assess phonological processes, but the ability to read legal nonwords (pseudowords) has become a benchmark for decoding ability. Both the phonological and orthographic components of word reading appear to have a significant heritable component (Olson et al. 1999).

The relationship between orthographic and phonological knowledge may be complex. As indicated in Fig. 1, an orthographic processing event is also a phonological processing event, perhaps implying that the two knowledge sources develop in tandem. However, the effects of reading experience may be a bit different for these two components. Early encounters with words may be crucial for establishing (phonological) decoding knowledge with orthographic knowledge acquired more gradually, as individual words are ‘practiced’ in reading situations. Stanovich and West (1989) found that reading experience accounts for variance in word reading skill beyond that accounted for by decoding. However, this does not mean that word knowledge gained through experience is independent of phonology. Indeed, early in reading experience, decoding unfamiliar words may be the primary mechanism for establishing orthographic representations of specific words (Share and Stanovich 1995). More generally, phonological processes that link graphic inputs to word identities may enable gains from further experience in reading, increasing the quality and quantity of word representations (Perfetti 1992).

Referring to ‘orthographic’ and ‘phonological’ as separate sources of knowledge does not entail a commitment to two separate routes in word identification. Whether as defects in the phonological component of single mechanism or as impairments of a distinct processing route, phonological problems may be the central deficiency for both specific reading disability and nonspecific low reading skill (Stanovich 1988). They can arise at many levels, including speech processing mechanisms and sensitivity to phonological structures, as well as written word identification. Becoming aware that spoken words consist of meaningless sound segments (phonemes) is important in learning to read, and children who do not become

aware are at risk for failure (see Liberman and Shankweiler 1991). Moreover, phonological processing problems may have consequences throughout the reading process: learning to read words, remembering just-read words, and understanding phrases and sentences are all processes operating on partly phonological representations.

2.2 Processing Lexical Meaning

The ability to get context-appropriate meaning from words is central to reading skill. At one level, this is a question of vocabulary. The more words a reader knows, the better the comprehension. At another level, it is the ability to select the right meaning of a word in a given context. Although the description of lexical meaning selection has become complex, a still widely (not universally) shared conclusion is this: the selection of word meaning proceeds in two-stages: (a) a general activation stage in which a word is accessed and its associated meanings nonselectively activated, and (b) a selection stage in which the meaning appropriate for context is selected while meanings inappropriate for context are suppressed.

One hypothesis is that less skilled readers are less effective in selecting a contextually appropriate meaning. According to the structure building framework (Gernsbacher 1990), readers build a coherent framework for a text by enhancing concepts required by the text while suppressing those that are irrelevant. The suppression hypothesis is that less skilled readers have deficient suppression mechanisms. To illustrate, in the sentence, ‘He dug with the spade,’ the final word has two meanings, but only one fits the context of the sentence. However, when adult readers are immediately asked to decide whether a following word is related to the meaning of the sentence, their decisions are initially slow for the word ‘ace’ (related to the inappropriate meaning of spade). Both appropriate and inappropriate meanings may be activated at first. With more time before the appearance of ‘ace,’ skilled readers show no delay in rejecting it; i.e., they ‘suppress’ the irrelevant meaning. However, less skilled readers continue to react slowly to ‘ace,’ as if they have not completely suppressed the irrelevant meaning of ‘spade.’

Whether ineffective use of context is a source of reading problems has become a complex issue. Pre-scientific beliefs on this question seemed to be that poor readers failed to use context in reading words. However, research on children’s word identification led to the opposite result: less skilled readers use context in word identification at least as much and perhaps more than do skilled readers (Perfetti 1985, Stanovich 1980). Less skilled comprehenders are good users of context, which they use in compensation for weaker word identification skill to identify words. The suppression proposal adds to the poor reader’s basic

word identification problem a specific suppression problem, the effect of which is a less productive use of context.

2.3 Processing Syntax

Less skilled readers often show a wide range of problems in syntax and morphology. The question is whether such problems, which are found across a wide age range, arise from some deficit in processing syntax or from some other source that affects performance on syntactic tasks. One possibility is that syntactic problems reflect a lag in the development of linguistic structures. An alternative hypothesis is that the syntactic problems reflect constraints of working memory or lexical processing limitations.

To illustrate one class of syntactic problem, consider two sentences with relative clauses below. (a) is the easier subject relative; (b) is the more difficult object relative.

(a) The girl that the boy believed understood the problem.

(b) The girl that believed the boy understood the problem.

The greater difficulty of (b) compared with (a) can arise from different degrees of interference they produce in the attempt to assigning a subject for 'understood.' For syntactic reasons, the interference produced by the object relative (b) is greater than the interference produced by the subject relative (a). Details aside, if syntactic deficits rest on lack of specific syntactic knowledge, then one might expect some qualitative differences between skilled and less skilled readers. However, in the relative clause example and in other syntactic examples, patterns of difficulty for skilled and less skilled young readers tend to be the same, suggesting quantitative rather than qualitative differences (Crain and Shankweiler 1988).

Research with adults suggests that both reading problems and spoken language problems (aphasia) may arise from processing limitations rather than structural deficits (Carpenter et al. 1994). In a study of subject and object relatives similar to those in (a) and (b), King and Just (1991) found that readers with low working memory have problems with object-relative sentences. More interesting, these problems (as seen in reading times for words) were most severe where the processing load was hypothesized to be the greatest—at the second verb in the object relative, i.e., 'understood' in (b). Comprehension difficulties may be localized at points of high processing demands—whether from syntax or something else. If this analysis is correct, then the problem is not intrinsic deficits in syntax, but the processing capacity to handle complexity.

Another perspective on this issue is the opportunity for practice. Because some syntactic structures are more typical of written language than spoken lan-

guage, the opportunity for practice is limited by the ability to read. Thus, continuing development of reading skill as a result of initial success at reading—and the parallel increasing failure as a result of initial failure—is undoubtedly a major contributor to individual differences in reading.

2.4 Processing Text

The source of skill differences in text comprehension can lie in the lexical and working memory processes considered above or in higher-level processes specific to reading texts. In addition to identifying words, parsing sentences, and encoding context-sensitive meanings, readers must use knowledge from outside of the text and apply procedures that derive intended meanings. The overall complexity of text comprehension implies several possibilities for processing failure, any resulting in a less integrated or less coherent representation of the text.

One example is the processes of coreference, by which a reader maps a pronoun to an already established referent, as illustrated in (c).

(c) Jane saw Margaret shopping in the grocery store. She was buying bread.

When there are two possible referents for a pronoun, less skilled readers may take longer to assign the pronoun to the intended referent (Margaret in this case), at least when the referent is not the first noun in the preceding sentence (Frederiksen 1981). If this simple pronoun case is general, less skilled readers may be less adept at integrating referential information across sentences.

The text processing component that has received the most attention as a source of individual differences is making inferences. Oakhill and Garnham (1988) summarize evidence suggesting that less skilled readers fail to make a range inferences in comprehension. When skill differences are observed, an important question is whether they occur in the absence of problems in lexical skills, working memory, or other lower-level factors. The answer to this question remains unclear in much of the research. However, there has been some success in identifying a group of children whose problems can be considered comprehension-specific, although highly general across reading and spoken language (Stothard and Hulme 1996).

Another example is comprehension monitoring, the reader's implicit attempts to assure a consistent and meaningful understanding. Skilled readers can use the detection of a comprehension breakdown (e.g., an apparent inconsistency) as a signal for rereading and repair. Less skilled readers may fail to engage this monitoring process (Baker 1984, Garner 1980). Low skilled comprehenders seem especially poor at detecting higher level text inconsistencies that negate a coherent model of the text content—for example, whether successive paragraphs are on unrelated topics.

However, it is not clear whether these higher level differences are independent of the reader's ability to construct a simple understanding of the text. Some evidence suggests that less skilled comprehenders fail accurately to represent or weight the propositions in a text (Otero and Kintsch 1992). The general interpretive problem here is that comprehension monitoring, like inference making, both contributes to and results from reader's text representation. This makes it difficult to attribute comprehension problems uniquely to failures to monitor comprehension, as opposed to more basic comprehension failures.

Because no texts are fully explicit, the reader must import knowledge from outside the text. Thus, a powerful source of comprehension differences is the extent to which a reader has such knowledge for a given text (Anderson et al. 1977). However, readers of high skill can compensate for lack of knowledge to some extent (Adams et al. 1995). It is the reader who lacks both knowledge and reading skill who is assured failure. The deleterious effect of low reading skill (and its motivational consequences) on learning through reading creates readers who lack knowledge of all sorts.

3. Summary

With a focus on the component processes, individual differences in reading skill become a matter of understanding how these processes and their interactions contribute to successful reading outcomes. Where the successful outcome is reading individual words, the processes are localized in knowledge of word forms—both general and word-specific phonological and orthographic knowledge—and word meanings. Inadequate knowledge of word forms is the central obstacle to acquiring high levels of skill. Severe problems in word reading reflect severe problems in phonological knowledge. Where the successful outcome is comprehension, the critical processes continue to include word processes, and problems in comprehension are associated with problems in word processing. In addition, processes that contribute to basic sentence understanding and sentence integration become critical. Processes that provide basic propositional meaning, including word meaning selection and parsing, and those that establish coherent text representations (integration processes, inferences, monitoring, conceptual knowledge) become critical to success. Less skilled readers, as assessed by comprehension tests, often show difficulties in one or more of these processes. Less clear is how to understand the causes of observed failures. A processing model helps to see the relationships among component processes and to guide studies of skill differences. The candidate causes of skill variation cannot be equally probable, when the output of lower-level processes are needed by higher level processes. The identification of basic

causes, as opposed to derivative symptoms that reflect the costs of low reading skill, requires research that can provide strong causal tests.

See also: Dyslexia (Acquired) and Agraphia; Dyslexia, Developmental; Dyslexia: Diagnosis and Training; Dyslexia, Genetics of; Eye Movements in Reading; Letter and Character Recognition, Cognitive Psychology of; Lexical Processes (Word Knowledge); Psychological and Neural Aspects; Reading, Neural Basis of; Reading Nonalphabetic Scripts; Speed Reading; Word Recognition, Cognitive Psychology of

Bibliography

- Adams B C, Bell L C, Perfetti C A 1995 A trading relationship between reading skill and domain knowledge in children's text comprehension. *Discourse Processes* **20**: 807–23
- Anderson R C, Reynolds R E, Shallert D L, Goetz E T 1977 Frameworks for comprehending discourse. *American Educational Research Journal* **4**: 367–81
- Baker L 1984 Spontaneous versus instructed use of multiple standards for evaluating comprehension: Effects of age, reading proficiency and type of standard. *Journal of Experimental Child Psychology* **38**: 289–311
- Carpenter P A, Miyake A, Just M A 1994 Working memory constraints in comprehension: Evidence from individual differences, aphasia, and aging. In: Gernsbacher M A (ed.) *Handbook of Psycholinguistics*. Academic Press, San Diego, CA, pp. 1075–122
- Coltheart M, Curtis B, Atkins P, Haller M 1993 Models of reading aloud: Dualroute and parallel-distributed-processing approaches. *Psychological Review* **100**: 589–608
- Crain S, Shankweiler D 1988 Syntactic complexity and reading acquisition. In: Davison A, Green G M (eds.) *Linguistic Complexity and Text Comprehension: Readability Issues Reconsidered*. Erlbaum, Hillsdale, NJ, pp. 167–92
- Frederiksen J R 1981 Sources of process interactions in reading. In: Lesgold A M, Perfetti C A (eds.) *Interactive Processes in Reading*. Erlbaum, Hillsdale, NJ, pp. 361–86
- Garner R 1980 Monitoring of understanding: An investigation of good and poor readers' awareness of induced miscomprehension of text. *Journal of Reading Behaviour* **12**: 55–63
- Gernsbacher M A 1990 *Language Comprehension as Structure Building*. Erlbaum, Hillsdale, NJ
- Harm M W, Seidenberg M S 1999 Phonology, reading and dyslexia: Insights from connectionist models. *Psychological Review* **106**: 491–528
- King J, Just M A 1991 Individual differences in syntactic processing. *Journal of Memory and Language* **30**: 580–602
- Lieberman I Y, Shankweiler D 1991 Phonology and beginning reading: A tutorial. In: Rieben L, Perfetti C A (eds.) *Learning to Read: Basic Research and its Implications*. Erlbaum, Hillsdale, NJ, pp. 3–17
- Oakhill J, Garnham A 1988 *Becoming a Skilled Reader*. Blackwell, New York
- Olson R K, Forsberg H, Gayan J, Defries J C 1999 A behavioral-genetic analysis of reading disabilities and component processes. In: Klein R M, McMullen P A (eds.) *Converging Methods for Understanding Dyslexia*. MIT Press, Cambridge, MA, pp. 133–53

- Otero J, Kintsch W 1992 Failures to detect contradictions in a text: What readers believe versus what they read. *Psychological Science* **3**(4): 229–35
- Perfetti C A 1985 *Reading Ability*. Oxford University Press, New York
- Perfetti C A 1992 The representation problem in reading acquisition. In: Gough P B, Ehri L C, Treiman R (eds.) *Reading Acquisition*. Erlbaum, Hillsdale, NJ, pp. 145–74
- Perfetti C A 1999 Comprehending written language. A blue print of the Re. In: Hagoort P, Brown C (eds.) *Neurocognition of Language Processing*. Oxford University Press, Oxford, pp. 167–208
- Plaut D C, McClelland J L, Seidenberg M S, Patterson K 1996 Understanding normal and impaired word reading: Computational principles in quasi-regular domains. *Psychological Review* **103**: 56–115
- Share D L, Stanovich K E 1995 Cognitive processes in early reading development: Accommodating individual differences into a model of acquisition. *Issues in Education* **1**: 1–57
- Stanovich K E 1980 Toward an interactive-compensatory model of individual differences in the development of reading fluency. *Reading Research Quarterly* **16**: 32–71
- Stanovich K E 1988 Explaining the differences between the dyslexic and the garden-variety poor reader: The phonological-care variable-difference model. *Journal of Learning Disabilities* **21**: 590–604
- Stanovich K E, Siegel L S 1994 Phenotypic performance profile of children with reading disabilities: A regression-based test of the phonological-core variable-difference model. *Journal of Educational Psychology* **86**(1): 24–53
- Stanovich K E, West R F 1989 Exposure to print and orthographic processing. *Reading Research Quarterly* **24**: 402–33
- Stothard S, Hulme C 1996 A comparison of reading comprehension and decoding difficulties in children. In: Cornoldi C, Oakhill J (eds.) *Reading Comprehension Difficulties: Processes and Intervention*. Erlbaum, Hillsdale, NJ, pp. 93–112
- Van Orden G C, Goldinger S D 1994 The interdependence of form and function in cognitive systems explains perception of printed words. *Journal of Experimental Psychology: Human Perception and Performance* **20**: 1269–1291.

C. A. Perfetti

Real Estate Development

1. What is Development?

Real estate development is the process of *adding value* to real estate. While most people think of development in terms of new buildings and subdivisions, development refers to any situation where value is added to a property. Developers take ‘development risk’ which ranges from leasing-risk associated with buying buildings with high vacancy to buying raw land with no utilities or zoning, and undertaking all the risks associated with obtaining public approvals, financing, construction, leasing, and operations of a new development project.

Developers initiate, plan, and coordinate the entire development process—buying the land, designing the

product, arranging financing, obtaining public approvals, constructing the building, leasing, managing, and ultimately selling the completed project. Development is distinguished from owning and managing property because it includes at least some element of adding value. While most developers also own and manage their buildings, many real estate investors and owners buy buildings that are completed and fully leased and which do not involve adding value through some form of development.

1.1 What Distinguishes Good Development From Bad?

While there is no absolute standard to distinguish good development from bad, the test is a simple one: Does the development enhance the social welfare of the community? (Peiser 1990). Providing new housing and office space enhances the community, but if it does so at the expense of environmentally precious resources, then on balance it may reduce the social welfare. Unfortunately social welfare is difficult to measure—especially the indirect costs and benefits of externalities associated with the development. Obtaining approvals for development is a political process that often is long and costly. One of the biggest challenges facing city planners and council members is to create an approval process that is unbiased and gives the various stakeholders appropriate voice in the outcome. At the same time, they must do their best to streamline the process and prevent unnecessary delays and costs.

1.2 Academic and Professional Antecedents

Real estate development relates to a number of academic fields including economics, urban planning, architecture, finance, law, political science, civil engineering, and geography. Urban economics provides the foundation for understanding real estate value and how cities change over time (Mills 1972). Real estate finance, public finance, land use, planning, and environmental law and economics, infrastructure, market analysis, demography, architecture, and landscape architecture all have significant bodies of research that contribute to our understanding of urban development and its impact on cities and countryside (see Brueggeman and Fisher 1997, DiPasquale and Wheaton 1996, Case and Shiller 1989).

Developers are generalists. They need to be conversant with many different disciplines. Because they are responsible for conceiving, planning, and executing the entire process, developers typically are strongest in finance and marketing. Because they determine what is built and how the urban landscape appears, they should be intimately familiar with cities—what works and what does not, both from a

private development perspective and from a government perspective (see Garvin 1996).

Successful development requires knowledge of all of the real estate disciplines—brokerage, finance, law, political approvals, design, construction, and management. Developers usually have several years of real estate experience before they undertake their own projects, most often in brokerage or real estate finance and investment. Legal, architectural, planning, and construction backgrounds also provide excellent training.

2. Industry Structure

2.1 Ownership

The real estate development industry is an industry in transition. Historically, it consisted primarily of private entrepreneurs who operated in a single market. Today, large publicly held national companies are emerging as significant owners of property, although their role and competitive position as developers is still being defined.

The different segments of the real estate industry comprise almost 20 percent of the nation's gross domestic product (Urban Land Institute 1997). The real estate industry is composed of many different players (see Fig. 1). Developers are the entrepreneurs who create and produce real estate. They may continue to own and manage the completed buildings or they may sell them to other parties. While developers have traditionally come from the private sector in the USA, redevelopment agencies and public housing departments act as public developers, often through public-private partnerships. Many corporations and public agencies including the armed forces also develop property for their own use, which is later recycled into the inventory of space available for anyone.

Each of the players is involved in a different phase of producing or managing real estate. Developers, architects, lenders, planners, and city council workers are more involved at the front end. Owners, property

Developers
Contractors
Lenders
Brokers
Property managers
Architects/Engineers
Planners/City council
Government agencies

Figure 1

Real estate industry players

Location	Property type
<i>CBD</i>	<i>Residential</i>
<i>Inner city</i>	<i>Office</i>
<i>Suburbs</i> — <i>Wealthy</i> <i>Graying</i>	<i>Industrial</i>
<i>Exurbs</i>	<i>Retail</i>
<i>Rural</i>	<i>Hospitality</i>
	<i>Land</i>

Figure 2

Real estate markets are defined by location and property type

managers, building departments, police and fire departments, and brokers tend to be more involved with the real estate after it is built.

Owners of property include the following:

Corporations
REITS (Real Estate Invest Trusts)
Banks—*REO* (Real Estate Owned)
Partnerships/Joint Ventures (JVs)
Individuals

2.2 Property Type

Figure 2 provides a breakdown of the real estate industry by location and property type. Each location represents a different and distinct market. All property types may be found in all locations. Certain property types, however, such as major office buildings and hotels are more likely to be developed in the Central Business District (CBD) and wealthy suburban nodes—typically located at the intersections of major freeways whereas industrial buildings, apartments, and shopping centers are more ubiquitous.

The real estate industry is divided into six main property types: residential, office, commercial, industrial, hospitality, and land (see Fig. 2). Each property type is virtually a separate industry, serving different purposes, markets, and customers. Although industrial warehouses, for example, may have completely different development requirements from apartments, they share a common institutional framework with respect to land, financing, legal aspects, and even construction. Still, each product is distinct, and value is created from different features. Apartments and industrial both benefit from locations near major highways. Industrial developers, however, prefer to avoid residential areas where homeowners complain about truck traffic. And apartment developers are attracted to areas with good schools, good retail services, recreation, and other amenities that are much less important to industrial tenants.

Developers, lenders, brokers, professional service providers, and most suppliers tend to specialize in one property type. While there is considerable cross-over

between property types, such as air conditioning subcontractors who work on both office buildings and apartments, professionals develop a network of relationships over the years that favor those who concentrate on a single property type.

2.3 For-Sale vs. Income-Producing Property

A critical distinction between property types is whether they are 'for-sale' or are 'income-producing.'

<i>For-sale</i> property	<i>Income-producing</i> property
Land	Apartments
Single-family homes	Office
Condominiums	Retail
	Industrial
	Hotels

For-sale property is sold to a buyer after it is developed whereas income-producing property is held for the cash flow it generates. For-sale property is valued based on how much buyers are willing to pay for enjoying the benefits of ownership—such as the consumption value of living in a home in a good school district near shopping and recreation.

Income-producing property includes five main property types. Values for income properties are based on the cash flow they generate from leases. Office buildings, for example, have tenants who sign 3 to 10 year leases. The developer uses the lease income to obtain long-term mortgage financing. When the office building is sold, investors pay a premium for leases signed by major corporations who have excellent credit ratings. A lease signed by a major corporation like IBM with a AAA credit rating is like purchasing a bond.

Income properties are valued by capitalizing their Net Operating Income (NOI). Capitalization rates are inversely proportional to prices. For example, if a building produces an NOI of \$100,000 per year, it is worth \$1.25 million on an 8 percent cap-rate (\$100,000 divided by 0.08) vs. \$1.0 million on a 10 percent cap-rate (\$100,000 divided by 0.10). Investors may be willing to purchase a building occupied by a major corporation such as IBM using an 8 percent capitalization rate whereas they would use a much higher cap-rate, say a 10 percent, for the same property occupied by a small non-credit tenant. The rental payments of a major corporation are like a bond—carrying the same credit rating as the corporation itself. In fact, the owner can sue the corporation for lost rent if the tenant defaults on the lease payment.

2.4 Design and Density

In addition to for-sale vs. income property, the development industry may be subdivided by design and density—expressed in units per acre for residential and Floor Area Ratio (FAR) for commercial buildings. Densities for residential property range from one

Design

Hi rise

Low rise/garden

Single family attached

Single family detached

Density

80 + DUs/acre

30 – 80 DUs/acre

18 – 30 DUs/acre

8 – 16 DUs/acre

2 + acres/DU–

12 DUs/acre

Figure 3

Residential design and density

unit/2+ acres for single-family detached homes to 80–200 units/acre for a high-rise apartment building. Residential developers tend to focus either on for-sale single-family homes or income-producing apartments. Figure 3 shows the range in densities for different types of residential construction.

3. The Development Process

Figure 4 shows the time line of development for income-producing property. The two main periods—development and operations—involve dramatically different risks, types of financing, and activities (see Peiser 1992).

The 'Development period' runs from the inception of the deal through stabilized occupancy (90–95 percent occupancy). This period is financed by a construction loan, which covers the construction and lease-up of the building. The building is technically placed in operation when the 'Certificate of occupancy' is received. Since lease-up represents a major activity of development and risk, however, it is shown in Fig. 4 as part of the development period. The 'operating period' runs from the point in time when the building reaches stabilized occupancy to the time of sale. Once stabilized occupancy is achieved, the developer is able to obtain a long-term permanent mortgage which is used to pay off ('take out') the construction loan (see Sect. 3.5).

3.1 Stages of Development

The development period in Fig. 4 includes predevelopment, construction, and leasing activities. The pre-development period runs from project initiation up to the start of construction. It includes feasibility studies, site acquisition, design, and finance.



Figure 4

Time line of development

3.2 Feasibility Studies

Many of the critical development decisions are made even before a developer acquires a site. Market feasibility studies help the developer identify the best location and site characteristics for a particular development (see DiPasquale and Wheaton 1996). Market feasibility differs for each product type. It includes demographic projections, income, employment, retail sales, and growth projections. The market study should conclude with estimates of demand for the product expressed in units or square feet/year in a particular location for a particular product-product type, as well as estimates of current and projected supply. The market study tells the developer how much space he can build, what rent to charge, and how quickly it can be absorbed by the market place. The market study identifies competitive properties, and helps the developer determine the best location in which to develop the product. It also tells the developer what the design features and amenities should be in order to attract the greatest number of customers.

For site selection, developers will often consider as many as 50–100 different sites in order to find five sites which they examine in detail. From the five they may pick one site to purchase. Development is said to be a local business because experienced developers will know information about most of the potential development sites, competition, rents, and costs in their market area.

3.3 Site Acquisition

Property acquisition occurs in three main stages: (a) putting the property under contract ('opening escrow'), (b) going hard on the earnest money deposit, and (c) closing. Purchase contracts typically run from 60 to 120 days or longer. When the buyer makes an offer, he places an 'earnest money deposit' of 3–5 percent of the total purchase price. Purchase contracts have a number of contingencies, which the developer wants to satisfy before he consummates the purchase. 'Closing' is the transfer of title to the buyer and the payment of all money to the seller.

After a site is placed under contract, the developer has a 15 to 30 day contingency period to do feasibility studies. The developer will hire consultants to determine the site's suitability for development. The analysis should cover site characteristics, environmental status, topography, drainage, utilities (water, sewer, storm, electricity, phone, gas, and cable), as well as neighboring properties, and the chance of obtaining necessary zoning and building approvals. Once the developer is satisfied that these conditions are suitable for development, he will go hard on the contract. 'Going hard' means that the earnest money becomes non-refundable—when the major contingencies such as good title and finding no environmental problems are satisfied. Since it may take

longer than the typical 15–30 day contingency period for the developer to perform the full range of studies necessary for development, the developer will often need to ask for an extension. The developer may also do much of the investigation before the property is even placed under contract or while the purchase contract is being negotiated. In a hot market, sellers are usually unwilling to extend the contingency period because other buyers are waiting in the wings.

For a large project, the feasibility period can extend several years and cost millions of dollars before the developer receives all of the necessary approvals. In a large project, the developers almost always has to purchase the property before many of the feasibility studies are done and long before the public approvals are in place. Such projects carry full 'entitlement' risk and represent the most risky (and potentially the most profitable) form of development.

3.4 Design

Design includes land planning, architecture, engineering, and consulting services that go into producing completed working drawings from which the building is built. The developer will often perform considerable preliminary design work on a property even before 'going hard' in order to determine how much space can be built, and what it will cost based on preliminary construction bids. Design drawings are used by the developer to obtain financing and political approvals. The architect prepares working drawings after the developer approves the final design. The general contractor and the subcontractors use the working drawings to bid the project and to build it.

3.5 Finance

If developers have a distinct expertise, it is in their ability to raise equity and debt money to finance their development activities. Development financing has two main pieces—construction financing to build the project, and permanent financing for operating it. Developers obtain construction financing primarily from commercial banks. Construction loans have short terms—1–2 years—and higher interest rates—2–3 percent above 'prime'—than permanent mortgages. Permanent mortgages are long term—30 years, and have lower interest rates than construction loans because the completed property, especially if it has high-quality tenants, offers better security to the lender than one under construction. Historically, pension funds, insurance companies, and savings and loan associations provided most of the permanent mortgage financing, but today Wall Street is heavily involved in it through securitization, commercial mortgage-backed securities, and conduit lending. As a result, permanent financing is highly competitive in the market place and developers and owners of

Development cost		Annual operating income	
Land	\$ 2,000,000	Gross rent	\$ 1,500,000
Building	\$ 8,000,000	Vacancy	5 percent \$ 75,000
Total	\$10,000,000	Effecting gross income	\$ 1,425,000
		Less operating expenses	\$ 425,000
Financing		NOI	\$ 1,000,000
Equity	\$ 2,500,000	Less debt service	\$ 660,388
Mortgage	\$ 7,500,000	Cash throw-off	\$ 339,612
Total	\$10,000,000		
		*Mortgage assumptions	
		Mortgage amount	\$7,500,000
		Interest rate	8 percent
		Term	30 years
		Annual debt service	\$660,388

Figure 5
Development cost and operating income

property have many more sources for permanent debt than they used to.

Equity financing is the most difficult and expensive money that developers raise. It represents the developer's own cash investment in a deal, or the money raised from 'equity investors.' It is much riskier than debt financing because equity investors receive cash flow only after the mortgage payments are made during operations and after the debt is paid off when the property is sold.

For example, if a building costs \$10 million altogether, the developer would typically invest 25–30 percent cash equity and the rest would be covered by debt. Suppose the building at stabilized occupancy (90–5 percent occupancy) generates \$1,000,000 in Net Operating Income (revenues minus expenses—\$1,000,000 NOI divided by \$10,000,000 purchase price equates to a 10 percent capitalization rate). Figure 5 shows the Development Cost, Financing, Annual Operating Income, and Mortgage assumptions for the sample building.

In Fig. 5 the simple return measures are the overall capitalization rate and the cash on cash return. The overall capitalization rate equals the NOI/Total Cost = \$1,000,000/\$10,000,000 = 10 percent. The *cash on cash return* equals the Cash Throw Off/Equity = \$339,612/\$2,500,000 = 13.58 percent.

Sophisticated investors focus on before-tax and after-tax internal rates of return (IRR) to determine whether or not they should invest in a particular real

estate deal. IRRs represent annual rates of return on equity that are comparable to returns from investing in stocks or bonds. IRR requirements depend on risk. Investors require IRRs on equity that are higher than 20 percent for new properties under development whereas they may require returns as low as 10–12 percent for investing in a fully leased office building with high-quality tenants (see Brueggeman and Fisher (1997) for a detailed discussion of IRRs).

3.6 Construction

Developers hire general contractors to build most nonresidential buildings. Homebuilders and many apartment developers, however, often do their own construction. The general contractor (or homebuilder) hires subcontractors who specialize in constructing various parts of the building such as foundations, framing, glass, granite, electrical, plumbing, heating, ventilating, and air conditioning (HVAC). Legal arrangements between the developer and the contractor may take several forms depending on who bears the risk for delivering the completed building within budget and on time. Common contract arrangements include fixed price, negotiated bid, design-build, fixed-price with a savings clause (where savings are split between the general contractor and the developer), and time and materials. Larger contractors are often unionized while small contractors and

especially those who work on residential projects and in certain 'right-to-work' states such as Texas are typically not unionized and tend to be lower cost. During the construction phase, the developer is most concerned about delivering the building on time and under budget. Since working drawings are often incomplete or have mistakes, a nonconfrontational relationship between the developer and contractor is key to success.

3.7 Leasing and Sales

Marketing for a new project begins even before the developer purchases the land. In order to obtain financing for office and retail projects, the developer often needs to demonstrate that tenants are interested in the property by obtaining letters of intent or signed leases on 30–50 percent or more of the property.

Developers may have in-house marketing personnel or use outside brokers to lease and sell their property. Exclusive brokerage arrangements are common, whereby a brokerage firm is the developer's primary agent. Brokers work on commission which ranges from 4–6 percent of the lease or sale value for smaller deals down to 1–2 percent for larger transactions in the millions of dollars. For leasing office or warehouse space, the developer will typically pay the broker who he has hired as the exclusive agent for his property a 3–4 percent commission for signing a lease if there are no other brokers involved. The tenant often has his own broker, in which case, a higher commission will be split between the two brokers.

Successful leasing involves more than having the lowest rent for the nicest space. The developer typically gives the tenant an allowance for tenant improvements (TIs) which covers interior finish. For law offices in which tenant finish can run \$70–\$80 or more per square foot (compared to standard finish of \$15–\$25 per square foot), the TI allowance can make the difference in getting the law firm to sign a lease.

3.8 Operations

Larger developers usually have their own property management departments to handle the ongoing leasing and management of a building after it is completed. Smaller developers may hire outside firms that specialize in property management. Property management activities include repair and maintenance of the property, leasing, relations with tenants, security, building and payment of all capital and operating expenses. Good developers give careful attention to property management because it sets the tone of relations with tenants and determines their firm's image in the marketplace.

The operating period ends with the sale of the property to a third party. Most real estate deals

generate a significant part of their financial returns from appreciation in the value of the property relative to its development or purchase cost. IRRs are highly sensitive to the length of the holding period. The shorter the holding period, say 3 years or less, the higher the IRR. It is standard practice when preparing a financial proforma to assume a 7 or 10 year holding period. The actual holding period depends on the owner's personal or institutional cash flow, tax, and life-cycle issues.

4. Current Issues

The real estate industry is highly cyclical. Because development can take from 1 to 5 or more years to complete a new project, depending on the difficulty of obtaining public approvals, buildings are often started when the market is hot, and completed when it is not. The real estate industry collapsed in the late 1980s–early 1990s in the deepest real estate recession since the Great Depression. The collapse was caused by a number of factors including over-building, tax policy that contributed to over-investment, too much careless money, and a major recession in the general economy due to downsizing of many corporations and national defense. Developers and investors who bought property at bargain prices in the early 1990s have made fortunes as rents and prices have recovered and surpassed pre-recession peaks. At the beginning of the twenty-first century, real estate markets in the USA are relatively 'in balance'—the supply of new space is in balance with demand for new space. Cycles, however, are endemic to real estate (Grenadier 1995). Those who learn how to play them well are the ones who will be most profitable.

Sects. 4.1–4.5 detail some of the most important issues in real estate development at the beginning of the twenty-first century.

4.1 Institutionalization of the Industry

During the recession of the late 1980s–early 1990s, many developers went bankrupt, along with the Savings and Loan Industry and many other financial institutions. Two powerful forces emerged from the recession—a vast expansion in the presence of Wall Street in real estate financing, primarily through securitization and Commercial Mortgage-Backed Securities, and the re-emergence of Real Estate Investment Trusts (REITs) as a major owner of property (Sanders 1998). REITs grew to more than \$140 billion in stock market capitalization—exceeding the 1 percent barrier of total market capitalization that represents the point at which Wall Street analysts pay serious attention to an industry. While REITs and other publicly financed companies, especially national homebuilders, will play a more important role in real

estate development and ownership than before, entrepreneurial developers will continue to play a significant if less dominant role in the creation of new product.

4.2 Future of Cities

Central Business Districts and inner-ring suburbs have gone through ups and downs over the last 50 years as many companies and higher-income residents have moved to outer-ring suburbs and smaller towns at the urban fringe. The constant outward migration has created enormous expansion of rapidly growing cities such as Atlanta, Phoenix, Dallas, Denver, Los Angeles, San Francisco, and Seattle. The outward growth has put pressure on open space and natural resources and placed demand on the public sector to build highways and other infrastructure to serve the new development.

At the same time, central cities and inner suburbs have suffered under declining municipal tax revenues and increasing demands for services from a poorer population, often augmented by new immigrants. While many cities such as Manhattan and Los Angeles are seeing signs of a return to the CBD by investors and residents, outward migration is likely to continue. Urban containment boundaries are a popular though controversial, technique for containing outward expansion. Developers have historically tended to prefer green-field suburban sites because they are easier to build, easier to finance, and more attractive to tenants and homeowners. Whether central cities and inner suburbs will be able to attract new investment and redevelopment is one of their most important challenges.

4.3 Smart Growth/Urban Sprawl

The rapid expansion of suburban growth in the last half of the twentieth century has fueled a backlash against development and *NIMBYism* (not in my backyard). 'Smart growth' is the latest approach to solving the problems associated with urban sprawl—congestion, loss of farmland, lack of functional open space, monotonous and boring development, and inadequate infrastructure (see Gordon et al. 1989, Peiser 1989).

4.4 Public–Private Development

As the regulatory process for major development projects has become more time consuming and costly, more developers are joining forces with cities, redevelopment agencies, school districts, armed forces, and other government agencies to develop property. The public–private partnerships help developers to obtain lower cost financing and necessary public approvals

while giving public agencies a share in the profits and the ability to deliver more public benefits such as parks, plazas, roads, sewer plants, and other infrastructure (Kotin and Peiser 1997).

4.5 Technology

The internet and other new technologies are having dramatic impacts on all aspects of real estate, from office building construction to shopping center functionality resulting from Internet commerce. As old shopping centers and industrial buildings become obsolete faster, new industrial distribution centers and 'entertainment' shopping centers are replacing them. While home offices and telecommuting have caused alarm for future office space demand and spawned new work patterns such as 'hot desks' (where employees share office space, plugging their lap top into the company computer network as they move from office to office) workers still appear to want the 'social' aspects of coming to work. Suburban corporate campuses with high tech communications, on-site recreation and sporting facilities, and high security are being built to respond to new workplace needs and employee preferences.

5. Social Responsibility

Real estate development touches on every aspect of people's lives. It determines not only the quality of the built-form environment, but also affects the life-style and social wellbeing of residents and workers in every community. What developers build may last a lifetime or longer. Because of their far-reaching role, developers bear a special responsibility. Most developers care deeply about the quality of their projects. High quality development depends as much on public sector support and services as it does on the private activities and financial resources of the developer. Since real estate involves very long-term investment, the long-term stability of the community and continuing attractiveness to new residents and firms is essential to long-term profitability for the developer. While developers are often vilified for destroying the environment and hastening neighborhood change, good development is responsive to neighborhood concerns and is sensitive to maintaining and improving environmental conditions. As public funds have been reduced for open space infrastructure improvements, development is often the instrument through which open space is preserved and new infrastructure is built. As their role expands, developers need to be ever more knowledgeable about the urban context in which they build and about the multitude of forces with which they must deal. The Urban Land Institute's mission statement sums up the developer's charge: 'to provide responsible leadership in the use of land in order to enhance the total environment.'

See also: Development and Urbanization; Neighborhood Revitalization and Community Development; Planning Ethics; Planning Issues and Sustainable Development; Planning, Politics of; Property Rights; Urban Geography; Urban Growth Models; Urban Planning; Central City Revitalization; Urban Sprawl

Bibliography

- Brueggeman W, Fisher J 1997 *Real Estate Finance and Investments*. Irwin/McGraw-Hill, New York
- Case K E, Shiller, R J 1989 The efficiency of the market for single-family homes. *American Economic Review* **79**: 125–37
- Corgel J, Smith H C, Ling D 1998 *Real Estate Perspectives*. Irwin/McGraw-Hill, New York
- DiPasquale D, Wheaton W 1996 *Economics and Real Estate Markets*. Prentice Hall, London
- Garvin A 1996 *The American City: What Works, What Doesn't*. McGraw-Hill, New York
- Gordon P, Kumar A, Richardson H 1989 Influence of metropolitan structure on commuting time. *Journal of Urban Economics* **26**(2): 138–51
- Grenadier S R 1995 The persistence of real estate cycles. *Journal of Real Estate Finance & Economics* **10**(2): 95–119
- Kotin A, Peiser R 1997 Public-private joint ventures for high volume retailers: who benefits? *Urban Studies* **34**(12): 1971–86
- Mills E S 1972 *Studies in the Structure of the Urban Economy, Resources for the Future*. The Johns Hopkins Press, Baltimore
- Peiser R 1989 Density and urban sprawl. *Land Economics* **65**(3): 194–204
- Peiser R 1990 Who plans America: Planners or developers? *Journal of the American Planning Association* **Autumn**: 496–503
- Peiser R 1992 *Professional Real Estate Development: The ULI Guide to the Business*. The Urban Land Institute, Washington, DC and Dearborn, Chicago
- Sanders A B 1998 The historical behavior of REIT returns: A capital markets perspective. Garrigan R T, Parsons F C (eds.) *Real Estate Investment Trusts*. pp. 277–305
- Urban Land Institute 1997 *America's Real Estate*. Washington, DC

R. Peiser

Realism/Neorealism

Realism is a tenet that views the world as it is, and not in terms of an unrealized ideal. It emphasizes the constraints of the real world and the need for pragmatism and prudence. Applied to international politics, realism sees a conflict-ridden world of states concerned pre-eminently with their security and pursuing power as the means to assure their survival. Realists therefore see the distribution of power as the major determinant of state behavior and international stability. Within this broad rubric, realists have developed alternative formulations, emphases, and distinctions.

1. The Realist Vision

The realist vision of international politics holds that states interact in an anarchical international environment in which survival is uncertain. Because power is the key requisite for survival, states define their interests in terms of power and constantly assess their strength relative to that of others. What emerges from their interaction is a balance of power that is the sole basis for order, regularity, and stability in international politics.

Realism emphasizes the competitive and conflictual nature of international politics. States either constantly prepare for war or are at war with one another. Fearing adverse shifts in the balance of power, they respond with internal mobilization and, should that prove inadequate, with external alignment. Self-reliance is preferred in a system of self-help, but allying with others to confront a common threat may be temporarily necessary. Cooperation is thus short-lived, lasting only as long as the common threat, and is a function of common interests and not common values or ideology. Realism can therefore readily explain alliances, such as the Nazi–Soviet pact of 1939, between wholly antithetical political systems. The result, mobilization and alignment in response to adverse shifts in the balance of power, returns the system to equilibrium. Hence, balances of power recur and reemerge in international politics.

1.1 Realist Assumptions

Realism is a statist perspective, for it assumes states to be the key actors in the international system. It excludes multinational corporations, nongovernmental organizations, and a variety of transnational and subnational entities that matter only to the extent that they affect states.

The theory assumes that states face recurrent threats and challenges in the international system. Power, especially military power, is the basis of both threats to a state and to its ability to assure its survival.

Realism presumes survival and its requisites to be the highest priority of states, indeed, their very reason for being (*raison d'état*). States value security considerations above all and respond to the imperatives of the international system. The primacy of security also implies that other policies, including foreign economic ones, reflect considerations of international power politics.

1.2 Realism Deemphasizes

For realists, international law and international organizations are largely irrelevant to international politics. Both reflect the interests of powerful states, for in international politics might makes right, and international institutions do not constrain great powers but merely provide additional venues for the

exercise of power. At most, both play a role only in matters of low politics, in regulating transportation, communication, and economic exchange, for example, and not in matters of high politics, especially military security.

Realists also consider the requisites of survival to dominate any moral considerations. Indeed, realists transform necessity into a moral imperative.

Realism downplays the role of domestic politics. It treats states as unitary actors that are functionally equivalent to one another, and the constraints imposed by the international system generate similar behavior even by dissimilar states. The international system once consisted largely of monarchies and then of authoritarian and military governments, and is now increasingly composed of democratic states. For realists, however, it continues to function in the same way.

Self-reliance precludes states from allowing the development of any division of labor or any interdependence regarding the requisites of state survival. Small states may have no choice but to rely on others, but great powers do have such choices and thus do not allow the emergence of any meaningful interdependence.

1.3 Realism as Equilibrium Theory

The balance of power is an equilibrium theory. The security policies of nations result in a balance of power among states. The equilibrium can be disturbed, but only by exogenous shocks, ones outside the model. The disturbance then sets in motion forces that return the system to a new equilibrium. The implosion of the Soviet Union, for example, disturbed the balance of international power, but was an event outside the model and inexplicable by it. But realism does purport to explain the set of forces that then would come into play to return the system to equilibrium.

Put this way, balance of power theory functions like Adam Smith's invisible hand in creating an unintended outcome. For Smith, the self-interested behavior of producers in a competitive market generated the best price for consumers. In international politics, the pursuit of national interest and the struggle for power in an anarchical system create a peaceful stable order. It is not the search for peace but the search for security that creates this stability.

1.4 Realism—its Historical Roots

Realism traces its lineage to Thucydides, who argued that his *History of the Peloponnesian War* (1996) provided the one and true explanation of the war: the growth of Athenian power and the fear this struck in Sparta. Thucydides argued that preceding incidents and specific complaints notwithstanding, the true cause was a shift in the balance of power.

As important to current realist formulations is the work of Thomas Hobbes, who translated Thucydides into English and argued (in *Leviathan*) that the state of nature, which he characterized as a war of all against all, is one in which war and preparation for war are a constant. Conflict is limited within societies by the creation of the Leviathan, a coercive state. For modern realists, who distinguish international from domestic politics, international relations remains an anarchic realm; there is no Leviathan and the war of all against all endures.

Realism dominated the study of international relations in the middle of the twentieth century as the world came to grips with the very occurrence of the Second World War and the failure of both international law and international organizations to prevent it. Hans Morgenthau's (1948) influential formulation of realism drew upon psychology and rooted realism in a human nature presumed to be inherently aggressive and in a view of individuals as power seeking.

Beginning in the 1950s and increasingly in the 1970s, realists (led by Waltz 1959, 1979) linked their arguments to systems theory and structuralism. This variant rejected basing realism in any theory of human nature or any internalized commitment on a balance of power. This neorealist formulation did not posit any individual or collective propensity toward power maximization *per se*. Rather, it posited a minimalist concern with survival in an anarchic setting as the only necessary conditions. Even good people and good states caught in an anarchic structure have no choice but to arm themselves in preparation for the possibility of war and must assiduously guard their relative power. Neorealism or structural realism firmly rooted the problem of international politics in the international system and not in the nature of individuals or states.

2. Disagreements

Despite the consensus in emphasizing states as the central actors, security as their prime concern, and power as their key instrumentality, there is some confusion and disagreement in the core of realist thought. First, there is confusion whether the theory is positive or normative. Realists typically argue that theirs is a positive theory of how the system actually works (an iron law of international politics), yet many have also used realist precepts to criticize the foreign policies of specific countries. Realists have argued, for example, that states do only what is in their national interest but have criticized specific actions, such as US intervention in Vietnam, as not being in the national interest.

That politicians adopted the argument has also led to confusion. First, the term balance of power has come to be used loosely and applied to virtually every configuration of power. Second, there is a disagree-

ment between scholars as to whether realism operates through intention (states must aim for a balance of power) or irrespective of it (as an unintended consequence).

Despite agreement that the distribution of power determines stability, scholars disagree about the relative stability of different configurations of power. Within the balance of power tradition, some scholars find bipolarity more stable, others find multipolar systems more stable. Although most agree that the international system tends towards balance (though more predictably and more quickly in some settings than others) and that hegemony and unipolarity are not possible or sustainable, some hold that the system tends towards power preponderance rather than balance, and that hegemony is stable but that a balance of power is not.

3. Problems

Realism is criticized for leaving too much out. A pristine realism that only says that balances recur cannot explain most foreign policies of most states most of the time. Moreover, the theory is static and cannot explain change over time. Important transformations of recent centuries, such as the rise of nationalism, the industrial revolution, the rise of ideological politics, the spread of democracy, the revolutions in communication, transportation, and warfare, realists judge irrelevant to international politics, which they treat as unchanged as long as anarchy continues to characterize international politics. Realists deem these factors to matter only to the extent that they affect the relative power of states (whose alteration is the only source of change in international relations) and are viewed as exogenous shocks. The only transformation that some realists accept as affecting world politics is the nuclear revolution, which they consider to undercut fundamentally the importance of relative power in a classical military sense.

The theory is also vague and imprecise. It is not clear what factors enter into a calculation of the balance of power, nor how balance is ascertained. It is difficult to test assertions of the theory when there can be wide disagreement over whether a particular distribution of power is balanced or not. Scholars dispute whether stability during the Cold War derived from US military superiority, a balance of power between the US and USSR, or the existence of nuclear weapons.

More important, the theory is indeterminate. It cannot predict which particular distribution of power will emerge, only that balance will. Nor can it predict the process by which balance is restored. US containment of the Soviet Union during the Cold War can be explained as the working of the balance of power, but so can the US roll back of German and Japanese

power during World War II. A US decision simply to contain German and Japanese power in the 1940s would have been as consistent with realism as the actual US decision to wage war and undo the consequences of German and Japanese aggression.

The theory makes precise predictions only when the constraints of the international system are presumed to be tight. Although the theory holds that it is great powers that structure global politics, the international system is most constraining of small powers and least constraining of large ones. Hence, the theory is typically indeterminate regarding the behavior of the great powers, which are presumed to matter most for international politics.

Formal studies have shown that realist premises are consistent with international cooperation, the construction of international organizations, and even the emergence of international norms. These works challenge the deductive logic of realism, arguing that the premises are insufficient to lead to realist conclusions.

Realists have had some difficulty in recent years in explaining significant international developments. The disjuncture between military and economic power has been sustained for an extended period of time. During the later stages of the Cold War, the world remained militarily bipolar, dominated by the might of the United States and the Soviet Union. Yet economically the world was multipolar, with Germany (and the EU more broadly) and Japan as significant economic powers that did not translate their economic prowess into military strength. Such sustained disjuncture between the distribution of economic and military power is problematic for realism.

4. Modifications

Realist thought continues to develop, in part to deal with such anomalies. One variant now posits that it is the perception of power rather than the actual distribution of power that matters. Another decouples changes in the distribution of power from the perception of threat, and holds that states respond to threats and not just shifts in power. Some see realism as characterizing particular eras rather than international politics generally. Eras in which territorial acquisition brings greater rewards than does commercial exchange are ones to which the logic of realism is more applicable; eras in which the nature of warfare gives the offense an edge are ones in which a realist logic is more likely to prevail.

Realism has a core strength: its emphasis on power and interest and an unregulated international system. Yet its most pristine formulation is rejected by many, if not most, scholars who articulate their positions relative to a core realist argument.

See also: Balance of Power, History of; Balance of Power: Political; Cold War, The; Conflict and War,

Archaeology of; Conflict Sociology; Geopolitics; Globalization: Political Aspects; National Security Studies and War Potential of Nations; Political Science: Overview; State, History of

Bibliography

- Baldwin D A (ed.) 1993 *Neorealism and Neoliberalism: The Contemporary Debate*. Columbia University Press, New York
- Brown M E, Lynn-Jones S M, Miller S E (eds.) 1995 *The Perils of Anarchy: Contemporary Realism and International Security*. MIT Press, Cambridge, MA
- Doyle M W 1997 *Ways of War and Peace: Realism, Liberalism, and Socialism*. Norton, New York
- Frankel B (ed.) 1996 *Realism: Restatements and Renewal*. Cass, London
- Jervis R 1978 Cooperation under the security dilemma. *World Politics* 30: 167–214
- Jervis R 1997 *System Effects: Complexity in Political and Social Life*. Princeton University Press, Princeton, NJ
- Keohane R O (ed.) 1986 *Neorealism and its Critics*. Columbia University Press, New York
- Morgenthau H J 1948 *Politics Among Nations: The Struggle for Power and Peace*. Knopf, New York
- Stein A A 1990 *Why Nations Cooperate: Circumstance and Choice in International Relations*. Cornell University Press, Ithaca, NY
- Thucydides 1996 *The Landmark Thucydides: A Comprehensive Guide to the Peloponnesian War*. [trans. Crawley G; Strassler R B (ed.)]. Free Press, New York
- Waltz K N 1959 *Man, the State and War: A Theoretical Analysis*. Columbia University Press, New York
- Waltz K N 1979 *Theory of International Politics*. Addison-Wesley, Menlo Park, CA

A. Stein

Realisms and their Opponents: Philosophical Aspects

In everyday usage, ‘realism’ is often used as a name for a practically or epistemically low-ambition attitude, while ‘idealism’ is often taken to denote a high-ambition—if not utopian—attitude. In philosophical usage, mostly, it is the other way around: those who are called realists tend to claim more than their opponents—they are the philosophical optimists.

Within philosophy itself, ‘realism’ adopts a variety of interrelated and contested meanings. It is used as the name for doctrines about issues such as perceptual access to reality, the existence of universals, the goals and achievements of science, the nature of truth, the objectivity of morality, and many other things. Given this variety, no single shorthand definition of the term ‘realism’ can be provided.

One manifestation of this variety is that the opponents of realism about these issues are not called

uniformly by a single label (other than the rather uninformative ‘antirealism’): the labels used include idealism, phenomenism, instrumentalism, conventionalism, fictionalism, noncognitivism, constructivism, relativism, irrealism, and others. Sometimes such labels vary from one dispute or domain to another even when similar ideas are being expressed. Sometimes they are used to make claims that are specific to the domain at hand and distinct from other antirealist theses.

In what follows a brief tour will be taken along some of the representative philosophical highways and lanes, from ontology through semantics to epistemology. Along the way, realism will be confronted with some of its various opponents and internal conflicts. The philosophical landscape in the neighborhood of realism is nowadays much broader than it used to be: in a typical older encyclopedia, ‘realism’ was taken to name doctrines about universals or about perception, or both, to the exclusion of much of its present coverage, and many of the current opponents were missing from view altogether.

1. Ontology

As an ontological doctrine, realism is a claim about the existence of something. Versions of ontological realism differ from one another in regard to what this something is—the domain of existents (things that exist)—and what kind of claim is being made about items in this domain. If we take ontological realism to have the form of the statement ‘*X* exists,’ then in order to identify different versions of realism and its denial we need to examine the various ways in which its two components are interpreted. We need to consider the two questions: ‘what exists?’ and ‘what is existence?’

If we just claim that the world exists, without providing further specifications of the constituents and structure of the world, our realism will remain rather uninteresting. Philosophers of the nominalist persuasion add that particulars, and nothing but particulars in time and space, exist. The world consists of objects such as you and me and the particular copy of this encyclopedia that you are reading now. A radical nominalist will claim that only ‘bare’ or propertyless particulars exist, while properties do not. Whereas nominalists can be realists of sorts, they are opposed by those who have been called the realists in the debate over the existence of universals. In this debate, realists claim that universals—kinds, properties, relations—exist, that they are, or are among, the constituents of the world. Not only does this particular encyclopedia exist, but among the furniture of the world are books as a cultural kind, properties such as weight and colour, and the relation of being owned by someone. Such universals can be multiply instantiated in particular objects in space and time. Platonic realism takes universals to constitute reality: redness, ‘book-

ness' and ownership inhabit a separate realm of ideas, independent of the particulars that exemplify them. Aristotelean realism takes both universals and particulars to be real, and claims that universals only exist when exemplified in particulars. In the philosophy of mathematics, the view that numbers, real-valued functions, and sets exist as 'abstract entities' is called Platonism, while the opponents denying their existence are known as intuitionists, formalists, and logicians.

Common sense takes the world to be furnished with medium-sized material objects that are the possible objects of sense perception. Considering these objects to exist leads to realist doctrines about the ontology of perception. The opponents of realism in the theory of perception include the idealists (to whom it is 'ideas' that constitute the world) and phenomenologists (who try to construe the world out of 'sense data'). *Direct realism* says that perception is a direct awareness of material objects that exist in reality and that nothing else exists between perception and perceptible objects. Naïve realism, entertained by philosophers as a target of criticism, is a version of direct realism that takes sensible qualities to be the intrinsic properties of material objects and these objects to have all the properties they are perceived as having. *Indirect realism* states that perception is directly about mental representations (such as bodily sensations and after-images) and only indirectly about the external world and that both the direct and indirect objects of perception exist.

An encompassing commonsense realism takes objects in the commonsense domain, whether material or mental, to exist. Its list of existents includes my bathroom and my belief that it is vacant, as well as ducks and the desire to eat one. Anyone ontologically committed to folk psychology, holding the commonsense view that beliefs, desires, hopes, and fears exist, is a realist about the mental. Common sense also seems to suggest that there are moral facts about the world. To say that something is morally good or morally bad is to describe aspects of objective reality. Moral realists deny a sharp distinction between facts and values, holding that moral values are among the facts of commonsense experience. In contrast to these views, an eliminative materialist denies that mental and moral facts exist and takes only matter or physical objects to exist, thus counting as a realist about the material but not about the mental or the moral.

Scientific theories characteristically discuss unobservable entities such as quarks and electromagnetic fields. If we take these to exist, we hold the ontological position of *scientific realism*. The opponents of realism about the ontology of scientific theories are the phenomenologists and fictionalists. For them, unobservables do not exist—only observables do. Rather than being among the fully authorized inhabitants of the real world, the unobservables of science are nothing but fictions of imagination in the scientists' minds.

One may combine scientific realism with a commonsense realism that takes both scientific objects and the objects of folk physics, folk psychology, morality, and folk sociology to exist: the full set of existents includes plans and plaitiveness, plantations and plebeians, plants and planets, platelets and plastids. This is disputed by a radical scientific realism that holds that only scientific objects exist; thus, trees do not exist, but the atomic and cellular structures that constitute them do. A radical commonsense realist will deny existence to scientific objects while granting existence to commonsense objects: plants exist while plastids do not. One can also be more discriminate among the various classes of commonsense objects, granting or denying existence to material, mental, and moral facts in different permutations, for example, holding that only the material exists; or that the material and the mental exist; or that all three exist.

One's notion of scientific realism has to be qualified in relation to specific scientific disciplines. The standard formulations are in terms of unobservable entities that lie far beyond the commonsense domain, with physics serving as the paradigm science. But suppose one wants to be a scientific realist about the explanatory posits of a sociological theory that discusses community and custom, role and authority, motive and reciprocity? If one takes these commonsense items to exist, the scientific realism that ensues will not be strictly modeled after the realism about photons and neutrinos and other noncommonsense posits of physical theories.

One can also choose to be a realist about modality. Modal realists argue that science and common sense deal with possibilities and necessities as aspects of the world. In decision theory, the decision-maker is viewed as considering possible choices and possible outcomes with probability values. In experimental science, the experimentalist considers various possible designs and possible outcomes with estimated probabilities. A modal realist takes such possibilities as real. An actualist denies the existence of possibilities and only regards actual things in a realist fashion. Yet, among the actual existents (in contrast to merely possible), one may include dispositions, propensities, capacities, and causal powers: they are potentialities that, under certain conditions, manifest themselves.

Having thus listed some possible domains of existents, and having made a selection from among them, the realist might leave it there. Of some *X*, the realist about *X* just says that *X* exists. Yet, pressed by antirealist arguments, the realist may feel forced to take a few more steps. This is because antirealists can go along with the realist to accept that *X* exists, but just interpret the meaning of '*X* exists' differently. Thus the realist has to raise the further question of what is claimed about the existence of the existents, or what is meant by '*X* exists'.

This question has a number of variants. One is about whether one reads existence claims literally or

reductively. A phenomenalist may say that electrons and elm trees exist, and add that what she means to say is that certain constellations of sense data exist; claims about electrons and elm trees and other such entities can be reduced to claims about sensory experiences. A realist would say that electrons and elm trees exist, period—signaling that the claim is to be understood literally. An eliminative materialist may say that a given mental property exists, and then add that this is just another way of saying that a certain neurophysiological configuration exists. A mental realist would resist such a reduction. In general, when considering the existence of *X*, the issue is whether the statement '*X* exists' is to be taken literally or to be reduced to another claim: '*Y* exists.'

Once we have fixed a literal answer to the question of what is claimed about the existence of the existents, another version of the question can be raised. This concerns the possibility and actuality of existence. The weakest realist claim is that it is *possible* for *X* to exist: the claim that *X* exists is intelligible and coherent. One is a realist about *X* not by claiming that *X* actually exists, but by claiming that *X* is the sort of thing that might exist. This is consistent with having doubts about its actual existence or a later discovery that it does not exist after all.

Many realists make stronger claims than just possible existence. They include statements about *actual* existence in their definition of realism. There are different versions of this, depending on the quantifier that is used in relation to the sets of items that are claimed to exist. No realist would like to claim that *all* alleged universals, particulars, commonsense objects, and/or scientific objects exist (this would imply that unicorns and ether are as real as magnolia trees and DNA molecules). Many realists would say that *some* of them, or *most* of them, exist.

Another version of the question (of what is claimed about the existence of existents) arises as we recognize that such claims are customarily formulated in terms of independence. Such claims take on the form, '*X* exists independently of *Z*'. Two further questions arise from such a claim: 'what is *Z*?' and 'what is independence?'

The ordinary formulation is that something exists mind-independently or independently of the mental, with 'mind' and 'mental' designating facts about the human mind. The paradigm meaning of 'exists' is based on an image of physical reality that existed before there were any humans around. Physical particles and their systems, such as molecules and galaxies as well as natural laws, existed independently of human perception and thought, and now that humans are around, these things continue to so exist—and were human life to disappear, they would still exist. One issue is whether 'exists' in this paradigm meaning applies to anything at all. Another issue is whether it is appropriately general to be extended to cover all important cases of existence claims, including

those about things that have emerged along with or since the arrival of humankind into the world. But what is independence?

If we take it to mean logical independence, we may want to say that physical objects are logically mind-independent since statements about them do not imply the existence of any human minds. This will require reading '*X*' literally as denoting physical objects: otherwise it would be open to a nonrealist to argue that *X*'s existence is logically mind-dependent for the simple reason that '*X*' is just another name for a bundle of sense data or some such. But if '*X*' denotes a human mind or a mental property itself, we cannot say that its existence is logically independent of the mental, because it entails the existence of the mental.

We may also say that something exists in the sense that it is causally independent of the mental: the human mind does not causally produce or reproduce it. Were this to apply to physical objects, 'causal production' itself would have to be literally understood as denoting a real connection in the world rather than a function of our minds. Whether it applies to human minds themselves depends on whether we take the reproduction of the species of sentient creatures like us to be causally mind-dependent activity. But it certainly does not apply to the many things in the world that are causally mind-dependent, such as material artifacts and social institutions.

Other notions of independence may be more successful in taking care of logically or causally mind-dependent things. We may choose to say that *X* exists independently of its recognition or any evidence we may have for its existence, or that it exists independently of any particular inquiry addressing it, or that it exists independently of any particular linguistic or other representation of it. Such formulations may enable realism about mental matters as well as about artifacts and many social entities that are causally dependent on the mental. The question of whether a given mental property or social entity exists is an objective question in the sense that it is separate from the question of whether anybody has thoughts, beliefs, or representations about it, has inquired into it, or has evidence for it.

Many realists hold the view that the world has a definite structure independently of our conceptual accounts of it. One may think that the world consists of propertied and related objects, such as natural kinds that are linked with one another by causal relations, and that these causal relations have a specific character, such as involving causal powers, causal transmission, and natural necessity. A nominalist would object to this by arguing that the world itself has no intrinsic structure, even though it may exist as an unstructured collection of propertyless particulars. Whatever structure the world is taken to have, it is rather our conceptualizations that impose it upon the world. The world is conceptually constructed so as to grant it a structure, and the origins of that structure

are in our ideas and conversations. This gives us an idealist and constructivist view on this matter. Thinkers are capable of 'worldmaking' as Nelson Goodman's 'irrealism' puts it, or of the 'social construction of reality' as the radical wing of social constructivism alleges. Insofar as our ideas are permitted to legitimately vary, this results in a form of ontological relativism as well: to each different communally-held or justified conceptualization corresponds a different structure of the world. Thomas Kuhn's suggestion that after a scientific revolution from one paradigm to another, scientists 'live in a different world,' can be—contestably—read as implying this idea.

The social world poses a special challenge to the realist: the social world *is* constructed and *is* dependent on people's ideas. There can be no communities and conventions, authorities and audiences without people acting on their ideas. Yet the realist wants to say they exist. These things may be taken to exist independently of, say, scientific theories about them, in the sense that the construction of a theory about a social entity is not a matter of constructing the entity.

2. *Semantics*

Ontological realism is about the existence of *X*. One may also raise questions about linguistic expressions '*X*' and '*X* is such-and-such' in relation to *X*. Indeed, 'realism' has become increasingly defined, especially by many antirealists, as a name for certain semantic views concerning such things as reference, meaning, and truth. It is characteristic of many antirealists to take semantic issues pertaining to language as primary, whereas realists often give priority to ontology and view semantic theses as derivative of, or motivated by, ontological positions.

The claim that scientific theories and the terms they include refer to real existents is part of the semantic thesis of scientific realism. Another part is the view that at least some of the major sentences contained in scientific theories can be used to make genuine statements about the real world, and as such, they may be true or may be false. More generally, semantic realism about some subject matter or domain maintains that statements about that domain are capable of being true or false—that they are capable of describing facts about the domain truly or falsely.

Some philosophers, such as Michael Dummett, have taken bivalence as a defining notion of realism: one is a realist about a statement or theory if one takes it to be either true or false. The ontological ramification of this is as follows: if there were gaps between these truth-values, this would indicate that the world itself is somehow indeterminate. If 'N is bald' is not clearly true or false but somewhere in between, then it would seem there is no determinate fact of the matter about N's baldness. This may be taken to motivate anti-realism. The realist should not reason along these

lines. If bivalence has a difficulty with statements like 'N is bald,' this may be nothing but an indication of the vagueness of the language of baldness, not a matter of some indeterminacy about N's head. For a realist, the properties of language have no implications for the properties of reality in such cases.

Realist theses about possible reference and possible truth in scientific theories are often complemented with claims about actual properties of actual science. One such claim—made by Richard Boyd, Michael Devitt and others—is that the theoretical terms of most current (or the best of 'mature') scientific theories typically refer and that their lawlike statements are at least approximately true. Another related claim is the convergence thesis: as science develops, its theories get progressively closer to the truth. Both of these are empirical claims about actual science and should not be made part of the concept of scientific realism. Their truth is dependent on contingent matters such as the institutional structure and other resources of scientific research as these happen to be in any given society and time period. This is not to deny that a realist may need a concept of approximate truth, truth-likeness, or closeness to the truth in order to defend the possibility of truth acquisition in science. Nor is it to deny the obvious idea that the notion of convergence towards the truth may be used to give content to the notion of scientific progress.

As for the concept of truth itself, a realist will hold the view that truth transcends any conceivable evidence. In this sense, truth is characterized negatively as nonepistemic, as being independent of whether, and how, we may have epistemic access to it. This runs counter to neopragmatist suggestions to define truth as provability (as does Dummett) or ideal acceptability (as does Hilary Putnam). For a realist, truth is one thing, and it is quite another to prove it, discover it, confirm our belief in it, or find evidential warrants for our beliefs. Since there is no necessary connection between truth and these things, the realist thinks that evidence-transcendent truth is not necessarily attainable. This implies that even if the epistemic warrants for a statement or theory were as good as possible, or in some sense ideal, our theory might be mistaken.

'Realism' is also used in connection with—but is not defined by—a specific view of what truth is: the correspondence account of truth. This view allows for many alternative formulations. A generic formulation is to say that a sentence (or proposition, or utterance, or whatever the appropriate truth-bearer is taken to be) is true in virtue of its correspondence with certain facts about the world, the truth-makers. This can be linked with ontological matters. On the one hand, truth presupposes an ontology within which truth-makers find a place. On the other hand, there is the route from an account of truth to ontology, from claims of truth to claims of existence. If it is up to the truth-makers to make the truth-bearers true, and if the truth-makers are objective facts about the world, then

the truth-makers of a true truth-bearer must exist. Thus, if mathematical statements can be true, then there must be mathematical objects of which they are true. If moral statements can be true, then there must be objective moral facts that make them so. Some realists have severed the connection between realism and the correspondence account of truth: they want to endorse the former without endorsing the latter. Most of these realists hold a deflationary or minimalist account of truth.

Nonrealists can also choose from different options. An instrumentalist about scientific theories can insist that no truth-values can be ascribed to theories as they make no claims about the world. Theories are no more than instruments for organizing the data, predicting future events, or helping to manipulate the world. They are tools of management or tickets of inference, and it is in the nature of tools that one should not ascribe truth or falsehood to them. Tools are characterized as useful or useless, convenient or inconvenient, relevant or irrelevant, powerful or powerless, and so on—all these attributions being relative to the purpose to which a given theory is put. This is the traditional version of instrumentalism that has been inspired by the familiar feature of physical theories that they appear to postulate unobservable entities. In case one has empiricist suspicions about such unobservables, it is tempting to construe theories as neither referring to them nor making any claims about them, whether true or false.

Likewise, in case one is ontologically suspicious about intentional properties, there is still an option to use the intentional vocabulary of folk psychology instrumentally: even though there really are no intentional properties on top of the physical ones, folk psychology can be usefully employed for predictive purposes provided it is construed nonreferentially. Daniel Dennett's 'intentional stance' serves as an example. One may adopt this stance as a perspective from which to treat systems *as* intentional without asserting that they really *are* intentional. No truth claims are being made about a system *really* possessing beliefs and intentions, but the predictive benefits created or lost by treating any given system as intentional is an objective matter.

Another instrumentalist possibility is to view a theory as a false tool, rather than as a truth-valueless tool. This is in line with Hans Vaihinger's fictionalism: falsehood is a characteristic of scientific theories regarded as fictions. Even though Vaihinger did not discriminate between natural and social sciences in applying his fictionalism—Isaac Newton's gravity and Adam Smith's economic man served as his examples—this view may be able to accommodate a feature of many social scientific theories: they do not seem to postulate unobservables in the same sense in which one finds them in physical sciences. Many social theories, and especially economic theories, involve idealizations and simplifications about familiar com-

monsense objects: immensely informed consumers with transitive preferences over all options of choice, and international economies with two countries with identical production functions, two goods, two fully-employed and internationally-immobile factors of production. Since many such idealizations are wide of the mark, one may be tempted to construe theories involving them as false, but nevertheless useful, instruments.

A popular view among the opponents of moral realism is to claim that apparent statements about some domain are not really genuine statements at all, but that they are rather used to express emotions or commitments or to persuade audiences. Thus, one may argue that moral claims of the form, '*X* is morally good' are not statements of some objective moral facts capable of being true or false, but express the utterer's emotions or persuasive intentions. This nonrealist position is called a noncognitivist position in contrast to its realist opponent, the cognitivist view, which takes moral claims to be truth-valued. This is an example of a reductive, nonliteral account: a descriptive claim that '*X* is good' is not supposed to be taken literally, but rather to be reduced to recommendations or approvals.

Yet another nonrealist possibility is to retain the vocabulary of truth and refute the correspondence account. Truth can be viewed as essentially involving pragmatic features, such as instrumental efficiency, ideal consensus, or warranted assertability. Whatever works for this or that purpose and meets certain constraints is true. These views make truth dependent on things such as our interests and goals, our thoughts and recognitional capabilities, our linguistic and other social conventions. Realists tend to think of truth as independent of such things.

Radical relativism regards all truths as indexical: '*P* is true in culture or community *C* but not in *D*,' '*P* is true for me but not for her,' and so on. Some formulations of constructivist sociology of scientific knowledge result in relativism of this sort by equating truth with acceptability within a scientific community. Equating truth with *ideal* acceptability is supposed to avoid the pitfalls of such relativism: the ideal conditions are supposed to fix truth uniquely. Realists claim to avoid them by not equating truth with anything epistemic.

3. Epistemology

Many of the issues above can be turned into issues of belief, justification, and knowledge. Some formulations of scientific realism refer to epistemic attitudes: scientists accepting a theory believe it to be true; and it is the goal of science to pursue true theories. Both formulations are problematic. A scientist with ontologically and semantically realist convictions may, with no inconsistency, accept for some purposes a

theory not believed to be true or not yet so believed. And whether or not truth is actually pursued as a goal is not implied by the notion that scientific theories include sentences that are capable of being true or false—or that their posits are capable of existence. It is also questionable whether science is the sort of entity to which goals can be attributed. Whether truth is among the goals of a particular body of scientists, and whether science as a collective endeavor will actually deliver truths as an outcome, depends on the actual prevailing institutional organization of science. Claims about scientists' actual beliefs, goals, and achievements are empirical claims whose truth depends on various contingencies; they should not be included in one's core definition of 'realism'. Of course, to accommodate such ideas, one may stipulate complex labels such as 'axiological veristic scientific realism,' which says that scientists had better pursue true accounts of the world.

Forms of epistemological realism are best formulated in terms of 'knowability'—of believability and justifiability. Supposing the world, and these or those of its constituents exist, can we acquire knowledge about them? Realist theories of perception share the view that there is no veil of perception, that perception is able to provide us with (perhaps very complex and indirect) access to at least some parts of reality. More generally, an epistemological realist holds that the world is knowable; for example, the theoretical claims of science may be justifiably believed to be (approximately) true.

With respect to scientific theories and the entities and relations they postulate, one may hold the view that the ontological question of their existence is irrelevant. They may exist, and theories may be true or false about them, but what is crucial is that we *cannot know* that they exist, nor that claims about them are true or are false. Therefore all that matters for science—insofar as its goals and criteria of success are concerned—is empirical adequacy, that is, the truth of the claims science makes about observable phenomena. This empiricist combination of ontological realism and epistemological antirealism has a long tradition, including Auguste Comte's positivism and, presently, Bas van Fraassen's constructive empiricism. People like Ernst Mach have been closer to taking the more radical step from epistemological to ontological antirealism, from denying epistemic access to the unobservable posits of science to denying their existence.

Many contemporary versions of antirealism, including Putnam's 'internal realism,' hold that insofar as truth is concerned, the semantic and epistemological issues cannot be kept distinct: truth is epistemic. Truth consists in warranted assertability, or in ideal acceptability. This means that a theory that is epistemically ideal cannot be false. An epistemically ideal theory is a fully justified theory: it is a product of our very best efforts in the most favorable circumstances, one that is

perfectly supported by all available evidence and meets all relevant theoretical constraints. While the realist will distinguish between the notions of being true and passing for truth in any circumstances, the antirealist will not acknowledge a difference between truth and passing for truth in ideal circumstances. The antirealist view makes truth conditions identical to the conditions of the recognition of truth, thus there is no recognition-independent or evidence-transcendent truth to be sought. Truth becomes epistemic; semantics and epistemology become fused. This may be taken to have ontological import. If an ideal theory suggests that quarks exist, then the statement 'quarks exist' is true, and therefore quarks exist. In this picture, ontic matters are derivative from epistemic matters.

This antirealist view is often presented as providing protection against skepticism about truth acquisition: making truth attainable by definition is supposed to help avoid the misery of truth possibly escaping us in all circumstances. Semantic realism decouples truth and justification by viewing truth as evidence-transcendent and otherwise nonepistemic. This may be taken to create room for skepticism: since even epistemically ideal theories may be false, we may be doomed to be mistaken forever, even though most realists would say that, as a matter of fact, we are not. Indeed, epistemological realists believe that we are able to acquire justified true beliefs about the world, thus they deny skepticism. Whether one formulates scientific realism as suggesting that most (or some) current scientific theories are recognizably and justifiably (approximately) true, or as the implied weaker view that it is in the character of scientific theories that they can be recognized as being true or false, one thereby subscribes to epistemological realism. Yet, given that truth and justification have been conceptually decoupled, the realist has to face the difficult problem of showing how these two things are related empirically so as to ensure that our cognitive efforts are somehow conducive to truth-acquisition.

The antirealist who holds truth to be epistemic is not necessarily much better equipped against skepticism. While actual justification will not guarantee truth, ideal justification does, or so some antirealists suggest. The challenge is to give a detailed account of what ideal justification consists of, and to do this in purely epistemic terms without invoking a prior grasp of the notion of truth itself. This may be difficult or impossible. If we say that ideal justification takes place in circumstances where all sources of error have been removed, and take this to mean circumstances where there are no obstacles to the attainment of truth, we will have failed to meet the challenge.

Bibliography

Armstrong D M 1978 *Universals and Scientific Realism*, Vols. I-II. Cambridge University Press, Cambridge, UK

- Cohen R (ed.) 1996 *Realism and Anti-realism in the Philosophy of Science*. Kluwer, Dordrecht, The Netherlands
- Cornman J W 1975 *Perception, Common Sense, and Science*. Yale University Press, New Haven, CT
- Devitt M 1984 *Realism and Truth*. 2nd edn. Princeton University Press, Princeton, NJ
- Hacking I 1983 *Representing and Intervening*. Cambridge University Press, Cambridge, UK
- Haldane J, Wright C (eds.) 1993 *Reality, Representation, and Projection*. Oxford University Press, Oxford, UK
- Hooker C A 1987 *A Realistic Theory of Science*. SUNY Press, Albany, NY
- Leplin J (ed.) 1984 *Scientific Realism*. University of California Press, Berkeley, CA
- Maddy P 1990 *Realism in Mathematics*. Clarendon Press, Oxford, UK
- McCormick P J 1996 *Starmaking. Realism, Anti-realism, and Irrationalism*. MIT Press, Cambridge, MA
- Niiniluoto I 1999 *Critical Scientific Realism*. Clarendon Press, Oxford, UK
- Psillos S 1999 *Scientific Realism. How Science Tracks Truth*. Routledge, London
- Putnam H 1981 *Reason, Truth and History*. Cambridge University Press, Cambridge, UK
- Sayre-McCord G 1988 *Essays on Moral Realism*. Cornell University Press, Ithaca, NY
- Sellars W 1963 *Science, Perception, and Reality*. Routledge, London
- van Fraassen B 1980 *The Scientific Image*. Clarendon Press, Oxford, UK

U. Mäki

Reasoning with Mental Models

Reasoning is the mental process of drawing a conclusion from a set of premises. The premises may be statements, perceptions, beliefs, or items of general knowledge. The conclusion may be a statement or a thought that guides action. Reasoning is a central component of human intelligence, and without it, there would be no laws, civilization, or science (see *Scientific Reasoning and Discovery, Cognitive Psychology of*). Laws would have no application because individuals would be unable to make the following sort of inference:

Pat has a license or else Pat is disqualified from driving;

Pat does not have a license;

∴ Pat is disqualified from driving.

This inference is a valid deduction, that is, its conclusion must be true given that its premises are true. Logic is the science of valid deductions, but not a theory of how people reason. Such theories are the province of psychology, and psychologists have proposed a variety of theories (see *Logic and Cognition, Psychology of*). Human beings can reason about topics for which they have no general knowledge, and theories of this ability fall into two main categories.

Some theorists postulate that reasoners rely unconsciously on formal rules of inference akin to those of a logical calculus (see *Natural Concepts, Psychology of*). The preceding inference depends on a rule of the form: A or else B, not A, therefore B. Other theorists propose that reasoners rely instead on their grasp of meaning and of principles akin to those for the semantics of a logical calculus. These theories rely on mental models, which are internal representations mirroring the structure of the external world. This article describes their role in reasoning and evidence corroborating it.

1. Mental Models

The idea that humans construct models of the external world goes back to the Scottish psychologist, Craik (see also *Mental Models, Psychology of*). He wrote:

If the organism carries a 'small-scale model' of external reality and of its own possible actions within its head, it is able to try out various alternatives, conclude which is the best of them, react to future situations before they arise, utilize the knowledge of past events in dealing with the present and the future, and in every way to react in a much fuller, safer, and more competent manner to the emergencies which face it (Craik 1943, Chap. 5).

Another antecedent is Wittgenstein's (1922) thesis that propositions represent reality in a similar way to pictures. What the modern theory of mental models adds to these programmatic proposals are three main assumptions:

(a) Each model represents a possibility. Thus, the assertion: either Pat has a license or else Pat is disqualified from driving calls for two models to represent the two possibilities (shown here in simplified form on separate horizontal lines):

License

Disqualified

where 'License' denotes a model of the possibility in which Pat has a license, and 'Disqualified' denotes a model of the possibility in which Pat is disqualified from driving.

(b) Models have a rich internal structure (not shown in the preceding diagram). Like an architect's plan, the parts of a model correspond to the parts of what it represents, and so the structure of a model corresponds to the structure of the world (Wittgenstein 1922, Propositions 2.13–2.17). Visual images are based on models, though many models are not visualizable (see *Imagery versus Propositional Reasoning*).

(c) The principle of truth: models normally represent what is true according to the premises, but not what is false. Hence, the preceding models of the disjunction represent only the possibilities that are true. Likewise, for true possibilities, models represent clauses in premises only when they are true in the possibility. For instance, the first model in the set above represents the possibility that Pat has a license,

but it does not represent explicitly that in this case it is false that Pat is disqualified from driving; likewise, the second model does not represent explicitly that in this case it is false that Pat has a license. Individuals make 'mental footnotes' about these matters, but normally they soon forget them. The principle of truth makes for parsimonious representations: reasoners do not have to bother with what is false.

Models are constructed as a result of perception, the comprehension of discourse, and imagination (see *Vision, Psychology of; Situation Model: Psychological*). They can represent real, hypothetical, or imaginary situations. They may also reside in long-term memory as a representation of knowledge (see *Mental Models, Psychology of*). They are accordingly a form of mental representation advocated by many psychologists as the way in which the mind represents reality, conceives alternative possibilities, and simulates the workings of the world. In principle, they could also underlie reasoning.

2. Deduction with Mental Models

The model theory of reasoning is due originally to the present author and his colleagues, notably the Irish psychologist Ruth Byrne (see Johnson-Laird and Byrne 1991, and for a recent alternative, Polk and Newell 1995). The theory postulates that models are based on the meaning of each premise, and that reasoners formulate a conclusion by describing something in the models that was not explicit in the premises. A conclusion that holds in all the models is necessary given the premises. A conclusion that holds in at least one model is possible given the premises. If models represent equiprobable possibilities—as reasoners often assume—the probability of a conclusion depends on the proportion of models in which it holds, as shown by the Italian psychologists Paolo Legrenzi, Maria Legrenzi, and Vittorio Girotto, and the French psychologist Jean-Paul Caverni (see Johnson-Laird et al. 1999). The model theory accordingly unifies logical reasoning, reasoning about possibilities, and probabilistic reasoning.

To illustrate the theory, consider again the inference:

Pat has a license or else Pat is disqualified from driving.

Pat does not have a license.

Therefore Pat is disqualified from driving.

The disjunctive premise calls for the models above. The second premise eliminates the model representing the possibility that Pat has a license. The remaining model is:

Disqualified

It yields the conclusion that Pat is disqualified from driving. This conclusion is a valid deduction because it holds in all the models—in this case, the single model—of the premises.

If individuals reason with models, then the more models that they have to construct—the more possibilities they have to envisage—the harder the task should be. Many experiments have corroborated this prediction (for a review including the studies described in this section, see Johnson-Laird 1999). For example, consider the following problem in spatial reasoning:

The cup is on the right of the knife.

The plate is on the left of knife.

The fork is in front of the plate.

The saucer is in front of the knife.

What is the relation between the fork and the saucer?

The theory predicts that reasoners should construct a two-dimensional model of the spatial layout of the items of the sort shown in the following plan:

plate	knife	cup
fork	saucer	

This model yields the conclusion: the fork is on the left of the saucer. In contrast, consider a problem in which the first premise is instead:

The knife is on the right of the cup

and the remaining premises are the same as above. This problem calls for two distinct models because the premises do not fix the relation between the plate and the cup:

plate	cup	knife	cup	plate	knife
fork		saucer		fork	saucer

The two models yield the same conclusion as before, but the problem should be harder because reasoners have to envisage two models. They do indeed draw a smaller percentage of correct conclusions to such multiple-model problems than to the one-model problems (Johnson-Laird and Byrne 1991).

Schaeken et al. have demonstrated similar differences between one-model and multiple-model problems about the temporal relations between events, using such premises as: John takes a shower before he drinks his coffee. Another group of Belgian psychologists (Vandierendonck et al.) have confirmed the difference in both temporal and spatial problems, and also shown that a task which preoccupies working memory for visual and spatial relations interferes with these inferences. Madruga et al. in Madrid have shown that inferences based on a conjunction, which calls for a single model, become harder when that premise is replaced by a disjunction, which calls for multiple models. Likewise, the Italian psychologists, Girotto et al. have shown that an inference based on a conditional premise of the form, If A then B, which calls for at least two models, becomes easier when a premise that eliminates one of these models is presented first rather than after the conditional premise. Two other Italian psychologists at Turin University, Bara and Bucciarelli, have shown that the difficulty of inferences known as syllogisms can also be predicted from the number of models that they require. Thus, the following syllogism is easy:

Some of the artists are beekeepers.

All of the beekeepers are chemists.

What follows?

Seven year-old children can draw the valid conclusion: some of the artists are chemists. In contrast, the next syllogism is hard even for adults:

None of the athletes is a banker.

All the bankers are chefs.

What follows?

Most reasoners construct a model in which the athletes are distinct from the chefs, and conclude that none of the athletes is a chef. To reach the correct conclusion, they need to realize that there could be chefs who are not bankers, and that each of the athletes could be such a chef. The only valid conclusion about chefs and athletes is accordingly that at least some of the chefs (i.e., those who are bankers) are not athletes. Few people reach this conclusion.

Another corroboration of the model theory comes from the study of children's understanding of assertions. Various experimenters, notably the French psychologists Barrouillet and Lecas, have demonstrated a trend in how children develop an understanding of conditional assertions of the form, If *A* then *B*. Young children treat them as though they were compatible with just a single model (in which both *A* and *B* occur), slightly older children treat them as compatible with two models (*A* and *B*, and not-*A* and not-*B*), and adolescents correctly treat them as compatible with three models (*A* and *B*, not-*A* and *B*, and not-*A* and not-*B*). Similarly, Sloutsky et al. at Ohio State University have observed that children tend to ignore the second clause of a disjunctive premise in order to ensure that the premise calls for just one model. This tendency to minimize the number of models has been corroborated by Ormerod et al. in their studies of how people paraphrase conditionals and other assertions.

A computer program implementing the model theory made a striking prediction. It showed that if reasoners abide by the principle of truth, then they should make systematic fallacies. These errors should arise because reasoners fail to represent what is false. Several recent studies have corroborated the occurrence of these 'illusory' inferences. Here is an example from an experiment carried out by Goldvarg and Johnson-Laird. Suppose that only one of the following assertions is true about a particular hand of cards:

There is a king in the hand or there is an ace in the hand, or both.

There is a queen in the hand or there is an ace in the hand, or both.

There is a jack in the hand or there is a ten in the hand, or both.

Is it possible that there is an ace in the hand?

In the experiment, 99 percent of the participants responded 'yes,' and the answer seems obvious. Yet, it is an illusion. It is impossible for an ace to be in the hand, because both of the first two assertions would then be true, contrary to the rubric that only one

assertion is true. Other experiments have corroborated the occurrence of illusions in causal and conditional reasoning, and Yang, a Chinese psychologist, has shown that they occur with syllogistic premises.

The model theory and theories based on formal rules, which were mentioned earlier, run in parallel for many inferences, but illusory inferences are a case where the two accounts diverge. The model theory predicts the illusions, but the formal rule theories rely solely on valid principles of inference and so they are unable to explain them.

3. *Models, Induction, and Reasoning in Daily Life*

Many of the inferences that you make in daily life are not deductive, that is, even granted the truth of your premises, there is no guarantee that your conclusion is true. You leap to a conclusion that goes beyond the information in your premises. The conclusion rules out some additional possibilities over and above those that the premises rule out, and so it may be false. By definition, such inferences are known as inductions.

The model theory extends naturally to inductive inferences including the informal inductions of daily life (see *Informal Reasoning, Psychology of*). Induction consists in the addition of information to models, sometimes with the consequence of eliminating a model (of a possibility). Often the process is hardly distinct from the normal business of making sense of the world. For instance, you use a voltmeter to examine an electrical circuit and you discover that the following proposition is true:

The battery is dead or the voltmeter is faulty (or both).

You then find out that:

The voltmeter is faulty.

And so you infer: the battery is not dead. Your inference is an induction, because your conclusion could be false even though your premises are true. Your disjunctive premise has three models:

dead

faulty

dead faulty

where 'dead' is a model of the battery as dead and 'faulty' is a model of the voltmeter as faulty. Your discovery that the voltmeter is faulty is consistent with two of these models, and in one of them, the battery is dead (contrary to your conclusion). Hence, your conclusion goes beyond your premises. Yet, it is highly plausible. You may have based it on an unstated belief that the chances are remote of both a dead battery and a faulty voltmeter. This belief leads to you to discount a model so that only the single model yielding your conclusion survives.

Reasoning in real life almost always depends on background knowledge—a knowledge of the meaning of words, of the world, and of the situation at hand.

This knowledge enables you to leap to plausible conclusions. In logic, when a conclusion follows validly from premises, no subsequent information can invalidate it. Logic is monotonic in that additional premises lead merely to additional conclusions, and never to the withdrawal of an earlier conclusion. Logic means never having to be sorry about a valid conclusion. In daily life, however, you do withdraw a conclusion when it conflicts with subsequent evidence. Sometimes, you do so because the conclusion was based on an assumption that you made by default—an assumption warranted only if no evidence exists to the contrary, for example, Quakers are pacifists. If a President is a Quaker, then you may infer that he is a pacifist, but when you learn that he is war-mongering, you withdraw this conclusion. The model theory allows for the withdrawal of the consequences of default assumptions (Johnson-Laird and Byrne 1991).

Many other sorts of inference in daily life lead to conflicts with reality. Suppose you know, for example:

If I turn on the ignition then the engine will start.

You turn on the ignition, and so it follows validly:

The engine will start.

Unfortunately, nothing happens. There is a conflict between a valid consequence of your beliefs and the facts of the matter. Something has to give. At the very least, you have to withdraw your conclusion. Researchers in artificial intelligence have developed various systems of nonmonotonic reasoning to try to deal with such cases (see Brewka et al. 1997), but no one knows what mental processes underlie human nonmonotonic reasoning. At its heart, however, there appears to be the generation of diagnostic possibilities. You try to imagine a scenario—a mental model—that explains why the engine did not start.

4. Conclusions

The theory of mental models postulates that human beings reason by imagining the circumstances described in the premises. They envisage mental models of these possibilities, focusing on what is true at the expense of what is false. Unlike other theories of reasoning, this principle correctly predicts the occurrence of certain systematic fallacies. The evidence accordingly confirms that human reasoners do rely on mental models. They construct them from perception, memory, and imagination, manipulate them to make inferences, and base their actions on the results.

See also: Imagery versus Propositional Reasoning; Informal Reasoning, Psychology of; Logic and Cognition, Psychology of; Mental Models, Psychology of; Practical Reasoning: Philosophical Aspects; Problem Solving and Reasoning: Case-based; Problem Solving and Reasoning, Psychology of; Problem Solving:

Deduction, Induction, and Analogical Reasoning; Scientific Reasoning and Discovery, Cognitive Psychology of

Bibliography

- Brewka G, Dix J, Konolige K 1997 *Nonmonotonic Reasoning: An Overview*. CLSI Publications, Stanford University, Stanford, CA
- Craik K J W 1943 *The Nature of Explanation*. Cambridge University Press, Cambridge, UK
- Evans J St B T, Newstead S E, Byrne R M J 1993 *Human Reasoning: The Psychology of Deduction*. Lawrence Erlbaum, Hove, UK
- Johnson-Laird P N 1983 *Mental Models: Towards a Cognitive Science of Language, Inference, and Consciousness*. Cambridge University Press, Cambridge, UK and Harvard University Press, Cambridge, MA
- Johnson-Laird P N 1999 Deductive reasoning. *Annual Review of Psychology* **50**: 109–35
- Johnson-Laird P N, Byrne R M J 1991 *Deduction*. Erlbaum, Hillsdale, NJ
- Johnson-Laird P N, Legrenzi P, Girotto V, Legrenzi M, Caverni J-P 1999 Naive probability: A mental model theory of extensional reasoning. *Psychological Review* **106**: 62–88
- Polk T A, Newell A 1995 Deduction as verbal reasoning. *Psychological Review* **102**: 533–66
- Web page on mental models and the bibliography maintained by Byrne et al. (http://www.tcd.ie/Psychology/Ruth_Byrne/mental_models/)
- Wittgenstein L 1922 *Tractatus Logico-Philosophicus*. Routledge & Kegan Paul, London

P. N. Johnson-Laird

Rechtsstaat (Rule of Law: German Perspective)

‘*Rechtsstaat*’ is a term of the German language and is not easily translated. ‘*Recht*’ stands for ‘law’ and ‘*Staat*’ for ‘state’; in English and most other languages grammar or semantics stand in the way of combining both elements into a compound. The theme evoked by the term ‘*Rechtsstaat*’ is often referred to as ‘rule of law,’ although the latter expression possesses a tradition of its own and carries connotations different from the German ‘*Rechtsstaat*.’

1. ‘Rechtsstaat’: an Introduction

Rechtsstaat concerns the age-old question of how to achieve order and freedom within a state. It connotes a balance between creating government authority powerful enough to keep peace both internally and

externally, to protect the state from violence or intervention on the one hand, and to secure a maximum of individual freedom on the other.

Rechtsstaat is a modern coinage which gained currency in German-speaking Europe in the early nineteenth century, when political thought from the French Revolution and the struggle for independence in British North America found wide support among the middle classes in Germany. English concepts for reining in the exercise of governmental power through law thus also count among the sources which influenced the evolution of '*Rechtsstaat*.' From this point of view, a '*Rechtsstaat*' is more than mere statehood characterized by an exercise of government power based on, and limited by, law. *Rechtsstaat* is closely connected to democracy, as democracy is a mode of legitimizing the exercise of power. If a democratic system is intended to be a *Rechtsstaat*, it needs to be governed by legal rules on the sources of such legitimacy.

The German-language term *Rechtsstaat* does not represent a phenomenon unique to Germany, but is the semantic and conceptual reaction to a fundamental issue of modern statehood. In the course of efforts to establish the rule of law in Germany, *Rechtsstaat* had been a political slogan to which various contents and political demands attached. It was only later that the term became textually included in written constitutions. With its many facets and unclear limits, it was misused during the Nazi period. After the end of Nazism, *Rechtsstaat* substantially informed constitutional thinking and is now considered to be a key element characterizing the German constitution of 1949, the Basic Law (*Grundgesetz*); its significance in that text will be further explored in Sect. 2.2. For the moment, we wish to concentrate on how and with what kind of implications *Rechtsstaat* describes a type or a model of state organization (see Sect. 2.1).

Counterbalancing authority and freedom, *Rechtsstaat* is a task not only for today's states, but also for organizations transcending the state level. This is true in particular of integrative processes that have progressed quite far, as in the case of the European Community or similar processes that have been launched in South America and other regions. In a wider perspective, it applies likewise to the task of consolidating the United Nations. Evidence for this may be seen in references to a European Constitution and in increasing references to a 'constitutional order' in international law, which has been considerably shaped by the United Nations (see Sect. 3).

2. The Modern Meaning of '*Rechtsstaat*': The State Level

The present-day meaning of *Rechtsstaat*—in general theory of the state—on the one hand stands for a category or type of state and, on the other hand, a legal

principle. We may speak of '*Rechtsstaat*' as a type, if the law of a state exhibits certain characteristics; these will be presented in Sect. 2.1. They derive above all from the constitution of the state in question and may be identified therein. It is immaterial in this regard whether the constitution is embodied textually in a document, as will most often be the case, or whether it is founded wholly or partly on unwritten law. The question of whether a given state may be considered a *Rechtsstaat* is a theoretical as well as a practical matter. It is a matter of theory insofar as political and legal doctrine seeks to comprehend and to differentiate hypothetical as well as existing states through systematic concepts. It is a matter of practice in the sense that the tradition of the term *Rechtsstaat* as a slogan continues to exert its influence: states whose laws wholly or partly fail to satisfy the requirements of *Rechtsstaat* will be forced to justify these failures in domestic political controversy and even in international relations. The more consensus exists on which core elements are necessary to a reasonable organization of a state which optimally balances the contrasting factors of effective government authority and individual freedom, the more will states be inclined, in their mutual relations, to demand that the concept of *Rechtsstaat* be made reality.

Traditional international law treated such demands as an illegal interference in internal affairs, contradicting the rule of the sovereign equality of states. The principle of non-interference in internal affairs has, however, been much eroded during the latter half of the last century. This concerned initially the area of fundamental human rights. International law today no longer categorically rules out demands from one state that another respect fundamental human rights. Moreover, there is a tendency to widen the ambit of human rights that are thus internationalized and to include aspects of domestic state organization: this pertains in particular to human rights which are directly related to democracy, such as freedom of expression. For example, states may make the establishment of economic or general political relations conditional on beneficiary states satisfying, in their municipal laws, the requirements of *Rechtsstaat*. While formal international recognition of another state used to be a merely declaratory act, this might change in future: recognition of another state is increasingly being predicated on substantive criteria. In the wake of the transformation in Europe during the last decade of the twentieth century, the Member States of the European Union, for example, demanded that the newly emerged states establish governmental structures of a *Rechtsstaat* type before the EU would put relations with these states on a formal footing.

While *Rechtsstaat* as a type covers all states, '*Rechtsstaat*' as a principle of constitutional law is confined to a particular state within whose constitutional law this principle may operate. *Rechtsstaat* as a principle is a legal norm. It governs state authority

and may serve as a yardstick in the jurisprudence of a constitutional court. This case is briefly described under Sect. 2.2 with reference to German constitutional law.

2.1 '*Rechtsstaat*' as a Type

Consensus may be assumed today that certain elements have to be present for a state to be considered a *Rechtsstaat*. Regarding aspects of state organization (so called formal *Rechtsstaat*) it is first and foremost the separation of powers, the independence of the judiciary, the need for administrative behavior to be law-based, and recourse to the courts against acts of government. These elements, however, are closely interrelated with substantive matters. A substantive *Rechtsstaat* refers to a state aiming at safeguarding justice through law-conforming actions of its legislative bodies, i.e., legislative enactments must respect higher-ranking constitutional law. More recently the notion of substantive *Rechtsstaat* has come to include the constitutional guarantee of individual rights, including fundamental and human rights. Central to all this is the need for government authority to be bound by law. Thus the distinction between formal and substantive *Rechtsstaat* is merely a rough one, and the following three aspects may better illustrate the commonly accepted constituent components of *Rechtsstaat*. *Rechtsstaat* formulates requirements for the make-up of a constitution; *Rechtsstaat* is a state whose objective is the realization of human dignity; and *Rechtsstaat* formulates requirements for the make-up of law in general (not merely constitutional law).

The idea of *Rechtsstaat* implies precedence of the constitution over other law. The legislative power is likewise bound by law, in the tradition of most states through a constitutional instrument, i.e., a written constitution, although this is not a necessary feature. Binding the legislative power through higher-ranking law does not mean that legislation to amend the constitution may not be passed. A *Rechtsstaat*, however, will be characterized by formalizing any changes to its constitution and by erecting barriers to its modification, usually by requiring especially enhanced majorities. A *Rechtsstaat*-type constitution will also be guarded against constitutional change through stipulating—in the case of written constitutions—that any amendments shall be expressly made in the text, i.e., by modification of the wording itself or appendix to the constitutional document. This avoids transgressing the bounds of the constitution and prevents confusion which may be caused by the coexistence of texts of equal legal standing. A *Rechtsstaat* does not require that a constitution purport to declare some particularly important component parts or substantive elements as immutable, thus fixing them 'in perpetuity,' although this has indeed been done in some constitutions.

The distinction between higher-ranking constitutional and other law in the concept of *Rechtsstaat* entails the precedence of the higher-ranking law in cases of conflict. Law incompatible with the higher-ranking law is void or, if possible, has to be interpreted in conformity with constitutional law. *Rechtsstaat*-type constitutions often spell out which judicial body is qualified to rule with binding effect on the constitutionality or unconstitutionality of ordinary law, in order to protect the authority of the law from those applying it (executive or judicial branches of government) who might otherwise be tempted to avoid applying the law by alleging its unconstitutionality. It is part of the logic of a *Rechtsstaat* that, for this reason among others, special courts are created for constitutional matters; however, this is not an essential element of a *Rechtsstaat* as long as other courts remain competent to safeguard the constitution and maintain its precedence over ordinary law.

The classical doctrine of separation of powers, which distinguishes between legislative, executive and judicial branches of government, forms the core of the modern concept of *Rechtsstaat*. It should be noted, however, that this distinction is only a basic one, with creation of legal norms by the executive branch or administrative bodies not being alien to the concept of *Rechtsstaat*. Such creation of legal norms does, however, need legal boundaries: laws not enacted by parliament will be acceptable in a *Rechtsstaat* only if the body creating such laws is by itself sufficiently legitimized. If such laws generate rights and duties of individuals, their creation needs to be backed up by an act of the parliamentary legislature. This in turn requires a sufficiently defined scope of the law-making power so delegated—parliament may not transfer its primary legislative responsibility summarily and without limitation to other governmental bodies, or even private norm-creating institutions. *Rechtsstaat* also demands that political parties—as associations designed to gain political power and which are legitimately also oriented towards particular group interests—are placed outside the ambit of government authority.

A *Rechtsstaat*-type constitution purports to make fundamental decisions on values regarding the organization of the state—but only of the state: it does not pretend to achieve comprehensive regulation of society as a whole. On the contrary, such a constitution will leave room for choice. Similarly, it will leave room for choice as regards ordinary domestic law, which must confirm to the constitution without, however, being predetermined by it. Legislating and applying laws does not amount to mere implementation of a constitution. The constitution of a state will not permit us to predict how political processes may unfold in terms of substance. This limits constitutional courts or other bodies entrusted with safeguarding the constitution—for them, interpreting the constitution means bearing in mind its openness. Safeguarding the constitution

thus means identifying where it draws strict limits and where it leaves room for choice and development. Modern constitutions characteristically use general terms, which may imply the prerogative of legislative choices yet to be made. Constitutional jurisprudence in a *Rechtsstaat* does bear political traits but remains confined to retracing the decisions already made in the constitution; it may not disregard the decisions of the constituent power.

Rechtsstaat includes also the guarantee of individual rights. The core of individual rights is the protection of human dignity. Rights to freedom and to equality alike are intended to afford such protection; they are defensive rights vis-à-vis the state which has a duty to prevent, or at least to make an effort to prevent, infringement of these rights in societal relations, i.e., between private persons dealing on an equal footing. Individual rights, whether manifested in international, domestic, constitutional, or ordinary law, have developed in great variety. While not all of these rights will count among the minimum of what a *Rechtsstaat* requires, a *Rechtsstaat* must include individual rights guaranteeing life and health, property and the freedom to choose one's occupation, freedom of expression, assembly, and association, of individual access to the courts, and of protection from discrimination. Access to the courts must not be understood in a formalistic sense: there needs also to be effective protection through granting a fair hearing and independence of the courts. Since *Rechtsstaat*, as has been explained, denotes a type, the concept of *Rechtsstaat* does not intend to impose a uniform level of fundamental rights. It does not prescribe, for instance, the scope of the protection to be afforded to property or the manner of safeguarding judicial independence.

Given the coexistence on an equal footing of numerous persons enjoying fundamental rights—i.e., all citizens—and given also the need to limit the latitude of individual behavior in the interests of the state as a whole, a *Rechtsstaat* needs to formulate rules on the limitability of individual rights and also to define the boundaries of such limitability. In this context, the principle of proportionality in a *Rechtsstaat* merits emphasis. Proportionality requires a balance between different rights and interests. The test of proportionality in a *Rechtsstaat* will be met only by such governmental action as is suitable to further a legitimate objective, which is necessary in the sense that it provides the least restrictive, yet equally suitable alternative available, and which, on consideration of all relevant circumstances, can be regarded as reasonable.

A third characteristic of *Rechtsstaat* is the quality of its law, encompassing the entire legal order and not merely those norms primarily concerned with state-individual relationships; the quality of civil and criminal law are also included. The whole legal system must possess certain qualities. For example, it must meet requirements as to normative clarity to provide legal

certainty, because indeterminate standards are incapable of lending themselves to judicial scrutiny. Moreover, law has to satisfy certain immanent criteria in order to serve as an instrument of guidance: it has to be free of contradictions and designed to retain continuity without unduly impeding social change. It has to be capable of acceptance by those bound by it, not perhaps in every detail but in its entirety. Law whose only prospect of implementation would lie in coercive measures or the threat of such would be doomed to fail. Law in a *Rechtsstaat* needs to be founded on a basic consensus of what is just.

2.2 '*Rechtsstaat*' as a Constitutional Principle

As explained above, one needs to differentiate between using the term *Rechtsstaat* for denoting a type and, on the other hand, for denoting a legally operative norm of constitutional law. To qualify a state as a *Rechtsstaat* does not require such a constitutional principle; its existence is not required if the sum total of constitutional norms already suffices to safeguard the objectives of *Rechtsstaat*, as described in Sect. 2.1. Some constitutions spell out expressly a principle of *Rechtsstaat*; in others the principle is derived from interpreting the constitution. In the Basic Law of the Federal Republic of Germany, for example, *Rechtsstaat* appears as a legal norm in its own right and is frequently referred to by the Constitutional Court. While there is disagreement as to which provision of the Basic Law should serve as the basis for the principle of *Rechtsstaat*, it is by most considered to be Article 20 para. 3 of the *Basic Law*, which reads that legislation shall be subject to the constitutional order and that the executive and the judiciary shall be bound by law and justice. This covers only a segment of the objectives of *Rechtsstaat*. Beyond this, the provision is taken to be a directly operative precept to ensure legal certainty, such as by avoiding indeterminacy in ordinary law or retroactivity of enactments; the principle of proportionality is likewise derived from this precept. State acts not in conformity with this precept are unconstitutional, being incompatible with the principle of *Rechtsstaat*. Laws may be null and void for this reason.

A general principle of *Rechtsstaat* as a constitutional norm in its own right is not without risks, due to its own indeterminacy. *Rechtsstaat* shares this risk with all blanket clauses. Workable civil law could not be imagined if it were limited to stating that behavior in legal relations should be governed by the principle of good faith alone, and constitutional law may similarly not retreat to merely stating a principle of *Rechtsstaat*, as this would leave decisions on what has to be considered in conformity with *Rechtsstaat* entirely to the judicial branch. Law dominated by blanket clauses concedes considerable decision-making powers to the courts and could thus not be safe

from arbitrariness. This is why a blanket-clause style of the *Rechtsstaat* principle can only have secondary relevance. Translating a *Rechtsstaat* into reality is better achieved by detailed norms that specify relevant objectives and that furnish a safer foundation for applying the law.

3. 'Rechtsstaat' and International Organizations

The notion of *Rechtsstaat* could be a useful concept also in the sphere of structures transcending individual states, as international organizations need to strike a balance between overarching interests affecting the whole organization and those regarding individual members. In the case of traditional international organizations, the issue has hardly ever been put forward. These organizations group together equal subjects of international law, i.e., states, in order to pursue jointly certain aims which as a rule remain individual objectives; the sovereignty of participating states remains unaffected. The European integrative community had its starting point in such a structure, but from the beginning it was designed to include transfer of sovereignty, and this community has increasingly acquired legal powers that justify looking at it as a new type of organization. Here, as in other organizations which initially aimed at economic integration but which have the potential to expand beyond this particular sphere, the question of arrangements addressing requirements of *Rechtsstaat* needs to be raised. The basic distinction between higher-ranking constitutional law and other law would apply in such a case. Matters of the quality of the law, such as the absence of systemic logical inconsistency or its indeterminacy, would also need to be addressed, as would issues of safeguarding compliance with the law and judicial control. Derived from the municipal laws of its member states, the objective of building a *Rechtsstaat* legal environment has already been realized in the European Union to a considerable degree, especially in the judicial branch. The European Court, while not a constitutional court, may be considered as a guardian of the fundamentals of Community law. Preserving the municipal *Rechtsstaat* heritage in supranational structures is an important task for the future that will grow more urgent as creation, enforcement, and control in the area of law is transferred to the integrated, supranational level.

Experiences gained in connection with *Rechtsstaat* are of interest also to international organizations to which member states have not transferred sovereign rights, insofar as such organizations are competent to issue legally binding decisions, to define legal norms, or to impose sanctions. Special organizations such as the World Trade Organization (WTO) or the United Nations (UN) (the global international organization designed to secure peace, to renounce the use of force, and, beyond that, to secure basic human needs), come to mind in this regard as their charters comprise

normative potentiality. The UN Security Council has powers that would enable it to fashion international relations in a legally binding manner far more than it has to date. The UN Security Council may be viewed as an organ of law enforcement. Apart from the right of self-defence of individual states, the Council enjoys a kind of monopoly of the use of force. Measuring this position against the yardstick of *Rechtsstaat* will reveal numerous shortcomings. The Council acts on the strength of norms whose determinacy may be open to question. It enjoys a monopoly on the use of force without legal safeguards, which would ensure that it acts to protect subjects of law who are threatened with violence by individual states. Being dominated by a privileged group of states, its legitimacy is in doubt. Judicial control is not ensured either, as the International Court of Justice does not exercise the role of a constitutional court in respect of the UN Charter as an international constitutional order. Any step in the direction of transforming the United Nations into a global state-like structure therefore imperiously calls for expansion of *Rechtsstaat* elements. This naturally cannot be achieved by simply imposing regulatory patterns tried and tested in *Rechtsstaat*-type national jurisdictions, but the tradition of *Rechtsstaat* offers a rich lode of lessons that will have to be heeded in the efforts to find an adequate legal framework for the modern world system and its segments.

See also: Constitutional Courts; Constitutionalism; Constitutionalism, Comparative; Democracy; European Union Law; Fundamental Rights and Constitutional Guarantees; International Organization; Rule of Law; State, History of; State, Sociology of the

Bibliography

- Bähr O 1864 *Der Rechtsstaat*. Kassel, Göttingen, Germany
Franck Th M 1999 *Democracy, Legitimacy and the Rule of Law: Linkages*. New York University Law School, Public and Legal Theory Working Paper No. 2, New York, Social Science Research Network Electronic Library (SSRN), Working Paper Series. <http://papers.ssrn.com>
Hofmann H 1995 Geschichtlichkeit und Universalitätsanspruch des Rechtsstaats. *Der Staat* 34: 1–32
Kunig Ph 1986 *Das Rechtsstaatsprinzip*. J. C. B. Mohr (Paul Siebeck), Tübingen, Germany
Kunig Ph 1995 Völkerrecht als Öffentliches Recht—Ein Glasperlenspiel? In: Randelzhofer A, Scholz R, Wilke D (eds.) *Gedächtnisschrift für Eberhard Grabitz*. C. H. Beck'sche Verlagsbuchhandlung, München, Germany, pp. 325–46
Neuman G L 1999 U.S. Constitutional conception of the rule of law and the Rechtsstaatsprinzip of the Grundgesetz. In: Battis U, Kunig Ph, Pernice I, Randelzhofer A (eds.) *Das Grundgesetz im Prozess europäischer und globaler Verfassungsentwicklung*. Nomos-Verlagsgesellschaft, Baden-Baden, Germany, pp. 253–68
Schmidt-Abmann E 1987 Der Rechtsstaat. In: Isensee J, Kirchhof P (eds.) *Handbuch des Staatsrechts der Bundesrepublik Deutschland*. C. F. Müller Juristischer Verlag, Heidelberg, Germany, p. 24

- Shapiro I (ed.) 1994 *The Rule of Law*. New York University Press, New York
- Sobota K 1997 *Das Prinzip Rechtsstaat, verfassungs- und verwaltungsrechtliche Aspekte*. J. C. B. Mohr (Paul Siebeck), Tübingen, Germany
- Voigt St 1998 *Making Constitutions Work—Conditions for Maintaining the Rule of Law*. Social Science Research Network Electronic Library (SSRN), Working Paper Series. <http://papers.ssrn.com>

P. Kunig

Copyright © 2001 Elsevier Science Ltd.
All rights reserved.

Recognition Memory (in Primates), Neural Basis of

The term *recognition* refers to the knowledge or feeling that an object has been encountered previously. Two different mental processes may contribute to recognition judgments: a general feeling of 'familiarity' and a specific 'recollection.' These two aspects of recognition may depend on different brain systems, but are treated together here.

1. Neural Systems Mediating Visual Recognition

For most sensory modalities (e.g., vision and touch), receptors convert sensory information into neural signals, and modality-specific cortical fields, such as the primary visual cortex ('V1' in Figure 1), process that information further. These early sensory processing stages are a prerequisite for recognition memory. However, the core recognition system must function more generally. If a blind person fails to recognize a visually presented object, we say that he or she has a visual impairment, not a failure of recognition memory. Therefore, this article will focus on general aspects of recognition and its neural basis, neglecting modality-specific ones.

1.1 Methods for Assessing Recognition Memory in Nonhuman Primates

Because both human and nonhuman primates rely heavily on visual information in their daily activities, recognition is best understood for visible objects. Visual recognition memory in monkeys has been studied by making selective brain lesions and by disconnecting one brain region from another. These studies identify whether a given brain structure is critical for recognition. Monkeys, of course, cannot tell us whether they remember seeing an object. For this reason, methods have been developed that allow them to reveal that information. One commonly used method is the *delayed nonmatching-to-sample task* (DNMS). In DNMS, each trial has two parts: a

sample presentation followed by a choice test. During the sample presentation, the monkey sees a single object, which covers the central well of a three-well test tray. The monkey then displaces the object to obtain a small piece of food hidden in the well underneath. A few seconds later the monkey sees the same, now-familiar object plus a novel one, with one object over the left well of the test tray and the other over the right well. The monkey can obtain more food by displacing the novel object but not by pushing aside the familiar one. Thus, on the choice test, the monkey solves the problem by applying a 'nonmatching' rule, i.e., by choosing the object that does not match the sample. If the monkey consistently chooses the novel object on the choice test, one can infer that the monkey *recognizes* the sample. In the laboratory, there are many variations of this basic method. For example, monkeys can be trained using a 'matching' rule (the *delayed matching-to-sample task*, or DMS) rather than the nonmatching one. It is also common to use a computerized apparatus that presents two-dimensional pictures on a touch-sensitive video screen rather than using actual objects on a test tray.

1.2 Role of the Medial Temporal Lobe in Monkeys

It has been known for many years that a limited set of brain structures, including components of the medial temporal lobe (MTL), are necessary for storing information about facts, objects, and events from all sensory modalities (Squire 1987). Early investigations into the neural basis of recognition memory pointed to two structures buried deep in the MTL: the amygdala and hippocampus (Mishkin 1982, see Figure 1). However, those experiments involved the mechanical removal of brain tissue, a technique that produced unintended damage to other parts of the MTL. More recent studies have used stereotaxic injection of neurotoxins to make more selective tests. These experiments show that, within the MTL, the perirhinal cortex and the entorhinal cortex, rather than the amygdala and hippocampus, are essential for visual recognition (Murray 2000). Monkeys with an intact amygdala and hippocampus, but without the perirhinal and entorhinal cortex, cannot recognize objects if they have to remember them for more than a few seconds.

Experts accept a central role for perirhinal and entorhinal cortex in object recognition and also agree that the amygdala is not important for this type of memory. Controversy remains, however, about the potential contribution of the hippocampus. Damage to the hippocampus leads to impairments in recognition memory in some studies but not in others. For example, one test of recognition relies on the tendency of monkeys to spend more time looking at novel objects than at familiar ones. This innate tendency is called 'preferential viewing.' Although hippocampal damage fails to produce consistent and systematic deficits on DNMS, it does disrupt pref-

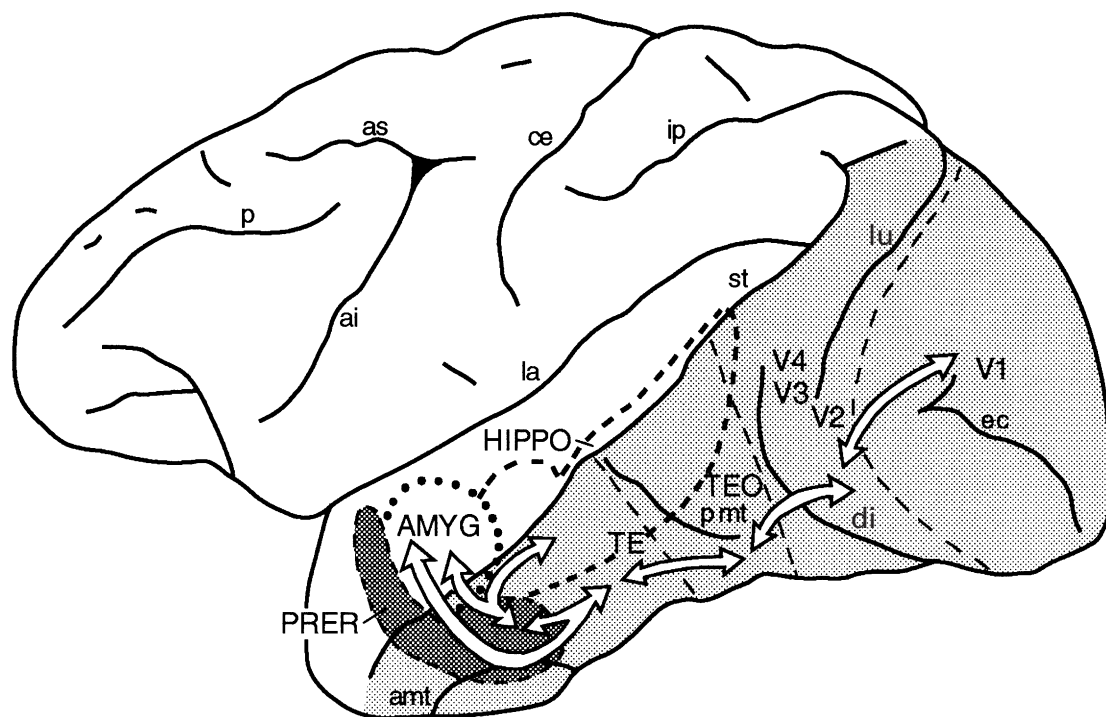


Figure 1

Diagram of the lateral view of a monkey brain showing some of the brain structures mediating visual recognition memory. The perirhinal and entorhinal cortex (PRER) are located on the ventromedial surface of the temporal lobe, and so are only visible by looking through a 'transparent' temporal lobe. The amygdala (AMYG) and hippocampus (HIPPO) are buried deep in the medial temporal lobe, and the hippocampus is shown slightly higher than its real location, for ease in viewing the lateral temporal cortex. Arrows indicate two-way information flow from early (V1–V4) to late (TEO and TE) cortical visual processing areas. These regions then transfer information into the amygdala and into the perirhinal and entorhinal cortex, which provide the main route by which information reaches the hippocampus. Abbreviations: ai, inferior arcuate sulcus; amt, anterior middle temporal sulcus; as, superior arcuate sulcus; ce, central sulcus; ec, external calcarine sulcus; io, inferior occipital sulcus; ip, intraparietal sulcus; lu, lunate sulcus; la, lateral fissure; p, principal sulcus; pmt, posterior middle temporal sulcus; st, superior temporal sulcus

erential viewing. This could be because the preferential viewing test, unlike DNMS, does not rely on learned rules, or because the stimuli have no association with food. Interestingly, on DNMS, larger perirhinal and entorhinal cortex lesions cause greater memory deficits than smaller ones, which is what one would predict if a structure were playing a direct role in a particular function. The reverse relationship holds for the hippocampus. That is, larger hippocampal lesions are associated with better recognition memory than smaller lesions, suggesting an indirect relationship of the damage to recognition memory deficits. The fact that only the perirhinal and entorhinal cortex have the predicted relationship to recognition memory supports the view that they are the parts of the MTL mediating that function.

The same parts of the MTL that are essential for visual recognition memory subserve tactual recog-

nition memory as well. They are also critical for many types of stimulus–stimulus association memory. Accordingly, their function may best be characterized as one of object identification. Through this process, the visual, tactual, gustatory, olfactory, and auditory properties of objects are somehow linked or 'bound.' Thus, the perirhinal and entorhinal cortex are essential not only for visual recognition, but also for storage of information about objects generally (Murray 2000).

1.3 Role of the Thalamus and Frontal Lobe in Monkeys

The perirhinal and entorhinal cortex do not operate alone, but instead work in concert with other anatomically connected brain structures that, together, form a neural network mediating visual recognition

memory. At least two other brain regions participate: the medial thalamus and the ventromedial part of the prefrontal cortex (Mishkin and Murray 1994). Damage to either of these structures produces essentially the same effect on visual recognition memory as does damage to the perirhinal and entorhinal cortex. In addition, the thalamic and frontal regions noted above are connected with each other, and each is connected to the perirhinal and entorhinal cortex. These structures need to interact to mediate visual recognition; surgical manipulations that prevent information from one structure from reaching a second produce a deficit in recognition memory much like that seen after lesions to each individual site (Parker and Gaffan 1998). Precisely how these structures interact to produce recognition memory remains uncertain; it is possible that each structure underlies some unique aspect of recognition memory.

1.4 The Recognition Network in the Human Brain

The same neural structures that are critical for recognition memory in monkeys appear to be essential in humans as well. For example, memory impairments are found in humans following damage to the MTL as a result of either brain infections (e.g., Herpes encephalitis) or temporary interruption of the blood supply. Damage to the medial thalamus, as a result of Korsakoff's disease, and to portions of the ventromedial prefrontal cortex, as a result of aneurysms of the anterior communicating artery, yield similar effects. It must be kept in mind, however, that in clinical cases brain damage rarely respects the boundaries of individual structures, so that the kind of precision outlined above for monkeys is not often possible for the human brain. Nevertheless, there is evidence that in humans, as in monkeys, the perirhinal and entorhinal cortex are important for recognizing objects. And, as in monkeys, there is controversy about the role of the hippocampus. People with damage to the hippocampus and adjacent perirhinal and entorhinal cortex are much more impaired in stimulus recognition than those with damage restricted to the hippocampus. Whereas some studies have shown that humans with damage limited to the hippocampus and closely related structures have relatively intact recognition abilities (Aggleton and Brown 1999), others have reported that such subjects are, on average, impaired in recognition memory (Squire and Knowlton 2000). In humans, the hippocampal contribution to recognition memory appears to reflect 'recollection' as opposed to 'familiarity' judgments.

2. Neurophysiology of Visual Recognition

Although the neuronal mechanisms mediating visual recognition memory are not known, there is some

information about the way in which neurons signal that a visual stimulus has been seen previously (Brown 1996). As shown in Figure 1, the temporal lobe consists of several parts. The MTL has already been discussed. On the outer, lateral surface of the temporal lobe are located visual areas TE and TEO. These regions are called the *inferior temporal cortex* (IT). The perirhinal and entorhinal cortex may be considered part of either IT or MTL. Many neurons in IT, including the perirhinal and entorhinal cortex, show changes in sensory responses with stimulus repetition. Two main kinds of recognition-related responses predominate.

Some neurons respond less to a visual stimulus on a monkey's second or subsequent viewing of an object than they do on the first. Other neurons respond more. Similar phenomena occur in the medial thalamus and the prefrontal cortex, but not in the hippocampus. Thus, these types of activity signaling stimulus repetition occur in precisely those areas thought to be essential for performance on DNMS. In addition, there is a third type of neuronal activity related to recognition. In monkeys performing DMS (see Sect. 1.1), some neurons in both IT and the prefrontal cortex reflect information about the sample object during the interval between sample presentation and choice test, as though the cell were maintaining a memory of the sample (Desimone et al. 1995).

Furthermore, for prefrontal neurons, this type of activity persists even when other objects appear during a memory period. This kind of activity could be very useful for the short-term storage of information about an object and for directing attention to items of interest.

3. Cognition and Recognition

Recognition is only one among many kinds of memory and cognition. It means, literally, to 'know again.' In people, recognition can be normal, or nearly so, in the face of devastating memory impairments of other types—for example, an inability to remember personal daily activities. Conversely, recognition can be extremely impaired in people who are capable of learning and applying complex rules and strategies. Accordingly, recognition memory can be viewed as an aid to more complex cognitive operations, but is, itself, a modular function with adaptive advantages of its own. The nature of these advantages is uncertain, but probably involves the ability to direct attention to novelties in the environment (perhaps by reducing the sensory responses to recognized objects) or to familiar objects that are of importance to the behavior of the moment. It remains to be seen whether the mechanisms that underlie the recognition of objects, as discussed in this article, also subserve the recognition of thoughts and ideas, logical truth and falsity, esthetic beauty and ugliness, or other abstract aspects of human recognition that distinguish us from our fellow primates.

See also: Amygdala (Amygdaloid Complex); Declarative Memory, Neural Basis of; Hippocampus and Related Structures; Learning and Memory, Neural Basis of; Recognition Memory, Psychology of

Bibliography

- Aggleton J P, Brown M W 1999 Episodic memory, amnesia, and the hippocampal–anterior thalamic axis. *Behavioral and Brain Science* **22**: 425–89
- Brown M W 1996 Neuronal responses and recognition memory. *Seminars in Neuroscience* **8**: 23–32
- Desimone R, Miller E K, Chelazzi L, Lueschow A 1995 Multiple memory systems in the visual cortex. In: Gazzaniga M S (ed.) *The Cognitive Neurosciences*. The MIT Press, Cambridge, MA
- Mishkin M 1982 A memory system in the monkey. *Philosophical Transactions: Royal Society London B* **298**: 85–95
- Mishkin M, Murray E A 1994 Stimulus recognition. *Current Opinion in Neurobiology* **4**: 200–206
- Murray E A 2000 Memory for objects in nonhuman primates. In: Gazzaniga M S (ed.) *The New Cognitive Neurosciences*. The MIT Press, Cambridge, MA
- Parker A, Gaffan D 1998 Interaction of frontal and perirhinal cortices in visual object recognition memory in monkeys. *European Journal of Neuroscience* **10**: 3044–57
- Squire L R 1987 *Memory and Brain*. Oxford University Press, New York
- Squire L R, Knowlton B J 2000 The medial temporal lobe, the hippocampus, and the memory systems of the brain. In: Gazzaniga M S (ed.) *The New Cognitive Neurosciences*. The MIT Press, Cambridge, MA

E. A. Murray

Recognition Memory, Psychology of

Recognition memory is usually used to refer to both a memory measurement and a memory process that occurs in this measure of memory. When discussing recognition memory as a memory measurement, the difference between a recognition memory test and other memory tests such as recall tests can be easily distinguished. Recognition tests also produce certain effects that are particular to the processes underlying recognition memory. Even though the nature of these processes is very complicated, some models of recognition memory have nonetheless been proposed (see, for example, Raaijmakers and Shiffrin 1992). In this article, these two characteristics of recognition memory are examined in more detail.

1. Recognition Memory as a Memory Test

1.1 The Nature of a Recognition Memory Test

In a recognition memory experiment, to-be-remembered items (also called target items) are presented to

subjects. An item may refer to a nonsense syllable, a word, a number, a picture, and so on. In such tasks, subjects are given a recognition memory test after exposure to such items.

There are basically two types of recognition memory tests. In a yes/no recognition memory test, a subject is presented a series of items one at a time. The subject is then asked to say 'yes' or 'old' if the subject thinks the current item is a target item, and 'no' or 'new' if the subject thinks it does not qualify as a target one (nontarget items are called distractor items).

This type of a recognition memory test is similar to the sort of true/false test often encountered in school examinations. Another type of recognition memory test is a forced-choice test. In a forced-choice test, the subject is shown two or more items at a time. One of the items is usually a target item and the others are distractor items. In this type of recognition memory test, the subject is asked to pick out the target item from a set of items. If the subject is shown three items at a time, for example, it is called a three-alternative forced-choice test.

As such, the forced-choice test is similar to multiple-choice tests often encountered in school examinations. Of course, many additional variations on recognition memory tests may be considered. For instance, in addition to judging whether an item given on a recognition memory test is a target item or a distractor item, a subject may be asked to give a numerical confidence rating consisting of the range of, for example, 1 (least certain) to 5 (most certain), to indicate how certain the subject is that the judgment is correct.

In order to develop a better understanding of recognition memory tests, it is worthwhile contrasting them with other popular forms of memory test, such as the aforementioned recall tests. In recall tests of memory, a subject is asked to reproduce or generate all the items that had been presented during an earlier study phase. Thus, the basic difference between recognition and recall test procedures is that in a recognition memory test, subjects are given the target and distractor items during the test, whereas in a recall test, the subjects are given instructions to reproduce or generate the items without displaying them.

One of the important differences is that, in general, the subjects can recognize items in a study list much better than they can recall them. In fact, if the subjects are first given a recall test and then given a recognition memory test, it will be usually found that they can recognize many items that they were unable to recall.

1.2 Signal Detection Theory and Performance on a Recognition Memory Test

In a recognition memory test, instead of being assumed that subjects either memorize target items or don't memorize them, it is assumed that the subjects judge as

the target items on something like level of familiarity which varies in memory strength. When an item can be retrieved more readily than the other items, it will be said that the item is at full strength.

The memory strength of a test item is thought of as the degree of its familiarity in long-term memory (LTM). The stronger the test item is in LTM, the more familiar it will be.

To illustrate the assumption, one recognition memory experiment will be considered. Assume that a list of items is given to two groups of subjects, followed by a yes/no recognition memory test. The subjects are tested with a mixed sequence of target and distractor items one at a time. Prior to the recognition memory test, one group of subjects (called a guessing group) is instructed to say 'yes' if they think an item was on the list, and 'no' if they think it is a distractor item. Guessing is permitted in this group when the subjects receive a recognition memory test. The other group of subjects (called a strict group) receives the same instruction as the guessing group except the guessing instruction. The subjects in this group are instructed not to make a guess when they judge whether an item given on a recognition memory test is a target item or a distractor item. The subjects in the guessing group will of course use guessing strategies. So the results may show that the subjects in the guessing group will pick up more target and distractor items than those in the strict group. The results mean that the subjects in the guessing group judge the target and distractor items on a higher level of familiarity than those in the strict group. However, as can be seen, performance on a recognition memory test in the guessing group may be very similar to that in the strict group.

The signal detection theory which originates with the study of auditory detection tasks (see Green and Swets 1966) and has been applied to recognition memory explains the results of this experiment clearly. The signal detection theory enables researchers to estimate the amount of information stored in LTM on which subjects' judgments on a recognition memory test are based.

The signal detection theory assumes that the subjects estimate the probabilities that a test item came from the distributions for target and distractor items, respectively, on the basis of its memory strength (or familiarity).

In a typical yes/no recognition memory test, there are four cells that can occur in combination with a subject's 'yes' or 'no' responses and target-distractor items. First, the item is a target item and the subject says 'yes'; this is the correct response and is called a *hit*. Second, the item is a target item but the subject says 'no' to it; this is not the correct response and is called a *miss*. Third, the item is a distractor item and the subject says 'no'; this is the correct response and is called a *correct rejection*. Fourth, the item is a distractor item but the subject says 'yes' to it; this is not the correct response and is called a *false alarm*.

These four cells are not independent. As shown above, the hit rate and the miss rate mean the response rates to the target items and the false alarm rate and the correct rejection rate mean the response rates to the distractor items. When the hit rate is known, the miss rate can be understood automatically. Similarly, when the correct rejection rate is known, the false alarm rate can be understood. Generally, the hit rate and the false alarm rate are referred to on recognition memory.

Figure 1 shows both a relationship between the distributions of memory strength for target and distractor items and four cells in a yes/no recognition memory test. Figure 1 shows that the distributions for target and distractor items on the basis of their memory strength represent normal curves and that the mean strength of the target items is usually higher than that of the distractor items.

The distance between the means of the distributions for target and distractor items is called d' (called *d-prime*). More precisely, the d' value is obtained by dividing the difference between the two means by the common standard deviation of the distributions. We can easily get the d' value from some publications (e.g., Elliot 1964, Hochhaus 1972). When the d' value increases, the subject will discriminate between target and distractor items easily. When the d' value is small, the subject will find it difficult to discriminate between two sets of items.

In addition to the d' value, there is another value called β (beta). The β value means the level of the criterion strength on which the subject bases the decision. When the d' value remains constant and the β value goes up (e.g., the β value moves to the target distribution), both hit rate and false alarm rate go down. When the d' value remains constant and the β value goes down (e.g., the β value moves to the distractor distribution), both hit rate and false alarm rate go up. When the d' values are assumed to be the same, the β value in the guessing group stated earlier may be very low and the β value in the strict group may be very high. Figure 1 shows one example of d' and β values on recognition memory. As can be seen, the d' value is referred to recognition accuracy which means the strength difference between target and distractor items in LTM and the β value is referred to as decision criterion. The subject evaluates the memory strength difference between target and distractor items on the basis of decision criterion.

1.3 Some Effects in Recognition Memory

Some effects have been demonstrated in recognition memory, which are referred to as the mirror effect and the context effect. These effects in recognition memory are explained in brief.

First, there is a strong regularity of recognition memory called the mirror effect (see Glanzer and

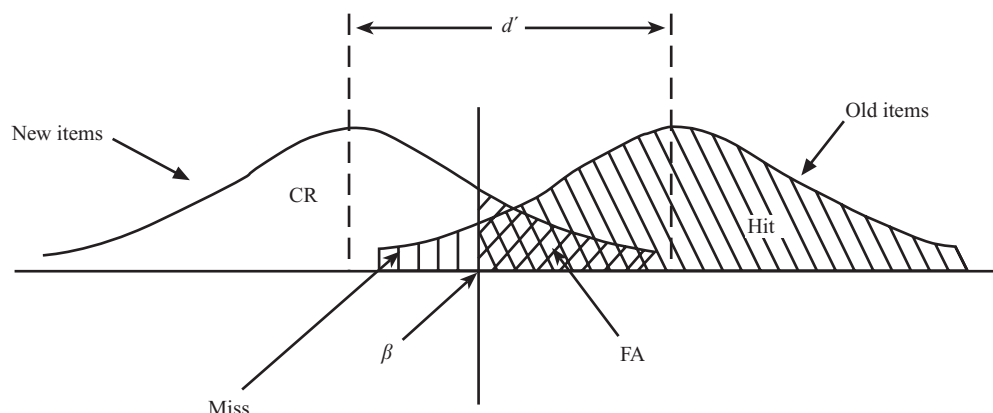


Figure 1

Four types of response results (hit, miss, false alarm (FA), and correct rejection (CR)), d' , and β in recognition memory

Adams 1985). The mirror effect refers to the regularity of recognition memory in which the type of items that is relatively easy to recognize as old when old is also relatively easy to recognize as new when new. Glanzer and Adams (1985) analyzed a total of 80 experimental findings with several different classes of items and found that performance on new items from each class is correlated with performance on the corresponding classes of old items.

Second, subjects can usually recognize target items better than they can recall them. However, in such test circumstances as recognition memory is measured under the changes in the context between study and test, it will be found that the subjects fail to recognize target items. It is known as the context effect in recognition memory. The context effect in recognition memory shows that even recognition memory depends on the relation between the context at the time of the study and at the time of its test. For detailed treatment of the context effect see *Encoding Specificity in Memory*.

2. Retrieval Processes in Recognition Memory

As stated earlier, recognition and recall are usually used to refer to not only memory measurements, but also memory (or retrieval) processes which occur in these memory measurements. There have been hot controversies about how retrieval processes in recognition and recall are similar and different. That question is very difficult to answer and is still controversial (Brown 1976, Tulving 1983).

Generally speaking, recognition memory seems to be easier than recall. Subjects who receive a recognition memory test only discriminate between target and distractor items, even though they do not know

the target items clearly. There are some reasons why recognition memory seems to be easier than recall. According to a classical threshold model of recognition memory, the level of the discriminative threshold in recognition is thought to be higher than that in recall. According to the two-process model of recall (Kintsch 1970), recognition memory and recall involve different retrieval processes. Recall involves an extra search (or generate) process in addition to a recognition (or decision) process. For more detailed treatment of retrieval processes in recall see *Memory: Organization and Recall; Memory Retrieval; Encoding Specificity in Memory*.

2.1 One-process Models or Two-process Models in Recognition Memory

How about retrieval processes in recognition memory? Recognition memory can usually be referred to by two types of models, one known as one-process models and the other known as two-process models.

One-process models include the Search of Associative Memory (SAM) model proposed by Gillund and Shiffrin (1984) and the Hintzman (1988) model. They assume that recognition judgments are based on the target item's familiarity involving a single retrieval (search-like) step. Two-process models include the Atkinson and Juola (1974) model and the Mandler (1980) model.

They assume that the retrieval processes in recognition memory are the result of two processes: familiarity and search. For example, Atkinson and Juola (1974) carried out response time experiments from well-learned lists. They obtained the results that items with high or low familiarity values are associated with a fast response of a target item or a fast response

of a distractor item, respectively, whereas items with intermediate familiarity values are associated with a slower search process to judge whether an item is a target item or a distractor item. This search process is generally supposed to be slower than the familiarity-based process in the two-process models of recognition memory. The controversies about whether there is a search process in recognition memory are still continuing.

2.2 Recognition Memory With and Without Recollective Experiences

In recent years, there has been considerable interest in comparing performance on a task that does not require conscious or intentional recollection of previous experiences with performance on a task that requires it (see *Implicit Memory, Cognitive Psychology of*). A recognition memory test is known as one of the tests that requires conscious recollection of previous experiences. However, Gardiner (1988) assumes that recognition memory may entail two processes concerning recollective experiences: one is a process with recollective experiences and the other is a process without recollective experiences. These two processes are related to differences in subjective experiences.

In a recognition memory test, when recognizing each item, subjects are instructed to indicate whether they are able consciously to recollect its prior occurrence in the study list (called 'remember') or whether they recognize it on some other basis (called 'know'). The results show that a few independent variables, such as levels of processing (deep vs. shallow processing), generating vs. reading target items, and divided vs. undivided attention, have been found to influence 'remember' recognition and to have little effect on 'know' recognition. For example, performance on semantic processing was better than that on shallow processing in 'remember' recognition. However, performance on semantic processing was the same as that on shallow processing in 'know' recognition.

The measures of a 'remember' judgment and of a 'know' judgment are introduced by Tulving (1985), according to which conscious recollection or 'remembering' is a by-product of retrieval from the episodic memory system. Feelings of familiarity or 'knowing' are characteristic of retrieval from the semantic memory system. Using 'remember'-'know' judgments in recognition memory, it may be possible to examine retrieval processes in recognition memory and memory systems.

See also: Learning and Memory, Neural Basis of; Memory Models: Quantitative; Memory Retrieval; Microsimulation in Demographic Research; Recognition Memory (in Primates), Neural Basis of; Short-

term Memory, Cognitive Psychology of; Short-term Memory: Psychological and Neural Aspects; Signal Detection Theory; Signal Detection Theory, History of; Working Memory, Neural Basis of; Working Memory, Psychology of

Bibliography

- Atkinson R C, Juola J F 1974 Search and decision processes in recognition memory. In: Krantz D H, Atkinson R C, Luce R D, Suppes P (eds.) *Contemporary Developments in Mathematical Psychology*. Freeman, San Francisco, CA, Vol. 1
- Brown J (ed.) 1976 *Recall and Recognition*. John Wiley & Sons, London
- Elliot P B 1964 Tables of d' . In: Swets J A (ed.) *Signal Detection and Recognition by Human Observers*. Wiley, New York
- Gardiner J M 1988 Functional aspects of recollective experience. *Memory and Cognition* **16**: 309–13
- Gillund G, Shiffrin R M 1984 A retrieval model for both recognition and recall. *Psychological Review* **91**: 1–67
- Glanzer M, Adams J K 1985 The mirror effect in recognition memory. *Memory and Cognition* **13**: 8–20
- Green D M, Swers J A 1966 *Signal Detection Theory and Psychophysics*. John Wiley and Sons, New York
- Hintzman D L 1988 Judgments of frequency and recognition memory in a multiple-trace memory model. *Psychological Review* **95**: 528–51
- Hochhaus L 1972 A table for the calculation of d' and β . *Psychological Bulletin* **77**: 375–6
- Kintsch W 1970 Models for free recall and recognition. In: Norman D A (ed.) *Models of Human Memory*. Academic Press, New York
- Mandler G 1980 Recognizing: The judgment of previous occurrence. *Psychological Review* **87**: 252–71
- Raaijmakers J G W, Shiffrin R M 1992 Models for recall and recognition. *Annual Review of Psychology* **43**: 205–34
- Tulving E 1983 *Elements of Episodic Memory*. Oxford University Press, New York
- Tulving E 1985 Memory and consciousness. *Canadian Psychology* **26**: 1–12

H. Tajika

Reconstruction/Disaster Planning: Germany

In 1945, with World War II over and the National Socialist dictatorship under siege by Allied troops, great parts of German cities lay in rubble, their populations often reduced to only a few tens of thousands. Despite great hardship, life reemerged unexpectedly fast and soon discussions began on rebuilding the cities. Reconstruction plans had been in preparation since 1943. There was therefore no 'zero hour' for urban planning. The following article will underline the continual development in planning

theory and practice, the inspirations for the concepts, and the debate about urban design before and after 1945.

1. Preliminaries

After World War I, town and country planning in Germany in general changed its focus from the policy of improvement—slowly introduced since the late-nineteenth century—to a strategic approach. This phase was hampered by the economic crisis in 1929. When the National Socialists came to power in 1933, they had in fact no specific policies for urban planning but were able to select from existing scholarship those strategies which were useful to the new dictatorship: 1933/34: Improvement of both traffic and housing conditions in the historic urban cores of German cities had been widely neglected since World War I in favor of building new housing schemes at the urban fringe. The first concepts for the renewal of historically valuable and attractive housing stock from pre-industrial periods date from around 1930. This new approach then was to maintain the pretty street layout by clearing the blocks' interiors (as occurred in Brunswick, Frankfurt, and Kassel). Lack of money due to armaments production led to such schemes being deferred. However, it was useful to early National Socialist policy to support them as work providing measures which brought beneficial improvements to local economies, as well as for ideological reasons (see von Petz 1990).

1937: Albert Speer's plan to redesign Berlin as a future 'Germania' became the model for the redesign of German provincial capitals. Showing conventional architecture of monumental style with axes appropriate to the parading party and public buildings within central urban areas, these plans were nevertheless designed according to existing planning knowledge which had already been developed widely in the preceding decade.

1940: The policy of armament and the war of conquest led to a stronger functional approach to planning. Growing individual mobility and the change from the detached settlements of the early 'blood and soil' period to compact housing concepts provoked a change to planning paradigms even in wartime. The advance to the East led to initiatives for redesigning the settlement structure of the conquered territories. Walter Christaller's 'central places' theory of 1933 was easily adaptable to the required hierarchical structural system. Political goals combined with military ambitions also led to marked improvements in planning methods and techniques, such as statistics, cartography, and aerial photography.

1943: The disastrous bombing of cities forced a return to the debate dating back to the CIAM discussion of 1933 on rebuilding towns. A decree by the Führer regarding the reconstruction of German

cities after the anticipated victory empowered Albert Speer to establish a team of reconstruction planners, with each member given responsibility for one or more cities. The plans were quite heavily influenced by comparisons with advanced design methods to international standard and some, indeed, served as first concepts for reconstruction after 1945 (Durth 1986).

2. Political Reconstruction

From 1871, national state pride and the desire for international reputation dominated the politics of the Imperial Reich. The Morgenthau doctrine aimed to prevent Germany from returning to such a powerful state. In 1945, therefore, the Allies promoted the decentralization of Germany, basing their strategy on a tradition dating from the Middle Ages to the nineteenth century of strong feudal states with attractive provincial capitals. The city of Bonn, a modest city among those former feudal seats, became the capital of the Federal Republic of Germany. However, many of the new West German states retained their former seats as capitals. Berlin was divided and the German Democratic Republic kept East Berlin as its capital. In the East, the former state capitals were promoted less. Thus two systems existed simultaneously between 1961 and 1990: East Berlin as the capital and center of power in the socialist state, with other cities as only regional centers, and in the western federal state a more modest capital with 11 quite strong state capitals and, under them, a tier of regional towns.

3. Rebuilding the Cities: The Western Approach

Before the war the modernization process in German cities was badly impeded by their historic heritage. West German postwar planners therefore used the opportunity of rebuilding the bombed cities according to modern international planning standards that, before 1933, had been developed continuously and which Germany, with its dreams of a thousand-year Reich, had for a while gambled away. Nevertheless a very few international links were maintained (Fischer 1990). The founding of the Federal Republic in 1948, the monetary reform urged by the Marshall Plan, and the Korean War, produced economic growth which helped to speed reconstruction. This focused on two goals: first, to bring sun, air, and light into the urban cores which, to most minds, had suffered so much from urban concentration and industrialization and, second, to provide motor car access to the cities.

But West Germany adopted different approaches for rebuilding the demolished core areas in the post-war period (for statistics on rubble, see Diefendorf 1993, pp. 3–17). The schemes ranged from radical modernization to a moderate reconstruction or tra-

ditional adaptation (also Beyme 1987, pp. 175–7). Munster in Westfalia or the Bavarian capital of Munich, for example, employed quite traditional lines, based on the historic layout and scale and, in cases of new constructions, a moderate adaptation of modern design for the facades. Monuments were rebuilt or restored. Cities such as Kassel, Dortmund, Cologne or Hanover, on the other hand, were proud of their rather radical approach to the modern town plan.

Recently some cities have given greater emphasis to re-evaluating historic structures and inner city areas. In Frankfurt, rebuilding from scratch the eastern row of buildings along Römerberg in medieval style in 1984 provoked an intense controversy in West Germany about reconstruction and the 'truth' of history. This is one of several examples of postmodern city design in Germany.

4. Urban Development Policy

The Athens Charter was published in Germany only in 1957 but its ideas were already 25 years old and, during these years, had more or less been the basis for urban development policies. Organizing the city according to four main functions (work, housing, leisure, transport) was from this point the planner's core strategy. The treatment of the neighborhood as the basic planning unit was introduced into German planning from 1940. News of Abercrombie's Greater London Plan of 1943/44 with its green belt strategy spread quickly, as did the reconstruction plan for Rotterdam. Thus, after 1948, a flourishing economy

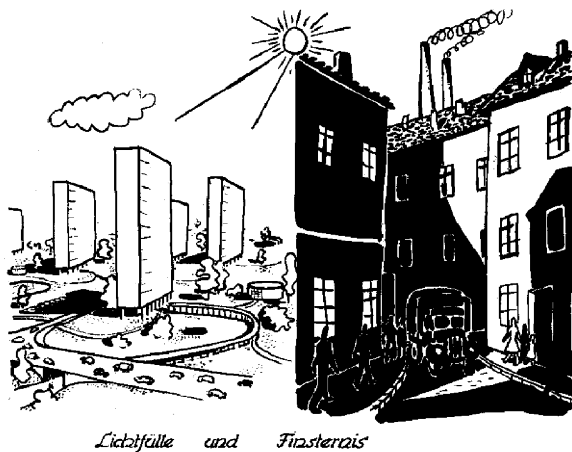


Figure 1
Marcel Lods, Werbung für die Forderungen der 'Charta von Athen.' Section du Plan (Mainz) 1947: Lichtfülle—Finsternis (Marcel Lods, promotion for the demands of the Athens Charter, (Mainz) 1947: richness of light—darkness)

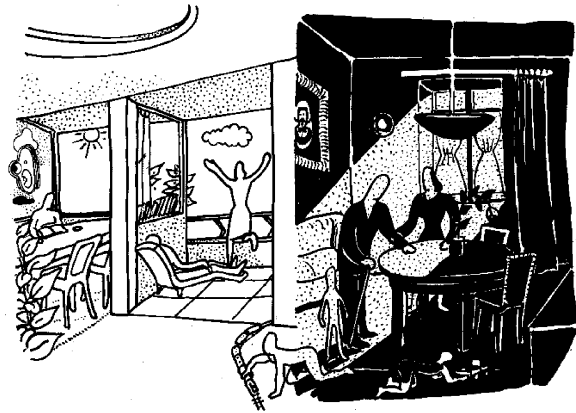


Figure 2
Marcel Lods, Werbung für die Forderungen der 'Charta von Athen.' Section du Plan (Mainz) 1947: modernes—traditionelles Wohnen (Marcel Lods, promotion for the demands of the Athens Charter, (Mainz) 1947: modern—traditional housing)

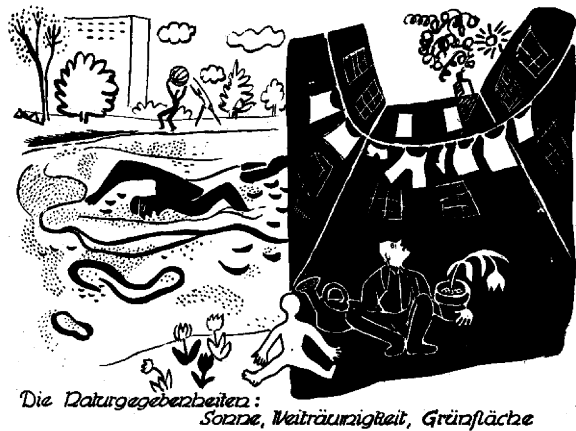


Figure 3
Marcel Lods, Werbung für die Forderungen der 'Charta von Athen.' Section du Plan (Mainz) 1947: Die Naturgegebenheiten—Sonne, Weiträumigkeit, Grünflächen (Marcel Lods, promotion for the demands of the Athens Charter, (Mainz) 1947: natural conditions—sun, space, green)

helped urban growth to develop to modern standards, based until the mid-sixties on the main concept of the 'organic art of town planning' (Reichow 1948), the 'composite and disaggregated city' (Göderitz, Rainer, Hoffmann 1957) or 'the car city' (Reichow 1959). The tenement house (*Mietskaserne*) and the noisy, crowded and dusty nineteenth-century industrial city were seen in a negative light (Figs. 1–3). The vision was of a green

city spread across the landscape, designed on a basic human scale as neighborhoods each with its own facilities, located in zones, suburbs or quarters. Motor car access was universal, but separated from walkways and rail systems, with a rich variety of open green spaces for leisure pursuits. The city center and central business district (CBD) were the dominant feature, incorporating reconstructed historic monuments and some high-rise buildings as the crown of the city.

The comprehensive urban renewal of nineteenth-century housing stock meanwhile aggravated the need for new housing, necessitating a change to a new policy of high-density housing in the periphery. Density outside the city also needed to be increased substantially to make public transport schemes efficient, and dreary suburban neighborhoods should, based on the perceptions of Jane Jacobs, or Hans Paul Bahrdt in Germany, be amended by 'urbanity,' by the transfer of urban design and dimensions to the periphery, without its lamentable deficiencies and defects.

Around 1975, severe economic changes stemming from the oil crisis and new technologies demanded an urgent review of planning aims and procedures. Several new arguments had already influenced planning—participation, ecology, the rediscovery of the nineteenth-century city and the attractiveness of its tenement house, and the ineffectiveness and inflexibility of large-scale planning. Planning was put under the microscope, necessitating reflection on, and alteration to, its theories, strategies, methods, and procedures. The new planning culture had to work with new technologies and economies, greater social mobility and the breakup of society and as a result new urban models developed. This discussion is still in progress.

5. *Rebuilding the Cities: The Eastern Approach*

Without a doubt, East Germany did things differently. Apart from Berlin, Dresden, Magdeburg, Leipzig and Chemnitz and a few others, many cities in the East were not touched by air raids. The building stock and inner areas of these cities therefore still date more or less from Hanseatic or court tradition as well as from the nineteenth and early twentieth centuries.

The German Democratic Republic (GDR) after its founding in 1948 abolished the private market, nationalized land and built structures, while the government took responsibility for their regulation. But the priority of the new socialist state was the improvement of its own economy to provide the population with basic goods and to make the productive sector independent from the Western market. Money for housing was scarce. Projects of the mid-1950s, like the *Stalinallee* in East Berlin (riposted in West Berlin with the International Building Exhibition, the 'Interbau'

at Hansaviertel, Tiergarten with buildings by internationally renowned architects such as Aalto or le Corbusier), or the construction of the new industrial town of Stalinstadt (later Eisenhüttenstadt) (May in press), were outstanding and proved to be excellent propaganda. As early as 1950, however, East Berlin developed a list of 16 principles of town planning as a socialist riposte to the Athens Charter, at the insistence of Moscow. According to these principles, a city had to be compact and of a traditional national, or regional style.

Only after Stalin's death was the GDR able to adopt a functional town plan and a modern language in architecture. This led to industrialization in the building sector. Lack of building materials and a dramatic shortage of housing forced the government to rationalize the building process radically. Nationally organized housing production units with regional competence, were from the early 1960s responsible for planning and construction. The planning competencies of local authorities were limited. Urban reconstruction under these circumstances meant erecting new prefabricated housing schemes in rows of multi-storey blocks on the urban fringe and moving people to them. The existing building stock was not renewed as no building material or a sufficiently skilled labor force were available. Private initiative was therefore more or less impossible. It was intended to replace the existing but decaying late-nineteenth century urban housing stock at a later date with new prefabricated blocks. Inspired by West Germany, a few test projects started around 1980 and some modified concrete slabs were produced to integrate better into the urban fabric. This proved too expensive. The policy to replace old housing stock as quickly as possible with blocks of prefabricated concrete units also met with limited success. The reprehensible treatment of the urban environment and the appalling decay of the cities was at the root of the people's movement in 1989.

See also: Accidents, Normal; Disasters, Sociology of; National Socialism and Fascism; Reconstruction/Disaster Planning: Japan; Reconstruction/Disaster Planning: United States; Transportation: Supply and Congestion

Bibliography

- Beyme von K 1987 *Der Wiederaufbau. Architektur und Städtebaupolitik in beiden deutschen Staaten*. Pieper, Munich, Germany
- Diefendorf J M 1993 *In the Wake of War: The Reconstruction of German Cities After World War II*. Oxford University Press, New York, pp. 3–17
- Durth W 1986 *Deutsche Architekten. Biographische Verflechtungen 1900–1970* (German Architects. Biographical Interlacings 1900–1970). Vieweg, Braunschweig/Wiesbaden, Germany

- Durth W, Gutschow N 1988 *Träume in Trümmern. Planungen zum Wiederaufbau zerstörter Städte im Westen Deutschland 1940–1950* (Dreams in rubble. Plans for the reconstruction of destroyed cities in West Germany 1940–1950) 2 vols. Vieweg, Braunschweig/Wiesbaden, Germany
- Durth W, Düwel J, Gutschow N 1998 *Ostkreuz, Aufbau, Architektur und Städtebau der DDR* (Architecture and town planning of the GDR) 2 vols. Campus, Frankfurt, Germany
- Fischer F K 1990 German reconstruction as an international activity. In: Diefendorf J M (ed.) *Rebuilding Europe's Bombed Cities*. Macmillan, Basingstoke, UK, pp. 131–44
- Göderitz J, Rainer R, Hoffmann H 1957 *Die gegliederte und aufgelockerte Stadt*. Wasuruth, Tübingen, Germany
- May R in press Planned city Stalinstadt: A manifesto of the early German Democratic Republic. *Planning Perspectives*
- Petz von U 1990 Urban renewal under National Socialism: Practical policy and political objectives in Hitler's Germany. *Planning Perspectives* 5(2): 169–87
- Petz von U 1997 Ruhr Basin 1920 In: Bosma K, Hellinga H (eds.) *Mastering the City*, 2 vols. NAI, Rotterdam, The Netherlands, vol. I, pp. 56–65, vol. II, pp. 184–91
- Reichow H B 1948 *Organische Stadtbaukunst, von der Großstadt zur Stadtlandschaft*. Westermann, Braunschweig, Germany
- Reichow H B 1959 *Die autogerechte Stadt. Ein Weg aus dem Verkehrschaos*. Otto Maier, Ravensburg, Germany

U. von Petz

Reconstruction/Disaster Planning: Japan

Japan has been prone to natural disasters, especially as it lies in the Pacific volcano zone, but among the most destructive have been frequent earthquakes. Moreover, its traditional urban dwellings, typically wooden buildings made with thatched roofs, are easily inflammable, making cities liable to major conflagrations. But it was not until the Meiji Restoration in 1868 and the subsequent era of rapid industrialization and urban growth that serious thought was given to constructing urban structures to survive such disasters. Since then, the development of reconstruction and disaster planning has been interrelated closely with that of modern urban planning, which in turn has been influenced by changes in the political, economic, and social environment.

1. Reconstruction/Disaster Planning in the Meiji Period

One of the earliest examples of reconstruction and disaster planning after the Meiji Restoration was the *Ginza district project* following the large 1872 fire in Tokyo that burnt out some 3,000 dwellings. A radical proposal was put forward to rebuild the entire district with Western-style buildings made of bricks. It was an

attempt to set a precedent in transforming the traditional urban structure of the capital into a modern one, reflecting the Meiji government's strenuous efforts to swiftly Westernize the country in order to facilitate the revision of the unequal treaties with the Western powers. Individual buildings were designed by expert foreign advisers employed by the government, and the project was completed in 1877. Although the financial difficulties allowed just half the planned area to be rebuilt with bricks, the newly reborn Ginza district quickly became popular as the most fashionable shopping center in the country.

However, there was a gradual realization that a more comprehensive and yet realistic approach than this kind of extravagant but essentially superficial Westernization of the capital was necessary. While this recognition led to the urban improvement projects for Tokyo, mainly consisting of road, water supply, and drainage works, that started in 1888 and lasted for 30 years, there were important prior developments in the form of reconstruction and disaster planning. In the wake of four big fires within a few months during the winter of 1880–81, the governor of Tokyo proclaimed regulations mandating fire-prevention measures in the central wards of Tokyo. Twenty-two streets and canals were designated as areas along which buildings had to be constructed of either bricks or stone, or as godowns made of clay, with roofs of incombustible materials. Extensive slum clearance was also carried out in the Kandahashimotocho district following one of the four big fires in 1880–81, though it meant the merciless uprooting of inhabitants from their old community, admittedly one of the major slums in Tokyo at that time. Meanwhile, the demand for national planning legislation intensified, and the *Urban Planning Act* was enacted in 1919. In practical terms, its prime concern was to check indiscriminate urban expansion by controlling future development through such measures as land readjustment projects, a principal feature of Japanese urban planning based on street widening and realignments. But an opportunity for its application to built-up areas soon came about in a tragic way.

2. Reconstruction after the Great Kanto Earthquake

On September 1, 1923, a severe earthquake of magnitude 7.9 hit the southern Kanto region including Tokyo and Yokohama. As it occurred just before noon, when people were cooking meals, the Great Kanto Earthquake caused massive fires that claimed more than 100,000 lives and destroyed 460,000 dwellings. The total damage amounted to approximately six billion yen, four times the amount of the national budget. The government set to reconstruction planning immediately, with Shinpei Goto, the Home

Minister, as the man in charge. He tackled the task enthusiastically, not least because he saw an opportunity to restore his reputation. The renovation plan for Tokyo made public five months earlier when he was the mayor of Tokyo had been criticized heavily as too extravagant, and he had been derided as a braggart. Within a week, he established basic principles for reconstruction, which included proposals for a powerful Ministry for Reconstruction and public purchase by the state of all land affected at an estimated cost of 3 billion yen, reflecting his determination to rebuild in a bold and comprehensive way. However, neither of these proposals was put into practice: opposition from other ministers and ministries to a mighty reconstruction ministry led to the abandonment of the idea. Public purchase of land was out of the question not only on financial grounds but also on political ones, for the landed interests had a strong say within the government as well as in the Diet. Moreover, the budget was reduced to 468 million yen at the 47th extraordinary Diet session in December, three months after the disaster.

Nonetheless, reconstruction projects with actual working expenses of 847 million yen over seven years between 1924 and 1930 meant grand engineering on an unprecedented scale. This becomes clear when they are compared with the earlier urban improvement projects for Tokyo which took 30 years but cost only 42 million yen. Approximately 80 percent of the afflicted area in Tokyo, 3,600 hectares, were covered by reconstruction land readjustment projects under the *Ad Hoc Special Urban Planning Act*, which made possible large-scale land readjustment in built-up areas. As a result, Tokyo's urban structure, of which the core had been Edo-era construction modified by the Meiji-era urban improvement projects, was transformed. This change laid the foundation for the infrastructure of the capital's central area until the 1960s. The reconstruction land readjustment projects gave rise to 52 principal roads (114 kilometers in total length), 122 subsidiary principal roads (139 kilometers), 55 parks (42 hectares), and a number of structures built of reinforced concrete to render them earthquake- and fire-proof. The latter included 121 schools and 424 bridges. Of particular interest were the activities of *Dojunkai*, an extra-departmental body to the Home Ministry set up in 1924. It provided, among other things, 2,501 apartments of reinforced concrete with the most up-to-date communal facilities, built on 15 estates between 1926 and 1933. These became the forerunner of the public apartments built to alleviate major urban housing shortages in the postwar years.

It should be also noted that reconstruction planning after the Kanto earthquake brought important advances in planning techniques and the training of technocrats and other planning personnel. This contributed to the substantial progress made in the development of Japanese urban planning in the 1930s and 1940s. However, the main cause of this progress,

the intensification of military preparations, was soon to lead to a war that would bring massive destruction to the country.

3. Air Defense Planning and Postwar Urban Reconstruction

During the 1930s, air defense became the prime concern of Japanese urban planning, as was duly shown in the *revised Urban Planning Act* of 1940, which made air defense an integral part of urban planning. Emphasis was placed upon the provision of open spaces and wider streets for fire-preventive purposes. Admittedly, urban planning benefited from this. For instance, the state purchase of an area of 13,730 hectares was decided in 1939 to provide a number of parks as air defense measures. Out of these, six large parks in Tokyo with a total of 640 hectares were established as part of the ceremony commemorating the 2,600th anniversary of the founding of the Japanese Empire. Evacuation of buildings to make room for air defense roads also led to the building of wide thoroughfares in cities such as Nagoya, Hiroshima, and Kyoto.

However, the sheer intensity of air raids during the closing months of WWII was far more devastating than these measures could handle, and Japan found its urban areas in ruins when it surrendered in August 1945. One hundred and fifteen cities and towns throughout the four main islands in Japan were officially designated in November 1945 by the government as war-damaged cities. Approximately 2,300,000 houses or 20 percent of the stock in those cities had been destroyed or seriously damaged, and 63,000 hectares of land burnt out. The damage was particularly extensive in the two largest cities, Tokyo and Osaka, and these accounted for nearly 40 percent of the total of land (26.8 percent for Tokyo and 12.1 percent for Osaka) and more than 45 percent of houses (32.9 percent for Tokyo and 13.8 percent for Osaka) affected. The area affected in Tokyo was more than four times as great as that afflicted in the Great Earthquake in 1923.

Despite the extent of damage and the fact that Japan was under American occupation, official policies and plans for postwar reconstruction were formulated speedily and implemented with determination. Evidently, such destruction was regarded as having provided an opportunity for drastic urban renovation. In November 1945, the War damage Rehabilitation Board (WRB; *Sensaifukkoin*) was set up as a responsible central body. Ichizo Kobayashi, a well-known businessman, was appointed as its first president and given ministerial status. The intention underlying his appointment was to introduce the spirit of private enterprise and thus to get rid of the bureaucracy that had dominated Japanese urban planning before the war. This reflected the wave of

democratization of Japan's political, economic, and social systems. The following month, the cabinet adopted the Basic Policy Principles for the Planned Reconstruction of War damaged Areas (*Sensaichi-fukkokeikaku Kihonhoshin*). It featured nine principles in planning, according to which the local authorities drew up plans very quickly. It was recommended that the damaged areas should be covered by land readjustment projects, and that principal roads should be quite wide, even as much as 100 meters for large cities.

However, the planning process continued to be dominated by civil servants. For one thing, the WRB officials simply ignored Kobayashi, their superior, when he argued for compact rather than grandiose reconstruction. The officials insisted that plans had to be bold and comprehensive. This was perhaps in line with what cities like Tokyo wished, but many other cities thought quite differently, only to be forced to accept the officials' arguments. Osaka, for instance, had no choice but to add a couple of 100-meter-wide thoroughfares to its original plan when it was told by the WRB officials that any approval of their plan would otherwise be out of the question. Osaka officials believed that such proposals were too extravagant given the country's economic situation. In fact, the plans were subject to constant cutbacks due to the economic difficulties, culminating in the Revision (*Saikento*) in 1949 of the basic principles for reconstruction drafted in 1945. The revision was made following the *Dodge Line*, a draconian anti-inflation program implemented by the occupation authorities in 1949. The *Dodge Line* forced the government to cut spending drastically. As a result, the area covered by land readjustment projects was reduced by approximately 30,000 hectares, or 50 percent over the country as a whole, and many ambitious road and park projects were eliminated. The cuts were especially severe for big cities. In the Tokyo plan, all seven of the proposed 100-meter-wide roads were simply abandoned, and the area for land readjustment projects was reduced by three-fourths. In the Osaka plan, proposals for 100-meter-wide roads were also dropped and the extent of land readjustment projects reduced by nearly 45 percent. Moreover, there was widespread public disapproval of reconstruction planning: in large cities, the open space proposals were often simply ignored by those who erected illegal houses on the designated sites, which thus became prejudicial to planned development in the future. In many provincial cities, there emerged organized opposition by citizens to the official plans, usually reflecting discontent that their opinions were not heard in the planning process.

4. Conclusions

Nevertheless, planning for disaster-proof urban structures made some progress from the 1950s onwards, thanks to a series of legislative measures aimed at enhancing public safety, such as the *Fireproof*

Building Promotion Act of 1952 (covering the total area of 64 hectares in some 80 cities), and the *Disaster-proof Building Blocks Construction Act* of 1961 (covering 253 hectares in 101 cities). Further, the *Urban Renewal Act* of 1969 led to the redevelopment of six disaster-prevention districts covering 526 hectares in Koto Ward, a heavily populated downtown area in Tokyo. But the country is still vulnerable to disasters, as shown in the Hanshin/Awaji Earthquake in 1995. Regrettably, there was considerable anger among local residents in Kobe, the worst-hit city, about the lack of consultation in the reconstruction planning and the manner in which plans were imposed from above. As Japan cannot be immune from earthquakes, there is still an enormous need to plan for disaster-proof urban structures throughout the country, and to promote greater public participation in the process.

See also: East Asia: Environmental Issues; East Asian Studies: Economics; East Asian Studies: Politics; Japan: Sociocultural Aspects; Reconstruction/Disaster Planning: Germany; Reconstruction/Disaster Planning: United States

Bibliography

- Fujimori T 1982 *Meiji no Tokyokeikaku*. Iwanamishoten, Tokyo
- Fukuoka S 1991 *Tokyo no Fukkokeikaku*. Nihonhyoronsha, Tokyo
- Hasegawa J, Mason T, Matsumura T, Tiratsoo N 2001 *Urban Reconstruction in Britain and Japan, 1945–1955*. Luton University Press, Luton, UK
- Ishida Y 1987 *Nihonkindaitoshikeikaku no Hyakunen*. Jichitai-kenkyusha, Tokyo
- Koshizawa A 1991 *Tokyo no Toshikeikaku*. Iwanamishoten, Tokyo
- Mikuriya T 1984 *Shutokeikaku no Seiji*. Yamakawashuppan, Tokyo
- Zenkokushigaichisaikaihatsukyokai 1991 *Nihon no Toshisai-kaihatsushi*. Jutakushinposha, Tokyo

J. Hasegawa

Reconstruction/Disaster Planning: United States

Reconstruction/disaster planning includes plans prepared specifically to guide reconstruction following disasters, plans prepared to mitigate losses from hazards through actions taken prior to the onset of disasters, and hazard-mitigation elements of more general urban and regional comprehensive plans. Although the content of each type of plan differs, the critical components and key features of the planning process are similar. In this article, the basic elements of reconstruction/disaster plans and the processes involved in preparing them are described.

1. The Problem

World-wide, the risks from natural and related technological hazards has been increasing at an accelerating pace. Population growth and migration toward at-risk areas, increased development and urbanization of hazard zones, long-term climactic trends, and other factors have combined to expose populations to losses of life and property. Although improved forecasting and warning systems have generally reduced the number of fatalities associated with disasters, particularly in developed countries, annual monetary losses everywhere are staggering.

This problem is particularly severe in the United States, where losses of property and income attributed to natural hazards are estimated at \$500 million per week (Mileti 1999). Large disasters, such as Hurricane Andrew in 1992, the Midwest floods of 1993, and the Northridge earthquake of 1994 have resulted in damages ranging from \$15 to \$44 billion (Board on Natural Disasters 1999). Worst case scenarios suggest that much larger disasters are probable. A \$200 billion earthquake in California and a \$50 billion hurricane in Florida are distinct possibilities. In the face of this large and growing exposure to economic loss, policy makers are slowly coming to realize that urban planning may be an effective way to foster greater resilience to natural disasters.

2. Evolution of Planning for Resilience

Throughout history, communities have been located so as to limit the potential for disaster and after disasters, governments have sought to rebuild in ways to reduce the likelihood of future catastrophes. Italian hill towns, for example, were located on high ground to avoid flooding and diseases such as malaria that were prevalent on the valley floors. After the Great Fire of London in 1666, Charles II issued a proclamation that called for rebuilding in ways that would reduce the potential for future disasters. In more recent times, however, the lessons of the past were forgotten. After the San Francisco earthquake and fire of 1906, for example, the city was quickly rebuilt with little attention to hazard mitigation. In fact, some of the largest losses in the Loma Prieta earthquake of 1989 occurred in the Marina District, which is located on debris pushed into the bay after the 1906 disaster.

Modern thinking about planning for resilience to hazards can be traced to the geographer Gilbert White, who first proposed the concept in 1936 in an article in the *Planners Journal* (1936). White's ideas were not acted upon until twenty years later, when the Tennessee Valley Authority prepared a flood-plain management plan for Bristol, Tennessee-Virginia in 1956. However, because few communities followed the TVA lead, state governments, beginning with California in 1972, began to require that local governments incorporate hazard mitigation provisions in their com-

prehensive planning. At last count, half of the states mention natural hazards in their planning enabling statutes, and eleven states, like California, actually mandate local government planning for the mitigation of natural hazards (Schwab et al. 1998).

Plans that address natural hazards take two forms. Post-disaster reconstruction and recovery plans focus on policies and actions to bring about rapid recovery from a disaster and to guide reconstruction of damaged structures and infrastructure to increase resilience to future disasters. Hazard mitigation plans focus on policies and actions that can prevent damage and loss of life before disasters occur. Each type of plan may be formulated as a stand-alone plan or as an element of a broader comprehensive plan, but the latter approach is probably best, since it embeds hazard mitigation in regular planning processes that have legal standing.

3. Post-disaster Reconstruction and Recovery Plans

The post-disaster period creates a window of opportunity to make changes that can improve resilience to future disasters and accomplish a number of other community goals. But, these opportunities are often lost due to the need to make decisions quickly in an atmosphere of confusion and strong political pressures to return quickly to normalcy. Post-disaster reconstruction and recovery plans provide a means to ensure that reconstruction occurs in ways that contribute to long-term sustainability. These plans focus on decisions, made in the short-term recovery period after a disaster, that have long-term repercussions. Schwab and his associates (1998) call attention to four groups of decisions that are particularly important to address in plans: (1) sites for temporary housing, relocation of damaged businesses, and dumping of debris; (2) closure of roads and bridges; (3) restoration or relocation of critical infrastructure; and (4) reconstruction or relocation of dwelling units and nonconforming uses that have suffered substantial damage. In addition, the plan should clarify roles and responsibilities in recovery, identify potential sources of financing, and minimize wasteful and conflicting efforts (Mileti 1999).

Experts such as Mileti (1999) and Schwab (1998) argue that postdisaster recovery and reconstruction plans will be most effective if they are prepared in accordance with the following principles. *Consensus building* through community involvement should be sought to avoid community conflict over reconstruction strategies. *Information* should be provided about the hazard, exposed populations, buildings, and infrastructure, resources available for recovery after the disaster, and governmental powers and programs to aid recovery. *Organization* should be addressed through identification of topic-specific groups that can

be put in place immediately after a disaster (in particular, designation of an official rebuilding/restoration group is critical). *Procedures* should vary as little as possible from those followed during pre-disaster periods, although permitting and code-review procedures should be streamlined in order to expedite rebuilding. *Damage evaluation* protocols should be in place, such as protocols for damage assessment by building inspectors and standards for post-disaster repair and construction. *Finances* for reconstruction should be identified including funding from both governmental and private sources. Finally, provision should be made to adjust the local government's *planning style* so that more flexible and cooperative procedures replace the formal, deliberative rule-oriented procedures that characterize pre-disaster decision-making.

4. Hazard Mitigation Plans

The American Planning Association recommends in its *Growing Smart Legislative Guidebook* (1998) that local government comprehensive plans include an element on hazard mitigation. A number of states (e.g., California, Florida, North Carolina for coastal areas) not only recommend this, they require it of every community (see Burby et al. 1997). Through a comprehensive plan's safety element, hazard mitigation is addressed directly and hazard considerations are also included in portions of the plan dealing with housing, transportation, protection of environmentally sensitive areas, natural resource protection, and capital improvements. Hazard mitigation plans, as stand-alone documents, also have been prepared to address a variety of hazards. In either case, the preparation of a plan involves similar steps.

The first step is to secure a resolution by the local governing body to put someone in charge of the planning effort and to establish a planning committee or task force. Official sanction for the planning process is essential to obtain the participation of other agencies and to tap into their resources and expertise. Participating agencies should include planning/community development, engineering/public works, finance, public safety (police, fire, emergency management), code enforcement, parks and recreation, public information, and a representative of the governing board.

One of the planning group's first jobs is to build a constituency for hazard mitigation and consensus about the need to take action. The lack of such a constituency, coupled with local officials' own lack of concern about hazards, has resulted in minimal attention to hazards in past planning (e.g., Burby 1998, Mileti 1999, Platt 1999). This is best accomplished by involving key stakeholders in the planning process from the beginning. The public involvement effort should target people and organizations exposed to hazards, neighborhood and civic organizations, the

business community, managers of critical facilities, land developers and others whose decisions affect urban development. In addition to gaining support for the planning effort, citizen involvement helps secure local knowledge about hazards, insures that information gathered in the planning process is relevant to issues citizens view as important, and helps to prevent misunderstandings and gain consensus for the plan's recommendations.

The next step in the planning process is to develop information about the problems hazards pose and the capacity of the community to build resilience to hazards. Hazard assessment includes identifying hazards (magnitude, severity, frequency, and causative factors) and mapping the areas of a community that are exposed to them. Losses experienced in past disasters are identified, and vulnerability to loss in future disasters is estimated by examining potential property damages in hazard events of various magnitudes.

Analysis of capacity helps in identifying hazard reduction measures that are feasible for a community and actions that are needed to increase capacity, such as adding staff with hazards expertise or developing dedicated revenue sources to fund mitigation. Other topics for capacity analysis include funding sources, legal authority for hazard mitigation actions, and evaluation of hazard mitigation measures that are already in use. The hazard assessment and capacity analysis lead to a statement of the problem facing the community and formulation of goals and objectives. It can be combined with a visioning process to develop, in collaboration with citizens and policy makers, an image of what a hazard resilient community would look like in the future. Visioning can help develop consensus about the goals of hazard mitigation and can be useful in tying hazard mitigation to other community concerns such as economic development and environmental protection.

Once the problem is understood and a vision of a hazard resilient community is agreed upon, preparation of the hazard mitigation plan involves making choices among possible strategies to build resilience. Six sets of possible measures should be reviewed.

(a) *Preventive* policies and actions are designed to limit the exposure of new development to losses from hazards (e.g., building code standards, zoning, and land acquisition).

(b) *Property protection* policies and actions limit the exposure of existing development to hazards on a parcel-by-parcel basis (e.g., retrofitting, relocation, acquisition and clearance, and insurance).

(c) *Structural protection* policies and actions limit the exposure of existing development to hazards (principally flooding) on an area-wide basis (e.g., flood control dams, levees, sea walls, and channel modifications).

(d) *Emergency services* policies and actions lessen the impact of a hazard after its onset (e.g., warning,

evacuation, flood fighting, fire fighting, and emergency sheltering).

(e) *Natural resource protection* policies and actions preserve and restore natural areas that also help lessen the magnitude of hazards (e.g., wetland protection, erosion and sedimentation control, and sand dune protection).

(f) *Public information* policies and actions build awareness of hazards and knowledge of individual actions to mitigate them (e.g., hazard maps, library resources, and technical assistance).

The strategies that have been identified as useful in creating a sustainable community that is resilient to hazards are evaluated in terms of their effectiveness, efficiency, equity, and feasibility (administrative, technical, financial, legal, and political). The goal is to formulate a program that draws wide support, addresses existing problems from hazards, and also prevents new ones from developing. The final plan includes, at a minimum, five components. These include

(a) a description of how the plan was prepared and who participated in its preparation,

(b) an assessment of hazards facing the community,

(c) the vision of a hazard-resilient community,

(d) recommendations of policies and actions, and

(e) identification when they will be undertaken, who is to undertake them, how they will be financed, and who will monitor performance (see Wetmore and Jamieson 1999).

5. Conclusion

The preparation of a post-disaster-reconstruction/hazard-mitigation plan provides the technical analysis, coordination among governmental agencies, and community involvement and consensus building necessary to increase resilience to natural and related technological hazards. Communities with coherent plans are able to build human settlements that are resistant to disasters and able to recover from disasters quickly. The plan-making processes and principles discussed in this article are proven to be effective. But state and local governments have been slow to prepare and adopt the needed plans. The challenge for governments at every level is to act on the knowledge that is available to reduce or even reverse the steady rise in destruction and human suffering from disasters.

See also: Accidents, Normal; Disasters, Sociology of; National Socialism and Fascism; Reconstruction/Disaster Planning: Japan; Reconstruction/Disaster Planning: United States; Transportation: Supply and Congestion

Bibliography

American Planning Association 1998 *Growing Smart Legislative Guidebook*. Chapter 7. Local Comprehensive Plans. American Planning Association, Chicago

Board on Natural Disasters, National Research Council 1999 Mitigation emerges as major strategy for reducing losses caused by natural disasters. *Science* **284**: 1943–7

Burby R J (ed.) 1998 *Cooperating with Nature: Confronting Natural Hazards with Land-use Planning for Sustainable Communities*. Joseph Henry Press, Washington, DC

Burby R J, Beatley T, Berke P R, Deyle R E, French S P, Godschalk D R, Kaiser E J, Kartez J D, May P J, Olshansky R, Paterson R G, Platt R H 1999 Unleashing the power of planning to create disaster resistant communities. *Journal of the American Planning Association* **65**(3): 247–58

Burby R J, May P J, Berke P R, Dalton L C, French S P, Kaiser E J 1997 *Making Governments Plan: State Experiments in Managing Land Use*. Johns Hopkins University Press, Baltimore, MD

Godschalk D R, Kaiser E J, Berke P R 1998 Integrating hazard mitigation and local land use planning. In: *Modernizing State Planning Statutes: The Growing Smart Working Papers*, Vol. 2. PAS Report 480–1, pp. 57–81. American Planning Association, Chicago

Mileti D 1999 *Disasters by Design: A Reassessment of Natural Hazards in the United States*. Joseph Henry/National Academy Press, Washington, DC

Platt R H 1999 *Disasters and Democracy: The Politics of Extreme Natural Events*. Island Press, Washington, DC

Schwab J, Topping K C, Eadie C C, Deyle R E, Smith R A 1998 *Planning for Post-disaster Recovery and Reconstruction*. PAS Report 483–484. American Planning Association, Chicago

Wetmore F, Jamieson G 1999 Flood mitigation planning: The CRS approach. *Natural Hazards Informer* **1** (July): 1–15

White G F 1936 Notes on flood protection and land use planning. *Planners Journal* **3**: 57–61

R. J. Burby

Reconstructive Memory, Psychology of

Reconstructive memory refers to the idea that retrieval of memories does not occur in some completely accurate form, as a video-recorder might replay a scene, but rather that recollection of memories is a process of trying to reconstruct (rather than replay) past events. Although the reconstruction can be quite accurate, the processes responsible can also introduce errors during retrieval. In fact, systematic errors in memory are the primary evidence for its reconstructive nature. This article provides an overview of the evidence that remembering is reconstructive (see also *Eyewitness Memory: Psychological Aspects*; *False Memories, Psychology of*; *Memory for Meaning and Surface Memory*).

1. Background

The hypothesis that remembering should be viewed as reconstructive dates to an important book by Sir Frederic Bartlett (1932). Bartlett contrasted reproductive memory (veridical, rote forms of memory, such as reproducing a telephone number) with recon-

structive memory and argued that the latter was more typical of our uses of memory outside laboratory and educational circumstances. That is, we are rarely faced with the task of remembering something exactly the way it happened, but more typically need only to get the essence of the event right. If a friend asks you, 'What did Kathleen tell you last night?' the request is not for a literal rendering of last night's conversation, but rather for the gist of what was said.

Bartlett argued that perceiving and comprehending events do not simply happen automatically, but that every event of comprehension involves the mental construction of one's understanding of the event in the world. We all struggle with the 'effort after meaning' in comprehending the events in the world around us. If encoding or perceiving is a construction, then when one wants to recall the events later, the attempt is to reconstruct the event.

Bartlett argued that recollection is guided by schemas, or general organizing structures, which aid encoding and retrieval. If you see a scene at the beach and are asked to recall it later, you might recall seeing a beach umbrella even if none was present in the actual scene itself, because it is consistent with the general schema of items that belong in a beach scene. Therefore, although schema can aid encoding and retrieval of information, they can also lead to errors.

A large amount of research is consistent with the idea that remembering is reconstructive. The standard textbook account holds that certain forms of remembering are reconstructive whereas others are reproductive. However, a strong case can be made that all remembering is reconstructive. What appears to be reproductive memory occurs in situations in which the reconstruction is quite accurate (Roediger and McDermott 1995). This article considers various forms of memory as they are experimentally studied and discusses evidence for reconstructive processes at work.

2. Sensory and Short-term Memory

The greatest challenge to the claim in the previous paragraph perhaps comes from studies of sensory memory, the borderline between perceiving and remembering. However, many theorists have argued that perceiving is itself a constructive activity, with fascinating phenomena of visual and auditory illusions used as evidence for this claim (see Gregory 1998, among many others). Sensory memory refers to the temporary persistence of information that has struck the senses, which lingers briefly as it is being comprehended. Visual persistence is called iconic memory and auditory persistence is labeled echoic memory. It would seem that iconic memory—essentially a fleeting after-image of the scene from the outside world—would surely be a form of reproductive memory. Yet even in these situations errors arise, showing that

at least the retrieval processes from this type of memory involve reconstruction. For example, in Sperling's studies of iconic storage, in which people had to report letters that they had briefly seen on a screen, a common error when people missed a letter was to report another letter than either looked like or sounded like the original letter (Sperling 1967). This type of error indicates that people may code even such simple items as single letters into visual patterns and associated sounds. When people miss a *b* in these experiments, they may substitute a *v* (which sounds like *b*) or a *p* or *d*, which share both similar appearance (a long line and a curve) and similar sound (the letters rhyme with *b*). Even reports from iconic memory may show reconstructive tendencies.

Short-term memory or working memory (see related entries on these topics) hold information for longer than sensory memories, but people are still ordinarily accurate in retrieving information from short-term stores, if no interference occurs. Does this accuracy reflect a rote, reproductive process? The answer seems to be no, because when the short-term memory system is challenged by having people operate under, for example, fast rates of presentation, errors occur. Errors are often (but not always) phonological in nature. That is, if someone tries to recall letter strings and misses a letter, similar sounding letters are confused (Conrad 1964). In the case of words, those that share visual and phonemic (sound-based) features are confused (Crowder 1976, Chap. 4). Therefore, even though short-term memory processes are often considered quite accurate (and they can be), recall in these situations typically occurs under conditions (short unfilled delays between study and test) that make for accurate reconstructions. Stress the system by presenting material fast, or by creating interference, and the characteristic error pattern indicative of a reconstructive process appears.

3. Long-term Memory

By far the greatest amount of research on reconstructive memory has occurred in the domain of long-term, episodic memory. Many different kinds of experiments have been conducted and we will consider some of the main types here, which have used a wide variety of materials: word lists, pictures and scenes, sentences, prose passages, and videotape. Research with all these materials reveals remembrance to be quite prone to error and to filling in the gaps.

3.1 Word Lists

Roediger and McDermott (1995) developed a paradigm in which people heard lists of 15 words that were generated as associates to another word, which was not itself presented. So, for example, subjects heard

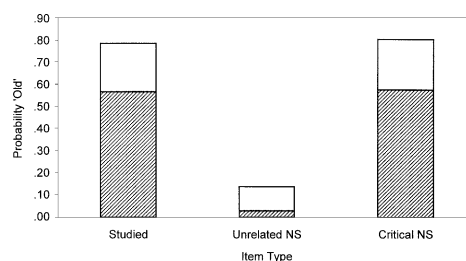


Figure 1

False recognition and false remembering. Mean probability of recognizing (total bar) and remembering (cross-hatched bar) studied words and lures. Unrelated lures were rarely recognized, but critical lures were recognized and remembered at about the same levels as studied words. People remember the occurrence of the critical items at the same level as words that were actually studied. Data are from Roediger and McDermott (1995, Experiment 2)

the list 'hot, snow, warm, winter, ice, wet, frigid, chilly, heat, weather, freeze, air, shiver, Arctic, frost.' All these words are associated to the word 'cold,' which was the generator of the list, but 'cold' was not presented. Immediately after hearing the list, subjects were asked to recall the items and were urged to be sure that the words they recalled had actually been presented in the list. Despite this warning not to guess, Roediger and McDermott (1995) found that their subjects recalled missing words like 'cold' at very high levels; in fact, probability of recalling the critical item often equals probability of recall of list items in this paradigm. After recall of many lists, subjects received a recognition test in which they were instructed to pick out only items that had actually appeared in the list. The test included words from the list ('winter,' 'chilly'), completely unrelated words ('automobile'), and the critical words such as 'cold.' Subjects examined each word and, if they judged it to be old (or studied), they were asked to make a further remember/know judgment. This judgment (Tulving 1985) asks people to report if they can remember the moment of the item's occurrence in the study list (a remember judgment) or if they just know it was on the list but cannot remember its moment of occurrence (a know judgment).

The recognition results are shown in Fig. 1, where it can be seen that subjects were good at distinguishing studied words from unrelated lures or distracters. However, the critical lures such as 'cold' were recognized at the same level as were the words in the list, about 80 percent. People recognize the related but nonstudied word at the same level as they do items in the list. In addition, subjects claimed to remember the moment of occurrence of the critical words (that, of course, were not actually presented) at the same level as for the words that were actually studied). In short,

in this paradigm subjects recall, recognize, and claim to remember the occurrence of words that were never presented at the same level as of the words that really were presented, which provides a compelling demonstration of illusory memory. The implication is that subjects reconstruct the list using their general knowledge and yet insert a quite specific memory—they remember 'cold' as having occurred—because it fits with the general theme or schema of the list.

3.2 Pictures and Scenes

Miller and Gazzaniga (1998) developed a similar procedure with common pictorial scenes, ones that had been covers of the *Saturday Evening Post* magazine. They showed people the scenes with a critical object from the scene removed. To use our earlier example, people might view a beach scene with a beach umbrella either present in the scene or absent from it. Miller and Gazzaniga (1998) found that when people were given recognition tests on objects in the scene, they were very likely to recognize the object as having been present when in fact it was not. Again, in reconstructing the scene, the object is inserted and seems part of the memory.

3.3 The Misinformation Paradigm

One of the most popular paradigms for studying reconstructive processes in memory was developed by Elizabeth Loftus and her colleagues (e.g., Loftus et al. 1978) and is called the misinformation paradigm. It is modeled after the events that an eyewitness to a crime might experience. In the lab, subjects view a set of slides or a videotape, then later they receive questions or a narrative. In Loftus et al. (1978), subjects witnessed an automobile accident that occurred at an intersection where a car failed to stop at a stop sign. In the later series of questions, subjects would be asked either one of three types of question: 'Did the car stop at the traffic sign?' (control condition), 'Did the car stop at the stop sign?' (consistent information condition) or 'Did the car stop at the yield sign?' (inconsistent or misleading information condition). This question served as the experimental manipulation for a question asked at a later point in time: 'What kind of sign was at the intersection, stop or yield?'

Results of this experiment are shown in Fig. 2, where it can be seen that, relative to the control condition in which the question was posed in a neutral manner, wording of the question that was consistent with the original information improved recognition whereas the conflicting information reduced recognition. This last finding is the misinformation effect—wrong information received after an event is often incorporated into one's memory for the event. This misinformation effect has been demonstrated in

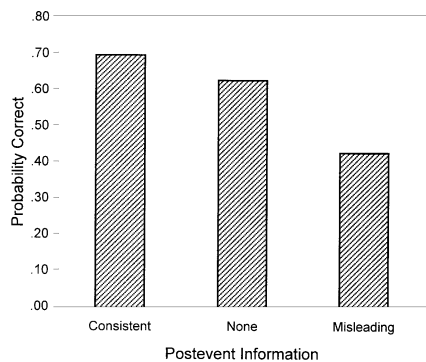


Figure 2

Probability correct (on a two-alternative forced choice recognition test) as a function of whether subjects had received information consistent with what they had previously seen, information inconsistent with what had been seen, or no specific postevent information. Misleading information after the event can reshape memory for the event. Data from Loftus et al. (1978, Experiment 2)

hundreds of experiments, although the magnitude of the effect depends on the exact nature of the test.

The general principle to arise from experiments on the misinformation effect is that memory for an event is not encapsulated in time the way the event itself is. Rather, information provided after the event can modify our memories for the event itself. Consistent information improves our later reconstruction, whereas conflicting or misleading information is harmful. When we are trying to reconstruct events from our past, we draw on all the information and resources at our disposal and the reported memory may include 'facts' that were not really part of the event, but incorporated into it later. Practically, if police or other people ask suggestive questions or embed wrong information into questions asked of witnesses after the event, then the witnesses' later recollections and testimony might be tainted (see *Eyewitness Memory: Psychological Aspects*).

3.4 Prose Recall

Perhaps the greatest body of research demonstrating reconstructive processes in retrieval has been carried out in prose recall. Brewer (1977) examined pragmatic implications in retention of sentences. A pragmatic implication is made when the person reading or hearing a sentence infers something that is neither explicitly stated nor logically implied by the sentence (see *Inferences in Discourse, Psychology of*). For example, 'The karate champion hit the cinder block' implies that the block was broken. However, this is a pragmatic implication; it is perfectly possible that the

block was struck but not broken. After people studied sentences of this type mixed with other sentences, they were given the first parts of the sentences as cues to recall the entire sentence. Brewer showed that in recall of sentences with pragmatic inferences, people were more likely to recall the implied verb than the verb actually used in the sentence. When reconstructing the action, they remember the champion as breaking the cinder block, not just hitting it. This result reflects a kind of inferential momentum, with people going beyond the literal truth of a statement to infer the events that probably occurred next. Often such inferences may be correct, but they are not necessarily so. When the inference is strongly made, but false, it can generate a false memory.

Sulin and Dooling (1974) also studied people's remembrance of implications as facts. In one case, they presented subjects with short paragraphs about a troubled girl and later tested for recognition both of ideas from the paragraph and of nonpresented ideas that were either consistent or inconsistent with the theme of the passage. The paragraphs studied by two groups of subjects were identical, with the exception of the name above the story. In one case it was Helen Keller and in the other case it was Carol Harris. One item on the recognition test asked subjects if the following sentence had appeared in the narrative: 'She was deaf, dumb, and blind.' Students who had read the story about Helen Keller frequently made the error of saying that this sentence had appeared in the passage, whereas those reading about Carol Harris almost never made this error. In trying to reconstruct whether the sentence may have been in the story, the knowledge about Helen Keller caused subjects to fill in what 'must have been' in the story and led subjects erroneously to conclude that the sentence was there.

Sometimes people remember the sentences that were in a prose passage no better than sentences they have never read before, so long as the new sentences are consistent with the material that was actually presented. Bransford and Franks (1971) gave subjects short sentences such as 'The rock rolled down the hill' or 'The rock crushed the hut.' There were four simple ideas like this that, when put together, would make up the complex sentence: 'The rock rolled down the hill and crushed the tiny hut.' The experimenters gave their subjects various sentences to build up the complex idea unit. They later tested subjects by asking them to pick out sentences they had actually studied from among other sentences. Two types of distractor sentences were used: some were consistent with the idea unit but had not been studied, whereas other sentences were unrelated to the idea unit. The test sentences themselves could vary in the number of idea units present, from one to four. Subjects rated their belief that the sentences had been previously studied on a 10-point scale, from -5 (sure the sentence was new or nonstudied) to +5 (sure the sentence was old or studied). The sentences that were not consistent

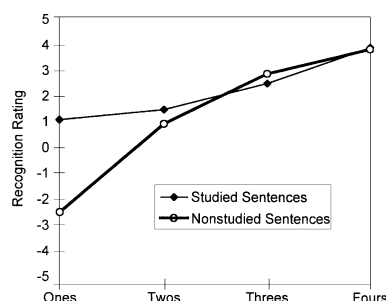


Figure 3

Mean recognition ratings for studied and nonstudied sentences as a function of number of propositions per sentence. Subjects could not distinguish sentences actually presented from those that had not been presented but that were consistent with the ideas. Data from Bransford and Franks (1971)

with the idea units were accurately rejected as not having been studied.

The results for the other two types of items are shown in Fig. 3 and reveal that the more idea units embedded in the test sentence, the higher the confidence rating that the sentence had been studied. Interestingly, this relation held both for sentences that had been studied and for new, nonstudied, sentences that were consistent with the overarching idea unit. Remarkably, except for sentences expressing only one idea, subjects were equally confident in the sentences that had not been presented in the study phase as for those that had been presented! Therefore, consistent with Bartlett's ideas, people seemed to retain the meaning of the sentences, but did not retain the literal wording of the sentences (see *Memory for Meaning and Surface Memory*).

Bartlett (1932), who began this tradition of work, was interested in repeated recollection of material. He presented subjects with a native American folktale, *'The War of the Ghosts,'* which he had them read twice. The story is a bit disjointed, has supernatural elements, and is rather difficult to understand. He found that subjects frequently made errors in recalling the story and that these errors increased over repeated retellings of the story. The stories got shorter, reflecting forgetting, but more importantly for present purposes, the story changed in systematic ways by subjects introducing errors that served to rationalize its bizarre elements. For example, people might 'remember' elements that had not been present to explain some of the unusual features in the story. Interestingly, this tendency to rationalization—to reconstruct the story more in line with the rememberers' own schema—seemed to grow over time and repeated retellings. Although Bartlett's original studies were rather informal, the work has been replicated under more systematic conditions (Bergman and Roediger 1999).

4. Conclusion

This article has reviewed selectively some key evidence that permits us to draw the conclusion that remembering the past should be viewed as reconstructing it. To draw on a metaphor from Hebb (1949), we can think of the process of remembering the past as we can conceive of paleontologists' reconstruction of a dinosaur from bone fragments and chips. The archeologist recovers a partial skeleton, but the finished product in a museum is shown as complete, with new bones added, old ones refinished or enhanced, and the entire skeleton constructed based on knowledge of what the animal must have looked like. Similarly, Hebb (1949) argued, remembering the past involves recollection of specific facts and details (the bones) that are woven together into a complete story of the event (like the skeleton). However, the story about the event might involve considerable constructive activity on the rememberer's part and the rememberer's present knowledge and goals may shape and determine how he or she remembers the past. A quote from Bartlett captures the essence of this reconstructive approach:

... the one overwhelming impression produced by this more "realistic" type of memory experiment is that human remembering is normally exceedingly subject to error. It looks as if what is said to be reproduced is, far more generally than is commonly admitted, really a construction, serving to justify whatever impression may have been left by the original. It is this "impression," rarely defined as with much exactitude, which most readily persists. So long as the details which can be built up around it are such that they would give it a "reasonable" setting, most of us are fairly content, and we are apt to think that what we have built we have literally retained (Bartlett 1932, pp. 175–6).

See also: Episodic and Autobiographical Memory: Psychological and Neural Aspects; Eyewitness Memory: Psychological Aspects; False Memories, Psychology of; Memory: Autobiographical; Memory Retrieval; Metamemory, Psychology of; Short-term Memory, Cognitive Psychology of; Working Memory, Psychology of

Bibliography

- Bartlett F C 1932 *Remembering: A Study in Experimental and Social Psychology*. Cambridge University Press, Cambridge, UK
- Bergman E T, Roediger H L 1999 Can Bartlett's repeated reproduction experiments be replicated? *Memory & Cognition* 27: 937–47
- Bransford J D, Franks J J 1971 The abstraction of linguistic ideas. *Cognitive Psychology* 2: 331–50
- Brewer W F 1977 Memory for the pragmatic implications of sentences. *Memory and Cognition* 5: 673–8

- Conrad R 1964 Acoustic confusions in immediate memory. *British Journal of Psychology* **55**: 75–84
- Crowder R G 1976 *Principles of Learning and Memory*. Erlbaum, Hillsdale, NJ
- Gregory R L 1998 *Eye and Brain*. Oxford University Press, Oxford, UK
- Hebb D O 1949 *The Organization of Behavior: A Neuropsychological Theory*. Wiley, New York
- Loftus E F, Miller D G, Burns H J 1978 Semantic integration of verbal information into a visual memory. *Journal of Experimental Psychology: Human Learning and Memory* **4**: 19–31
- Miller M B, Gazzaniga M S 1998 Creating false memories for visual scenes. *Neuropsychologia* **36**: 513–20
- Roediger H L, McDermott K B 1995 Creating false memories: remembering words not presented in lists. *Journal of Experimental Psychology: Learning, Memory, and Cognition* **21**: 803–14
- Sperling G 1967 Successive approximations to a model for short-term memory. *Acta Psychologica* **27**: 285–92
- Sulin R A, Dooling D J 1974 Intrusion of a thematic idea in the retention of prose. *Journal of Experimental Psychology* **103**: 255–62
- Tulving E 1985 Memory and consciousness. *Canadian Psychology* **26**: 1–12

H. L. Roediger III

Recovery of Function: Dependency on Age

When the brain is injured in adulthood there is a loss of function followed by a prolonged period during which there is partial recovery of some types of cognitive and motor functions. The extent of functional recovery varies with age and although injury to the infant brain can have less severe consequences than similar injury in adulthood, there are times when injury is especially devastating. These differences are related to the details of anatomical changes at different stages of brain development. The difference in functional outcome after injury at different ages is related to the injury-induced changes in neuronal structure, changes that are referred to as plastic changes. These plastic changes are similar to the neural and glial changes that are associated with learning in the normal brain. Various factors, such as experience, hormones, and drugs, modulate these plastic changes. Even these factors cannot completely restore lost functions because there are limits to brain plasticity.

1. The Nature of 'Recovery' after Brain Injury

There are many ways to damage the brain (e.g., stroke, disease, trauma) and it is rare indeed for a family not to contain a member who is affected by one of them. For example, it is estimated that about 400,000 people with new head injuries are admitted to hospitals each

year in the United States. Similarly, it is estimated that there are about 400,000 new cases of stroke every year in the United States and most surviving stroke victims experience significant cognitive loss. These figures do not include neurological disorders such as Parkinson's and Alzheimer's diseases, which together affect more than 1 percent of the population. The estimate in the United States is that about 48 million people have some sort of neurological disorder, excluding mental diseases such as schizophrenia, which are clearly a result of neural dysfunction.

The high incidence of brain injury and disease thus poses perhaps the most perplexing clinical neurological problem, which is the issue of how to repair the damaged brain. It is tempting to assume that if behavior improves after injury, then there is recovery of the original behavior. However, just because a behavior appears to recover does not mean that the lost behavior itself has returned. The problem is to determine when recovery has actually occurred. In fact, true recovery is a rare occurrence after brain injury.

We can identify at least three outcomes that could occur after brain injury. First, there may be compensation resulting from an adaptation to the lost function. Compensation could reflect a change in strategy or it may represent a substitution of a new behavior for the lost one. For example, a person who has a deficit in moving the eyes can be expected to have difficulty in reading. This difficulty can be solved, however, by moving the head, which allows restitution of function (reading) by substitution of one behavior for another. Similarly, if, following a stroke, a person can walk only with the aid of a cane, the ability to walk has returned but the original walking ability has not returned, just an adaptation to the behavioral difficulty.

A second outcome might be partial restitution of the original behavior. This could reflect recovery from some sort of nonspecific effect of the injury, such as swelling, or it may reflect a genuine return of function. Behaviors that return over a period of months are likely to reflect a partial restitution of function. This is most clearly seen in people with significant language disturbances (i.e., aphasia) in which the severe deficits in making the movements of the tongue and mouth necessary to speak may diminish leaving the person with more subtle language disturbances, such as in finding words (anomia).

Third, there is the possibility that there could be a complete restitution of the original behavior. Although it is at least theoretically possible for lost functions to return completely in a plastic brain, careful behavioral analysis shows that this is rarely, if ever, the case in either adult laboratory animals or human patients. In principle, however, it ought to be possible to reorganize the brain to allow at least some functions to completely recover. This process is most likely in the developing brain and it has been docu-

mented for over 100 years that children with damage to language areas can show an almost complete recovery of lost language functions.

2. Age-dependent Recovery

Broca first noticed in the late 1800s that damage to the speech zones of the left hemisphere during infancy did not interfere with the later development of language (for a review, see Finger and Almli 1988). Rasmussen and Milner (1977) later showed that the sparing of language functions was due to the shifting of the speech zones either to the opposite hemisphere or, in some circumstances, within the left hemisphere. In these later studies the critical factor was age: lesions of the left hemisphere language zones prior to 5 years of age changed the speech representation so that the affected language zone moved to the right hemisphere.

Similar lesions from the ages of about 6 to 12 still allowed recovery of language functions but this was accomplished by intrahemispheric reorganization (see also Krashen 1973). The studies of the recovery of language functions after infant brain injury led to the idea that damage to the infant brain led to better functional outcomes than damage after similar injury in adulthood. This conclusion was supported by the studies of Margaret Kennard in the late 1930s in which she showed that unilateral motor cortex impairments are less severe after motor cortex injury if the lesion occurs in infancy rather than adulthood (e.g., Kennard 1940). One difficulty with the studies of both Kennard and the language reorganization in children is that there are also many examples in which early brain damage results in rather severe functional loss, such as in cerebral palsy or hemiplegia resulting from unilateral cerebral injury at birth. In fact, Hebb (1949) noted that children with frontal lobe injuries in infancy showed little functional recovery in adulthood and for some behavior, such as social behavior, these children might actually have more severe deficits than adults with similar injuries.

The apparent contradiction between the recovery of language and some motor functions on the one hand, and the poor recovery of some executive, social, and other motor functions on the other hand, has been resolved by laboratory studies in which precise age at injury and the nature of the behavioral deficits have been systematically examined. A key element of these studies is that the age of the brain is measured not in relation to the birth day, but rather it is measured by its stage of anatomical development. For example, compared to humans, newborn rats are approximately equivalent to about 6-month old fetuses when they are born. Thus, if the newborn rat brain is damaged, we would anticipate functional symptoms similar to those of human infants whose brain is injured at the beginning of the third trimester. In order to compare across species we can separate the periods of brain development into five broad stages: neurogenesis, cell

migration, cell differentiation, synaptogenesis, and gliogenesis (see *Prenatal and Infant Development: Overview*).

Studies on both laboratory rats and cats have shown that if the cortex is injured bilaterally during neurogenesis, when neurons are formed, there is substantial recovery and, if the injury is not too extensive, there can be virtually complete functional recovery. In contrast, however, if the cortex is injured in the first few days after birth in the rat, which is a time of neural migration and cell differentiation, the effect is functionally devastating as animals show far more severe effects of injury than would be expected even in aging animals. This poor outcome is not a function of extent of injury, nor is it localized to particular cortical areas (see Kolb 1995).

Rather, there is something about the rat cortex shortly after birth that makes it especially vulnerable. For example, damage during this time may disturb the subsequent process of synapse formation or may even alter stem cell activity in some way. Once this phase of development is over, however, and brain development moves to the production of connections and glia, the brain shows a remarkable ability to compensate for injury. Rats or cats with cortical injuries during this phase show far better behavioral capacities in adulthood than animals with similar lesions at any other time. In fact, on some behavioral tests these animals show recovery that is virtually complete.

Importantly, this recovery is far more extensive in cognitive tasks, such as the learning of various spatial navigation problems, than it is on tests of species-typical or motor behavior (e.g., Bachevalier and Malkova 2000, Kolb and Whishaw 1989). In humans, spatial behavior appears to show less recovery than verbal behavior (Stiles 2000). This behavior-specific recovery likely reflects the relative ease of reorganizing cortical circuitry in supporting different types of behavior. The neural substrates for less plastic behavior are presumably more highly specified and constrained during development than the neural substrates for more plastic behavior. It also reflects a difference in the way the brain produces cognitive processes versus basic sensory and motor processing. Thus, the most extensive recovery is associated with lesions of the frontal areas and the least extensive recovery is associated with damage to the primary sensory areas.

3. Mechanisms Underlying Recovery of Function

The fundamental assumption here is that recovery of function can occur only if there is some type of modification of the structure of the remaining brain. This structural change is necessary if the remaining brain cells are to be recruited in some manner to control behaviors that preinjury were controlled by the lost neurons. In principle, there are three ways that the brain could make changes that might support

recovery. First, there could be changes in the organization of the remaining, intact, circuits in the brain. The general idea is that the brain could reorganize in some way to do 'more with less.' It is unlikely that a complexly integrated structure like the cerebral cortex could undergo a wholesale reorganization of cortical connectivity, but rather, recovery from cortical injury would be most likely to result from a change in the intrinsic organization of local cortical circuits in regions directly or indirectly disrupted by the cortical injury. Although it might be possible to produce significant reorganization of cortical connections in the young brain, the overwhelming evidence in experimental animals is that this is rare and it is just as likely to be associated with abnormal functioning as with recovery.

Second, there could be a generation of new circuitry. Cerebral reorganization can be stimulated either by some sort of behavioral therapy or the application of some sort of pharmacological treatment. In either event, the stimulus could influence reparative processes in the remaining brain or could enhance the production of new circuitry. Once again, it seems most likely that the induced neuronal changes would be in the intrinsic organization of the cortex. One might predict that induced neuronal changes might be more extensive than in the case of endogenous change, in part because the treatment can act upon the whole brain rather than just on affected regions.

Third, there could be a generation of new neurons and glia to replace at least some lost cells. Stem cells that give rise to the neurons and glia of the brain remain active in the brain throughout life (see *Cognitive Development in Infancy: Neural Mechanisms; Cortical Development: Transplantation and Rewiring Studies; Hypnosis, Neural Basis of*). It is thus possible that neurogenesis could be stimulated after injury, especially during development, and that these new neurons could replace those lost to injury or disease (e.g., Weiss et al. 1996). All three of these possibilities for cerebral plasticity do occur, but the likelihood of each type of change varies with age at injury.

4. Structural Correlates of Functional Recovery

When the brain is damaged, there is a loss of neurons in the injured area but, in addition, many neurons that are not directly injured lose their normal synaptic inputs. The immediate effect of this synaptic loss is (a) neuronal death, (b) atrophy of the dendritic and/or axonal fields that have lost their connections, and/or subsequent reorganization of the connections. This reorganization can be inferred by using a Golgi stain to visualize the dendritic fields of the remaining neurons. Because the dendritic branches of a neuron provide the substrate for most of a cell's synapses, a decrease (or increase) in the amount of dendritic space available is correlated with a decrease (or increase) in the number of synapses.

The hippocampal formation of adult rats has provided a good model to study synaptic replacement. When input to the hippocampus is reduced by damage to the major input pathway via the entorhinal cortex, there is a massive denervation of the hippocampal formation, which can be seen in the atrophy of the dendritic fields of the hippocampal cells in the week after injury. Beginning at about 7–10 days after the injury, there is then a slow regrowth of the dendritic fields, although the pattern of regrowth is different from the original pattern. This regrowth reflects a reorganization of the circuitry and it is correlated with a partial recovery of function. The restructuring of the synaptic organization of the hippocampus occurs primarily because the remaining inputs to the hippocampus expand to occupy the regions of lost inputs.

A similar process appears to occur after restricted neocortical injury in adulthood as well, although the changes are less dramatic than in the hippocampus and there is much less functional recovery (for a review, see Kolb 1995). In contrast, however, there is extensive reorganization after cortical injury during development, and this reorganization is correlated with functional outcome. Golgi analyses of cortical neurons of rats with perinatal lesions consistently show a general atrophy of dendritic fields and a drop in spine density across the cortical mantle. In contrast, rats with cortical lesions around 10 days of age show an increase in dendritic fields and an increase in spine density relative to normal control littermates. Thus, animals with the best functional outcome show the greatest synaptic increase whereas animals with the worst functional outcome have a decrease in synapses relative to control animals.

But infant animals not only show changes in local cortical circuits but also in more distant cortical connectivity. Perhaps the most extensive studies of changes in cortical connectivity are those showing that after unilateral motor cortex lesions in infant rats or cats there is a major expansion of the ipsilateral corticospinal pathway from the undamaged hemisphere, which is correlated with partial recovery of skilled forelimb use (e.g., Castro 1990). These expanded connections do not reflect the growth of new pathways, however, but rather reflect the failure of superfluous pathways to die off during development. That is, in development there is normally a period during which many pathways are eliminated but after some types of brain injury these pathways are maintained (for a review, see O'Leary 1992). The presence of these abnormal pathways may support some types of functional recovery but it may also interfere with the normal functioning of remaining cortical functions. This possibility has been termed 'crowding' to reflect the idea that the normal functions of a cortical region can be crowded out by the development of abnormal connections (e.g., Teuber 1975). Thus, there may be recovery of some types of behavior, such as

motor functions, at the expense of other functions (e.g., Kolb et al. 2000). This appears to be especially likely if the injury occurs during the period of neuronal migration and unlikely if it occurs during the period of maximal synaptogenesis.

As noted earlier, synaptic change may also occur after injury because the brain generates new neurons. Although this is likely to be rare in the adult, it does occur during development, but the generation of these neurons varies with precise embryonic age and location of injury. If it is damaged during the time of neuronal generation, the brain can compensate by generating rather large numbers of new cells that appear to support at least partial restitution of function (e.g., Hicks et al. 1984). It is not known if this regenerative process occurs after injury to specific regions of the brain but it is likely a general phenomenon. Thus, if the brain is injured while it is still producing neurons, it appears capable of generating replacements (e.g., Kolb et al. 1998a). In contrast, if the brain is injured during neuronal migration there appears to be little generation of new neurons. If the same injury occurs during maximal glial formation, there is neuronal generation. This generation is selective, however, and occurs only if the midline telencephalon is damaged, namely the olfactory bulb or midline cortex (e.g., Kolb et al. 1998b). The newly generated tissue is functional and animals regain at least partial function.

See also: Adolescence, Sociology of; Adolescent Behavior: Demographic; Adolescent Development, Theories of; Adolescent Health and Health Behaviors; Adolescent Vulnerability and Psychological Interventions; Adolescent Work and Unemployment; Adolescents: Leisure-time Activities; Adult Development, Psychology of; Antisocial Behavior in Childhood and Adolescence; Brain Damage: Neuropsychological Rehabilitation; Brain Implants and Transplants; Cerebral Cortex: Organization and Function; Childhood and Adolescence: Developmental Assets; Cognitive Development in Childhood and Adolescence; Friendship: Development in Childhood and Adolescence; Identity in Childhood and Adolescence; Neural Plasticity; Social Competence: Childhood and Adolescence; Socialization in Adolescence; Youth Culture, Anthropology of; Youth to Adulthood across Cultures, Transition from

Bibliography

- Castro A J 1990 Plasticity in the motor system. In: Kolb B, Tees R C (eds.) *The Cerebral Cortex of the Rat*. MIT Press, Cambridge, MA, pp. 563–88
- Bachevalier J, Malkova L 2000 Neuropsychological indices of early medial temporal lobe dysfunction in primates. In: Levin H S, Grafman J (eds.) *Cerebral Reorganization of Function after Brain Damage*. Oxford University Press, New York, pp. 27–48

- Finger S, Almli C R 1988 Margaret Kennard and her 'principle' in historical perspective. In: Finger S, Levere T E, Almli C R, Stein D G (eds.) *Brain Injury and Recovery*. Plenum, New York, pp. 117–32
- Hebb D O 1949 *The Organization of Behavior: A Neuropsychological Theory*. Wiley, New York
- Hicks S P, D'Amato C J, Glover R A 1984 Recovery or malformation after fetal radiation and other injuries. In: Finger S, Almli C R (eds.) *Early Brain Damage*. Academic Press, Orlando, FL, pp. 127–47
- Kennard M A 1940 Relation of age to motor impairment in man and in subhuman primates. *Archives of Neurol. Psychiatry* **44**: 377–97
- Kolb B 1995 *Brain Plasticity and Behavior*. Erlbaum, Mahwah, NJ
- Kolb B, Cioe J, Muirhead D 1998a Cerebral morphology and functional sparing after prenatal frontal cortex lesions in rats. *Behavioral Brain Research* **91**: 143–55
- Kolb B, Cioe J, Whishaw I Q 2000 Is there an optimal age for recovery from motor cortex lesions? I. Behavioral and anatomical sequelae of bilateral motor cortex lesions in rats on postnatal days 1, 10, and in adulthood. *Brain Research* in press
- Kolb B, Gibb R, Gorny G, Whishaw I Q 1998b Possible brain regrowth after cortical lesions in rats. *Behavioral Brain Research* **91**: 127–41
- Kolb B, Whishaw I Q 1989 Plasticity in the neocortex: Mechanisms underlying recovery from early brain damage. *Prog Neurobiol* **32**: 235–76
- Krashen S D 1973 Lateralization, language learning, and the critical period: Some new evidence. *Language Learning* **23**: 63–74
- Levin H S, Eisenberg H M, Benton A L 1989 *Mild Head Injury*. Oxford University Press, New York
- O'Leary D D 1992 Development of connectional diversity and specificity in the mammalian brain by the pruning of collateral projections. *Current Opinion in Neuroscience* **2**: 70–7
- Rasmussen T, Milner B 1977 The role of early left-brain injury in determining lateralization of cerebral speech functions. *Annals of the New York Academy of Sciences* **299**: 355–69
- Steward O 1991 Synapse replacement on cortical neurons following denervation. In: Peters A, Jones E G (eds.) *Cerebral Cortex*. Plenum, New York, Vol. 9, pp. 81–131
- Stiles J 2000 Spatial cognitive development following prenatal or perinatal focal brain injury. In: Levin H S, Grafman J (eds.) *Cerebral Reorganization of Function after Brain Damage*. Oxford University Press, New York, pp. 201–17
- Teuber H-L 1975 Recovery of function after brain injury in man. In: *Outcome of Severe Damage to the Nervous System*. CIBA Foundation Symposium 34. Elsevier, North-Holland, Amsterdam, The Netherlands
- Weiss S, Reynolds B A, Vescovi A L, Morshead C, Craig C G, van der Kooy D 1996 Is there a neural stem cell in the mammalian forebrain? *Trends in Neuroscience* **19**: 387–93

B. Kolb

Recovery of Function, Neural Basis of

Everyone has had the experience of having had an injury in one's life, from burning a finger to scraping an elbow to stubbing a toe, perhaps cutting oneself or

breaking a bone. The universal feature of this experience is that, with care and treatment and time, the injury heals and usually nothing more than a scar and a faded memory remain.

Such recovery is not common with an injury to the nervous system. One of the most important goals in the field of neuroscience is the issue of recovery of function following nerve, spinal cord, and brain injury. It is an unfortunate fact that injuries of the nervous system can produce profound deficits in the quality of life. Sometimes these deficits are obvious, as when an injured nerve affects motor or sensory functioning. Sometimes there are profound deficits without obvious signs of injury, as when someone suffers from memory loss from injury or developmental/aging effects, for example, autism or Alzheimer's disease. In either case we would like to be able to restore functioning.

Sometimes, however, there are instances of considerable recovery of function following brain injury. Stroke patients who initially have profound language disorders, for example, may in the long term have considerable restoration of function. Until recently, in the animal experimental literature, there has been difficulty in localizing neural correlates of functioning because the effects of damage were only temporarily observed with significant amounts of recovery of function over time. Ideally, if we had enough understanding of the nervous system then we might be able to promote recovery of function. The study of brain-behavior relationships, of localization of function, and of recovery of function, are intimately associated themes.

1. Theories

The four major theories that account for instances of recovery of function following brain injury are examined in this article. No one theory fully accounts for the spectrum of observed recovery, nor are the theories mutually exclusive—more than one factor may be involved in recovery. To a certain extent, these theories provide a heuristic outline of the factors that affect recovery. It is the hope that a better understanding of these factors will lead to more effective treatment of the injured.

1.1 Vicariation

Vicariation means second-handed. Recovery by vicariation occurs when an injury damages the part of the nervous system that is actually responsible for a function, and a second neural structure that had never been involved in that function now takes over for the injured structure and performs that previously lost function. A key feature of vicariation is that the structure that takes over the function was never involved in the original behavior. This distinction is lost in some treatments.

The best example of vicariation is the observation that injuries in young individuals are said to be less deleterious than similar damage when it occurs in adults. For example, damage to the left hemisphere in a stroke can cause profound loss of language functions in the adult. Children with left hemisphere damage, however, often do not show severe language deficits. This observation fits our concept of the plasticity of the developing organism.

This concept—that youngsters are less affected by brain injury—has been applied to laboratory experiments where it has generally been found that damage is less intrusive in young individuals. For example, damage to the frontal lobes does not produce delayed response deficits in monkeys, and damage to the septal region does not produce social cohesiveness in rats. However, it is not always true that an early lesion is beneficial. In these two examples, when the individual becomes an adult then delayed response deficits and increased social cohesiveness appear. Thus, the effects of the damage may not be obvious until the individual matures.

Of all the theories of recovery of function, the theory of vicariation is the least likely one to be true. This is because much of the recovery that has been attributed to vicariation is actually due to a region that already has some functioning. As a result, the following theories are more likely explanations, and thus more likely to cause recovery.

1.2 Diaschesis

Diaschesis refers to the effects of shock. Recovery by diaschesis occurs when a nervous system injury does not actually damage the tissue that is responsible for the lost function, but rather that the injury causes a shock effect which indirectly and deleteriously affects normal functioning, and function returns when the shock subsides. Shock can occur when there is temporary loss of the blood supply (loss of nutrients including glucose and oxygen, or loss of functions to remove waste metabolic products) or temporary loss of necessary modulating neural activity. The key feature is that when functioning returns it is mediated by the very same tissue that was responsible for the function in the first place, so that the recovered behavior is exactly the same as the original one. Thus, the idea of sparing is essential in the context of diaschesis.

Time appears to be the major manipulation to effect diaschesis. However, time itself is not a very profound or necessarily a helpful explanation. The real question is what happens that takes time to dissipate the effects of shock, or more to the point, what can be done to diminish the effects of shock. Steroid drugs have been given to prevent the effects of swelling, for example, and have some effect in mitigating loss of function from some spinal cord injuries. To be effective the treatment must occur soon after the injury. Since the

brain and spinal cord are enclosed within the cranium and spinal column, respectively, any swelling of the nervous system in these confined spaces can cause substantial damage that is secondary to the original injury. Thus, besides indirect effects of shock any injury can cause additional damage.

One example of diaschesis from the laboratory is known as the Sprague effect. Damage to one superior colliculus in cats causes a visual neglect in the opposite visual field; that is, objects are ignored if they are presented in that visual field. Subsequent damage to the visual cortex on the same side as the visual neglect immediately restores reactions in the affected visual field. This example represents diaschesis by loss of normal neural modulatory influences. Here subcortical visual regions in the tectal area receive three visual inputs, one from the eye through the optic nerve to the superior colliculus, one from the opposite superior colliculus, and the other through the visual cortex. By damaging the superior colliculus there is an imbalance in the surviving tectal region. By subsequently damaging the visual cortex on the same side as the visual neglect, the balance in the surviving tectal region is restored. This illustrates one of the major ways of thinking about diaschesis, which is to think of it as a balance of modulation of neural sources.

Another example from the laboratory is the loss of a simple visual discrimination for brightness. Rats trained on a black-white discrimination will appear to have lost that memory when their visual cortices are removed. With subsequent training the rats reacquire the ability to discriminate black from white. Since it takes about as many trials to relearn as it did to originally learn, this appears to be a case of vicariation where another part of the brain (presumably the remaining cortex) takes over functioning. If that is the case, however, that subsequent training is creating a new memory, then it should not matter whether training after the brain injury is made with the black or white stimulus as the correct choice. But reversing the correct choice profoundly interferes with subsequent training. This shows that although it appears that the brain injury has removed the original memory that in fact the memory remains to interfere with reversal training.

A number of manipulations have been found to affect recovery through diaschesis, each associated with diminishing the effects of shock and enhancing the expression of the spared tissue: extensive preoperative training, interoperative retraining (serial lesions), extensive postoperative retraining, cross-modal training on similar problems (i.e., training using other sensory modalities), drugs that diminish shock (e.g., steroids), drugs that enhance access to the spared tissue. As with strategies to improve performance on exams, a practical strategy to lessen the effects of future brain injury or the inevitable effects of aging would be to overlearn with multiple cues and learning methods.

Parkinson's disease provides an example from the clinical literature where the diminishing effects of diaschesis prevents the initial appearance of symptoms. Parkinson's disease is a movement and thought disorder caused by progressive degeneration of cells in the substantia nigra that project dopamine to the basal ganglia. Its symptoms only appear when so many nigra cells die that only about 20 percent of the normal amount of dopamine remains in the basal ganglia. Until that point, it appears that the basal ganglia can compensate for the loss of dopamine (see below). It is only when the dopamine projection to the basal ganglia falls below a critical point that the symptoms of Parkinson's show themselves. Presumably, slow progressive disorder does not produce the effects of shock that would be seen with one larger lesion.

Of the factors that account for recovery of function from brain injury, diaschesis is one of the most important.

1.3 Neural Reorganization

Neural reorganization generally refers to the body's own mechanisms for recognizing loss and for healing itself. However, man-made interventions can also be classified as reorganization. By far, neural reorganization receives the most attention of the four theories, although its practical usefulness has yet to be realized. Several means of effecting neural reorganization have been described. Importantly, these mechanisms are adaptations of means that the nervous system has for normal development and growth rather than specialized adaptations that might have evolved solely for the purpose of a complex brain being able to repair itself. That is, one questionable idea about brain organization in general is the idea that more complex brains, being more susceptible to injury of its delicate mechanisms, somehow developed a new method of dealing with brain injury. That seems not to be the case. Here the known mechanisms of reorganization are reviewed.

1.3.1 Reactive synaptogenesis. Synaptogenesis refers to new synapse formation. This mechanism is essential developmentally for correctly hooking up the nervous system (along with overproduction and selective cell death, as in programmed cell death or apoptosis). Reactive refers to an action in response to injury. New connections can be made by a number of activities, including new axonal growth (sprouting), new pre- and post-synaptic growth and terminal and dendrite formation, new receptor formation and/or distribution, for example. Furthermore, injury to the visual cortex causes degeneration of its projections to subcortical visual areas which leaves vacated synaptic space that is filled by sprouting of the optic nerve. In the example above (see Sect.

1.2) where rats were trained on a black–white discrimination, sprouting in subcortical visual areas could account for the recovery of function. It is generally thought that sprouting should be beneficial and attempts have been made to promote sprouting. However, in the case of brightness discrimination this idea is questionable since subsequent damage to the areas of sprouting has no permanent effect in preventing learning differences in brightness. In general, since rats with visual cortical damage do not recover perceptions of form and pattern, neural reorganization is not a significant factor for those functions.

Two examples illustrate where sprouting is actually less than helpful. First, spinal cord injuries cause sprouting that is accompanied by tremors. The tremors are alleviated when the sprouted connections are destroyed. Second, injury of one superior colliculus in hamsters causes sprouting to the opposite superior colliculus. The result is that the animals visually orient to the side away from the visual stimulus. Cutting the sprouted optic nerve to the remaining superior colliculus resolves the problem. Thus, in these two examples, blindly enhancing sprouting by generally giving nerve growth factors, for example, may not be such a good idea. The problem with sprouting is to have control over where and what type of new growth takes place.

1.3.2 Neurogenesis. We learn that we are born with the total number of neurons that we are ever going to have, and that we continually lose neurons as we grow older. The one celebrated exception to this rule is that certain birds which learn a new song each mating season show cell division and new neuron development as adults. Adult rats and recently adult humans have been shown to create new neurons in the hippocampus, a limbic system structure associated with certain kinds of memories. It remains to be determined whether there are sufficient numbers of new neurons and whether there are functional connections of these newly available neurons. Similarly, it is unknown whether these neurons have a role in repairing brain injury. However, on the face of it, this is an exciting development.

1.3.3 Transplants. Related to the idea of natural neural replacement (neurogenesis) is the idea that we can purposely implant neurons to replace the damaged ones. Transplants have been tried most often in the case of patients with Parkinson's disease, where the transplant donor is either the patient (dopamine producing cells from the patient's own adrenal glands) or an aborted fetus (dopamine producing cells from the substantia nigra; this source is controversial). Initial results of such operations were mixed at best,

with more recent limited successes. The ameliorating effects may not be due to the effects of dopamine or to neural connections formed by the implant, however. Nerve growth factors that are secreted by the implant are more likely candidates. Transplants themselves have been recently overshadowed by a return to surgical intervention, a procedure known as a pallidotomy which is a lesion of the globus pallidus.

1.3.4 Supersensitivity. Synapses that are deprived of their normal expected amount of input activity can become supersensitive whereby the synaptic transmission becomes unusually sensitive to small amounts of neurotransmitter. In Parkinson's disease the deleterious symptoms are delayed presumably because of the compensating supersensitivity of cells in the basal ganglia that are deprived of their usual amount of dopamine from degenerating neurons in the substantia nigra. It is only when there is too much loss of dopamine that the basal ganglia neurons cannot compensate and symptoms of Parkinson's appear. The actual mechanism of supersensitivity can involve increased sensitivity of the postsynaptic receptors, increased reactivity of postsynaptic receptors, increased number of postsynaptic receptors, changes in dendritic spines that reduce impedance (i.e., enhance transmission), increased transmitter release, and changes in intracellular events that enhance the signal.

1.3.5 Prostheses. Man–machine implants may be used to overcome deficits. For sensory loss, these involve a transduction device to convert an external stimulus into a neural-like signal with a neural stimulator. Currently, auditory implants into the cochlear nerve or nuclei are being tried to restore hearing. Visual implants using a primitive camera and stimulation of visual cortex have a history of limited success in effecting primitive sensations. Not all prostheses are surgically placed into the brain. One alternative for visual problems is to place the pattern of visual stimuli onto a skin surface where the visual stimulus is now interpreted as a tactile stimulus. Other approaches have not involved sensory transduction but rather replace the damaged element responsible for higher level processing. In this regard a recent attempt has been made to implement functioning in a silicon chip that is used to replace a damaged hippocampus. For motor loss, electrodes have been implanted to stimulate muscles, coordinated by a computer, to effect walking. This requires sensors for proprioception and balance.

1.3.6 Unmasking. It is now well appreciated that the nervous system has multiple representations for each

of the senses and motor functions. Moreover, it is becoming clear that representations may overlap each other, and that under normal circumstances some of the underlying representations may be suppressed. When a part of the main representation is damaged the underlying map may now express itself, either preventing any deficit or causing an apparent recovery. Unmasking is generally found in instances where the recovery occurs too quickly for sprouting or other types of reorganization to have been made.

1.3.7 Regeneration. Certain amphibians and lizards will grow a new limb to replace one that has been lost to a predator. Amphibians and lizards also have the ability to re-establish connections when a nerve is severed. This regeneration is also seen to varying degrees in the peripheral nervous system of mammals, aided by the supporting Schwann cells which guide the reconnection. Although the capacity for regeneration persists in the central nervous system of mammals, regeneration is not a significant factor in recovery of function because central glial cells create scar tissue which becomes a significant barrier to reconnection.

1.4 Substitution

Substitution refers to using an alternative means to achieve the same goal. If damage has actually removed that part of the brain that is responsible for a function (as opposed to Diaschisis, above, for example) then substitution is the most likely means of recovery. Within the animal literature one needs to be cognizant of factors like substitution because otherwise false conclusions can be made about the nature of the recovery. Yet substitution is often ignored because it does not involve surgical or chemical intervention. There are three significant means of substitution. Recovery by substitution occurs when another sense, motor act or strategy can be used.

1.4.1 Sensory. A sensory stimulus or property of a stimulus can be used as a substitute for the original sensation. Damage to visual cortex in rats makes them blind to form and pattern perception yet they are still capable of detecting primitive visual experiences such as brightness and contour. As a result, such rats appear to be able to solve discriminations for orientation (e. g., horizontal vs. vertical black and white stripes) because they can make use of local brightness and contour cues to solve the problem. In extreme conditions where the rats are prevented from using these local cues, they can no longer 'solve' the orientation problem, and thus they are blind. The goal for a therapist would be to design conditions

where a problem can be solved in multiple ways. For example, persons who are red-green color blind can learn to drive by noting the position of the active light on a traffic signal.

1.4.2 Motor. One motor act can substitute for the original motor act. Cats with dorsal rhizotomies (denervation of the sensory roots of the spinal cord) for the hind limb initially have difficulty walking a thin railing. After a period of time, however, the cats negotiate the task very well and appear to be completely normal. However, upon close examination with high-speed photography, it has been shown that these cats adopt a motor strategy of shifting their weight to properly place the hind foot. Clearly, functional recovery does not imply neural recovery.

Another example from the older literature would be the instance where a rat has been trained to run a maze. The same rat that has been trained to run can solve the very same maze when it is flooded and it has to swim. The same idea applies to conditions where brain damage prevents the original behavior.

1.4.3 Strategy. After ruling out that the senses and motor functions are not affected, it might be found that there may be a problem with the implementation of a sequence of perceptions and actions. This higher order activity can be referred to as a strategy. For example, a woman with global (diffuse) brain damage got easily distracted, so much so that a telephone ring would be enough of an interruption that she would forget she was cooking. Or she would spend her whole day watching TV. Her strategy to overcome this deficit was to write a list of things to do, checking them off as she did them, and to have a timer set to go off every 15 mins. At every alarm she would check her list to see what it was she was supposed to be doing.

From the animal literature comes another example. Rats with hippocampal brain damage have difficulty in alternately pressing one of two bars for a food reward. This task requires that the animals remember which bar was pressed last so that they can press the other bar the next time. Some clever rats figure out that if they hold onto the bar that they just pressed then they immediately know to press the other bar the next time.

2. Factors

The discussion above identifies the major theories that can account for instances of recovery of function. Within this discussion we have identified a number of factors that can be used clinically and within the laboratory to promote recovery of function. It is

reasonable at this time to simply list those factors, because although they were discussed in terms of theories above in truth many of these factors apply to more than one theory. Here is a listing of some factors that might enhance recovery of function: age at time of the injury, reduction in secondary effects of the injury (shock), the proficiency of preinjury training, training on alternative stimuli or strategies or acts, training between injuries (serial lesion effect), time after injury, amount of postinjury training on the same or similar tasks, substitutions. It would be misleading to suggest that any or all of these factors will lead to recovery in all instances. It is a sad fact that injuries to the nervous system produce profound and lasting deficits. The hope for further research is to find new applications of these factors and to develop new methods of treatment.

3. Implications

Clinically the importance of finding methods that enhance recovery of function is obvious. What is not so obvious is that the clinical and experimental literature about brain injury also give insight into the fundamental organization and operation of the brain and our mental life. The mind-body problem—the relationship of our physical being with our mental experience—continues to be a fundamental interest in psychology. Both the clinical and experimental literature make fundamental contributions to each other. Newer methods for imaging the brain during cognitive activity (fMRI, PET, neuromagnetic methods, along with the old standby EEG) have confirmed much of our understanding of the organization of the brain and what happens during injury, on the one hand, and given insights and further directions, on the other. The optimism of our times holds out great promise for sophisticated advances that may someday allow us to repair the damaged mind.

Importantly, we learn from the field of recovery of function the following lessons. First, localization of function is the norm in the nervous system. Second, there is a hierarchical organization to the nervous system, with basic functions organized at low levels. Third, even low levels of the nervous system are capable of seemingly sophisticated functions like learning and memory. Fourth, any one function is probably represented in only one place in the nervous system. However, fifth, similar functions can often accomplish the same goals.

See also: Behavioral Neuroscience; Brain Damage; Neuropsychological Rehabilitation; Brain Implants and Transplants; Memory Problems and Rehabilitation; Neural Plasticity; Neural Plasticity in Auditory Cortex; Neural Plasticity in Visual Cortex; Neural Plasticity of Spinal Reflexes; Recovery of Function: Dependency on Age

Bibliography

- Bach-y-Rita P (ed.) 1980 *Recovery of Function: Theoretical Considerations for Brain Injury Rehabilitation*. University Park Press, Baltimore, MD
- Finger S (ed.) 1978 *Recovery from Brain Damage*. Plenum Press, New York
- Finger S, Levere T E, Almlie C R, Stein D (eds.) 1988 *Brain Injury and Recovery: Theoretical and Controversial Issues*. Plenum Press, New York
- Johnson J A (ed.) 1990 *Occupational Therapy Approaches to Traumatic Brain Injury*. Haworth Press, Binghamton, NY
- Levin H S, Benton A L, Grossman R G 1982 *Neurobehavioral Consequences of Closed Head Injury*. Oxford University Press, New York
- Montgomery J (ed.) 1994 *Physical Therapy for Traumatic Brain Injury (Clinics in Physical Therapy)*. Churchill Livingstone, New York
- Scheff S W (ed.) 1984 *Aging and Recovery of Function in the Central Nervous System*. Plenum Press, New York
- Stein D G, Brailowsky S, Will B 1995 *Brain Repair*. Oxford University Press, New York
- Stein D G, Rosen J J, Butters N (eds.) 1974 *Plasticity and Recovery of Function in the Central Nervous System*. Academic Press, New York
- Uzzell B P, Stonnington H H (eds.) 1996 *Recovery after Traumatic Brain Injury*. Erlbaum, Hillsdale, NJ

D. Lavond

Reduction, Varieties of

Fundamental issues in the ontology of the sciences can depend on whether the elements of one domain turn out to be 'nothing over and above' a congeries of elements of some other, perhaps less mysterious or more basic, domain. It has been argued that natural numbers, for example, are nothing over and above certain sets, and their properties just the familiar properties of those sets; that temperature is nothing over and above the mean kinetic energy of molecules; that genes are nothing over and above strings of DNA molecules; that the mental is nothing over and above the physical; and that the social is nothing over and above an aggregate of individuals, their antics, and attitudes. Each such claim, whether correct or not, postulates a dependency of one domain on another, a dependency strong enough to justify regarding the dependent domain as ontologically derivative, the other as ontologically more basic. Given the more basic domain, the dependent domain can plausibly be regarded as an ontological 'free lunch.' The elements of the dependent domain are not denied existence or reality, but neither are they granted an independent reality.

Such dependency claims have been thought to have far-reaching ramifications not just for ontology but also for methodology. For example, the social sciences investigate social phenomena, involving distinctively

social entities, properties, facts, and laws. The individualism–holism debate concerns, in the first instance, the status of these items—whether they are anything over and above the totality of individual entities, properties, laws, and facts—but also the methodological autonomy of the associated sciences. (For an even finer-grained analysis which disentangles other conflated but independent components of this debate, see Pettit 1993.) According to Watkins (writing in the 1950s), individualism holds that social institutions ‘do not possess a quasi-divine autonomy’ (in O’Neill 1973, p. 149), and that ‘every complex social situation or event is the result of a particular configuration of individuals, their dispositions, situations, beliefs, and physical resources and environments’ (in O’Neill 1973, p. 168). He argued that, as a consequence, explanations of social phenomena should be couched in terms of the ‘situations, dispositions, and beliefs of individuals’ (in O’Neill 1973, p. 179). His adversary, Mandelbaum, states *holism* thus: ‘Societal facts are as ultimate as are psychological facts ... those concepts which are used to refer to the form or organization of a society cannot be reduced without remainder to concepts which only refer to the thoughts and actions of specific individuals’ (in O’Neill 1973, p. 223).

As Mandelbaum’s claim here illustrates, definability was once the Holy Grail of the reductionist enterprise. Bertrand Russell famously maintained that postulating entities, in contrast to defining (or ‘constructing’) them, ‘has many advantages; they are the same as the advantages of theft over honest toil’ (Russell 1919, p. 71). But just as when the quest for the Holy Grail looked misguided, some more realistic venture might engage a knight’s energies, so when definability failed to capture the interesting dependency claims, philosophers looked elsewhere. And they have discovered a plethora of dependency relations, some of which hold out the promise of prizing apart the methodological and ontological issues, enabling, for example, both the metaphysical strictures of the individualist and the methodological dispositions of the holist to be accommodated.

1. Explicit Definability

Explicit definition is the clearest way of demonstrating that *As* are nothing over and above *Bs*, guaranteeing that all truths about *As* are expressible by talking overtly only about *Bs*. A full specification of the *Bs* yields *ipso facto* a full specification of the *As*. An explicit definition of *A* in terms of $B_1, \dots, B_n$ has the following general form:

$$A =_{\text{defn.}} F(B_1, \dots, B_n)$$

The right-hand side, $F(B_1, \dots, B_n)$, may involve, in addition to $B_1, \dots, B_n$, only certain *permissible* operations. In the strictest case these are the purely logical connectives (*and*, *not*, *all*, and *identity*). More liberally

one might include the concept of *set* and *set membership*, or perhaps the whole apparatus of mathematics. At least as traditionally understood, the definition should be knowable *a priori*, through pure analysis of the concepts involved, and it should hold by logical *necessity*.

A striking illustration of the power of definition is *logicism*, the program to reduce mathematics to logic (Russell 1919). In the nineteenth century, the basic concepts of analysis had been successively defined in terms of the concepts of natural numbers and sets. It began with the elimination of infinitesimals via the definition of a limit. Irrationals were defined as sequences of rationals, and rationals as sets of pairs of natural numbers, leaving the natural numbers alone as primitive, undefined entities. The Russell–Frege definition of natural numbers is remarkable in that it reaches right outside the category of number. Numbers are defined using only the apparatus of logic—including the logic of sets. For example:

zero = the set of all empty sets

one = the set of all singleton sets

For this to be adequate, truths about numbers must follow from the definitions, together with definitions of the mathematical operations, and the logic of sets. This hope was almost vindicated—but for the discovery of troublesome paradoxes in the foundations of set theory. Despite the paradoxes, the program remained a paradigm of ontological reduction, inspiring much subsequent philosophical activity (Carnap 1928) and providing the framework for most debates about ontological commitments.

2. Reducibility

Definability proved too strong a requirement for typical claims of ontological dependence for reasons involving the constraints that definitions be known *a priori* and that they hold true by logical necessity.

As is well known, our understanding of the phenomena of heat, codified in thermodynamics, was transformed by the reduction of thermodynamics to statistical mechanics (Nagel 1961, Chap. 11). The reduction involves deducing the laws of thermodynamics from laws of statistical mechanics together with bridge assumptions connecting the concepts of the two domains (*temperature* and *heat* with *molecule*, *mass*, and *motion*). For example, the Boyle–Charles law for ideal gases (pressure times volume is proportional to temperature) can be shown to follow from the assumptions of statistical mechanics and the following assumption:

(1) *Temperature is mean kinetic energy of component molecules.*

Assumption (1) may look like a definition, and is sometimes touted as such, but it is not the result of a

priori analysis of its concepts, and it is not logically necessary. (Suppose everything were continuous stuff all the way down and temperature were, say, proportional to density of calorific fluid.) If true, (1) is not logically necessary, and if known, not known *a priori*.

Some maintain that (1), even if not true by definition or by logical necessity, is nevertheless a matter of identity, vindicating the reductive claim that temperature is nothing but motion of molecules. However, unlike definitional identities, it is a *contingent* identity, analogous to the contingent identity of *the morning star* and *the evening star* (Smart 1959). Analysis of concepts would not yield the fact that the brightest celestial body in the morning sky is identical with the brightest celestial body in the evening sky, and we can easily imagine a world in which they are not. But the morning and evening stars are nevertheless one and the same—Venus. So perhaps (1) states a contingent identity known *a posteriori*.

The idea that *identity* (the relation which each entity bears to itself necessarily and to no other thing) could be a contingent matter is still baffling. One way to reduce the baffle is to distinguish between intension (concept) and extension (item singled out). The intension is a certain way of singling out the extension. The extension of both *the morning star* and *the evening star* is a chunk of matter, but the intensions are different ways of determining that chunk. The intensions are *coextensive* in fact but not of necessity.

Property terms also come with an extensional and an intensional component. The extension of *black* is the class of things that are black. If all black things happened to be chunks of coal, and all chunks of coal happened to be black, then *black* and *coal* would be coextensive. They would single out the same class of objects. But is coextensiveness enough for reduction? The two terms are associated with different intensions, they pick out distinct features of the objects, and even if coextensive they would not be coextensive of necessity. Further, we would not be inclined to say that *black* reduces to *coal* (or vice versa) just because they were coextensive. So there must be more to reduction of properties than coextensiveness.

Temperature and *mean kinetic energy* are not properties but *magnitudes*. Unlike a property, the extension of temperature is not a class of objects. Rather, the *temperature* of an item can be any one of a number of different possible numerical values—relative to a unit of measure and an origin. The extension of *temperature*, like *mean kinetic energy*, thus associates an item with a number. We could construe (1) as the simple claim that the magnitudes *temperature* and *mean kinetic energy* are coextensive. However, as the argument concerning *black* and *coal* illustrates, a reduction requires a tighter connection. The extensional equivalence reported by (1) has to hold by *some kind of necessity* (not logical necessity).

A possible world is a complete way the totality of things might be. A state is logically necessary if it holds

in all possible worlds. Laws of nature rule out some but not all possibilities (some laws might have been different from what they are in fact) so they are not logically necessary. But what the laws of nature entail involves a kind of necessity—*nomological* necessity. A state is nomologically necessary if it holds in all worlds in which the actual laws of nature hold. For example, if the theory of relativity is true then in some sense it is impossible for an object to accelerate through the light barrier. This is not just an accidental fact, like the decay of some radium atom, but it is not *logically* necessary either. It is nomologically necessary.

For (1) to constitute a reduction it must be understood as the claim that *temperature* is coextensive with *mean kinetic energy* by nomological necessity. We can now exhibit the difference between definability and reducibility:

A is *definable* in terms of the *Bs*, just in case *A* is coextensive with $F(B_1, \dots, B_n)$ by *logical* necessity.

(Or: *A* is logically equivalent to $F(B_1, \dots, B_n)$.)

A is *reducible* to the *Bs*, just in case *A* is coextensive with $F(B_1, \dots, B_n)$ by *nomological* necessity (in the domain of the *Bs*).

(Or: *A* is nomologically equivalent to $F(B_1, \dots, B_n)$.)

It follows that *A* is reducible to the *Bs* if *A* is nomologically equivalent to something definable in terms of the *Bs*. Definability entails reducibility (since logical necessity entails nomological necessity) but not vice versa.

Just as (1) transformed our understanding of temperature, our understanding of a mental state like *pain* may be transformed by the following claim:

(2) *Pain is the stimulation of C-fibers in the brain.*

For this to be a reductive claim on the model of (1), *pain* must be nomologically equivalent to *C-fiber stimulation*. What play the roles here of reducing and reduced theory? The reducing theory would be neurophysiology, and the reduced theory could be our ordinary everyday conception of pain—collections of platitudes like:

Pain is typically brought about by bodily damage, and results in wincing, groaning, and avoidance behavior.

Such platitudes were central to the *behaviorist* hope that pain might be characterized, or *defined*, in purely behavioral terms (Ryle 1949). When this hope of definability was abandoned the platitudes were central to the *functionalist* hope that pain could be suitably *reduced* to physical states of the brain (Lewis 1993, essay 7, Armstrong 1968). Assumption (2), together with a suitably rich scientific theory, might well yield such platitudes. There is a puzzle here to do with the status of the bridging principles like (1) and (2). If we want to reduce the mental to the physical, for example, by exhibiting purely physical (nomological) equivalents of mental features, we would not want to have to acknowledge irreducible laws governing the psychological realm, or irreducible connections between the two realisms. So for reductionism to follow, the laws underpinning nomological equivalents stated in the

bridging principles would have to be laws of the lower-level domain—in this case purely physical laws. The puzzle is how purely physical laws could entail principles straddling the two domains—an answer will be presented in Sect. 5.

If analogous reductions were to hold for all mental features then *physicalism*—the thesis that the mental is nothing over and above the physical—would be vindicated. Undergirding the reduction would be a multitude of nomologically necessary extensional equivalences between mental terms on the one hand and logical products of purely physical terms on the other. A successful reduction would not, of course, entail the elimination of mental entities or of the laws within the mental realm. (The reducibility of thermodynamic laws to statistical mechanics is rather a *vindication* of the former.) And far from being the denial of systematic and lawlike mental–physical connections it is the very affirmation of them—in the form of bridge principles.

3. The Problem of Networking

For a successful reduction we must have appropriate logical or nomological equivalents. There is, however, an important objection to the existence of such equivalents in two central cases—the relation between the mental and the physical, and the relation between the social and the individual—namely, the heavily networked character of the concepts characteristic of different domains. Mental concepts are heavily networked not just with behavioral and physical concepts but also with other mental concepts. Social concepts are heavily networked not only with each other but also with psychological concepts via those very same social concepts.

Consider *pain* again. Pain is tied not only to physical inputs and behavioral outputs, but also to other psychological states. Those in pain tend to believe they are in pain, and desire to get out of it. These are platitudes about pain which any reductive account of pain should accommodate. So it looks as though there may be no simple way of giving logical or nomological equivalents of pain in purely physical terms.

The case of social concepts is even more acute. Consider the statement, ‘Jones cashes a check at the bank.’ The predicate involves a social property—making essential use of the social concept of a *bank*—and the claim itself entails the existence of a bank. According to Mandelbaum, we cannot eliminate the concept *bank* in explicating the content of this claim. An analysis in terms of psychological states (of Jones, the teller, the bank manager, etc.) would have to include beliefs and desires the contents of which will essentially involve the concept of bank.

... concepts involving the notion of status and role cannot themselves be reduced to a conjunction of statements in

which these or other societal concepts do not appear (O’Neill 1973, p. 225).

So here is an obstacle to finding a purely individualistic equivalent of the social property. While Mandelbaum is thinking in terms of definability, a similar argument might be mounted against reducibility.

4. Implicit Definability

The networking problem was solved by an elegant theory of implicit definition, initially discovered by Carnap in the course of his search for a reduction of the theoretical to the observable (Carnap 1966) and fully developed by Lewis into a general reductionist framework (Lewis 1993, essay 6).

Quine famously argued, against positivism, that theoretical terms gain their meaning from being heavily networked with each other and with observable terms only in the context of a rich and testable theory (Quine 1961). Carnap proposed an essentially reductionist account of such terms which nevertheless accommodated this holistic thesis. The essential ingredient was provided by a then little known result of Ramsey (1931). Ramsey showed how a theory about, for example, electrons and protons, $T(\text{electron}, \text{proton}, \dots)$ could be stripped of explicit mention of unobservable entities and replaced by a lean purely observational surrogate, its so-called *ramsification*. Theoretical terms—*electron*, *proton*—are first replaced with variables, E , P , (or ‘gaps’) to yield a ‘gappy’ sentence $T(E, P, \dots)$ which may or may not be true depending on what the gaps are filled with. Now consider the definite claim that the gaps can be indeed truly filled by something or other:

$$\begin{aligned} &\text{Ramsification of } T(\text{electron}, \text{proton}) \\ &= \text{there are items } E, P \text{ satisfying } T(E, P) \end{aligned}$$

This stripped-down version of the theory entails nothing nontrivial about electrons or protons (because it does not mention them at all), but does entail all the purely observational consequences of the full-blown theory. Since the point of the theory (according to the positivists) is to generate observational consequences, and the ramsification does this without reference to theoretical entities, Ramsey concluded that the theoretical ontology is redundant (Ramsey 1931).

Carnap drew a different lesson. He agreed with Ramsey that the factual component of the theory is delivered by its ramsification, but he noted that for the factual claim to be true there must be some items satisfying the gappy statement $T(E, P, \dots)$ —for that is exactly what the ramsification says. In that case, we may as well regard those items as the intended

extensions of the concepts *electron* and *proton*. The whole theory thus serves as a partial constraint on the meanings of its own theoretical terms. It is partial, both because it places no constraints in the case in which the theory is not observationally adequate, and because even when it is there may be a range of legitimate satisfiers of the gappy claim. However, in certain cases we may be lucky and the theory might be strong enough to single out a unique set of satisfiers. In that case we can use the theory to give an explicit definition (Lewis 1993, essay 6).

Something analogous can be sketched for mental properties. We begin with our ordinary everyday collection of platitudes about the mental ('folk psychology') in which, as noted, the mental concepts such as pain and desire are heavily networked with each other and with behavioral and physical concepts. Now ramsify this theory, eliminating the mental concepts and replacing them with variables (*P* for pain and *D* for desire). The ramsified theory will look something like this: there is a property *P* typically caused by bodily damage and which causes wincing, and if a something has *P* it typically bears *D* to getting rid of *P*.

This ramsification is stated purely in physicalistic and behavioral terms and might be true of a system which consisted only of physical constituents with purely physical properties. Neurophysiologists might discover that C-fiber stimulation is a feature typically caused by bodily damage and causes wincing; that C-fiber stimulation initiates D-fiber stimulation, which in turn causes C-fiber stimulation-ridding behavior, and so on. Enough of these would support the claim that pain is nomologically equivalent to C-fiber stimulation; the desire to get rid of pain nomologically equivalent to D-fiber stimulation, and so on.

If this is right, then the holistic and networked character of the mental is not itself an insurmountable obstacle to reduction in the case of the mental. Hopefully something similar might be achieved for the social domain. The reduction in the latter case would be to purely individual properties and relations—psychological, behavioral, and physical—which may in turn be reduced to the purely physical.

5. Multiple Realizability

Not long after functionalism had provided this systematic framework for reduction, philosophers began to worry that there may be no appropriate, lawlike extensional equivalences in the case of the mind (Putnam 1967, Davidson 1970). Suppose in a typical human being pain is 'realized' by C-fiber stimulation, whereas in some untypical humans it is D-fiber stimulation, and in Fido the dog it is E-fiber stimulation. Suppose, further, that there is no end to the number of different states which realize pain in different beings. Suppose in some distant galaxy silicon-based aliens have evolved, or could evolve,

with a very different physical structure. Would not they be capable of experiencing pain too? There may be no limit to the ways in which pain can be thus realized in physical beings. So there may be no single, purely physical, nomological equivalent of pain. Pain would not be reducible to the physical.

Or think of the economic institution of cash. Lots of different kinds of stuff in fact play the role of cash, and almost any suitably portable object could play the role of cash given an appropriate combination of beliefs, desires, and conventions. Now reducibility is a transitive relation: if *A* is reducible to the *B*s and the *B*s are reducible to the *C*s, then *A* is thereby reducible to the *C*s. So if the social is reducible to the physical and the psychological, and the psychological in turn to the physical, then the social is reducible to the physical. But, given that almost any physical item at all could be cash, is it plausible that there be a single, purely physicalistic, nomological equivalent of *cash*? Hardly.

Does multiple realizability render physicalism and individualism false? Alternatively, is reducibility as characterized necessary for ontological dependence? Suppose every property which realizes pain (or cash) in all nomologically possible circumstances is itself a purely physical property. Would that not be compatible with physicalism? This suggests the possibility of a relation between the mental and the physical weaker than full reducibility but strong enough to vindicate the idea that the mental is nothing over and above the physical.

6. Supervenience

The favored term for the desired relation, one which guarantees determination without entailing reducibility, is supervenience—a term first used to characterize the relation between the moral and the natural.

It is widely held that moral differences are undergirded by natural or nonmoral differences. Suppose Francis of Assisi is good in virtue of kissing a leper with putrefying flesh (Francis of Assisi 1903). Then if Clare kisses an exactly similar leper and she is not good then there must be some nonmoral difference between the two—perhaps, unlike Francis, Clare is suppressing revulsion. Sameness of nonmoral features guarantees sameness of moral features. A difference at the moral level requires a difference at the nonmoral level. The moral supervenes on the nonmoral.

In the social realm, the individualist's battle cry might be: no difference at the social level without some difference at the individual level. Sameness at the individual level guarantees sameness of social features. That kind of supervenience claim seems eminently reasonable, one which even a holist might be embarrassed to deny. If Jones cashes a check in situation A but not in situation B, then there must be *some* difference at the individual level between the two situations—either between Jones in A and B, or

between some other individuals in A and B. Kim generalizes the supervenience condition thus:

One family of properties is 'supervenient' upon another family of properties in the sense that two things alike with respect to the second must be alike with respect to the first (Kim 1978, p. 149).

So stated the principle is open to interpretation. The necessity involved ('must be alike') may be logical or nomological. Some supervenience claims—like the supervenience of the moral on the non-moral—are a matter of logical necessity (although see Oddie 1991). Other supervenience claims are a contingent affair. Most physicalists think that the supervenience of the mental on the physical is a contingent matter—for we can imagine worlds consisting of unembodied minds differing in their mental attributes but (lacking bodies) being physically indistinguishable. The supervenience of the social on the individual is disputed—some hold it to be a logical affair (Currie 1984) others a contingent affair (Pettit 1993).

Kim's statement of the supervenience principle requires individual by individual comparison: any two *B*-indiscernible individuals must be *A*-indiscernible. It has therefore been dubbed a kind of local supervenience. This may be too stringent for characterizing the supervenience of the social on the individual (Currie 1984) and even for the supervenience of the mental on the physical (Burge 1979). In both cases, it is claimed, the supervening properties involve not only the local properties of the individual, but the relations of the individual to its environment as well as to other individuals. (Whether Clinton is the President depends not just on Clinton's psychological and physical state but on psychological and physical states nationwide. Whether Clinton thinks that *that* woman is sexy requires a real woman, not just an hallucinated woman, to be there.) Furthermore, in the social case, some of the supervening entities (First Bank of Boulder, the Stock Market, France) are in any case of a different logical type from the individuals on which they are claimed to supervene, so we need to compare more than just the social properties of individuals. Finally, a number of arguments have shown that, given reasonable assumptions, local supervenience of the logical variety entails definability, and local supervenience of the nomological variety entails reducibility (Kim 1978, 1989, Bacon 1986, Oddie 1991), thereby failing the nonreductionist desideratum.

In order to avoid these problems global supervenience has been put forward as the salient relation. Call two worlds *B*-indiscernible if the *B*s have the same extensions in both (that is, the *B*s are identically distributed in the two worlds). Then *A* globally supervenes on the *B*s if any two (nomologically) possible worlds which are *B*-indiscernible (the *B*s have the same extensions in both) are also *A*-indiscernible (*A* has the same extension in both). This principle is a

desideratum of any concept of determination, and so it is the weakest relation that might serve the supervenience theorist's purposes.

Definability entails global supervenience of both logical and nomological varieties, and reducibility entails global supervenience of the nomological variety. Since global supervenience is the weakest principle guaranteeing determination, the crucial remaining question is whether it avoids reducibility. To illustrate this possibility let us define a very simple, albeit artificial, property, *G*:

(3) *G* = *being both poor and identical either to Clare or to Francis*.

Only Clare and Francis can have *G*, and they have it just in case they are poor. So any two worlds which are identical in the distribution of poverty will be identical in the distribution of *G*. That is to say, *G* globally supervenes on poverty.

Is *G* reducible to poverty? Let us suppose all four logically possible distributions of *G* are also nomologically possible. Is *G* nomologically equivalent to something definable solely in terms of poverty? (3) is not such a definition. On the right-hand side, in addition to *poor*, the logical connectives (*and* and *or*) and the identity relation, we have reference to two individuals. Identity, like the logical connectives, is typically counted part of the logical apparatus and so a permissible component of a definition. But the individuals Francis and Clare are not. If we disallow direct reference to Francis and Clare, or to any entity like the property of *being identical to either Francis or Clare* which could perform the same logical role, then *G* cannot be defined in terms of *poverty* alone. Given this, there is no nomological equivalent of *G* in terms of *poverty*. *G* is neither definable in terms of, nor reducible to, *poverty*.

Like nearly every philosophical argument, this one involves some disputed assumptions, but if it is right then it illustrates the possibility of global supervenience without reducibility (a substantial simplification of the point first made by Hellman and Thompson (1975)).

7. Conclusion

If global supervenience yields determination without reducibility then this is good news indeed for non-reductive physicalists and non-reductive individualists. The determination component articulates a clear sense in which the social and the mental are nothing over and above the physical. In creating the actual world God would not first have to create physical stuff and the laws governing it, then add a whole layer of extra mental entities, properties, and laws, and finish it up with social entities and laws. The physical by itself would suffice, guaranteeing its mental and social superstructure. The nonreductive component, how-

ever, would guarantee that mental and social concepts would still be essential for discourse and theorizing. The social sciences could enjoy methodological autonomy without having to apologize for a bloated ontology.

See also: Atomism and Holism: Philosophical Aspects; Causes and Laws: Philosophical Aspects; Functional Explanation: Philosophical Aspects; History of Science; History of Science: Constructivist Perspectives; Individualism versus Collectivism: Philosophical Aspects; Physicalism and Alternatives; Truth, Verification, Verisimilitude, and Evidence: Philosophical Aspects

Bibliography

- Armstrong D M 1968 *A Materialist Theory of the Mind*. Routledge, London
- Bacon J 1986 Supervenience, necessary coextension, and reducibility. *Philosophical Studies* 49: 163–76
- Burge T 1979 Individualism and the mental. *Midwest Studies in Philosophy* 4: 73–121
- Carnap R 1928 *Der Logische Aufbau der Welt*. Weltkreisverlag, Berlin
- Carnap R *Philosophical Foundations of Physics*. Basic Books, New York
- Currie G 1984 Individualism and global supervenience. *British Journal for the Philosophy of Science* 35: 345–58
- Davidson D 1970 Mental events. In: Foster L, Swanson J W (eds.) *Experience and Theory*. University of Massachusetts Press, Amherst, MA, pp. 79–101
- Francis of Assisi St 1903 *The Little Flowers of Saint Francis* [trans. Arnold T W]. Dent, London
- Hellman G, Thompson F 1975 Physicalism: ontology, determination, reduction. *Journal of Philosophy* 72: 551–64
- Kim J 1978 Supervenience and nomological incommensurables. *American Philosophical Quarterly* 15: 149–56
- Kim J 1989 The myth of non-reductive materialism. *Proceedings and Addresses of the American Philosophical Association* 63: 31–47
- Lewis D 1993 *Philosophical Papers*. Oxford University Press, New York, Vol. 1
- Nagel E 1961 *The Structure of Science*. Harcourt, Brace & World, New York
- Oddie G 1991 Supervenience, goodness, and higher-order universals. *Australasian Journal of Philosophy* 69: 20–47
- Oddie G, Tichy P 1990 Resplicing properties in the supervenience base. *Philosophical Studies* 58: 259–69
- O'Neill J 1973 *Modes of Individualism and Collectivism*. Heinemann Educational, London
- Pettit P 1993 *The Common Mind*. Oxford University Press, New York
- Putnam H 1967 Psychological predicates. In: Capitan W H, Merrill D D (eds.) *Art, Mind and Religion*. Pittsburgh University Press, Pittsburgh, PA, pp. 37–48
- Quine W O 1961 Two dogmas of empiricism. *From a Logical Point of View*. Harvard University Press, Cambridge, MA
- Ramsey F P 1931 *The Foundations of Mathematics*. Kegan Paul Tranch, Trubner, London

Russell B 1919 *Introduction to Mathematical Philosophy*. Macmillan, New York

Ryle G 1949 *The Concept of Mind*. Barnes & Noble, New York

Smart J J C 1959 Sensations and brain processes. *Philosophical Review* 68: 141–56

G. Oddie

Reference and Representation: Philosophical Aspects

Paintings, barometers, sentences, and minds are examples of things that represent—they represent things as being one way or another. In this sense, they can be thought of as being about, or as referring to, certain things: the putative subject of the painting, air pressure, or approaching weather, etc. This article addresses the connections between representation and reference; the notion of representational content; representation in language and mental representation; discusses some theories of reference for names; and discusses some issues about the representational content of mental states.

1. Representation: the Most General Notion

What is common to all cases of representation is a correspondence between various states of what does the representing and various states of what is represented. Thus, the number of tree rings represents the age of a tree by virtue of the fact of the possibility of matching different tree ring states with the different ages. In this case, the correspondence is a very simple one: n tree rings corresponds to an age of n years. Likewise, there is a simple correspondence between various states of petrol gages and various levels of gas in the tank to which the gage is connected. But there are highly complex cases as well. Inscriptions in secure codes of credit card numbers represent those numbers, but via extraordinarily complex correspondences. When there are pairwise correspondences between differing states, there is the possibility of carrying information. This is how the number of tree rings can give information about the age of a tree. But, of course, the nature of the correspondence needs to be known for this possibility to be realized.

In many cases, the relevant correspondence relation has a causal underpinning that runs from the various states being represented to the various states doing the representing. For example, the level of gas in the tank causes the reading on the gage, and the appearance of the subject causes the array of paint on the canvas (via the actions of the painter). But structures can represent how things *will* be; a drawing may represent where a playground will be located but is, obviously, not

caused by where it will be. There is representation in the wide sense when there are nonaccidental correspondences between states of affairs. There will be a causal explanation of these nonaccidental correspondences but its precise nature varies from case to case. (See Dretske 1995, Chap. 1 for more on the variety, but note that he ties representation to *purpose*.)

1.1 Representational Content and Reference in General

Things that represent are said to have representational content, that content being how things are being represented as being. The usual way of capturing representational content is set theoretically: content is given by a division of all the complete ways things might be—the event space in the language of statistics—into those that accord with how things are being represented as being and those that do not. Complete ways things might be are ways things might be in which *everything* is settled one way or the other. They are often called possible worlds. The content is then the set of possible worlds that accords with how things are being represented as being. For example, if a petrol gage represents a gas tank as half-full, its being half-full is the content, and that is identified with the set of possible worlds where the tank is half-full. There are, though, considerable differences of opinion about the nature of the possible worlds appealed to in this style of account. This article touches briefly on these in the discussion of representation in mind and language. (For an accessible discussion of the main views, see Bigelow and Pargetter 1990, Chap. 4.)

One advantage of the set-theoretic approach to content is that it allows a simple account of when things are as they are being represented as being. Things are as they are being represented as being—the representation is true or correct or accurate—if and only if the way things actually are is a member of the content. Or, to say it in terms of possible worlds, the representation is accurate if and only if the actual world is a member of the set of worlds which is the content of the representation. In still other words, states that represent say that things are a certain way in the sense of delimiting the possibilities for how they might be, and count as accurate when how things actually are is inside the relevant region of ‘logical space.’

A second advantage is that it allows a simple account of the widest sense in which states that represent are about, or refer to, this, that, or the next thing. Suppose that a barometer’s reading such and such represents that the surrounding air pressure is so and so. Intuitively, the reading is about the air pressure being so and so. This can be spelt out as follows. In every possible world where the reading is correct, in every world where things are as the barometer represents them, the air pressure is so and so. States which

represent are about or refer to what is the case in all the worlds where things are as they are being represented as being. (For more on ‘aboutness’ and the difference between being about an object and a property, see Dretske 1995, Chap. 1, Sect. 5.)

1.2 Mental Representation

A footprint can be thought of as representing the shape of the foot that made it an hour ago. But equally it can be thought of as representing the weight of the person who made it, the shape of the footprint itself half an hour ago, the shape of the foot that made it one day before the foot print was made, and so on. Almost any state that represents will have many representational contents, and it is a nonsense question to ask which is its real content, to ask what the state really represents.

Among the representational structures of especial interest to humans are their brains, and, especially, those parts of our brains that constitute our powers of mental representation. In particular, humans *believe* that things are thus and so, and do so by virtue of the processes in their brains which constitute their belief representing that things are thus and so. These processes *qua* states of brains will themselves have many contents including, for example, that the temperature of the surrounding brain fluid is thus and so, but only one will be the content of what is believed. To suppose otherwise would make what is believed absurdly indeterminate. Therefore, a major question in the philosophy of mind is what determines, out of all the contents of the brain structures that are beliefs, the one which is the content of those beliefs.

One popular answer is: the functional role the structure plays, where this is understood as the role the structure plays in mediating between: inputs to subjects, typically via their sense organs, outputs understood in terms of behavioral responses, and other internal states. However, there are many different ways of fleshing out this answer. (For a recent survey, see Braddon-Mitchell and Jackson 1996, Chaps. 3, 5.)

2. Linguistic Representation and Reference

2.1 Introduction

Language is an especially powerful representational device. With it, highly detailed claims are possible concerning how things are that far outrun, in terms of detail and precision, what can be done with barometers, petrol gages, etc. There is a syntactical distinction between sentences and parts of sentences that roughly corresponds to an important semantic or representational distinction.

A sentence makes a claim about how things are. This is how it gets to be true or false: true when the way the sentence represents things is the way they are, false otherwise. This explains the self-evident status of the biconditional 'Snow is white' is true if and only if snow is white. The left-hand side is true if and only if things are as 'Snow is white' represents them as being. Moreover, as a good way of stating how 'Snow is white' represents things as being is to use the sentence itself, the biconditional comes out true.

By contrast, parts of sentences like 'Snow' and 'is white' contribute in a systematic way to how whole sentences in which they appear represent things as being, but do not themselves represent how things are. This is why they are not themselves either true or false. They can, however, be thought of as referring to, respectively, snow and white things. This is because the contribution that the presence of the word 'snow' makes to the conditions under which 'Snow is white' is true, that is, to how the sentence represents things as being, is such that the sentence cannot be true unless there is some snow. Similarly, the contribution that the presence of 'is white' makes is such that the sentence cannot be true unless there are some white things.

Many recent discussions have focused on the referential properties of names, and many authors restrict the use of the term 'refer' to names. They prefer to talk of the relation between predicates (and, sometimes, common nouns) and things in terms of whether the predicates are *true of* the things or whether the things *satisfy* the predicates.

The next section considers the three main theories of reference for names.

2.2 Theories of Reference for Names

A word like 'London' is used for something taken to have certain properties. This is why it is not used for water or Paris—it is known that they do not have the right properties. On the other hand, the large city that is the capital of England does have the right properties. In short, certain properties or descriptions are associated with the word 'London' and this guides its use. This suggests a simple theory of reference for names in general: 'N' refers to that which satisfies the descriptions associated with 'N.' If we use 'D' for those descriptions, another way of stating this theory is by saying that 'N' is short for the definite description 'the D.'

A famous advantage of the *description theory of reference* is that it can explain why certain identity sentences are informative. Everything is, trivially, self-identical. How then can it be informative to be told that Hesperus is identical to Phosphorus? The answer is that the first name is associated with being the star first seen in the evening and the second name is associated with being the last star to disappear from the night sky as morning approaches, and it is

informative to be told that one thing has these two properties. (The classic source for claims of this kind is Frege 1952.)

Recent discussions in the philosophy of language have raised a number of objections to the description theory of reference. (A good account of the major objections can be found in Devitt and Sterelny 1987. A classic source is Kripke 1980.)

First, if 'N' is an abbreviation of 'the D,' it should make no sense to suppose that N might have failed to be D. But, in fact, it typically makes perfect sense. For example, the descriptions associated with 'Shakespeare' include as a major part that he was the author of certain plays. (Surprisingly little is known about Shakespeare's life.) But it makes perfect sense to suppose that Shakespeare did not write the plays associated with his name. This, indeed, is what those who think that Bacon wrote them do suppose. A similar point can be made for names of natural kinds. The description associated with 'water' is something like potable, clear liquid that fills lakes and falls as rain from the sky—this is roughly what is in a dictionary. However, if on some remote planet—traditionally called 'Twin Earth' by philosophers—there is found clear potable stuff in its lakes which falls from its sky, etc. that is not H₂O, it seems that the right thing to say would be that this stuff—XYZ, as it is often dubbed—is not water. As it would naturally be said, 'It has been discovered that water is not the only clear potable liquid that fills lakes, falls from skies etc. in the universe.' Conversely, if H₂O on Twin Earth were black and rather tar-like, it would still be water. As it would naturally be said, 'It has been discovered that water comes in a quite different form on Twin Earth from its form at room temperature on Earth.' (The classic source for Twin Earth is Putnam 1975.)

Second, if names are abbreviated definite descriptions, then their reference should in general be variable. By changing what is supposed to be the case, it should be possible to shift what they refer to. This is because the reference of definite descriptions is in general shiftable. 'The President of the USA in 1998' refers to Bill Clinton, but had he lost to George Bush, it would refer to Bush and not Clinton. But it seems that the reference of names cannot be shifted; they are, as it is said, rigid designators. 'Clinton' always refers to the very same person.

Description theorists have proposed a number of replies to these (and other) objections. For example, they have urged that it is wrong to conflate the description that settles the reference of a name with the description (s) most famously associated with a name. Although Shakespeare is most famous for the plays, the key description that settles the reference of the name is something like 'the person who is a certain kind of causal origin of the use of the term "Shakespeare."' The debate over whether or not Shakespeare wrote the plays is the debate over whether the person who is a certain kind of causal origin of the occurrences

of the name on the covers of plays etc. did indeed write the plays. (For a recent catalogue of replies that the description theorists might make to various objections, see Jackson 1998.)

The causal theory of reference is the best known alternative to the description theory. The fundamental idea is that 'Aristotle' in our mouths refers to the person it does because of an initial act of baptism that tags that person with a name which then got passed on from speaker to speaker as the name to use for that tagged person until it ended up in our mouths. Of course, the name itself underwent various changes as it passed from the ancient Greek to Modern English. A similar account is suggested to explain why 'water' refers to H_2O . H_2O was baptized with the word 'water' (in ignorance of the fact that it was H_2O that was being baptized) and then the word was passed on as the word to use for that which had been baptized. This account explains why the clear potable stuff that fills the lakes etc. on Twin Earth is not water, is not what is referred to when the word 'water' is used, by appealing to the fact that it was not the natural kind baptized.

There are a number of questions to ask about the causal theory. Is it superior to a description theory that gives a prominent role to a description like 'was baptized ... and subsequently used for what had been so baptized'? Is it supposed to apply to the naming of as yet unborn children? Remembering that causal interactions are invariably with a very large number of factors, should a descriptive element be added to ensure the 'right' reference? (Some of these questions are addressed in Devitt and Sterelny 1987.)

Finally, there is the direct reference theory of reference. On this view, names like 'Aristotle' refer directly to Aristotle. The only semantic property of a name is its bearer. It is hard to discuss this view because it steadfastly refuses to address what seems a perfectly legitimate question, namely, what makes it true that 'Aristotle' refers to Aristotle and not to Plato or my dog? It is also hard to see what it can say about empty names. It has sometimes been speculated that Socrates never existed. 'He' was merely an expository device of Plato's. Although this is almost certainly false, it makes sense to suppose that it is true. But is this possible if 'Socrates' has no meaning apart from its bearer? (For references and more discussion and dissent, see Devitt 1996, Chap. 4.)

3. Mental Representation and Linguistic Representation

3.1 How are they Connected?

As noted above, our minds represent; in particular, to believe is to represent that things are thus and so. Usually, the easiest way for humans to express how they believe things are is to use a sentence. For example, the sentence 'Clinton will not be impeached'

was used in February 1999 by many commentators on American politics to express what they believed. It is obvious that there is an intimate connection between mental and linguistic representation. What is controversial is the extent to which mental representation requires linguistic representation.

One view is that the phenomenon of mental representation in general, and belief in particular, is independent of language; dogs, for example, really do have thoughts and beliefs. Language is simply a device for telling about how things are taken to be. This view can though allow that there are some beliefs one could not have without having a language. It is, for instance, hard to make sense of a creature believing that there are a prime number of apples in the bowl in front of it in the absence of that creature having mastery of a complex coding system, i.e., a language.

However, some hold that belief in general requires language, where language means something substantially more sophisticated than the kind of signaling systems that, e.g., dogs have. On this view, although contentful utterance requires *mental* representation—telephone answering machines do not *mean* anything by the sentences they produce—in order to be a believer, you must have a language. Sometimes, this seems to be, at bottom, a verbal stipulation about the use of the term 'belief': dogs have quasibelief which is very like belief except that it is not expressed in a public language. But sometimes it is clear that the view is that dogs do not have anything like belief at all. A major question for this view is the very close connection between perception and belief. Perceptual experience seems to be shot through with belief-like representational content: the experience of it's raining seems to be a state that by its very nature points to believing that it is raining, though it may be overridden by collateral information, for example, about the operation of a nearby sprinkler system. It is, accordingly, hard for advocates of the more radical view to avoid the implausible view that animals like dogs do not see or hear. (For a defense of the view that belief requires language, see Davidson 1984, Chap. 11, see also McDowell 1994.)

3.2 Capturing the Content of Mental and Linguistic Representation

It was noted earlier that the standard way of capturing representational content is in terms of divisions among complete ways things might be, possible worlds. The problem of the content of logically equivalent sentences and beliefs raises an important question for how to think of these possible worlds. Intuitively, the sentences 'X is an equilateral triangle' and 'X is an equiangular triangle' differ in meaning. Likewise, intuitively, the belief that X is an equilateral triangle and the belief that X is an equiangular triangle are different beliefs. In both cases, the intuition rests on

the fact that it was a *discovery* that, necessarily, every equilateral triangle was equiangular, and conversely. But, on standard conceptions of possible worlds, 'X is an equilateral triangle' and 'X is an equiangular triangle' are true in exactly the same possible worlds precisely because they are logically equivalent.

There have been, broadly speaking, two reactions to this problem. One has been to bite the bullet and insist that the contents are the same. (The most worked out version of this response can be found in Stalnaker 1984.) The other has been to argue that, for the purposes of capturing content, a more fine-grained conception of the entities among which the division is being made is needed. For example, possible worlds might be thought of as constructions out of certain concepts so that, for example, the constructions at which 'X is an equilateral triangle' is true differ from those at which 'X is an equiangular triangle' is true by virtue of the former having the concept *side*, where the latter have the concept *angle*. (See Peacocke 1983, for an overview of the issues see Braddon-Mitchell and Jackson 1996.)

3.3 Mental Representation and the Language of Thought

The functionalist answer to how a belief gets to have the belief content that it does—and, more generally, how a mental state gets to be the mental state that it is—was described earlier, and it was noted that there are many views about how to spell out the key functional roles. However, a popular answer holds that an important part of any functionalist account is the way belief and desire work together to produce behavior: roughly, subjects behave in a way such that what they desire is satisfied if what they believe is true (Stalnaker 1984, Braddon-Mitchell and Jackson 1996).

An alternative approach—the language of thought hypothesis—appeals to the idea that creatures capable of representational thought, creatures capable of having intentional mental states, have a language of thought. When subjects believe that snow is white, they have in their heads a sentence in 'mentalese' that says that snow is white. (This approach allows that dogs may have beliefs, because it is possible to have an internal language of thought, while lacking a public language in which to express thoughts.) This approach gives functional role a place in determining the difference, for example, between belief and desire, but content is treated via a 'semantics for mentalese.' Consider, for example, subjects who believe that snow is white. They have in their heads a token that stands for snow (the mentalese word for snow), a token that stands for whiteness (the mentalese predicate for whiteness), put together in a structure that makes up the mentalese sentence *snow is white*. What makes it true that they believe rather than desire that snow is white is the functional role this structure plays: it

serves more of a 'responding to the world' role and less of a 'changing the world' role. (A classic source is Fodor 1988.)

This approach raises many important issues. One is how we should spell out the crucial relation between the words of mentalese and what they stand for? What makes it true that a structure in the brain is the mentalese for snow and not water, for example? One proffered answer is some kind of causal cum informational link between brain token and what is tokened, but the details are controversial. A second issue is whether neurological investigations might refute the language of thought hypothesis. Some cognitive scientists have argued that the way the brain represents the environment is more like the way maps represent than the way sentences represents (see Braddon-Mitchell and Jackson 1996).

4. Representation and Context

Representational content is often thought of as context dependent. For example, what a petrol gage represents about the level of gas in the tank depends on the nature of the connection between gage and tank. Connected in one of the more usual ways, the gage having its pointer in the far left position indicates that the tank is empty, but connected differently that pointer position might represent that the tank is full. That is to say, identical gage configurations can have very different representational contents depending on the gage's 'context.' Or, as it is often put, representational content with respect to the amount of gas in the tank does not *supervene* on a petrol gage's configuration. As it is said, the content is broad or wide or context dependent.

It is equally possible, however, to think of any given reading as saying: either the gage is connected to the tank in way W_1 and the gas level is L_1 , or the gage is connected in way W_2 and the level is L_2 , or In this way of thinking, a way naturally adopted when the manner of connection between gage and tank is unknown, the content is not dependent on the manner of connection.

In other words, and speaking more generally, there are two essentially equivalent ways of thinking of the interplay of representational content and context. In the first, it looks like this:

In C_1 , S represents that P_1
 ...
 In C_i , S represents that P_i
 In the second, it looks like this:
 S represents that if C_1 then P_1
 ...
 S represents that if C_i then P_i

In the first way of thinking, content comes out as context dependent or broad; in the second, it comes out context independent or, as it is said, narrow.

A pressing, currently much debated question about both mental and linguistic content is whether we should think of it in the context dependent or in the context independent way. Or, equivalently, is what a person believes and what a person claims about how things are when they speak and write a broad or narrow property of them?

The traditional view is that mental and linguistic content is narrow. A principal source for this view is a powerful intuition often exploited by science fiction writers. The idea is that if the stimuli actually impinging on subjects via their environments is duplicated *exactly*, say, by a computer, the subjects will not be able to tell the difference. This is the core idea behind flight simulators and virtual reality machines, of course. Although these devices fall short of perfectly duplicating the impacts of their target environments on their subjects, their makers take it for granted that were perfect duplication achieved, subjects would be unable to tell whether, for example, they were really flying or were in a simulator. This suggests that subjects' mental lives, including how their mental states represent things as being, supervene on their nature from the skin in (or, perhaps, from their central processors in). How their mental states represent things as being is, therefore, a narrow property of subjects: two subjects which are identical from the skin in are identical in how they take, desire etc. their worlds to be. Moreover, if the content of sentences is given by the content of the mental states they serve to express, it will follow that linguistic content is narrow.

However, many philosophers, often influenced by the Twin Earth example discussed earlier, repudiate this intuition. Consider, runs their argument, Mary and Mary's duplicate from the skin in, Twin Mary, on Twin Earth. They both use the word 'water' for the clear, potable etc. stuff around them, but the conditions under which what is said is true, and equally those under which what is thought is true, differ. What Mary says and thinks when she uses the sentence 'There is water nearby' is true if and only if there is H₂O nearby, but what Twin Mary says and believes will be true if and only if there is XYZ nearby. Hence, runs the argument, the contents of the words and thoughts of Mary and Twin Mary differ; ergo, content is context dependent.

The issue is complicated, however, by the fact that it is plausible that there is an indexical element to the use of a great many words including 'water.' It is plausible that the word 'water' is used in part to mean the stuff that the users of the word have come across which is clear, potable etc. If this is right, a plausible explanation of why Mary and Twin Mary refer to different stuff in thought and talk when they use the word 'water' is that they and their language communities interact with different stuff when the word 'water' is used. That is, Mary refers to the clear, potable stuff that *she* and *her* community interact with, whereas Twin Mary refers to the stuff *she* and *her*

(different) community interact with, and the reason they refer differently is that the stuff they interact with is different. This means that there is *a* sense in which what they say and think is the same when they say and think 'There is water nearby.' They both say and think that some of the clear potable etc. stuff that *she* herself interacts with is nearby. This does not affect the point that the truth conditions of what they say and think differ, but they differ in the same way that the truth conditions of 'I am bald' differ depending on who is speaking. There is a good sense in which each person who says or thinks 'I am bald' says or thinks the same thing, namely, that I myself am bald. (There is a large literature on this important question; see Pessin and Goldberg 1996 and Braddon-Mitchell and Jackson 1996.)

See also: Intentionality and Rationality: A Continental-European Perspective; Intentionality and Rationality: An Analytic Perspective; Meaning and Rule-following: Philosophical Aspects Cognitive Science: Overview; Intentionality and Rationality: A Continental-European Perspective; Intentionality and Rationality: An Analytic Perspective; Meaning and Rule-following: Philosophical Aspects; Methodological Individualism: Philosophical Aspects; Perception: Philosophical Aspects

Bibliography

- Bigelow J, Pargetter R 1990 *Science and Necessity*. Cambridge University Press, Cambridge, UK
- Braddon-Mitchell D, Jackson F 1996 *The Philosophy of Mind and Cognition*. Blackwell, Oxford, UK
- Davidson D 1984 *Inquiries into Truth and Interpretation*. Oxford University Press, Oxford, UK
- Devitt M 1996 *Coming to our Senses: A Naturalistic Program for Semantic Localism*. Cambridge University Press, Cambridge, UK
- Devitt M, Sterelny K 1987 *Language and Reality*. MIT Press, Cambridge, MA
- Dretske F 1995 *Naturalizing the Mind*. MIT Press, Cambridge, MA
- Fodor J 1988 *Psychosemantics*. MIT Press, Cambridge, MA
- Frege G 1952 Translations from the philosophical writing of Gottlob Frege. In: Geach P, Black M (eds.) *Translations from the Philosophical Writings of Gottlob Frege*. Philosophical Library, New York
- Jackson F 1998 Reference and description revisited. *Philosophical Perspectives* 12: 201–18
- Kripke S 1980 *Naming and Necessity*. Harvard University Press, Cambridge, MA
- McDowell J 1994 *Mind and World*. Harvard University Press, Cambridge MA
- Peacocke C 1983 *Sense and Content*. Oxford University Press, Oxford, UK
- Pessin A, Goldberg S (eds.) 1996 *The Twin Earth Chronicles*. M E Sharpe, Armonk, NY

Putnam H 1975 The meaning of 'meaning'. In: Putnam H (ed.) *Mind, Language and Reality*. Cambridge University Press, Cambridge, UK
 Stalnaker R 1984 *Inquiry*. MIT Press, Cambridge, MA

F. Jackson

Referendums

The referendum allows citizens to vote on a public issue in an election at the local, subnational, or national level. The vote may be advisory or mandatory, statutory or constitutional. Most democracies have voted on a referendum, and the incidence of referendums increased in the last quarter of the twentieth century. The referendum agenda does not have a distinctive ideological slant (Butler and Ranney 1994). National and subnational referendums on matters like entry into international agreements, new constitutional arrangements, and changes in monetary systems have been frequent topics put to voters. Recently in the USA conservative issues have driven the process, but liberal concerns dominated in other decades (Magleby 1994).

Several notable referendums from the last quarter of the twentieth century include the vote in the UK on entry into the European Economic community, the separatist votes in Quebec, the devolution votes in Scotland and Wales, the votes in several European countries on a common currency, and a growing interest in the process in the new democracies of Africa, South America, or Eastern Europe. The USA is one of the few democracies that never experimented with a national referendum, although the process has occasionally been proposed. One notable proposal for a national referendum in the USA was the failed Ludlow Amendment to the US Constitution, which would have required a national referendum before war could be declared (Magleby 1984). More recently, forms of national referendums have been proposed by former Representative Jack Kemp, among others (Kemp 1979).

Referendums can be proposed by the elected representatives and put before the voters in a special or regular election. When such referendums are optional they are called facultative (Bogdanor 1994). The popular referendum permits citizens to petition to place a law or constitutional change before the voters that has previously been enacted by the legislature.

Referendums are sometimes called plebiscites, a reflection of the early use of the process in the Greek city-states. More recently, critics of the process point to plebiscites called by Hitler and other totalitarian leaders as examples of how the process can be used for undemocratic ends. Referendums were found to be consistent with representative democracy in the USA

by the Supreme Court (*Pacific States Telephone and Telegraph Company vs. Oregon* 1912).

In their expansive review of referendums around the world, David Butler and Austin Ranney (1994) find that the subject matter of referendums falls into four main categories: constitutional issues especially after a revolution, territorial breakup, or when changes in the rules are needed; territorial issues; moral issues on matters that reflect deep divisions and which may best be resolved by a popular vote (e.g., alcoholic beverage regulation, divorce, or abortion); and finally a range of miscellaneous issues.

A more common form of citizen petitioning is the initiative process. Under the initiative, citizens may petition to place before the voters statutes (statutory initiative) or constitutional changes (constitutional initiative). As with the popular referendum, access to the ballot is based on the proponents obtaining a specified number of petition signatures. Some governments permit the direct initiative whereby petitioners go directly to the ballot, while some require the proposal to be considered by the legislature before it may go to the voters (the indirect initiative).

At the national level, the initiative process is more identified with Switzerland. Swiss cantons also use the initiative with as many as four elections per year where voters decide initiatives or referendums (Kobach 1994).

Imitating the Swiss, several American states adopted the initiative and popular referendum process—sometimes called direct legislation—shortly after the turn of the last century. As with other Progressive reforms like the direct election of US Senators and the direct primary, reformers hoped to expand the role of voters and limit the role of elected officials and intermediary institutions like political parties, legislatures, or elected executives. The United States experienced a second wave of reform during the 1970s and since; these reforms again expanded suffrage, pressed for more expansive use of direct primaries especially in presidential elections, and renewed interest in and use of initiatives and popular referendums.

In the United States, the initiative and popular referendum occur most frequently in western states (see Fig. 1). These states were in their formative political years during the peak of the movement, and this region was a stronghold for the Progressives (Price 1975). Only five states have adopted the process since the 1920s: Alaska (1956), Wyoming (1968), Illinois (1970), and Florida (1972). Since the mid-1970s, several states have considered adopting the process but only Mississippi has done so. Direct legislation is even more common at the local level in the United States, although precise counts of the numbers of jurisdictions with the process nor the extent of use do not exist.

As in Switzerland, the most common form of initiative and popular referendum in the USA requires

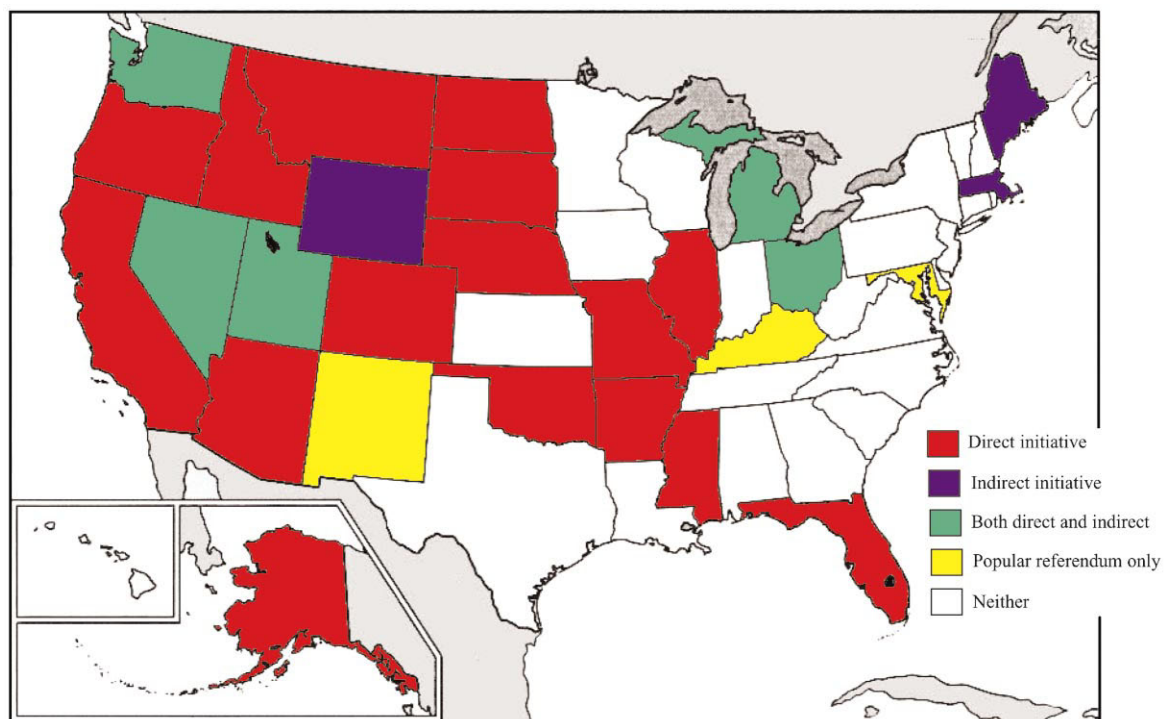


Figure 1
Provisions for initiative and popular referendum in the USA

voters to vote for or against specific statutes or constitutional amendments. As a result, ballot measures are lengthy and written in technical and legal language. Official summaries provided by state election officials are also written at a sophisticated level, making it difficult for voters to often obtain a simple neutral summary of the proposition (Magleby 1984). The length and complexity of the process in the USA makes the process less accessible to less educated voters, having a policy bias against the poor and racial minorities. Initiatives are preferred over popular referendums because petitions may be in circulation longer, and initiatives permit their sponsors to write the language of the measure rather than the legislature. An example of an initiative that was in effect a referendum is the 1964 vote on open housing in California (Wolfinger and Greenstein 1969).

Referendums put on the ballot by state legislatures or changes to the state constitution occur frequently. Every state except Delaware requires that changes in the state constitution be approved by voters. Voters are much more likely to vote affirmatively on these legislative referendums than they are on initiatives. More than 60 percent of legislative referendums pass while the same percent of initiatives fail (Magleby 1994).

Since the mid-1970s, the initiative has experienced

renewed usage (see Fig. 2). The growth in initiative usage includes an even more dramatic increase in the number of proposed initiatives in the petition circulation process that fail to reach the ballot. Why the increase? Initiative activists have learned that the process is a powerful agenda-setting device. A vote in a single state can push an issue onto the national agenda, as votes on tax reduction, elimination of affirmative action, a nuclear weapons freeze, and legislative term limitation has demonstrated. The news media give substantial attention to contentious initiatives, calling attention to the sponsors of the measures and, in some instances, advancing political careers. California gubernatorial candidates make frequent use of the process for self-promotion and governors in noninitiative states advocate the process to establish populist credentials. The large sums of money spent on initiatives have also called attention to the renewed use of the process. Not surprisingly, an initiative industry has developed that helps with petition circulation, legal appeals, campaign management, and media advertising. The process has become an integral part of the governance of several states, most notably California (Schrag 1998).

Absent from the initiative process are the normal checks and balances of the US political system. There is no bicameral check except for Nevada's requirement

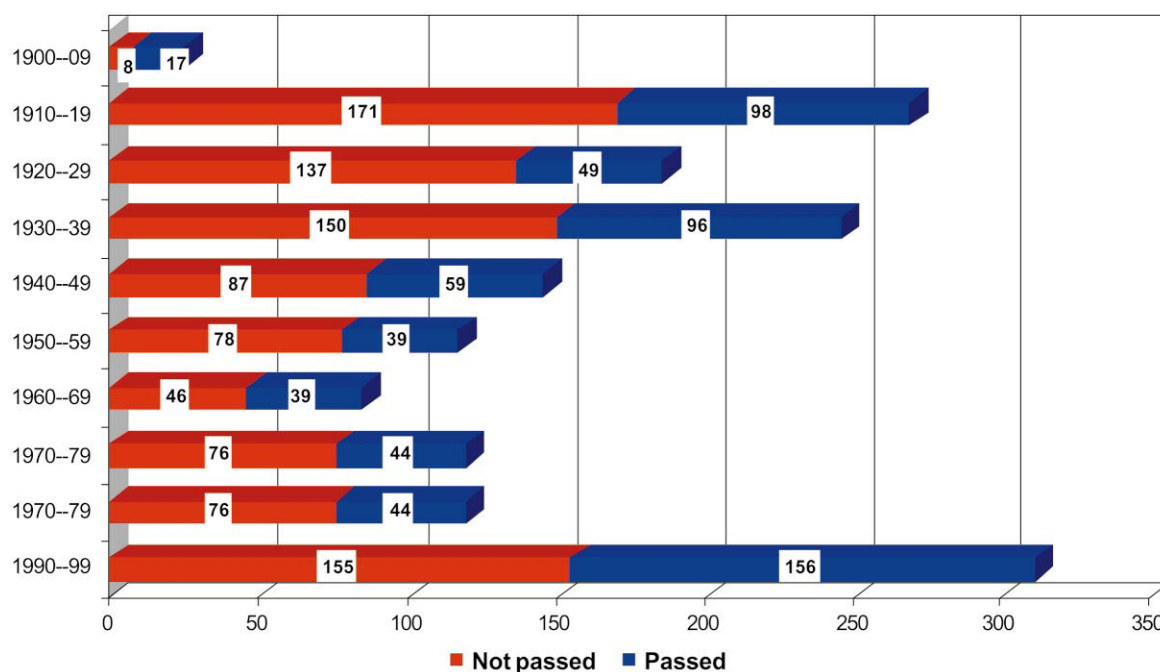


Figure 2

History of initiatives and referendums USA 1900-97. Source: Initiatives and Referendums Institute

that constitutional initiatives must pass in identical form in two consecutive elections. There is no gubernatorial veto, no hearing process, and no ability to amend or modify. The only institutional check on the initiative is the courts. Judicial review of initiatives occurs often before the election on matters having to do with signature collection, ballot title, and summary, and whether the initiative contains more than one subject. Postelection judicial review is more concerned with constitutionality, and federal courts have not been timid about overturning the people's votes. State courts, like the California Supreme Court, faced a voter backlash in the 1980s because of the frequency with which that court had declared initiatives unconstitutional.

With the growth in democracy around the world, there has followed renewed interest in referendum devices as a mode of popular consultation. Indeed, some analysts point to a coming clash between direct and representative democracy in the USA, in part precipitated by technology (Broder 2000). States like Oregon have experimented with statewide referendums entirely with mail balloting and some see the Internet as a way to facilitate frequent referendum voting (Becker and Slaton 2000).

As the process has grown in importance, so has research on it. Much of the research focuses on voter decision making in the absence of party cues. Ideology and elite cues are important to voting in this context

(Magleby 1984, Lupia 1994). Other research has focused on the role of consultants in the process, finding that these campaign professionals find the process more lucrative than candidate elections, and prefer it because it gives them greater creative license and means they can avoid the problems associated with candidate scandals and intrusive spouses (Magleby and Patterson 2000).

Another expanding research emphasis is the policy consequences of initiatives. Scores of books have been written on the impact of California's Proposition 13 property tax reduction and similar measures in other states (Sears and Citrin 1985, O'Sullivan et al. 1995). Other books examine the policy and institutional implication of the process (Gerber et al. 2001).

One major methodological challenge to researchers is the lack of a centralized database on local initiatives. Administration of local elections is left to counties and rarely are records kept for more than a few years. Given the large number of counties and the frequency of referendums at the local level, the task of assembling this database has proven difficult.

Given the likely ease of electronic referendum voting, research on how voters behave in this environment will be important. Do electronic voting schemes limit some voters' ability to participate? Would voting electronically further disenfranchise poor and less educated voters from a process that they already find difficult to navigate?

See also: Constitutionalism; Democracy; Democracy, History of; Electoral Systems; Legislatures: United States; Participation: Political; Political Representation

Bibliography

- Becker T, Slaton C D 2000 *The Future of Teledemocracy*. Praeger, Westport, CT
- Bogdanor V 1994 Western Europe. In: Butler D, Ranney A (eds.) *Referendums Around the World: The Growing Use of Direct Democracy*. AEI Press, Washington, DC
- Broder D S 2000 *Democracy Derailed: Initiative Campaigns and the Power of Money*. Harcourt, New York
- Butler D, Ranney A 1994 Practice. In: Butler D, Ranney A (eds.) *Referendums Around the World: The Growing Use of Direct Democracy*. AEI Press, Washington, DC
- Gerber E, Lupia A, McCubbins M, Kiewit D 2001 *Stealing the Initiative: How State Government Responds to Direct Democracy*. Prentice-Hall, Upper Saddle River, NJ
- Kemp J 1979 *An American Renaissance: A Strategy for the 1980s*. Harper and Row, New York
- Kobach K 1994 Switzerland. In: Butler D, Ranney A (eds.) *Referendums Around the World: The Growing Use of Direct Democracy*. AEI Press, Washington, DC
- Lupia A 1994 Shortcuts versus encyclopedias: Information and voting behavior in California insurance reform elections. *American Political Science Review* **88**: 63–76
- Magleby D B 1984 *Direct Legislation: Voting on Ballot Propositions in the United States*. Johns Hopkins University Press, Baltimore
- Magleby D 1994 Direct legislation in the American States. In: Butler D, Ranney A (eds.) *Referendums Around the World: The Growing Use of Direct Democracy*. AEI Press, Washington, DC
- Magleby D B, Patterson K D 2000 Campaign consultants and direct democracy: Politics of citizen control. In: Thurber J E, Nelson C J (eds.) *Campaign Warriors: The Role of Political Consultants in Elections*. Brookings Institution Press, Washington, DC
- O'Sullivan A, Sexton T, Shefrin S 1995 *Property Taxes and Tax Revolts: The Legacy of Proposition 13*. Cambridge University Press, Cambridge, UK
- Pacific States Telephone and Telegraph Company vs. Oregon* 233 US 118 (1912)
- Price C M 1975 The initiative: A comparative state analysis and reassessment of a western phenomenon. *Western Political Quarterly* **28**: 243–62
- Schrag P 1998 *Paradise Lost: California's Experience, America's Future*. University of California Press, Berkeley, CA
- Sears D, Citrin J 1985 *Tax Revolt: Something for Nothing in California*. Harvard University Press, Cambridge, MA
- Wolfinger R E, Greenstein F I 1969 The repeal of fair housing in California: An analysis of referendum voting. *American Political Science Review* **63**: 755

D. B. Magleby

Reflective Equilibrium

A 'reflective equilibrium' is the end point of a deliberative process in which we reflect on and revise

our beliefs about an area of inquiry. The inquiry might be as specific as the question, 'What is the right thing to do in this case?' or 'Is this the correct inference to make?' Alternatively, the inquiry might be much more general, asking which theory or account of justice or right action we should accept, or which principles of inductive reasoning are justifiable. We can also refer to the process or method itself as the 'method of reflective equilibrium.' After briefly describing the method, we shall describe its history, the controversy surrounding it, and some implications it has for work in ethics.

1. The Method of Reflective Equilibrium

The method of reflective equilibrium consists in working back and forth among our considered judgments (some say our 'intuitions') about particular instances or cases, the principles or rules that we believe govern them, and the theoretical considerations that we believe bear on accepting these considered judgments, principles, or rules, revising any of these elements wherever necessary in order to achieve an acceptable coherence among them. The method succeeds and we achieve reflective equilibrium when we arrive at an acceptable coherence among these beliefs. An acceptable coherence requires that our beliefs not only be consistent with each other (a weak requirement), but that some of these beliefs provide support or provide a best explanation for others. We arrive at an optimal equilibrium when the component judgments, principles, and theories are ones we are uninclined to revise any further.

The method of reflective equilibrium has been advocated as a coherence account of justification in several areas of inquiry, including inductive and deductive logic as well as ethics and political philosophy. The key idea underlying this view of justification is that we 'test' various parts of our system of beliefs against the other beliefs we hold, looking for ways in which some of these beliefs support others, seeking coherence among the widest set of beliefs, and revising and refining them at all levels when challenges to some arise from others. For example, a moral principle or moral judgment about a particular case (or, alternatively, a rule of inductive or deductive inference or a particular inference) would be justified if it cohered with the rest of our beliefs about right action (or correct inferences) on due reflection and after appropriate revisions throughout our system of beliefs. By extension of this account, a person who holds a principle or judgment in reflective equilibrium with other relevant beliefs can be said to be justified in believing that principle or judgment.

Because there is revisability of our beliefs at all levels, this coherence view contrasts sharply with foundationalist approaches to justification. A foundationalist approach would take some of these beliefs—some theoretical claim, or moral principles or judgments—as fixed. As we shall see, however, the point of

controversy, especially in ethics, is not that reflective equilibrium allows for the revision of moral judgments, but rather that it involves giving some justificatory weight to them.

2. Recent History

This approach to the justification of rules of inductive logic was proposed by Nelson Goodman in his classic *Fact, Fiction, and Forecast* (1955). Goodman's idea was that we justify rules of inference in inductive or deductive logic by bringing them into reflective equilibrium with what we judge to be acceptable inferences in a broad range of particular cases. No rule of inference would be acceptable as a logical principle if it was not compatible with what we take to be acceptable instances of inferential reasoning. In this sense, our beliefs about acceptable rules of inference are constrained by the 'evidence' provided by what we believe to be good or correct examples or instances of inferential reasoning. At the same time, we should correct or revise our views about particular inferences we initially might think are acceptable if we come to see them as incompatible with rules that we generally accept and refuse to reject because they, in turn, best account for a broad range of other acceptable inferences.

Despite the fact that the origins of reflective equilibrium lie in mid-twentieth century discussions about the justification of inductive logic, its principal development through the rest of the century lies primarily in ethics and political philosophy. Specifically, the method was given prominence by John Rawls's description and use of it in *A Theory of Justice* (1971). Most of our discussion will concentrate on its use in ethics and political philosophy. (Although accounts of the justification of empirical knowledge were developed that share with reflective equilibrium its coherentist approach, they generally do not make explicit use of the terminology of reflective equilibrium.)

Rawls argues that the goal of a theory of justice is to establish the terms of fair cooperation that should govern free and equal moral agents. On this view, the appropriate perspective from which to choose among competing conceptions or principles of justice is a hypothetical social contract or choice situation in which contractors are constrained in their knowledge, motivations, and tasks in specific ways. Because this choice situation is fair to all participants, Rawls calls the conception of justice that emerges from this choice 'justice as fairness.' Under these constraints, he argues, rational contractors would choose principles guaranteeing equal basic liberties and equality of opportunity, and a principle that permitted inequalities only if they made the worst-off groups in society as well off as possible.

Instead of simply accepting whatever principles contractors would choose under these constraints on choice, however, Rawls imposed a further condition of

adequacy on them. The chosen principles must also match our considered judgments about justice in reflective equilibrium. If they do not, then we are to revise the constraints on choice in the contract situation until we arrive at a contract that yields principles that are in reflective equilibrium with our considered judgments about justice. In effect, the device of the contract must itself be in reflective equilibrium with the rest of our beliefs about justice. The contract helps us determine what principles we should choose from among competing views, but the justification for using it and designing it so that it serves that purpose must itself derive from the reflective equilibrium that it helps us achieve.

The method of reflective equilibrium thus plays a role in both the construction and justification of Rawls's theory of justice. Critics of Rawls's theory and his method of reflective equilibrium, especially utilitarians, challenge the prominence the method gives to moral judgments or intuitions. Indeed, some argue that there is no justificatory force to the contract itself if we 'rig' the contract so that it yields principles that match our intuitions (Hare 1973). Despite various criticisms, defenders of the method have elaborated it and extended its use in broad areas of ethics. Later in this entry we shall consider some of the criticisms of this method and some of its extensions in more detail. Before doing so, however, it is necessary to describe it more fully.

3. Narrow and Wide Reflective Equilibrium

3.1 Narrow Reflective Equilibrium

A reflective equilibrium may be *narrow* or *wide* (Rawls 1974). All of us are familiar with a process in moral deliberation in which we work back and forth between a judgment we are inclined to make about right action in a particular case, and the reasons or principles we offer for that judgment. Often we consider variations on the particular case, 'testing' the principle against them, and then refining and specifying it to accommodate our judgments about these variations. We might also revise what we say about certain cases if our initial views do not fit with the principles we grow inclined to accept.

Such a revision may constitute a moral surprise or discovery (Daniels 1996). Suppose, for example, that we are considering whether we should ignore age in the distribution of medical treatments. Many initially believe that age is a 'morally irrelevant trait,' just like race, and they would insist that rationing medical services by age is just as unacceptable as rationing by race. On considering a variety of cases, however, it might become apparent that we all age, but that we do not change race, and that this means that the different treatment of people at different ages, if systematically applied over the life span, does not create inequalities

between persons, as it would in the case of race. We might be led by this realization to think that age rationing might be acceptable under some conditions, when race rationing would never be, and this would be a moral surprise for many who changed their view.

To the extent that we focus solely on particular cases and a group of principles that apply to them, we are seeking only *narrow* reflective equilibrium. Presumably, the principles we arrive at best 'account for' the cases examined. Others, however, may arrive at different narrow reflective equilibria, containing different principles and judgments about justice. As a result, we still face an important question about justification unanswered by the method of narrow reflective equilibrium: which set of beliefs about justice should we accept?

Because narrow reflective equilibrium does not answer this question, it may seem to be a *descriptive* method appropriate to moral anthropology, not a *normative* account of justification in ethics. In fact, Rawls (1971) at one point suggested that arriving at the principles that match our moral judgments in reflective equilibrium might reveal our 'moral grammar' in a way that is analogous to uncovering the grammar that underlies our syntactic ability as native speakers of a language to make judgments about grammatical form. In support of the analogy, some contemporary theorists who systematically examine our moral intuitions, often through hypothetical as opposed to real cases, believe they are uncovering a moral structure of principles that is *a priori*.

Uncovering a syntax, however, is a descriptive and not a justificatory task. Once we can identify the grammar or rules that best account for a person's syntactic competency, we do not ask the question, 'Should that person have this grammar?' We are satisfied to have captured the grammar underlying a person's idiolect. In ethics and political philosophy, in contrast, we must answer that justificatory question, especially since there is often disagreement among people about what is right, disagreement that is not resolved simply by pursuing narrow reflective equilibrium.

3.2 Wide Reflective Equilibrium

Rawls's proposal is that we can achieve real justification in the choice of principles of justice by broadening the circle of beliefs that must cohere. In a *wide* reflective equilibrium, for example, we broaden the field of relevant moral and nonmoral beliefs to include an account of the conditions under which it would be fair for reasonable people to choose among competing principles, as well as evidence that the resulting principles constitute a feasible or stable conception of justice, that is that people could sustain their commitment to such principles.

For example, the constraints on choice in Rawls's contract situation, such as the veil of ignorance that

keeps us from knowing facts about ourselves and our specific preferences, require justification. Rawls needs to show that these constraints are 'fair' to all contractors (thus 'justice as fairness' is his label for this procedural account of how we find out what is just). To provide such justification, Rawls appeals to beliefs about the fundamental 'moral powers' of 'free and equal' agents (they can form and revise their conceptions of what is good and they have a sense of justice). He also appeals to the ideal of a well-ordered society in which principles of justice play a particular role as public principles for reconciling disputes. Rawls must even provide us with an account of 'primary social goods,' necessary if agents who do not know their own actual preferences are to decide what principles would be better for them to choose. The device of the contract is thus in reflective equilibrium with certain background theories that themselves contain moral beliefs. The contract is not simply based on uncontroversial assumptions about human rationality (although it describes a rational choice problem within the constraints it imposes). Without acceptance of this wider circle of moral beliefs, Rawls's construction lacks support.

Our beliefs about justice are justified (and, by extension, we are justified in holding them) if they cohere in such a wide reflective equilibrium. Obviously, the method of wide reflective equilibrium is here only illustrated by appeal to the detail of Rawls's use of it. If we abstract from that detail, we see that there is a complex structure and interaction of beliefs, at many levels of generality, that bear on the construction of an account of justice. We shall later return to this point when we talk about implications of this method for work in ethics.

In *A Theory of Justice*, Rawls seemed to think that all people might converge on a common or *shared* wide reflective equilibrium that included 'justice as fairness,' the conception of justice for which he argues. Wide reflective equilibrium thus played a role in the construction of the theory, helping 'us' to articulate its key features in ways that led to principles that matched 'our' considered judgments. At the same time, wide reflective equilibrium constituted an account of justification. Shared agreement on that wide equilibrium would produce a well-ordered society governed by principles guaranteeing equal basic liberties, fair equality of opportunity, and the requirement that inequalities be arranged to make the worst-off groups as well off as possible.

In his later work, *Political Liberalism* (1993), Rawls abandons the suggestion that all people might converge in the same, shared wide reflective equilibrium that contains his conception of justice. This change comes about because Rawls believes that justification must accommodate the unavoidable 'fact' that human reason, exercised under conditions of liberty, will yield a pluralism of reasonable comprehensive moral views. Because of 'reasonable pluralism,' Rawls recasts jus-

tice as fairness as a 'free standing' political conception of justice on which people with different comprehensive views may agree in an 'overlapping consensus.' The *public* justification of such a political conception involves no appeal to the philosophical or religious views that appear in the comprehensive doctrines that form this overlapping consensus. Instead, we might think of this process of working back and forth among the key shared ideas in the public, democratic culture and the articulated features of the political conception of justice as a *political* reflective equilibrium (Daniels 1996).

Although there is no convergence on a shared wide reflective equilibrium that contains the political conception of justice, wide reflective equilibrium still plays a critical role in justification. For individuals to be *fully justified* in adopting the political conception of justice, the conception articulated in the political reflective equilibrium, they must incorporate it within a wide reflective equilibrium that includes their own comprehensive moral or religious doctrine. It will count as a reasonable view for them only if it coheres with their other (religious and philosophical) beliefs in reflective equilibrium. Over time, people modify their comprehensive views so that they may accommodate the overlapping consensus, a process that involves both philosophical reflection and the moderating influence of living under institutions that are governed by the shared conception of justice.

4. Criticisms of Reflective Equilibrium

4.1 Challenges to Moral Intuitions

Central to the method of reflective equilibrium in ethics and political philosophy is the claim that our considered moral judgments about particular cases carry weight, if only initial weight, in seeking justification. This claim is controversial. Some of the most vigorous criticism of it has come from utilitarians, and it is instructive to see why.

A traditional criticism of utilitarianism is that it leads us to moral judgments about what is right which conflict with our 'ordinary' moral judgments. In response, some utilitarians accept the relevance of some of these judgments and argue that utilitarianism is compatible with them. Thus Mill argued for a utilitarian foundation for our beliefs about the importance of individual liberty. Some utilitarians have even argued that key features of our 'common sense morality' approximate utilitarian requirements, and that we have acquired these beliefs just because they do, unconsciously, reflect what promotes utility—they reflect the wisdom of a heritage.

An alternative utilitarian response to the claim that utilitarianism conflicts with certain ordinary moral judgments is to dismiss these judgments as pretheoretical 'intuitions.' As such, they are likely to result from cultural indoctrination and thus reflect superstition,

bias, and mere historical accident. On this view, moral intuitions or judgments should have no evidentiary credentials and should play no role in moral theory construction or justification. Indeed, as the prominent utilitarians Richard Brandt and Richard Hare argued against Rawls, simply making 'coherent' a set of beliefs that have no 'initial credibility' cannot produce justification, since coherent fictions are still only fictions. Indeed, when Rawls describes the conditions under which we might solicit considered moral judgments, namely that people be calm and have adequate information about the cases, they do not by themselves do anything to assuage the utilitarian worries.

This criticism has some force, since two standard ways of supplying credentials for initial judgments are not available. One traditional way to support the reliability of these judgments or intuitions is to claim, as eighteenth-century theorists did, that they are the result of a special moral faculty that allows us to grasp particular moral facts or universal principles. Modern proponents of reflective equilibrium reject such mysterious faculties. Indeed, they claim moral judgments are revisable, not foundational.

A second way to support the initial credibility of considered judgments is to draw an analogy between them and observations in science or everyday life. For example, what counts as observational evidence in science depends on theory, and theory may give us reasons to reject some observations as not constituting counterevidence to a scientific law or theory. In this way we might see an analogy between the revisability of moral judgments and observations.

Developing this analogy, however, seems to require that we also tell some story about why moral judgments are reliable 'observations' about what is right. Perhaps we might need something like the causal story that some theorists of knowledge offer to explain the reliability of observations. Since no such story is forthcoming, opponents argue, proponents of reflective equilibrium must reject the requirement or give up the analogy. Proponents of reflective equilibrium might reject the requirement by suggesting it is premature to ask for such a story in ethics or by claiming that we can provide no analogous causal story for credible judgments we make in other areas, including mathematics or logic.

It might seem that the burden of argument has shifted to advocates of reflective equilibrium to show why 'initial credibility' should be ascribed to moral judgments or intuitions. Defenders of reflective equilibrium may nevertheless reject this burden, arguing that critics, especially utilitarian critics, actually face the same problem. For example, Richard Brandt argues that 'facts and logic' alone, and not moral intuitions, should play a role in moral theory construction and justification. On Brandt's view, we should choose moral principles when they are based on desires that have been subjected to maximal criticism by facts and logic alone (he calls it 'cognitive

psychotherapy'). We should avoid any appeal to moral intuitions that might infect this critique.

The desires that Brandt appeals to, however, are themselves shaped and influenced by the very same social structures that utilitarians complain have biased and corrupted our moral judgments. Nothing in the process of critique by facts and logic alone can eliminate this source of bias. If this claim is right, it undercuts the suggestion that we can step outside our beliefs to arrive at some more objective form of justification. Instead, we may be better off recognizing that our process of critique—in the method of wide reflective equilibrium—is explicit about exposing the sources of bias and historical accident to criticism and revision, rather than fooling ourselves into thinking that desires are some sort of morally uninfluenced layer of 'facts.'

4.2 Other Criticisms of Reflective Equilibrium

Other criticisms of reflective equilibrium in ethics have focused on points familiar from discussions in epistemology more generally. One important complaint concerns the vagueness of the concept of coherence. If we simply take logical consistency as the criterion for coherence, we have much too weak a constraint. What account can we give of a stronger notion?

More has to be said about how some parts of the system of beliefs 'support or explain' others than is provided in the account above. Critics of Rawls's requirement that the contract situation be adjusted, if necessary, to yield principles that are in reflective equilibrium with our judgments about justice complained that this was a 'rigged' contract and that it did no justificatory work beyond reflective equilibrium. If, however, the moral judgments that play a role in discussions of fair process, the well-ordered society, and the moral powers of agents are somewhat independent of the considered judgments about justice, then we get the kind of independent support for the principles that add justificatory force. Little work has actually been done, however, to flesh out a stronger account of coherence in ethics as opposed to epistemology more generally, where coherentism has been defended and elaborated.

A second line of objection derives from antioherentist accounts of justification rooted in the theory of knowledge. Some insist, for example, that coherence accounts of justification cannot be divorced from coherence accounts of truth. On the other hand, Rawls's view fits with the claim by some contemporary theorists of knowledge that a coherence account of justification is distinguishable from a coherence account of truth, and defensible when so separated.

This separability is important to Rawls, even more so in his later work than in *A Theory of Justice*. Rawls never took a stand on the question of whether there are 'truths' about justice that correspond to a conception of justice that we justify through reflective

equilibrium. In his early work, the method of reflective equilibrium suggested ways in which convergence might be achieved among those who began with disagreements about justice. By drawing attention to the many features of a comprehensive theory on which argument and evidence could be brought to bear, the method of wide reflective equilibrium held out promise that more convergence might result than if people had only considered judgments and principles to agree and disagree about. Convergence did not imply that truth was reached, but under some conditions, it could be taken as evidence for truth, just as it might be in non-normative areas of inquiry.

Once Rawls took the 'burdens of judgment' and reasonable pluralism seriously, that is once he 'politicized' justice in his later work, it became much less plausible to talk about convergence on a shared wide equilibrium that might contain moral truth. Instead, even if we converged in an 'overlapping consensus' on a conception of justice, it would have to be justified from the perspective of distinct wide reflective equilibria. It became more important for Rawls to suggest how claims about justice could be 'objective' without presupposing moral truth. These developments in Rawls's account thus appear to move him farther away from those who would seek to give a realist account of moral truth.

One further group of criticisms of reflective equilibrium derives largely from claims about human rationality. A starting point for this line of argument is the complaint that people who have different starting points as their initial set of beliefs, say with different degrees of credence given to their beliefs, may arrive at different reflective points of equilibrium. This possibility undercuts the suggestion that critical pressures alone can produce convergence. A more general form of this worry is that the model of reflection involved in reflective equilibrium, in which we are to produce coherence among all our beliefs, overemphasizes and idealizes human rationality. The suggestion is that we would be better off presupposing a more minimal form of rationality. A specific version of this complaint is that the method involves an information burden that cannot be met. Yet another version of this criticism is that we should allow for much less 'rationalist' forms of modification of our views, recognizing that 'conversion' experiences such as those involved in 'paradigm shifts' may affect any account of coherence.

A full defense of reflective equilibrium as a method would require a more developed response to many of these lines of criticism than exists in the literature.

5. Applications and Implications of Reflective Equilibrium

Despite these criticisms, some philosophers have argued for a broader understanding of the relevance of the method of reflective equilibrium to practical ethics.

In thinking about the course of right action in a particular case, we often appeal to reasons and principles that are notoriously general and lack the kind of specificity that make them suitable to govern the case at hand without committing us to implications we cannot accept in other contexts. This requires that we refine or specify the reasons and principles if we are to provide appropriate justifications for what we do and appropriate guidance for related cases. Philosophers who have focused attention on the importance of specification have drawn on the method of reflective equilibrium for their insights into the problem.

In practical ethics, especially bioethics, there has been a vigorous debate about methodology. Some argue that we must root all claims about specific cases in specific ethical theories, and that the hard work is showing how these theories apply to specific cases. Others argue that we may disagree about many aspects of general theory but still agree on principles, and that the hard work of practical ethics consists in fitting sometimes conflicting principles to particular cases. Still others argue that we must begin our philosophical work with detailed understanding of the texture and specificity of a case, avoiding the temptation to intrude general principles or theories into the analysis.

A grasp of the method of wide reflective equilibrium suggests a way around the exclusionary nature of this debate. Wide reflective equilibrium shows us the complex structure of justification in ethics and political philosophy, revealing many connections among our component beliefs. At the same time, there are many different types of ethical analysis and normative inquiry.

This suggests a more eclectic view of the debate about method. Work in ethics requires all levels of inquiry proposed by the disputants, not always all at once or in each case, but at some time or other. Sometimes it is true that we cannot resolve disputes about how to weigh conflicts among principles unless we bring more theoretical considerations to bear (these considerations need not involve comprehensive ethical theories). Sometimes we can agree on relevant principles and agree to disagree on other theoretical issues, still arriving at agreement about the rightness of particular policies or actions and their justification in light of relevant reasons and principles. Sometimes we must see what is distinctive about particular cases and revise or refine our reasons and principles before we can arrive at understanding of what to do. Finally, and of the greatest significance, the method of wide reflective equilibrium should make it clear that work in ethical theory cannot be divorced from work in practical ethics. We must test and revise theory in light of our considered judgments about moral practice. The condescending attitude of many who work in ethical theory toward work in practical ethics thus is incompatible with what wide reflective equilibrium establishes about the relationship between these areas of ethical inquiry.

The method of wide reflective equilibrium makes it plausible that all such 'methods' should be seen as appropriate to some tasks in ethics and are but parts of a more encompassing method.

Bibliography

- Daniels N 1979 Wide reflective equilibrium and theory acceptance in ethics. *Journal of Philosophy* 76: 256–82
- Daniels N 1996 *Justice and Justification: Reflective Equilibrium in Theory and Practice*. Cambridge University Press, New York
- DePaul M R 1993 *Balance and Refinement: Beyond Coherence Methods of Moral Inquiry*. Routledge, New York
- Goodman N 1955 *Fact, Fiction, and Forecast*. Harvard University Press, Cambridge, MA
- Hare R M 1973 Rawls' theory of justice. *Philosophical Quarterly* 23: 144–55, 241–51
- Rawls J 1971 *A Theory of Justice*. Harvard University Press, Cambridge, MA
- Rawls J 1974 The independence of moral theory. *Proceedings and Addresses of the American Philosophical Association* 47: 5–22
- Rawls J 1993 *Political Liberalism*. Columbia University Press, New York
- Sencerz S 1986 Moral intuitions and justification in ethics. *Philosophical Studies* 50: 77–95
- Stich S P 1990 *The Fragmentation of Reason*. MIT Press, Cambridge, MA
- Van der Burg W, Van Willigenburg T (eds.) 1998 *Reflective Equilibrium: Essays in Honour of Robert Heeger*. Kluwer Academic Publishers, Dordrecht, The Netherlands

N. Daniels

Reflexivity in Anthropology

Critiques of the social sciences as practiced in the United States emerged in the 1980s in which reflexivity became a keyword, standing for both the means of the critiques as well as the form that alternative practices might take. The consequences for disciplinary practices of reflexivity as means of critique and alternative form of inquiry were nowhere as great among the social sciences as in social and cultural anthropology.

According to one of the early statements in this critical trend, reflexivity is '... the capacity of any system of signification to turn back upon itself, to make itself its own object by referring to itself: subject and object fuse' (Myerhoff and Ruby 1982). Myerhoff and Ruby go on to stress that 'significant distinctions exist between reflexivity and related attitudes such as self-regard, self-absorption, solipsism, self-reference, self-consciousness, and autobiography. Reflexivity does not leave the subject lost in its own concerns; it pulls one toward the other and away from isolated attentiveness toward oneself.'

Reflexivity as a practice in the social sciences is, thus, both a distinctive mode of thinking and a strategy

of writing that concerns itself with the process of inquiry as an integral part of the claims to knowledge that the inquiry might produce. As a concept, it is related to the idea of feedback in cybernetic models, of metacommunication, and of such functions as deixis and indexicality inherent in the faculty of language use. However, it is the more ideological uses of reflexivity in debates about the revision of anthropology's distinctive methods of fieldwork and ethnography that are of concern here. Debates about the role and purpose of reflexivity in the process of research as well as in the changing form of the ethnographic text became the major specifiable way of assessing the impact upon anthropology of the broader intellectual movement of the 1970s and 1980s known loosely as postmodernism.

In the broad sweep of European philosophy, reflexivity as a category of interest in social theory begins with Kant's distinction between analysis and an analytic in which an analytic is the kind of inquiry where the subject of inquiry is also the object of inquiry. One genealogical stream from this originary idea of reflexive inquiry leads through Hegel to phenomenology and eventually to Heidegger and existentialism in France. The other, and for us, more relevant genealogical stream leads to hermeneutics in the work of Hans-Georg Gadamer and Paul Ricoeur by way of the neo-Kantian writings of Dilthey and the work of Max Weber, in which reflexivity, while implicit, is a strong element of his methodological considerations. In the work of Gadamer and Ricoeur, reflexivity becomes an explicit and crucial intellectual operation in the activity of interpretation whereby the individual ethical consciousness expands reflectively through encounters with the foreign or the different. Thus, while long a potential existing in a sort of minor key to the predominant positivist tendency of social theory which oversaw the post-World War II expansion of sociology and anthropology in the United States, philosophical works of hermeneutics, in which the operation of reflexivity was highlighted, only developed explicitly a strong appeal among sociologists and anthropologists from the early 1980s onward.

Of all of the social sciences, the importance of reflexivity as a style of inquiry has been most natural to precisely those methods of anthropology which have distinguished it as a discipline. Pioneered by Bronislaw Malinowski and Franz Boas, the pursuit of fieldwork investigation through long-term participant observation, working directly in native languages, and the constant interpretation and translation of behaviors and speech encountered on a daily basis, inherently involves on the part of the investigator a keen consciousness of oneself and one's relations with specific others as the medium of all claims to knowledge from this research practice.

For a long time, anthropologists only came to terms with the complexities and problems of this distinctive form of knowledge production informally as a pastime

of professional culture and folklore, as something to be communicated between mentors and students. It was only at the end of the 1960s, and then again, and more definitively in the 1980s, that anthropologists examined and developed explicitly the varieties of meanings and implications of reflexivity as a major style of knowledge production for the future practice of their discipline.

Since the 1980s, reflexivity has become a 'keyword' indicating both a self-conscious, mainstream style of work in anthropology and part of its own enduring disciplinary politics of knowledge. Here, it is important to note that reflexivity has only served as such a central category, standing in turn for the most concrete aspect of the controversial impact of postmodernism on anthropological practice, in the recent evolution of the discipline in the United States. It has been discussed and debated from the 1980s on in other anthropological establishments around the world, but mainly through the stimulus of the prestige and diffusion of trends in United States anthropology.

1. Reflexivity in Two Moments in the Critique of Anthropology

During the late 1960s, US anthropologists became preoccupied openly with a long simmering critique of their discipline. This critique had theoretical and methodological dimensions, but it was primarily political in nature, stimulated by the ferment of domestic resistance to the Vietnam War, renewed discussions of anthropology's historic development within the expansion of European colonialisms, and the interest in new work in Marxist theory. At that time as well, Geertz was developing a broad and deep theoretical foundation for cultural anthropology in hermeneutic philosophy, work that was eventually to serve as both the primary foil and influence for later discussions of reflexivity in anthropological method. (See Geertz 1968 for his early position on reflexivity in fieldwork.) But in the *de facto* reflexive attempt to undertake a critical anthropology of anthropology in the politicized atmosphere of the 1960s, the issue of reflexivity itself as the major dynamic of what it was to practice critical analysis was little addressed. An important exception is the paper by Scholte (1969) which summarized this period of critique. Important as a figure who introduced currents of European social theory into US anthropology, Scholte foreshadowed what sort of theoretical work and changes in practice would be necessary to transform anthropology into a discipline with a sustained critical purpose beyond that moment of heightened political critique.

In the interval between the two periods of internal disciplinary critique, perhaps the most important work that served as the raw material on which the eventual surfacing of explicit considerations of reflexivity in

anthropology depended was the burgeoning of a literature on fieldwork experience, written alongside the production of more formal ethnographies from which such reflections about the process of inquiry were largely absent. This literature, fueled by the publication of Malinowski's diaries in 1967, incorporated the spirit of self-critique of the late 1960s and was produced by anthropologists who were then students and young professors. There emerged a more theoretical strain of this literature in which the reflexive character of fieldwork became the major preoccupation (e.g., Rabinow 1979) and that led eventually to experiments with the form of the ethnography itself (e.g., Crapanzano 1980). These works, combined with an interdisciplinary movement, variously characterized as postmodernism and poststructuralism and led by scholars of literature who were experiencing their own disciplinary crisis, stimulated the theoretically deeper, more wideranging, and sustained internal critique of anthropology of the 1980s.

Incorporating all of the dimensions of the earlier critique, the 1980s critique primarily examined the central practices of the discipline, especially in its character as a hermeneutic enterprise, a view for which Geertz had laid the foundation. In particular, it did so by a focus on the rhetoric and forms of the signature textual product of anthropology, the ethnography. (See Clifford and Marcus 1986, the volume that has stood for this critique in the way that the Hymes volume did for the earlier one.) This critique was, thus, the efflorescence of a comprehensive discussion of all of the issues relating to reflexivity that had long been so essential to the practice of anthropology. Influenced by the circulation of much more diverse and sophisticated theoretical resources, anthropologists looked with an unprecedented intensity and thoroughness at the relationships in their work between reflexive modes of knowing, critical and ethical awareness, the political contexts of inquiry, and alternative possibilities for practice.

The basic act of reflexivity which initiated this critique were the probing discussions of the relationships of anthropologists to informants—of the self to other—and the dependence of the production of anthropological knowledge on these relationships which had so often been effaced in the ways that ethnographies had in fact been written. Evoking alternative modes of representing others, questioning the act of representation itself, suggesting the ways ethnographic knowledge emerges from collaboration and dialogue, and probing the political conditions which have inhibited such possibilities, this moment of critique served both to stimulate degrees of experiment with the ethnographic form and to encourage ramifying discussions about the critical purpose of anthropological analysis (e.g., Marcus and Fischer 1999). This undertaking was itself an umbrella act of reflexivity, and reflexive practice became the common frame for thinking about past ethnographies, future

possibilities, and the form and purpose of anthropological knowledge itself. This field of discussion was highly contested: some felt that the critique by its nature was too hermetic, too literary, and not sufficiently committed to anthropology's study of social struggles in the contemporary world; some thought the attention to reflexive issues was an aesthetic distraction and a major blow to the production of objective knowledge that could be shared by a scholarly community and on which disciplinary progress depended; and contrastingly, for some, the kinds and levels of reflexivity encouraged by the critique were neither reflexive nor critical enough, leading to a spiralling array of discussions about insufficiencies and exclusions that had even more ideal reflexive practice in view.

The immediate effect of this period of critique were the changes in the ways that ethnographies were both written and discussed. But while the critique was developed through attention to texts and the way that anthropologists rhetorically established authority for them, the larger stake was the effort finally to justify them theoretically and ethically and expand anthropology as a critically reflexive field of inquiry. There can be little doubt that this much has been established as at least a long-term mainstream tendency in the discipline. This tendency has placed anthropology in much closer relation to such parallel interdisciplinary fields, themselves rooted in varieties of reflexive inquiry, as media studies, feminist studies, postcolonial studies, and most broadly, cultural studies, than to its historic partners in the other social sciences.

2. *The Incorporation of Reflexivity in the Mainstream of Anthropological Research*

Discussions about reflexivity, or at least the evaluation of ethnographic claims to knowledge in terms of the specific conditions and relations of inquiry that produced them from fieldwork, are not only highly visible and commonplace in anthropological discourse of the 1990s, they are *de rigueur*. Further, some form and degree of reflexive commentary is worked prominently into virtually every ethnography written these days. If not enacted in experimental text design, there is the presumption that all ethnography is produced collaboratively, and there are far richer and more probing discussions than ever before about the kinds of contexts which influence ethnographic representations of subjects and shape arguments from ethnographic data (see Marcus 1994). Concern for reflexivity has had a special appeal for ethnographic filmmakers, and some of the most cogent writing on the importance of reflexivity in the representation of other forms of life have come from this arena (e.g., MacDougall 1998). Current productions of ethnographic film and media demonstrate most clearly that the practice of reflex-

ivity crucially redefines and adapts the classic conditions of ethnographic inquiry to new conditions of fieldwork where anthropological research projects must negotiate different identities and roles in relation to the highly articulated projects of their subjects in the overlapping contexts of a globally interconnected world.

2.1 Variant Ideologies in the Practice and Form of Reflexivity

In sociology, discussion of reflexivity has tended to be integrated formally as a methodological tool or technique (e.g., see Bourdieu 1990), limiting its ability to affect fundamentally the forms and purposes of sociological research. In anthropology, the outcome has been quite different, generating a number of changes in the ways that ethnographies are written and fieldwork is conceived. This in turn has obviously had considerable implications for changes in both the form and content of the knowledge that anthropology's core research process generates.

In engagement with the terms of the 'Writing Culture' critique of the 1980s, the largest body of literature stimulated explicitly by reflexive inquiry examines in great detail the self-other relation in classic conditions and specific sites of fieldwork. While some have characterized this literature as narcissistic or merely self quests, in its most accomplished examples (e.g., Behar and Gordon 1995, Reed-Danahay 1997, Taylor 1999, Tsing 1993), it has addressed cogently issues of contemporary changes in concepts of self, personhood, and identity among many of the peoples who had been the subject of study by previous generations of ethnographers. At the microlevel of fieldwork, it has also clarified aspects of the history of anthropology as an academic enterprise and in relation to the broad range of effects that inquiry itself has on the peoples and subjects studied (e.g., Rosaldo 1989). While somewhat unruly and less governed by genre rules of form in social science writing, this literature has, in sum, been very useful in deepening and critiquing the complexities of constructs of subjectivity and otherness within their conventional usages in Western intellectual discourse—a long-term, continuing objective of anthropology as a discipline. It has also probed critically ethnographers' positions in relation to initial objects of study so as to transform the objects in novel ways and reconfigure the compass of research in response to radical contemporary changes in both the worlds of the academic and who is identified as subjects of study.

There has also been a strain of reflexive inquiry in anthropology for which reflexivity has been more distinctly a means by which to define an object of study or to circumscribe new temporal and spatial boundaries for a fieldwork project (Marcus 1998). Key to these works is clarity of the positioning or

location of ethnographic inquiry, which reflexivity enables, in relation to other historic and geographical agents in the same social field. Reflexivity as positioning not only serves the creation of new forms of objectivity in anthropology, sensitive to the issues of critique that discussions of reflexivity have raised, but redefines both the scope of ethnography and its arguments. In these works the limits and reach of once very locally focussed ethnographic arguments expand in unpredictable and intriguing ways (for a now classic experiment of this sort, see Taussig 1987; for more recent examples, see Martin 1994, Fortun 2000).

Finally, there is the presence of reflexivity in ethnography because of the evolution of certain realities on the ground over the course of career-long scholarly involvements with certain peoples and places. In recent decades, literally self-conscious cultural awareness has increased markedly and even been mobilized politically in many locales and regions such that the conditions of ethnographic research require a rethinking and reorientation toward one's subjects, now organized into social movements, nongovernmental organizations, and other groups involved in objectifying and representing themselves. Here, reflexive inquiry is the means to a complex and necessary politics of research which not only shapes fieldwork but makes its continuation possible at all (see Warren 1999 for an exemplar of this type of increasing variant of reflexive inquiry).

2.2 Reflexive Modernization and Reflexivity in the Changing Conditions of Anthropological Fieldwork

Broader changes in contemporary social conditions worldwide that stimulate the trend of reflexive inquiry in anthropology is encompassed in the concept of reflexive modernization (Beck et al. 1994), a development in social theory of the 1990s that responds to the postulation of a present era of postmodernity by trying to specify the characteristics of such an era for the pursuit of empirical inquiry. This construct proposes that the modernity characteristic of industrial society is succeeded by a modernity that involves living with irreducible contingency, living in a more complex and less controllable world. Most importantly, reflexivity as a self-monitoring process on the level of both institutions and individual persons becomes not only a pervasive aspect of society, but also a key focus of social thought itself.

This version of social theory suggests for the proliferation of styles of reflexive inquiry in anthropology that these are not driven only as a result of the internal critique of anthropology, as a seminal countertrend to a preceding era of positivist tendencies, but as a more generally shared condition with the societies and cultures which anthropology makes its subjects of study. This suggestion in turn has a potentially profound and enduring effect on the research proto-

cols of anthropology in recognizing certain affinities and identifications between the reflexive predicaments of ethnographers and those whom they study. Reflexivity is relevant not because it provides truer or more powerful explanatory models with which anthropologists can treat social facts among their subjects, but rather because reflexivity is also practiced differentially 'out there' on a widespread basis (see Marcus 1999). This fact suggests the breaking of the frame of conventional scholarly work altogether, at least in the anthropological tradition, rather than the demise and replacement of specific concepts and characteristics of this academic frame. Thus, reflexivity is no longer only a moment in the critique of disciplines but becomes more generally the object and medium of social and cultural analysis as well.

See also: Cultural Critique: Anthropological; Ethnography; Feminist Theory: Postmodern; Hermeneutics, History of; Hermeneutics, Including Critical Theory; Interpretation in Anthropology; Postmodernism: Methodology

Bibliography

- Beck U, Giddens A, Lash S 1994 *Reflexive Modernization: Politics, Tradition and Aesthetics in the Modern Social Order*. Stanford University Press, Stanford, CA
- Behar R, Gordon D A (eds.) 1995 *Women Writing Culture*. University of California Press, Berkeley, CA
- Bourdieu P 1990 *The Logic of Practice*. Stanford University Press, Stanford, CA
- Clifford J, Marcus G E 1986 *Writing Culture: The Poetics and Politics of Ethnography*. University of California Press, Berkeley, CA
- Crapanzano V 1980 *Tuhami: Portrait of a Moroccan*. University of Chicago Press, Chicago
- Fortun K 2000 *Advocacy After Bhopal: Environmentalism, Disaster, New Global Orders*. University of Chicago Press, Chicago
- Geertz C 1968 Thinking as a moral act: Ethical dimensions of anthropological fieldwork in the new states. *Antioch Review* 28: 139–58
- Hymes D (ed.) 1972 *Reinventing Anthropology* 1st edn Pantheon Books, New York
- MacDougall D 1998 *Transcultural Cinema*. Princeton University Press, Princeton, NJ
- Marcus G E 1994 On ideologies of reflexivity in contemporary efforts to remake the human sciences. *Poetics Today* 15: 383–404
- Marcus G E 1998 *Ethnography Through Thick and Thin*. Princeton University Press, Princeton, NJ
- Marcus G E (ed.) 1999 *Critical Anthropology Now: Unexpected Contexts, Shifting Constituencies, Changing Agendas*. School of American Research Press, Santa Fé, NM
- Marcus G E, Fischer M J 1999 [1986] *Anthropology as Cultural Critique: An Experimental moment in the Human Sciences*. University of Chicago Press, Chicago
- Martin E 1994 *Flexible Bodies: Tracking Immunity in American Culture from the Days of Polio to the Age of AIDS*. Beacon Press, Boston

- Myerhoff B, Ruby J 1982 Introduction. In: Ruby J (ed.) *A Crack in the Mirror: Reflexive Perspectives in Anthropology*. University of Pennsylvania Press, Philadelphia
- Rabinow P 1977 *Reflections on Fieldwork in Morocco*. University of California Press, Berkeley, CA
- Reed-Danahay D E (ed.) 1997 *Auto/Ethnography: Rewriting of the Self and the Social*. Berg, Oxford, UK
- Rosaldo R 1989 Imperialist nostalgia. In: *Culture & Truth: The Remaking of Social Analysis*. Beacon Press, Boston, pp. 68–87
- Ruby J (ed.) 1982 *A Crack in the Mirror: Reflexive Perspectives in Anthropology*. University of Pennsylvania Press, Philadelphia
- Scholte B 1969 Toward a reflexive and critical anthropology. In: Hymes D (ed.) *Reinventing Anthropology*. Random House, New York, pp. 430–57
- Tausig M 1987 *Shamanism, Colonialism and the Wild Man: A Study in Terror and Healing*. University of Chicago Press, Chicago
- Taylor J 1999 *Paper Tanguos*. Duke University Press, Durham, NC
- Tsing A L 1993 *In the Realm of the Diamond Queen*. Princeton University Press, Princeton, NJ
- Warren K B 1999 *Indigenous Movements and Their Critics: Pan-Maya Activism in Guatemala*. Princeton University Press, Princeton, NJ

G. E. Marcus

Reflexivity, in Science and Technology Studies

Considerations of reflexivity have salience in diverse intellectual practices and traditions: literature, philosophical logic, the visual arts, cybernetics, as well as social theory and sociology. Indeed, it is an issue with potential relevance for all modes of representational and performative practice. In the most general terms, considered as a feature or style of inquiry, reflexivity is a concept that expresses a concern to turn 'subject' into 'object,' to turn the activity of inquiry into a topic for (its own) inquiry. To 'be reflexive' is to bring the process of doing an activity into the purview of that activity as a feature of it. Thus, it implies a certain self-consciousness, and a certain paradoxical, even Zen-like, mode of operation: bootstrapping; attempting to see the flashlight that illuminates the scene as a part of the scene it illuminates; taking account of this moment as one writes of 'this moment.' This article details the ways in which reflexivity has featured in science and technology studies (STS), and speculates on the current and future status of reflexive study in STS.

1. Alternative Introduction

This heading is not to declare, exactly, that there is anything 'wrong' with that first introductory paragraph. It does the job, and it does it correctly; that is,

in accord with the 'Instructions for Manuscript Preparation' for this volume. It is just that in doing 'just that,' it is unable to acknowledge the conditions under which it does so. Thus, it leaves all those things out of (its) account; which is normal, and normally unobjectionable. Perhaps no other article in this *Encyclopedia* makes any reference to such things, preferring instead, perfectly 'obviously,' to concentrate on the primary task of speaking authoritatively (and, importantly, briefly) about its topic. Thus, the conditions of production of these particular texts are constructed as irrelevancies for, and obstacles to, the proper work. Social science inquiry that focuses on conditions of production has always acted (intentionally or not) to 'make trouble' of this kind: Marxist economics for capitalism; sociology of education for teaching; STS for science. Reflexivity adds a twist to this dynamic: in collapsing the external/internal, outsider/insider, on/in relation, the source of trouble, being as much 'here' as it is 'there,' is no longer clearly located or easily attacked. This, then, is the primary role for reflexive inquiry: to make things difficult; to enrich the process of inquiry; to go round and round in circles; to fill out the picture; to frustrate all action; to act authentically and without self-contradiction; to change the subject.

2. Reflexivity as a Problem

In STS, the impetus for explicitly reflexive studies has been felt most strongly in the important sub-field of STS known as the sociology of scientific knowledge (SSK) (see *Sociology, Epistemology of; Technology, Social Construction of*). Here, the reflexive tie between its subject (knowledge about scientific knowledge) and its object (scientific knowledge) is very marked. That is, SSK exhibits a strong 'X of X' structure. Because of this, and because SSK's epistemology is explicitly relativist, reflexivity in SSK has a 'bite' akin to classic philosophical skepticism.

Extremely crudely, and for illustrative purposes only, SSK's relativism with respect to scientific knowledge can be, and indeed has been, taken as equivalent to a statement of this kind: 'Scientific knowledge does not constitute truth because it is socially constructed.' Read this way, SSK then becomes vulnerable to the standard anti-relativist critique, the argument from self-refutation: 'If this statement is asserted as a (scientific) truth, then it suffers from performative self-contradiction, and thus refutes itself' (see, for example, Hollis and Lukes 1982). Note that the self-refutation argument uses reflexivity: it is through the process of turning its target onto itself that the demonstration of self-refutation is effected.

This is indeed one of the commonest modes in which reflexivity operates: as a weapon in the agonistic game of criticism. Thus, for those threatened with attack-by-reflexivity, it becomes a source of vulnerability to

be abjured and defended against. Reflexivity becomes, that is to say, a 'problem' to be 'managed.' The various ways that practitioners of SSK have done this have been one of the preoccupations of 'reflexivist' studies (see Sect. 2.3).

2.1 Logical Consistency: Bloor's Strong Programme

David Bloor's *Knowledge and Social Imagery* (1991), is one of the earliest texts to set out the elements of what has become known as the SSK. It contains an endorsement of reflexivity as the fourth and last of the 'tenets' for the 'strong programme in the sociology of knowledge' (see *Strong Program, in Sociology of Scientific Knowledge*):

In principle [the strong programme's] patterns of explanation would have to be applicable to sociology itself. Like the requirement of symmetry, this is a response to the need to seek for general explanations. It is an obvious requirement of principle because otherwise sociology would be a standing refutation of its own theories (1991, p. 7).

Here, Bloor's positive support for reflexivity is a matter of logical consistency and a creative way of managing the self-refutation charge. Rather than deny the reflexive applicability of the strong programme, he emphasizes it, secure in his understanding that the socially constructed character of all knowledge, including his own, does not count against its validity. Bloor's reflexivity, however, remains at this 'in principle' level. His negative attitude to any empirical instantiation of his reflexive programmatics is documented in a text that claims to be precisely that (Ashmore 1989). Thus for Bloor (as well as for his former 'Edinburgh school' colleague Barry Barnes 1974, Epilogue), reflexivity is an argument used to defend a position; it is certainly not a practice.

2.2 Harry Collins: Special Relativism and the Ban on Reflexivity

Not so for, perhaps, the most influential writer through the heyday of SSK (ca. 1975–90), Harry Collins (e.g., 1985). Collins's doctrine of 'special relativism' separates the natural world from the social world, and the distinctive tasks and 'natural attitudes' of the natural scientist from those of the social scientist. The sociologist of scientific knowledge should treat the natural world as socially constructed, while treating the social world as 'real' and a source of sound data. However, this sense of 'realism' is taken from a reading of how natural scientists act towards their realm of inquiry, namely, the natural world. In effect, what Collins is doing here is to deny the practical relevance of the purported reflexive similarity between subject and object in SSK for purposes of the conduct

of research. The result is the short-circuiting of attempts to apply Collins's conclusions about, for example, the replication of experiments to those conclusions themselves (nevertheless see Mulkay 1985, Chaps. 4 and 5, Ashmore 1989, Chap. 4). In short, Collins's 'management strategy' for reflexivity is to ban it altogether. This radical solution to the problem of reflexivity has laid Collins open to charges of inconsistency from both wings of the 'realist-relativist' continuum. The recommended moves for repairing this perceived inconsistency differ, however, with the 'realist' suggesting he abandon his problematic relativism about nature, and the 'more-radical-relativist' proposing the dropping of his social realism (see *Technology, Social Construction of*).

2.3 Relativist Consistency, Radical Reflexivity: The Reflexivists

In the late 1980s, a small number of 'full-blown' reflexive texts were published (Ashmore 1989, Mulkay 1985, Woolgar 1988). This group of writers became known as 'reflexivists.' Ashmore's *The Reflexive Thesis* is an SSK study of SSK, a general exploration of reflexivity, and an experiment in 'new literary forms.' Mulkay's *The Word and the World* is a series of 'explorations in the form of sociological analysis' that draws upon analyses of biochemists' discourse. Woolgar's *Knowledge and Reflexivity* is a collection of papers from a small group of regular contributors (including Ashmore, Mulkay, and Woolgar) to the Discourse and Reflexivity Workshops held in the UK in the early 1980s. What links all these texts together, apart from their authors' close socio-cognitive relation, is the common attempt to treat reflexivity as an opportunity; even, for celebration. What happens when one stops evading reflexivity, when one looks 'the monster' in the eye, when one gives it a place at the heart of the research process? Self-destruction? Massive irrelevance? Third apocalyptic simile?

Bruno Latour's contribution to *Knowledge and Reflexivity* (Latour 1988) is its exception. Latour argues that the 'meta-reflexive' writing strategies of the reflexivists is a solution to a false problem: the naiveté of readers who believe too much and too easily. On the contrary, the real problem that (scientific, factual) writers have is being believed, or indeed, read at all. Latour further claims that the reflexivists are searching, despite their own gestures at weakening their authorial privilege, for the final meta-level that leaves no remainder and from which vantage point, everything may be taken into (their) account. As an antidote to all this, Latour recommends a strategy of 'infra-reflexivity': just to tell stories, drawing upon the richness and variety of all the 'ordinary' resources available. Latour's *Aramis* (1992, 1996)—despite its appearance as an experiment in 'new literary forms'—is, perhaps, an example.

3. Criticisms: Pitfalls and Pratfalls

Reflexivity gets you nowhere. It is not a useful thing to do and its results cannot be used. It is impractical. As an activity, it immobilizes you. It leads to idle navel-gazing and foolish self-absorption. It is 'all pots and pans and no pudding' (Bloor, cited in Ashmore 1989, p. 20). Logically, it leads to the abyss of the infinite regress or the never-ending search for the perfect meta-level. Reflexivity is politically and morally irresponsible games playing, fiddling while Rome burns. It has no soul (to bare) and no wisdom (to distribute), just an overdose of clever-clever irony. It disables criticism by taking that role on for itself; in its attempt to play all the roles of the academic community at once, and all by itself, it betrays itself as deeply, and ironically, anti-social. Indeed, it appears to seek nothing from its readers but a kind of 'aesthetic' admiration marked by a knowing, and smug, and boyish, amusement.

However, for the larger STS community, reflexivity has had positive functions and effects. The reflexivist fulfils the important role of the jester at the court of STS. Moreover, the presence of reflexive work as part of the 'canon' of STS acts as a corrective to a kind of methodological overconfidence resulting from the recent success of the field. Reflexivity reminds us that the fundamental epistemic questions and quandaries that originally animated SSK are permanent features of the practice of 'inquiry into inquiry' epitomized by STS scholarship. The text that aspires to 'transparency,' that is blithely unconcerned with the conditions of its own production, is now, at the very least, 'visible,' a member of a marked, rather than unmarked category.

4. The Current Situation

If the work of the reflexivists is taken as the model for 'adequately' reflexive work in STS, then it has to be said that this style has been taken up on very few, if any, occasions between 1990 and 2000, Latour's *Aramis* (1992, 1996) being a possible exception (see Sect. 2.3). Nevertheless, interest in reflexivity has not vanished. Explicit discussions still appear (e.g., Pels 2000, Lynch 2000). Indeed, it can be argued that some new currents of work in STS appear animated by reflexive concerns, though couched in a new register. For example, the emphasis on the performativity of social objects, on the multiplicity of ontologies, and on the essential situatedness of (their and our) knowledge production, found in the work of John Law, Anne-marie Mol, and others attempting to move beyond actor network theory (Law and Hassard 1999) is implicitly deeply reflexive.

See also: Actor Network Theory; Sociology, Epistemology of; Strong Program, in Sociology of Scientific Knowledge; Technology, Social Construction of

Bibliography

- Ashmore M 1989 *The Reflexive Thesis: Wrighting Sociology of Scientific Knowledge*. University of Chicago Press, Chicago
- Barnes B 1974 *Scientific Knowledge and Sociological Theory*. Routledge and Kegan Paul, London
- Bloor D 1991 *Knowledge and Social Imagery*, 2nd edn. University of Chicago Press, Chicago
- Collins H M 1985 *Changing Order: Replication and Induction in Scientific Practice*. Sage, London
- Hollis M, Lukes S (eds.) 1982 *Rationality and Relativism*. Blackwell, Oxford
- Latour B 1988 The politics of explanation: An alternative. In: Woolgar S (ed.) *Knowledge and Reflexivity*. Sage, London
- Latour B 1992 *Aramis ou l'amour des techniques*. La Découverte, Paris
- Latour B 1996 *Aramis or The Love of Technology*. Harvard University Press, Cambridge, MA
- Law J, Hassard J (eds.) 1999 *Actor Network Theory and After*. Blackwell, Oxford, UK
- Lynch M 2000 Against reflexivity as an academic virtue and source of privileged knowledge. *Theory, Culture and Society* 17(3): 26–54
- Mulkay M 1985 *The Word and the World: Explorations in the Form of Sociological Analysis*. George Allen and Unwin, London
- Pels D 2000 Reflexivity: One step up. *Theory, Culture and Society* 17(3): 1–25
- Woolgar S (ed.) 1988 *Knowledge and Reflexivity: New Frontiers in the Sociology of Knowledge*. Sage, London

M. Ashmore

Reflexivity: Method and Evidence

Unity of Science, sometimes referred to as logical positivism, is the dominant paradigm in social science. It assumes a fundamental unity between the physical and social world, and the applicability of the same model of science and scientific inquiry. This unity derives from the belief that reality is ‘out there’ waiting to be described. The goal of science—physical and social—is the discovery of universal laws that allow us to predict and explain according to causes that are contained in the laws or some less demanding ‘mechanism.’ Reflexivism is a catchall term for a variety of approaches that deny that the main task of a social science is prediction or the discovery of universal laws. The world, and especially the social world, is not a container of things that have to be captured by our concepts, but is to a large degree constituted by our descriptions and theories. For this reason, most reflexivists see fundamental differences between the physical and social worlds and believe that our task in the latter is understanding (*verstehen*), and only secondarily, or derivatively, explanation (*erklären*).

According to Hans-Georg Gadamer, the principal difference between the natural and human sciences, is

not one of method but of the aims of knowledge (Gadamer 1987, Schutz, 1973a). Despite this fundamental difference in purpose, reflexivism and the ‘mainstream’—our shorthand for social scientists who embrace the unity of science position—share a significant number of methodological practices. These include a clear differentiation of premises from conclusions, respect for the canons of inference, recognition that some standards of validation must be established for data and source materials, differentiation of correlation from cause; the recognition of spuriousness problems when we rely on correlations, reliance on logic in establishing warrants for our assertions, and the belief that all findings can be contested. The deeper division about method is *within* the reflexivist community, between phenomenological, interpretivist, and hermeneutic approaches on the one hand and some postmodern or critical approaches on the other which do not subscribe to the above criteria. Jean-François Lyotard (1987), a representative of the latter tradition, asserts that ‘narrative knowledge certifies itself in the pragmatics of its own transmission without having recourse to argumentation and proof.

Because reflexivists are concerned with the social order and its reproduction through the behavior of actors, they have to be attentive to questions of context and meaning within which these actions take place. Context and meaning are principally shaped by language, and reflexivists accordingly recognize the importance of norms and problems of interpretation. This recognition has important implications for the constitution of facts and their evidentiary value, the problem of truth and its role in ‘theory- building.’

1. First: What is Evidence

Alfred Schutz (1973a) observed that all facts are created by cognitive processes. John Searle (1995) distinguished between ‘brute,’ or observable facts (e.g., a mountain), and ‘social,’ or intentional and institutional facts (e.g., an encyclopedia article). For the mainstream, facts are created by interpretation because they need to be detached from their context by intellectual abstraction. For reflexivists, they need to be considered in their particular setting. Neither the mainstream nor reflexivists can describe, let alone explain and theorize, without importing meaning. Reflexivists accept this as an epistemological given, not a corrupting intrusion to be ignored, controlled for, or explained away. Mainstream approaches, especially quantitative ones, often seem oblivious to the complications that arise from the impossibility of a pure empiricism. James Coleman has nevertheless shown that every measurement procedure that assigns a numerical value to a phenomenon has to be preceded by a qualitative comparison. While the assignment of

Table 1

What is evidence?

A	Mainstream—the observed, including the effects of social structures on behavior
B	Phenomenologists—subject's observations of self and structure
C	Interpretivists—B + observations of structure and subjects through each
D	Hermeneuticians—C + observer's relationship with subject and structure

numbers may permit powerful mathematical transformations, it is illicit to make such assignments if the antecedent qualitative comparison has not or cannot be completed (Coleman 1964). It is here that reflexivist and mainstreamers clash. The latter trying—illegitimately in the eyes of reflexivists—to apply the traditional measurement techniques and scientific operations to the social world, while the mainstream assert that even if the preconditions for successful measurements or causal modeling are not present, the 'scientific method' should serve as a regulative idea (see Table 1).

For the mainstream, the reality that matters most is usually that of the theorist who looks at the facts from an 'objective' point of view. Phenomenology, interpretivism and hermeneutics emphasize the subject's story, but treat this evidence differently. Interpretivists may or may not situate themselves in their own account of the evidence, whereas hermeneuticians are consistently reflexivist. Hermeneuticians may treat the social structure as a more or less autonomous evidence, and may or may not become evidence themselves as a result of changes they experience through the act of interpreting the subject.

The first move of the interpretivist is phenomenological. It is to gather inductively evidence that allows [the researcher] to evaluate the problem as understood by the actor (Ricoeur 1984, Berger and Luckmann 1966). This is not significantly different from applying a garden variety decision making approach to a social science problem. But the reconstruction of an actor's perspective and understanding may be only the first step. Treatments of evidence acquire validity when the subjects themselves recognize that understanding as adequate (Schutz 1973b, p. 64). Researchers must read back their interpretations to those who have been interpreted. This standard of validation is not uncontested, as it too is based on imputation. Paul Ricoeur insists that such interpretations as 'configurational acts'—might *not* be intelligible to the actors themselves, since they represent constructions that no actor or witness could have put together when the events occurred (Ricoeur 1984, p. 157).

Why not just ask people what their statements and behavior mean? Interpretivists and hermeneuticians

insist that unobservable social structures exist in all social contexts, and make the meaning of a peoples actions and words not theirs alone to interpret or control (Ricoeur 1971). The researcher's task is to reproduce both the subject's actions *and* the encompassing frames and social structures within which this action took place. Imagine asking a white woman why she moved her handbag from one side of her body to the other when a group of black men were approaching on the other side of the sidewalk. She would probably deny her action was racist. But irrespective of her actual intentions, her behavior may have reproduced a particular racist structure by typecasting those young black men as potentially dangerous felons.

Hermeneuticians go a step further and insist that the role of researchers must be included in accounts of actions. (That this is a big step away from and beyond phenomenology is obvious in Schutz's (1973a, 1973b) uncritical treatment.) Researchers must comprehend the evidence through their own conceptions, while simultaneously realizing the perspectives of actors (1973a, pp. 104–11). Gadamer describes this process as placing oneself in the position of the other not to generate empathy or to better apply one's own criteria, but to attain a higher universality that overcomes, not only our own particularity, but also that of the other. This is the fusion of horizons of observer and subject (Habermas 1994). The hermeneutic assumption of intertextuality calls for an understanding of an actor's story in its complete relationship to all other stories available in a site, as well as in its relationship to the observer.

2. *Second: Where is Evidence Found?*

In a certain sense it is unfair to compare main stream and interpretivist evidence selection strategies, because they have different objectives. The mainstream wishes to test theories, interpretivism intends to generate understandings. Reflexivism wants to expand the field of inquiry. Instead of focusing only on 'why' questions that automatically privilege causal accounts, reflexivists insist that 'how possible' and 'what questions are also legitimate objects of inquiry. 'Why' questions direct attention to the antecedent causes in a phenomenon, whereas 'how possible' and 'what' questions are what makes a phenomenon what it is. (For a fundamental discussion of the differences between 'what' and 'why' questions, see Alexander Wendt 1999.) Reflexivists who study sovereignty are not interested, at least not primarily, in what brought it about but rather in the generative logic of the concept, its various instantiations in institutions, and the constraining and enabling force of its norms that allow actors to reproduce domestic and international structures.

To the extent that mainstream rules are designed to promote the assessment of alternative theoretical

accounts of a phenomenon, whether in a single case, or across some universe, they do not address interpretivist concerns. Because reflexivists do not usually engage in the competitive testing of theories, they are on the whole uninterested in the method of difference, crucial cases, and statistical significance. While the mainstream generally treats the knowledge of what constitutes a case as self-evident, reflexivists contend that case selection pre-judges 'facts' not yet in evidence.

This difference in orientation has implications for the choice of methods. While the mainstream stresses the deductive rigor in the models they employ, or the somewhat more problematic justifications for inductive inference, reflexivists stress the role of analogies and of conceptual frames that are antecedent to any causal argument or inductive inference. Given its choice of methods, the mainstream must consider counterfactual alternatives to which it rarely assigns 'weight' to the several causes considered relevant and apply the standard statistical controls to large enough samples whose cases are independent of one another. In practice, the mainstream decision rules that establish a warrant for a particular interpretation resemble more a procedure of sequential exclusion of alternatives than a 'deductive' demonstration or a generalizing inference.

Reflexivists do not believe it possible to establish complete analytical or theoretical control over a subject and consequently regard the 'bivalence' principle of logic (either something is so, or not so, and no third possibility exists) as generally inapplicable. The best our explanations can do is offer plausible accounts, even though some accounts might be more persuasive than others. Here a variety of criteria might interact and any proposed solution is not only 'preliminary' but contestable from the outset. By reversing the default drive of the mainstream theorists, who believe that a case or subject is theoretically identical until shown to be meaningfully different, reflexivists *assume* such differences among subjects until they can be persuaded otherwise. Reflexivists expect to find heterogeneity in their evidence; this is not a problem to be solved, but justification for emphasizing interpretative methods. They do not expect to find evidence that would allow predictions about future behavior within a case, let alone about a class of subjects across the historical universe. They argue that their stance allows for a more appropriate heuristics and thereby for a stronger warrant for their assertions. Reflexivists also believe that the traditional trade-off between rigor and detail misstates the underlying epistemological problem.

To elaborate on this last point, let us consider the problem of predictive success as a warrant for 'science.' Some mainstream approaches assert that the only warrant for a theory is predictive success, and that concern for the appropriateness of characterization is misguided. Reflexivists—scientific realists as well as

logical positivists—counter this claim by pointing out that accurate predictions are sometimes made by demonstrably false theories, and that making predictions the only criterion for warranted knowledge discourages *systematic* search for mistakes when predictions fail. To that extent we are in danger of substituting *ad hocery* for disciplined inquiry. In short, reflexivists recognize control as a problem that is very hard to solve, and so reduce their claims about their theoretical constructs. The mainstream tends to see control as a technical problem, and so does not adjust its aspirations and expectations about evidence. (See, for example, the high level of attention 'omitted variable bias' receives in King et al. 1994, pp. 168–82, but how easily it is expected to succumb to simple techniques. On p. 172, we are reassured that it is possible to account for omitted variables because 'fortunately, in most cases, researchers have considerable information about variables outside their analysis.' Interpretivists recognize that such information is often impossible to obtain.)

3. Third: Evidence and Theory-building

The secret to mainstream confidence in using evidence to *test theories* and develop generalizations for use in other places and times is the implicit assumption of uniformity. This assumes that things, concepts, and people are meaningfully identical across time and space. Votes are votes, wars are wars, and power is power—everywhere and for all time. (This problem has long been recognized in political science. In 1970, Giovanni Sartori pointed out the perils of doing comparative research. He warned against 'stretching' the meanings of variables across contexts in order to achieve some contrived generalization.) Some mainstream scholars acknowledge that unit homogeneity is often impossible to attain, but argue that understanding the degree of unavoidable heterogeneity helps to estimate the degree of uncertainty or likely biases of inferences (King et al. 1994, pp. 93–4). They allow the theory being tested to determine if a feature is meaningful. But what if the theory—and in social science this is virtually always the case—is not perfectly specified so as to avoid omitting a meaningful source of variance?

For reflexivists, this misstates the problem and so offers a misguided solution. If we knew which untheorized features of the field were relevant to understanding social outcomes, presumably we would have accounted for them in the *a priori* theory. Reflexivists see no solution to this problem because it is impossible, logically, empirically, and theoretically, to specify a causal model that could possibly account for everything that is excluded outside its boundaries, and everything that is purportedly causally effective within those boundaries. Nevertheless, reflexivism is subject to the same tension between homogenization and

uniqueness. It objects to mainstream homogenization without often explicitly acknowledging that it also uses forms of homogenization to make sense of the social world. *Intersubjectivity* is really a way of homogenizing various parts of society or of historical developments through common webs of meaning. This is why common understandings and the practices based on institutional rules are so important. If people understand the same social practices in the same way, then they offer an example of sameness, of homogeneity or at least of comparability. It should be clear that we are not dealing with 'constant conjunctions' à la Hume, or antecedent (efficient) causes as in mainstream analysis. Common understandings or norms that are constitutive of practices (institutions) do not cause 'behavior.' They may provide weighty reasons, guide or inspire behavior, rationalize or justify it, or they may be ignored. They do not effect cause in the sense that a bullet through the heart causes death or an uncontrolled surge in the money supply causes price inflation (Kratochwil and Ruggie 1986, p. 767).

Reflexivists need a methodological technique to deal with both difference and similarity. (The difference between positivist and reflexivist homogenization of social reality is that positivism simply assumes homogeneity or uses prior theoretical categories to impose it on difference, while reflexivism theorizes about and uncovers homogeneity, treating it as a social phenomenon to be discovered and understood, not as one to be assumed and inflicted; see Habermas 1994, p. 108.) The image of a text, requiring both interpretation and an audience to which it is addressed and which brings to it certain pre-understandings that frame it, provides here their heuristic metaphor. Instead of viewing actors as simple 'throughputs' for norms and institutional rules, reflexivists point to the constraining and enabling functions of norms that cannot be reduced to a structure that determines behavior. Here the term 'actor' who has to act out or perform something rather than serve as a 'mechanism' or conveyor belt of forces is an alternative metaphor with considerable heuristic power. Reflexivist interrogation of mainstream homogenization is not always successful, but encourages us to guard against ad-ducing identity where meaningful difference exists and vice versa.

The differences between the mainstream and reflexivity on this issue are easily observed when we consider how each handles the anomaly, a piece of evidence that confounds its theory or understanding of some phenomenon. The mainstream response is similar to how nineteenth century naturalists and anatomists dealt with physical 'monstrosities.' Rather than indicating systematic chaos or contingency, an interpretation with which reflexivists would feel comfortable, these anomalies instead provided opportunities to reaffirm the natural order and to define it with more precision. They were incorporated into more elaborate and powerful taxonomies (Ritvo 1997, p.

137). (Based on an empirical evaluation of the history of science, Pierre Duhem 1962) pointed out the power of prior theories, methodological commitments, and other elements of the scientific project to ensure the rejection of anomalous findings. Lakatos, Kuhn, and Feyerabend have also offered different accounts of why theories are so sticky in the face of countervailing evidence. As if these were not sufficient, there are excellent motivated and cognitive reasons to restore balance and persevere in ones prior beliefs.) While interpretivists try to place any given historical event within a narrative, within a contextualized bounded social structure, the mainstream places that same event within the framework of an opposition between the particular and the universal (Ricoeur 1984, p. 112). It treats the historical event according to whether it is part of a more general story to tell about the world, or an idiosyncratic 'outlier' that needs to be explained away.

Mainstream theories are sticky, in the Lakatosian or Kuhnian sense. Single pieces of contrary evidence do not and must not be used to change the observer's account of the situation (Kuhn 1970). But, ironically, and perhaps only in this regard, reflexivism is truer to Popperian (logical) positivism than the mainstream which adopts the practical and sociological defenses of Lakatos and Kuhn. Popper offers an account of falsification that resembles that of an open market or Darwinian system. Explanations fail when they are no longer competitive, according to established falsification criteria, relative to the other available explanations. In other words, one's understandings are far more responsive to new evidence, no matter how anomalous, than would be the case in either a Lakatosian or Kuhnian world. The product could be called a relative working truth, or, as Popper called it, 'situational certainty' (Popper 1979, pp. 78–81). It is a relative truth in that its validity exists only in comparison to other possible accounts. It is a 'working truth' in the sense that its validity is acknowledged to be a pragmatic convention, rather than an absolute fact that is simply part of the world.

Since we never test our theories against a reality, but only against other theoretical interpretations of the world, two additional problems arise. The first is that truth is no longer a property of 'the world out there,' but of theoretically informed assertions for which we need warrants. Consequently, the bivalence principle that links data and theories is no longer applicable, since most questions in science remain undecidable. This makes all scientific statements subject to revisions and introduces a second important element of pragmatics into theorizing. A statement, or 'fact,' is true only in the sense that can be accepted in daily practice until something more satisfactory comes along. It is a 'truth' in the sense that it is believed to be contextually valid, within the prior two constraints.

Reflexivism is doubly Popperian. It resists both theoretical priors, sociological and institutional con-

sensus, and motivated and cognitive impulsion toward premature closure. It treats every anomaly as if it is, at least potentially, meaningful. (Of course, Popper was never the naïve falsificationist who argued that a single disconfirmation meant the abandonment of the imperfect theory. Disconfirmation had to be accompanied by an alternative theory that could account for all that the failed theory had accomplished, and then some.) The mainstream, on the contrary, rejects treating every disconfirming instance as a serious challenge, instead appearing to don the protective belts offered by a Lakatosian research program, or the Kuhnian mantle of normal science toiling on within a settled paradigm.

Bibliography

- Berger P L, Luckmann T 1966 *The Social Construction of Reality. A Treatise in the Sociology of Knowledge*, 1st edn. Doubleday, Garden City, NY
- Coleman J S 1964 *Introduction to Mathematical Sociology*. Free Press of Glencoe, New York
- Duhem P 1962 *The Aim and Structure of Physical Theory*. Atheneum, New York
- Gadamer H-G 1987 Foreword. In: Baynes, Bohman, McCarthy (eds.) *Truth and Method, After Philosophy*, 2nd edn. MIT Press, Cambridge, MA, p. 340
- Habermas J 1994 *On the Logic of the Social Sciences*. MIT Press, Cambridge, MA, pp. 108, 151
- Lyotard J-F 1987 The postmodern condition. In: Baynes K, Bohman J, McCarthy T (eds.) *After Philosophy. End or Transformation?* The MIT Press, Cambridge, MA, p. 80
- King G, Keohane R O, Verba S 1994 *Designing Social Inquiry: Scientific Inference and Qualitative Research*. Princeton University Press, Princeton, NJ, pp. 93-4; 168-82
- Kratochwil F V, Ruggie J G 1986 International organization: a state of the art on the art of the state. *International Organization* 40: 753-76
- Kuhn T 1970 *The Structure of Scientific Revolutions*. 2nd. edn. University of Chicago Press, Chicago
- Popper K 1979 *Objective Knowledge: an Evolutionary Approach*. Oxford University Press, Oxford, UK, pp. 78-81
- Ricoeur P 1971 The model of the text: meaningful action considered as a text. *Social Research* 38(4): 534
- Ricoeur P 1984 *Time and Narrative*. University of Chicago Press, Chicago, Vol. 1, pp. 129-130, 157
- Ritvo H 1997 *Platypus and the Mermaid, and Other Figments of the Classifying Imagination*. Harvard University Press, Cambridge, MA, p. 137
- Sartori G 1970 Concept misformation in comparative politics. *American Political Science Review* 64(4): 1033-53
- Schutz A 1973a Common-sense and scientific interpretation of human action. In: Natanson M (ed.) *Collected Papers, The Problem of Social Reality*. Martinus Nijhoff, The Hague, Vol. I, pp. 6; 36-8; 104-11
- Schutz A 1973b Concept and theory formation. In: Natanson M (ed.) *Collected Papers, The Problem of Social Reality*. Martinus Nijhoff, The Hague, p. 64

Searle J R 1995 *The Construction of Social Reality*. The Free Press, New York

Wendt A 1999 *Social Theory of International Politics*. Cambridge University Press, Cambridge, UK

T. Hopf, F. V. Kratochwil, and R. N. Lebow

Reform: Political

In its most ordinary sense, political reform is any departure from the status quo that a proponent believes or represents as an improvement. But if reform is defined as those changes that democratic political systems have adopted in order to improve the legitimacy, accountability, and efficiency of modern government, it is possible to identify some common core values and reform actions in advanced democratic states. A recent survey of the 29 member states of the Organization for Economic Cooperation and Development (OECD) found that the most frequently stated public service values were (in order): impartiality, legality, integrity, transparency, efficiency, equality, responsibility, and justice (OECD 2000). Since to date less than half the OECD countries have explicitly adopted the last five of these core goals, they constitute the political reform frontier.

Democracies incorporate core values into their legal frameworks in the form of statutes, constitutions, civil service regulations, and public service acts. Reformers seek to change the legal framework in order to better realize one or more of the core values. The three main sources of these core values are general democratic principles (i.e., values consonant with systems in which citizens ultimately control governments), professional norms (i.e., standards for those who assume roles in the government), and prevailing ethical systems (i.e., societal conceptions of appropriate behavior).

The OECD survey discovered that impartiality is the most common core value held by its member states. This is defined as acting in a neutral or objective manner when dealing with various citizens and groups. The central idea embodied by this value is that states should not be partial towards specific persons in the administration of their laws. Examples of modern reforms that aim to increase impartiality are the creation of a neutral civil service, merit reviews for public employees, rules for bidding and winning public contracts, and judicial independence. It is not surprising that this is the most commonly stated value in advanced democracies. A government that administered its laws in a biased or partial fashion could not possibly act in the public interest. Partiality undermines democracy (by favoring the interests of the few over the majority's will), administrative professionalism (by violating the trust that citizens have in the government to act on their behalf), and conventional ethical norms (fair treatment as practiced in civil society).

Granting the importance of impartiality to modern governments, there is a serious debate as to how far impartiality should extend in public life. Elected officials and their top policy advisors are expected to represent the groups and individuals who put them in office so some forms of 'partiality' are routinely accepted in modern democracies. The key reform questions occur over when partiality is permissible and at what levels of government. In general, those who make policy are allowed to be partial to parties, constituents, and key support groups, while those who implement (i.e., public administrators below the highest levels) and those who interpret the laws (i.e., the judiciary or regulators) are expected to be more strictly impartial. But there is considerable cross-national variation in this rule as exemplified by the differences between the American and British Civil Service traditions (Kaufman 1956).

Closely related to the first value is a second: legality (a stated value in 22 of the 29 OECD countries). Legality means that decisions are incorporated into published laws and codes, and that there is due process protection for individuals when governments make decisions. It would be rare to find unpublished laws in advanced democracies although many, for instance, provide for relatively unlimited emergency powers in crises. More commonly, the reform focus is on due process—i.e., what constitutes a fair hearing for individuals who might be affected by government decisions. The problem with due process is that it can create highly complex and slow proceedings, greatly adding to the expense and untimeliness of government decisions. In politics like the USA where judicial due process is highly developed, reform means finding ways to adjudicate matters more informally and expeditiously (Kagan 1991), and in politics where due process is lacking, it might mean adopting an administrative procedures act (i.e., Greece and Portugal).

Integrity, the third most commonly stated value among OECD nations, is the subject of the most heated reform debates. It means that public office holders should not place their own interests above the public interest in the performance of official duties. When they do, this is usually considered corrupt behavior (Rose-Ackerman 1999). The most common forms of corruption are bribery, conflict of interest, extortion, and personal enrichment. Bribery is the exchange of something of personal value for the performance of an official act (Lowenstein 1985). It is almost universally proscribed in advanced democratic societies, but can take increasingly subtle forms. Hence, some bans on gifts, honoraria, and post-public service employment are attempts to prevent potential quid pro quo exchanges that might distort the judgment of public officials.

In the USA, this has led to increasing conflict of interest regulation, the purpose of which is to remove interests that might potentially seem to be corrupting (McAlevy 1988). Hence, officials have been required to

recuse themselves from participating in decisions that might directly and materially affect them. Alternatively, they might be asked to place their assets in a blind trust or even sell off potentially problematic holdings to prevent potential corrupt motives. Extortion and personal enrichment occur when public officials exploit the monopoly that the state has over some decision or service in order to extract rents. Some reforms that deal with this type of abuse include prohibitions against the unauthorized use of official or personal information, criminal penalties for those who misuse their office resources, anti-nepotism laws, limitations on accepting gifts, penalties for the failure to perform prescribed duties, bans against illicit enrichment, and laws against impeding or interfering with the public procurement process.

Corruption is a critical area of political reform. In Mexico, it has been a key issue for decades, and ultimately played a role in the *Partido Revolucionario Institucional's* demise in the 2000 national Presidential election. In the USA, anti-bribery logic has been widely incorporated into the campaign finance reform debate. As the modern state has assumed a wider regulatory and service provider role, the potential to use the state for corrupt purposes has grown. Nations that seek to enter into closer trading and political associations with OECD nations are under considerable pressure to address their unresolved corruption problems.

Another hotly contested area of political reform is transparency. Transparency (meaning the openness of the government to citizen observation and participation) is valued both for its own sake and for its instrumental value in promoting other core values. Full disclosure of government information helps to produce a more informed citizenry, which is a good in itself if it improves the quality of electoral decision-making. At the same time, full disclosure fosters greater accountability in the sense that it makes it harder for public officials to act in particularistic and corrupt ways without the public finding out and voting them out of office. Similarly, public participation can be beneficial to citizens, but also to the agency and decision making bodies that must make choices affecting real lives.

The most common transparency reforms are open meeting laws and freedom of information requirements (Thompson 1987, pp.116–122, Pupillo 1993). In the case of the former, so-called 'sunshine' laws require that the public and press be notified of impending meetings in which decisions might be taken. Citizens are given time to comment on the contemplated actions, and votes and discussions related to the meeting must be held in the open. 'Freedom of information' statutes require the government to release documents and information upon public request. Almost all advanced democracies have some restrictions on the release of confidential information that might compromise national security, police

actions, or personal privacy. As before, one can find considerable variation among advanced democracies. The Official Secrets Act in Great Britain gives government ministers much more ability to resist the disclosure of information that would be potentially damaging or embarrassing to the government. At the other extreme, some states in the USA have placed the right to request government information into their state constitutions.

The fifth value is efficiency. It is fair to say that this is a more contemporary reform goal. In an earlier period, public sector behavior was focused on the impartial administration of rules and programs, while private sector behavior was driven by market forces and profit. In the latter half of the twentieth century, private sector norms have been imported into public sector thinking. Government has been 'reinvented' to be more consumer-oriented and efficient. Efficiencies can be achieved by fostering competition through the contracting out of service activities to private firms, the promotion of inter-agency competition, and the benchmarking of government performance linked to outcome based rewards. One of the large issues this trend raises is whether accountability can be retained when activities are given over to private sector agents (OECD 1999).

Accountability is the ability to hold public officials responsible for their actions and decisions. Transparency is essential to a system of accountability in the sense that the public needs to know what officials are doing in order to hold them responsible for what the government does. While necessary, however, transparency is not sufficient for accountability; i.e., the public needs to be able to reward and punish officials for the consequences and quality of their actions and decisions. In democracies, this means that voters must have the ability to vote out incumbents who fail them (i.e., hence the need for truly competitive elections) and elected officials must be able to control the actions of those who administer the laws.

The problem that arises in advanced democracies is that incumbent officials and administrative officials often enjoy electoral advantages and legal protections that insulate them from the need to be highly responsive to public demands. In the USA, term limitations for legislators and campaign finance reform are seen as means to increase turnover and responsiveness. Critics charge that this greater responsiveness is purchased at the price of lower expertise and weaker judgment (Cain and Levin 1999).

The last two core values are equity and justice. Equity can mean either procedural or substantive equality. In the US for instance, the equity principle led to the 'one person, one vote' doctrine of the Supreme court—essentially requiring that all districts that elect officials for general governmental bodies must be equally populated to avoid diluting the vote of one individual vs. another. It also informs the desire to place limits on campaign contributions so that the

voice of one constituent is not unequal to that of another because the former has more wealth and resources than the latter (Cain 1995). The preoccupation with equity is stronger in the US than in most other advanced democracies.

Justice, on the other hand, refers to the substantive fairness of decisions that the government makes. Reform at this level begins to conflate with ideology, focusing on the government's obligation to make laws that provide positive opportunities for its citizens. Parties of the left and right will see these issues differently, and reform for one party can mean undoing the reforms of the other. The most general formulation that can be given to justice is that the laws accord with the common ethical and moral norms in civil society. Hence, where there is consensus in civil society about these norms, justice can be elevated beyond the merely political to broadly accepted goals.

1. Perspectives on Reform

This review of core values among OECD nations and the reforms that track those values is certainly not comprehensive. It merely illustrates the most common concerns. Apart from the debate over each of them, there are general differences among advanced democracies in reform orientations. They relate to the three sources of core goals: democratic, professional, and private ethical values.

With respect to the first, democratic values, there are important differences between those who interpret more democratic to be more majoritarian and those who resist such a formulation. American populists for instance prefer reforms that put more power directly in the hands of voters and less in the hands of elected and unelected officials. Neo-Madisonians in the US see popular control as lacking in the requisite checks against majority tyranny and in the expertise that representative government affords. Advocates of parliamentary systems prefer leaders to be chosen by (and therefore tied more directly to) other elected officials, and that the public show more deference to those in public office. Clearly, the position a society takes on this dimension strongly influences reforms that affect accountability, transparency, and equity.

Perspectives on professional values also shape modern governmental reforms. Systems that rely heavily on the socialization and training of public officials will tend to rely more on codes and internal monitoring (e.g., the British Civil Service). Politics that view professionalism as elitist or that are skeptical that people can be educated to high levels of public service, will rely on external checks and legal regulations (e.g., the American system).

There will also be important differences in the perspectives of those who believe that public ethics derives directly from private ethics vs. those who believe that the two differ in important ways. In the

view of the latter group, public officials need to be held to a higher standard and to ever growing scrutiny. In the view of the former, the two sectors can be analyzed within the same framework, and lessons learned from one sector can be imported to another. The shape of reforms to come will be determined by how future reformers situate themselves on these three continuums.

See also: Accountability: Political; Corruption: Political and Public Aspects; Democracy; Democracy, History of; Democratic Theory; Equity; Interest Groups; Justice and Law; Justice: Political; Social Justice

Bibliography

- Cain B E 1995 Moralism and realism in campaign finance reform. *The Legal Forum*, pp. 111–40
- Cain B E, Levin M 1999 Term limits. *Annual Review of Political Science* 2: 163–88
- Kagan R 1991 Adversarial legalism and American government. *Journal of Policy Analysis and Management* 10(3): 369–406
- Kaufman H 1956 Emerging conflicts in the doctrines of public administration. *The American Political Science Review* 50(4): 1057–73
- Lowenstein D W 1985 Political bribery and the intermediate theory of politics. *UCLA Law Review* 32(4): 784–851
- McAlevy K 1988 Conflicts of interest and state legislatures. *The Journal of Law and Politics* 5(1): 209–37
- OECD 1999 *Managing Accountability in Intergovernmental Partnerships*. OECD Publications, Paris
- OECD 2000 *Trust in Government: Ethics Measures in OECD Countries*. OECD Publications, Paris
- Pupillo T D 1993 The changing weather forecast: Government in the sunshine in the 1990s—An analysis of state sunshine laws. *Washington University Law Quarterly* 71(4): 1165–87
- Rose-Ackerman S 1999 Political corruption and democracy. *Connecticut Journal of International Law* 14(2): 363–78
- Thompson D F 1987 *Political Ethics and Public Office*. Harvard University Press, Cambridge, MA

B. E. Cain

Reformation and Confessionalization

1. The Reformation or 'Temps des Réformes'

The term 'Reformation' traditionally refers to those events and changes which in sixteenth-century Europe originated in church and theology and then took hold of the social, political, and cultural order, reorienting the thinking and behavior of individuals as well as social groups. From this point of view, the originator of all these changes was the German Augustinian friar Martin Luther (1483–1546), who published his 95 theses against the system of indulgences on October

31, 1517. This date was for the first time celebrated by Protestants as a secular holiday in 1617, and has since the nineteenth century been celebrated yearly as 'Reformation day' in countries all over the world. This chronological, geographical, and theological restriction to Luther and 1517 corresponds neither to the historical facts nor is it confessionally or historiographically undisputed. Rather, it is the result of a double interpretation: on the one hand, the contemporary heroization of Luther by his adherents or his demonization as arch-heretic by Rome and Roman Catholic countries like Spain (the Reformation as *pestis Germaniae*), and on the other hand, the monopoly on interpretation by the Lesser German (*klein-deutsch*) Lutheran historiography, which gained worldwide acceptance, particularly in the United States. This interpretation continued to influence the so-called Luther Renaissance of the 1920s and 1930s.

After World War II Reformation historians became less deeply rooted in confessional traditions, and a pluralism of methods and theories gained acceptance in historiography. As a consequence, the traditional 'Lutheran restriction' (*lutherische Engführung*, Bernd Moeller 1998) lost influence, and the Reformation increasingly appeared as a chronologically and geographically wide-ranging phenomenon, as a 'temps des Réformes' (Jean Delumeau 1997), a period of change from the medieval *societas christiana* to the differentiated, multi-confessional society of the early modern period. It increasingly appeared as a period of several centuries, in which the Reformations of Luther, Zwingli, Calvin, and all the other sixteenth-century reformers represented an important phase which accelerated a long-term development. Under the specific historical and historiographical conditions of English history, there was recently also developed a revisionist historiography, which on the one hand proceeds from the assumption that there were a number of 'Reformations' in sixteenth-century England and which on the other hand stresses the continuity of late medieval conditions in church and religion, particularly in popular religion (Eamon Duffy 1992, Christopher Haigh 1993).

To describe the meaning of 'Reformation' in current scholarly discussions as well as from a broader cultural perspective, three main dimensions should be distinguished: first, the historical events; second, the myth of the Reformation in Protestant and 'Western' self-understanding and self-assurance; and third, the historiographical concepts and the main issues of current research.

2. The Historical Context

On the one hand, there were new beginnings in theology, piety, spirituality, and the social manifestations of religion in the highly urbanized zones of late medieval Europe, above all in Italy, in the

Netherlands, in central Europe including Bohemia, in England, France, and Spain (the reform crisis around archbishop Cisneros, 1436–1517). In the context of the Humanism of the *Devotio moderna* and similar currents of thought, the laity acquired a more important role in the church, rational forms of church service (preaching) and religious behavior were strengthened, and a ‘normative centeredness’ (*normative Zentrierung*, Berndt Hamm) as well as an organizational and spiritual formation set in.

On the other hand, decisive changes occurred in the organization and administration of the church as well as in the relationship between state and church. The development of early modern national and territorial states led to a regionalization—but not yet to a division—of the universal church into ‘national’, territorial, or town churches, in which the secular governments gained control over the church of their respective territories either through concordats with the Pope or through a revolution (as the Hussites in Bohemia).

The Roman curia had also changed dramatically since the fourteenth century, because of the formation and bureaucratization of its administration and because of the triumph of monarchical papalism over conciliarism after the Council of Basel. Recent research on the papacy therefore considers Rome to have been more advanced on the path of modernization, and sees Luther’s Wittenberg Reformation as the result of a crisis of modernization in a remote, backward area of Christendom. Through Luther and the other reformers the religious and ecclesiastical changes received a new direction, legitimacy, and dynamism, which since the middle of the sixteenth century also took hold of the Roman Catholic church and changed it further. In particular the *sola* formulas—only through faith, grace, and the reading of the Bible could human beings be justified before God (*sola fide, sola gratia, sola scriptura*)—focused the different reform attempts in a new and unique way. The widespread church criticism of the late Middle Ages was put on a new foundation. Medieval, accidental anticlericalism developed into early modern fundamental anticlericalism. The popes and the clergy were no longer criticized only for the neglect of their duties, but it was claimed that the priests and the pope were superfluous as intermediaries between God and human beings: indeed, that they were harmful to the salvation of the individual, the community, and society as a whole. In this way, the pope and the Roman church were—in the eyes of their opponents—for the first time completely de-legitimized and unmasked as the Antichrist.

This also meant that the state or other secular governments (town councils, estates, etc.) gained more power over the church. When there was no longer a difference between clergy and laity, the prince could claim to be *praecipuum membrum ecclesiae* (privileged member of the church) and administer and govern the

church as *summus episcopus* or ‘supreme head of the church.’ In the late Middle Ages, there existed territorial or national churches within the universal church. With the Reformation, Protestant territorial or state churches which no longer belonged to a universal church came into being.

Luther’s appearance (above all at the Imperial Diet of Worms in 1521), Zwingli’s appearance in Zurich, and Calvin’s appearance in Geneva in the 1540s, but also the appearance of several other reformers who had their own ideas about theology and church policy, thoroughly changed the intellectual and religious climate and opened up new perspectives in state and society. The printing press became decisive for the quick dissemination of the new teachings: not only were the writings of Luther and other reformers published in print runs of hundreds of thousands, but the mostly illustrated broadsheets constituted a new medium of mass communication. In this way, not only the intellectuals, the political elites, and the princes were seized by the new visions and ideas, but also the lower nobility, the burghers, and the peasants. In the Holy Roman Empire, the Reformation was a mass movement which climaxed in the rebellion of the imperial knights (Sickingen, Hutten), the Peasant’s War of 1525/26 and—until the 1540s—in burgher movements in the imperial and territorial towns.

Because of the close connection between religion and society in old Europe, the Reformation of the church inevitably led to a reformation of the political and social order. In this context, the idea of a communal church, decisively influenced by the laity, had a strong appeal. In the secular sphere, this corresponded with corporative (estate), communal, or republican (urban) concepts of order defended by the knights or the urban and rural communities (*Gemeinderreformation*, communal Reformation). Later, comparable reform movements occurred in the Netherlands, France (nobility and towns, above all in the south), Poland (nobility), with considerable delay also in England (in the seventeenth century, above all in London and among the gentry). These communal movements only had a lasting success in regions with favorable social, political, and constitutional preconditions, above all in Switzerland and the northern Netherlands.

In most of the states where the Protestant Reformation took hold, including the German territories, the princes and the officials of the central state defeated these Reformation movements of the nobility and the common man ‘from below.’ They took the Reformation into their hands and utilized it for their own interests, especially the formation of the early modern state and of an early modern society of disciplined subjects (*Fürstenreformation*, princely Reformation). As a consequence, early modern territorial churches (as in the Empire) or national churches (as in Scandinavia, England, etc.) came into being. These were increasingly staffed by a university educated and

married clergy, whose area of responsibility was identical with the territory of the respective state and who in numerous ways supported early modern state-building.

In this context, important developments were: First, the increase in funds and property assets through secularization; second, the extension of the activities of the state upon the so-called *res mixtae*, i.e., education, marriage and family, poor relief, and social welfare; third, the mutual strengthening of religious and political identities among the people; fourth, the increased legitimacy of the prince who as supreme head of the church or *summus episcopus* had a quasi-sacral quality; and fifth, the instruments of the state church (church discipline, visitations, sermons, and catechesis) which aimed at influencing the behavior and thinking of the people and which effectively supplemented secular forms of disciplining.

However, the climax of this connection between religious and political change was only brought about by the process of confessionalization between ca. 1550 and ca. 1650, which was a fundamental process of society on the basis of explicitly formulated confessions of faith. On the Protestant side these were the *Confessio Augustana* of 1530 and especially the *Book of Concord* of 1580 for Lutheranism, the *Helvetic Confessions* of 1536 and 1566 as well as the other 'national' confessions, but especially the *Canons of Dordrecht* of 1619 for Calvinism, the *Book of Common Prayer* of 1549, and the *Thirty-nine Articles* of 1563 for Anglicanism. By committing the subjects to a fixed set of norms concerning dogma, belief, and behavior, the different confessional churches became clearly delimited and this corresponded to a high degree of integration on the 'inside.' At the same time, the processes of formation in state and society became more dynamic and were given a new legitimacy.

As the Catholic church also went through such a process of confessionalization (based on the doctrinal canons of the Council of Trent, especially the *Professio Fidei Tridentina*), the Catholic and re-Catholicized countries of Europe (besides the German territories, e.g., Austria, Bavaria, and the prince bishoprics, above all Spain, Poland, France, and the Italian states) participated in the above-mentioned renewal of state and society in the wake of the reformation of the church. In contrast to Max Weber and the sociology of religion influenced by him, the historiography on confessionalization no longer assumes that only one of the early modern confessional churches—the Calvinist church—was the driving force of modernization, but that modernization resulted from the Calvinist, the Lutheran, and the Catholic confessionalization: in this respect, all three confessionalizations were functionally equivalent (Wolfgang Reinhard 1977).

However, in Catholic societies in particular there were often extra-confessional factors like geographical location, structure of the social elites, etc. that hampered the potential for modernization of the Catholic

confessionalization. In general, the state authorities, which in the course of the Reformation and the processes of confessionalization aimed at social, political, and cultural concentration and standardization as well as at the formation of an early modern society of disciplined subjects, were again and again confronted with the opposition of these subjects. Popular religion, including the belief in magic, remained widespread as did older forms of behavior which contradicted the norms of state and church. Changes in society more often than not lagged behind the normative aims of state and church. Moreover, it should not be forgotten that the differentiation of Latin Christendom had from the beginning overshot the Protestant Reformations accepted by the state. Already during the sixteenth century Baptists, Spiritualists, Anti-Trinitarians, etc. formed a big group of '*Chrétiens sans église*' (Kollakowski 1969) outside of the established confessional churches, and there were also some atheists. After the age of confessionalization in the second half of the seventeenth century, such trends became more and more important.

3. The 'Reformation' as Myth and Concept in Universal History

Just like the French Revolution, the Reformation also became a *lieu de memoir* and a myth: first of all for the Germans, but eventually with a universal claim. Luther was already made a hero by his contemporaries, for example as *Hercules Germanicus*, who fights against the clerical Hydra with a mace (woodcut by Hans Holbein, 1520). Especially, the nailing of the 95 theses to the door of the Castle Church in Wittenberg became an icon of modern times. However, this event appears only one generation later in one of Melanchthon's writings, and therefore probably never happened. But for generations of history teachers and the educated bourgeoisie the blows of Luther's hammer on the church door were a signal for the beginning of the modern period.

A critical distance to the religious meaning of the Reformation notwithstanding, the Enlightenment prized Luther and his rebellion against Rome as an important steppingstone towards liberty of conscience and the emancipation of the individual from authority, an interpretation further elaborated by Hegel and the liberals of the nineteenth century. Hegel identified the Reformation with the rise of modernity and saw it as '*die alles verklärende Sonne*,' which follows upon '*die Morgenröte des Mittelalters*.'

Leopold von Ranke's interpretation of the Reformation was even more influential in historiography. In his *Deutsche Geschichte im Zeitalter der Reformation*, which was first published in 1839, widely read and translated into several languages, Ranke interpreted the German Reformation of 1517 as a signifi-

cant turning point in universal history. Hegel's and Ranke's interpretations of course included the Calvinists and Western European Protestants, who called themselves 'églises réformées' or 'ceux de la religion réformée.' In a secularized form, this interpretation became one of the kernels of the concept of 'Western civilization,' particularly influential in the USA, as well as of the Marxist concept of human progress by revolutions (see Sect. 4, *frühbürgerliche Revolution*, early bourgeois revolution). In current historiography the term 'Reformation' remains relevant especially in the English-speaking world, where handbooks and encyclopedias are given the title 'of the Reformation,' even if they are equally concerned with the Renaissance and the period of confessionalization and cover much more than the history of religion and the churches.

4. Historiography and Issues of Current Research

In the research and interpretation of the Reformation and the period of confessionalization four main currents can be identified at present: first, church history in the stricter sense of the word; second, societal history (*Gesellschaftsgeschichte*), which modifies the classic sociology of religion in the tradition of Max Weber and Ernst Troeltsch and adapts their findings to the historiography of the early modern period; third, neo-Marxist interpretations; and fourth, post-modern approaches like cultural, gender, and ethno-history.

Its commitment to modern scholarly standards notwithstanding, church historians in theological seminaries are still oriented towards the meaning of the Reformation for the history of salvation as well as the historical background of the respective scholar's own confessional existence. In contrast to this, the other approaches emphasize the secular preconditions and consequences of the Reformation. Ranke's interpretation of the Reformation as a turning point in universal history pointed the way and was further developed by Max Weber, who brought the potential of the modern period for rationalization, discipline, and modernization in connection with the Reformation (*Die calvinistische Ethik und der Geist des Kapitalismus*).

The Marxist interpretation of the Reformation as an 'early bourgeois revolution' (*frühbürgerliche Revolution*, F. Engels, M. M. Smirin, M. Steinmetz) interpreted the role of the Reformation in universal history in a revolutionary, progressive sense. Moreover, in the historiography of the German Democratic Republic, the German nationalist perspective of the nineteenth century was retained: Reformation and 'early bourgeois revolution' were seen as compensations for the lack of a secular bourgeois revolution in Germany. At present these macro-historical interpretations continue to make themselves felt in two research para-

digms developed by German historians: in the thesis of 'the Revolution of 1525' and the interpretation of the Reformation as communalism by Peter Blickle and Thomas A. Brady on the one hand, and in the paradigm of confessionalization by Wolfgang Reinhard and Heinz Schilling on the other hand. The paradigm of confessionalization neutralizes Weber's fixation on Calvinism and instead sees the dynamism of change in the overarching process of confessionalization be it Lutheran, Catholic, or Calvinist. In France and in other Romance countries the history of the Reformation was and is not at the center of interest.

In the countries of Eastern central Europe a consciousness for the general importance of the early modern history of church and religion is only gradually developing. English historiography follows its pragmatic traditions also in this respect, and—apart from the revisionism of the English Reformation mostly lead by Catholic historians—it is little interested in the theoretical and methodological discussions about the macro-historical role of the Reformation. By contrast, the approaches of micro-history and cultural history, which claim to supplement or replace the traditional macro-historical interpretations and which have recently become the focus of attention, have their roots especially in Anglo-American historiography. On the one hand, the advocates of these approaches stress the continuous meaning of magic and popular religion (Bob Scribner 1981), which also in Protestant countries resisted the rationalization of the interpretation of the world as well as the disciplining of belief and behavior according to the norms of the respective confessional churches. On the other hand, they stress the long-term change in the meaning of rites and images, in the concepts of order, thinking, and behavior, etc., from the late Middle Ages until the eighteenth century, and they thereby open up new sources and further knowledge to the discussion about the character of the Reformation and its historical meaning.

See also: Catholicism; Cultural History; European Religion; Fundamentalism (New Christian Right); Media and History: Cultural Concerns; Political Parties, History of; Poverty in History; Protestantism and Gender; Renaissance

Bibliography

- Archiv für Reformationsgeschichte: Internationale Zeitschrift zur Erforschung der Reformation und ihrer Weltwirkungen* [*Archive for Reformation History: An International Journal concerned with the history of the Reformation and its significance in world affairs*] 1903/1904
Blickle P 1992 *The Communal Reformation*. Humanities Press, Atlantic Highlands, NJ

- Chaunu P 1975 *Le Temps des réformes: Histoire religieuse et système de civilisation: La Crise de la chrétienté. L'Éclatement, 1250–1555*. Fayard, Paris
- Delumeau J, Wanegffelen T 1997 *Naissance et affirmation de la Réforme*. P.U.F., Paris
- Duffy E 1992 *The Stripping of the Altars: Traditional Religion in England c. 1400–c. 1580*. Yale University Press, New Haven, CT
- Guggisberg H R, Krodell G G (eds.) 1993 *Die Reformation in Deutschland und Europa: Interpretationen und Debatten [The Reformation in Germany and Europe: Interpretations and Issues]*. Gütersloher Verlagshaus, Gütersloh, Germany
- Haigh C 1993 *English Reformations: Religion, Politics, and Society under the Tudors*. Clarendon Press, Oxford, UK
- Hillerbrand H J (ed.) 1996 *The Oxford Encyclopedia of the Reformation*, 4 vols. Oxford University Press, Oxford, UK
- Kolakowski L 1969 *Chrétiens sans église: La Conscience religieuse et le lien confessionnel au 17^e siècle*. Gallimard, Paris
- Moeller B (ed.) 1998 *Die frühe Reformation in Deutschland als Umbruch*. Gütersloher Verlagshaus, Gütersloh, Germany
- Oberman H A 1994 *The Reformation: Roots and Ramifications*. W B Eerdmans, Grand Rapids, MI
- Reinhard W 1977 Gegenreformation als Modernisierung? Proleg zu einer Theorie des konfessionellen Zeitalters. *Archiv für Reformationsgeschichte* 68: 226–52
- Reinhard W 1989 Reformation, Counter-Reformation, and the early modern state: A reassessment. *The Catholic Historical Review* 75: 383–404
- Schilling H 1992 *Religion, Political Culture and the Emergence of Early Modern Society: Essays in German and Dutch History*. E J Brill, Leiden, The Netherlands
- Schilling H 1995 Confessional Europe: Bureaucrats, la bonne police, civilizations. In: Brady Th A Jr, Oberman H A, Tracy J D (eds.) *Handbook of European History 1400–1600. Late Middle Ages, Renaissance and Reformation*. Vol. II, E J Brill, Leiden, The Netherlands, pp. 641–81
- Scribner R W 1981 *For the Sake of Simple Folk: Popular Propaganda for the German Reformation*. Cambridge University Press, Cambridge, UK

H. Schilling

Refugees: Geographical Aspects

Migrants are commonly differentiated based on their motivation for moving. The term 'refugee' is used to denote an individual who is forced flee in search of safety in contrast to someone who chooses of their own accord to relocate, i.e., a voluntary migrant. This entry examines geographical aspects of forced migration, a condition that applies to more than 34 million of the estimated 100 million permanent migrants worldwide.

1. The Term 'Refugee'

Legal distinctions regarding refugees can be traced to the United Nations Conference on the Status of

Refugees and Stateless Persons. Article 1 of the 1951 Geneva Convention Relating to the Status of Refugees applied the term to

[Any person who] ... owing to a well-founded fear of persecution for reasons of race, religion, nationality, membership of a particular social group, or political opinion is forced to migrate involuntarily across an international boundary and remain outside his country of nationality.

Cause, motive, and political boundary thus became the classic benchmark determiners of refugee status. The 1951 Refugee Convention specifically addressed the status of Europeans displaced by WWII.

An amended 1967 Protocol established legal standards for refugees in all world regions at any time. Although the 133 signatories are not required to provide asylum to refugees, they are prohibited against *refoulement*—expelling or returning refugees to countries where their life or freedom would be threatened (Hamilton 1999).

Under international law, final authority to grant refugee status rests in the hands of a given host government. Practically speaking, however, a branch of the United Nations bears the primary task of protecting refugees. In 1951, the UN General Assembly created the Office of the United Nations High Commissioner for Refugees with a brief three-year life span in mind. In 1954, however, the Assembly extended UNHCR's mandate for another five years. Successive five-year renewals have followed ever since.

Headquartered in Geneva, UNHCR is charged with refugee protection and the promotion of lasting solutions to their displacement. According to its statute, UNHCR is competent to assist:

[Any person who] ... owing to well founded fear of being persecuted for reasons of race, religion, nationality, or political opinion, is outside the country of his nationality and is unable or, owing to such fear or for reasons other than personal convenience, is unwilling to avail himself of the protection of that country; or who, not having a nationality and being outside the country of his former habitual residence, is unable or, owing to such fear or for reasons other than personal convenience, is unwilling to return to it (UNHCR 1993).

Protection remains UNHCR's *raison d'être*, but the complexities of post-Cold War conflicts have forced the agency to broaden its scope considerably. Whereas the Office of the United Nations High Commissioner for Refugees once restricted its concern to people outside their country of origin, it has become increasingly involved in assisting and protecting returnees who repatriate to their home countries as well as groups of people who face refugee-like situations within their country of origin (UNHCR 1993). As the dimensions and dynamics of forced migration change so the term 'refugee' has come to apply to other forms

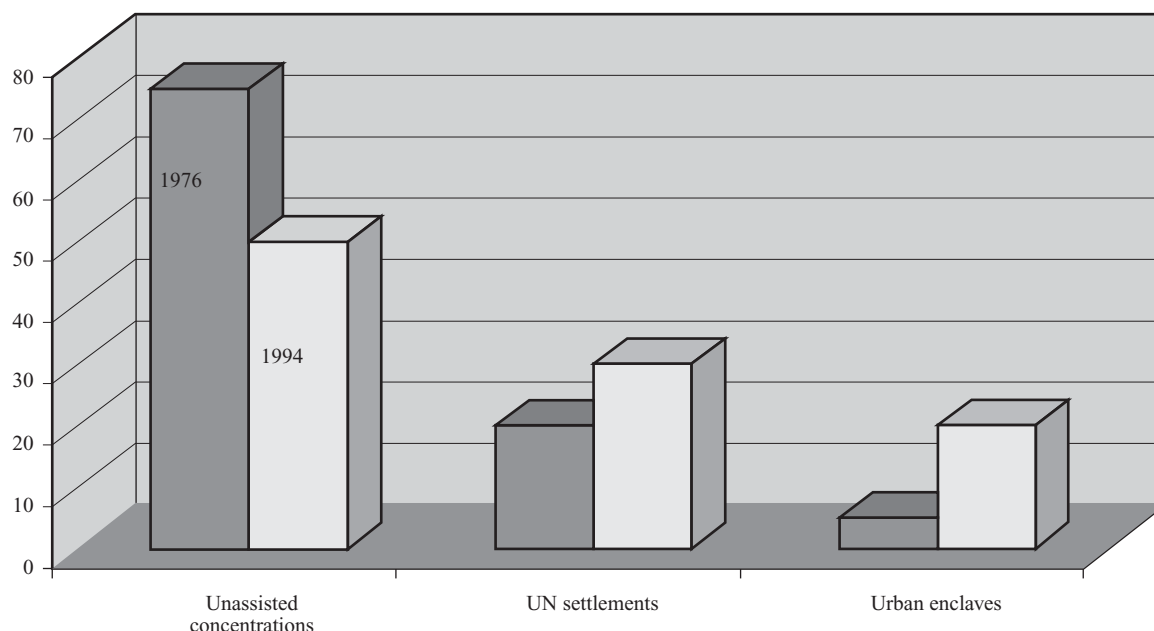


Figure 1
Principal forms of refugee settlement in Africa, 1976 vs. 1994

of displacement. The discussion to follow examines 'classic' international refugees who cross a political boundary and then, several types of 'internal' refugees.

1.1 International Refugees

Although geographers, starting with Ravenstein's work in the late 1800s, have long studied voluntary migration, their analysis of forced migration lagged far behind. At the 1950 meetings of the Association of American Geographers, Proudfoot introduced a nine-fold categorization of refugee population. That paper constitutes an early example of geographic work on the specific topic of refugees. Geographers thus began documenting the flight and distribution of specific refugee populations, and, they have continued to do so ever since.

Kunz (1973) offered a set of conceptual postulates structured around the dimensions of space and time. That a sociologist provided the model of refugee flight patterns reflected that fact that geographers still gave limited attention to forced migration. In 1977, an article published in *The Professional Geographer* summarized the state of the refugee literature. Rogge concluded that geographers had produced a considerable volume of literature on human mobility, but that '... one form of mobility, however, namely the forced or impelled migration of refugees, has received only peripheral attention by the discipline.'

Refugee research, including that of geographers, has grown substantially during the last 25 years

(Hyndman 1999). Refugee study programs now exist at universities in Canada, Egypt, Great Britain, Kenya, Sweden, and the United States. Academic interest has correlated with the growing magnitude of refugee crises worldwide wherein the global total grew from less than 3.5 million in 1977 to a peak point of 18.2 million in 1992 (UNHCR 1993). Although media coverage tends to reinforce the perception that all refugees are in camps, most refugees—at least in the case of Africa—are still more likely to reside outside UN settlements (see Fig. 1).

During the 1980s, refugee research largely focused on the first and second stages of the refugee experience—the processes of flight and resettlement. Resettled Central Americans and Indochinese in the United States were the focus of study in the context of more developed countries. That East Africa had become the epicenter of refugee crisis in the developing world was reflected in books on Eritrean and Ugandan resettlement (see Rogge 1985, Harrell-Bond 1986, and Bulcha 1988). A basic distinction was frequently made between assisted refugees housed in UN-sanctioned camps or settlements and unassisted refugees who either self-settled in rural areas or in urban centers (Wijbrandi 1990).

In 1990, the locus of the global refugee population was predominantly in the global South. Developing countries hosted nearly 97 percent of all refugees. Conflicts in Afghanistan, Ethiopia, and Mozambique produced populations that each numbered in the millions apiece. The Palestinian community (2.48

million) had long since become permanent stateless exiles dispersed throughout the Middle East. Together, these four groups accounted for more than 13 million of the 16.6 million refugees worldwide. Yet, refugees could be found among more than one hundred different countries. And exile periods were extending longer and longer.

During the decade of the 1990s, geographers diversified their approach to match the growing complexities of the refugee picture worldwide (Black 1991 and Black and Robinson 1993). Refugee settlement was treated less and less often as if it were a self-contained process in isolation, thus moving beyond a 'refugee-centric' perspective (Chambers 1986). Bascom (1993), for example, related the larger processes of rural transformation and agrarian change to that of refugee settlement. Although media images give the impression that refugees are mass 'flows,' it became abundantly clear that ethnic, age, gender, and class differentiation all exist among uprooted populations (Bascom 1998).

Repatriation became a 'lead' research theme in the 1990s (Wood 1989 and Allen and Morsink 1994). As early as 1984, UNHCR had identified repatriation as one of three 'durable' solutions for refugees. The *de facto* solution is long-term integration of refugees in the country of first asylum. The exceptional solution, reserved for only a small percentage of asylum applicants, is resettlement to a third country of asylum. The optimum solution of the three is repatriation—the return migration of refugees from exile to their homeland. The UN High Commissioner for Refugees proclaimed the 1990s as the 'decade of repatriation.'

Repatriation research focused first on legal or logistical matters, then shifted to the socioeconomic and psychological dimensions of repatriations (Rogge 1991). Mozambique, Eritrea, and Afghanistan were some of the principal countries of concern. In theory, the repatriation decision is meant to be voluntary, fully informed, and only to occur when the conditions that prompted flight no longer exist. Practically, however, repatriations can involve different scenarios such as refugees choosing to return despite ongoing conflict or being forced to return home by host governments. Forced deportation, or *refoulement*, of refugees is the worst possible violation of the 1951 Refugee Convention.

The face of refugee studies changed dramatically during the last decade of the twentieth century. It began with the dissolution of the Soviet Union, whose fragmentation prefigured a pattern to come; states under siege amid competing national interests and thus producing complicated flows and counter-flows of displaced people. Wars in the Persian Gulf and Yugoslavia spiked the refugee total up to its record high. The year of 1994 marked the largest and fastest exodus in the modern history of refugee migration. More than 250,000 Rwandans fled the country in only 24 hours. Yet the era of mass movement across discrete

political boundaries was already giving way to more 'complex humanitarian emergencies' (Hyndman 1999).

The painful saga of intervention in Somalia followed by the Rwandan debacle have made the international community increasingly aware of the relationship of refugees to peacekeeping, peacemaking, political stability, and humanitarian action. The UN began emphasizing the need to:

(a) incite and expedite conflict resolution in source countries so refugees can return,

(b) take vigorous action—including emergency response to crises and the provision of international protection on site—to eradicate the causes of flight before refugees flee, and

(c) initiate more proactive intervention to prevent the development of conditions that might impel people to leave in the first place (UNHCR 1993).

The United Nations forged cross-mandate approaches to refugee crises. In the case of Somalia and Mozambique, for example, UNHCR sought collaboration with governments, the UN Department of Humanitarian Affairs (DHA), other concerned United Nations agencies, and NGOs to establish means of protection and assistance for all affected populations, i.e., returnees, internally displaced persons and other impoverished nationals. UNHCR has also increased assistance for refugees outside formal camps and settlements as well as to local populations in resettlement areas.

Meanwhile, the balance of the globe's refugee population began to swing toward Europe. People had come to think of refugees as a Third World problem, but the shattering of Yugoslavia forced Europeans and North Americans to engage with refugees much more directly. First, warfare sites moved closer. Then refugees turned to neighboring countries for protection, assistance, and asylum. And the need for political stability in southern Europe as well as ethnic links to Bosnians, Croatians, Slovenians, and Serbs could not be ignored. Whereas Europeans and North Americans enjoyed the luxury of modest refugee resettlement programs during the 1980s, they were inundated with asylum requests during the 1990s. The United States, Canada, and 17 European nations all received refugees from Bosnia. Meanwhile, the plight of Kurds entrapped within Iraq initiated a new debate about 'internal refugees.'

1.2 Internal Refugees

During the 1980s and 1990s, the character of war shifted in balance from conflict between sovereign states to civil conflict. International refugees now represent the smaller portion of the globe's displaced population. Internally displaced persons (20 million) outrank the number of 'classic' refugees—asylum-seekers who have involuntarily crossed an interna-

Table 1

Significant populations of internally displaced civilians, as of early 1999

Country	Total number
Sudan	4,000,000
Angola	1,000,000–1,500,000
Columbia	1,400,000
Iraq	1,000,000
Afghanistan	540,000–1,000,000
Burma	500,000–1,000,000
Turkey	400,000–1,000,000
Bosnia–Herzegovina	836,000
Azerbaijan	576,000
Sri Lanka	560,000
India	520,000
Burundi	500,000
Rwanda	500,000
Lebanon	400,000–500,000
Uganda	400,000
Russian Federation	350,000
Peru	340,000
Congo-Kinshasa	300,000
Sierra Leone	300,000
Georgia	280,000
Cyprus	265,000
Yugoslavia	257,000
Congo-Brazzaville	250,000
Somalia	250,000
Guinea-Bissau	200,000
Kenya	200,000
Ethiopia	150,000
Syria	125,000
Philippines	122,000
Algeria	100,000–200,000
Eritrea	100,000
Liberia	75,000
Croatia	61,000
Armenia	60,000
Bangladesh	50,000

tional boundary and remain outside their country of nationality. Sudan (4 million) tops the list of countries with large numbers of civilians displaced by persecution or armed conflict (see Table 1).

'Internal refugees,' more commonly known as internally displaced persons (IDPs), became a central focus of refugee research (see Bennett 1998, Cohen and Deng 1998, and Hampton 1998). In 1999, the IDP table in the *World Refugee Survey* included 41 countries as compared to only 25 in 1990. IDP growth reflects the pandemic of civil conflict as well as paramilitary activity in settings of chronic social anarchy like Somalia and Columbia. Most IDPs are refugees in every respect except that they have not, or, cannot cross an international boundary. Although civilians are more frequently targeted as a deliberate tactic of war, they have no right to protection or

humanitarian assistance as long as they remain inside their countries. Estimates of the number of internally displaced persons are often fragmentary and unreliable because they remain largely inaccessible to outside monitors and non-governmental relief organizations (Hamilton 1999). The total number of IDPs may be much higher than 17 million.

'Internal refugees' have become a central policy concern. In 1992, the secretary-general of the UN appointed a Representative on Internally Displaced Persons. Representative Francis Deng continues to parlay the status associated with his title into a limited protection framework for IDPs. The current working definition of IDP is that of a:

Person or group of people who have been forced to flee or to leave their homes or places of habitual residence as a result of, or in order to avoid, in particular, the effects of armed conflict, situations of generalized violence, violations of human rights or natural or human-made disasters, and who have not crossed an internationally recognized border (Bennett 1998).

Clearly, the locus of concern is limited to those forcibly uprooted within a given national territory. The working definition is broadly inclusive, combining several forms of internal migration, which can also be are treated as separate entities.

Most IDPs are forced to move due to explicit forms of persecution or armed conflict, but others migrate because of natural or man-made disasters. Victims of clear-cut natural catastrophes such as a hurricane in Honduras, massive mud slides in Venezuela, or repeated earthquakes in Turkey are sometimes referred to as 'environmental refugees.' Yet, famine-induced migrations often involve man-made factors. Drought, for example, was the immediate, observable cause of large population movements in Ethiopia and Sudan during 1984 and 1985. Poor rainfall may have prompted flight, but displacement within Ethiopia was amplified by the deliberate obstruction of food aid into stricken areas. In Sudan, more than 100,000 people left their homes. Largely at fault was the structure of the Sudanese political economy that set the stage for migration out of rural regions. For many who left, migration was but the culmination of an impoverishment process associated with the appropriation of labor and land by mechanized agricultural schemes, obstructionist tactics employed by merchants, and the state's long-term neglect of the rural sector. The flight of 'environmental refugees' like these is a symptomatic manifestation of oppressive economic and political conditions.

Two other forms of forced migration bear mention—migrants in refugee-like conditions and development-induced displacement. In its highly regarded annual *World Refugee Survey*, the US Committee for Refugees, a public information program of Immigration and Refugee Services of America, reserves a separate table entitled 'For people in refugee-like

Table 2
Significant groups of people in refugee-like situations, as of early 1999

Host country	Place of origin	Total number
Russian Federation	Former USSR	1,000,000
Jordan	Palestinians	800,000
Thailand	Burma	350,000
Bangladesh	Pakistan (Biharis)	238,000
Jordan	Iraq	150,000
Syria	Stateless Kurds	142,000
Kuwait	Kuwait (Bidoon)	117,000
Iraq	Kuwait (Bidoon)	100,000
Tanzania	Burundi	95,000
Ukraine	Stateless Tartars	91,000
Syria	Palestinians	64,000
Venezuela	Colombia	60,000
Eritrea	Ethiopia	35,000
Cameroon	Chad	30,000
Congo-Kinshasa	Rwanda	30,000
Ecuador	Columbia	30,000
Kyrgyzstan	Tajikistan	26,000
Syria	Iraq	23,000
Lebanon	Palestinians	22,000
Yemen	Iraq	20,000
Yemen	Somalia	15,500
Turkmenistan	Tajikistan	11,000
Slovenia	Former Yugoslavia	8,000
Kenya	Rwanda	7,000
Panama	Columbia	7,000
Congo-Brazzaville	Rwanda	5,000
Tanzania	Rwanda	5,000
Yemen	Sudan	5,000

conditions' (see Hamilton 1999 and Table 2). The 1999 report identifies more than 3.5 million such persons found within 28 countries. Some of these refugee-like people are stateless, denied the protection afforded by citizenship, e.g., Kurds in Syria or Tartars in Ukraine. Others are regarded by host governments as illegal aliens, or simply tolerated, or ignored (Hamilton 1999). Development-induced displacement is a final form of forced migration. Revolutionary governments in China and Ethiopia launched ambitious development programs that required resettlement of hundreds of thousands of people, some voluntarily, many not. Today in China, work continues on the largest dam resettlement project in history. Thus far, more than 1.2 million persons have been displaced in the wake of the Three Gorges Dam project.

2. Future Directions

At the end of the 1990s, the locus of the world's refugee population remained predominantly in the South, but 2.38 million of the world's 13.4 million refugees (18 percent) are settled in the First World.

Although the total number of refugees has decreased since 1992, the number of countries that host asylum-seekers increased to 125. Many countries host refugees from several countries simultaneously; thus 161 different populations of exiled refugees exist worldwide. Palestinians (3.8 million) and Afghans (2.6 million) still constitute the two largest refugee communities (see Table 3). Iraq, Sierra Leone, Bosnia and Herzegovina, Somalia, Sudan, Eritrea, Liberia, Croatia, Angola, Burundi, Vietnam, and El Salvador are also important source countries. At least 250,000 citizens from each one have fled into exile. Meanwhile, migration policies are tightening in Europe and North America. Geo-policing of migration poses new challenges to refugees who desire asylum and permanent residence in fortress Europe. In addition, the United States has placed some of the most severe restrictions on asylum procedures in its history.

Contemporary refugee flows are subject to far more complexity than prevailed in the earlier European context of World War II. The case of the Kurdish population entrapped within Iraq initiated a new debate about human rights, national sovereignty, and new modalities for international intervention. Massive upheaval in Kosovo and NATO's unprecedented level

Table 3
Principal locations of international refugees and asylum seekers, as of early 1999

Country of asylum	Place of origin	Total number
Jordan	Palestinians	1,463,000
Iran	Afghanistan	1,400,000
Pakistan	Afghanistan	1,200,000
Russian Federation	Former USSR	1,000,000
Jordan	Palestinians	800,000
Gaza Strip	Palestinians	773,000
West Bank	Palestinians	555,000
Syria	Palestinians	365,800
Guinea	Sierra Leone	350,000
Iran	Iraq	531,000
Lebanon	Palestinians	364,600
Thailand	Burma	350,000
China	Vietnam	281,000
Bangladesh	Pakistan (Biharis)	238,000
Armenia	Azerbaijan	218,000
Sudan	Eritrea	320,000
Yugoslavia	Croatia	279,000
Tanzania	Burundi	260,000
United States	El Salvador	242,500
Yugoslavia	Bosnia–Herzegovina	201,000
Azerbaijan	Armenia	180,000
Ethiopia	Somalia	180,000
Uganda	Sudan	170,000
Guinea	Liberia	160,000
United States	Guatemala	145,500
Congo–Kinshasa	Angola	140,000
Thailand	Burma	140,000
Zambia	Angola	140,000
Russian Federation	Former USSR	129,500
Côte d'Ivoire	Liberia	125,000
Saudi Arabia	Palestinians	122,500
Liberia	Sierra Leone	120,000
India	China (Tibet)	110,000
India	Sri Lanka	110,000
Nepal	Bhutan	100,000
Germany	Bosnia–Herzegovina	100,000

of intervention intensified the debate further. 'Safe havens' for Iraqi Kurds, 'zones of tranquility' for returning Afghan refugees, 'open relief centers' for would-be Sri Lankan refugees, and 'safe corridors' to Muslim enclaves in Bosnia exemplify 'new' spaces and discourses associated with refugees (Hyndman 1999).

Social scientists face major challenges and opportunities on many fronts. Longitudinal studies are needed to trace the refugee experience and associated processes through the 'full circle' of flight, resettlement, repatriation, and reintegration. The fourth stage of the refugee cycle—reintegration of former refugees after their return home – is especially in need of analysis. The use of remote sensing is a promising new research technique that geographers have begun em-

ploying to document displacement and to monitor environmental degradation around refugee sites (see Lodhi et al. 1998 and Bjorgo 2000). 'Refugeeism,' diaspora experiences, ethnic identity, de-territorialization, post-colonial expressions of place, and sense of home are important refugee related themes that social theorists in geography are well equipped to address.

The era of discrete refugee definitions is over. The traditional distinction between the refugee and voluntary migrant continues to be eroded in the face of new human dramas and political controversies. One need look no farther than the recent case of a Cuban six-year old rescued after his mother died in search of asylum or that of Chinese stowaways in container ships bound for America.

See also: Diaspora; Ethnic Groups/Ethnicity: Historical Aspects; International Migration by Ethnic Germans; International Migration by Ethnic South and Southeast Asians; Migration into the Americas and between American Countries; Migration: Sociological Aspects; Migrations, Colonizations, and Diasporas in Archaeology; Refugees in Anthropology

Bibliography

- Allen T, Morsink H (eds.) 1994 *When Refugees Go Home*. Africa World Press, Trenton, NJ
- Bascom J 1993 The peasant economy of refugee resettlement in eastern Sudan. *Annals of the Association of American Geographers* **83**: 320–46
- Bascom J 1998 *Losing Place: Refugee Populations and Rural Transformations in East Africa*. Berghahn Press, New York
- Bennett J 1998 Forced migration within national borders: The IDP agenda. *Forced Migration Review* **1**: 4–6
- Bjorgo E 2000 Using very high resolution multispectral sensor imagery to monitor refugee camps. *International Journal of Remote Sensing* **21**: 611–16
- Black R 1991 Refugees and displaced persons: Geographical perspectives and research directions. *Progress in Human Geography* **15**: 281–98
- Black R, Robinson V (eds.) 1993 *Geography and Refugees: Patterns and Processes of Change*. Belhaven Press, London
- Bulcha M 1988 *Flight and Integration: Causes of Mass Exodus from Ethiopia and Problems of Integration in the Sudan*. Scandinavian Institute of African Studies, Uppsala, Sweden
- Chambers R 1986 Hidden losers? The impact of rural refugee programs on poorer hosts. *International Migration Review* **20**: 245–63
- Cohen R, Deng F (eds.) 1998 *The Forsaken People: Case Studies of the Internally Displaced*. The Brookings Institution, Washington, DC
- Hamilton V 1999 *The 1999 World Refugee Survey*. US Committee for Refugees, Washington, DC
- Hampton J (ed.) 1998 *Internally Displaced People: A Global Survey*. Earthscan, London
- Harrell-Bond B 1986 *Imposing Aid: Emergency Assistance to Refugees*. Oxford University Press, London
- Hyndman J 1999 A post-cold war geography of forced migration in Kenya and Somalia. *The Professional Geographer* **51**: 104–14
- Kunz E 1973 The refugee in flight: Kinetic models and forms of displacement. *International Migration Review* **7**: 125–46
- Lodhi M, Echavarria F, Keithley C 1998 Using remote sensing data to monitor land cover changes near Afghan refugee camps in northern Pakistan. *GeoCarto International* **13**: 33–40
- Rogge J 1977 A geography of refugees: Some illustrations from Africa. *Professional Geographer* **29**: 186–9
- Rogge J 1985 *Too Many, Too Long: Sudan's Twenty-year Refugee Dilemma*. Rowman and Littlefield, Totowa, NJ
- Rogge J 1991 Repatriation of refugees: A not-so-simple 'optimum' solution. In: Allen T, Morsink H (eds.) *When Refugees Go Home*. Africa World Press, Trenton, NJ
- UNHCR 1993 *The State of the World's Refugees 1993: The Challenge of Protection*. Penguin, New York
- Wijbrandi J 1990 Organized and spontaneous settlement. In: Khulman T, Tieleman H (eds.) *Enduring Crisis: Refugee Problems in Eastern Sudan*. African Studies Centre, Lieden, The Netherlands

Wood W 1989 Long time coming: The repatriation of Afghan refugees. *Annals of the Association of American Geographers* **79**: 345–69

J. Bascom

Refugees in Anthropology

People being forced to leave their homes to settle in new environments is nothing new. However, 'refugees' as a legal category and as the object of academic research are more recent phenomena. This article will outline the development of refugee studies, in response to global refugee crises of the twentieth century. It will also present some key concepts and theoretical influences and discuss some of the challenges to anthropology of refugee studies.

1. The Emergence of 'Refugees'

The historical context in which refugees emerged as an internationally recognized category was provided by the massive displacements in Europe following two world wars. The Geneva Convention of 1951 is the critical landmark of the establishment of institutions and legal instruments to protect and assist such populations. A convention refugee is an individual who has crossed a national boundary and has a well-founded fear of persecution in the country of origin for reasons of race, religion, nationality, membership of a particular social group, or political opinion. Subsequently, other, regionally based legal conventions, notably in Africa and Latin America, have adopted more inclusive definitions of refugees which better correspond to the manifold types of mass displacements that often follow war and social disruptions in their regions today. Categories outside of legal provisions but equally of concern to agencies of protection are internally displaced people, people who flee within their own countries. Others are humanitarian refugees—like those offered temporary protection, these are persons who are not personally persecuted but are victims of war or war-related conditions. Another category of displaced people with important parallels to the experience of refugees, but without a global agency to protect their interests, are those forcibly relocated by their governments for the purpose of development. Such development-induced displacement may result, for instance, from the acquisition of land to make way for commercial forestry or the construction of a dam.

Refugee studies developed in response to the changing realities of the global refugee crisis in the early 1980s. The escalation of violent conflicts and mass disruptions in many areas of the world had then, within three years, dramatically doubled the global

refugee population. New increases followed the political changes at the ending of the Cold War. At the end of the twentieth century, large numbers of refugees from national conflicts and upheavals in postcolonial states and from crumbling multiethnic federations in Europe still dominated the asylum claims. According to estimates of the United Nations Commission for Refugees (UNHCR), there were 12 million recognized refugees in the world in 1999. Adding the millions of internally displaced people and asylum seekers (persons whose petitions for asylum are still pending), the figure was 19 million. The majority of refugees are found in the regions of the southern hemisphere, in particular in Africa and Asia, where many seek refuge in neighboring states. Only a trickle of displaced people make their way to the more affluent but more restrictive host countries of the north, in Europe or in North America. Even fewer are allowed to settle there.

Until the early 1980s, when refugee studies emerged as an academic field of study, with academic degrees and international journals, 'international migration' had been the main academic context in which research in the social sciences on refugees was usually found. The field is largely interdisciplinary, drawing together a range of research from the behavioral and social sciences, but with a strong presence since its inception of international law and anthropology. The interdisciplinary nature of the field has also meant the fruitful confrontation of anthropology with other theoretical and methodological approaches and the need to clarify the nature of its specific contribution. Furthermore, with its clear relevance for the acute issues of many refugee situations, much anthropological research in the field has a strong focus on practical application and advocacy. However, depicting the severing of a community and culture from their place, refugees have also implied a challenge to some basic theoretical assumptions in anthropology.

1.1 Who is a Refugee? Moving Beyond Legal Definitions

The answer to the question 'Who is a refugee?' is a crucial one in today's world. For those being classified, it may be a matter of life or death, because of the exclusive rights and entitlements of humanitarian concern and international protection that come with the status. In the academic field, the question is a continuing concern as a matter of theoretical and conceptual delineation of its object of study. The definition of 'refugees' as given in the Geneva Convention is too narrow, both as a protection instrument (many would argue), and as a theoretical concept. There is an agreement that, as a useful academic tool, the concept needs to take account of the many different situations which force people to leave their homes, and the proliferation of other legal statuses that have developed in different countries for people moving in

'refugee-like circumstances' (Escalona and Black 1995). Another basis of definition on more theoretical grounds has to do with the cause of flight. Today, causes are often complex, involving multiple factors and the blurring of distinctions, such as that between economic and political motives (for instance, economic collapse may often follow from political change). It is clear that, given the complexity, monocausal explanations are not very useful, because one status may blend into another. For instance, the Palestinian guest workers in Kuwait, once ousted from their homeland, were again displaced during the Gulf War in the early 1990s. Such explanations tend to be simplistic and obscure important similarities in the experiences of the victims of these forces (Shami 1994). A broader definition, well accepted in the field today and better attuned to current realities of displacement in the world, is to consider as refugees those people who are forced to abandon their countries of origin due to violent conflict or insupportable living conditions (Zolberg et al. 1989). These new realities have also questioned the relevance of differentiating between refugees and economic migrants. Rather than being seen as dichotomous, categories can be defined along a continuum, based on the degree of force involved. This would range from voluntary, such as labor migrants moving to better their situation, to involuntary, those forcibly expelled from their home areas (Hansen and Oliver-Smith 1982). While comparative approaches to kinds of migrations are increasingly emphasized by researchers in the field, so is the importance of maintaining that which is specific to refugees and other forced migrants, in terms of human rights and other humanitarian provisions made in their name (Shami 1994). From a somewhat different position, Elizabeth Colson makes a similar plea: 'Despite their similarities with other migrants, refugees have a unique status and force themselves upon the world's attention, because they emerge from and provoke acute crisis' (Colson 1987, p. 4).

The burning question of who and what a refugee is, is also a matter of 'the politics of space and otherness' (Gupta and Ferguson 1992). Answers are sought by examining the powerful institutional contexts and discourses, national and international, in which the category of 'refugee' is construed and public policy shaped, often in ways that misrepresent and exclude refugees. Further, eschewing overgeneralizations of 'the refugee experience' that have emerged out of such contexts (Malkki 1995a), anthropologists today would rather ask what it means to be a refugee, given the specific history of displaced groups and the circumstances in which they find themselves.

2. The Dynamics of Displacement

As refugees, people and their communities are not only forcibly removed, they are also remade. Forced

migration is an important aspect of social change, and any theoretical approach to refugees necessarily engages with process. While the study of refugees shares some common ground with international migration, it must also respond to the realities of sudden, involuntary change and the different forms of power which act on refugees' lives. For those displaced, change is often multifaceted and compressed in time, thus also providing a theoretical challenge to anthropologists, more used in the past to study change as slow and cumulative. However, emerging theoretical frameworks can draw insights from a growing body of research in anthropology on war and violent conflicts, social collapse and reconstruction. As displacement affects most aspects of people's lives, it addresses core issues in the anthropological study of communities, such as patterns of subsistence and livelihood, kinship and family, social and political organization, as well as the systems of cultural meaning which produce a sense of coherence and predictability in everyday life.

These issues are embedded in the wider dynamics of social and political change which form important contexts for the analysis of refugee communities. Within these contexts, anthropological case studies typically focus on uncovering the perspectives and strategies through which displaced groups deal with their predicament, renegotiating livelihoods, relations, and identities in interaction with their hosts. Anthropologists also examine the making of refugee communities, exile cultures, and identities in the places of settlement, including the dilemmas of those refugees whose attachments to the homeland and to fellow countrymen in other parts of the world remain stronger than to the society in which they reside (Fuglerud 1999). For a political movement such as the Chilean left, exiled in the 1970s, this may be manifested as a conflict between political ideology and the realities of everyday life, between 'life as it is' and 'life as it should be' (Eastmond 1997).

3. The Diversity of Refugee Situations and Research Emphases

The great diversity of conditions, in particular the marked contrast between southern and northern host states, has meant the development of different emphases in the concerns and theoretical approaches applied to the problem of refugees in different parts of the world.

3.1 Northern Perspectives: Refugees in Resettlement

Issues of resettlement and integration are common in studies of refugees in North America or Western Europe, reflecting the legal provisions for permanent residence of most refugees into these societies. Much research on resettlement in the USA, for instance, seems to focus on sociocultural adaptation, which is a

conventional framework in the study of the immigrant experience. However, the analyses also draw fruitfully on current theories of ethnicity and identity, emphasizing agency and cultural construction in the process (e.g., DeVoe 1992). These concerns reveal that many exile communities are concerned with continuity but also that, despite their claims, exile cultures typically represent dynamic forms. In these fields, American anthropologists, like other Westerners in the profession, increasingly turn to study the institutions of their own societies and their interaction with refugees. Being a refugee often means something different in Western Europe, where large influxes of newcomers are relatively recent and where integration is often preceded by a long asylum-seeking process. An emphasis in much of the current research is the vicissitudes of integration in relation to policies and processes of social exclusion. Importantly, perspectives from the inside of diasporas, groups which maintain strong ties to the home country and to members in other parts of the world, indicate that full integration into the host country may not be a primary concern (Fuglerud 1999). However, more systematic and comparative studies of the experience of different groups in different countries and circumstances of reception in northern countries are called for (Escalona and Black 1995).

3.2 Southern Frameworks: Refugees and Development

In southern countries, particularly in Africa, displacement and refugees have frequently been studied as a problem of development. A general framework for analyzing displacement and involuntary resettlement with a focus on the responses of dislocated people and the role of the state in resettlement is provided by Hansen and Oliver-Smith (1982). This has been an important starting point for some of the comparative work in southern countries.

One theme is that of development schemes and modernization displacing large numbers of people within their countries. An early contribution from anthropology is Colson's (1971) study of the social consequences of the relocation of the Gwembe Tonga due to the building of the Kariba Dam in 1958. More recent analyses take account of power relations and local forms of resistance by those ousted from their home areas (McDowell 1996). Another theme is the critical evaluation of humanitarian assistance and the potential of more long-term development that also benefits the host population. Harrell-Bond's (1986) illuminating study analyzed the implications of relief operations on refugees and their hosts in Uganda. Neglecting refugees' own concerns and resources, the interventions of relief agencies were found to create conflicts and fragmentation both within refugee collectivities and with the local host population. Other critical studies have focused on camps as organized

solutions to refugee crises. Comparing the experience of camp-settled to self-settled Angolan refugees in Zambia over many years, Hansen (1992) found a preference for staying with kin even in meager and less secure circumstances, rather than submitting to the rigidly controlled but well-provisioned camps. While camp refugees remained strangers in the new country and wanted to return, most of the others did not. These studies highlight the problem of defining assistance in narrow economic terms without taking account of the perspectives and resources of refugees themselves. Malkki's ethnography (1995b) of displaced Hutu refugees from Burundi settled in Tanzania compared identity and historical memory of refugees in camps with those who had settled in a nearby town and relied on their own creative strategies for survival. Whereas the latter were pragmatically juggling social identities, in the diversity typical of many African towns, camp refugees were passionately concerned with creating their mythico-history as members of the Hutu nation.

4. Refugees as People Out of Place: Challenges to Anthropology

Camps, as a form of containment, also reveal something about the workings of power within the international refugee regime and the deeply cultural yet global notions of order which underpin it. In a world where nation-states are seen as the 'natural' basis of belonging, according to Malkki, refugees occupy an in-between or liminal position and as such challenge the dominant 'national order of things.' Drawing on theories of social classification and of the cultural constructions of nation-ness, such analyses of displacement have opened up new theoretical spaces of anthropological inquiry into the relationship between people and place, nationalism and history. Inquiries into the significance of place and the creative ways in which people come to see different places as 'home' also provide important theoretical frameworks for the study of repatriation of refugees. This applies both to international policy, in which going home is construed as the naturally desired end of 'the refugee cycle,' and to return as lived experience. Images of home, which are a central premise in much exile ideology and upheld in memories and claims to the homeland, cannot always be reconciled with the realities of life upon return (Black and Koser 1999).

Clearly, refugees present a challenge to nation-states, but they do so also to the social sciences and some of their cherished concepts. In anthropology, until recently, refugees and displacements were rather marginal as research themes, in interesting ways reflecting the position of displaced populations themselves. This neglect may be attributed to a theoretical bias in anthropology that has tended to downplay violence and suffering (Harrell-Bond and Voutira

1992). Other commentators, like Malkki (1995a) refer to a 'sedentary bias' in anthropology, meaning that communities and cultures have been assumed to be naturally located in a particular place. Studies of displacement thus require anthropologists to re-examine their concepts of culture and community as bounded and territorialized units. Also, with the growing theoretical interest in anthropology in globalization and transnational phenomena, nationalism and conflict, violence and social memory, the study of refugees and displacement is gaining new relevance and fresh theoretical insights. Taken together, these new challenges will contribute to the coming of age of the study of refugees as an academic field in anthropology with a more solid theoretical foundation. If, as it is often claimed, refugees are symptomatic of the state of the world, with global mobility, fluidity, and fragmentation, but also claims and allegiances to particular lands and cultural identities, then they may well occupy a central position in future anthropology.

See also: Diaspora; Globalization, Anthropology of; Human Rights, Anthropology of; Human Rights, History of; Human Rights: Political Aspects; Refugees: Geographical Aspects

Bibliography

- Black R, Koser K 1999 *The End of the Refugee Cycle? Refugee Repatriation and Reconstruction*. Berghahn Books, Oxford, UK
- Colson E 1971 *The Social Consequences of Resettlement*. Manchester University Press, Manchester, UK
- Colson E 1987 Introduction: Migrants and their hosts. In: Colson E, Morgan S M (eds.) *People in Upheaval*. Center for Migration Studies, New York, pp. 11–16
- DeVoe P 1992 *Selected Papers on Refugee Issues*. Committee on Refugee Issues (CORI), American Anthropological Association, Washington, DC
- Eastmond M 1997 *The Dilemmas of Exile: Chilean Refugees in the United States*. Gothenburg Studies in Social Anthropology 13, Gothenburg, Sweden
- Escalona A, Black R 1995 Refugees in Western Europe: Bibliographic review and state of the art. *Journal of Refugee Studies* 8(4): 364–89
- Fuglerud O 1999 *Life on the Outside. The Tamil Diaspora and Long Distance Nationalism*. Pluto, London
- Gupta A, Ferguson J 1992 Beyond 'culture': Space, identity and the politics of difference. *Cultural Anthropology* 7(1): 6–23
- Hansen A 1992 Some insights on African refugees. In: DeVoe P (ed.) *Selected Papers on Refugee Issues*. Committee on Refugee Issues (CORI), American Anthropological Association, Washington, DC, pp. 100–9
- Hansen A, Oliver-Smith A (eds.) 1982 *Involuntary Migration and Resettlement: The Problems and Responses of Dislocated Peoples*. Westview Press, Boulder, CO
- Harrell-Bond B E 1986 *Imposing Aid: Emergency Assistance to Refugees*. Oxford University Press, Oxford, UK
- Harrell-Bond B E, Voutira E 1992 Anthropology and the study of refugees. *Anthropology Today* 8(4): 6–10

- Malkki L 1995a Refugees and exile: From 'refugee studies' to the national order of things. *Annual Review of Anthropology* **24**: 495–523
- Malkki L 1995b *The Purity of Exile. Violence, Memory, and National Cosmology Among Hutu Refugees in Tanzania*. University of Chicago Press, Chicago
- McDowell C 1996 *Understanding Impoverishment: The Consequences of Development-induced Displacement*. Berghahn Books, Oxford, UK
- Shami S 1994 *Population Displacement and Resettlement: Development and Conflict in the Middle East*. Center for Migration Studies, New York
- Zolberg A, Zuhrike A, Aguayo S 1989 *Escape from Violence, Conflict and the Refugee Crisis in the Developing World*. Oxford University Press, Oxford, UK

M. Eastmond

Region: Geographical

Region is a concept of great significance that has a long and storied history in geography. When the discipline was becoming an organized field of study during the late nineteenth and early twentieth centuries in Europe and North America, regions were its principal subject matter. Two conceptions of the geographical region vied for legitimacy. The region's importance was eclipsed during the 1960s by new systematic and quantitative approaches. Regional synthesis did not disappear, but ceased to provide geography's unifying framework. By the end of the twentieth century, geography lost confidence in spatial science and unifying concepts and frameworks of most sorts. The region was rehabilitated and enjoyed something of a vogue as geographers and others refocused attention on the nature and meaning of space and place.

1. Regions for Everyone

Region is a concept both familiar and useful to fields of knowledge ranging across the arts and sciences. Distinctive resource endowments, climatic activity, and plant and animal life, along with widespread patterns of human occupancy all differentiate the earth's surface. That differentiation is the basis for identifying and naming regions: alpine forest; coal fields; wine country. Areas of the human body, too, are commonly referred to as regions; one of the most frequent physiological uses of region is in reference to the human brain, with distinct regions said to control discrete neurological functions. The essence of regions is the areal continuity (whether on the earth, the human body, or some other object of study, such as a painting) of some specified facts of interest to the observer (whether it be agricultural practices, digestion, or

brush stroke). Regions of all sorts are identified on the basis of areally associated common characteristics.

Despite its widespread parlance, region enjoys a special relevance in geography because of the concept's dictionary definition of and popular association with the earth's surface. Regions, *ipso facto*, are geographical regions. As a field of knowledge widely perceived to study the earth's natural and human variation, geography is similarly understood to rightly 'own' regions as its logical and meaningful object of study. Geography certainly has devoted itself to the continuing transformation and differentiation of global nature and society. Yet the discipline has had a problematic relationship with the region: an on-again, off-again love affair with its putative object of study.

2. The Region in Geography

The region has been such a troubled concept owing to geography's hybrid status as, at once, a natural and social science. When the great disciplines were borne during the late nineteenth century, geography came into being with its shoulders rubbing kindred disciplines of geology and anthropology. Geology laid claim to the scientific study of the origin, history, and structure of the earth, while anthropology claimed an equally scientific status for its study of the origin, cultural development, and behavior of humans.

Geography was the study of the earth and life on earth, especially human life and its capacity to change the face of the earth. More than a little geology, a little anthropology, geography on the continent, British Isles and in North America identified itself as the study of the relationships and interactions between nature and culture. Halford Mackinder, a founder of British academic geography and great champion of the regional approach, lamented in 1887 'one of the greatest of all gaps ... between the natural sciences and the study of humanity.' He went on to insist that it was 'the duty of the geographer to build one bridge over [this] abyss ...' (cited in Livingstone 1992, p. 190).

This early history of the discipline bore direct consequences for the conceptualization of regions, for what better place, quite literally, to examine such connections than in areas ostensibly characterized by common natural and human features—which is to say, than in regions? Geography neither invented the region nor wrested it away from a rival discipline. Rather, the region fell into the discipline's lap as the obvious focus for its pursuits.

The study of regions seemed a natural fit with the new discipline of geography not merely by default, however. Geographers attracted to the study of what were then called 'man-land' relations tended to focus, if not lavish, their attention on rural areas in an often deliberate rural bias. Well into the twentieth century, most people around the world lived in rural areas or small towns and to some extent, surely, geography was

simply following demography in its preoccupation with the countryside. But more willful considerations were also at play. The countryside offered an ostensibly more transparent scrim for observing interactions between nature and culture: it was easier to parse a relationship between soils and climate on the one hand, and cultivation, settlement, and even religious practices on the other in the countryside than to try to make analogous sense of the nature–culture nexus in burgeoning cities of the day. Livingstone (1992, p. 264–6) further suggests that geographers seeking a moral compass for their work deliberately sought out rural areas because of the perceived superiority of rural values in a rapidly changing world. Anti-urbanism was a conscious strategy, often fraught with political overtones, as was the case with regional geography in Germany and France. From the start, regional and rural geography were synonymous.

Early in the twentieth century, when geography was being established as a university discipline in Western Europe and the United States, the study of regions took place in service to geography's larger purpose of bridging the natural and human sciences. Regions were thought to disclose nature–culture relations with quintessential clarity and were a basis for the most hotly contested debates of the age, namely over the relative weights of nature and culture in determining the character of people and places. Regional geography supported a variety of positions about the degree and direction of what was somewhat delicately called environmental influence. Taking cues from a stringent reading of Darwin's principles, Huntington (1924) found evidence of a determining environmental influence on the fate of world regions through the instrumentality of race and intelligence, a determinist position staked out earlier in the influential work of German geographer Ratzel (1896–8). Elsewhere, a more nuanced approach was developed in France by Vidal de la Blache (1926). 'To him, the regional articulation of life-style consisted in the material expressions of human–land relationships' (Livingstone 1992, p. 267) which he explored in a series of essays about French *pays*. Vidal emphasized society's variable role in modifying nature and was concerned with investigating how different natural realms gave rise to what he called *genres de vie*, or different ways of living.

2.1 The Region is Geography

A subtle shift took place in the early twentieth century in the status accorded geographical regions. Debates over environmental influence waned with the inevitable rejection of crass determinism and the emergence of a possibilist consensus built around *Lamarckian* notions of human adaptation as an active rather than passive response to nature. Under these changing conditions, regions were no longer the mere

handmaidens of research but came to be the discipline's *raison d'être*. Regions held a newly intrinsic interest. In Germany, Alfred Hettner promulgated '... the principle of *chorology*—the explanatory investigation of terrestrial reality divided into a series of component regions—' as the only way 'geography could retain both its physical and human components because it was precisely in their dialectic that particular places acquired regional character' (Livingstone 1992, p. 263).

In the United States, two very different regional concepts competed for acceptance and legitimacy. The first was contained in the work of Sauer and the so-called Berkeley school of (cultural) geography. Borrowing from Germanic anthropology, Sauer translated the anthropologist (and Berkeley colleague) Alfred Kroeber's 'culture area' into the geographers' 'cultural landscape'—synonymous with then-contemporary meanings of region. In Sauer's conception, regions existed as organic wholes, awaiting discovery and interpretation through intensive geographical field work. Also presumed was the inherent uniqueness of regions in terms of the presence and combination of natural and cultural features. The Sauerian tradition rejected grand theorizing in favor of a particularism that found perfect expression in richly textured and seemingly exhaustive cultural landscape studies (Sauer 1963).

The rival regional concept was associated with areal differentiation as enunciated by Hartshorne in his exegetical *The Nature of Geography*, published in 1939. *Contra*-Sauerian particularism, Hartshorne rejected the heretofore common notion that regions existed as organic wholes. Regions might be observed and perhaps even experienced in a harmonious, holistic fashion, but otherwise such holism was 'not inherent in the world which the geographer studies—neither in the world of nature nor in the actual world which nature and man together have made' (Hartshorne 1961, p. 362). While denying regions an independent ontological status, Hartshorne sought to guide their study on to firmer ground. '(F)or him, the regional project consisted in determining how "particular elements and complexes of elements are related to those in others"' (Livingstone 1992, p. 310)—such particulars including relief, vegetation, climate, land use, etc. Hartshorne called for a formal process of *areal differentiation* to define geographical regions based on one or another consciously chosen geographical particular. Such a scheme would lay the systematic basis for areal classification (e.g., the agricultural/climatic/religious/botanical regions of the world), comparison, and mapping.

Until the mid-twentieth century, these two regional concepts uncomfortably co-existed with each other. 'In the first (Hartshornian) view, regions were a form of areal classification, and in the second (Sauerian), regions were relatively unique areas of the earth that possessed a geographic individuality' (En-

trikin 1991, p. 16). Hartshornian regions enjoyed a wider purchase in the discipline because they insinuated themselves easily in the discipline's burgeoning systematic subfields of cultural, social, and economic geography.

By all accounts, during the 1940s and 1930s regional geography turned formulaic, with material synthesized and presented in a numbingly predictable format, beginning with the physical environment (landforms and climate) and then moving through several layers, as it were, of human occupance (settlement patterns, house types, farming practices, barn and fence types, etc.). While some scholarship rose above this narrow groove, sadly much did not. '(A)s clearly illustrated by many of the regional texts that appeared during those decades, the result was a rather tedious catalogue of factual material within an uninspiring framework' (Johnson 1991, p. 40). Like an overly tended hothouse flower, regional geography withered at the very moment it enjoyed the discipline's central place.

2.2 *The Region is Dead! Long Live the Region!*

By the 1960s, many geographers openly questioned whether the region could any longer provide an integrative framework of study. The criticism leveled at regional geography was often upsparring, as when Peter Gould characterized it as 'shabby, parochial, and unintelligent' (cited in Johnston 1991, p. 40) and easily caricatured. Gould's comments contributed to a wider assault on regional geography by a new generation of geographers seeking to redefine the discipline as a spatial science. Through the very real weaknesses of the traditional regional approach, and the sheer hubris of geography's so-called young turks (such as Gould), spatial analysis decisively eclipsed regional synthesis as the discipline's heart and soul. Suddenly the strenuous disagreements about particularism and areal classification sounded oddly antiquarian compared with the clarion and oh-so-modern call of the quantitative revolution.

Amidst this upheaval, regional geography did not vanish. Quantitative methods lent the Hartshornian region something of a new lease on life. Those working in the Sauerian tradition simply ignored the spatial science gold rush and continued advocating their cultural landscape approach in what became something of a lively, though curious, intellectual backwater. But the region generated little in the way of theoretical or methodological debate until the late 1970s. By that time, the euphoria surrounding spatial science had run its course and quantification was finally understood as a tool, albeit a powerful one, rather than geography's *raison d'être*. A new methodological pluralism overtook geography and the region once again began inspiring fierce discussion and debate—the news of its death having been vastly overstated.

The 1980s witnessed a resurgence of interest in regional geography emanating from two disparate sources (Pudup 1988). The first was practitioners of traditional regional geography who shared an unwavering belief that geography's *modus vivendi* remained regional synthesis. Emblematic of the revived tradition was the election of two of its chief contemporary practitioners, John Fraser Hart and Peirce Lewis, to the Presidency of the Association of American Geographers (AAG) during the 1980s. Both used their AAG presidential addresses to celebrate traditional regional approaches. Hart called regional geography 'the highest form of the geographer's art' and the discipline's long neglected unifying theme: 'I know of no other theme that is even remotely so satisfying as the idea of the region. All of the different strands of geography converge when we try to understand regions, and we need the concept of the region in order to understand why we need the diverse and variegated systematic sub-fields of geography.' (Hart 1982, p. 18). Three years later, Peirce Lewis echoed Hart's sentiment in another spirited defense of synthetic description which he considered 'a kind of rough and ready definition of geography: describing the earth's surface and trying to make sense of it.' (Lewis 1985, p. 47).

Such 'back to basics' calls could be dismissed as pure nostalgia were it not for the equally spirited discussions about regions arising from a very different quarter—geographers who sought to rehabilitate the concept and create a 'reconstructed regional geography.' The new regional scholarship was pioneered, perhaps unsurprisingly, by a generational changing of the guard. Skeptical about both spatial science and regional synthesis, geographers coming of age in the 1980s brought to the study of regions a deep engagement with theoretical debates and philosophical discussions ranging across the social sciences (Thrift 1983). They also brought to the study of regions a consciousness borne from being living witnesses to dramatic changes wrought by globalization (Massey 1995, Allen et al. 1998, Storper 1997).

The study of regions underwent a profound transformation. Regions ceased to be a topic interesting chiefly to geographers and historians and became an object of fascination in mainstream social sciences. Economists, political scientists, sociologists and, of course, geographers began asserting 'that the region might be a fundamental basis of economic and social life "after mass production."' That is, since new successful forms of production—different from the canonical mass production systems of the postwar period—were emerging in some regions and not others, and since they seemed to involve both localization and regional differences and specificities (institutional, technological), it followed that there might be something fundamental that linked late 20th century capitalism to regionalism and regionalization' (Storper 1997, p. 3). The apparent *fin de siècle* tendency

toward regionalization of economic and social life, in places like 'the third Italy,' Baden-Württemberg, and Silicon Valley, was wildly at odds with post-World War II era expectations of global integration and homogeneity. The global village seemed to be giving way to a world of regions.

3. The Regional World

Reconstructed regional geography radically departs from traditional regional synthesis. Gone are assumptions of organic holism and naturalism, or that geographers take unmediated cues from the physical environment in discovering neatly packaged regions. Rural bias has given way to a plurality of spatial possibilities, including rural and urban but also sub- and ex-urban regions, as in the case of 'edge cities' like Orange County, California. Economic processes of trade and investment are accorded far greater importance in regional transformation, if not central importance, in contrast to traditional regional geography's emphasis on culture and its tendency to either ignore or collapse most aspects of social life into the material expressions of culture. Indeed social relations, understood as 'always inevitably spatialized' (Allen et al. 1998, p. 138) take on an unprecedented significance in the new regional geography. Regions are socially constructed and open to deconstruction through social relations both in- and outside the region. The brave new regional world is an unstable one: '... that process of construction is constantly evolving. The spatialities of our lives are the product of continual negotiation, the outcome of the articulation of differentially powerful social relations.... Any settlement of social relations into a spatial form is likely to be temporary. Some settlements will be longer lasting than others' (Allen et al. 1998, p. 138). Finally, regions are also discursively produced through cultural representations that imbue meanings differentiating one region from another, e.g., a 'core' vs. a 'peripheral' region. 'What is said' about a region can powerfully shape the form and content of its social relations and what literally takes place there. Another obvious change is the new regional geography's nomenclature. Perhaps because 'region' carries such bruised and battered baggage, concepts like space, place, locality, and even territory are typically used interchangeably with region. The long-standing question of the region's appropriate geographical scale appears moot.

But not all the hoary questions have gone away. As the twentieth century turns into the twenty-first, geographers continue to wrestle with how to understand regional difference and uniqueness. That places differ from one another—and not just in their landforms and climate—can no longer be assumed, as was the earlier luxury, but has to be explained in a world governed by recognizably systemic processes of global

capitalism. In this regard, regions continue to be human places where geographers explore relations between the general and the particular. And because such questions by definition defy agreement, the regional enterprise is guaranteed a long and storied future.

See also: Cultural Geography; Geography; Globalization: Geographical Aspects; Local Economic Development; Local-Global Linkages: Environmental Aspects; Regional Geography; Rural Geography; Urban Geography

Bibliography

- Allen J, Massey D 1998 *Rethinking the Region*. Routledge, London
- Entrikin J N 1991 *The Betweenness of Place*. Johns Hopkins University Press, Baltimore, MD
- Hart J F 1982 The highest form of the geographer's art. Presidential address. *Annals of the Association of American Geographers* 72: 1–29
- Hartshorne R 1939 The nature of geography. *Annals of the Association of American Geographers* 29: 1–469 (1961 *The Nature of Geography*. Edwards, Ann Arbor MI)
- Huntington E 1924 *The Character of Races, as Influenced by Physical Environment. Natural Selection and Historical Development*. C Scribners Sons, New York
- Johnson R J 1991 *A Question of Place*. Blackwell, Oxford, UK
- Lewis P F 1985 Beyond description. *Annals of the Association of American Geographers* 75: 465–78
- Livingstone D N 1993 *The Geographical Tradition*. Blackwell, Oxford, UK
- Massey D 1995 *Spatial Divisions of Labour: Social Structures and the Geography of Production*, 2nd edn. Routledge, New York
- Pudup M 1988 Arguments within regional geography. *Progress in Human Geography* 12: 369–90
- Ratzel F 1896–8 *The History of Mankind*. Macmillan, London
- Sauer C O 1963 *Land and Life: A Selection from the Writings of Carl Ortwin Sauer*. University of California Press, Berkeley, CA
- Storper M 1997 *The Regional World*. Guilford Press, New York
- Thrift N 1983 On the determination of social action in space and time. *Society and Space* 1: 23–57
- Vidal de la Blache P 1926 *Principles of Human Geography*. Lowe and Brydon, London

M. B. Pudup

Regional Geography

1. Introduction

Regional geography is a field of study which strongly contrasts with systematic geography. While the latter is concerned with the distribution of certain characteristics, such as landforms, climate, vegetation, and

human activities, over the surface of the earth, regional geography looks into the relationships these phenomena have within the different regions. The prime interest in the systematic fields of geography is to describe the patterns certain phenomena display over the entire globe and to elucidate the (general) rules that govern these patterns. Regional geography, on the other hand—at least in the mainstream thinking of the nineteenth and early twentieth centuries—is not interested in general rules but, on the contrary, in the singularity or the specificity of regions. Thus, traditional regional geography is characterized by a synthetic view of the world, i.e., it tries to put things together in a regional context, and by the (epistemological) conviction that it needs to view the world as a mosaic of idiosyncratic, singular entities—the regions.

This understanding is deeply rooted in most people's everyday consciousness of the ways in which the world is structured. Thus, when we think of the American Midwest we conjure up images of commercial agriculture, endless cornfields, livestock production, and a predominance of white middle-class Protestants, whereas when we think of the Sahara or Central Australia we probably imagine little vegetation, heat, and dryness, and an environment hostile to most life-forms. The fact that people tend to be regional geographers in their everyday consciousness may be attributable to the fact that most of them were taught geography at school. It can also be seen as a strategy of the individual to simplify a complex world by structuring it along the lines of regional entities. Whichever the reason, the fact that we do tend to think in idiosyncratic ways helps to explain the lasting success of regional geography—at least in the form of regional monographs. In our age of extensive travel over the entire globe these monographs are useful backdrops for our encounters with strange environments, since they allow us quickly to attune to (rapidly) varying circumstances.

As an academic endeavor, determining geographers' shared conviction as to what constitutes the essence of the discipline, regional geography was not a lasting success story. However, after a period of strong rejection in the 1960s and 1970s there were signs of a renaissance in the last decades of the twentieth century. The following sections of this entry deal with the mixed press regional geography encountered in the past, with recent developments, and with the—possible—future(s).

2. Historical Development

Although regional geography can be traced back to ancient Greece, the 'classical' period was that of the late nineteenth and early twentieth centuries. The external factors explaining the success of geography as

a discipline and of regional geography as its center at the time were the rapid exploration (and exploitation) of the interior of continents, the setting up of a regular or more efficient administration in many countries, and the stronger ties that were forged between countries and places. All this necessitated the generation and dissemination of (geographical) knowledge, much of which was of a place-specific nature. Before geography was set up as a university discipline later in the nineteenth century, naturalists such as Alexander von Humboldt (together with Aimé Bonpland) ventured to study hitherto unexplored parts of the world. Von Humboldt supported a holistic understanding of (his) academic explorations: in his extensive research papers on South America, which covered more than 9,000 pages, for instance, natural and human phenomena were seen in context.

Thus, regional geography is associated with the early beginnings of geography as a university discipline and it was, in the case of von Humboldt, informed by the Enlightenment, as von Humboldt concerned himself critically with the situation in the South American colonies. However, later in the nineteenth century a holistic understanding of the (natural) sciences became increasingly marginal because they were undergoing a process of continuing specialization. In geography, on the other hand, the holistic approach experienced a sustained success as it influenced much of the mainstream methodological thinking well into the twentieth century.

In the German-speaking countries the Austrian Viktor Kraft (1929) offered his very influential epistemological analysis of the object, the aims, and the methods of geography. Following a neo-Kantian line of argument Kraft subdivided the discipline into 'general,' i.e., systematic, geography (as a nomothetic science), and regional geography ('Länderkunde'—as an idiographic science). In line with this view, it was maintained that the study of regions is the prime object of geography (an object truly exclusive to the discipline), that therefore regional geography is at the core of the discipline, and that, consequently, systematic geography and its branches are ancillary sciences. Moreover, regional geography had an integrating role to play as it incorporated physical as well as human geography under one epistemological roof. The fact that regional geography thus gave unity and an exclusive object to geography explains why it remained the hegemonic methodological discourse well into the twentieth century.

In Germany, Hettner (1927), amongst others, had successfully codified regional geography along the lines of Kraft's neo-Kantian epistemological reflections so that it occupied a pivotal position within geography from the 1920s onwards. In the USA, Richard Hartshorne's *The Nature of Geography* (1939) gave a similar importance to regional geography. The view that regional geography had an overarching, integrating role to play, however, came under in-

creasing attack in the 1950s and 1960s when the spread of an analytical-pragmatic and nomothetic paradigm reached the community of geographers.

In essence, practitioners of regional geography were accused of unscientific working practices, of gathering—and not analyzing—facts, and of emphasizing the ‘individuality’ of spatial units, the regions. The critics whose discourse on the philosophy of science was informed by logical positivism (cf. Harvey 1969) and critical rationalism pointed out that regional geography was lacking a theoretical foundation on which an explanation of empirical evidence could be grounded. Part of the criticism was directed at what the critics regarded to be a premodern frame of mind of regional geographers, who saw the world as a well-ordered mosaic (of regions) which the regional geographer’s ‘art’ had to put together, at the implicit geo-determinism and at the exclusion of social practices from academic work.

By the 1960s and early 1970s a nomothetic discourse had asserted itself in geography which was to form the mainstream methodological underpinning of the discipline for the next decades (Hanson 1999). As a consequence, from then on regional geography only played a minor role until the last decades of the twentieth century. This minor role, however, mainly expressed itself in the low esteem that regional geography enjoyed in the eyes of those geographers whose frame of mind was shaped by an understanding of academic work deeply rooted in logical positivism and critical rationalism. This state of affairs was sharply contrasted with the fact that at the same time many regional studies were being published (Gregory 1988). This is why it is not possible to infer from the (then) dominant paradigmatic understanding of the discipline the performance and the relevance of regional geography in the wider society.

The peripheral position of regional geography in academia during the 1970s did not change, despite the attempts of individuals to put it on a (new) epistemological foundation. In Germany, Wirth (1978), for example, in his endeavor to resurrect regional geography takes up certain key elements of critical rationalism as developed by Popper. According to Wirth, regional geography should be treated as a historical science which deduces singular events from general laws (which in the case of regional geography can be taken from the systematic fields of geography) and special antecedent conditions (which result from the unique situation under investigation). This approach will not only render a description but also provide the academic with an *explanation* of empirical phenomena. This view challenged other academics, notably Bahrenberg (1979). The dispute that evolved between the two antagonists did not, however, have a widespread impact. The fact that a broad discussion on the significance of regional geography did not take place at the time may have its roots in the repercussions of the ‘quantitative and theoretical revolution’ of the

1960s and 1970s, as well as in the growing inner diversity of the discipline as a whole.

3. Recent Trends: The Emergence of a ‘New Regional Geography’

In the 1980s there was a development, primarily taking place in the French- and English-speaking countries and in The Netherlands, which can be characterized as a trend back to the study of the specific, a renaissance of regional (human) geography (Gilbert 1988, Pudup 1988). This tendency can be seen as a reflection of a number of developments.

3.1 The Background

To begin with, a positivist mode of explanation in general and spatial analysis in particular were seen to be insufficient tools for geographical studies. The criticism leveled at the mainstream geography of the 1960s and 1970s was that it neglected humans as intentionally acting beings (‘structure–agency–debate’) and that it tended to reduce space to an entirely social construct (‘spatial exorcism’). Second, ‘space’ and ‘place’—key geographical concepts—gained in importance in the neighboring disciplines. This had repercussions on geography. Now geographers were able to point out that distinguished scholars, such as the sociologist A. Giddens, were borrowing from geography. And third, in the light of the crisis of the ‘grand narratives’ of humankind, such as religion, communism, science, etc. (and of ‘grand theories’ in the academic world), in the age of postmodernity there was an overall social trend which gave more importance to the special/specific (ideographic), and which acknowledged a greater diversity in regional cultures, ways of life, etc. Hence, a totalizing view as to the ways in which societies, economies, cultures, etc. should be seen and understood came under increasing pressure.

Regional specificity had reinstated itself on the methodological agenda because it was recognized by a growing number of academics that, as Massey (1984, p. 120) points out,

just as there are no purely spatial processes, neither are there any non-spatial social processes. Nothing much happens, bar angels dancing, on the head of a pin. ... The fact that processes take place over space, the facts of distance and closeness, of geographical variation between areas, of the individual character and meaning of specific places and regions—all these are essentials to the operation of social processes themselves.

3.2 The Different Approaches

According to Gilbert (1988) there are at least three different approaches dealing with regional specificity

that can be distinguished in the recent endeavor to promote a 'new regional geography': the political-economic, the humanistic-phenomenological, and the structurationist.

The political-economic approach is very much inspired by the tradition of dialectical materialism in Marxist geography. It focuses on the ways in which (global) mechanisms of capitalism take effect on the ground, in particular on the ways in which the division of labor is spatially structured in capitalist societies. This understanding of the spatial operation of capitalism is in essence based on two main premises: it rejects the notion of a linear realization of systemic processes in different spatial contexts ('regions') and it assumes a dialectical relationship between the systemic framework (of economic development) and the locally and regionally specific development trajectories. Hence, regions are not just spatial backdrops against which overall developments take place, they are the specific sociospatial, historically contingent context ('layer of history,' Massey 1984) which determines the specific pathways of capitalist development in the varying particular and the overall contexts.

According to Gilbert (1988, p. 219) the term 'region' is given a substantially different meaning and therefore a 'new regional geography' differs markedly from the traditional understanding: 'The traditional concern with the people/nature relation has widened to include society as the prime agent in region formation. Thus, the substance of regional geography has become the triangular relations between people, society and nature.'

The humanistic-phenomenological approach can best be understood as a multilevel change of perspective. It is a break with the established positivist methodology and it marks a different approach of the academic toward social reality. In their critique, Van der Laan and Piersma (1982, p. 416f.) emphasize the reductionism of quantitative geography which in essence meant that human beings were treated as 'things, as non-intentional materialistic research objects.' As a consequence—and in direct opposition to the positivist notion of *explanation*—the proponents of a humanistic approach place the intentionally acting human being at the center of academic inquiry. Furthermore, academics are to discard their 'classical' perspective as outsiders and to assume an insider's view of social reality in the sense of comprehending, *understanding* the meaning of social acts and the intentions of agents (Buttimer 1979). Only through this shift will social practice be rendered intelligible.

The humanistic approach looks into the cultural meaning of the 'region' which is being conceived of as the expression of a specific—culturally embedded—relationship between people and their spatial environment. The distinguishing quality of this relationship is the awareness of a shared culture among the population, and the associated distinct identity of that population as it is noticeable to outsiders. This

relationship also encompasses the resultant feeling within the population of belonging to 'their' region.

According to the programmatic considerations of its proponents, a humanistic-phenomenological approach has as its focal point individuals, their everyday view of the world, and the meaning which they attribute to their environment. However, the individual is not seen as a solitary agent but as a carrier of a shared cultural identity.

The most striking characteristic of the structurationist perspective, which is mainly based on Giddens's (1988) theory of structuration, is the dialectical understanding of the 'region' as the outcome and the producer of social practice. The 'region' is regarded as a specific spatial and temporal constellation of social relationships, and at the same time as a medium for these social relations. Pohl (1993) very aptly speaks of a 'co-reacting catalyst.' Behind this view is a comprehensive definition of the 'cultural region' in which elements of the political-economic approach are also incorporated.

3.3 *Their Conceptual Connections*

These different conceptional approaches do not in any way form coherent edifices of knowledge, nor are they as clearly demarcated as these few observations may suggest. Additionally, they are not at all canonized by the scientific community of geographers, particularly outside the French- and English-speaking countries. Notwithstanding these qualifications, there are a number of issues which link the different approaches on a conceptional level: it is the understanding of the term 'region' and of the central epistemological dimensions 'individual' and 'history.'

One of the outstanding arguments in the discussion is the ontological status being ascribed to the 'region.' According to a number of proponents of a new regional geography, regions are not only empirically important—as analytically relevant constructions of the academic—they are also the outcome of human history and therefore of an ontological nature (Pickles 1986). This is why the individual and history are at the core of a new, reconstituted regional geography. Another point which many advocates of a new regional geography agree with is the assumption that 'regions' are a persistent feature of societies. Contrary to the modernist conviction that regional differences will be eroded in the process of social modernization, these differences are assumed to remain, as spatial differentiation is seen as a central element of society (in the sense of social relations being spatially and temporally contingent phenomena). Such a notion of the 'region,' however, has little in common with the chorological view according to which the 'region' is an expression of the opposition against (or the inertia inherent in) the homogenizing trends of modernization. The third notion which runs through the debate

is the structural analogy of the three approaches. What is meant here is the mentioned dialectical concept of the regions being both the result and the 'medium' of social relations. As a co-reacting catalyst, the region has emergent powers as 'social relations within the region develop because of the specific way individuals and groups relate within the particular regional space' (Gilbert 1988, p. 215).

One distinctive feature of the debate about a new regional geography is the fact that it has a heavy theoretical and conceptual slant, whereas empirical studies are scarce. The existent empirical work does not exhibit a uniform approach toward method, but rather a certain methodical pluralism or voluntarism. There are basically two empirical approaches. Whereas one group of academics works historically by reinterpreting regional developments which have taken place (cf. for instance Gregory's outstanding work *Regional Transformation and Industrial Revolution: A Geography of the Yorkshire Woollen Industry*, 1982), the focus of the other group ('locality research'; cf. Cooke 1989, Harloe et al. 1990) lies on the present and on the examination of the consequences of (global) economic restructuring on 'localities.'

3.4 Unresolved Issues

After a fairly intensive (conceptual) debate in the 1980s and early 1990s the interest in a reconstituted regional geography waned somewhat in the late 1990s. The reasons behind this trend may have been a certain fatigue after a period of intensive debate, the fact that many geographers who have migrated to other fields of interest in the discipline do have strong reservations about (prejudices against) 'regional geography,' and the problematic or unresolved issues of the debate. All three are inter-related in various ways, and they need to be taken seriously if regional geography is to have a future as a stimulating, academically rewarding pursuit.

The interest of geographers—like that of other academics—usually is arrested by methodological debates which have the potential to renew the cognitive and the social system of which academics are members. However, fatigue sets in if these debates fail to live up to the expectations of the scientific community. In the case of a new regional geography this may be due to the fact that the promise of modernization was not convincing enough, both to those who took part and to the grant-giving bodies (who are operating on a system of academic self-control; thus colleagues with reservations about regional geography decide on submitted research proposals). As the new methodological direction is characterized not by one but by several approaches, it is hard to find sufficient collective support within the scientific community of geographers. What is more, despite their innovative

and stimulating new ideas, these approaches have inherent shortcomings which weaken the case for a new regional geography.

To begin with, there is, typical for a 'young' paradigm, the need to clarify the usage and meaning of key concepts and terms. This is particularly so in the case of the 'locality' debate, in which the usage of the term 'locality' leaves much to be desired. Most importantly, the usage of that term reproduces the casual handling of concepts that was typical of the 'old school' and that had received so much—deserved—criticism. Terminological constructs are taken as starting points without confronting these with the terms relevant for both the life world ('Lebenswelt') and the social systems. It was not the process and the outcome of the socially constructed spatial coding which was on the agenda but rather the examination of social interaction within the boundaries of the previously demarcated study areas ('localities'). This, however, bears the danger of reifying (predetermined) spatial concepts and thus of wasting a chance for epistemological progress.

This problem refers to another unresolved issue: the inadequate interaction between theory and empirical evidence. In the case of the locality studies, this becomes apparent due to the exclusive focus on the level of the locality and the exclusion of national or global economic development trajectories and their analytical framework. Because a new regional geography—rightly—assumes that phenomena cannot be explained by themselves, a stronger effort needs to be made to interrelate empirical description and theoretical abstraction.

The last point to be discussed here is as fundamental as the previous issues. What are the implications of a regional geography which is conceived of as a social science, thus relinquishing the former claim of giving unity to geography (by integrating physical and human geography)? Will a social scientific understanding of regional geography therefore lead to a loss of disciplinary unity? There can be little doubt that this issue is an important one as geographers have to compete with academics from other disciplines for funds, etc. Showing disciplinary unity can help to underscore the importance of the discipline and thereby demonstrate the importance of academic subjects as social systems which can effectively organize funds, support, and the like. However, as a cognitive system, geography has to approach the question of disciplinary unity via a different route. Unity is no value in itself, of course, so the issue at stake is whether there are social developments which suggest a holistic approach. One key area to be dealt with in such a way is the crisis of late modern societies effected by ecological and other human-made problems. Because of the complexity of these issues a comprehensive approach has to be taken in order to resolve the problems involved. Serious as these matters may be, they offer a chance for academics to cooperate

in various ways: amongst colleagues of the same subject, on an interdisciplinary level, and with the wider public. Only a concerted effort can help to come to grips with the numerous dangers inherent in the modern world we live in and which we helped to create.

4. *The Prospects of Regional Geography*

The future of geography as a whole and of regional geography as one of its branches will very much hinge on the ability of geographers to contribute to these and other socially relevant issues. Such an argument is based on pragmatic and normative considerations and not, as was the case with traditional regional geography, on the methodological neo-Kantian assumptions outlined above (the latter, however, having the normative implication as to how the subject, the methods, etc. of regional geography should be seen).

After the paradigmatic shifts in the twentieth century which first brought traditional regional geography into disrepute, and then questioned the hegemony of a positivist mode of explanation ('spatial analysis'), we now live in an era where the grand, all-encompassing theories have lost their authority and where different analytical frameworks (methodologies) coexist. It has become clear that there is not *one* way of coming to grips with the complexity academics (have to) deal with. It is understandable that the proponents of a positivist mode of explanation, in their wish to be accepted as equals by their colleagues from other (positivist) disciplines, broke with the regional tradition in the 1950s and 1960s. However, from the vantage point of the early twenty-first century, this development can now be seen as a certain stage in the development of the discipline which has evolved as the disciplinary discourse progressed. As Hanson (1999, p. 136) points out, there is now

a greater willingness to acknowledge complexity and to engage multiple approaches to understand that complexity. There is now a greater acceptance of the notion that people create knowledge rather than discovering objective truths ... and that taking advantage of what is useful in each of a number of approaches can yield greater understanding than does adhering only to one.

What are the implications of these considerations for the future of regional geography?

Taking up the line of argument presented by Hanson, an enlightened methodological understanding of a modern regional geography cannot afford to dismiss lightly any conceptual framework, as different perspectives may yield different facets of knowledge. If regional geography is to be taken seriously as a competitive academic subject, it needs to engage in the methodically controlled generation of knowledge, and it needs to do that by probing the relevance of different methodological approaches. This is no easy task as different understandings of what science is (or should

be) about rest upon different epistemological and methodological assumptions, for instance on contrasting concepts of knowledge or causality. But the options are much less attractive. Additionally, it needs to be pointed out that along with such a postmodern methodological understanding comes the inherent danger of warranting an 'anything goes' attitude (including a regional geography without or with implicit theory). A competitive regional geography, however, needs to take on board an explicit reference to theory, and it needs to take seriously the reciprocal relationship between theory and empirical evidence, as this is an important mechanism for expanding the theoretical foundation of the discipline.

At this point the above-mentioned pragmatic and normative considerations need to be reconsidered again. This is so because any definition of competitive research would be incomplete if it did not reflect the social contexts in which research agendas are being drawn up or in which research results are being applied. It is important that the value-free stance of traditional regional geography is given up because it would be a self-deception to believe that regional geography can do without explicitly referring to or incorporating normative issues. As academic disciplines are an intrinsic part of society, they need to reflect the origin and the legitimacy of research topics and the use of research results. But which questions are to be raised and which problems are to be tackled by regional geographers? A systematic discussion of this issue has not yet taken place. However, it needs to be addressed in the context of a concerted effort to bring to life a 'new' new regional geography as a theoretically informed empirical project (Thrift 1998). The debate of the 1980s and 1990s broke new ground and generated a number of stimulating (conceptual) ideas from where the debate could be taken.

See also: Globalization: Geographical Aspects; Region: Geographical; Regional Integration

Bibliography

- Bahrenberg G 1979 Anmerkungen zu E. Wirths vergeblichem Versuch einer wissenschaftstheoretischen Begründung der Länderkunde. [Some remarks on E. Wirths failed attempt to justify Lauderkunde Epistemologically.] *Geographische Zeitschrift* 67: 147–57
- Bassand M, Guindani S 1983 Maldéveloppement régional et luttes identitaires. *Espaces et Sociétés* 42: 13–26
- Buttimer A 1979 'Insiders', 'outsiders' and the geography of regional life. In: Kuklinski A, Kultalahti O, Koskiahio B (eds.) *Regional Dynamics of Socioeconomic Change*. Finn publishers, Tampere, Finland, pp. 157–78
- Claval P 1998 *An Introduction to Regional Geography*. Blackwell, Oxford, UK
- Cooke P (ed.) 1989 *Localities. The Changing Face of Urban Britain*. Unwin Hyman, London

- Giddens A 1988 *Die Konstitution der Gesellschaft. Grundzüge einer Theorie der Strukturierung*. Campus, Frankfurt, Germany
- Giddens A 1991 *Modernity and Self-identity: Self and Society in the Late Modern Age*. Stanford University Press, Stanford, CA
- Gilbert A 1988 The new regional geography in English and French-speaking countries. *Progress in Human Geography* 12: 208–28
- Gregory D 1982 *Regional Transformation and Industrial Revolution: A Geography of the Yorkshire Woollen Industry*. MacMillan, Basingstoke, UK
- Gregory D 1988 Regional geography. In: Johnston R J, Gregory D, Smith D M (eds.) *The Dictionary of Human Geography*. Blackwell, Oxford, UK, p. 293
- Hanson S 1999 Isms and schisms: Healing the rift between the nature-society and space-society traditions in human geography. *Annals of the Association of American Geographers* 89: 133–43
- Harloe M, Pickvance C G, Urry J (eds.) 1990 *Place, Policy and Politics: Do Localities Matter?* Unwin Hyman, London
- Hartshorne R 1939 *The Nature of Geography: A Critical Survey of Current Thought in the Light of the Past*. Association of American Geographers, Lancaster, PA
- Harvey D 1969 *Explanation in Geography*. Edward Arnold, London
- Hauer J, Hoekveld G 1993 Introduction: Why regional human geography? In: Hauer J, Hoekveld G (eds.) *Moving Regions*. Koninklijk Nederlands Aardrijks-Kundig Genootschap, Utrecht, The Netherlands, pp. 13–24
- Hettner A 1927 *Die Geographie. Ihre Geschichte, ihr Wesen und ihre Methoden*. F Hirt, Breslau, Poland
- Holmen H 1995 What's new and what's regional in the 'new regional geography'? *Geografiska Annaler, Series B* 77: 47–63
- Johnston R J, Hoekveld G A, Hauer J 1990 *Regional Geography. Current Developments and Future Prospects*. Routledge, London
- Kraft V 1929 Die Geographie als Wissenschaft. In: Kraft V, Lampe F (eds.) *Methodenlehre der Geographie*. Franz Deuticke, Leipzig, Germany, pp. 1–22
- Massey D 1984 *Spatial Divisions of Labor. Social Structures and the Geography of Production*. MacMillan, Basingstoke, UK
- Pickles J 1986 Geographical theory and educating for democracy. *Antipode* 18: 136–54
- Pohl J 1993 Kann es eine Geographie ohne Raum geben? *Erdkunde* 47: 255–66
- Pudup M B 1988 Arguments within regional geography. *Progress in Human Geography* 12: 369–90
- Sayer A 1989 The 'new' regional geography and the problems of narrative. *Environment and Planning D: Society and Space* 7: 253–76
- Smith N 1984 *Uneven Development: Nature, Capital and the Production of Space*. Blackwell, New York
- Swyngedouw E A 1989 The heart of the place: the resurrection of locality in an age of hyperspace. *Geografiska Annaler, B* 71: 31–42
- Thrift N 1998 Towards a new new regional geography. *Berichte zur deutschen Landeskunde* 37–46
- Van der Laan L, Piersma A 1982 The image of man: paradigmatic cornerstone in human geography. *Annals of the Association of American Geographers* 72: 411–26
- Villeneuve P 1986 Les théories sur la région: Polarisation et capillarité. In: De Koninck R, Landry L (eds.) *Les genres de vie urbains: essais exploratoires*. Université Laval, PQ, Canada, pp. 17–33

Wirth E 1978 Zur wissenschaftstheoretischen Problematik der Länderkunde. [Towards an Epistemological foundation of regional geography.] *Geographische Zeitschrift* 66: 241–61

G. Wood

Regional Government

Regional government analysis has traditionally been restricted to state politics in federal systems or to center-periphery relations in the integration of the nation-state. We focus in this entry on Western Europe and North America to show: (a) that the political and legal status of regional institutions is extremely diverse; (b) that the rationales for regionalization have to be referred to state traditions; (c) that these rationales converge to develop both regional politics and public policies; and (d) that transnational integration gives a new relevance and meaning to regionalism.

1. The Ambiguous Status of Regional Government

Regional government is among the topics in comparative politics that clearly reveal the contingency of political institutions and the diversity of political processes behind them. Some countries exhibit very strong subnational governments, such as states in federal political systems (e.g., the USA or Switzerland), while others do not have any regional institutions at all (in centralized states such as Greece or France until the Fifth Republic). When such institutions do exist they may range from weak administrative units without much social significance to real representative institutions with powerful political capacities. To enhance the complexity of the picture, regions often coexist with other subregional units (counties, provinces, and *départements*), and display very different legal and fiscal authorities over the lower levels of territorial governments.

This variance suggests that the analysis of regional government is anything but simple, and that the idea of a universal form or of a common feature of regional institutions, even in the limited area of Western democracies, is misleading. Regional government cannot be analyzed independently of the political context within which it is taking place. It has to be referred to the general process of state building and state tradition, and understood as the product of different trends of regionalization.

In a restrictive meaning, regional government can be defined as the activity of political authorities specifically designed as an intermediary level between the state and local governments, and grounded in legitimacy vested in regional forms of political rep-

resentation. Therefore, it involves the scale of political authority (the level where sovereignty is defined and bounded), the aims of government (whether it takes care of regional issues or not), and the tools used to fulfil these goals (centralized vs. decentralized means of authority). It is conceivable to observe a government of regional issues without corresponding mechanisms of political representation at the regional level (e.g., with the Scottish National Office before the devolution reforms in the United Kingdom or the regional *Préfets* before the decentralization reform in France). Nevertheless, most Western democracies have evolved since the 1980s towards a greater empowerment of regional levels of political representation (most dramatically in Spain, France, Belgium, Italy, Canada, and more recently the UK), while federal systems turned to more interdependent patterns of relations between states and federal governments (notably in Germany, the USA, and Switzerland). In a more extensive definition, it is therefore more appropriate to conceive regional government as a set of complex institutional interactions and social mobilizations around political representation and public policies defined at an intermediary level between the state and local governments. If all Western democracies do not display distinctive political institutions at the regional level, all of them nevertheless show strong tendencies towards social and political mobilizations, and adjustments in institutional settings at this mesolevel of government (Sharpe 1993). Thus, this extensive definition is more appropriate to capture the actual issues of regional government in its contemporary context.

2. State Traditions and Rationales for Regionalization

From this definition we can infer that different state traditions, that is, differences in historical paths of state building and related administrative cultures, will matter in defining regional government. In exploring this topic, Loughlin and Peters have identified four main state traditions. (i) The Anglo-Saxon tradition (including countries such as the USA, the UK until 1998, Ireland, and Canada with the exception of Quebec), where there is no (in the UK) or a limited (in the USA) legal basis for the state, and where regional government can be either the 'state power' of the federal system (the USA), or shadowed by the strength of local government (in the UK). (ii) The Germanic tradition (covering countries such as Germany, Austria, Spain after 1978, and Belgium after 1988), rooted in a strong legal status of a Weberian state and a co-operative federalism giving large powers to regional government. (iii) The French State tradition, which is Jacobean and centralist, and where regions appear as technocratic creations progressively building a regionalized unitary state (for example, France,

Spain until 1988, Portugal, Greece, Belgium until 1988, and Quebec). (iv) Finally, the Scandinavian state tradition of Sweden, Norway, and Denmark, which, like the Germanic one, is rooted in a strong legal basis of the state, but in which the territorial organization is based more on a strong local autonomy than on the development of regional government (Loughlin and Peters 1997).

State tradition conceptualized in this way does not explain everything regarding regional government. For instance, there are strong discrepancies in the Anglo-Saxon model between European and American countries, and The Netherlands appears as an exception in the Germanic tradition. But it helps nevertheless to understand the different territorial policy paradigms (Balme et al. 1994) and political rationales for regionalization, understood as the forces driving the creation and development of regional government.

The first of these rationales is strictly political and linked to the integration of the nation-state. Regionalism here is close to nationalism, and regional government can be seen either as a tool to sustain the integrity of the state or as the 'slippery slope' leading first to larger claims to regional autonomy and ultimately to separatism and independence. This issue is dependent on social and territorial contiguity, on the degree of economic integration of regions to the country, and on the use of political violence between protagonists. However, the meaning of institutional development at the regional level will be different in an Anglo-Saxon context, where it appears as explicitly political with devolution (Elcock and Keating 1998), from a continental one, where regional decentralization can also be justified by administrative arguments and in this respect be more easily acceptable.

The second form of regionalization is federalism. Here the rationale is not so much political as constitutional-legal. It refers not to its integrity or to the coexistence between different cultural communities as in the previous case, but to the structure of the state, to its territorial organization normatively defined. Again, the strengthening of regional government has different meanings in the Germanic and in the Anglo-Saxon tradition. When states gain in power in the co-operative federalism of Germany, it is understood as more political interlocking (*Politikverflechtung*) between federal and state levels of government. This process leads to complex patterns of interdependence, often sub-optimal and sometimes unable to avoid 'the joint decision-trap', that is, the impossibility to make any decision other than the smallest core for agreement (Scharpf et al. 1976). The governing capacity of the system is in this context limited by vetoing powers of territorial governments, but the general picture is nevertheless a context of 'big government', increasing public expenditure, and governmental overload. It is noticeable that this kind of analysis has been expanded by this author to European governance as a whole

(Scharpf 1988). In the European context, federalism has been transformed by European integration (Jeffery and Sturm 1993, Jeffery 1997).

The increase in regional government in an Anglo-Saxon 'disjoined' federal system would take a very different meaning, and does not appear likely in the contemporary context. In such a political system, states do not have an institutional access to federal government other than through the Senate. Federal and state policies are supposed to be distinct areas of government, and there is in principle no co-operation between the two levels of government in defining public policies, other than through judicial review. In fact, the 'marble cake' concept of federalism has long been violated by numerous overlaps, redundancies, and competition between federal and state administrations through the development of public policies (Gray et al. 1990, Wright 1988). Here the state tradition is liberal and rests on the idea of a minimal central government. Consequently, the empowerment of state governments would mean resisting the continuous development of federal policies, and saving the political system both from centralization and from 'big government'. Empirically, however, despite a lower level of public spending than in Europe and more limited welfare policies, the US political system has gradually evolved towards higher levels of public intervention and more overlap and interdependence between levels of government.

Finally, the third rationale for regionalization lies in public policies. Even in centralized states such as France, the development of the public sector during the decades after World War II, as well as changes in the economy associated with increased mobility, have often led to recognition of the need for a regional scale of decision making. In France, regions were first created as administrative bodies aimed at improving the territorial fitness of public investments and only gradually became political institutions, with the first election of their councils in 1986. Again, in a centralized state, the reform is mainly administrative (at least at the beginning), whereas in a federal system it is openly political with the transfer of the burden of public policy from federal to state governments.

3. Regional Politics and Policies

These three rationales have converged since the 1970s to strengthen in one way or another the political salience of regional government. This in no way means more autonomy at the regional level. It means, on the contrary, more interdependence in complex institutional settings, where regional specificities, be they economic, social, cultural, or political, have come to be recognized. This has also led to altogether fewer conflictual and hierarchical modes of integration between center and peripheries, defining new forms of public policies and political authority.

Regarding regional politics, political mobilization around regionalist parties was noticeably sustained in Quebec, Catalonia and the Basque Country, Corsica and Overseas French regions, Wallonia and Flanders, in Italy with the Northern League, and to a lesser extent in Wales and Scotland. More generally speaking, independently of quasinationalist movements, politics at the regional level gained in saliency with the development of public policies at the regional level. Regions more specifically develop public programs in education and professional training, culture, economic development, transportation, and regional policy. They answer in a way to a functional logic of public policy, where sectoral policies are selected by the relevance of the scale of government, but also and mainly by the opportunity structure defined by the existing institutional design and potential political support. Regional public policies are thus linked to mobilizations of social interests, sometimes through neo-corporatist modes of organization, and in any case to some forms of political exchange structuring political leadership (Koller-Koch 1997, Négrier and Jouve 1998). More controversially, the performance of regional institutions also refers to social capital and civic traditions (Putnam 1993). Without overestimating such a causality, the relation between regional sociality and identity on the one hand and regional governance on the other hand offers stimulating perspectives for analysis. Generally, societal mobilizations at the regional level tend to correspond to inter-institutional modes of regionalization of public policy.

4. Regional Government and Transnational Integration

Finally, regional government is affected by transnational integration in two different ways. First, economic globalization increasing the mobility of economic factors defines an opportunity for the development of regional policies aimed at managing the territorial consequences of these changes, both in attracting investments and in defending regional firms, industries, and social sectors. The creation of transnational markets such as the European community, NAFTA, and MERCOSUR transforms socioeconomic conditions and needs, most dramatically in the case of border territories and monoindustrial regions affected by international competition. Second, political integration at a supranational level, as it is best exemplified with the European Union, defines new issues for regions to get access to regional policy programs, and more generally to preserve their interests in the new political process. Getting involved in new public policies, regions also invent new channels of influence progressively designing a new political game made of transnational arenas and forums. The creation in 1992 by the Maastricht Treaty of the Comity of the Regions

reveals this logic of recognition of regions by European institutions and their inclusion in this political process. Even if their influence is extremely diverse (Keating and Jones 1985) and needs not to be understood to be an alternative to the power of nation-states (Le Galès and Lequesne 1998), it is nevertheless part of the more polycentric and multilevel forms of governance (Hooghe 1996) characteristic of the European Union. In the same way, the development of foreign relations (Keating and Aldecoa 1999) and of transnational co-operations between regions (Balme 1996) contributes to the bypassing of the Westphalian and exclusively interstate system of international relations, and defines regional collective action as an answer to globalization. This context defines a new relevance for regionalism, but also transforms its meaning (Letamendia 1999), as this new regionalism turns away from parochialism and tradition, and combines economic openness and cultural identity to sustain socioeconomic development in a more cosmopolitan way.

See also: European Integration; Federalism; Interest Groups; Regional Integration

Bibliography

- Balme R (ed.) 1996 *Les Politiques du Néo-régionalisme*. Economica, Paris
- Balme R, Garraud Ph, Hoffmann-Martinot V, Ritaine E 1994 *Le Territoire pour Politiques: Variations Européennes*. L'Harmattan, Paris
- Elcock H, Keating M 1998 *Remaking the Union. Devolution and British Politics in the 1990s*. Cass, London
- Gray V, Jacob H, Albritton R B (eds.) 1990 *Politics in the American States. A Comparative Analysis*. 5th edn. Scott, Forsman, Glenview, IL
- Hooghe L (ed.) 1996 *Cohesion Policy and European Integration. Building Multi-level Governance*. Oxford University Press, Oxford
- Jeffery C (ed.) 1997 *The Regional Dimension of the European Union. Towards a Third Level in Europe*. Cass, London
- Jeffery C, Sturm R 1993 *Federalism, Unification and European Integration*. Cass, London
- Keating M, Aldecoa F 1999 Paradiplomacy in action. The foreign relations of subnational governments. *Regional and Federal Studies* 9(1): special issue
- Keating M, Jones B (eds.) 1985 *Regions in the European Community*. Oxford University Press, Oxford
- Koller-Koch B (ed.) 1997 *Interactive Politik in Europa: Regionen im Netzwerk der Integration*. Leske und Budrich, Opladen, Germany
- Le Galès P, Lequesne C (eds.) 1998 *Regions in Europe*. Routledge, London
- Letamendia F 1999 *Nacionalidades y Regiones en la Union Europea*. Editorial Fundamentos, Madrid
- Loughlin J, Peters B G 1997 State traditions, administrative reforms and regionalization. In: Keating M, Loughlin J (eds.) *The Political Economy of Regionalism*. Cass, London, pp. 41–62
- Négrier E, Jouve B 1998 *Que Gouvernent les Régions d'Europe? Echanges Politiques et Mobilisations Régionales*. L'Harmattan, Paris
- Putnam R D, Leonardi R 1993 *Making Democracy Work. Civic Traditions in Modern Italy*. Princeton University Press, Princeton, NJ
- Scharpf F W 1988 The joint-decision trap. Lessons from German federalism and European integration. *Public Administration* 66: 239–78
- Scharpf F W, Reissert B, Schnabel F 1976 *Politikverflechtung*. Scriptor-Verlag, Kronberg
- Sharpe L J (ed.) 1993 *The Rise of Meso Government in Europe*. Sage, London
- Wright D S 1988 *Understanding Intergovernmental Relations*. Brooks, Cole, Pacific Grove, CA

R. Balme

Regional Identity

1. Introduction

The concept 'identity' is concerned with the answer to the question: 'Who am I?' 'Place identity' and 'regional identity' are answers to the question: 'Where am I?' Sometimes this question is more differentiated: 'Where do I come from' and 'Where do I belong to?' The phenomenon regional identity has an enormous range: it stretches from the personal home to territorial politics (geopolitics), from self-consciousness up to the problem, how to cope with the impacts of globalization.

2. Definition

(a) Regional identity is part of the self-identity of the human being. Regional identity results from the close ties of an individual to a region. It may consist of different intensities: it reaches from (i) a vague sense of belonging to a particular region, (ii) a close attachment to a region, (iii) a deliberate confession to and identification with a region, up to (iv) an active engagement for the region. In each of these cases areas, which are defined emotionally or rationally, serve as an objective, relatively stable source of orientation for the behavior of the individual.

Regional identity is very close to personal identity. Personal identity does not exist for itself but is built—and permanently rebuilt—in social interactions. So there is a strong relation to the (social) environment of a person. In the same way regional identity is not only a personal feeling towards a place but is part—or counterpart—of the collective. There is no self-identity without embeddedness into a community and there is no identity of the collective without individuals.

Therefore regional identity is not only part of the personality which is bounded to the region, but also and particularly consists of the knowledge of others having similar feelings for the region as oneself.

(b) Regional identity is not always regarded as an

attitude or a characteristic of a person (or a group of persons). Sometimes it means a particular region as a whole. Then 'region' is regarded as an entity with special characteristics so far different from other regions. Although it obviously bases on individual attitudes, regional identity is regarded as a relatively autonomous social structural formation, which neither can be seen as the total of individual attitudes nor can be dissolved into society.

To this point there are different positions according to the underlying epistemological background. To declare regional identity as a structure of its own derives from the tradition of the 'physique sociale' founded by Emile Durkheim (1893) and—in a lower degree—from communitarism. Following the position of action theory this second definition of regional identity is seen as an improper naturalism or essentialism. But nevertheless it can be pointed out that a specific characteristic of regional identity obviously is the oscillation between the level of action of the individual and the level of structural formation of the society.

3. Special Aspects of Regional Identity

3.1 Characteristics of Regional Identity

Regional identity is not an isolated status quo of an individual, but is bounded to the development of social interactions as well as to the individual biography. Regional identity develops from different sources, from individual experiences during childhood, from the perception of the personal space, and other experiences. However, even more important are socially mediated aspects, because identity does not exist in individuals for itself, but develops, grows, and changes in social interactions. Therefore culture, tradition, landscape, and history of the region form regional identity and are part of it.

A decisive element of regional identity is the close relation between the region and the historical dimension. The common history is not only referred to as the history of a political territory, it may also refer to the affiliation to an unified confessional territory, a linguistic boundary or similar aspects. Normally there is a bundle of cultural elements important for the identity of a region and the identification of individuals with the region in question. Such elements can be:

(a) the economic basis, e.g. the production of coal and steel in the Ruhr area in West Germany. The popular name of the region 'Ruhrpott' (melting pot at the river Ruhr) reminds one of the process of melting ore with coal;

(b) the language or specific dialect, e.g. 'Languedoc' or 'Occitanie' in the south of France (where people say 'hoc' instead of 'oui' in the north of France), 'Catalunia' in the east of Spain, 'Friul' in the northeast of Italy;

(c) a particular landscape like an island, a big river, or a mountain area (Corsica, The Rhine Valley, the 'Padania' around the river Po, The Highlands);

(d) even specific regional dishes (Cornish pastry and Cornish cream, Tirolian 'Knoedel') can be part of the shared frame of reference. Of course regional identity rarely develops from that kind of particularity alone.

Abstract phenomena as well as objective elements can stand as objects of identity. Normally it is not a single element alone but several characteristics of a specific region which contribute to regional identity.

These shared symbols are normally grounded in the history and development of the region. But the historical dimension also may be an artificial 'myth' or an imagined community in the sense of Anderson (1996). What is really important is the belief in the common history and the existence of a collective. Following Weber (1922) this belief is called 'Gemeinsamkeitsglaube' (for details see Sect. 4). If these shared symbols become an important element of communication in the lifeworld or even in political discourse then there is a switch from regional identity (a) to (b).

3.2 Regional Identity and Scale

Regional identity is a kind of spatial identity on a certain scale called the meso-level. Regional identity is located between the national and the local level. To illustrate regional identity spatially the term 'county' should most likely fit.

Scale is helpful to distinguish regional identity from place identity. Although in some cases place identity is used as the overall term for the different levels of scale, place identity is normally associated with the directly perceived space.

An aspect which distinguishes regional identity from place identity is a higher degree of abstraction from the concrete space. To a lesser extent regional identity is bounded to the concrete perceived space, the personal perception and experience, and the co-presence of individuals. 'Local identity' is bounded to the network of friends and relatives, to the personal socialization of the individual and therefore people are deeply attached to particular places. Regional identity, however, is more abstract: it does not need an actual or the remembrance of a former co-presence of people at the same place. The person with close ties to the region normally knows the region, the physical space, and even more its inhabitants, only in part by personal experience. Particularly against this background symbols which are known and shared by a collective are quite important to establish a sense of community.

3.3 Regional Identity and Regionalism

Regional identity is different from localism as well as from nationalism or patriotism. Above all regional identity is a cultural, social, and psychological phenomenon and not a political one, but this depends on the

degree of the intensity described above. As far as regional identity includes a political dimension, it tends to keep cultural traditions or to blame economic disparities.

If the interest for regional issues leads to a strong political engagement the term 'regionalism' is used to describe this attitude which leads to a larger political movement. Regionalism is the political movement with the aim to preserve or to gain cultural or economic self-reliance (or both). Political autonomy is not the first aim but tends to develop during the history of a political movement. Regionalism often is marked by clear aims and strategies. If the intention is political autonomy one even has to talk about nationalism at least under the perspective of its supporters. These regional political movements can be organized like political parties. Examples are the 'Lega Lombarda' in Northern Italy, the 'Partis Quebecois' in Canada in a more traditional political sense on the one hand, and the liberation movements in Corsica or the Basque provinces in the north of Spain on the other hand.

The line between vague regional identity and active political regionalism is not sharp and the motifs are mixed together. See for instance the Occitanian case: on the one hand in the late 1960s the inhabitants of this region felt disadvantaged. The starting point of the protest was the bad situation of a region depending economically on the production of wine. Most people in that region were affected by this and it had little to do with regional culture and so forth. On the other hand, the regional protest led to much solidarity in the heart of France, especially in the intellectual milieu around the politically left students and academics. The engagement of these people for Occitanian, drinking Herault wine and spending their holidays at the Causse de Larzac, had little to do with their place of origin or the feeling-at-home in that region. The Occitanian identity of these intellectuals expressed their (general) political engagement in favor of disadvantaged regions and the utopia of an intact world. Regionalism unites individual interests and focuses people on common ends. Therefore, regional identity is to be seen as an element of the social system in the sense of the second definition given above.

3.4 Regional Identity and 'Real' Region

Talking about the Occitanian identity of the intellectuals at Paris has already expressed that the region as an area on the earth's surface and the region of the identity are not always congruent. Of course there must be a relation of a particular regional identity to a specific region. But this does not mean identical maps. For example, an administrative region like 'Bretagne' in France or 'Andalusia' in Spain is not congruent with the area of people calling themselves 'Bretons' or 'Andalusians' and is even less congruent with their regional identity. If people draw their mental maps of

the region in favor, the results differ considerably. Normally it can be assumed that the regional identity is more distinctive in the heartland than in the border area. In the center the regional language or dialect prevails and the specific traditions and rules of the lifeworld are cultivated. But even the contrary is possible: if the neighboring region seems to be strange or even hostile regional identity will be strongest at the border, e.g., this phenomenon can be found in the case of South Tyrol.

Another important point in this context is the personal absence from or presence in the region in question. Regional identity is not attached to the physical presence in that particular region. On the contrary, regional identity is sometimes stronger in foreign surroundings than at home: Friulians in Canada, people from Sicily in Torino, Bavarians in Hamburg, and so forth often feel and think more Friulian, Sicilian, and Bavarian than people who see their region as the natural place of the everyday life.

Finally, the Occitanian example shows that regional identity is not always a particular regional identity. Regional identity is an universal value too and not attached to a specific region. To prefer federalism to centralism in a political system expresses this attitude.

4. The Formation of Regional Identity

How does regional identity develop? According to the two aspects of the notion mentioned above in the definition, at least the answer to this question has to make a difference between two levels: the first is the anthropological aspect (in close relation to self-identity), the second is the social theory with special respect to social structural formation.

4.1 Regional Identity as a Basic Need of Human Beings

Regional identity is part of self-identity and therefore stretches from the unconsciousness to the rational consciousness. Especially an active political engagement for the reason in favor is already part of the discursive consciousness.

Above all regional identity is part of the practical consciousness in the sense of Giddens (1984). Among others practical consciousness means living and working in everyday life by routines. Therefore the surroundings and experiences in the places where people live form a stable frame for individual decisions and behavior. One can say that regional identity is built and permanently rebuilt in the social and spatial context of the individuals. The well-known region and its people give trust and security for the behavior of the individuals.

From a phenomenological point of view region has a similar function: at the heart of the lifeworld is the home world in contrast to the strange world outside.

Regional identity as a social process is constructed on different levels: in kindergartens, schools, clubs, and other institutions. Communication and experiences in everyday life form the attitude towards the region.

4.2 Regional Identity as an Element of Social Cohesion

In social theory the topic 'regional identity' is part of the discussion about inclusion and exclusion starting with the book of Ferdinand Tönnies *Gemeinschaft und Gesellschaft* (*Community and Society*) at the end of the nineteenth century (1887). Max Weber says that the formation of community (*Vergemeinschaftung*) and the formation of society (*Vergesellschaftung*) are two of the essential sociological notions. If people believe that there is a traditional or emotional feeling of togetherness then community can arise. To understand regional identity two points are of special interest. First, there must be a real starting point of a regional community (e.g. language, religion, origin) but it is just the belief in the force of this shared characteristic that can lead to the formation of a community. Therefore, even regional identity is socially constructed. Second, there is no need for the co-presence of individuals. The feeling of belonging together can be a powerful motor for the formation of inclusion, especially if there are no other characteristics easy to recognize and to manipulate. So the notion regional identity is not very different from the modern notion of ethnicity. The modern theory of ethnology too regards 'ethnicity' as a construct, while the traditional (primordial) notion of ethnicity is essentialistic and has a strong affinity to race.

From a systematic point of view social cohesion needs rationality as well as emotion. In a similar way to Weber, Emile Durkheim (1893) illustrates a difference between 'mechanical solidarity' and 'organic solidarity' in human societies. In the traditional 'Physique sociale' as well as in the modern 'New Social Physics' social integration is to be seen as a result of collective consciousness (on the basis of emotion, religion, and so forth) in the first case and of treaties between agents in the second. It helps people to avoid making arduous decisions if they follow the idea of 'natural cohesion'. Or as B. Anderson points out: 'In this way, nationness is assimilated to skin-colour, gender, parentage and birth-era—all those things one can not help. And in these "natural ties" one senses what one might call "the beauty of *gemeinschaft*"' (Anderson 1996, p. 143).

5. Regional Identity Between Modernism and Postmodernism

Is regional identity a relic of premodern societies disappearing during the process of modernization? Following the ideas of Giddens (1984), people are

more and more disembedded and this means that their activities and their thinking, that social relations in general, are no longer bound to the space-time-context. Time and space get more and more divided. The boundaries of the region and the self coincide no longer. The space of reference for the individual actor in the time of globalization finally is the whole earth. Therefore it is declared that regional identity loses its importance during the process of modernization. A similar position can be derived from the General Systems Theory of N. Luhmann (1994). According to Luhmann, the whole modern world is inter-related and regional worlds of their own no longer exist. Differentiation in the world is based on functions. Normally the same person belongs to different systems in different places depending on their role as a consumer, employee, or lover. M. Weber had a strong historical interest and regarded social change and the process of modernization but there is no clear statement that he regarded community as a—more and more disappearing—historical phase during the formation of (modern) society.

On the one hand one can be sure that there are no longer closed up regional worlds like in premodern times. On the other hand 'region' is a very simple and therefore very practical code to produce cohesion and identification. Even more: the more complicated and differentiated the modern world becomes the more we need simple possibilities to discern people, political subsystems, and so forth. Therefore regional identity is one of the powerful social bases for political movements, starting from regional protest based on environmental issues up to most of the wars in the late twentieth century. 'Belonging to'—not only same region, but also to the same religion, the same 'race' and so forth—is much more relevant than class differences or other social aspects to include or to exclude people. Such kinds of inclusions—and the corresponding exclusions—are responsible for most of the recent (violent) political developments especially in the east of Europe, but also in Africa or in the Near East.

Regional identity is embedded into the more general debate on the role of identities in relation to the universal values (of the Enlightenment). It is part of the debate which of these two basic values is of more importance: to respect individual historically grown traditions or to assert the general human rights all over the world. There is a strong political dimension connected to this subject: it stretches from the apparently simple question if Islamic girls are allowed to join school lessons with a scarf around their head in laizistic French schools up to the question if western politicians are allowed to claim the recognition of human rights in the People's Republic of China. The popular considerations of Finkelkraut (1989) in France and Germany and of Said (1978) and Bauman (1991) in the English speaking countries give new importance to this subject. The position of the—now

as accidental recognized—values of the Enlightenment has diminished and pluralism had to be put on the agenda. There are different ‘language games’ (in the words of Ludwig Wittgenstein) which are incomparable.

So it has to be pronounced that in the end regional identity is part of the postmodern debate. In a postmodern context regional identity can be submitted to the leading category of difference (and unity). In (post-)modern guise questions come back which were important already at the end of the eighteenth century: Immanuel Kant defended the achievements of the (French) Enlightenment, while John Gottfried Herder so preparing romanticism and historicism declared that these values are not valid for the rest of the world: in different regions develop different cultures which cannot be compared and ranked.

6. Relevance of Regional Identity Today

Following the theoretical considerations, regional identity as ‘a sentiment of belonging to’ always exists to a higher or lower degree at the individual level. Sometimes it is just folklore, but very often it gets at the heart of social and political issues. Regional identity is part of the general tendency in societies to construct social cohesion. Some examples are ecological and political movements, emigration colonies, and economic developments. For instance, political movements focused on ecological topics or resistance against nuclear power plants often affirm to defend ‘their’ region. As already shown in the example of Occitania there is often a mixture of political interest and traditional emotions: without doubt regional identity serves as symbolic arm in political discussions but without regional identity such political movements would not exist—or only for a short time.

People who are forced to emigrate try to keep their former identity in Little Sicily, in the Arabian ‘banlieue’ of Paris, in the Inner City Ghettos of the ‘Guest workers’ in Western Germany. Home is what you feel when you have lost it, so people try to keep their identity by staying together. In globalization, regional identity has a new dimension. On the one hand, people lose control to their surroundings more and more, so they try to keep regional traditions, on the other hand, it is much easier to get involved in the process of globalization and to keep its regional identity within a transnational space at the same time. Even more, jobs are more and more insecure, the firm is no longer the only center of thinking of the employees, while place gains new importance (Sennett 1997). So regional identity becomes more important.

In a similar way regional identity works in regional economics. The moral economy precedes the market economy and is still of importance. The success of industrial districts is sometimes an outcome of regional embeddedness of the entrepreneurs as well as

the employees of the region where it is located and where the network works. Shared rules and common values produce trust, confidence, and reciprocity which are regarded as the surplus making such specialized regions more successful than others. On the other hand it is also possible that there is an economic decline because of too much regional identity and following lock-in-effects: the Ruhr area in the 1970s and 1980s or—in a broader sense—Sicily and its whole Mafia problem may be seen as examples. Regional identity is therefore not a sentimental relic of premodern times. Regional identity is one of the geographical dimensions of the construction of modern society.

See also: Identity in Anthropology; Identity Movements; Identity: Social; Language and Ethnicity; Region: Geographical

Bibliography

- Agnew J 1993 Representing space: Space, scale and culture in social science. In: Duncan J, Ley D (eds.) *Place, Culture, Representation*. Routledge, London, pp. 251–71
- Anderson B 1996 *Imagined Communities: Reflections on the Origin and Spread of Nationalism*, rev. and extended 2nd edn. Verso, London
- Bauman Z 1991 *Modernity and Ambivalence*. Polity Press, Cambridge, UK
- Durkheim E 1893 *De La Division du Travail Social*. Alcan, Paris
- Entrikin J N 1994 Place and region. *Progress in Human Geography* **18**: 227–33
- Entrikin J N 1996 Place and region 2. *Progress in Human Geography* **20**: 215–21
- Entrikin J N 1997 Place and region 3. *Progress in Human Geography* **21**: 263–8
- Finkelkraut A 1989 *Die Niederlage des Denkens*. Rowohlt Taschenbuch Verlag, Reinbek [*La Défaite de la Pensée*]
- Giddens A 1984 *The Constitution of Society, Outline of the Theory of Structuration*. Polity Press, Cambridge, UK
- Hard G 1987 ‘Bewußtseinsräume’. Interpretationen zu geographischen Versuchen, regionales Bewußtsein zu erforschen. *Geographische Zeitschrift* **75**: 127–48 [‘Regions of regional consciousness’. Geographical approaches to regional identity]
- Luhmann N 1994 Inklusion und Exklusion. In: Berding H (ed.) *Nationales Bewußtsein und kollektive Identität*. Suhrkamp, Frankfurt a. Main, pp. 15–45, [*National Consciousness and Collective Identity*]
- Pohl J 1993 *Regionalbewußtsein als Thema der Sozialgeographie*. Münchener Geographische Hefte 70. LaBleben, Kallmünz [*Regional Consciousness as Subject of Social Geography*]
- Said E W 1978 *Orientalism*. Pantheon Books, New York
- Sennett R 1997 The new Capitalism. *Social Research* **64**: 161–80
- Short J R 1991 *Imagined Country: Environment, Culture and Society*. Routledge, London
- Tönnies F 1887 *Gemeinschaft und Gesellschaft. Abhandlung des Communismus und des Socialismus als Empirischer Culturmasse*. Fues, Leipzig, Germany

Weber M 1922 *Wirtschaft und Gesellschaft*. Mohr, Tübingen
[Weber M 1978 In: Roth G, Wittich C (eds.) *Economy and Society*. University of California Press, Berkeley, CA]

J. Pohl

Regional Integration

1. Definition

The term 'regional integration' refers to the process whereby sovereign nation-states voluntarily establish common institutions for collective governance. This process is multifaceted. At the political level, regional integration involves the creation of supranational institutions, with executive, legislative and judicial powers, and the delegation of policy competences from domestic governance institutions to these institutions. On a legal level, regional integration involves the establishment of a new hierarchy of legal norms, where supranational law is not only an integral part of domestic law (the doctrine of 'direct effect') but is also sovereign over domestic law (the doctrine of 'supremacy'). At the economic level, regional integration involves the development from a common market (in parallel with existing national markets), through a single market (with the removal of barriers to the free circulation of goods, services, capital and labor), to a common or single currency (in parallel with, or replacing, national currencies), and eventually to full economic union (with common macroeconomic and fiscal policies). At the sociological level, regional integration involves the creation of new supranational identities, loyalties, interest groups, and social movements, in addition to and sometimes in opposition to existing national identities, interests, and movements.

2. The Early Study of Regional Integration

The first steps towards regional integration occurred in Europe in the immediate aftermath of the Second World War. This led to the establishment of the European Coal and Steel Community (ESRC) in 1952 by six states (Germany, France, Italy, and the Benelux countries), who then created the European Economic Community and the European Atomic Energy Community in 1958. Hence by the end of the 1950s, regional integration was more advanced in Europe than in any other region in the world. Nevertheless, modeled on the European process, inter-state agreements and supranational economic regimes began to emerge in South East Asia, Latin America and North America.

As a result, in the 1950s, the process of regional integration was regarded by most social scientists as a generalizable phenomenon. For example, Karl Deutsch and his colleagues (Deutsch et al. 1957)

developed a general 'transactions theory' of integration: as cross-border economic and social transactions increase, new regional identities and communities are created, which facilitate new institutions, and so on. The zenith of the early study was the 1970 special issue of the journal *International Organization*, which tried to develop and test theories that could explain the varying stages of regional integration throughout the world (Lindberg and Scheingold 1971).

3. European Integration: A Sui Generis Phenomenon

However, with the global economic crises in the early 1970s, it became clear that the experiments in regional integration outside Europe were not as deeply entrenched as the process within Western Europe. Ernst Haas, the founding father of 'regional integration theory,' consequently proclaimed that the generalizable study of regional integration was now 'obsolete' (Haas 1976). From the launch of European integration in the 1950s to the relaunch in the mid-1980s, there have been two dominant explanations of European integration.

3.1 Neofunctionalism

Neofunctionalism was first proposed by Haas (1958) to explain the establishment of the ECSC Treaty, and was developed by his disciples to explain subsequent integration steps. The basic argument of neofunctionalism is that European integration is a deterministic process, whereby 'a given action, related to a specific goal, creates a situation in which the original goal can be assured only by taking further actions, which in turn create a further condition and a need for more, and so forth' (Lindberg 1963, p. 9).

This 'spillover' process is promoted by two sets of actors. First, at the domestic level, societal and sectional interests promote integration of different policy sectors to achieve political advantage over their opponents or secure benefits already accrued by the integration of other policy sectors. For example, in the 1980s, multinational business interests lobbied for the establishment of a European-wide single market, to secure greater economies of scale and lower transactions costs in cross-border trade. Second, at the European level, the supranational institutions promoted further integration to increase their ability to shape policy outcomes. For example, in the 1960s the European Court of Justice consciously made rulings and established doctrines to promote legal integration in Europe, and in the mid-1980s the European Commission was instrumental in persuading governments to adopt the program for completing the single market.

3.2 Intergovernmentalism

Neofunctionalism's failure to explain the slowdown of European integration in the 1970s, and the subsequent strengthening of the 'intergovernmental' elements of the European Communities, led to the emergence of a starkly opposing theory of European integration, known as intergovernmentalism. In direct contrast with neofunctionalism's emphasis on nonstate interests at the domestic and supranational levels, intergovernmentalism argues that European integration is driven by the interests and actions of the European nation-states. In this interpretation, the main aim of nation-states is to protect their geopolitical interests, such as national security and 'sovereignty.' Also, interstate bargaining is a zero-sum game, where 'losses are not compensated by gains on other issues: nobody wants to be fooled' (Hoffmann 1966, p. 882). Consequently, against the neo-functional 'logic of integration,' intergovernmentalists see a 'logic of diversity [which] suggests that, in areas of key importance to the national interest, nations prefer the certainty, or the self-controlled uncertainty, of national self-reliance, to the uncontrolled uncertainty of the untested blunder' (Hoffmann 1966, p. 88). From the intergovernmentalist perspective, then, European integration was relaunched in the mid-1980s because the governments in the large member states shared the same policy preference for a single market and the governments in the periphery states were bought off through 'side payments,' such as regional transfers (e.g., Moravcsik 1991).

4. Reintegrating Regional Integration with General Social Science

Ironically, as European integration gathered pace at the end of the 1980s, regional integration began again to be understood as a general phenomenon in social science. There were three main reasons for this shift. First, with the Single European Act in 1987, the establishment of the EU in 1993, and the launch of European economic and monetary union in 1999, European integration began to have a significant impact on domestic political systems and on global political and economic relations. European integration and the EU hence became a major focus of research in a number of social science disciplines.

Second, in the 1990s, the success of the European single market encouraged regional integration in other parts of the world—where governments breathed new life into moribund regional organizations (such as the Association of South East Asian Nations) or created new regional governance regimes (such as the North American Free Trade Area). Once again, regional integration was a general rather than a *sui generis* phenomenon.

Third, and related to these two aspects, with this renewed interest in the European process came specialization, where scholarship focused on specific aspects of the integration process or the EU rather than on the broad historical development of regional integration. This specialization enabled scholars of European integration to apply general social science theories, e.g., of interstate bargaining, interest group behavior, legislative-judicial interactions, and monetary integration. Specialization consequently increased the subtlety and sophistication of our explanations of regional integration. It also enabled scholars of regional integration to be part of the mainstream study of these cross-systemic social science phenomena. This integration of regional integration and social science occurred in political science, legal studies, economics, and sociology.

4.1 Regional Integration in Political Science

In the International Relations subdiscipline of political science, regional integration is a particular type of supranational regime building between nation-states. For example, in his 'liberal intergovernmentalist' (LI) theory of regional integration, Andrew Moravcsik (Moravcsik 1998) deliberately places regional integration at the intersection between the two main approaches in international relations: liberalism and realism (see *International Organization*). From liberalism, LI argues that state preferences are formed through a process of domestic competition between rival economic interests in society. From realism, LI argues that once state preferences have been formed, states bargain to protect and promote their private interests, and cooperate to design institutions to reduce transactions costs, overcome collective action problems, and establish credible commitments (for example, by delegating agenda-setting and enforcement powers to supranational agents).

In the subdiscipline Comparative Politics, the EU, despite its peculiarity, possesses all the main attributes and processes of a political system. This enables the different elements of regional integration to be explained using general theories from comparative politics (Hix 1999). For example, the EU governments have delegated executive powers to the European Commission, and have designed institutions to control these powers, in the same way that legislative principals delegate to and control the powers of executive agencies in the domestic arena. Similarly, on the 'inputs' side of the political system, mass attitudes towards regional integration are shaped by the same utilitarian calculations and affective beliefs as the public show towards domestic institutions. Further, on the 'outputs' side, regulation (rather than redistribution) emerged as the central policy instrument of regional integration in parallel with the expansion of regulation and independent regulatory agencies in the domestic arena.

4.2 Regional Integration in Legal Studies

In legal studies, with the establishment of the doctrines of direct effect and supremacy, and the hierarchical relationship between the European Court of Justice and national courts (where national courts apply EU law in the domestic arena), the EU legal system is akin to a federal legal system (Weiler 1991). However, since the German Constitutional Court judgement on the legality of the 1992 Treaty on European Union (the Maastricht Treaty), there is not a clear hierarchy of norms in the EU system. EU law, in its primary form in the EU Treaty and in secondary legislation, is supreme over conflicting domestic statutes and legislation. However, according to the German Constitutional Court and similar constitutional authorities in other member states, EU law remains subservient to basic fundamental rights, as set out in domestic constitutions and the European Convention on Human Rights.

In the intersection between political science and legal studies, the interaction between the EU governments and the European Court of Justice is similar to the interaction between domestic legislative and judicial authorities. On one side, EU governments have delegated judicial power to the European Court of Justice to establish a rule of law, without which there would be little incentive for governments to implement EU legislation. On the other side, protected by the constitutional guarantees in the EU Treaty, the European Court of Justice has pursued the political strategies of many domestic constitutional courts, by reinforcing its legal authority versus lower courts and shaping political outcomes through the policy adjudication process.

4.3 Regional Integration in Economics

In economics, regional economic integration is a particular case of international economics and monetary economics. In international economics, states have liberalized trade to promote comparative advantage and economies of scale. From this perspective, the United States (in the nineteenth century), the European Union, the North American Free Trade Area and the World Trade Organization are all examples of the same phenomenon. In all these cases, the removal of barriers to the free circulation of factors of production was then followed, to a greater or lesser extent, by the creation of common regulations to address market failures, such as unfair competition (through merger rules), asymmetric information in market transactions (through rules on consumer protection), and negative externalities (through rules on environmental protection).

In monetary economics, the process of economic and monetary union in Europe, with the creation of the single currency, has led to the resurrection of the old 'Optimum Currency Area' (OCA) theory of

monetary unions (Eichengreen and Frieden 1994). In the OCA theory, nations decide to form a monetary union when the benefits of keeping independent exchange rates and interest rates (to address asymmetric demand shocks) are lower than the benefits of establishing a common currency (of lower transactions costs and greater certainty). However, through the empirical and theoretical study of monetary integration in Europe, this theory has been refined to take account of both political and economic calculations, how costs and benefits change with different levels of trade integration, and how different assumptions about the trade-off between inflation and unemployment change these calculations.

4.4 Regional Integration in Sociology

Although to a lesser extent than these other social science disciplines, regional integration has also become an important topic in sociology. For example, Soysal (1994) argues that supranational human rights norms—through the European Convention on Human Rights and the policies of the European Union—were as important as domestic factors in transforming the conception and operation of citizenship norms in Germany and France in the 1980s.

See also: Cultural Assimilation; Cultural Geography; International and Transboundary Accords, Environmental; International Law and Treaties; International Trade; Economic Integration; Regional Government; Regional Identity; Regional Science; Regionalism and Cultural Expression; Sovereignty; Political

Bibliography

- Deutsch K W, Barrell S A, Kann R A, Lee M Jr, Lichterman M, Lindgren R E, Lowenheim F L, Van Wagenen R W 1957 *Political Community in the North Atlantic Area*. Princeton University Press, Princeton, NJ
- Eichengreen B, Frieden J (eds.) 1994 *The Political Economy of European Monetary Unification*. Westview Press, Boulder, CO
- Haas E B 1958 *The Uniting of Europe: Political, Social and Economic Forces, 1950–57*. Stevens & Sons, London
- Haas E B 1976 *The Obsolescence of Regional Integration Theory*. Institute of International Studies, Berkeley, CA
- Hix S 1999 *The Political System of the European Union*. Macmillan, London
- Hoffmann S 1966 Obsolete or obsolete? The fate of the nation state and the case of Western Europe. *Daedalus* 95: 862–915
- Lindberg L N 1963 *The Political Dynamics of European Economic Integration*. Oxford University Press, Oxford, UK
- Lindberg L N, Scheingold S A (eds.) 1971 *Regional Integration: Theory and Research*. Harvard University Press, Cambridge, MA
- Moravcsik A 1991 Negotiating the Single European Act: national interests and conventional statecraft in the European Community. *International Organization* 45: 19–56

- Moravcsik A 1998 *The Choice for Europe: Social Purpose and State Power from Messina to Maastricht*. Cornell University Press, Ithaca, NY
- Soysal Y N 1994 *Limits of Citizenship: Migrants and Post-national Membership in Europe*. University of Chicago, Chicago
- Weiler J H H 1991 The transformation of Europe. *Yale Law Journal* 100: 2403–83

S. J. Hix

Regional Planning, History of

1. The Term 'Regional Planning'

'*Gallia est omnia in partes tres.*' All Gaul is divided into three parts. This command, from the reign of Julius Caesar in the first century BC, no doubt is one of the earliest example of regional planning. A region is a district or area defined by generally recognizable attributes, physical cultural, economic, or political. The word 'region' derives from the Latin '*regio*,' meaning direction or boundary, and from '*regere*,' to direct. Thus the concept of region embodies the interaction among place, people, and work or economy. A region may or may not be formally recognized as a political entity or sub-division. More likely a region is simply a shared mental construct of its constituents.

Regional planning then is the effort of the society or elements in the society to intervene in the integrated developmental evolution of the area in question. In ancient kingdoms, the wishes and wisdom of the ruler constituted the regional planning conception for the society. Values and objectives underlying regional planning and foresight particularly included protection from external threats, but also the provision of infrastructure—roads, water, the enhancement and protection of productive fields and forests, provision for population growth and economic development activities such as industry and trade, cultural enrichment, and provision for societal continuity in succeeding generations. These values and practices were often codified in religious beliefs and scripts. These values still underlie and are used to justify contemporary efforts to define and carry out regional planning.

At this point it will be helpful to discuss a few related words or concepts. We need first to distinguish regional from city planning. The concepts are related but city planning implies a *polis* or civic entity governed by elected or other recognized official bodies. A region does not necessarily have such official status. Indeed, often it does not.

Regional analysis is the study of the interaction of key variables—physical, social and economic—in a

regional context. Regional science, a relatively recent evolution founded upon economic location theory, seeks to describe and predict systemic regional phenomena.

We need also to distinguish types of regions. The European Community is a region comprised of nation states seeking to develop a commonality of purpose and policy, which presumably will result in superior outcomes for its constituent members. The European Union thus practices a kind of multi-state regional planning, much as the federal government in the United States does for the aggregate of its fifty constituent states. Parts of nation-states also may constitute regions. One thinks of Scotland and Wales in the United Kingdom or the Appalachian Region or the Great Plains of the United States.

Another distinction among regions is between mono-nodal, multi-nodal, and non-nodal regions. Metropolitan regions are examples of mono- and sometimes multi-nodal regions. Natural resource regions are often non-nodal. River basin regions are another category: linear regions.

Regions can also be distinguished as growth regions or as lagging regions. Regions may lag in development behind national averages, due to lack of natural resources, limited accessibility, or cultural or educational factors. Much attention has been given to assisting lagging regions to 'catch up,' with mixed results.

Today, we hear much about 'integrated regional resource management,' and 'sustainable development.' These concepts, while imprecise and difficult to measure, are increasingly shaping the directions towards which regional policy planning is moving.

We should also distinguish between 'regional planning,' and 'regional plan.' A regional plan is a specific document which lays out in physical detail a future desired end-state. It may also contain social and economic policy statements. The regional plan is typically based on a regional survey of social, economic, and physical factors. 'Regional planning' is the process by which regional plans are made. 'Regional policy' is a related concept. Regional policies are, in effect, guidelines to realize or implement regional plans.

2. Origins and Practice of Regional Planning

The origins of modern regional planning doctrine can be traced to reactions to the oppressive conditions created by the Industrial Revolution in nineteenth century England and France. Robert Owen and Charles Fourier proposed utopian schemes for alternative workers' housing and communities. Piotr Kropotkin's *Fields, Factories and Workshops* (1901) offered insights into the connections of people, land and labor. Charles Booth's *Life and Labor in London* (1892) systematically surveyed the conditions of work-

ing class people. Ebenezer Howard's 1902 *Garden Cities of Tomorrow* presented an influential alternative to unplanned metropolitan expansion. Patrick Geddes, a Scottish biologist, published his *Cities in Evolution* in 1915. Geddes' synthesis of the salient connections among people, land, economic activity, and natural systems greatly influenced the American essayist and critic, Lewis Mumford. Mumford, in turn promoted Geddes' integrative approach to social and public policy through such books as *The Culture of Cities* (1938). Through Mumford, Geddes' ideas influenced the Regional Planning Association of America (RPAA), a progressive private group concerned with development which was active in the 1920s and 1930s.

2.1 The Regional Planning Association of America

The RPAA embraced Ebenezer Howard's idea to siphon off and channel new growth at the suburban edges of existing large cities into planned garden cities. The group also espoused the notion that economic growth should be decentralized to smaller cities and towns and diverted from large metropolitan areas. The RPAA was especially influential in New York State, where Governor Al Smith appointed Stein and his fellow RPAA colleague, Henry Wright, to a newly-created Commission on Housing and Regional Planning. In 1926, the Commission issued a regional plan for the state that called for a concentrated grid of lines of transportation and communication along major corridors, separation of cities and highways, and the creation of independent communities away from the centralizing tendencies of its major cities. Protection of forest, farmlands and recreation areas were also proposed. The plan had considerable influence on policy and some of its proposals were ultimately carried out. What is notable is that these policies and issues are still under serious debate some 75 years later (Spann 1996).

The RPAA's conception of regional planning inevitably brought the group into conflict with the well-funded effort to create a Regional Plan for the New York metropolitan area begun in 1921 under the direction of Thomas Adams, a British town planner who had worked on plans for garden cities of Welwyn and Letchworth, and the Greater London region.

2.2 The Regional Plan of New York and Its Environs

The monumental Regional Plan of New York and Its Environs, begun in 1921 and published in 1929 (Committee on the Regional Plan of New York and Its Environs 1929). The plan brought together many of the development trends and proposals of the time and was underwritten by generous grants from the Russell Sage Foundation totaling some \$1.2 million. In 1990,

this would be equivalent to roughly \$20 million. The Regional Plan and Survey was and probably remains the best funded metropolitan regional plan ever undertaken (Hays 1965).

The plan, encompassing 22 counties surrounding the Port of New York and targeted on the year 1965, incorporated proposals of the Port Authority rail plans, including those for belt line freight rail lines. Schemes for rapid-transit improvements, highway construction, zoning rationalization, and parks and parkways were assembled from a variety of sources and agencies and synthesized into a unified conception. The plan called for a 're-centralization' of industry and business, involving, among other proposals, a shift of manufacturing outward from Manhattan and the core of the region. Whatever its flaws, the 1929 Regional Plan of New York and Its Environs left a significant imprint on Metropolitan New York. The plan's basic physical planning conceptions were reflected in the subsequent development of the region.

The Regional Plan of 1929 clearly had a significant effect on the physical development of the region. But it also brought important changes in the institutional and public bodies concerned with regional development. Numerous planning and zoning agencies were established at the county and municipal levels as a direct result of field work by the staff of the Regional Plan. State policies for regional development were also stimulated by the plan and its advocates. The plan also influenced the Depression-era New Deal policies of the Roosevelt Administration. President Roosevelt was well-aware of the ideas of both the RPAA and the Regional Plan of New York.

Roosevelt spoke at a Roundtable on Regionalism held at the University of Virginia on July 11, 1931 on the need for regional planning and development, aided by the initiatives of the federal government to lead the country out of the economic stagnation it had found itself. Thus Roosevelt was instrumental in getting the thought of Southern Regionalists, such as Howard Odum and Rupert Vance of the University of North Carolina, and the ideas of the RPAA group onto the agenda of the New Deal. This may have been the origin of the idea for a Tennessee Valley Authority (TVA) in Roosevelt's thinking. On a national scale, the New Deal also created a National Resources Committee to survey the needs and resources of the country.

2.3 The Tennessee Valley Authority

The Tennessee Valley Authority (TVA) was the first of the river basin or other large regional agencies to attempt to bring together the then-new concepts of integrated resource development, which was central to the conservation ideas of the time, and regional planning which was a part of the new city and area planning movement.

Roosevelt's early speeches and papers clearly reveal a deep-seated and long-standing concern with farm problems, rural poverty, and the inadequate methods of providing services to rural areas. He also saw these problems as coupled with those of overcrowding and poor living conditions for the working people in the big cities, which he believed were fueled in part by the migration from farms to cities. The emerging concepts of regional planning seemed to Roosevelt to provide the linkage between the two problems. It is within this context that the regional planning ideas of Roosevelt were appended to earlier ideas for a multi-purpose plan for the Tennessee River, which had focused on flood control, navigation, and power generation. The main body of the TVA Act, as introduced on March 9, 1933, consisted of a comprehensive river improvement program plus two sections which provided for surveys of regional development problems and for recommendations to Congress on necessary legislation to solve these problems. The river improvement sections of the TVA Act were very explicit and provided for a navigation channel from Knoxville to the Ohio River, flood control to protect valley cities and to help control floods on the Mississippi River, and power generation. In contrast, the regional planning sections were very general and thus susceptible to either broad or narrow interpretation. TVA's mandate to conduct and implement regional planning was never clear or unambiguous.

The National Resources Committee (later the National Resources Planning Board) envisioned TVA as the first of a nationwide array of river basin planning regions. TVA was viewed as a model for the rest and as a testing ground for new communities as originally proposed by the RPAA group. The first of these was Norris, Tennessee, built near the first hydroelectric dam, Norris Dam, originally to house construction workers, and then to be converted to housing for the local populations dislocated by the rising waters of Norris Lake. Roosevelt's grand vision for river basin regional planning initiatives was thwarted by the interventionist scale of the vision, in a country dominated by unregulated market forces, and by the outbreak of World War Two. TVA nevertheless became a model for numerous grand multi-purpose hydro-electric and irrigation projects around the world, some misconceived and contrary to sound, integrated regional resource planning and management principles. Sadly, TVA itself has relinquished any serious commitment to regional planning leadership and has become, in effect, primarily a publicly-owned wholesale electricity producer.

2.4 Appalachian Regional Commission

Besides TVA, America's other important multi-state regional planning effort has been the Appalachian Regional Commission (ARC). Like TVA, the ARC was created to deal with an economically lagging

region, the mountainous, relatively inaccessible chain of counties stretching from Alabama and Georgia to the Southern Tier of Upstate New York. The ARC began as an idea of President John Kennedy, who toured West Virginia's poverty-stricken mining towns in his campaign for the White House. The ARC was established in 1965 in the Johnson Administration. The ARC developed a plan for improving access and social services for isolated communities in parts of thirteen states. Highway improvement and infrastructure programs were central elements of ARC regional development. Thirty years later, assessments of the effectiveness of federal investments in the region are mixed. But according to one comprehensive study, the federal investment infusion seems to have been effective. Counties in the ARC program were noted on average to have developed faster economically than comparable counties not in the program (Bradshaw 1992, Isserman and Rephann 1995). ARC was still functioning in the year 2000 but other federal regional development efforts such as the eight Department of Commerce Title V regional commissions setup in other lagging regions of the US were not so fortunate. All were closed down in 1980.

2.5 Metropolitan Regional Planning

The post-war era in the United States, with its spreading suburbs and growing automobile dependency, raised new issues for regional planners, notably in the large metropolitan areas. Federal aid for metropolitan planning became available with the enactment of Section 701 of the Housing Act of 1954, stimulating the creation of planning commissions for the nation's metropolitan areas. By 1960, two-thirds of US metropolitan areas had some form of area-wide planning and, by 1971, thanks to federal grants and statutory requirements, all of the nation's 247 metropolitan areas had adopted official regional planning, mostly through elected, voluntary, Councils of Government (COGs). The A-95 review process, which required regional review and coordination to receive funding for large scale, federally-assisted projects, and Section 208 of the federal Clean Water Act of 1972 further stimulated the creation of regional planning bodies in metropolitan areas.

In the New York Metropolitan Region, the privately-supported Regional Plan Association produced a Second Regional Plan for the New York metropolitan area, released in November 1968. The Second Regional Plan emphasized the creation of regional sub-centers linked by fast rail to Manhattan to expand job opportunities for those trapped in the central cities. For the suburbs, these centers would provide an organizing focus for social and civic life, employment opportunities, and easy access to and from Manhattan. The centers would be composed of 'regional activities' which would have to be built in any event to

accommodate the expected additional twelve million people who would join the 18 million already in the 31-county region (expanded from the original 22-county definition) by the year 2000.

Forecasts for a large shift in employment away from manufacturing to service and white-collar jobs suggested the potential for significant office growth, some of which could go to centers outside of Manhattan. The plan proposed massive new office development for Manhattan, both in downtown and midtown but the plan also proposed that regional centers be built, both as redeveloped inner-city sub-centers such as Downtown Brooklyn, Jamaica and Newark, as older city centers in suburban cities, or as entirely new centers on open land. The centers idea took hold in a number of places, notably Jamaica, Queens, Downtown Brooklyn, and Fordham Road in the Bronx. Downtown Newark also experienced new central place development and expanded centers were developed in the outlying areas of White Plains, Stamford, and New Brunswick.

The Second Regional Plan draft addressed other issues as well. Its housing policies called for increased densities, a more compact region, better related to accessibility to jobs and activities, and more housing opportunities for lower-income and minority families. The draft also called attention to the continuing need to capture more open space and to deal with the increasingly difficult problem of waste management. The plan called for completion of the region's share of the interstate highway network but it also made a strong case for the preservation and enlargement of existing rail transit systems.

In the 1980s, regional planning in the United States suffered a severe setback when the federal government withdrew virtually all its support for regional planning agencies, resulting in the demise of many programs. Only the transportation planning function survived in most metropolitan areas. TVA and the Appalachian programs managed to continue, though their planning functions were greatly reduced. Some states replaced federal assistance for regional planning with state funds. Other regional agencies survived by becoming more entrepreneurial, focussing on service delivery. A few states such as Florida, Georgia and Massachusetts, concerned about the negative aspects of growth, have developed growth management or 'Smart Growth' initiatives, which require regional planning as a framework for development policies, in some cases.

Regional cooperation in politically fragmented metropolitan areas has generally proved difficult to realize in the United States. A number of metro areas, however, have managed to achieve success in facilitating regional planning consensus. The Twin Cities region of Minneapolis/St. Paul, and the Portland Oregon region are often cited as examples of successful regional initiatives. The Minnesota Twin Cities area created in 1967 through state legislation a seven-

county Metropolitan Council. The Council was created to address specific region-wide problems such as sewers and transportation. The Council has adopted a development guide for regional and local decision-making that identifies planned urban areas and growth centers. The Twin Cities regional initiative has enjoyed some major success, though its planning function has been scaled back in recent years.

The Portland Oregon three-county region in 1979 authorized by referendum a metropolitan service district. The district provides technical assistance and reviews local plans. In 1992, the Metro, as it is called, was given approval by charter to adopt a regional framework plan to guide regional development. The framework plan will address land use, development densities, housing, transportation, natural areas, and water supply. The Portland process is generally viewed as a successful example of the application of effective regional planning to control the negative aspects of rapid growth, though some concern has been raised about the impact of controls on the housing costs.

Metropolitan regional planning has experienced a resurgence in the 1990s in a number of American urban areas such as San Diego, California and Chattanooga, Tennessee, while, in the New York metropolitan region, the Regional Plan Association released in 1995 a Third Regional Plan, successor to the 1929 Regional Plan of New York and Its Environs, and to its Second Regional Plan of 1968 (Yaro and Hiss 1996). The Third Plan has focused on ways to help the region recover its momentum so as to maintain its faltering position in the global marketplace. To address the problems facing the region, the Third Regional Plan proposes five ambitious campaigns focussed on open and green space, transportation improvements, strengthening urban centers, improving workforce skills, and streamlining governance structures. It remains to be seen whether the Third Regional Plan will have as much impact as its two predecessors.

2.6 British and European Regional Planning

The trajectory of regional planning in Britain and Europe has somewhat paralleled that of the United States. Regional planning in post-World War II Britain found new support and momentum. The New Towns idea was re-embraced as a means to rebuild a Britain exhausted by war. Regional plans for the utilization of the old, lagging industrial regions of the North of England, Wales, and Scotland were adopted. These provided for decanting of populations into a new generation of expanded and new towns. Older urban areas would be contained within protected greenbelts.

In the 1960s, regional planning was further spurred by commission reports concerned about fast growing areas, notably in the Southeast of England. Metro-

politan regional planning, mainly advisory, was undertaken by county and regional councils. Their mandate was to:

(a) work out broad objectives for each region and provide a framework for detailed sectoral decision-making;

(b) advise on national policy-making where it could significantly affect regions;

(c) advise on the application of national policy in their region; and,

(d) help create a consensus within a region as to its problems and possible solutions.

Regional planning tended to be more favored by the Labour Party, less so by the Conservatives. The Conservatives tended to view taxation for social schemes and infrastructure as a burden on business and regulation and control of development as disincentives to market responses to problems of growth. The Thatcher government in the 1980s reconfigured regional planning bodies and responsibilities: the Greater London Council and other regional councils were abolished and regional planning was transformed into a mechanism for assisting large-scale private redevelopment schemes.

With the return of a more liberal-left government in the 1990s, there has been a modest return to regional planning initiatives. These initiatives, however, have continued the private/public partnership approach promoted by the Thatcher government. The central government places more emphasis on local government bodies. Grand regional planning schemes are not in vogue, as pragmatism and incremental and market approaches are more the order of the day. Whether this approach will enable the UK to meet the challenges of the first decades of the new century remains to be seen. With continuing population growth, a deteriorating environment, limited local resources, and continued un- and under-employment in some regions, together with a need for further integration into the European system, Britain faces hard choices. Robust, effective regional planning and implementation intermediate between the central government and local governments would appear to be a prerequisite for meeting these challenges. Whether the political climate will permit and support an articulated interventionist approach remains to be seen.

On the European continent, regional planning has been practiced with considerable success in both France and the Netherlands. France, with its strong provincial history and identities, has been able to focus its regional planning efforts on channeling excessive growth away from the Paris region towards second- and third-order city-regions. Meanwhile, the Paris region has benefited from metropolitan region planning and development that has focused on transit-based development and suburban centers. Regional planning in the Netherlands has emphasized a '*Randstad*' plan which focuses development on the ring of cities surrounding a protected green center in

the middle of the country. The European Union has fostered regional cooperation among regions that share common characteristics. An example is the Atlantic Arc Region, the grouping of the regions of EU member country fronting the Atlantic Ocean, which share technologies and cooperative approaches to managing planning and development.

3. Future Prospects for Regional Planning

To summarize, it is reasonable to say that regional planning in western industrialized countries has historically had a checkered history of meeting the challenges of growth and decay at the sub-national level. But it is also accurate to point out that those areas that adopted and tried to carry out regional planning schemes have often achieved better results than those which simply drifted into incremental, piecemeal development. It is almost always an advantage to have some systematic framework and political consensus on which to build, even if all relevant variables cannot be predicted or controlled.

There is reason to believe that regional planning will enjoy new vitality and acceptance. Powerful new forces, needs, and demands are compelling a renewed look at the desirability of regional planning approaches. Globalization trends are fostering the integration of economies on an international basis. Metropolitan economies are increasing seen as the building blocks of a world trading system. The communication and air travel revolutions have reduced the accessibility advantages enjoyed by a few large cities. Competitiveness will require regions to make the most of their assets, and grow new ones. This will take conscious, deliberate, smart planning. Environmental pressures and the deterioration of air, water, and land have reached serious proportions in many countries. Regional approaches will be needed to address the pollution of air and water sheds, which rarely conform to the boundaries of local jurisdictions. Global warming and the need to reduce carbon emissions will inevitably require regional approaches to the transportation and land-use interface. Sustainability is becoming the norm by which regional development decisions increasingly are being judged. More attention is being paid to the protection of natural and cultural amenities, natural habitat, green areas, wetlands and waterways, which, by their scale, mandate regional approaches. At the same time, the digital revolution has provided policy-makers and the public with new tools to collect, manage, and display vast amounts of relevant data, in ways only dreamed of in the early years of regional planning. Powerful computer-based geographic information systems (GIS) now permit mapping and analysis of trends, projection of impacts, and improved understanding of alternatives by regional bodies. This new capacity will surely enhance regional planning possibilities and effectiveness in many areas and countries.

See also: Planning Issues and Sustainable Development; Planning Theory: Interaction with Institutional Contexts; Regional Geography; Regional Integration; Resource Geography; Urban Planning: Central City Revitalization; Urban System in Geography

Bibliography

- Boardman P 1978 *The Worlds of Patrick Geddes: Biologist, Town Planner, Re-educator, Peace-warrior*. Routledge & Kegan Paul, London
- Booth C (ed.) 1892 *Life and Labour of the People of London*. Macmillan, London
- Bradshaw M 1992 *The Appalachian Regional Commission: Twenty-five Years of Government Policy*. The University Press of Kentucky, Lexington, KY
- Committee on the Regional Plan of New York and Its Environs 1929 *Regional Plan of New York and Its Environs*. The Committee on the Regional Plan of New York and Its Environs, New York
- Cullingworth B 1999 *British Planning: 50 Years of Urban and Regional Policy*. Athlone Press, Somerset, NJ
- Downs A 1994 *New Visions for Metropolitan America*. Brookings Institution and the Lincoln Institute of Land Studies, Washington, DC
- Friedmann J, Weaver C 1979 *Territory and Function: The Evolution of Regional Planning*. University of California Press, Berkeley and Los Angeles
- Geddes P 1915 *Cities in Evolution*. Williams and Norgate, London
- Hansen N 1968 *French Regional Planning*. Indiana University Press, Bloomington, IN
- Hays F 1965 *Community Leadership*. Columbia University Press, New York
- Isard W 1960 *Methods of Regional Analysis*. John Wiley & Sons, New York
- Isserman A, Rephann T 1995 The economic effects of the appalachian regional commission: An empirical assessment of 26 years of regional development planning. *Journal of the American Planning Association* 61(3): 345–64
- John D et al. 1998 *Building Stronger Communities and Regions: Can the Federal Government Help?* National Academy of Public Administration, Washington, DC
- Johnson D A (ed.) 1996 *Planning the Great Metropolis: The 1929 Regional Plan of New York and its Environs*. E & F Spon, London
- Katz B (ed.) 2000 *Reflections on Regionalism*. Brookings Institution, Washington, DC
- Kropotkin P 1901 *Fields, Factories and Workshops*. S. Sonnenschein, London
- MacFadyen D 1970 *Sir Ebenezer Howard and the Town Planning Movement*. MIT Press, Cambridge, MA
- MacKaye B 1962 *The New Exploration: A Philosophy of Regional Planning*. University of Illinois Press, Urbana, IL
- Meck S (ed.) 1999 *Planning Communities for the 21st Century*. American Planning Association, Chicago
- Mumford L 1938 *The Culture of Cities*. Harcourt Brace Jovanovich, New York
- So F, Hand I, McDowell B D (eds.) 1986 *The Practice of State and Regional Planning*. International City Managers Association and the American Planning Association, Chicago

Spann E K 1996 *Designing Modern America: The Regional Planning Association of America and its Members*. Ohio State University Press, Columbus, OH

Wannop U A 1995 *The Regional Imperative: Regional Planning and Governance in Britain, Europe, and the United States*. Jessica Kingsley Publishers, London

Yaro R, Hiss T 1996 *A Region at Risk: The Third Regional Plan for the New York, New Jersey, Connecticut Metropolitan Area*. Island Press, Washington, DC

D. A. Johnson

Regional Science

Economists pay too little attention to the location of economic activity and the effects of distance on economic systems. Geographers focus on the spatial dimension but lack rigorous tools to study the effects of location and distance on human behavior or explain why activities are located where they are. Urban and regional planners need a stronger analytical foundation to go beyond the physical and aesthetic aspects of urban form to consider social and economic dimensions of city and region. With such premises, 60 economists, geographers, political scientists, sociologists, and planners gathered at the annual meeting of the Allied Social Science Associations in 1954 and founded the Regional Science Association. Its objective was ‘to promote studies focusing on the region.’ Envisioned were the use of theoretical frameworks and methods of the social and other sciences and the development of new ones specifically for regional analysis.

This interest in the region as a subject of study had important and immediate political and intellectual antecedents in the United States. In 1927 the regional plan for New York considered economic factors in metropolitan growth and spatial form, largely through the work of Robert Haig. In 1933 the Tennessee Valley Authority became the nation’s first major regional development program, arguably since the western expansion. In 1935, asserting that policies of national development depend on regional factors, the US National Resources Committee asked ‘modern students of regional science’ (10 regional geographers and two regional sociologists) to identify the best type of region to use for regional planning and development. In 1942, Alvin Hansen, perhaps the leading American Keynesian economist and a Harvard professor, advocated a comprehensive postwar expansionary program of regional development on the premise that the effective utilization of resources in any region makes the whole nation richer.

Two organized, interdisciplinary efforts with very different intellectual roots had focused on regions. In the 1920s Lewis Mumford and his colleagues in the Regional Planning Association of America saw cities

as congested and unmanageable. They advocated a mode of thinking and procedural methods aimed at seeing the region as a whole—integrating city and region, and utilizing, not destroying, the natural environment. Closer to the regional scientists because of their emphasis on both science and the region and their scholarly, university base were the southern regionalists. Centered at the University of North Carolina, they were well established by 1931 in terms of research achievements, agendas, and infrastructure. Howard Odum, the acknowledged leader of the effort and a sociologist, argued in 1945 that the nation owes the fullest development of the South not only to the South but also to the nation itself. Similar in ambition but predating the Regional Science Association by a quarter century, Odum advocated a regional approach to the scientific study of human society, arguing that it would entail the integration of method and theory, require the joint effort of the traditional social science disciplines, and create fundamental new understanding of all society, not just its regions.

Members of the Regional Science Association from the outset debated whether regional science would become a new discipline or an interdisciplinary activity. Walter Isard, widely recognized as the father of regional science, was the leading proponent of the disciplinary view. In 1956, he became an economics professor at the University of Pennsylvania, where he formed the Department of Regional Science in 1958. By then he had published 23 papers in the leading economics journals (acknowledging the help of Alvin Hansen in the first five). Yet the Regional Science program would go far beyond economics in its scope and impact. In 1960 the first Ph.D. in Regional Science was awarded to William Alonso, later regional planning professor at Berkeley and Saltonstall Professor of Population Studies at Harvard. Subsequent doctoral recipients took positions in economics, engineering, geography, planning, and other university faculties as well as public agencies and consulting firms.

Isard's prediction that regional science programs would emerge from geography departments and achieve wide recognition before 2000 was not realized. In the 1970s Cornell University established the second Regional Science Ph.D. program, and the University of Illinois initiated a multidepartmental program requiring completion of the Ph.D. in a participating department. By 1976, regional science degree programs or specializations existed at 14 universities internationally. The lone department devoted exclusively to regional science remained the one at Pennsylvania until it was closed during university retrenchment in 1994.

Among the early challengers of the disciplinary view was Lloyd Rodwin, a planning professor at the Massachusetts Institute of Technology. He envisioned the social sciences developing more systematic spatial studies, geography becoming more analytical, and regional science becoming 'a generic name for regional

studies.' This vision has become reality, but not fully. Regional science has become a generic name for a certain kind of regional studies, one whose origin is in mathematical abstraction and quantitative modeling.

Today regional science is an international, interdisciplinary endeavor. It brings together economists, geographers, planners, and others whose research has a regional dimension. It is the confluence of many streams, but arguably also the origin of a few. Each discipline brings different theoretical and methodological approaches to regional science, and each typically has different substantive concerns within a given topic. Migration, for example, interests several disciplines. An economist might focus on migration behavior as a way of measuring quality of life in cities, a demographer on regularities in the age pattern of migration, a geographer on turnarounds in core-periphery migration, a planner in forecasting migration for land use and transportation modeling, a sociologist on the effect of kinship links on migration destinations, and a real estate specialist on the relationship between housing prices and migration.

The numerous scholarly regional science journals and annual conferences create effective forums for exchanging ideas among disciplines. Although dominated by the abstraction and modeling inherited from economics, these forums are eclectic and open. They have been helpful incubators for scholarly groups who went on to create their own forums, be they input-output analysts, spatial social theorists, urban economists, or transportation modelers. Collectively regional scientists actively borrow ideas from a wide range of sources and combine, adapt, and extend them to regional research.

The recognition Isard anticipated has materialized in many ways other than through a proliferation of university departments. The Regional Science Association became the Regional Science Association International (RSAI), an umbrella organization overseeing three major supraregional organizations, the North American, European, and Pacific coordinating councils. Most countries in Asia, Europe, and North America have active regional science associations, and there are several multinational ones, including the British-Irish, French-language, German-speaking, and Australia-New Zealand organizations. Most sponsor annual meetings, and several have journals.

A widely accepted notion is that almost all regional scientists have dual identities, economist and regional scientist, geographer and regional scientist, and so on. For many, the regional science meetings and journals provide the best audience for honing and distributing their research. Others' participation ebbs and flows with changes in their immediate research agendas. No complete census of this fluid regional science community exists. The associations, including those in China, India, and Brazil, may have over 10,000 members combined. RSAI itself has about 2,000 dues-paying members annually.

Regional science journals have become numerous and well established. Five international ones with regional science in their titles are now published by major commercial presses: *Annals of Regional Science*, *International Regional Science Review*, *Journal of Regional Science*, *Papers in Regional Science*, and *Regional Science and Urban Economics*. Echoing the debate a half century ago about the name 'regional science,' some regional science associations leave it out of their journal titles, for example, *Journal of Regional Analysis and Policy*, *Review of Urban and Regional Development Studies*, or *Review of Regional Studies*.

Book series, position announcements, honorary degrees, and external research prizes are other marks of regional science's role within the social sciences. Noteworthy, too, are the National Science Foundations program in Geography and Regional Science in the US and similar ones in other countries. Perhaps the most spectacular recognition of regional science occurred when the Emperor and Empress of Japan personally welcomed participants to the RSAI World Congress in 1996.

1. Subject Matter

The international regional science community studies a wide range of topics, sometimes ambitiously defined as those social, economic, political, and behavioral phenomena that have a spatial dimension. A common, narrower view groups regional science research under three themes: location theory, methods of regional and spatial analysis, and regional development and policy.

Location theory is the study of what social and economic activities are located where and why. Traditionally, emphasis has been on the location of firms and households. The latter considers inter-regional and international migration and labor markets, as well as intrametropolitan residential mobility and housing markets. Early contributions focused on the role of transportation costs and proximity to suppliers and markets. Recognizing interdependencies of behavior focuses attention on agglomerations or clusters in space, as well as questions of land use, urban form, housing markets, city size, hierarchies among and systems of cities, market thresholds, and market boundaries. Recognizing interactions among activities and individuals distributed over space raises questions about commuting, trade flows, congestion, transportation, other infrastructure, facility location, taxes, user fees, local public finance, and diffusion of innovations and growth.

The phrase, methods of regional analysis, describes a growing set of techniques used to identify, analyze, or forecast key regional characteristics or changes. Among the methods are three invented to study national economies but adapted to regions. The inventors received Nobel Prizes in Economic Sciences:

Wassily Leontief (1973) 'for the development of the input-output method and for its application to important economic problems,' Lawrence Klein (1980) 'for the creation of econometric models and their application to the analysis of economic fluctuations and economic policies,' and Sir Richard Stone (1984) 'for having made fundamental contributions to the development of systems of national accounts.' Regional scientists have adapted and extended techniques from demography. The cohort-component model focuses on the components of population change and their probabilities (births, deaths, and migration), grouping the population into cohorts by age, sex, and often race or ethnicity. Regional scientists borrowed the gravity and entropy models directly from physics and found that they described certain aggregate human behavior well. Migration, trade, transportation flows, and other interaction between two places tend to increase with the product of their masses (e.g., populations) and decrease with distance between them. Regional scientists have also adapted and applied optimization techniques to contexts in which spatial interaction is important, such as transportation networks, location of facilities, and land use.

Regional scientists use statistical methods extensively. Yet here, too, important adaptations and extensions permit the study of spatial and regional phenomena. Geographers, in particular, have extended standard statistical concepts to situations where observations have locations, for example, to draw random samples over space or to determine the likelihood that an observed locational pattern could have occurred by chance. Spatial statistics, spatial econometrics, geographical information systems, and other methods of spatial statistical analysis have become major parts of regional science—for economists, planners, engineers, and others as well as geographers. In the case of statistical methods widely used to study travel demand, residential location, migration, and other behavior with a spatial dimension, the pioneer was an economist, Daniel McFadden, who won the Nobel Prize in 2000 'for his development of theory and methods for analyzing discrete choice.'

Regional development and policy focuses on why some regions grow and prosper while others lag and even decline. Research ranges from theories of growth and development to evaluation of specific policies. Persistent questions deal with the divergence and convergence of regional growth rates and incomes, the relationships between core regions and peripheral regions, and the geography and timing of the spread of growth. Other long-standing questions deal with regional labor markets, including the relationship between growth and unemployment and the persistence of wage differences among regions. Key policy issues concern people prosperity vs. place prosperity (should the focus be on helping people directly, for example, through education or out-migration incen-

tives or on helping people indirectly by making the region more competitive through infrastructure, tax incentives, or other actions), ways of making lagging regions more competitive and prosperous, the rationales for regional policy, and the roles of different levels of government.

Policy evaluation has attracted much attention as well. Commonly the objective is to measure the effects of various (nonregional) national policies on regions, for example, free trade treaties, defense spending, or tax reductions. Other times the objective is to evaluate regional policies, for example, by estimating what would have happened to lagging regions in the absence of the regional support programs. Sometimes the emphasis is on comparing alternative regional plans or policy instruments before they are implemented, generally using cost-benefit, cost-effectiveness, or goals achievement methods augmented to include policy objectives, distributional issues, and hard to quantify outcomes.

2. Intellectual Progress

Two easily contested, unavoidable decisions must be made before discussing the intellectual progress of regional science. The first is whose contributions should be counted. The open, fluid nature of the regional science community makes this decision difficult, and the early claims to disciplinary status make it controversial. The perspective here is that the idea itself is more important than the primary tribal self-affiliation of its originator. The key criterion is whether the idea is part of the scholarly conversation in regional science meetings and journals. Mention of an individual's contribution does not imply that the person considers him or herself a regional scientist, although each person mentioned has participated in regional science meetings or published in regional science journals. Debates about whether particular intellectual contributions were initiated by regional scientists, borrowed by regional scientists, or adapted by regional scientists, and who is really a regional scientist are fruitless. The gravity model provides an early case in point. One observer described it as work that may clearly be termed regional science, while another argued that it had been pioneered by outsiders and applied in many disciplines. By far the most interesting fact here is how understanding and use of an idea have evolved.

The second decision concerns the scope of regional science. Something narrower than human activities and phenomena with an important spatial dimension must bound this discussion, but many offer too incomplete a view. Some describe regional science as a toolbox for practical analysis that lacks rigorous theory, while others decry it as a theoretical never-never land with too little realism. The path chosen here recognizes Isard's deliberate efforts to nurture the field

of regional science. He edited several of its journals, commissioned papers, supported dissertations, organized meetings, and shaped an influential book series. Three of those books frame a useful way of discussing regional science: Isard's *Location and Space-Economy* (1956), the Isard-led collaborative effort, *Methods of Regional Analysis* (1960), and a reader edited by John Friedmann and William Alonso and dedicated to Isard, *Regional Planning and Development* (1964). The books' foci—theory, method, and policy—frame the discussion of intellectual progress well.

2.1 Location and the Spatial Economy

The quest here is to model mathematically and show graphically both the determinants and consequences of individual location decisions. Regularities in spatial patterns are readily observable, among them, the emergence and persistence of large cities, clusters of related industries, relatively homogenous residential areas within cities, and regularities in city size distributions. The goal, chiefly of economists who work in this tradition, is to show how these agglomerations materialize from the interaction of self-serving decisions of individuals and firms. Isard's book drew on the pioneering German contributions by Johann Heinrich von Thünen, Alfred Weber, August Lösch, and Walter Christaller.

Two works from the birth period of modern regional science were extraordinarily seminal. Alonso's dissertation, *Location and Land Use*, focuses on transportation costs and its role in residential decisions and the resultant land use patterns within a metropolitan area. Concentric rings of land use emerge from the behavior of utility maximizing individuals. Alonso's starting point like von Thünen's is the existence of an agglomeration, in this case the central business district of a monocentric city. Charles Tiebout, a founding member of the Regional Science Association, argued that 'people vote with their feet' by choosing locations within the metropolitan area that offer the combination of revenue and expenditure patterns that appeal most to them. Thus, spatial mobility among diverse communities within an urban area provides a solution to the collective consumption problem posed by Samuelson and another mechanism by which individual preferences create spatial variation. Much of today's field of urban economics builds on these two contributions.

A contemporary contribution of this genre is *The Spatial Economy* (1999) by Masahisa Fujita, Paul Krugman, and Anthony Venables, which they describe as a continuation and possible validation of Isard's project. It derives agglomerations instead of positing them as a starting point (e.g., Alonso's central business district around which commuters organize themselves). In the 'new economic geography,' as Krugman

calls it, agglomerations result from the tension between centripetal forces that promote concentration of economic activities and centrifugal forces that oppose it. Of key interest is when a spatial concentration of economic activity is sustainable, and when a system without spatial concentration breaks down and concentration emerges. Combining formal theoretical models that include increasing returns to scale and computer simulations to study nonlinear dynamic properties, this work links urban, regional, and trade economics in a way that exceeds the dreams of the formal modelers of a half-century ago. As was the case then, however, many geographers and others within the regional science community do not embrace this formal mathematical, theoretical mainstream economics approach. They are far more interested in the social, cultural, and institutional factors involved in spatial development and in the study of real places and real regional histories in contrast to abstract ones. The new *Journal of Economic Geography* hopes to encourage dialogue among the groups.

2.2 *Methods of Regional Analysis*

A very different early quest of modern regional science was the creation of practical tools for the study of real places, and here regional science has had spectacular success. The input-output model attracted early attention. A series of regional input-output articles appeared in the *Review of Economics and Statistics* between 1951 and 1971. Their foci ranged from the theoretical specifications of the models and the definitions of concepts to studies of the nature of metropolitan economies, the location of a major steel complex, and the effects of arms reduction. This initial period culminated in the creation of two landmark regional input-output models constructed from survey-based data, the Philadelphia model of Isard and Thomas Langford and the West Virginia model of William Miernyk. The next three decades focused on building and using models, with attention paid to the treatment of households and consumption, economic-demographic linkages, and strategies for building the models in a cost-effective way. Here too progress exceeded the dreams of the 1950s, with regional input-output models now widely used for impact studies in the private and public sectors in many countries of the world. In fact, this method of regional analysis has become an industry with data and software readily available from commercial firms, most notably the IMPLAN system within the United States.

Econometric modeling similarly saw astonishing progress from the efforts of pioneering scholars to the commonplace of professional practice. Lawrence Klein sketched out how regional econometric models might be developed as satellites of national models, and his student Norman Glickman developed a fully

operational landmark model of the Philadelphia economy. For roughly a decade, many others built regional econometric models as doctoral dissertations, one becoming the blueprint for the National Regional Impact evaluation system of the US Department of Commerce. Today most states and many companies maintain such forecasting models. This method, too, has become an industry with many companies selling forecasts to private subscribers. A particular noteworthy case is the eclectic REMI model of George Treyz, which grew from a design at a regional science meeting in 1978 into a widely used commercial product.

Modeling land use and transportation flows within a metropolitan region proved to be arguably the most ambitious search for practical methods, and the most spatially explicit. Its intellectual scope is daunting: understanding the determinants of the location of industrial, commercial, residential, and other activities and of the movement of people among those locations. It combined five elements: the theoretical urban economic framework of Alonso and Lowdon Wingo, mathematical programming methods to make them operational starting with the Herbert-Stevens model of land use, gravity models of spatial interaction eventually enriched by the entropy formulations of Alan Wilson, discrete choice statistical methods eventually incorporating McFadden's work, and the transportation forecasting framework established by Robert Mitchell and Chester Rapkin. For four decades, university research groups, notably those led by Britton Harris, Alan Wilson, and David Boyce, have produced a sustained flow of dissertations and other contributions. Yet, unlike the cases of input-output and econometric models, the pioneering research phase did not lead to widespread adoption in routine public and private practice. Dominant practice remains a set of incremental steps—predict population and employment change, decide (usually unaided by models) where those activities will locate, and then use transportation modeling software to forecast the flows among the locations. The research quest continues, large-scale operational models continue to be built (about a dozen were in use in the early 1990s), and the research leads to improvements in the transportation modeling software and other elements of the incremental approach.

2.3 *Regional Development*

The basic quest here is to understand differences in regional growth and development. During the early years of modern regional science it attracted the attention of two more Nobel Laureates in Economics. Gunnar Myrdal was honored in 1974 for his 'penetrating analysis of the interdependence of economic, social and institutional phenomena.' His work on economic theory and underdeveloped regions centered

on the principle of cumulative causation, a 'vicious circle' of economic and noneconomic factors that led activities to cluster in certain regions and caused regional inequalities. Douglass North was honored in 1993 for 'for having renewed research in economic history by applying economic theory and quantitative methods in order to explain economic and institutional change.' He studied regional economic growth using exportable commodities or the export base as the key concept, pointing out the incentives regions have to reduce their processing and transportation costs to better the competitive position of their exports. An influential book by Michael Porter, *The Competitive Advantage of Nations* (1990), builds on these notions of clusters of activities and competitiveness of export industries. He argues that national advantage derives from successful firms concentrating geographically within a country, their competitive success stimulated by their local environment, including government policy at the regional and local level. An earlier formulation remains popular in regional policy. Based on the growth pole research initiated by Francois Perroux, governments concentrate investment in selected urban centers to stimulate the economies of their surrounding areas or draw population away from the crowded, larger metropolitan areas.

2.4 Other

Brief mention of some other work completes a broader picture of regional science. RSAI has awarded its Founder's Medal to Martin Beckmann for research in transportation network theory, location theory, and spatial economics; William Alonso for urban economics, spatial demography, and regional development; Jean Paelinck for methods of regional analysis, spatial econometrics, and regional policy; and David Boyce for transportation network models and urban transportation planning methods. The three most cited articles in the *International Regional Science Review* are on population dispersal from major metropolitan regions internationally (D. Vining 1978), intra-urban residential mobility (J. Quigley and D. Weinberg 1977), and the shift of the US population to nonmetropolitan areas (C. Beale 1977). In the *Journal of Regional Science* they are on the distribution of residential activity in urban areas (J. Herbert and B. Stevens 1960), determinants and consequences of migration (M. Greenwood 1985), and an efficient algorithm to solve the Weber location problem (H. Kuhn and R. Kuenne 1962). In *Regional Science and Urban Economics*, they are on the baby boom and the housing market (N. Mankiw and D. Weil 1989), the demand for housing characteristics (J. Follain and E. Jimenez 1985), and publicly provided inputs and economic growth (T. Garcia-Mila and T. McGuire 1992). The supplementary Bibliography includes well-cited books and recent overviews of aspects of regional science.

See also: Cities, Internal Organization of; Ecological Economics; Location Theory; Migration, Economics of; Region: Geographical; Regional Geography; Space and Social Theory in Geography; Spatial Interaction Models; Spatial Pattern, Analysis of; Transportation Geography; Transportation Planning; Transportation: Supply and Congestion; Urban System in Geography

Bibliography

- Anselin L 1988 *Spatial Econometrics*. Kluwer Academic Publishers, Boston
- Fujita M 1989 *Urban Economic Theory*. Cambridge University Press, Cambridge, UK
- Friedmann J, Weaver C 1979 *Territory and Function*. University of California Press, Berkeley, CA
- Hansen N, Higgins B, Savoie D J 1990 *Regional Policy in a Changing World*. Plenum Press, New York
- Henderson J V 1985 *Economic Theory and the Cities*, 2nd edn. Academic Press, Orlando, FL
- International Regional Science Review* 1995 **17**(3) and **18**(2) (Regional Science at 40 issues)
- Isard W, Azis I J, Drennan M P, Miller R E, Saltzman S, Thorbecke E 1998 *Methods of Interregional and Regional Analysis*. Ashgate Publishing, Aldershot, UK
- Journal of the American Planning Association* 1994 **60**(1): 3-44 (Large Scale Models symposium)
- Journal of Regional Science* 1985 **24**(4) (Silver Anniversary issue)
- Krugman P 1991 *Geography and Trade*. MIT Press, Cambridge, MA
- Lahr M, Miller R (eds.) 2001 *Regional Science Perspectives in Economic Analysis*. Elsevier, Amsterdam
- Markusen A 1985 *Profit Cycles, Oligopoly, and Regional Development*. MIT Press, Cambridge, MA
- Miller R E, Blair P D 1985 *Input-Output Analysis Foundations and Extensions*. Prentice-Hall, Englewood Cliffs, NJ
- Nijkamp P (ed.) 1986 *Handbook of Regional and Urban Economics*. North Holland, Amsterdam
- Plane D A, Rogerson P A 1994 *The Geographical Analysis of Population*. Prentice-Hall, Englewood Cliffs, NJ
- Wilson A G 1970 *Entropy in Urban and Regional Modeling*. Pion Press, London

A. Isserman

Regionalism and Cultural Expression

A region is a geographic concept that denotes a set of places understood as having something in common. While this common ground is typically geographic, political, and/or economic, it gives rise to shared forms of cultural expression as well; conversely, shared culture may encourage political or economic linkage, even perceived geographic unity. Regional culture refers to these shared cultural patterns most generally, while regionalism is reserved for a highly favorable evaluation of the culture of place. Research has

addressed questions about regionalism's political implications, its persistence under conditions of modernity, and its capacity to ground identities under conditions of postmodernity.

1. Definitions

When people refer to a region, they are indicating a set of places that they view as having some common characteristic. Regions can be specific to a small group or organization; for example a regional sales manager might cover a specific territory that no one outside the firm would think of as a region. They can be individual, as in 'the region where I grew up.' And they may involve an exquisite local knowledge. This essay concerns regions having a wider application than these.

Usually the places included in a region are contiguous; if they are not, the plural form gets used, as in the 'polar regions.' The basis for defining a region may be purely geographic (the Mediterranean region) or historical (the Spanish empire), but the places within any single region usually come to share more than geography or history. Regions are or tend to become economic (the Ruhr (*Ruhrgebiet*)), ethnic (Yorubaland in West Africa), or cultural (the Bible Belt of the American South). Political boundaries may or may not contain a region, which can include all or parts of different nation states. Bound by their common characteristics, regions are often at odds with the political unit(s) that claims sovereignty over them (e.g., the Kurdish region of Turkey and Iraq).

Even in the case of regions defined strictly by geography or political necessity, some degree of common culture emerges. Climate, topography, and relationships to the sea will influence agricultural and commercial activities. Thus, in the most obvious illustration of this point, political antagonists like the Palestinians and Jews of the Middle East prepare similar food in similar ways. For this reason scholarship on such regional forms of cultural expression once took them as natural, the straightforward products of material circumstances and history. Such shared characteristics are weak versions of regional culture.

Cultural regions typically derive from a preexisting geographic, economic, or political region, but in some cases the cultural common ground can come first. Nations that share the same language become trading partners, for example, and elaborate other ties; even here, however, there is usually the prior political connection of colonialism. Religion would offer another example of this potential, for example in the Organization of the Islamic Conference. Stretching from Indonesia to Nigeria, this is hardly a territorial region but is sometimes envisioned as such, and it constitutes the potential for economic and political links.

Regional culture comes in a strong version as well. Geographer D. W. Meinig (1986) defined regional culture as, 'that which is characteristic of a group of people who are deep-rooted and dominant in a particular territory, who are conscious of their identity as deriving from a common heritage, and who share a common language and basic patterns of life' (p. 80). He contrasts the seventeenth-century Welsh region, which shares geography, ethnicity, language, and a sense of self-identity, with the West Country, a much vaguer region with its own traditions, dialect, and ancestral families but no unmistakable 'ethnic' distinctions. The West Country was a 'relatively subdued form of regional culture, one which its members are conscious of but not normally militant about, though strong feelings of identity may be aroused under special circumstances' (p. 81).

In light of this distinction between strong and weak regional cultures, manifest versus latent, regionalism is a strong and self-aware form of regional culture, and sectionalism is an intolerant and divisive form of regionalism. The term regionalism generally has a positive connotation, referring to things that win local and often external approval; other strong but destructive manifestations of regional culture, such as the vendetta tradition in Corsica, are not usually designated as regionalism. Regionalism aggressively promotes its own symbolic expression, which can reinforce, or even generate, social and political bonds.

2. Real Regionalism, Debatable Impact

Considerations of regionalism used to begin from a bimodal model: urban culture versus folk culture; cosmopolitan versus provincial; center versus periphery. In this view regionalism was often conflated with localism, the residual cultural forms that hegemonic centralized culture had not yet obliterated. Regional culture, although it made meaningful the lives of people in the region and perhaps enriched the larger cultural pool, was one of those things that was likely to fade away under industrial capitalism and centralized political authority. Research mapped (and sometimes celebrated) regional differences, and it also addressed the impact of these cultural differences on the larger political units.

For example, in the 1930s, sociologist Howard W. Odum argued that the American South remained backward because it lacked social institutions that could facilitate social change while protecting traditional culture; the pro-business New South could not reconcile change with continuity. Odum concluded that the South must reinvent its debilitating local culture by blending it with national culture into regionalism, whereby (in his somewhat idiosyncratic definition) the local was part of the larger unity. In Odum's view, sectionalism was a pathological form of localism, the result of cultural inbreeding (Odum and

Moore 1938, p. 43), although many contended that this regional/sectional distinction was hard to maintain. Odum saw literary and aesthetic regionalism as an undeniable fact—he catalogued over two thousand regionalist titles from two decades of *Publishers' Weekly*—but rather than local color provincialism, which would smack of sectionalism, he maintained that 'regional portraiture is only fundamental units in the great American literary fabric.' True regionalism, in other words, contributes to the health of both region and nation.

As the New South has become both established and even normative on a national level, the regionalism/sectionalism distinction has faded. Social scientists like John Shelton Reed (1986) have continued to assess the degree of Southern regional variation that persists. Analyzing the economic and political differences among a nation's regions is a standard activity on the part of both social scientists and the popular press.

A contemporary political scientist who has made a strong case for the political impact of regional culture is Robert Putnam (1993). Using data from Italy's twenty administrative regions, Putnam shows how 'institutional performance'—effectiveness of and satisfaction with regional government—is high in Italy's north and low in the south. He connects this with presence or absence of civic traditions, which he traces back to medieval period; 'the astonishing tensile strength of civic traditions testifies to the power of the past' (p. 162). Why is the past so powerful? Putnam draws on theories of social capital to show how norms, whether of 'never cooperate' or of trust and reciprocity, produce stable cultures, with only the latter—'civic equilibrium'—able to solve the collective action problem. Culture and structure reinforce each other.

Sociology and urban studies have renewed attention to place as a variable, as in Douglas Massey and Nancy Denton's study of the devastating impact that inner-city spatial isolation—'American apartheid'—has on African-Americans (see also Gieryn 2000). Ethnographers study the symbolic and aesthetic worlds such conditions produce, as when Elijah Anderson analyses the intricate set of signs and behaviors that black youth must master to steer between trouble with the police and with their peers (E. Anderson 1999). While this type of research is often at the level of neighborhood, some employ a broader, regional canvas.

A good example of regional cultural analysis is Francis Violich's 'urban readings,' in which he offers a phenomenological account of the intersection of ecology, urban functions, and cultural history on the Dalmatian coast, now in Croatia (1998). Ecologically, the region possesses rich sea resources along with a poor, mountainous land where the limestone foundation makes water collection a societal preoccupation. Isolation and the need to manage their water supply and harbors has produced strong urban cul-

tures, independent of one another and of whichever regime—in the past thousand years these have been Hungarian, Venetian, French, Austrian, Marxist, and Croatian—was currently in power. A regional culture of autonomy, tradition, and resistance to change has been the result, making Dalmatian cities like Dubrovnik both elegant and backward. Violich is not just describing regional culture but is also making the case for regionalism, for he is concerned about the human cost of losing our 'rootedness to place.' Older towns that reflect regional aesthetics and local building materials become 'treasures of the past,' standing in sharp contrast with the anesthesia of today's cities, 'hodge-podges of synthetic eclecticism and tiresome look-alikes, grown too large to allow the common citizen to claim a sense of identity with them or pride of belonging' (p. 51). This association of identity with territory, the spiritual nourishment that comes from being 'at home' somewhere, is characteristic of regionalism. It is surprisingly parallel to the inner-city gang members' desire to be with their 'homies' in a secure and familiar territory.

3. *Debatable Regionalism, Real Impact*

Scholars of regionalism like Putnam or Violich assume that regions are both naturally based and historically configured. They are materialist in that they see the natural features of the land as encouraging certain economic formations, and these in turn give rise to political and cultural patterns distinctive to the region. Yet even these scholars recognize that human ties to any particular region have become far more optional than in the past. In his discussion of identity, Violich points out that people seek environments compatible with their desired self-images in a way that was not necessary in earlier, more homogeneous cultures that provided a secure sense of belonging. 'Today, in contrast, with great individual opportunities for regional mobility, one can choose among diverse physical and cultural environments in various parts of a large city or metropolitan area as a way of individualizing one's cultural expression' (p. 59).

This is not something new, some scholars have asserted. Spatial boundaries and their cultural representations have always been subject to decisions made in the pursuit of interests. Beginning in the 1970s the view emerged that politics and the power of dominant groups, not geography, has always shaped humanity's mental and physical maps. Much of this literature centered on the origins of nationalism, as in the theses of 'the imagined community' (B. Anderson 1991) or 'the invention of tradition' (Hobsbaum and Ranger 1983), although smaller communities were similarly fictitious (Laitin 1986, Griswold 1992). Once in place, the arbitrary or fanciful boundaries could have powerful and lasting consequences, and indeed some like 'the nation' became unavoidable, as the 'world culture'

thesis pointed out. Cultural expression played a role in such inventions, both encouraging them, as in Benedict Anderson's discussion of the role of newspapers, and in sustaining them, as in critiques of heritage movements.

Revisionist historical and sociological research, such as that of the New Western Historians, has focused on the manipulation of collective memory (Ayers et al. 1996). Regional representations, in this view, are highly suspect. For example, a universally recognized regional symbol is the New England village with its white steeples and town commons. Stephen Nissenbaum (1996) contends that New England towns were originally not different that other towns in British North America, for there was no reason for farmers to live in villages. The compact village arrangement was a consequence of the commercial revolution in the late eighteenth and early nineteenth centuries, when it made sense for merchants and professionals both to set up business in and live in villages. And the 'pastoralization' of town commons—previously they had been bare patches, rather like parking lots—didn't take place until after the Civil War where local Village Improvement Societies, 'usually dominated by businessmen or newcomers to the community,' began to plant grass and elm trees. Litchfield, Connecticut actually hired Frederick Law Olmstead, designer of Central Park, to 'restore' their common.

Once region became seen as subject to human agency, the door was open to a lively debate on the relationship between regionalism and modernism, no longer assuming that the latter simply came along and eclipsed the former. Yatsuka (1998) describes how internationalist architects believed modernism to be 'globally shared beyond a regional understanding.' In architecture, unlike in politics, the international (modernist) preceded the regional: 'contrary to the simple notion of "regional things," the idea of "regionalism" arose due to the appearance of "internationalism"' (p. 167). Regional reaction—'a romantic nationalism'—arose in response to modernism, as in the case of the Finnish modernist Alvar Aalto. Yatsuka cites 'critical regionalism'—Kenneth Frampton's term for resistance against corporate modernism on the one hand and the postmodern reaction (compensatory façade) on the other—critical regionalism couldn't have been conceived of without modernism. He looks at Japanese architects such as Kenzo Tange, who both used and altered traditional (regional) forms such as the *sukiya* (sixteenth-century teahouse) and *shinden-zukuri* (nobleman's palace), grafting elements onto modernist layouts; by asserting 'we inherit spirit, not individual idioms' (p. 177), Tange avoided nostalgia just as he avoided placeless internationalism. Along similar lines James Dennis (1998) shows how the regionalist painters of the thirties, despite being given to bouts of antimodernist rhetoric, demonstrated strong modernist influences. The dichotomy between regionalist and modernist, like that between provincial

and cosmopolitan, has been not just overdrawn but misconceived.

During the 1970s and 1980s some theorists of postmodernity argued that in the world of electronic communications and unrestricted cultural flows, place-based notions like 'region' and 'regional culture' were irrelevancies, at best productions for a tourist market intent on collecting scenic memorabilia (MacCannell 1976). The 'no sense of place' thesis maintained that electronic media has severed the link between social place and physical place. '[A]s place and information access become disconnected, place-specific behaviors and activities begin to fade' (Meyrowitz 1985, p. 148). This was something of a return to the bi-model model, only instead of industrialization, capitalism and imperialism, now electronic media and globalization were bleaching regional cultures out of the postmodern fabric.

More recent research has drawn back from this position, in part due to the undeniable assertiveness of place-based social formations, traditional or new-mint (Logan and Molotch 1987). Indeed, globalization itself, to the degree that it makes state boundaries less significant, gives rise to place-specific formations like the much discussed 'Europe of the regions.' Proliferation of communications channels encourages formation of new possibilities for the assertion and communication of cultural identities, and in some cases regionalism seems to find new life through these media (Griswold and Engelstad 1998).

4. Future Research Directions

Research at the end of the twentieth century accepts the constructed nature of regional cultures but takes a W. I. Thomas view that such regional cultures are real in their consequences, regardless of their origins. Making space into place, 'mapping the invisible landscape' through words and symbols (Ryden 1993), is a cultural process that, while by no means innocent of politics and economics, seems to fulfill a human need for meaningful spatial orientation. Regionalism and the culture of place represent, following this line of thought, a psychologically fundamental dimension of cultural expression. Thus, the regionalist impulse is not likely to atrophy, regardless of changes in mobility or communications technology.

Scholars have been addressing the lines along which regionalism gets configured (Ayers et al. 1996), the circumstances under which regionalism expands (Griswold and Engelstad 1998), the ability of regional culture to offer a viable alternative to class and racial hierarchies (Stack 1996), and the degree to which regionalism influences social and political behavior (Putnam 1993). Particular attention is being paid to how a regional culture is apt to be less the result of shared local experience, as celebrators of regionalism have maintained, and more the consequence of elite

conflicts brought about by external social changes. Anthony Cosgrove (1993) describes how the sixteenth-century Venetian elite still had its fortune tied to maritime power, and felt increasingly insecure due to French pressure. One response was a rash of estate construction in the rural interior, 'terra firma.' While these were working estates, they were ideological as well, representations of the harmony of various spheres: macrocosm, microcosm, divine will, world, theatre, human body, buildings, and lands. In an era when professionals were rising in status, architects like Andrea Palladio (1508–1580) created this regional culture of harmony and control. Palladio's use of classical forms and his emphasis on harmonizing buildings with landscape combined 'Roman imperial grandeur [with] Venetian sensitivity to atmosphere, site, and situation' (p. 20). Thus, the Palladian rural landscape both expressed the unease of the urban regional elite and reassured them of their centrality within an orderly cosmos.

This type of analysis, while recognizing the fundamental human desire to connect to place, looks for how political and economic struggles shape the way this connection gets articulated as regional cultural expression. This is one of the liveliest lines of research at the beginning of the twenty-first century and seems likely to continue. A second direction for the future is a renewed interest in the relationship between technology and regionalism. Studies of how elite conflicts and the pursuit of economic interests mold regional culture have tended to downplay the impact of new communications media, but the question of technological change is being revived by the Internet. The impact of the Internet on regionalism is unclear, with possibilities ranging from the theory that it will constitute the final blow to 'place' and other material forms of identity to the theory that it will encourage new 'tribal' groupings, including those of region. A new generation of researchers, unconstrained by former assumptions about how modernization would devastate regional culture but also largely free from a romanticist reverence for the traditional, will undoubtedly explore these possibilities.

See also: Ethnonationalism: Cultural Concerns; Locational Conflict (NIMBY); Neighborhood: General; Neighborhood Revitalization and Community Development; Place in Geography; Postmodernism in Geography; Regional Geography; Regional Identity; Residential Segregation: Sociological Aspects; Space: Linguistic Expression

Bibliography

- Anderson B 1991 *Imagined Communities: Reflections on the Origin and Spread of Nationalism*, rev. edn. Verso, London
- Anderson E 1999 *Code of the Street: Decency, Violence, and the Moral Life of the Inner City*, 1st edn. W. W. Norton, New York
- Ayers E L, Limerick P N, Nissenbaum S, Onuf P S 1996 *All Over the Map: Rethinking American Regions*. Johns Hopkins University Press, Baltimore
- Cosgrove D 1993 *The Palladian Landscape: Geographical Change and Its Cultural Representation in Sixteenth-century Italy*. Pennsylvania State University Press, University Park, PA
- Dennis J M 1998 *Renegade Regionalists: The Modern Independence of Grant Wood, Thomas Hart Benton, and John Stuart Curry*. University of Wisconsin Press, Madison, WI
- Gieryn T F 2000 A space for place in sociology. *Annual Review of Sociology* **26**: 463–96
- Griswold W 1992 The writing on the mud wall: Nigerian novels and the imaginary village. *American Sociological Review* **57**: 709–24
- Griswold W, Engelstad F 1998 Does the center imagine the periphery?: State support and literary regionalism in Norway and the United States. *Comparative Social Research* **17**: 129–75
- Hobsbaum E, Ranger T (eds.) 1983 *The Invention of Tradition*. Cambridge University Press, Cambridge, UK
- Laitin D 1986 *Hegemony and Culture: Politics and Religious Change Among the Yoruba*. University of Chicago Press, Chicago
- Logan J R, Molotch H L 1987 *Urban Fortunes: The Political Economy of Place*. University of California Press, Berkeley, CA
- Maffesoli M 1996 *The Time of the Tribes: The Decline of Individualism in Mass Society*, trans. Smith D. Sage, London
- Massey D S, Denton N A 1994 *American Apartheid: Segregation and the Making of the Underclass*. Harvard University Press, Cambridge, MA
- MacCannell D 1976 *The Tourist: A New Theory of the Leisure Class*. Schocken Books, New York
- Meinig D W 1986 *The Shaping of America: A Geographical Perspective on 500 Years of History. Volume 1 Atlantic America, 1492–1800*. Yale University Press, New Haven, CT
- Meyrowitz J 1985 *No Sense of Place: The Impact of Electronic Media on Social Behavior*. Oxford University Press, New York
- Nissenbaum S 1996 New England as region and nation. In: Ayers E L, Limerick P N, Nissenbaum S, Onuf P S (eds.) *All Over the Map: Rethinking American Regions*. Johns Hopkins University Press, Baltimore, Chap. 2
- Odum H W, Moore H E 1938 *American Regionalism: A Cultural–Historical Approach to National Integration*. Henry Holt, New York
- Putnam R D 1993 *Making Democracy Work: Civic Traditions in Modern Italy*. Princeton University Press, Princeton, NJ
- Reed J S 1986 *The Enduring South: Subcultural Persistence in Mass Society*. University of North Carolina Press, Chapel Hill, NC
- Ryden K C 1993 *Mapping the Invisible Landscape*. University of Iowa Press, Iowa City, IA
- Stack C 1996 *Call to Home: African Americans Reclaim the Rural South*. Basic Books, New York
- Violich F 1998 *The Bridge to Dalmatia: A Search for the Meaning of Place*. Johns Hopkins University Press, Baltimore
- Yatsuka H 1998 Internationalism versus regionalism. In: Ferguson R (ed.) *At the End of the Century: One Hundred Years of Architecture*. Museum of Contemporary Art, Los Angeles

W. Griswold

Copyright © 2001 Elsevier Science Ltd.
All rights reserved.

12939

Regression Discontinuity Design

1. Definition

The regression discontinuity (RD) research design is a quasi-experimental design that can be used to assess the effects of a treatment or intervention. Unique to the RD design is that participants are assigned to groups solely on the basis of a pretreatment cutoff score. The name ‘regression discontinuity’ indicates that a treatment effect appears as a ‘jump’ or *discontinuity* at the cutoff point in the *regression* function linking the assignment variable to the outcome. In its simplest form, the design has two groups (those scoring above and below the cutoff), a pretest (the assignment variable) and a post-test (the outcome). More complex variations are also possible.

To many, the RD design seems counterintuitive. Using a cutoff to assign participants to the treatment and comparison groups creates a pretreatment group nonequivalence that seems like it should lead to selection bias. But while the design does induce initial nonequivalence, it does not necessarily lead to biased treatment effect estimates. Instead of assuming pretreatment equivalence in measured and unmeasured means and variances—as with the randomized experiment—in the absence of a treatment effect the RD design requires the two groups to be equivalent in their pre–post regression functions. This assumption can be tested.

The major attraction of RD designs is that they can be used to estimate the effects of treatments given to those who most need or deserve them, provided that need or merit is determined as a qualification score on the assignment variable (i.e., the cutoff) and nothing else. Because the design does not require some needy individuals being assigned to a no-treatment group, it has ethical advantages over experiments for assessing treatment effects.

2. History

The RD design was first used in psychology by Thistlethwaite and Campbell (1960) and was first discussed in detail in Campbell and Stanley (1963). It has been repeatedly reinvented—by the economist Goldberger (1972), the statistician Rubin (1977), the educators Tallmadge and Horst (1976), and by the public health researchers Finkelstein et al. (1996). Despite such independent reinvention, the RD design has not been used much outside of compensatory education where students scoring below some predetermined cutoff value on an achievement test are then assigned to remedial training (Trochim 1984). But it has also been used to study how unemployment benefits influence recently released prisoners (Berk and Rauma 1983) and how different levels of in-

carceration affect those in prison (Berk and de Leeuw 1999). A variation was also used to study how class size affects test scores (Angrist and Lavy 1999).

The low frequency of published examples may be due to several factors. One is its recency. Another is that RD designs require administrators to assign participants to conditions *solely* on the basis of quantitative indicators, restricting the role of judgment, discretion, or favoritism. Finally, understanding the design depends on some familiarity with regression analysis, making the strategy difficult to convey to nonstatistical audiences.

3. The Basic RD Design Structure

Simulated data for an RD design are shown in Fig. 1. It depicts the example of a compensatory treatment—perhaps a reading program designed to help students who initially score poorly (defined as below the cutoff) on a standardized reading measure. The figure depicts the standard bivariate plot of post-test and pretest for 500 simulated participants. Data points for the treatment recipients are represented with ‘+’ while the comparison cases are ‘o.’ The cutoff value in this example occurs at the middle of the pretest distribution (pretest = 50) and is indicated by the vertical line. All students scoring below 50 are defined as needing compensatory training and get the treatment, while the remainder (i.e., those to the right of the cutoff) are assigned to the comparison group.

The solid lines represent the straight-line regression of post-test on pretest for each group. In the absence of a treatment effect, the critical assumption is that the regression line in the comparison group would continue to the left of the cutoff into the treatment group region, as indicated by the heavy dashed line. That is,

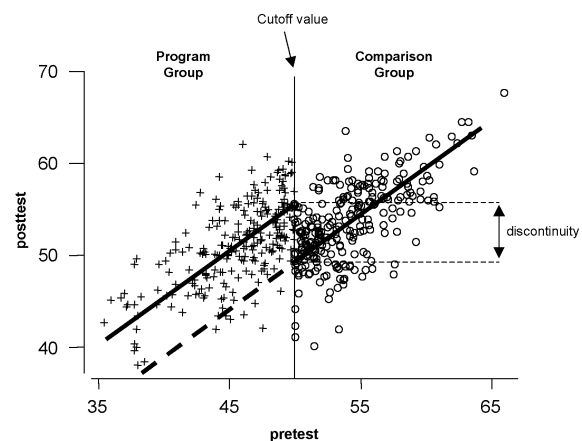


Figure 1

Basic regression discontinuity design for a compensatory treatment showing a positive effect

if the reading program does not work, then the observed treatment group regression line would be where the dashed line is. The fact that the observed treatment group line is displaced from this 'expected' line suggests a treatment effect—treatment group participants scored higher on the post-test (i.e., vertically) than would have been predicted from the comparison group pre-post relationship. The size of the treatment main effect is estimated as the vertical jump or discontinuity between the regression lines at the cutoff. Assuming that higher scores on both measures indicate better performance, one can conclude from the figure that the treatment improved participant performance by the amount of the vertical discontinuity.

In experimental designs one assures through random assignment or otherwise that the treatment and comparison groups are initially equivalent. Otherwise, post-treatment differences can be attributed to the intervention. In the RD design one obviously does not expect this kind of equivalence. Instead, it is assumed that the *relationship* between the assignment variable and the outcome is equivalent for the two groups—that is, the same continuous regression line describes both groups. Thus, interpretation of the RD design depends mostly on two factors: (a) that no alternative cause operates at or near the cutoff; and (b) that we can perfectly model the pre-post relationship. These issues constitute the major problems in the statistical analysis of the RD design, particularly the second one.

4. Design Considerations and Variations

4.1 Selection of the Cutoff

The choice of cutoff value can be made solely on the basis of available resources. If a program can only handle 25 persons and 70 apply, then a cutoff point can be selected that distinguishes the 25 most 'needy' persons from the rest. The cutoff can also be chosen on substantive grounds. If the assignment measure is an indication of severity of illness measured on a 1–7 scale, relevant experts might contend that all those scoring 5 or more are ill enough to deserve treatment. So, a theoretically justified cutoff of 5 is used.

4.2 Assignment Variations

While the use of a pretreatment cutoff value distinguishes the RD design, it is often difficult to implement. In many situations it does not appear to allow room for professional judgment or discretion. In fact, the design does not preclude incorporating such judgment so long as it can be quantified or explicitly accounted for when specifying the cutoff value.

Violation of the assignment rule leads to 'mis-assignment.' It seems to be most prevalent at scores around the cutoff and is often labeled as the 'fuzzy' RD design. Standard analyses will yield biased estimates of treatment effect in the fuzzy RD case. While statistical procedures for dealing with mis-assignment have been suggested (Trochim 1984), mis-assignment is better avoided if possible.

4.2.1 Multiple cutoff points. RD designs are not limited to a single cutoff value. In the absence of treatment the assumption is that the pre-post relationship can be described as a single continuous regression line extending over the entire range of the assignment scores. Even when multiple cutoff points are used, the comparison group's pre-post relationship is still used as the counterfactual for projecting where other groups' regression lines should be if there is no treatment effect.

The simplest multiple cutoff case involves two cutoff scores. In this variant, persons scoring above the higher cutoff might be assigned to one treatment, those scoring below the lower cutoff to another, and those scoring within the cutoff interval might be controls. The assignment of treatments to each part of the assignment distribution can be made in many ways, including by random assignment.

Considerable work has been done on another way of 'coupling' the RD design to the randomized experiment. This is when individuals scoring in a range just above or below an originally selected cutoff value are randomly assigned to conditions, while those scoring further below and further above the cutoff are analyzed as from an RD design (Boruch 1973). Trochim and Cappelleri (1992) discuss four other ways of coupling random assignment to RD design features.

Multiple cutoff points are also involved in designs where the treatment is implemented a series of times (i.e., in phases), perhaps beginning with the most needy before successively giving it to those less needy. In the first phase, one uses a single cutoff value, assigning those most in need to the treatment and then measuring post-treatment performance. Then the cutoff value would be moved to include the next most needy group on the assignment measure. Eventually, everybody could receive the treatment, but in groups ordered by need and phase. This is particularly useful in organizations seeking to 'roll-out' a treatment over time, but eventually serving all.

4.2.2 Multiple assignment measures. The requirement of strict adherence to the cutoff criterion entails a single quantitative indicator. When this does not capture the degree of pretreatment need well, Trochim (1990) discusses two multivariate strategies.

The first involves the use of several separate measures, each having its own cutoff value. This is most feasible when a prospective participant must meet the different cutoff criteria on measure 1 ‘and’ measure 2 ‘and’ measure 3, etc. Less feasible is an ‘or’ rule when assignment depends on meeting a fixed number of criteria from a larger set—say, 5 out of 10. The second involves assignment according to a new variable that is a composite of several individual measures. Using multiple measures this way can help incorporate more judgment into a quantitative assignment strategy.

4.3 Treatment Variations

When treatment comparisons are absolute, the control group receives no formal treatment or a placebo treatment. When they are relative, the comparison group receives an alternative treatment, very often the current treatment to which a novelty is to be compared.

An advantage of the RD design is that it enables assignment of progressively riskier treatments to those in greater need. Assume three treatment programs: the standard, plus two increasingly riskier experimental ones. It is then possible to have two cutoff points and to assign the riskiest treatment to the neediest patients, the least risky (i.e., standard) treatment assigned to the least needy, and the moderately risky treatment assigned to those falling between the two cutoffs. (If no *a priori* basis exists for making risk judgments, an alternative would be to use a single cutoff and to assign those least in need to standard treatment while randomly assigning those on the other side of the cutoff into one of the two experimental protocols.) RD designs are flexible for examining treatment variations.

5. Post-treatment Measurement Variations

In RD designs, pre- and postmeasures do not have to be on the same or equivalent measures. For instance, one might assign persons to a health education treatment on the basis of household income but then examine its effects on attitudes towards health. The assignment variable is an economic indicator and the outcome is a cognitive one.

The RD design is not restricted to a single outcome, and for each post-treatment measure there can be a separate RD design and analysis. But it will often be useful to create an aggregate outcome or outcomes. For instance, we may have many knowledge items that are scaled to give a total and subtest scores for mathematical reasoning, computational skills, verbal reasoning, analogies, and so on. We might then conduct separate analyses for both the total score and each subtest.

Outcome measures need not be continuous normal variables. While such distributions facilitate RD anal-

ysis, they are not necessary for it. Thus, Berk and Rauma (1983) evaluated the effects of a California law that extended unemployment benefits to released prisoners who had not previously been eligible. In order to qualify, prisoners had to earn at least \$1,500 working in the prison over the 12 months prior to release. Would extending these benefits reduce subsequent recidivism—a dichotomous outcome where 1 indicated recidivism and 0 staying out of prison? To analyze this outcome, Berk and Rauma (1983) relied on a linear random utility model related to the binary logit model. When continuous normal outcomes are not feasible, the analyst must specify a statistical model to account for the specified distributional form of the outcome.

5.1 Internal Validity

Internal validity refers to the degree to which causal inference is reasonable because all alternative explanations for an apparent treatment effect can be ruled out. The RD design is very strong with respect to internal validity because the process of selection into treatments is fully known and only factors that would serendipitously induce a discontinuity in the pre-post relationship at the cutoff can be considered threats. However, validity also depends on how well the analyst can model the true pre-post relationship so that underlying nonlinear relationships do not masquerade as discontinuities at the cutoff point.

5.2 Measurement Error

Stanley (1991) has argued that the RD design yields biased estimates of the treatment effect when the assignment variable is measured with error. Cappelleri et al. (1991) refuted this for situations with no interaction effects and Trochim et al. (1991) for the interaction case. The reason such error does not induce bias here although it does in quasiexperiments is that assignment to treatment is by the fallible assignment variable and not by some latent underlying selection process, as happens in quasiexperiments.

5.3 Statistical Power

It takes approximately 2.75 times as many participants in an RD design to achieve statistical power comparable to a simple randomized experiment (Goldberger 1972). When RD is combined with random assignment within an interval around this cutoff, 2.48 times the number of respondents are needed if 20 percent of participants are randomized; 1.96 times with 40 percent randomization; 1.46 with 60 percent; and 1.14 with 80 percent (Cappelleri and Trochim 1994, 1995).

6. Analysis

6.1 Assumptions of RD Analysis

Five central assumptions must be made in order for responsible analysis of data from an RD design:

(a) *The cutoff criterion.* The cutoff criterion must be followed without exception. Otherwise, a selection threat arises and estimates of the treatment effect are likely to be biased.

(b) *The pre-post distribution.* The true pre-post distribution must be known and correctly specified as linear, polynomial, logarithmic, or the like. Reichardt et al. (1995) point out the difficulties in specifying the appropriate model, especially when some curvilinearity exists in the pre-post relationship.

(c) *Comparison group pretest variance.* There must be a sufficient range of pretest values in the comparison group to enable adequate estimation of the pre-post regression line for that group. Variability in the treatment group is also desirable, though not strictly required because one can project the comparison group line even to a single treatment group point.

(d) *Continuous pretest distribution.* Both groups must come from a single continuous pretest distribution. In some cases one might find intact groups that serendipitously divide on a measure so as to imply some cutoff (e.g., pents from two different geographic locations). But such naturally discontinuous groups must be used with caution.

(e) *Treatment implementation.* It is assumed that the treatment is uniformly delivered to all recipients—that is, they all receive the same dosage, length of stay, amount of training, or whatever. If not, it is necessary to model explicitly the treatment as implemented, thus complicating the analysis considerably.

6.2 A Model for the Basic RD Design

The model presented here is for the basic RD design. Given a pretest assignment measure, x_i , and a post-treatment measure, y_i , the model can be stated as follows:

$$y_i = \beta_0 + \beta_1 \tilde{x}_i + \beta_2 z_i + \beta_3 \tilde{x}_i z_i + \dots + \beta_{n-1} \tilde{x}_i^s z_i + e_i$$

Where $\tilde{x}_i$ = pretreatment measure for individual i minus the value of the cutoff, x_0 (i.e., $\tilde{x}_i = x_i - x_0$); y_i = post-treatment measure for individual i ; z_i = assignment variable (1 if treatment participant; 0 if comparison participant); s = the degree of the polynomial for the associated $\tilde{x}_i$; β_0 = parameter for comparison group intercept at cutoff x_0 ; β_1 = linear slope parameter; β_2 = treatment effect estimate; β_n = parameter for the s th polynomial or interaction terms if paired with z ; e_i = random error.

The major hypothesis of interest is: $H_0: \beta_2 = 0$ tested against the alternative: $H_1: \beta_2 \neq 0$. This model estimates both main and interaction effects at the cutoff point in order to examine changes in means and

slopes. It postulates a pre-post relationship that is polynomial at the cutoff. It also requires subtracting the cutoff score from each pretest score. The term $\tilde{x}_i$ has a superscript tilde to indicate this transformation of the pretest x_i . Finally, the model allows for any order of polynomial function (although certain restrictions are made in specifying the function as described below). Thus, the true pre-post relationship can in theory be linear, quadratic, cubic, quartic, and so on, or any combinations of these.

6.3 Model Specification

The key analytic problem is correctly specifying the model for the data—in this example, as a polynomial model. No simple or mechanical way exists to determine definitively the appropriate model. As with any statistical modeling, the RD analysis requires judgment and discretion and conducting multiple analyses based on different assumptions about the true pre-post relationship.

6.4 Steps in the Analysis

Recall what data the basic RD design provides on each unit. There is a pretest assignment value x_i in the model. Knowing this and the cutoff value allows creation of a new variable, z_i , which is equal to 1 in the treatment group and 0 if not. Finally, there is a post-test score labeled y_i in the model. Given the variables x_i , z_i , and y_i , the steps to be followed in the polynomial analysis are:

(a) *Transform the pretest.* The analysis begins by subtracting the cutoff value from each pretest score, thus creating the term as in the model. This sets the intercept equal to the cutoff so that estimates of effect are made at the cutoff rather than at $x_i = 0$.

(b) *Examine the relationship visually.* It is important to determine whether there is any visually discernable discontinuity at the cutoff. It could be a change in level, in slope, or both. If a discontinuity is visually clear at the cutoff one should not be satisfied with analytic results that indicate no effect. However, if no discontinuity is apparent, variability in the data may be masking an effect and one must attend carefully to the analytic results.

The second thing to look for is the degree of polynomial that may be required as indicated by the overall slope of the distribution, but particularly in the comparison group part. A good approach is to count the number of flexion points (how often the distribution ‘flexes’ or ‘bends’). A linear distribution implies no flexion points, while a single flexion point could indicate a quadratic function. This information is used to specify the initial model.

(c) *Create higher-order terms and interactions.* Depending on the number of flexion points, create transformations of the transformed assignment variable, $\tilde{x}_i$. The rule of thumb is that one goes two orders

of polynomial higher than indicated by the number of flexion points. Thus, if the bivariate relationship appeared linear (i.e., no flexion points), one would want to create transformations up to a second-order $(0 + 2)$ polynomial. The first-order polynomial already exists in the model ($\tilde{x}_i$) and so one would only have to create the second-order polynomial by squaring $\tilde{x}_i$ to obtain $\tilde{x}_i^2$. For each transformation of $\tilde{x}_i$ one also creates the interaction term by multiplying the polynomial by z_i . In this example there would be two interaction terms: $\tilde{x}_i z_i$ and $\tilde{x}_i^2 z_i$. If there seem to be two flexion points in the bivariate distribution, one would then create transformations up to the fourth $(2 + 2)$ power and their interactions. This rule of thumb errs towards overestimating the true polynomial function needed, for reasons outlined in Trochim (1984).

(d) *Estimate the initial model.* The true analysis can now begin. One simply regresses the post-test scores, y_i , on $\tilde{x}_i$, z_i and all higher-order transformations and interactions created in step (c) above. The regression coefficient associated with the z_i term (i.e., the group membership variable) is the estimate of the main effect of the treatment. If there is a vertical discontinuity at the cutoff it will be estimated by this coefficient. One can test the significance of the coefficient (or any other) by constructing a standard t -test. If the analyst correctly overestimated the polynomial function required to model the distribution at step (c), then the treatment effect estimate will be unbiased. However, by initially including terms that may not be needed in the true model, the estimate is likely to be inefficient—that is, standard error terms will be inflated and statistical significance underestimated. But if the coefficient for the effect is highly significant, it would be reasonable to conclude there is an effect. Interaction effects can also be examined. A linear interaction is implied by a significant coefficient for the $\tilde{x}_i z_i$ term.

(e) *Refining the model.* The procedure described thus far is conservative, designed to reduce the chances of a biased treatment effect estimate even at the risk of increasing the error of the estimate (Trochim 1984). So, on the basis of the results of step (d) one might wish to attempt to remove apparently unnecessary terms and re-estimate the treatment effect with greater efficiency. This is tricky and should be approached with caution lest it introduces bias. One should certainly examine the regression results from step (d), noting the degree to which the overall model fit the data, the presence of any insignificant coefficients, and the pattern of residuals. One might examine the highest-order term in the current model and its interaction. If both coefficients are nonsignificant, and the goodness-of-fit measures and pattern of residuals indicate a good fit, one might then drop these two terms and re-estimate the model. One would repeat this procedure until either: (i) all of the coefficients are significant; (ii) the goodness-of-fit measure drops appreciably; or (iii) the pattern of residuals indicates a poor fit.

6.5 Analyses with Multiple Cutoff Points

The basic model described above can be applied directly when multiple cutoff points are used and random assignment is used within cutoff intervals. If assignment within intervals is nonrandom, one must also address the potential for selection bias in the analysis. When multiple cutoffs are used to distinguish separate treatments (i.e., multiple treatments are not assigned within the same cutoff interval), one would have to construct multiple treatment assignment variables for the analytic model (e.g., z_1, z_2, z_3) and all necessary interaction terms. Clearly, as more cutoffs and groups are added, model specification becomes more complex.

6.6 Analyses with Multiple Assignment Measures

Imagine having multiple measures, each with its own cutoff. Each assignment measure is then transformed by having its own cutoff value subtracted from it to create $\tilde{x}_i$. The analysis would then include all transformed assignment measures, group membership, higher-order terms, and interactions. For the simple first-order (or linear) case, one could use the model:

$$y_i = \beta_0 + \beta_1 z_i + \beta_2 \tilde{x}_{1i} + \beta_3 \tilde{x}_{2i} + \beta_4 \tilde{x}_{3i} \\ + \beta_5 \tilde{x}_{1i} z_i + \beta_6 \tilde{x}_{2i} z_i + \beta_7 \tilde{x}_{3i} z_i + e_i$$

This model does not include any two-way assignment variable interactions terms (e.g., $\beta_i \tilde{x}_{1i} \tilde{x}_{2i}$ or $\beta_i \tilde{x}_{1i} \tilde{x}_{2i} z_i$) or any three-way terms (e.g., $\beta_i \tilde{x}_{1i} \tilde{x}_{2i} \tilde{x}_{3i}$ and $\beta_i \tilde{x}_{1i} \tilde{x}_{2i} \tilde{x}_{3i} z_i$) but such an assumption may be reasonable where primary interest is in estimating β_1 , the treatment effect, and not interactions. Clearly, though, the use of multiple assignment variables with higher-order polynomial models will quickly lead to an unwieldy analysis.

The situation is simpler when multiple assignment variables are rescaled to reflect ‘number of assignment variables on which a person meets the criterion’ (e.g., on 5 of 10). This number then becomes the x_i variable and one would conduct the basic analysis described earlier. When multiple assignment variables are combined into a single index, the analysis is also a straightforward application of the basic procedures described initially.

7. Conclusions

From a methodological point of view, inferences drawn from a well-implemented RD design are considered comparable in internal validity to conclusions from randomized experiments. However, the lower statistical efficiency of RD designs, and the resulting sample size demands, may limit the design’s utility. From an ethical perspective, RD designs are com-

patible with the goal of getting the treatment to those most in need. It is not necessary to deny the treatment from potentially deserving recipients simply for the sake of a scientific test. From an administrative viewpoint, the RD design may be directly usable when allocation formulas are the basis for assigning treatments or programs.

With all of these considerations in mind, the RD design must be used judiciously. In general, the randomized experiment is still the first method of choice when assessing causal hypotheses (Cappelleri and Trochim 2000). However, where they are ruled out as impractical, the RD design should be considered as a practical alternative.

See also: Linear Hypothesis: Regression (Basics); Linear Hypothesis: Regression (Graphics); Linear Hypothesis: Regression (Quantile); Quasi-Experimental Designs

Bibliography

- Angrist J D, Levy V 1999 Using Maimonides' Rule to estimate the effect of class size on scholastic achievement. *Quarterly Journal of Economics* **114**: 533–75
- Berk R A, de Leeuw J 1999 An evaluation of California's Inmate Classification System using a generalized regression discontinuity design. *Journal of the American Statistical Association* **94**: 1045–52
- Berk R A, Rauma D 1983 Capitalizing on nonrandom assignment to treatments A regression discontinuity evaluation of a crime control program. *Journal of the American Statistical Association* **78**: 21–8
- Boruch R F 1973 Regression-discontinuity designs revisited. Unpublished manuscript. Northwestern University
- Campbell D T, Stanley J C 1963 Experimental and quasi-experimental designs for research on teaching. In: Gage N L (ed.) *Handbook of Research on Teaching*. Rand McNally, Chicago (Also published as *Experimental and Quasi-experimental Designs for Research*. Rand McNally, Chicago, 1966)
- Cappelleri J C, Trochim W M K 1994 An illustrative statistical analysis of cutoff-based randomized clinical trials. *Journal of Clinical Epidemiology* **47**(3): 261–70
- Cappelleri J C, Trochim W M K 1995 Ethical and scientific features of cutoff-based designs of clinical trials: A simulation study. *Medical Decision Making* **15**(4): 387–94
- Cappelleri J C, Trochim W 2000 Cutoff designs. In: Chow S-C (ed.) *Encyclopedia of Biopharmaceutical Statistics*. Marcel Dekker, New York, pp. 149–56
- Cappelleri J C, Trochim W M K, Stanley T-D, Reichardt C S 1991 Random measurement error does not bias the treatment effect estimate in the regression-discontinuity design: I. The case of no interaction. *Evaluation Review* **15**(4): 395–419
- Finkelstein M O, Levin B, Robbins H 1996 Clinical and prophylactic trials with assured new treatment for those at greater risk. 2. Examples. *American Journal of Public Health* **86**: 696–705
- Goldberger A S 1972 Selection bias in evaluating treatment effects: Some formal illustrations. Discussion Papers, 123–72. University of Wisconsin, Madison, WI
- Reichardt C S, Trochim W M K, Cappelleri J C 1995 Reports of the death of regression-discontinuity analysis are greatly exaggerated. *Evaluation Review* **19**(1): 39–63
- Rubin D B 1977 Assignment to treatment on the basis of a covariate. *Journal of Educational Statistics* **2**: 1–26
- Stanley T D 1991 'Regression-discontinuity design' by any other name might be less problematic. *Evaluation Review* **15**(5): 605–24
- Tallmadge G K, Horst D P 1976 A Procedural Guide for Validating Achievement Gains in Educational Projects. Office of Education, Washington, DC
- Thistlethwaite D L, Campbell D T 1960 Regression-discontinuity analysis: An alternative to the *ex post facto* experiment. *Journal of Educational Psychology* **51**: 309–17
- Trochim W M K 1984 *Research Design for Treatment Evaluation: The Regression-Discontinuity Approach Program*. Sage, Beverly Hills, CA
- Trochim W 1990 Regression-discontinuity design in health evaluation. In: Sechrest L, Perrin E, Bunker J (eds.) *Research Methodology: Strengthening Causal Interpretations of Non-experimental Data*. US Dept. of HHS, Agency for Health Care Policy and Research, Rockville, MD, pp. 119–39
- Trochim W M K, Cappelleri J C 1992 Cutoff assignment strategies for enhancing randomized clinical trials. *Controlled Clinical Trials* **13**: 190–212
- Trochim W M K, Cappelleri J C, Reichardt C S 1991 Random measurement error does not bias the treatment effect estimate in the regression-discontinuity design: II. When an interaction effect is present. *Evaluation Review* **15**(4): 571–604

W. M. Trochim

Regulation and Administration

Regulation is the effort to shape significant institutions—industry, universities, immigration—to serve some social concerns that might not be taken into account in the ordinary process of doing business. According to one definition, it is the imposition of legal orders on those engaged in economic activity (Kagan 1978). States regulate capitalist enterprises within capitalist or mixed economies.

1. Emergence and Justification

Bureaucratic regulation of industries emerged in the nineteenth century with the rise of industrial capitalism and the bureaucratic division of labor. Britain established a factory inspectorate for the textile industry as early as 1832. The use of inspectorates to address the more dangerous conditions at work—in match factories, in tenement labor—expanded dramatically throughout the Western world in the 1890s (Rodgers 1998, pp. 254–9).

Regulation has proven to be at least partly reversible: central trends in capitalist economies during the 1970s and 1980s included both privatization of

state-owned industries (see e.g., Graham and Prosser 1988, Pistula and Rasmussen 1990) and political pressure toward deregulating already regulated industries (Derthick and Quirk 1985, Harris and Milkis 1996).

Economists have thought it important to explain why regulation is necessary within a capitalist economy, given that neoclassical economic theory holds that social purposes are best accomplished by markets. The importance of neoclassical theory has meant that advocates and critics of regulation have also expended effort understanding the justifications (see e.g., Breyer 1982). Justification has sometimes centered on the concept of 'externalities.' A market requires that business pays for some components, for example, steel in manufacturing a car. However, the manufacturer does not have to buy the air that it pollutes in the process of making cars, and in turn, neither do consumers who drive the cars. In the United States in the 1960s and 1970s in particular, concern about the inability of companies to take into account the social costs of doing business was used to justify regulating products and processes in manufacturing.

In addition to regulating products and processes, governments regulate price and entry: the price that an industry may charge and the requirements to enter into the business. The justification of price and entry regulation does not rely upon the concept of externalities. Rather, the argument is that some industries take on the characteristics of monopoly, whether as a natural monopoly or because there are very high barriers to entry. For example, it could be inefficient to have multiple railways competing against each other since the costs of laying railroad track are very high. What is more, because it is expensive to build a railway, it is extremely difficult to enter the business, which means that the competition that is supposed to ensure that prices and services are responsive to consumer demand does not exist. Without competition, railroads could charge what they would for services. Thus railroads, the first industry subject to national bureaucratic regulation, in the United States would charge more for short-haul than for long-haul trips because there was much less competition in short haul (Skowronek 1982, pp. 124–5, 141–50). In air transport, too, the barriers to entry are high because airspace is limited and in any one airport runways and gates are limited. Someone must allocate the space and ensure that once the airline has the space it does not ignore what customers want. Economists also note the importance of 'network effects' in monopoly. If having compatible computer systems facilitates the exchange of information, each person buying a computer system has an incentive to buy one compatible with those people already have, whether or not that is the best system. Advocates assert that professions like medicine and law require regulation because expertise is required to assess whether the profession is being practiced correctly. Consumers cannot assess whether

a surgeon knows what she is doing until it might be too late.

These justifications for regulation have been subject to serious critique from those who believe that markets should be unregulated and from those who believe that regulation is ineffective because industries have special access to the state not shared by their workers or customers (Derthick and Quirk 1985). In this view regulatory bureaucracies, designed to incorporate public concerns into capitalist economies, become 'captured' by the industries they are to regulate. The regulatory bureaucracies then act largely to insulate the industries from competition and protect their interests. Sometimes that has happened by design: railroad regulators meant to maintain and steady the railroad industry. In the late nineteenth century railroads in the United States wanted to be regulated to moderate the incentives to cut prices that were putting railroads out of business (Skowronek 1982). In France *notaires* welcome regulation to preserve their monopoly on drawing up legal contracts; opening that monopoly would have meant decreasing income as a result of competition (Suleiman 1987).

Economists call the effort to gain protection from the market by legislative or bureaucratic intervention 'rent-seeking behavior'—that is, an effort to gain a return beyond what is required to keep one in the business—and many have argued that rent-seeking is the heart of regulation. For example, lawyers benefit from the licensing of lawyers; it restricts the numbers of lawyers available and thereby maintains a higher price than would be the case if anyone could practice law. The claim that regulation protected inefficient industries from competition and used state power to keep prices artificially high provided a substantial part of the support for deregulation in the United States. In the late 1970s and 1980s (Derthick and Quirk 1985, Harris and Milkis 1996). In other nations, critiques of state ownership of industries led to privatization, which in turn entailed regulation (Graham and Prosser 1988). Across the industrialized Second World, industries differed in their ability to maintain monopolies or other advantages regulation provides (Derthick and Quirk 1985, Suleiman 1987).

In the United States, deregulation seems to have promoted consolidation in the airline industry and the trucking industry just as understandings of 'natural monopolies' might have predicted. Perhaps surprisingly, the strength of the pro-environmental regulation lobby meant that environmental deregulation did not go far (Harris and Milkis 1996). Those advocating a more market-based approach in pollution rights gained recognition in Clean Air Act Amendments, but government officials were still deeply involved in regulation (Harris and Milkis 1996). Presidents elected on antiregulation platforms do not completely control the political environment in which they work. The American federal system also provides access in the states when people have lost at the federal level. For

example, when the federal government tried to limit the regulation of advertising, activists turned to the more sympathetic state governments.

States approach regulation relying on different institutions that have, in turn, different styles. Courts, legislatures, local governments, marketing boards, and bureaucratic agencies all can (and have) imposed orders on capitalist enterprise. In the United States before the Civil War, rather than relying on negotiations between business and bureaucrats, or on rules from bureaucrats that were enforced in the courts, the United States relied on courts—largely state courts—to enforce statutes or common law duties. In addition, state legislatures granted charters to businesses and in return exacted some form of regulation: limits on what entrepreneurs could charge to cross a bridge or use a road (Hartz 1948), or requirements that those who owned docks contribute toward keeping harbors dredged and open (Hartog 1983). Such regulation is not a matter of general rules implemented through a bureaucracy, but it was an effort to require businesses to take into account public concerns not readily addressed in the market. By the late nineteenth century most states had instituted some bewildering array of boards, commissions, and special courts to instead ensure that a broader public interest was taken into account.

2. Institutions that Regulate

The first American federal regulatory agency, the Interstate Commerce Commission, was created in 1887. It was unusual in that it established an agency independent of immediate control by ministers, responsible for making a framework of policy through trial-type procedures, with decisions that could be appealed to general jurisdiction courts (Baldwin 1985, p. 7). The United States' official regulatory procedure is more rule-oriented than those in other Western industrialized states, and courts are more deeply involved in making policy. The American system also allows for much more public participation and opportunity to argue against regulations both within an administrative agency and through courts after administrative agencies have decided.

After World War Two regulation in the United States became open to wider public participation. Administrative agencies must hold public hearings after announcing the topic of a rule they are considering. They must then justify the rule they issue according to the evidence they have. Often an organization concerned with the rule will bring the regulatory agency to court. Federal courts usually review the processes by which the rules were promulgated and decide whether the rule or decision is justified according to the evidence. What lawyers have won in the United States has been observed sometimes with horror, sometimes with envy in other nations.

Other nations do not rely so extensively on the equivalent of generalist courts to assess the procedures by which rules are enacted. The French have a specialized court that hears challenges to administrative decisionmaking, the *Conseil d'Etat*, or Council of State. The *Conseil d'Etat* is staffed with people who have built careers within administration; the purpose is to have a court that will take into account public purposes, or the interests of the state, in deciding cases brought by individual challengers. In Britain, practitioners in regulation were frustrated enough with how their general jurisdiction courts supervised administrative decisions that the state created a version of a specialized administrative court in the 1970s (Sterett 1997). Particular judges specialize in hearing complaints about administrators. The political justification for both courts is that specialization in administrative complaints will lead to sensitivity to public purpose and an awareness of what can go wrong in administration. Germany, also, relies on a specialized administrative court staffed by a trained civil service judiciary that lies above a system of specialized tribunals.

A further contribution to the complexity of regulation in the United States is the extent of interest group litigation. Interest group challenges to the general rules under which administrators work are far more numerous than in any other nation. However, social welfare groups have in recent years organized to challenge rules concerning immigration and environmental regulation in Western Europe. The expansion of standards via the European Court of Justice and directives via the European Union have made it more possible to challenge domestic regulation, or refusals to regulate, in Europe. Finally, cross national treaties or other agreements also shape domestic regulation, for example in ozone regulation (Braithwaite and Drahos 2000; Canan and Reichman 1993).

In most countries there are specialized courts between generalist courts or the upper level administrative court and front line administrators. These specialized courts usually include representatives of relevant sectors (e.g., employers and union representatives in labor tribunals) in their staff. Members of specialized courts are not always legally trained. Their procedures are often less formal than those of generalist courts, and those before them do not always have legal representation. The informality and discretion allowed most tribunals provides substantial opportunity for a slip between formal statement of government policy and its meaning in particular cases.

3. What Does it Cost?

Line items in government budgets for regulatory agencies do not actually tell what regulation costs. To comply with regulatory requirements, businesses incur the costs of improving safety or documenting that they

are trying to. To the extent that regulation requires businesses to internalize their externalities, to pay for pollution abatements, for example, there is no one thing to measure. Business representatives have argued that the processes of American regulation are more difficult than those in other Western postindustrial states: compliance requires more documentation, inspectors more readily impose fines, businesses can expect more public interest group confrontation, hearings, and possibly court battles (Kagan and Axelrad 2000). The outcomes in Western states—the regulatory standards that are imposed—tend to converge. (Kagan and Axelrad 2000). It is unclear whether this means that other states achieve comparable standards without the process costs incurred in the United States or that the United States through its extensive procedures sets standards that are copied elsewhere.

Regulation involves trade-offs between effectiveness and other considerations. For example, we might want nursing homes to provide good care for people, but that is hard to measure. To the extent that we try to enforce it, we might direct effort in an organization toward documenting practices that do not quite get at caring, such as providing meals on time (Braithwaite 1993). A more substantive mandate to care might better capture what we want, but it would be unenforceable. In addition, different substantive regulations are acceptable to businesses and broader publics in different countries. For example, American commitments to nondiscrimination, which critics see as more extensive than those in most other countries, impose costs through requirements of documentation of fair procedures. Such provisions do not require any particular substantive outcome. Other countries, however, employ approaches that would be unacceptable in the United States. For example, Sweden tried to implement employment equality in a way that could keep administrative costs low while resulting in real change, by setting employment quotas for women in jobs in which they were underrepresented (Brown and Wilcher 1987). Britain, on the other hand, has relied on loose and open-ended definitions of nondiscrimination implemented through tribunals whose members are not legally trained and whose primary jurisdiction is not nondiscrimination law. The result seems to be a relatively inexpensive system that has done little to challenge discriminatory practices in the workplace (Kenney 1992).

Globalization has increased the importance of legal rules and lawyers in regulation in national states outside the United States, raising questions about the costs of more formal styles of regulation (Kagan 1997, Kagan 2000). Despite concerns about the expense of regulation, enterprises sometimes choose to pursue the goals of regulation beyond what governments require, because compliance offices sometimes employ people who take the goals of regulation to be their mission, or because large corporations concern themselves with

their public image as good citizens (Edelman 1990; Rees 1988). Compliance is not always a matter of direct orders.

4. *Explaining Regulation: Political Institutions and Political Culture*

Explanations for the different approaches to regulation—contentious, and emphasizing public procedures and rules as against consensual, and without formal rules—across postindustrial states are of two different types. The first explanation emphasizes the historical emergence of governance in different countries. The second explanation emphasizes cultural predilections such as individualism as against faith in authority.

A historical-institutional approach holds that political institutions have shaped the emergence of different regulatory regimes. Those political institutions are in turn the result of large-scale historical processes. In particular, in much of Western Europe, a career civil service insulated from politics was created in the nineteenth century (Kelman 1981). Administrative arms of government often started as the personal servants of the monarch; the political environment within which they worked did not include mass electorates. Governments could develop bureaucracies insulated from political pressure, setting in place systems difficult to disrupt, no matter what cultural preferences there might be for a broader or more accessible or more aggressive system of regulation. Alternatively, the incorporation of political fragmentation also has a history necessary, for example, to understand affirmative action policy in India (Parikh 1997).

Parliamentary systems, in which the government controls the legislative agenda, provide the government substantial control over how a regulator bureaucracy will be organized and the forms accountability will take. Should a government believe it is best to include only responsible-seeming or moderate interest groups in their discussion of regulation, they will find little institutionalized opposition (Vogel 1986). Parliaments see little reason to institute checks on administrators that can be invoked directly by interest groups, such as a right to sue. From the point of view of government ministers, both in charge of government departments and sitting in parliament, such a capability only makes regulation more difficult.

In addition to administrators and their ministerial supervisors, courts participate in regulation, to a different extent in different national states. In civil law systems, judging is a career to which people are trained. The highest level administrative court in France, the *Conseil d'Etat*, is staffed with top-level civil servants. Their charge is quite clearly to attend to complaints against administrators with an eye to maintaining the

effectiveness of a regulatory regime; it is not a system designed to allow confrontation over the fundamentals of public policy. Even within British-style common law systems, judges in the general legal system are selected from the elite of the bar, who are narrowly trained and disinclined to confront administrators. The insulation from the electoral politics that so profoundly shapes the civil services includes the judiciary.

The historical-institutional explanation points us toward the very different institutional development in the United States in order to understand differences in the contentiousness of regulation. In the American presidential and federal system, the legislatures—Congress and the states—are not wholly within the control of the executive. Those who wish to insist on visions of regulation different from that of the executive can gain leverage in the legislature to do so, not least through mobilizing popular mistrust of closed procedures. The American system of formal rulemaking procedures and some openness in administration resulted first from the American Bar Association's mistrust of regulation and then from the politically distinct mistrust emergent in the 1960s and 1970s. Both gained leverage through Congress and through the courts, which had more generous standards for allowing people to sue than did courts in other nations.

In contrast, cultural explanation for differences in regulation among states focuses more on shared meanings and less on the institutional sources of those meanings. American culture is individualist and assertive according to this interpretation, while cultures in other Western countries are more deferential, in part based on a monarchical past and strong social hierarchies (Vogel 1986). Deference, however, hardly characterizes the upheavals of the 1960s and 1970s, pervasive throughout Western industrialized states. Demands for participation were institutionalized in law in the United States but not in other nation-states in that period. In other nation-states attempts to make the government more responsive included instituting ombudsmen, state institutions to listen to complaints about administration in particular cases. Responses did not include a fundamental change in how rules were instituted, though they did in the United States.

Further evidence raises questions concerning the completeness of the cultural explanation. In Britain, for example, a country cited as having informal regulatory practices in comparison to the United States (Vogel 1986), land-use planning has been contentious, rule-bound, and full of litigation. A cultural explanation arguing that the British are more deferential than Americans cannot account for that contentiousness. An alternative institutional and historical explanation is relatively simple. The legal profession long made most of its income in land. By the time bureaucratic land-use regulation emerged, the

government had to take into account the interests of an already organized legal profession accustomed to litigating land disputes. In attending to national cultures as a basis for explanation, tensions within cultures require scholars to attempt to explain why some cultural preferences have been enacted into law and some have not.

Supranational courts shape regulation within the European Union. Both the European Court of Human Rights and the European Court of Justice make decisions concerning the rules that nation-states use in admitting immigrants, limiting sex discrimination, and determining the right to a family life or the right to travel. The European Court of Justice also decides cases concerning environmental regulation, based on directives from the Council of Ministers. The context for regulatory bodies therefore includes not only the businesses and practices that they regulate, alongside the attentive publics following regulation and their domestic political institutions. Increasingly, it includes supranational institutions harmonizing requirements cross-nationally. A perspective that takes nation-states as monolithic entities would see the supranational courts as an imposition on national autonomy. Since, however, not everyone within a nation-state agrees—some political activists want stricter regulation of pollution, or more openness to the families of settled immigrants—the introduction of supranational institutions makes for more complex alliances. State officials themselves are often not united, so supranational decisions can strengthen the hand of one administrative office over another.

5. *Constituting Regulation: Rules and Discretion*

National legislatures and executives may state general rules for regulation: that sex discrimination in employment is illegal, that businesses must reduce air pollution. Regulation is a complex task in which goals are often in tension—for example, to clean the air while ensuring that businesses that pollute do not have to close, or without severely curtailing driving. Multiple and conflicting goals for complex organizations mean that what counts as compliance requires interpretation. Even in the United States, where the courts are in the business of continually clarifying the rules, the facts of a situation can be complicated enough that compliance with a rule is something that will entail negotiation between the regulator and regulated, which a court then only ratifies. An order to emit less smoke can easily become subject to negotiations regarding the kind of scrubbers to install (Melnick 1983). Distinguishing between general rules and decision making in individual cases is difficult. In regulation, general rules often emerge out of particular decisions.

Where officials have substantial discretion in determining compliance, the meaning of regulation emerges

in negotiations with the regulated company. Case-level administrators interpret their instructions in light of what regulated businesses ordinarily do. For example, in Britain, specialized courts (or tribunals) that were responsible for hearing labor complaints interpreted discrimination law so to prohibit *unreasonable* unfavorable treatment of women, even though the sex discrimination statute did not rely upon a standard of reasonableness. 'Reasonableness' was a standard in unfair dismissal cases outside of sex discrimination. Common sense told tribunal members that reasonableness must be part of the sex discrimination statute; in practice the tribunals ratified most of what businesses had been doing (Kenney 1992, pp. 104–8).

Changing the ordinary practices of an organization is difficult. Where there are applicable rules or direct court orders against an organization, the organization must construe what it means to comply. Organizations have every reason to interpret decisions so that they fit with organizational priorities. For example, a court order to pay women equally with men in comparable jobs still leaves open what count as equivalent jobs, how equal the pay must be, and over how long a period it must be equalized, all of which will be up to an employer to determine (McCann 1994). Administrative agency decisions or court interpretations of them gain meaning in their application.

The professional training of decision-makers shapes what regulations mean. In employment discrimination law in the United States, personnel administrators interpret the law so as to require what they see as good management practices, which requires that people who work together get along. They do not often interpret the law as a matter of legal right, that is, as an enforceable claim that someone can make against his or her employer. Rather, personnel administrators are more likely to interpret problems as matters of personality conflicts (Edelman 1993). Thus the presence of enforceable rights does not necessarily make regulation very different from, for example, enforcement of water pollution laws in Britain, where there are no clear rules and enforcement officials negotiate and bluff (Hawkins 1984, see also Graboski and Braithwaite 1986). A substantial difference is whether officials can enlist courts in enforcement efforts.

Recent scholarship focuses on both business organizations and administrative officials working in complex environments in which they mutually shape meanings of the law (Edelman 1993, Harris and Milkis 1996). Charting variation across sectors requires attending to that environment in which regulation works, including the variation in attentive publics as well as the organization of the state. While national states sometimes do have national approaches shared across regulatory sectors, the emergent supranational context is shaping regulation so as to require analysts to attend to reasons for convergence as much as differences between nation-states.

Differences could result less from differences between postindustrial states; the global context of operation might lead to convergence among First World states (Draho and Braithwaite 2000), while Third World countries produce goods, often in conditions much worse for workers. Regulation there could depend as much on global political forces as on internal political dynamics.

See also: Broadcasting: Regulation; Bureaucratization and Bureaucracy, History of; Employment and Labor, Regulation of; Environment Regulation: Legal Aspects; Environmental Policy: Protection and Regulation; International Communication: Regulation; Occupational Health and Safety, Regulation of; Public Administration, Politics of; Regulation, Economic Theory of; Regulation: Empirical Analysis; Regulation: Family and Gender; Regulation: Sexual Behavior; Regulation Theory in Geography; Regulation: Working Conditions; Regulatory Agencies

Bibliography

- Baldwin R 1985 *Regulating the Airlines: Administrative Justice and Agency Discretion*. Clarendon Press, Oxford, UK
- Boyle E 1998 Political frames and legal activity: The case of nuclear power in four countries. *Law and Society Review* 32(1): 141–74
- Braithwaite J 1993 The nursing home industry. In: Tonry M, Reiss A J (eds) *Beyond the Law: Crime in Complex Organizations*. *Crime and Justice* 18: 11–54
- Braithwaite J, Draho P 2000 *Global Business Regulation*. Cambridge University Press, Cambridge, UK
- Breyer S 1982 *Regulation and its Reform*. Harvard University Press, Cambridge, MA
- Brown C, Wilcher S J 1987 Sex-based employment quotas in Sweden. In: Brown C, Pechman J A (eds) *Gender in the Workplace*. Brookings Institution, Washington, DC
- Canan P, Reichman N 1993 Ozone partnerships, the construction of regulatory communities, and the future of global regulatory power. *Law and Policy* 15(1): 61–74
- Derthick M, Quirk P J 1985 *The Politics of Deregulation*. Brookings Institution, Washington, DC
- Edelman L B 1990 Legal environments on organizational governance: the expansion of due process in the American workplace. *American Journal of Sociology* 95(6): 11401–440
- Edelman L B 1993 Internal dispute resolution: The transformation of civil rights in the workplace. *Law and Society Review* 27(3): 497–514
- Edelman L B, Suchman M C 1997 Legal rational myths: The new institutionalism and the law and society tradition. *Law and Social Inquiry* 22: 906–41
- Einhorn R 1991 *Property Rules*. University of Chicago Press, Chicago
- Graboski P, Braithwaite J 1986 *Of Manners Gentle: Enforcement Strategies of Australian Business Regulatory Agencies*. Oxford University Press, Melbourne, Australia
- Graham C, Prosser T 1988 *Waiving the Rules*. Open University Press, Philadelphia

- Harris R A, Milkis S M 1996 *The Politics of Regulatory Change: A Tale of Two Agencies*. Oxford University Press, New York
- Hartog H 1983 *Public Property and Private Power: The Corporation of the City of New York in American Law, 1730–1870*. University of North Carolina Press, Chapel Hill, NC
- Hartz L 1948 *Economic Policy and Democratic Thought: Pennsylvania: 1776–1860*. Harvard University Press, Cambridge, MA
- Hawkins K 1984 *Environment and Enforcement: Regulation and the Social Definition of Pollution*. Clarendon Press, Oxford, UK
- Kagan R A 1978 *Regulatory Justice: Implementing a Wage-price Freeze*. Russell Sage Foundation, New York
- Kagan R A 1997 Should Europe worry about adversarial legalism? *Oxford Journal of Legal Studies* 17(2): 165
- Kagan R A 2000 The consequences of adversarial legalism. In: Kagan R A, Axelrad L (eds.) *Regulatory Encounters: Multinational Corporations and American Adversarial Legalism*. University of California Press, Berkeley, CA
- Kelman S 1981 *Regulating America, Regulating Sweden: A Comparative Study of Occupational Safety and Health Policy*. MIT Press, Cambridge, MA
- Kenney S J 1992 *For Whose Protection? Fetal Protection Policies in the United States and Great Britain*. University of Michigan Press, Ann Arbor, MI
- McCann M 1994 *Rights at Work*. University of Chicago Press, Chicago
- Melnick R S 1983 *Regulation and the Courts*. Brookings Institution, Washington, DC
- Parikh S 1997 *The Politics of Preferences*. University of Michigan Press, Ann Arbor, MI
- Pistula J M, Rasmussen K 1990 *Privatizing a Province: The New Right in Saskatchewan*. New Star Books, Vancouver, Canada
- Rees J 1988 *Reforming the Workplace: A Study of Self-regulation in Occupational Safety*. University of Pennsylvania Press, Philadelphia
- Rodgers D 1998 *Atlantic Crossings: Social Politics in a Progressive Age*. Harvard University Press, Cambridge, MA
- Ruggie M 1984 *The State and Working Women*. Princeton University Press, Princeton, NJ
- Skowronek S 1982 *Building a New American State*. Cambridge University Press, New York
- Sterett S 1997 *Creating Constitutionalism?* University of Michigan Press, Ann Arbor, MI
- Suleiman E 1987 *Private Power and Centralization in France: The Notaires and the State*. Princeton University Press, Princeton, NJ
- Vogel D 1986 *National Styles of Regulation: Environmental Policy in Great Britain and the United States*. Cornell University Press, Ithaca, NY

S. Sterett

Regulation, Economic Theory of

Regulation is defined as ‘the act of controlling or directing by a rule, principle, [or] method’ (Webster 1989). This definition encompasses many forms of control by a variety of actors. The ensuing discussion will focus on the important form of economic regu-

lation in which the government controls the activities of producers in order to enhance the well being of the customers they serve.

1. Regulating with Accurate Information

When the average cost that each firm incurs in serving customers (i.e., the firm’s cost per customer) declines as the firm serves more customers, industry costs will be minimized when only one producer serves all customers. For example, the cost of providing local telephone service to a community is often lower when a single telecommunications network is constructed and operated than when multiple networks are operated. Similarly, because of declining average costs, the local distribution of electricity, water, and gas is often provided at minimum cost by a single firm. But if a single profit-maximizing firm faces no competition, consumers will have no alternative source of supply, and so an unregulated sole producer may have considerable freedom to charge high prices for its services. Regulation of the prices charged by a single (monopoly) supplier can help to protect consumers while permitting the industry configuration that minimizes the costs of serving customers.

Consumer protection can take many forms, particularly when the regulated supplier provides more than one service. To illustrate this fact, consider the mandate of the US Postal Rate Commission. When it sets the prices that the US Postal Service charges for the services it delivers, the Commission is required to ensure that the pricing structure is simple and that prices are fair and equitable. The Commission must also ensure that established prices reflect the value of the service rendered and cover the costs of providing each service (Postal Reorganization Act 1970, § 3622(b)).

To determine how a regulatory body (hereafter, a regulator) might attempt to satisfy such mandates, consider first, the case of an omniscient regulator who knows both how highly consumers value the services produced by the regulated firm and the minimum possible costs that the firm incurs in serving customers.

1.1 Linear Pricing with a Single Output

Initially suppose that the regulated firm produces only one product. Also suppose that the regulator desires to set the price for this product so as to maximize the aggregate welfare of all consumers, while delivering sufficient profit to the firm to guarantee that it will voluntarily serve customers at the established price. Consumer welfare measures the total value that consumers derive from their purchase and consumption of the product. A firm’s profit is the difference

between the revenues it receives and the costs it incurs. Profit is a measure of producer welfare.

To maximize consumer welfare while delivering at least the minimum required level of producer welfare, the regulator must set the price that consumers pay for each unit of the product they purchase so as to deliver the largest possible sum of consumer and producer welfare. This welfare-maximizing price is the one that equals the extra (or marginal) cost the firm incurs in producing the last unit of the product it sells. This is because consumers will continue to buy more of the product as long as the extra value they receive from doing so exceeds the price they must pay for an additional unit of the product. Therefore, when the price reflects the producer's marginal cost of production, consumers will purchase additional units of the product whenever the value they derive from consumption exceeds the cost of allowing that consumption. Consequently, a regulated price that reflects the producer's marginal cost of production will ensure the level of output that maximizes the combined welfare of consumers and the firm.

However, when unit production costs decline as the firm produces more output, a price set equal to the marginal cost of producing the last unit of output will cause the firm to suffer a financial loss. For all but the last unit produced, the compensation the firm receives (i.e., the price) will be less than the incremental production cost it incurs. To ensure that the total revenue the firm receives does not fall below the total production cost it incurs, the regulator must raise the authorized price above the level of marginal cost. A price equal to the firm's average cost of production at the prevailing level of production will deliver the highest possible level of welfare to consumers while ensuring that the regulated firm's costs do not exceed its revenues.

1.2 Nonlinear Pricing

A regulator might contemplate implementing a pricing structure in which customers do not simply pay the same unit price for each unit of the product they purchase. For example, each customer might be charged both a fixed participation fee that does not vary with the amount of the product purchased and a constant unit price for each unit that is purchased. Under such a pricing structure, the unit price might be set equal to the regulated supplier's marginal cost of producing the last unit of output it sells and the participation fee might be set to recover exactly the deficit the firm incurs (i.e., the difference between the total revenue it receives from sales and the total production costs it incurs) when it sells its product at the specified unit price. This pricing structure would deliver the maximum possible level of consumer welfare while ensuring non-negative profit for the firm, provided the participation fee is not so onerous

as to compel some customers to terminate all purchases of the regulated product (Wilson 1993). In practice, large participation fees are often viewed as unduly onerous for consumers with limited resources. Consequently, particular concern for the well being of these consumers or general concerns with equity and fairness can limit the use of participation fees and other alternatives to constant unit prices.

1.3 Linear Pricing with Multiple Outputs

Now consider the more common setting in which the regulated firm produces multiple products (for example, both rural and urban telephone services). Also suppose the regulator chooses to implement only a constant unit price for each regulated product. Since the prices charged for different products can vary, there are many different ways in which the regulator can raise prices above marginal production costs while ensuring that revenues match total operating costs exactly. To maximize the combined welfare of all consumers, the regulator will implement proportionate mark-ups ($(p_i - c_i)/p_i$) of price (p_i) above marginal cost (c_i) that vary inversely with the sensitivity of consumer demand to price (Ramsey 1927, Baumol and Bradford 1970). The regulator will set the price of a product close to its marginal cost of production when a larger price increase would cause consumers to severely curtail their purchases of the product. In contrast, the regulator will raise the price of a product well above its marginal cost of production when doing so does not cause consumers to reduce their purchases of the product significantly. This pricing policy secures the greatest combined welfare for consumers by limiting the aggregate reduction in consumption that arises as prices are raised sufficiently far above marginal production costs to generate revenues that match production costs.

1.4 Quality Deterioration and Cream Skimming

As noted above, regulators do not always seek solely to maximize the aggregate welfare of consumers. Consequently, in practice, regulated prices are not always raised above marginal costs by greater amounts the less sensitive consumer demand is to price. In fact, some prices (e.g., those for rural, residential telephone service) may be set below marginal production costs while other prices (e.g., those for telephone service to urban business customers) may be set well above marginal production costs. Such prices may arise from fairness and equity concerns, for example. Pricing patterns of this sort can introduce two important problems: quality deterioration and cream skimming.

When the price of a product is set below its marginal cost of production, each extra unit of the product that the firm sells causes its profit to decline. Consequently,

the firm would secure greater profit if consumers purchased fewer units of the product. One way to induce consumers to purchase fewer units of a product is to reduce the quality of the product. Therefore, when prices are set below marginal production costs, the regulator may need to impose penalties on the firm if the level of service quality it delivers is found to be inadequate.

When prices are set above cost for some products and below cost for other products, competitors may find it profitable to supply only those products for which price exceeds cost and leave the incumbent regulated firm to supply those services for which price is set below cost. For example, unless they are prohibited from doing so, competitors may provide telecommunications service only to large businesses in urban areas. This practice is known as cream skimming. A regulator can prevent cream skimming by barring potential competitors of the incumbent regulated firm from providing services in the regulated industry. Such entry prohibitions, though, can hinder industry performance by precluding the operation of firms that have lower operating costs than the incumbent supplier and by removing the discipline that the threat of entry can impose on the incumbent firm. Furthermore, competitors are sometimes able to circumvent prohibitions on entry by developing and marketing substitute products. A regulator can limit incentives for cream skimming by setting prices that more closely reflect the incumbent firm's operating costs. When prices reflect costs in this manner, entry will typically be encouraged only when a competitor can satisfy consumers' needs at lower cost than the incumbent supplier can.

2. Regulating With Limited Information: Incentive Regulation

When, as presumed above, the regulator is well informed about the firm's operating capabilities, the regulator can set prices to generate revenues that exactly match production costs when the firm operates efficiently (i.e., when it incurs only the minimum possible cost required to serve customers). When the regulator is not perfectly informed about the firm's capabilities, the regulator cannot simply mandate that the firm operates efficiently. Instead, the regulator must induce the firm to minimize its operating costs. There are a variety of techniques the regulator can employ to do so. These techniques constitute elements of what is termed incentive regulation.

2.1 Regulatory Lag

When prices are set to match realized production costs only periodically (e.g., once every three or four years) rather than continually, the firm derives some financial

benefit from reducing its operating costs. The larger the period of time before prices are reduced to match lower realized operating costs, the more pronounced is the firm's incentive to reduce its costs. On the other hand, the longer prices are permitted to remain above operating costs, the longer consumer welfare is held below the level that could be secured given the firm's realized costs. This tradeoff determines the optimal length of time between price revisions, i.e., the optimal regulatory lag.

2.2 Price Cap Regulation

Price cap regulation has enjoyed increasing popularity in recent years. It incorporates a form of regulatory lag while delegating some pricing discretion to the regulated firm. The typical price cap regulation plan specifies a price cap period (often four or five years) and a maximum rate at which the regulated firm's inflation-adjusted prices can rise annually, on average, during this period (Sappington in press). The firm's realized costs and earnings are usually reviewed at the end of the specified price cap period, and often employed to update estimates of the pricing restriction that would allow the regulated firm to secure a reasonable level of earnings in the next price cap period. When the firm's realized performance is employed to set future standards, the firm's incentives for exceptional performance are dulled. However, price cap regulation provides some incentives for cost reduction by instituting an explicit regulatory lag.

By controlling the rate at which regulated prices can increase on average, price cap regulation affords the firm some discretion in setting prices for individual services. This discretion can help the regulated firm to limit undesirable cream skimming by realigning prices to more closely match production costs. However, this discretion can also enable the firm to raise certain prices that the regulator prefers to remain low (e.g., the price of local telephone service to rural residential customers). To limit the firm's ability to implement undesired pricing structures, price cap regulation plans often prohibit the firm from raising significantly the prices of certain specified products.

2.3 Earnings Sharing Regulation

By providing incentives to reduce operating costs, price cap regulation can simultaneously provide incentives to reduce service quality, since superior quality is often more costly to provide. By allowing prices to diverge from realized costs for considerable periods of time, price cap regulation can also admit earnings that depart significantly from expected levels. To mitigate these potential problems, an earnings sharing plan is sometimes appended to price cap regulation. An earnings sharing plan requires the firm to share with its customers a fraction of realized earnings above (and

often below) specified thresholds. Such sharing helps to insure the firm and the regulator against extremely high (and low) earnings, but at the expense of reduced incentives for exceptional performance (Lyon 1996).

2.4 Yardstick Regulation

In certain settings, yardstick regulation can be employed to provide strong incentives for superior performance while limiting the risk of extreme earnings. Under yardstick regulation, the standards to which one firm is held are based upon the performance levels achieved by other firms. To illustrate, in the British water industry, the prices and service quality standards that are imposed on the monopoly provider in one jurisdiction are linked to the costs and service quality levels achieved by monopoly producers in other geographic regions (Cubbin and Tzanidakis 1998).

When firms face similar operating circumstances, yardstick regulation can enable the regulator to impose a reasonable but challenging performance standard on the regulated firm without basing this standard on the firm's own historic performance. Consequently, yardstick regulation can ensure reasonable earnings for the regulated firm while providing the firm with strong incentives for superior performance. However, these desirable outcomes can only be secured via yardstick regulation if the firms are known to face similar operating circumstances or if all relevant differences in their operating circumstances are well understood (Shleifer 1985). In the absence of this knowledge, the standards imposed under yardstick regulation can result in extreme earnings, much like under price cap regulation. Yardstick regulation may also encourage collusion among the firms that are implicitly being compared. By agreeing to reduce their performance below achievable levels, the firms can effectively relax the constraints imposed by yardstick regulation.

2.5 Franchise Bidding

Even when there is a single regulatory jurisdiction and a single producer in the industry, a regulator may still be able to harness the discipline of competition by structuring competition for the industry. The regulator may allow potential producers to bid for the right to serve as the sole producer in the industry. When the potential producers have similar abilities and similar information about operating circumstances, intense bidding competition for the market can secure for consumers the low prices and high service quality that competition among producers in the market might otherwise provide. However, successful franchise bidding requires that all relevant performance requirements be specified completely in advance, and that the

regulator and firm both adhere precisely to the terms of the franchise contract once a producer has been selected. If the regulated firm can renege on its responsibilities after it is installed as the sole producer, or if the regulator can demand more services from the selected producer than it originally agreed to provide, franchise bidding can fail to deliver the maximum possible attainable level of consumer welfare (Williamson 1976).

2.6 Regulatory Options

When the regulated firm has better information than the regulator about its operating capabilities, consumers are often best served when the regulator allows the firm a choice among regulatory plans. A careful structuring of the options afforded the firm can induce it to employ its superior information to choose a plan that best serves consumers (Laffont and Tirole 1993, Armstrong and Sappington in press).

To illustrate this point, suppose that a regulator decides to impose price cap regulation on a firm rather than allow the firm to choose between price cap regulation and cost-plus regulation, whereby prices are re-set frequently to match realized production costs. To limit the likelihood of financial distress, the regulator may not insist that prices decline rapidly. In contrast, when the firm has the option to choose cost-plus regulation, the regulator can impose more stringent limits on price increases under price cap regulation without the fear of causing financial distress. If the firm believes that it will be unable to realize adequate earnings under the proposed price cap regime, it will choose to operate under cost-plus regulation instead. But if the regulator has not overestimated the firm's capabilities in setting the challenging requirements of the price cap regime, the firm will choose this regime and deliver its pronounced benefits to consumers.

3. Regulation in Vertically Related Industries

Sometimes, a regulated firm may have the ability to supply both inputs and final products. Inputs are products that firms employ to make other products. Final products are those that customers consume, and that are not used to make other products. For example, a local telecommunications network operator may sell to long-distance telecommunications firms the local network access (an input) they must have to complete the long-distance messages they transport. In addition, the local network operator may have the ability to deliver long-distance telephone calls (a final product) itself. Two fundamental questions arise in such settings: (a) should the monopoly supplier of an essential

input be permitted to compete downstream, i.e., against the firms to which it sells the input? (b) how should the price of the essential input be regulated?

3.1 Regulating Industry Structure

One potential benefit of allowing an input supplier to operate downstream is the additional competitive pressure that the input supplier may impose on other downstream producers. The additional competitive pressure can serve to reduce industry prices, and thereby enhance consumer welfare. An input supplier may be a particularly strong competitor when economies of scope are present. Economies of scope exist if multiple products can be supplied at lower cost when they are produced simultaneously by a single firm rather than when they are produced in isolation by distinct firms (Baumol et al. 1982 p.71).

A potential cost of authorizing downstream participation is that it may encourage the input supplier to disadvantage its downstream rivals unfairly. For example, the input supplier may intentionally reduce the quality of the essential input that it sells to its downstream competitors in order to make it more difficult for the competitors to provide high quality service at low cost. However, the input supplier may not have an unambiguous incentive to disadvantage its downstream rivals, because their demand for the essential input typically declines as the quality of the input declines. This reduced demand can reduce the profit derived from supplying the essential input (Weisman in press). The magnitude of the profit the input supplier foregoes by reducing the demand for the input it supplies is greater the higher the price the firm is permitted to charge for the input, *ceteris paribus*. Consequently, raising the price of an essential input above its marginal cost of production can reduce the incentive of an input supplier to disadvantage its downstream rivals.

3.2 Regulating Input Prices

Raising the price of the essential input above its marginal cost of production can also help to limit cream skimming and ensure that downstream production is carried out by the least-cost suppliers. To illustrate this point, suppose that if the input supplier operates downstream, the prices it charges there are regulated. Further suppose that the price of one downstream product (called product A) is set at a level (p^A) that exceeds the supplier's marginal cost of production (perhaps to facilitate a price below marginal cost for another product). If the price of the essential input is set equal to its marginal cost of production in this setting, a downstream producer may be able to serve all consumer demand for product

A profitably even though its marginal cost of production ($\tilde{c}^A$) exceeds the corresponding cost (c^A) of the input supplier. The higher-cost producer can set a price below the regulated price for product A and still secure positive profit if its cost disadvantage ($\tilde{c}^A - c^A$) is smaller than the regulated profit margin on product A ($p^A - c^A$). To preclude operation by higher-cost competitors, the price of the essential input can be raised above its marginal cost of production by the amount of profit ($p^A - c^A$) the input supplier foregoes when it provides the input to competitors (Laffont and Tirole 1994, Baumol et al. 1997, Armstrong in press). One potential drawback to this pricing policy, though, is that it provides little incentive for the input supplier to reduce its operating costs.

4. Additional Forms of Regulation

Regulation of producers' activities can help to enhance consumer welfare even when industry cost minimization does not require operation by a single producer. To illustrate this fact, consider the potential value of environmental regulation and health and safety regulation.

4.1 Environmental Regulation

In certain instances, production activity can impose costs on society in addition to the direct physical costs of the inputs used in production. For example, some production processes may expel harmful pollutants into the surrounding land, water, or air. If no restrictions are placed on producers' rights to expel pollutants into the environment, producers may produce more than the surplus-maximizing level of pollution. This is the level of pollution at which the social damage that would be caused by an additional unit of pollution is exactly equal to the social cost that must be incurred to eliminate that additional unit of pollution. Pollution in excess of the surplus-maximizing level imposes social damage that exceeds the cost of eliminating the extra pollution. In contrast, the cost of reducing pollution below the surplus-maximizing level exceeds the social benefits derived from the reduction in pollution.

Various regulatory policies can be employed to control environmental pollution. For instance, direct taxes (effluent fees) can be imposed on pollution when it is readily observed and measured. By setting the tax on pollution equal to the marginal social damage caused by the pollution, the producer can be induced to realize the surplus-maximizing level of pollution. When the level of pollution produced by a firm is difficult to measure, taxes can be imposed instead on the products whose production causes pollution, or on inputs (such as gasoline) to the activity (automobile driving) that causes pollution. The use of specific

pollution abatement technologies can also be mandated.

Recently, some governments have introduced tradable pollution permits. Each permit entitles its bearer to emit one unit of pollution (e.g., one ton of sulfur dioxide). Polluting firms are able to buy and sell permits at prices determined by the available supply of and demand for these permits. A key potential benefit of tradable permits is that they promote pollution abatement at minimum cost. Producers that are able to reduce the pollutants they expel at a cost below the prevailing price of a permit will find it profitable to do so and sell any permits that they own. In contrast, producers for whom pollution abatement is more costly than the price of a permit will continue to expel pollutants but will pay for the right to do so (Schmalensee et al. 1998).

4.2 Health and Safety Regulation

Pollution abatement regulation is often intended to improve the health and well being of citizens. Other forms of health and safety regulation can serve this same purpose. The regulations come in many forms, including information requirements, standards, and prohibitions.

Regulations that require manufacturers to provide relevant product information to consumers can help consumers choose among available products. These regulations include requirements that product contents be fully specified and that the potential health risks associated with consumption be stated clearly. In contrast, direct regulation of product standards and outright prohibitions on the sale of certain products can substitute information collection and decision-making by a regulatory body for decision-making by consumers. For example, when the government imposes minimum home construction standards or when it bans certain drugs from the marketplace, the government eliminates the option a consumer might otherwise have to purchase below-standard homes or banned drugs. One rationale for such regulations is that they can implement the decisions that consumers would make if they were fully informed about all relevant benefits and costs of consumption, without requiring consumers to develop the expertise and obtain and evaluate the information required to assess the benefits and costs (Viscusi 1992).

5. Competing for Regulatory Benefits

The preceding discussion illustrates some of the ways in which government regulation of producers can be designed to benefit consumers. It is important to note, however, that regulation need not only serve this role in practice. Since regulators typically have the power to deliver benefits to some parties and impose costs on others, the affected parties will have an incentive to try

to secure favorable treatment from regulators. When industry producers are better organized and have more at stake than industry consumers, regulation that benefits producers at the expense of consumers can arise (Stigler 1971, Peltzman 1989). For example, new competitors may be barred from an industry even though their entry and operation would be in the best interest of consumers. These realistic possibilities merit careful consideration when evaluating the benefits and costs of existing and proposed regulations.

See also: Business Law; Environmental Health and Safety; Social Aspects; Environmental Planning; Policy Process: Business Participation; Regulation and Administration; Regulation: Empirical Analysis; Regulation Theory in Geography; Regulatory Agencies

Bibliography

- Armstrong M (in press) The theory of access pricing and interconnection. In: Cave M, Sumit M, Vogelsang I (eds.) *International Handbook of Telecommunications Economics*. North-Holland Publishers, Amsterdam
- Armstrong M, Doyle C, Vickers J 1996 The access pricing problem: a synthesis. *Journal of Industrial Economics* **44**(2): 131–50
- Armstrong M, Sappington D (in press) Recent developments in the theory of regulation. In: Armstrong M, Porter R (eds.) *Handbook of Industrial Organization*. Elsevier Science Publishers, Amsterdam
- Baumol W, Bradford D 1970 Optimal departures from marginal cost pricing. *American Economic Review* **60**(2): 265–83
- Baumol W, Ordover J, Willig R 1997 Parity pricing and its critics: a necessary condition for efficiency in the provision of bottleneck services to competitors. *Yale Journal on Regulation* **14**(1): 145–64
- Baumol W, Panzar P, Willig R 1982 *Contestable Markets and the Theory of Industry Structure*. Harcourt Brace Jovanovich, Inc., New York
- Cubbin J, Tzanidakis G 1998 Regression versus data envelopment analysis for efficiency measurement: an application to the England and Wales regulated water industry. *Utilities Policy* **7**(1): 75–85
- Laffont J, Tirole J 1993 *A Theory of Incentives in Procurement and Regulation*. MIT Press, Cambridge, MA
- Laffont J, Tirole J 1994 Access pricing and competition. *European Economic Review* **38**(9): 1673–710
- Lyon T P 1996 A model of sliding-scale regulation. *Journal of Regulatory Economics* **9**(3): 227–47
- Peltzman S 1989 The economic theory of regulation after a decade of deregulation. *Brookings Papers on Economic Activity: Microeconomics*. The Brookings Institution, Washington, DC, pp. 1–59
- Postal Reorganization Act 1970 *United States Public Law Number 91-375, 84 Statute 719*
- Ramsey F 1927 A contribution to the theory of taxation. *Economic Journal* **37**: 47–61
- Sappington D (in press) Price regulation and incentives. In: Cave M, Majumdar S, Vogelsang I (eds.) *International Handbook of Telecommunications Economics*. North-Holland Publishers, Amsterdam

- Schmalensee R, Joskow P L, Ellerman A D, Montero J, Bailey E 1998 An Interim evaluation of sulfur dioxide emissions trading. *Journal of Economic Perspectives* 12(3): 53–68
- Shleifer A 1985 A theory of yardstick competition. *Rand Journal of Economics* 16(3): 319–27
- Stigler G 1971 The theory of economic regulation. *Bell Journal of Economics and Management Science* 2(1): 3–21
- Viscusi K 1992 *Fatal Tradeoffs*. Oxford University Press, New York
- Webster's Encyclopedic Unabridged Dictionary of the English Language* 1989 Portland House, New York
- Weisman D (in press) Vertical integration. In: Madden G, Savage S (eds.) *Handbook of Telecommunications Economics*. Edward Elgar Publishers, Aldershot, UK
- Williamson O 1976 Franchise bidding for natural monopolies—in general and with respect to CATV. *Bell Journal of Economics* 7(1): 73–104
- Wilson R 1993 *Nonlinear Pricing*. Oxford University Press, Oxford

D. E. M. Sappington

Regulation: Empirical Analysis

The economic regulation of firms and industries constitutes an extreme form of government intervention in markets. Government antitrust, competition, tax, and trade policies, as well as most environmental, health, safety, and information regulations, shape but do not replace the market, although the most severe of these policies may blur that distinction. Economic regulation stands in sharp contrast. In industries subject to economic regulation, government agencies or departments exercise considerable control over firms' entry, pricing, investment, and product choice decisions, and administrative decisions replace market outcomes. Intervention may go as far as replacing private ownership with government-owned and -operated enterprises.

The modern field of empirical regulatory economics developed during the 1960s and early 1970s, as economists began 'to apply the familiar principles of microeconomics in comprehensive fashion to the subject' (Kahn 1988, p. xv). George Stigler, whose contributions to the field were recognized in his 1982 Economics Nobel prize citation, sparked much of this growth. Through the 1980s, economists focused on quantifying the effect of regulation on firm behavior, industry outcomes, and consumer welfare. The diversity of regulatory forms and industry characteristics across the transportation, energy, utility, financial, and communication sectors provided a rich set of institutions for investigation. Significant policy reforms, even deregulation, in structurally competitive sectors during the 1970s and 1980s provided new sources of variation in the regulatory environment and

invigorated empirical regulatory research. More recent attention has turned to the traditional 'public utility' sector, particularly the telecommunication and electric utility industries. Market-based institutions have begun to replace state-ownership and regulation, as technological change called into question the natural monopoly assumption in many of these industries and policy interest in privatization raised questions of how best to restructure and sell-off vertically-integrated monopolies. Analyzing the operation of these redesigned markets has become a main focus of empirical regulatory research.

This article focuses on empirical analysis of the effects and reform of economic regulation. Section 1 describes the dominant empirical methodologies. Section 2 sketches the evidence on the effects of traditional economic regulation. Section 3 discusses recent analyses of market design and regulatory scope. To keep the reference list manageable, readers are referred to surveys and collective volumes wherever possible. These guide readers interested in delving more deeply into particular issues to a wealth of primary studies.

1. Empirical Methodologies

Measuring the effect of economic regulation requires the establishment of a counterfactual benchmark: what would firm or industry performance look like in the absence of regulation? A variety of approaches to this have been used. The most compelling lesson of the existing literature is that the impact of economic regulation depends critically upon its particular institutional form and the characteristics of the industry under study. Measuring the impact of regulation requires as much attention to the details of how regulators operate in practice as to the prescribed legal form of the regulation. Paul Joskow's (1974) study of electric utility regulation in the US provides a persuasive exposition of this argument.

The most common methodology used to measure regulatory effects compares outcomes in regulated markets with those in unregulated markets. Time series variation may facilitate this comparison; the wave of US deregulation in the late 1970s and early 1980s generated numerous 'before-and-after' studies of regulation. Differences across political jurisdictions, e.g., across states within a country or across countries, may provide another source of variation in regulatory regimes. Empirical studies also may use differences in the intensity or form of regulation to infer its effects, particularly if regulation is likely to bind only under certain well-defined circumstances.

Two conditions limit the usefulness of this approach. First, unregulated comparisons may not exist. Industries that are regulated in one political jurisdiction tend to be regulated or government-owned in other political jurisdictions. Regulation may be so

long-lived that comparing current regulated performance with historical unregulated performance is meaningless. Second, this method implicitly assumes that differences in regulatory regimes are exogenous to firm or market performance. This is often false. If, for example, jurisdictions impose electricity regulation in response to high or rapidly increasing local electricity rates, comparing price levels in regulated and unregulated locales might suggest no effect even if regulation does, in fact, constrain prices. Using panel data to estimate effects from 'differences in differences' econometric models may be useful in some cases, as may developing more complete models of the determinants of regulation.

A second methodology estimates a structural model of the market, forms assumptions about behavior in the absence of regulation, and compares simulated unregulated outcomes with those observed under regulation. This technique may be useful even in the absence of regulatory variation. It is most appropriate when regulatory constraints can be clearly specified. Suppose regulators constrain prices but do not influence investment or factor use decisions. Structural estimates of underlying demand and cost functions permit one to predict unregulated profit-maximizing prices and compare these with current regulated prices. If regulated firms do not minimize costs, however, or if regulatory intervention alters investment decisions in a systematic way, then the regulated cost function may be a poor estimate of the unregulated long-run cost function, and the implied price comparison will be misleading.

A third approach uses experimental data to measure regulatory effects. Field experiments, such as the time-of-use electricity pricing studies conducted in the US during the early 1970s, can reveal responses to alternative regulatory or market designs. In the best of circumstances, these can provide information on how market outcomes respond to exogenous variation in regulatory mechanisms, holding other economic conditions constant. Their cost limits their widespread use, however, and their temporary nature may preclude an assessment of long-run equilibrium responses to alternative policies. Laboratory experiments, which present volunteers (often students) with a set of objectives, operating guidelines, and payoffs, also have been used in a limited way to simulate responses to alternative regulatory designs.

2. The Effects of Traditional Economic Regulation

It is impossible to describe succinctly the full effects of economic regulation across industries, over time, and through space. Regulation may influence prices, costs, and the pace of technological change in an industry. This may lead to the creation or suppression of

economic rents, and change the distribution of rents across various stakeholders or interests in the industry. These effects are likely to vary considerably across markets and regulatory forms, and preclude simple generalizations. This article highlights some illustrative results. Further detail is provided in the many reviews referenced at the end of this article and the primary sources they cite.

2.1 Empirical Analysis of Natural Monopoly Regulation

Industries such as electricity supply, telecommunications, natural gas transmission and distribution, water and sewer service, cable television service, and to some extent railroads have historically been considered 'natural monopolies.' This applies when total production costs are minimized by providing service to all customers in a given market through a single firm. As economists since Adam Smith have recognized, minimizing production cost may not minimize price, since monopolies use their market power to mark up price above competitive levels and earn monopoly rents. This historically engendered one of two policy responses. Government ownership in natural monopoly sectors was typical outside the USA, and was used frequently in the USA for water and sewer service, but less so for electric utilities and cable television service. Government regulation of privately-owned firms was the dominant US response in most other cases. In this scheme, regulators typically grant a legal, 'franchise' monopoly to a single provider in a geographic market, but retain control over price to mitigate the monopoly mark-ups that an unconstrained legal monopolist would choose. Prices are determined generally by commissions that review costs, and often investment decisions, and set prices to cover allowable costs and a 'fair rate of return' on 'prudent' capital investments.

Early economic research on the regulated public utility sector was generally theoretical and focused on how to perfect regulatory pricing, treating as given the desirability and general effectiveness of government regulation. Optimal pricing questions also arose for state-owned utility enterprises, and French and British economists were in the vanguard of research in this area (Kahn 1988, Laffont 1994, and *Poverty: Measurement and Analysis*). Stigler and Friedland's (1962) study of the early electricity industry was one of the first empirical analyses of the actual effect of utility regulation, and concluded that regulation did little to reduce prices below unregulated monopoly levels. While the particulars of this research generated controversy at the time, and subsequent work suggests that electricity prices in the Friedland-Stigler sample were, in fact, lower for regulated firms (Peltzman 1993), the notion that regulation might not improve

upon market outcomes even in those sectors with the strongest public interest rationale stimulated considerable additional research.

The literature subsequent to Stigler and Friedland (1962) suggests generally that price regulation, particularly in the electricity and cable television industries, has constrained average prices below the level an unconstrained legal monopolist would choose. Analysis of average price levels masks the effect of regulation on price structures, however. Price structure analyses in electricity and telephone industries suggest redistributive rather than Ramsey-like efficient pricing objectives, often with considerable transfers from customers with relatively elastic demands to those with relatively inelastic demands. For example, regulatory cross-subsidies from long-distance telephone rates to basic residential telephone service and from urban to rural telephone customers were a major force driving new entry into the US long-distance sector, leading ultimately to the 1984 divestiture of AT&T and the regulatory reform of the telephone industry. The tenacity of these cross-subsidies more than 15 years later testifies to the strength of political interests supporting these policies. Some economists have estimated the social welfare cost of the residential service cross-subsidy to be as high as \$7.0 billion per year (in Peltzman and Winston 2000, p. 78).

The effect of natural monopoly regulation on static production costs is less well-measured. One of the most developed theoretical and empirical literatures on incentives for excessive capital investment in rate-of-return regulated industries proved largely inconclusive (Averch and Johnson 1962, Joskow and Rose 1989). There is some evidence of higher wages in firms that expect regulators to flow increased labor costs through to higher consumer prices, though wages do not tend to be higher overall in regulated sectors (Hendricks 1975, Peoples 1998). In the railroad industry, low regulated profit rates led to underinvestment in track maintenance and repair on many lines, reducing efficiency and increasing average transit time over the track.

Of particular concern is the effect of regulation on the pace of technological innovation. This may occur from direct regulatory constraints on new process or product introductions, such as the Federal Communications Commission's delay in allocating radio spectrum to wireless telephone systems in the USA and its restrictions on enhanced telephone service offerings. Indirect regulatory effects may also influence the pace of technological change by altering expected payoffs to investment in new technology. While often difficult to measure, these costs can be enormous. For example, Hausman estimated lost surplus from the delayed introduction of wireless phone service in the USA at more than \$30 billion annually (in Peltzman and Winston 2000, p. 80).

The ambiguity of cost and innovation effects may arise in part from the difficulty of establishing an

appropriate empirical counterfactual benchmark to regulation, an essential prerequisite to measuring either static or dynamic efficiency effects. Nevertheless, ongoing policy concerns about the efficiency of production, investment incentives, and the pace of technological change in natural monopoly sectors have fueled much of the recent movement toward market-based reforms, discussed in Sect. 3.

2.2 Empirical Analysis of Regulation in Multifirm Industries

A substantial portion of the regulatory apparatus built during the twentieth century involved multifirm industries, such as trucking, airlines, natural gas and oil production, property and liability insurance, and banking. Regulation in many of these was rationalized by the deleterious effects of excessive, rather than insufficient, competition, particularly during the depression of the 1930s. In others, such as airlines, regulation was advocated to promote the orderly development of an emerging commercial industry. Many of these regulatory systems were dismantled or reformed substantially between 1975 and 1985, reflecting a broad consensus on its social costs.

Multifirm price regulation usually requires all firms serving a given market to adhere to a standard tariff. In some industries, this tariff was set by the regulatory commission, as the Civil Aeronautics Board did for the US airline industry. In others, tariffs were set collectively by the firms and ratified by the regulatory commission, as the Interstate Commerce Commission did for trucking rate bureau filings. Regulation generally discourages entry or market extension. By eliminating the ability of existing firms to discount rates and the ability of new firms to enter the market, regulation limits competition and tends to raise prices significantly. For example, economists estimate that regulated prices in the late 1970s were at least 15–20 percent above competitive levels in the trucking industry and 25–30 percent above unregulated prices in the airline industry.

High prices generate significant potential rents. The value of these rents is embodied frequently in the market value of regulatory assets, such as taxicab medallions, operating certificates for regulated trucking firms, or state liquor licenses. Asset values may reflect only a portion of the rents generated by regulation, however. In some sectors, organized labor was able to garnish a substantial share of regulatory rents (Rose 1987, Peoples 1998). Members of the Teamsters Union, for example, were a major beneficiary of trucking regulation, with wages 50 percent above those for nonunion truck drivers and drivers outside the regulated industry. In other industries, regulatory rents were dissipated through various rent-seeking behaviors. This seems to have been particularly

acute in the airline industry, where nonprice competition repeatedly eroded load factors and profit margins despite increasing regulated prices. In the rare situations in which regulatory policy effectively constrains maximum prices and exit from markets, rents may be negative. In the short-run, this can transfer quasi-rents from owners of regulated capital to consumers of their products. Over the longer-run, its adverse effect on investment tends to erode both the quantity and quality of service provided. The rash of bankruptcies in the US railroad industry during the 1970s and early 1980s, and natural gas shortages during periods of below-market ceiling prices in the 1960s and 1970s, were consequences of this type of policy.

As in the natural monopoly sectors, regulation influences the structure of prices as well as the average price level. Economic regulation often leads to a more uniform price structure than in unregulated markets. Prices tend to be less responsive to cost differences across consumers or over time, due in part to cross-subsidies for favored higher-cost consumer groups and to aversion to congestion-based or peak-load pricing. This distorts both capacity utilization and firms' investment incentives. These effects have been analyzed in the context of traffic mix distortions between railroads and trucking firms, provision of air service to small communities, and excessive pooling in insurance markets.

Postderegulation experience suggests that regulation often suppressed variation due to price discrimination, relative to the unregulated market. Airlines are perhaps the best example of this. It is important to recognize that price discrimination, while controversial, may enhance overall welfare relative to a uniform price structure; see the literature on Ramsey-Boiteaux pricing (Kahn 1988). In airline markets, for example, the pervasive use of low discount fares to fill seats that would otherwise depart vacant increases substantially the population for whom air travel is viable, and raises airline revenue per available seat-mile. The net welfare effect of current levels of price discrimination is difficult to measure, and has not been directly calculated for any deregulated market.

Economic regulation in multifirm industries has been associated with both direct and indirect cost effects. Regulatory restrictions and rigidities may prevent firms from optimizing production in both static and dynamic contexts. For example, entry restrictions in the trucking industry led to high rates of partial or empty backhauls and circuitous routings, raising the cost of transporting a given commodity flow. Unit banking regulations, which restrict a firm to a single physical location in one state, both limit realizations of economies of scale in provision of service and insulate less efficient banks from takeovers by more efficient competitors. Inefficiencies such as these generate potentially huge regulatory deadweight losses (Winston 1998).

The welfare effects of cost increases may be difficult to quantify when associated with offsetting changes in the product provided to consumers. For example, the suppression of price competition may encourage nonprice competition, increasing both costs and quality relative to outcomes in unregulated markets. In the regulated airline industry, increased flight frequency and rapid adoption of new aircraft increased both costs and value to passengers. Measuring cost or productivity effects without controlling for the value of higher service quality leads one to overstate the adverse welfare effects of regulation in situations like this. Recent advances in structural demand estimation may help to quantify the value of particular quality attributes, though both data and modeling limitations continue to make this one of the weaker empirical links in most assessments of the overall effects of regulation.

3. The Effects of Restructuring Regulated Natural Monopolies

The focus of regulatory policy and research changed sharply in the late 1980s and 1990s, with the ascendancy of market-based institutions over state-ownership and regulation in even traditional 'natural monopoly' sectors. Restructuring, requiring firms to divest along vertical segment lines, has become common. Segments viewed as structurally competitive, such as electricity generation, have been increasingly deregulated. Network segments that continue to operate under natural monopoly cost conditions, such as electricity transmission grids or distribution systems, generally are required to provide access to competitive suppliers, often at terms subject to continued regulatory oversight. Where economic regulation has been retained, it has become more market-oriented. This reflects both dissatisfaction with recent performance under regulation and a response to theoretical research emphasizing attractive properties of incentive-based regulatory schemes (see *Poverty: Measurement and Analysis*)

These changes raise new questions about market design and regulatory interfaces. The emerging empirical literature has focused on three main issues: the impact of privatization, the performance of restructured sectors, and the effects of incentive-based regulation. Policy in this area is very much in flux, as the initial wave of enthusiasm for market solutions confronts unanticipated consequences from their implementation.

3.1 The Impact of Privatization

Privatization has been motivated by a range of concerns including the cost of service, pace of technological innovation, and impact of continuing subsidies on government budget deficits. A review of the

voluminous literature on privatization is beyond the scope of the present article; see Vickers and Yarrow (1991), La Porta and López-de-Silanes (1999), and their references. Because privatization has been so closely linked with restructuring and regulatory reform initiatives over the past quarter-century, however, it is useful to highlight key empirical results.

One of the most robust empirical results from this literature is the dramatic impact privatization typically has on cost reduction. This is particularly evident in labor costs. Substantial labor shedding has accompanied privatization almost everywhere, with labor force reductions of 50 percent or more a common result. Where comparisons with privately-owned firms are feasible, studies suggest that postprivatization labor rates approach those at firms that were never government-owned. This suggests considerable excess labor in state-owned enterprises. This may reflect the use of state-owned enterprise to promote certain social policies as well as direct operating inefficiencies, suggesting that excess labor costs may include both transfers as well as social dead-weight loss.

Privatization often leads to both price and output increases. Rising output despite higher prices suggests the elimination of artificial constraints, such as rationing, or higher quality service for which consumers are willing to pay more. In numerous studies of utility sectors, waiting time for service falls and investment and service quality increase following privatization. This seems most common in sectors that were capital-starved or in which service was underpriced under state-ownership. Even in countries where state-owned utilities were perceived to be functioning relatively well, however, increased investment and innovation may be important benefits of privatization.

One caveat is in order: the frequent concurrence of privatization with changes in the competitive environment can make it difficult to empirically disentangle the relative contributions of ownership change, hard budget constraints, and increased competition to improved performance. Moreover, it appears that the benefits of privatization may be greater when accompanied by increased competition or deregulation. Studies vary considerably in their recognition and successful treatment of this problem.

3.2 *Performance in Restructured Natural Monopoly Sectors*

Vertical restructuring has become common in natural monopoly industries worldwide. In telecommunications, equipment provision and long-distance service were among the first to be carved out of vertically-integrated telecommunications firms and opened to competitive suppliers. Power generation has been divested from transmission and distribution functions and opened to competition in many electricity systems. Much of this policy movement appears stimulated by

concerns that prices charged by monopoly providers were too high, relative to those attainable by privately-owned firms operating in competitive sectors. In some cases, this reflects real operating inefficiencies and reduced innovation under monopoly. In other cases, apparently high prices may in part reflect perverse regulatory pricing schemes. Action on these concerns was facilitated by technological innovations that reduced scale economies at certain production stages and enabled decentralized operations to achieve many of the economies of scope previously attained by vertically integrated firms.

Economists have been especially interested in assessing the competitiveness of restructured markets. Empirical research suggests these markets have yet to converge to competitive ideals. In the US long-distance telephone market, for example, AT&T lost its sanctioned monopoly provider status in 1984, following the earlier technological erosion of its 'natural monopoly.' While long-distance prices have declined, the convergence to competitive norms has been relatively slow, with average per-minute prices more than double estimated long-run incremental costs as late as 1998 (Peltzman and Winston 2000, p. 77). Some economists have suggested that this reflects relatively weak competition among the small and decreasing number of existing long-distance providers.

Analyses of electricity markets echo this theme. In most restructured electricity markets, unregulated generators bid to supply electricity to the system operator, which determines the dispatch order of generators to supply the electricity demanded at each moment in time. There is considerable evidence of generator market power in many markets, with generators bidding above their marginal generation costs and strategically withholding capacity from the market to increase prices (see Wolfram 1999, Borenstein et al. 2000, Joskow in Peltzman and Winston 2000, and Joskow and Kahn 2001).

While many restructured markets have been prone to supracompetitive prices, the extent to which price exceeds cost varies tremendously across different settings. There is growing evidence that performance variations are heavily dependent upon the institutional details of the market's design. These include the way the competitive sector is structured, the terms on which suppliers interface with their customers and the network, and the management of network investments. In electricity markets, for example, the relative sizes of different suppliers, their inframarginal capacity, the use of long-term hedging contracts between generators and their customers, and the degree to which customers see and can respond to real-time price variation have all been demonstrated to influence the exercise of market power.

While competition may enhance cost efficiency in the deregulated sector, restructuring may increase other costs. For example, substitution of market mechanisms for internal decision-making or 'com-

mand and control' by a vertically integrated firm may increase coordination costs in electricity markets (Joskow in Peltman and Winston 2000). In California, for example, reserve margins and hence system operating costs have increased as a result of moving to a dispatch market. Local distribution companies may face difficulties when their requirement to serve, which was imposed in exchange for franchise monopoly in many utility sectors, is maintained, but restructuring eliminates their direct control over electricity supplies, as evidenced by California utilities in 2000–2001.

In telecommunications, access issues focus on pricing and interconnection requirements for the 'local loop,' the set of wires connecting individual subscribers to the local network. These problems are magnified by the potential for local exchange carriers to compete with other firms for services beyond the local loop. The determination of access terms has substantial effects for investment incentives by incumbent exchange companies and potential competitors, entry into local, long-distance, and enhanced services, and the performance of competitive markets where entry occurs. This is the focus of much current regulatory policy and some empirical analysis (Crandall and Hausman in Peltzman and Winston 2000).

3.3 Effects of Incentive-based Regulation

While regulatory reform over the past two decades has been substantial, economic regulation continues to play an important role in many sectors. Economic regulation in the new millennium is more market-based than its counterpart of 1970, but less market-driven than current rhetoric would suggest. Policy innovation has been particularly notable outside the USA, perhaps because these systems were being designed *de novo* rather than patched on to existing regulatory practices. Regulation has moved away from traditional rate-of-return regulation, which determined prices through cost-based mark-up rules (Kahn 1988). More common are higher-powered incentive schemes such as price caps, which fix price levels or rates of change, and allow firms to keep the surplus generated from unanticipated improvements in efficiency, at least until the next review of the cap or its adjustment factor. The models behind these schemes are discussed elsewhere (see *Poverty: Measurement and Analysis*). It is important to recognize that implementation, however, generally deviates considerably from the theoretical construct.

Many researchers have attempted to measure the impact of incentive-based regulation, particularly in the telecommunications sector. This literature broadly suggests that incentive regulation is accompanied by lower prices. Kridel et al. (1996) suggest that these effects may reflect coincident policies that increase competition more than the incentive effects of incentive regulation itself, however. This interpretation

is particularly suggested by the difficulty in finding consistent effects of incentive regulation on costs or productivity, although the short time span of many studies may impede detection of such effects. As earlier empirical work in regulation suggests, one should exercise great caution in affirming the superiority of incentive-based policies based on the theoretical properties of idealized schemes.

4. Conclusion

The reforms of the 1990s have created a vast new set of institutions and regulatory policies. This is an area in which empirical analysis greatly lags behind the theoretical literature, and the returns to additional careful work are likely to be significant. Empirical research in regulatory economics, argued by some in the 1980s to belong to the domain of economic history, remains vital to guiding, understanding, and evaluating policy initiatives today.

See also: Cost–Benefit Analysis; Law and Development; Policy Knowledge: Advocacy Organizations; Poverty: Measurement and Analysis; Regulation, Economic Theory of; Regulation, Political Economy of; Socialism

Bibliography

- Averch H, Johnson L L 1962 Behavior of the firm under regulatory constraint. *American Economic Review* **52**: 1052–69
- Borenstein S, Bushnell J, Wolak F 2000 *Diagnosing Market Power in California's Restructured Wholesale Electricity Market*. National Bureau of Economic Research Working Paper 7868, Cambridge, MA
- Hendricks W 1975 The effect of regulation on collective bargaining in electric utilities. *Bell Journal of Economics and Management* **6**: 451–65
- Joskow P L 1974 Inflation and environmental concern: Structural change in the process of public utility price regulation. *Journal of Law and Economics* **17**: 291–327
- Joskow P L, Kahn E P 2001 *A Quantitative Analysis of Pricing Behavior in California's Wholesale Electricity Market During Summer 2000*. National Bureau of Economic Research, Cambridge, MA
- Joskow P L, Noll R G 1994 Economic regulation during the 1980s. In: Feldstein M (ed.) *Economic Policy During the 1980s*. University of Chicago Press, Chicago
- Joskow P L, Rose N L 1989 The effects of economic regulation. In: Schmalensee R, Willig R (eds.) *Handbook of Industrial Organization*. North-Holland, Amsterdam, Vol. 2
- Kahn A E 1988 *The Economics of Regulation: Principles and Institutions*. MIT Press, Cambridge, MA
- Kridel D J, Sappington D E M, Weisman D L 1996 The effects of incentive regulation in the telecommunications industry: A survey. *Journal of Regulatory Economics* **9**: 269–306
- La Porta R, López-de-Silanes F 1999 The benefits of privatization: Evidence from Mexico. *Quarterly Journal of Economics* **114**: 1193–242
- Laffont J-J 1994 The new economics of regulation ten years after. *Econometrica* **62**: 507–37

- Peltzman S 1993 George Stigler's contribution to the economic analysis of regulation. *Journal of Political Economy* **101**: 818–32
- Peltzman S, Winston C (eds.) 2000 *Deregulation of Network Industries: What's Next?* AEI-Brookings Joint Center for Regulatory Studies, Washington, DC
- Peoples J 1998 Deregulation and the labor market. *Journal of Economic Perspectives* **12**: 111–30
- Rose N L 1987 Labor rent-sharing and regulation: Evidence from the trucking industry. *Journal of Political Economy* **95**: 1146–78
- Stigler G J 1971 The theory of economic regulation. *Bell Journal of Economics and Management* **2**: 3–21
- Stigler G J, Friedland C 1962 What can regulators regulate? The case of electricity. *Journal of Law and Economics* **5**: 1–16
- Vickers J, Yarrow G 1991 Economic perspectives on privatization. *Journal of Economic Perspectives* **5**: 111–32
- Winston C 1998 US industry adjustment to economic deregulation. *Journal of Economic Perspectives* **12**: 89–110
- Wolak F 2001 *An Empirical Analysis of the Impact of Hedge Contracts on Bidding Behavior in a Competitive Electricity Market*. National Bureau of Economic Research, Cambridge, MA
- Wolfram C D 1999 Measuring duopoly power in the British electricity spot market. *American Economic Review* **89**: 805–26

N. L. Rose

Regulation: Family and Gender

There are few more politically sensitive areas than the role the state should play in regard to 'the family.' It is not easy for liberal democratic states to justify intervention in the private sphere of personal relationships and family life. In the first place memories of extreme examples of state intervention during the twentieth century, most notably in Nazi Germany, where the state dictated who married and who had the right to reproduce, make this difficult territory. Second, even when governments have little by way of explicit family policies, they usually make implicit assumptions about what families should look like, which in pluralist societies are bound to be at odds with the way in which some people organize their affairs. Not only have modern Western states tended to assume that the family will consist of two parents and children, but historically both government policies and the professional experts employed by the state have also made gendered assumptions about the part played by men and women in family life, particularly in respect of responsibilities for earning and for caring, which are increasingly at odds with the social reality.

Feminists have been one of the most vocal groups criticizing government policies on the family, but the relationship between parents, children, and the state has been one of the most hotly debated issues of the twentieth century, as important as the more common touchstone of individualism vs. collectivism. Gender,

understood as a fundamental organizing principle in society (Scott 1988), is crucial to the analysis of the relationship between the state and the family. The concept of 'gendered divisions,' particularly in respect of resources, labor, and investments, is key to understanding the development and impact of the public policies and family law that regulate the family.

1. Ideas about the Nature of State Regulation and the Family

The debate about state regulation, the family, and gender in the last quarter of the twentieth century has revolved around the extent to which states have propped up a particular family form, but the positions taken are complicated and do not follow a simple political division between Left and Right.

Feminists have argued that for governments, family policy all too often meant ways of bolstering the traditional two-parent family, with a male breadwinner and an economically dependent female carer, and that policies rarely impacted on different family members in the same way. For example, in the 1970s, Scandinavian feminist writers proposed that the modern welfare state amounted to a new form of 'public patriarchy,' the argument being that the dramatic increase in women's employment, which began in Sweden and Denmark in the 1970s, involved for the most part women leaving their homes to engage in similar caring work, often for children, in the service of the state. In the UK, Wilson (1977) suggested that the postwar welfare state amounted to nothing less than the state organization of domestic life, by which she meant that state policies, for example in respect of cash benefit entitlements, sought to reinforce the male breadwinner family model. A few years later, a similar kind of criticism was made from an entirely different quarter.

Arguing from a 1980s New Right position, Mount (1982) suggested that the family had been in 'permanent revolution' against the unwarranted intrusions of the state. In his view, even apparently helpful public visitors, such as home nurses, were essentially seeking to impose standardized modes of care on families and were able to invoke a wide range of powers if they deemed the family to have 'failed' in some respect. Mount's view keyed into the long-standing suspicions of working people, especially in the English-speaking countries, of officialdom. However, this does not mean that nonintrusive, nonstigmatizing state intervention has not been welcomed. The panoply of Swedish collective social provision has been referred to by the citizens of that country as 'the people's home.'

The idea that state officials practice some kind of surveillance over the family also found favor with followers of Foucault, although Donzelot (1979) stressed the extent to which the professionals in emerging welfare states formed alliances with women

in order to do so. In her history of domestic violence, Gordon (1990) also stressed the extent to which social workers had sided with the weaker family members against the more powerful, thus questioning the extent to which state regulation worked to reinforce traditional power relationships within the family. Indeed, most feminists, while criticizing the kinds of assumptions about the family inherent in social policies, have also looked to the state to work for gender justice (Okin 1989). Few have followed Elshtain (1981) in defending the family as a separate, private sphere, outside the reach of the state.

In recent years the very different idea that the state has played a part in deregulating family life has gained in strength. The argument is that rather than working to uphold traditional family patterns, the state has effectively underpinned a trend towards individualistic behavior that is essentially selfish. This view has been formulated in the context of the dramatic postwar statistics on family change. Marriage rates have fallen in most Western countries; people are marrying later; fertility rates (especially in some Western European countries) have fallen well below replacement rates; divorce rates have increased since the 1970s and remain high and stable in Northern Europe and the US; rates of extramarital childbearing have increased; and behind these changes lies the huge increase in cohabitation. In addition, all Western states have seen an increase in women's employment, albeit predominantly part-time in many countries. These changes in behavior appear to signal greater sexual and economic autonomy on the part of women and, in the view of some, more irresponsible behavior on the part of men. Together with a shift in attitudes towards individualism, which has been documented for a large number of Western countries (Inglehart 1997), the changes in behavior have been widely interpreted as signaling that men and women are less prepared to commit to, and to invest in, family life.

The optimism of academics in the 1960s and early 1970s about the fundamental stability of the traditional two-parent family has given way to profound pessimism at the end of the twentieth century. Bellah et al.'s (1985, p.85) influential study of middle America concluded that 'if love and marriage are seen primarily in terms of psychological gratification, they may fail to fulfil their older social function of providing people with stable, committed relationships that tie them into the larger society.' The causes of behavioral and attitudinal change are extremely complex, but the fear of many commentators is that both public policy and family law have encouraged rather than countered them.

In respect of public policy, Murray (1984) proved particularly influential in the English-speaking world in arguing that the provision of welfare benefits, far from sustaining the male breadwinner model family, allowed women a measure of financial autonomy and thus promoted family breakdown. In respect of family

law, Glendon (1981) argued strongly that legal ties among family members in the postwar period were loosened. In particular, the move that took place in most Western countries towards the implementation of no-fault divorce in the 1960s and 1970s has been interpreted in terms of the withdrawal of government from the attempt to impose an external moral code in favor of allowing the parties to a marriage to work out their own destiny. A similar view has been taken by leading commentators in non-English speaking countries with very different legal traditions. Thus Thery (1994) has suggested that marriage has become a private choice and that marriage as an institution has become disengaged from wider social structures, the law included.

In whatever way the role of the state in regulating the family is interpreted, the solutions are far from easy to determine. Those who are suspicious of the state, whether because they feel that it has perpetuated traditional family forms and gender roles, or because they feel that these have been effectively undermined by the state, have difficulty in trusting government to do better, and in any case will demand radically different kinds of action. However, the picture of a homogenous state acting either to support or to undermine traditional family patterns is in all probability too simple. In the first place, the relationship between law and behavior is very difficult to determine. For example, Phillips' (1988) comparative study of divorce has shown that levels of marital breakdown are not dependent on legal change, and may be high in the absence of a liberal divorce law. Nor should people's resistance in the face of legal change be underestimated. Mount's account of the family resisting state interference is supported by findings from comparative studies that show legal change to have less effect on family behavior than on other areas of human activity. Second, the various departments of Western liberal governments rarely act coherently. Different departments may pursue very different agendas in respect of an issue such as childcare, for example.

Indeed, it may be more helpful to think of government regulation in the area of the family as layered. In all probability, Western governments in the twentieth century have tended to be Janus-faced in regard to family policy and law. While most have had a clear idea of the (traditional, two-parent) family form they wish to promote, policies have often pulled in a different direction. Thus, for example, while most income maintenance systems have a cohabitation rule, which prohibits the payment of welfare benefits to lone mothers who have a 'man in the house' (in accordance with the assumptions of the male breadwinner model family), they have also served to facilitate and legitimate, if not actually cause, the formation of lone mother families (Kiernan et al. 1998). The means of regulating the family have been many and various and certainly include the huge and

vexed area of birth control, abortion, and reproduction (see *Regulation: Sexual Behavior*). Historically, public policy and family law have sought to impose two powerful prescriptive frameworks on the family: that of the male breadwinner model, which has had a considerable impact on the gendered division of labor, and a strict moral code, which resulted in the stigmatizing of divorce and extramarital births. These frameworks were embodied in legislation but were also internalized by ordinary men and women. Their erosion in the late twentieth century has meant that increasingly people must negotiate their own solutions.

2. Public Policy

Given that in modern societies independence derives primarily from wage-earning, the assumption that women are located mainly in the private sphere as wives and mothers, supported by breadwinning husbands, has also meant that women have only been partially individualized for the purposes of public policy making (Pateman 1989). The liberal dilemma first described by Okin (1979), whereby for the purposes of determining political citizenship 'individual' effectively meant 'male head of household,' persisted. Thus most Western states have subscribed to some degree to the idea of a male breadwinner model family and have therefore treated women as dependants of men, although this set of assumptions has been modified to very different degrees in different countries (Lewis 1993). In its pure form, the male breadwinner model assumes that married women will not engage in paid work but will rather undertake the work of caring for children and other dependants at home without the support of collectively provided services, and that they will be subservient to their husbands for the purposes of social security entitlements and tax. In reality, such a model never existed; working class women have always gone out to work as and when the family economy demanded it. However, the model matched the social reality of middle class women fairly closely at the turn of the century. For the late nineteenth century social theorist Herbert Spencer, society was 'progressing' towards a position whereby all women would be able to stay at home in their 'natural sphere.'

It is no coincidence that the foundations of most Western European and Australasian welfare states were also laid down at the turn of the century, when the ideology of 'separate spheres' for men and women was at its height. Skocpol (1992) has argued that the core program of modern welfare states—social insurance against sickness and unemployment—is fundamentally 'paternalist.' The program sought to protect the full-time, respectable male worker. Women were rarely covered as workers in their own right and were usually eligible for benefits only as the wives of

male employees, the justification being a division of labor that was perceived to follow 'naturally' from their capacity for motherhood. Women have thus tended to make (smaller) contributions and draw (smaller) benefits via their husbands. Furthermore, in welfare regimes such as that of the US and early and mid-twentieth century Britain, where the social security systems developed dual social insurance and social assistance benefits, this provision was gendered, with first class (insurance) benefits going mainly to men as workers, and second class (assistance) benefits, such as aid to families with dependent children in the US, to women, usually in their capacity as mothers.

Nevertheless the male breadwinner logic that informed most early twentieth century social protection was broadly accepted, not just by men and by policy makers who regarded the responsibility for dependants as a guarantor of male work incentives, but by women themselves. In the context of hard household labor and frequent pregnancy the male breadwinner model was a shared ideal. This changed in the postwar period as fertility rates fell and housework became less onerous. Assumptions regarding the existence of a male breadwinner model were eroded to different degrees in different countries, but the assumptions informing public policy usually lagged behind behavioral change. In France, because of the very different occupational structure (of family farms and firms), women tended to gain entitlements both as wives and mothers on the one hand, and as paid workers on the other.

Given the additional importance of pronatalism in France, the former tended to be more generous, but abundant public child care provision made possible a long tradition of full-time work for women. The Scandinavian countries moved furthest away from the male breadwinner logic in terms of public policy provision, deliberately pulling women into paid employment from the 1970s by the introduction of separate taxation and parental leaves and by increasing child care provision, to the point where the dual-breadwinner family became the norm. The contrast with Britain and Germany is striking. The latter has no independent taxation, which works against female employment when there are high marginal rates of tax, and a strong social insurance welfare system, in which women's eligibility remains tied to that of their husbands. Until very recently neither country made a significant commitment to collectively provided care services. In Germany female labor market participation rates have been low in the postwar period, while British women have tended to work short, part-time hours. In the US, where collective provision for able bodied adults did not develop beyond assistance benefits for lone mothers, state policy has played little part in determining how male and female members of the household decide the division of paid and unpaid work. However, the absence of explicit policies does not mean the absence

of a regulatory framework. The policy logic operates to constrain choices: adult women who are also mothers must depend on either men or on their own efforts in the labor market.

Policies towards lone mothers provide a particularly striking illustration of the extent to which public policy frameworks do have an impact on behavior. In the context of assumptions regarding the traditional two-parent family, the problem of lone motherhood becomes that of women with children and without men. In the absence of a male breadwinner, the state has to decide to what extent and under what conditions it will step in to fill the place of the missing husband; how far it will exert pressure on an able-bodied absent father to support his child and possibly its mother; and how far it will expect the lone mother to support the child by herself becoming a breadwinner. Policies in respect of each of these have varied enormously between countries and over time. But broadly speaking there has been a pendulum effect.

In the nineteenth century, lone mothers were treated as workers and were expected to support as many children as they could; by the mid-twentieth century they were more likely to be treated as mothers and permitted to draw benefits for as long as they had children under a particular age, and by the end of the century policy in most Western countries has veered again towards treating them as workers, although only in the US has the right to draw benefit been removed (see *Lone Mothers in Affluent Nations* and *Lone Mothers in Nations of the South*).

3. Family Law

The family has been seen both as bedrock and yet as fragile. Historically, the stability of marriage and the family has been seen as fundamental to the stability of the *polis*. Sociologists have suggested that marriage has moved from institution to relationship. In fact it remains both, but the regulation of the institution in accordance with an externally prescribed moral code has been substantially eroded. In the last quarter of the twentieth century Western states have largely ceased to regulate the role of men and women as husbands and wives. In respect of family law, the attention of the state has turned to regulating their roles as fathers and mothers.

Whereas the erosion of the male breadwinner model lagged behind social change, many would argue that changes in family law have run ahead of the social reality, certainly in terms of mentalities if not behavior. The reasons for abandoning the attempt to impose a strict moral code have been attributed to changing moral beliefs with an increased tolerance for pluralism, the rise of a psychological view of personal affairs, and the increased emphasis on individual rights. Nevertheless, most of those campaigning for reform of the

divorce laws in the 1960s were convinced that they were proposing a higher form of morality that would be based on love and that would come from within, rather than being imposed from outside. The removal of fault from divorce law made it impossible to allocate blame and then, as a matter of course, to award long-term financial support to 'innocent' female spouses. Reform of family law meant that the state ceased to regulate the dissolution of marriage and concentrated on adjustment between the parties. Under a no-fault regime, a marriage without children is effectively terminable at will. The import of the process of liberalization has been to let men and women as husbands and wives order their own affairs, in other words to assume the existence of a degree of individualization, while also putting in place measures that seek to regulate them as fathers and mothers. The problem is that it has been difficult to make a neat separation between these two sets of roles, not least because of the need of children for caretakers and the tendency for women to perform more of this unpaid work.

The idea that divorcing couples should be treated in exactly the same way, thereby ignoring the different nature of their contributions to the marriage and the inequalities resulting from women's greater investment in childbearing and childrearing, followed the introduction of no-fault divorce. By the 1980s, in the English-speaking countries, the presumption of female dependence had been replaced in family law by the equally problematic assumption that husbands and wives were equal members of a partnership. This ignored the substantive inequalities which meant both that the view of the more powerful party as to the desirability of continuing the relationship might prevail, and that the outcomes for the economically weaker partner were likely to be poor. As academic research sounded the alarm about the latter (Weitzman 1985), and about the effects of divorce on children (Wallerstein and Kelly 1980), so the state made greater efforts to regulate the roles of men and women as fathers and mothers, focusing particularly on fathers.

In the face of rapid family change, the growing numbers of lone mother families and the resulting rise in the burden on public expenditure, governments sought to enforce the responsibility of men for their biological children. In the US, Australia, and the UK in the 1980s and 1990s new child support legislation was passed to this effect. Continental European countries had a much longer history of efforts to secure maintenance from absent fathers. A government-appointed commission in the UK examined these in the early 1970s, but rejected them as too intrusive, which indicates how far opinion moved in the last quarter of the century. If husbands and wives were to be left alone, parents were not. Fathers were also increasingly expected to have contact with their children, in part because the psychological evidence showed this was important for children, in part

because policy analysis showed that fathers who maintained contact were also more likely to pay for their children, and in part as a response to fathers' rights campaigners.

However, it has not been easy to deal with children separately from their carers. No-fault divorce removed the link between the grounds for divorce and maintenance without substituting any alternative principle to justify the latter. The relaxation of the divorce law took place in Western liberal democracies which felt that it was neither possible nor desirable to regulate the marriage system such that men's right to remarry and further reproduce was curtailed. This made it difficult to enforce the responsibilities of men in relation to their biological children, and by extension, those children's carers.

4. Future Prospects

The assumptions lying behind traditional marriage constructed husbands and wives as equal-but-different partners, something public policy accepted when it assumed the existence of the male breadwinner model family. In so doing, it constructed women as dependants and ran the risk of failing to recognize either the nature of their contribution as both paid and unpaid workers or the entitlements that were their due. Family law, on the other hand, moved towards a construction of marriage as a partnership of formal equals after the 1960s, which ran the risk of ignoring the substantive inequalities between men and women. Thus the problems of the relationship between the state and the family in regard to gender hinge on the larger problem of equality and difference.

Faced with family change that has been both dramatic and rapid in the last quarter of the twentieth century, Western governments have two basic options: to recognize the social reality and work with it or to seek to put the clock back. Just as it is possible to see the state acting both to bolster the traditional family and to transform it, so it is possible to identify ways in which the state is recognizing and resisting change. To move towards an emphasis on the responsibilities of men and women as parents is to recognize the changes that have taken place in the marriage system, but as Smart and Neale (1999) have argued, at the same time public policy and family law have attempted to perpetuate the traditional responsibilities of men for maintenance and women for care beyond the life of their intimate relationships.

See also: Family and Gender; Family, Anthropology of; Family as Institution; Family Law; Family Planning Programs: Development and Outcomes; Family Planning Programs: Feminist Perspectives; Father-

hood; Fertility Control: Overview; Gender and the Law; Reproductive Rights in Affluent Nations

Bibliography

- Bellah R, Madsen R, Sullivan W, Swidler A, Tipton S M 1985 *Habits of the Heart: Individualism and Commitment in American Life*. University of California Press, Berkeley, CA
- Donzelot J 1979 *The Policing of Families* (trans. by Robert Hurley). Pantheon, New York
- Elshtain J B 1981 *Public Man Private Woman: Women in Social and Political Thought*. University of Princeton Press, Princeton, NJ
- Glendon M A 1981 *The New Family and the New Property*. Butterworths, Toronto, ON
- Gordon L (ed.) 1990 *Women, the State and Welfare*. University of Wisconsin Press, Madison, WI
- Inglehart R 1997 *Modernization and Post-modernization: Cultural, Economic and Political Change in 43 Societies*. Princeton University Press, Princeton, NJ
- Kiernan K, Land H, Lewis J 1998 *Lone Motherhood in Twentieth Century Britain: From Footnote to Front Page*. Oxford University Press, Oxford, UK
- Lewis J (ed.) 1993 *Women and Social Policies in Europe: Work, Family and the State*. Edward Elgar, Aldershot, UK
- Mount F 1982 *The Subversive Family: An Alternative History of Love and Marriage*. Allen and Unwin, London
- Murray C A 1984 *Losing Ground. American Social Policy 1950–1980*. Basic Books, New York
- Okin S M 1979 *Women in Western Political Thought*. Princeton University Press, Princeton, NJ
- Okin S M 1989 *Justice, Gender and the Family*. Basic Books, New York
- Pateman C 1989 *The Disorder of Women: Democracy, Feminism, and Political Theory*. Stanford University Press, Stanford, CA
- Phillips R 1988 *Putting Asunder. A History of Divorce in Western Society*. Cambridge University Press, Cambridge, UK
- Scott J W 1988 *Gender and the Politics of History*. Columbia University Press, New York
- Skocpol T 1992 *Protecting Soldiers and Mothers. The Political Origins of Social Policy in the United States*. Belknap Press of the Harvard University Press, Cambridge, MA
- Smart C, Neale B 1999 *Family Fragments*. Polity Press, Malden, MA
- Wallerstein J S, Kelly J B 1980 *Surviving the Breakup. How Children and Parents Cope with Divorce*. Grant McIntyre, London
- Weitzman L J 1985 *The Divorce Revolution: The Unexpected Social and Economic Consequences for Women and Children in America*. Free Press, New York
- Wilson E 1977 *Women and the Welfare State*. Tavistock, London

J. Lewis

Regulation, Political Economy of

Government regulation of firms uses the 'coercive power' of the state to alter firms' pricing, entry, production, investment, or product choice decisions. An extensive empirical literature analyzes the effects of

'economic regulation' of price and entry as well as environmental, health, safety, and information regulation. These analyses amassed substantial evidence of high prices, excess profits, and production inefficiencies under economic regulation (see Joskow and Rose 1989, Joskow and Noll 1994; and *Regulation: Empirical Analysis*). Studies of environmental, health, and safety regulation revealed substantial deviations from least-cost solutions and dramatic variation in measures such as expected cost per life saved across regulatory interventions (see Viscusi 1994, Gruenspecht and Lave 1989). These outcomes, so systematic and pervasive, were difficult to reconcile with 'public interest' models of regulatory intervention. Examination of regulatory systems further suggested that they often deviated from cost minimization or efficiency objectives by design. This begged for an explanation beyond bureaucratic incompetence or malfeasance, common to earlier Marxian or 'capture' theories of government. Stigler's (1971) *Economic Theory of Regulation* provided the conceptual underpinnings of what has become a vast enterprise among economists and political scientists to understand the origin, objectives, maintenance, and reversal of regulatory policy in the context of rational choice models of political behavior. These simple insights stimulated a substantial theoretical and empirical literature (Noll 1989, Peltzman 1993).

Theories of legislative and bureaucratic behavior have become increasingly sophisticated since the 1980s. The literature has moved well beyond Stigler's simple model to incorporate the role of competing interests, policy entrepreneurs, majoritarian political institutions, legislative oversight of regulatory agencies and commissions, and asymmetric information (see Noll 1989, Joskow and Noll 1994, Laffont 1999, and *Regulatory Agencies*). The empirical literature has struggled to keep pace with these developments.

Common to virtually all empirical studies in this area is the importance, and difficulty, of defining well-specified alternatives to the political economy model under investigation. To test a given model of political behavior, or to interpret results as consistent with a given model, one must understand how predicted outcomes differ between competing hypotheses of behavior and the model being tested. For example, there is a natural tendency to identify rent conveyance with rent-seeking activity. In some cases, however, increased rents for regulated firms may be consistent with both successful rent-seeking behavior by those firms and elimination of some market inefficiency. It is important, though often quite difficult, for empirical analyses to identify and attempt to distinguish alternatives such as these. Studies vary considerably in the degree to which they meet this standard.

This article introduces the main empirical methodologies used to explore the political economy of economic, environmental, health and safety regu-

lation. Examples of their applications are briefly discussed. Interested readers are urged to pursue the surveys listed in the references for more complete discussions of this literature.

1. Measuring Distributional Effects

Empirical studies of political economy generally begin with estimates of the distributional effects of a given policy. While many empirical methods may be used (see Joskow and Rose 1989 and *Regulation: Empirical Analysis*), asset market 'event studies' have been particularly popular. These measure price movements in capital assets, such as company share prices, at the time policy changes are first announced. If the announcement is not previously anticipated, and asset markets efficiently process information, these price changes provide estimates of the policy's effect on the capitalized value of expected rent streams. This method may be used to measure the effects of the policy change on firms within a regulated industry, corporate customers and suppliers, and possible competitors. The distributional implications for individual or small business customers, employee groups, and other noncorporate interests usually must be measured using alternative methods.

The simplest political economy studies use measured distributional effects to infer the relative influence of different interest groups in the political process. These implicitly assume that the observed changes in rent distribution are anticipated and intentional, and that larger rent gains reflect greater political power. Consider, for example, the extensive debate over the origin of the first federal economic regulatory commission in the US, the Interstate Commerce Commission created in 1887 to regulate railroads. Was this legislation intended to solve an inefficiency in an industry characterized by large sunk costs and natural monopoly conditions? If the legislation was a response to interest group politics, did it respond to agrarian interests and other shippers' demands to constrain monopoly pricing by railroads, or was it a railroad-sponsored effort to establish a government-enforced cartel? Did legislation intend one outcome, only to be thwarted by the regulatory agency's subsequent behavior? There have been numerous empirical studies attempting to answer these questions, many using railroad share price responses to various regulatory 'events' (e.g., Prager 1989, Mullin 2000). In these, railroad share price gains in response to the Interstate Commerce Act, for example, have been interpreted as evidence that the regulation was intended to benefit railroads, supporting Stiglerian models of regulatory politics. While it may be reasonable to assume that *ex post* gains identify *ex ante* intended beneficiaries, this should be made explicit and considered against alternative explanations. To move convincingly from a positive

statement about the distributional consequences of a policy to an inference about its political motivation requires an understanding of distributional effects under leading alternative models.

2. Estimating Political Support for Regulation

A second empirical methodology estimates the determinants of regulatory incidence or political support as a function of underlying economic interests. These studies generally involve three steps. First, *ex ante* winners and losers of a regulatory policy are identified, using either prior estimates or a model of the regulation's effect. For example, Stigler's (1971) analysis of truck weight laws modeled state truck weight and size limits as raising the cost of trucking service, particularly over longer hauls. Truckers therefore lose from lower limits, as do their current customers, businesses using their own trucks, and potential customers for whom trucking rates constrain alternative transportation rates. Competitors to trucks, such as railroads, gain from tighter restrictions. There may be diffuse 'public interests' as well, for example, in limiting damage to roadbeds and reducing congestion, though these may well be endogenous.

Second, the importance of the identified economic interests within a given jurisdiction or at a particular point in time is measured. There is considerable variation in how this is done. For example, the number of trucks, scaled or unscaled, the dollar volume of activity in trucking, and the fraction of freight hauls susceptible to potential truck competition are all possible measures of the importance of trucking interests in a jurisdiction.

Finally, the incidence or intensity of regulation, or the political support for regulation, is estimated to be a function of the measured economic interests and any other relevant factors. Stigler used this method to explain differences in the intensity of truck weight limitations across states, for example. Other studies apply this technique to model variation in political support, such as voting records, as a function of constituent economic interests in the policy, *inter alia*.

Most studies in this area find that constituent interests are significant determinants of regulatory policy, though interests must be defined broadly to include constituent preferences (such as conservation or environmental preferences) as well as direct economic interests. Detailed studies of voting behavior suggest that the underlying models may be quite complex, however; see, for example, Joskow and Schmalensee (1998) on the US acid rain program. The significant impact of legislators' ideology on their voting behavior (e.g., Kalt and Zupan 1984) suggests possible slack in the agency relationship between constituents and their legislators, as well support for more sophisticated models of voting in majoritarian systems across issues and over time.

3. Modeling Regulatory Agency Behavior

A final strand of empirical work in political economy begins with the recognition that most regulatory legislation leaves considerable discretion over its implementation to regulatory agencies or commissions. Early work on regulatory 'capture' assumed that industry groups could influence regulators to use this discretion to benefit the industry in direct contravention to legislative intent. As economists and political scientists began to explore the extent to which apparent regulatory capture was instead a reflection of legislative intent, they confronted the questions of how much discretion regulators retain and how regulators use that discretion to form policy (e.g., McCubbins et al. 1987). This led to an extensive literature on the operation and oversight of the administrative process, most of it within political science. This area remains one of the least-developed empirical literatures in regulatory economics.

While some early analyses explored the responsiveness of regulatory agency decision-making to congressional preferences, most economics research in this area analyzes the responsiveness of agency decision-making to interest groups and various measures of efficiency or social welfare. This is methodologically quite similar to the legislative studies described in Sect. 2, with individual regulatory decisions as the unit of observation. As in those studies, the results suggest a role for concentrated economic interests to influence regulatory outcomes. This does not itself demonstrate that regulators respond to interest groups in contravention with broader social objectives, however. For example, studies have found that despite significant overall delays in the process by which the Food and Drug Administration (FDA) controls entry of new pharmaceutical products, the FDA responds more quickly to drugs with greater market potential (Dranove and Meltzer 1994). In this case, both producer and social interests are aligned with reducing delay of an innovation.

4. Conclusion

While research on the political economy of regulation has made significant progress since the 1970s, fundamental questions remain unanswered by either theoretical or empirical work (Peltzman 1993, Joskow and Noll 1994). One of the most basic is why some industries are regulated (or deregulated) and others are not. Another is why regulatory policy seems characterized by 'waves' or cycles: changes tend to occur across broad ranges of industries and sectors within fairly short periods, separated by longer periods of relative inactivity. Considerable scope remains for additional work in this area.

See also: Antitrust Policy; Broadcasting: Regulation; Environmental Economics; Industrial Relations and

Collective Bargaining; International Communication; Regulation; Occupational Health and Safety, Regulation of; Policy History: State and Economy; Policy Process: Business Participation; Privatization of the Economy; Regulation, Economic Theory of; Regulation: Empirical Analysis; Regulatory Agencies; Safety, Economics of

Bibliography

- Dranove D, Meltzer D 1994 Do important drugs reach the market sooner? *RAND Journal of Economics* **25**: 402–23
- Gruenspecht H K, Lave L B 1989 The economics of health, safety, and environmental regulation. In: Schmalensee R, Willig R (eds.) *Handbook of Industrial Organization*. North-Holland, Amsterdam, Vol. 2
- Joskow P L, Noll R G 1994 Economic regulation during the 1980s. In: Feldstein M (ed.) *Economic Policy During the 1980s*. University of Chicago Press, Chicago
- Joskow P L, Rose N L 1989 The effects of economic regulation. In: Schmalensee R, Willig R (eds.) *Handbook of Industrial Organization*. North-Holland, Amsterdam, Vol. 2
- Joskow P L, Schmalensee R L 1998 The political economy of market-based environmental policy: The US acid rain program. *Journal of Law and Economics* **41**: 37–83
- Kalt J P, Zupan M A 1984 Capture and ideology in the economic theory of politics. *American Economic Review* **74**: 279–300
- Laffont J-J 1999 Political economy, information and incentives. *European Economic Review* **43**: 649–69
- McCubbins M, Noll R, Weingast B 1987 Administrative procedures as instruments of political control. *Journal of Law, Economics and Organizations* **3**: 243–77
- Mullin W 2000 Railroad revisionists revisited: Stock market evidence from the progressive era. *Journal of Regulatory Economics* **17**: 25–47
- Noll R G 1989 Economic perspectives on the politics of regulation. In: Schmalensee R, Willig R (eds.) *Handbook of Industrial Organization*. North-Holland, Amsterdam, Vol. 2
- Peltzman S 1993 George Stigler's contribution to the economic analysis of regulation. *Journal of Political Economy* **101**: 818–32
- Prager, R 1989 Using stock price data to measure the effects of regulation: The Interstate Commerce Act and the Railroad Industry. *RAND Journal of Economics* **20**: 280–90
- Stigler G J 1971 The theory of economic regulation. *Bell Journal of Economics and Management* **2**: 3–21
- Viscusi K 1994 Health and safety regulation. In: Feldstein M (ed.) *Economic Policy During the 1980s*. University of Chicago Press, Chicago

N. L. Rose

Regulation: Sexual Behavior

The regulation of sexual behavior refers to the enactment of laws and legal codes at the federal, state, or local level whose intent is to define and restrict specific sexual conduct. Sexual conduct encompasses a range

of behaviors and activities, only some of which are regulated by law depending on the age, marital status, and legal residence of the individual. Sexual behaviors include those practiced between individuals—as in any intimate bodily contact such as kissing, touching, intercourse, oral and/or anal sex—and those practices, usually solitary, such as masturbation and nude public exposure. While all societies have enacted some form of a regulatory mechanism to prohibit sexual behaviors considered socially offensive or physically and/or psychologically harmful to its members, the regulation of sexual behaviors typically reflects ‘a society’s particular vision of sex, marriage, and the family’ (Francoeur 1991a, p. 583). In the case of the USA, sexual regulations have historically served to reinforce traditional moral values in the form of social and religious taboos proscribing non-procreative sexual behaviors, including those practiced within and outside of marriage. Changes in social relations and in the expanding configurations of coupling, marriage, and family structure since around 1950, have resulted in more tolerant and permissive attitudes about sexuality and more diverse sexual lifestyles (D’Emilio and Freedman 1988). Many of the existing legal regulations are either not enforced, in the process of reform or repeal, or are confronting considerable resistance. Moreover, cultural forces, such as changing gender roles and norms of sexual development, influences of mass media, popular culture, information technology, and Internet access, as well as alterations in the workplace have provided new issues with regard to the regulation of sexual behavior, which continue to be intensely contested in the political and legal arenas in the USA. Since around 1980, such issues as sexual harassment, the sexual abuse of women and girls, the role of sexually explicit materials in public and private life, government policy regarding sexual orientation and military personnel, and the prevention and treatment of HIV/AIDS have placed this issue at the center of social controversy (Gagnon 1991).

1. The Legal Sphere: Protection vs. Restriction

The multifaceted role of our legal system is particularly evident in this complex process of enacting, interpreting, and/or disputing the laws regulating sexual behavior. In this process, the legal system is intermittently supported or challenged by a myriad of cultural, social and political forces that view this system either as an appropriate and necessary force enacting restrictive laws in the interest of individual safety, family cohesion, and societal order, a system that arbitrarily applies laws to certain individuals or groups of individuals, or conversely, as an ameliorating one that can redress the loss of individual right to privacy and consensual sexual expression.

In the USA, sexual behaviors are primarily regulated by way of a legal process that restricts what it

defines as either harmful or socially unacceptable sexual conduct, both in private and public. More specifically, it is the general consensus that certain sexual behaviors merit prohibiting. In these instances, the laws are typically viewed as necessary and justifiable because they either provide parameters for sexual relations, as with regulations dictating the age of majority, or they prohibit coerced sexual contact between individuals, unsolicited sexual behavior on the part of the individual, and sexual exploitation, such as laws prohibiting rape, coerced nudity, sexual exploitation of an incapacitated individual, sexual contact between family members (incest), sexual contact between adults and children (pedophilia), sexual exposure to another individual (exhibitionist/indecent exposure), or the unsolicited sexual attention and/or pursuit by another individual (sexual harassment) (Leonard 1993). Other regulations are controversial because they restrict consensual sexual behavior and/or private sexual activity that is subjectively defined as socially offensive, inappropriate or immoral, as in sexual relations between: individuals married to other partners (laws prohibiting adultery); sexual relations between homosexual partners (laws targeting homosexual sodomy); oral or anal sex between heterosexual individuals, married or unmarried (those targeting heterosexual sodomy); and sexual relations exchanged for payment (prostitution). Resistance to these regulations and increasing controversy regarding their enactment has emerged from a growing awareness of one's right to privacy and the constitutional requirement that any restriction must be justified by compelling social need and benefit.

A related set of controversial regulations pertaining to sexuality but not directly involving sexual behaviors are those whose objective is to limit the production of and access to, sexual 'entertainment' and sexually explicit material. Operating in both the private and public sphere, these regulations focus on the right of the individual to access or 'consume' sexual entertainment and/or sexually explicit material in a variety of venues and mass cultural mediums—including various entertainment spheres, TV, movies, video, and via the Internet. Also restricting private, consensual behavior, these regulations range from censoring sexually explicit material in print to banishing live entertainment shows to peripheral locations in urban centers to confining sexually explicit content on cable TV to specific channels and hours (Communications Decency Act of 1995, Telecommunications Act of 1996).

2. Sexual Regulations that 'Protect'

Even regulations whose objective is to protect the individual from harm have a history of vague legal definitions and arbitrary enforcement. A case in point is the discrepancies in the penalties attached to

violations of sexual regulations and other criminal laws such as rape vs. murder, the disproportionate penalties given to minority vs. non-minority individuals charged with rape, and the increasingly subjective determination of incidents of sexual harassment.

More often, regulations pertaining to sexual behavior are applied to some but not others, and in many instances the determination is made on the basis of sexual orientation, sexual preferences or to behaviors considered outside of normative heterosexual activity. Individuals who fall into any one of these categories are then part of a 'sexual minority' with the unequal application of laws to these persons often questioned as to its constitutional legality.

3. Regulations that Restrict: A Normative View of Sexuality

In the legal enactment and interpretation of these regulations, rarely does the body of scientific and accurate knowledge about sexuality inform this process, a body of knowledge best represented by the repository of scholarly contributions on the subject from the social, biological and physical sciences, and from the humanities. As a result, knowledge of sexuality has had limited influence on developing a meaningful public and legislative discussion and coherent public policy regarding the regulation of sexual behaviors. Instead, the regulations of behaviors, especially those conducted in private between consenting adults, are typically based on a normative, procreative view of sexuality originating in English common law. This view approaches human sexuality as a potentially uncontrollable, inherently deviant and socially disruptive force in need of restraint, one rightfully subject to stigma, penalty, and criminalization, especially outside of marriage. In the USA it has resulted in very clear demarcations of appropriate and inappropriate sexual behavior in which fornication between married, heterosexual partners is considered a sexual behavior worthy of protection under the law. However, at the same time other sexual behaviors, such as those between homosexual partners or oral/anal sex between married or unmarried partners, are determined to be immoral and illegal. Some laws do not specify 'non-reproductive sex, but use laws against disorderly conduct, carnal abuse, lewdness, lascivious behavior, public indecency, loitering, and indecent exposure to prosecute crimes of oral and anal sex' (Franceour 1991a, p. 588).

Regulations based on a normative, procreative view of sexuality typically target adult behavior inherently private and consensual. As they become increasingly at odds with the sexual practices of the population and are enacted against an increasingly vocal sexual minority (more typically evident in the western world), their application and enactment becomes not

only controversial, but often arbitrary. Two examples stand out as particularly incongruent with actual sexual practices—the state sodomy laws and the military law regarding homosexual behavior among its members—and their application has become increasingly both controversial and unworkable. Forces supporting their enactment have confronted and continue to confront those seeking their revocation; a conflict played out in the courts, in the media, and in strategy meetings by opposing teams of legal experts and community activists. In both examples, the point of contention is whether their continued enforcement either ensures the ‘common good’ and protects someone from harm, or whether it promotes intolerance and violence against a sexual minority who are the target of their enactment. To illustrate this point, a brief history of these two regulations follows.

4. State Sodomy Laws

State sodomy laws criminalize certain sexual behaviors within heterosexual marriage and all sexual behaviors occurring between individuals of the same gender. More specifically, sodomy is often defined as deviate sexual intercourse, including genital–anal, genital–mouth, manual–anal, or manual–genital contact. Sodomy laws are any one of the many laws, which criminalize non-reproductive, non-commercial sexual behavior between adults in private. Since their enactment, restrictions against sodomy between unmarried individuals have typically served the purpose of dictating ‘normative heterosexuality.’ This is clear when reviewing their application, as sodomy laws are rarely enforced in the case of married, heterosexual couples engaging in oral or anal sex. Not only are they primarily enacted in those cases when sexual acts occur between members of the same gender, but they have been used to justify legal discrimination against homosexual people (Cain 1993, Copelon 1990) and to deny lesbians and gay men a range of other rights. This occurred when the laws were introduced as relevant evidence in adoption and custody cases and they have justified the taking of children away from gay and lesbian parents.

5. The History of Sodomy Law Repeal

The original 13 colonies continued the English tradition in which sodomy carried the penalty of death and every state adopted some form of a sodomy law as it joined the USA. By 1960, all states had some form of sodomy law in place, although their terms were often vague and they rarely stated clearly what the law prohibited. Throughout the long period of their enactment state by state, sodomy laws covered a variety of sexual behaviors, including oral sexual

relations between two men, two women or between a man and a woman; anal intercourse was included either under sodomy or buggery laws, without distinction as to whether the two persons are heterosexual or gay, married or unmarried (Franceour 1991a, p. 587).

It was not until 1961 that the first state repealed its law when the Illinois legislature revised their criminal code without prohibiting sodomy. Other states followed soon afterward, although this movement significantly tapered off in the 1970s and 1980s during which time a spate of ‘homosexual-only’ sodomy laws were introduced. These laws were quickly, and in a number of cases successfully, challenged for their constitutionality under equal protection and privacy rights at the state level. The brief movement toward repeal came to a standstill in the 1980s as it coincided with the early stages of the HIV/AIDS crisis and courts and lawmakers withdrew from [such] efforts in light of the potential criticism of promoting homosexual sex at a time that the HIV transmission modes were becoming known.

In 1986, the US Supreme Court upheld a Georgia anti-sodomy law in the case of *Bowers vs. Hardwick*, ruling that the constitution ‘does not protect homosexual relations between adults even in the privacy of their own homes’ claiming that (a) homosexual couples did not have a constitutional right to engage in acts of consensual sexual acts, and (b) oral and anal sex is a criminal offense ‘regardless of whether the parties who engage in it are married or unmarried, or of the same or different sexes’ (Franceour 1991a, p. 587). It was not until the 1990s that the HIV/AIDS transmission was more clearly understood and the changing political climate allowed for repeal legislation to move forward.

Repeal legislation has continued to slowly dismantle state sodomy laws one by one, and in the process often unveils the weak foundation upon which the pro-sodomy camp stands. A telling example is the recent case argued by the state of Montana in defending its ‘Deviate Sexual Conduct Law’ to the Montana Supreme Court. Here, the state argued that there is no right to privacy when it comes to sexual conduct; some of the justices, not persuaded by this argument asked, ‘what could citizens in the state expect to be covered under the privacy clause if it was not sexual activity?’ (ACLU Newswire, 1997).

At the beginning of the twenty-first century, 16 states plus Puerto Rico have sodomy laws; 11 states plus Puerto Rico have sodomy laws which apply to both heterosexuals and homosexuals, and five states have laws, which apply only to homosexuals. The growing movement for repeal has been successful in 34 states: 27 states have repealed the law through legislative action, and seven states have had their laws invalidated by courts. Sodomy laws are primarily being challenged in state courts on the basis that state constitutions can provide greater rights than does the

federal constitution. One brief in support of repeal states said: 'When considering a public morality justification for (sodomy laws), it is important for the Court to recall that prohibitions on interracial marriage, and the segregation of races generally, were values deeply imbedded in social morality and were values that invoked both God and nature.' But appeals to natural or theological ethics cannot constitutionally be used to legitimate laws whose sole function is to give effect to private citizens' prejudice or conviction (ACLU Newswire 2000).

6. Military Law Regulating Sexual Behavior

The most significant military law enacted to regulate the sexual behavior of military personnel in active service is the Pentagon policy of 'don't ask, don't tell' regarding the sexual behavior of homosexual service personnel. Enacted on July 19, 1993, this rule specifically forbade US military commanders from asking, pursuing, or harassing homosexual service members regarding their orientation. More significantly, it codified the same restrictions for homosexual service members as had existed before 1994 under Pentagon policy, requiring secrecy and celibacy as a condition of their military service, under penalty of being fired or imprisoned. It also openly barred homosexual individuals from serving in the military. Five years later, the 1998 discharge figures have shown a record high of 1145 individuals discharged under the current policy, a 92 percent increase in homosexual discharges since the law was implemented (Service members Legal Defense Network).

When challenged on the figures, Pentagon spokespersons indicated that the high number of discharges was due to an increase in 'voluntary' statements made by service members; what they failed to state was that the overwhelming majority of such statements were made to psychotherapists, in personal diaries, contained in anonymous online profiles, or, in direct violation of the law, constituted responses to direct questions about one's orientation. During that time, service members have indicated that they were coerced to make such statements due to hostile command climates characterized by fear, intimidation, and threat of assault, death and/or criminal prosecution. This was evidenced by the documented cases of physical assault and the brutal murder of Private Barry Winchell at Fort in Campbell, KY, which marked a turning point in the debate over the Pentagon's policy.

Coming under considerable pressure, the Pentagon initiated a series of measures in 1998 designed to 'improve' policy implementation by incorporating such measures as: prohibiting 'plea bargains' to obtain evidence of consensual homosexual conduct; protecting service members from investigations in order to encourage the reporting of anti-gay harassment or

lesbian-baiting without fear of reprisal; limiting the scope of investigations to prevent fishing expeditions; requiring commanders to consult with higher headquarters before initiating investigations into alleged homosexuality, and providing necessary and effective training on investigation parameters under the policy. Moreover, in December 1998, the Pentagon implemented a first-of-its-kind survey to determine the views of its service members about anti-gay discrimination evident in the military. It additionally required universal training on the 'don't ask, don't tell' policy of all military personnel with the intent to promote full compliance with military policy that anti-gay threats and harassment will not be tolerated and to educate all service members on policy's parameters.

In issuing the results of its survey, the Pentagon indicated that disparaging, anti-gay remarks, hostile attitudes and harassment are commonplace in the American military, with a majority of the service members stating that they believed that superiors and colleagues tolerated it. Such behavior has been confirmed by independent investigatory reports indicating those incidents of verbal abuse and physical assault in 1999 more than doubled from the year before. Placing the discharge figures in realistic terms, the report specified that on average, the policy has resulted in three persons per day being discharged after declaring they are gay, with a 142 percent increase in harassment reports for 2000 (SMDN, March 10, 2000).

The regulation of sexual behaviors is ultimately an issue of privacy, consent, and choice. In the US 'there is no shortage of opinions on what constitutes acceptable or unacceptable sexual behavior or what justifies regulating or prohibiting a particular behavior or activity' (Francoeur 1991a, p. 583). Yet the legal process of regulation enactment or revocation inevitably provides crucial opportunities for engaging public discourse, refining strategies to promote social change, and questioning ideological assumptions about sexuality and morality. A continuing challenge is to effectuate the appropriate balance between the common good and individual rights, privacy, and the social/personal cost exacted for laws that unnecessarily restrict sexual behavior.

See also: Child Abuse; Gay/Lesbian Movements; Heterosexism and Homophobia; Homosexuality and Psychiatry; Lesbians: Historical Perspectives; Lesbians: Social and Economic Situation; Prostitution; Queer Theory; Rape and Sexual Coercion; Rationality and Feminist Thought; Reproductive Rights in Affluent Nations; Reproductive Rights in Developing Nations; Sexual Harassment: Legal Perspectives; Sexual Harassment: Social and Psychological Issues; Sexual Orientation and the Law; Sexual Perversions (Paraphilias); Sexuality and Gender; Slavery as Social Institution; Transsexuality, Transvestism, and Transgender

Bibliography

- ACLU Newswire 1997 *Montana's Deviate Sex Law Challenged*. May 20, 1997
- ACLU Newswire 2000 *Louisiana May be Latest in Overwhelming Trend of State Courts Finding Sodomy Laws Unconstitutional*. April 10, 2000
- ACLU on Censorship: <http://www.aclu.org/library/pbp14.html>
- Bancroft J, Rosen R, Davis C, Weinstein D, Ruppel H 1991–95 The Society for the Scientific Study of Sexuality. *Annual Review of Sex Research* 1–6
- Cain P A 1993 Litigating for lesbians and gay rights: a legal history. *Virginia Law Review* 79: 1551–641
- Copelon R 1990 A crime not fit to be named: Sex, lies and the constitution. In: Kairys D (ed.) *The Politics of Law: A Progressive Critique*, rev. edn. Pantheon Books, NY, pp. 177–94
- D'Emilio J, Freedman E 1988 *Intimate Matters: A History of Sexuality in America*. Harper and Row, New York
- Federal Communications Commission, The Telecommunications Act of 1996: www.fcc.gov/telecom.html
- Francoeur R 1991a Sex and the law. In: Miller R (ed.) *Becoming a Sexual Person*. Macmillan Publishing, New York
- Francoeur R (ed.) 1991b *A Descriptive Dictionary and Atlas of Sexology*. Greenwood Press, New York
- Friedman S 1990 *Sex Law: A Legal Sourcebook on Critical Sexual Issues for the Non-lawyer*. McFarland and Co., Jefferson, NC
- Gagnon J H 1991 Sexual conduct and the life course. In: Turner C F, Gagnon J, Miller H, Moses L (eds.) *AIDS, Sexual Behavior and IV Drug Use*. National Academy Press, Washington, DC, pp. 500–53
- Lauman E, Gagnon J, Michael R, Michaels S 1994 *The Social Organization of Sexuality: Sexual Practices in the USA*. University of Chicago Press, Chicago
- Leonard A S (ed.) 1993 *Sexuality and the Law: An Encyclopedia of Major Law Cases*. Garland Publishing, New York
- Lichtenberg K E 1994 *A Research Guide to Human Sexuality*. Garland Publishing, Inc., New York
- Miller vs. California 1973 413 US 15
- NACLA 1998 Report on the Americas. *Sexual Politics in Latin America*, Vol. XXXI, January/February
- National Coalition Against Censorship: http://www.ncac.org/cen_news/cn71sex_censorship.html
- Parker R, Gagnon J (eds.) 1995 *Conceiving Sexuality*. Routledge, New York
- Portelli C 1998 *SIECUS Report*, Vol. 26, No. 5, p. 2
- Reno vs. ACLU 117 S. Ct. 2329, 138L. ed. 2nd 874 (1997)
- Roth vs. US 354 US 476 (1957)
- Sobell S, Wescott K, Benecke M H, Dixon Osburn C 2000 Don't ask, don't tell, don't pursue, don't harass. SLDN 6th Annual Report. March 9 2000. www.sldn.org/Templates/Law/Record.html?Record=22
- Sodomy Laws State by State: <http://www.sodomylaws.org/>; <http://www.aclu.org/issues/gay/sodomy.html>
- Service Members Legal Defense Network (SLDN) is comprised of a nationwide network of attorneys offers legal advice and assistance to serve US service members at risk of discrimination, persecution and dismissal on the basis of sexual orientation. This information has been culled from its 6th Annual Report.
- Sex and Censorship: Dangers to Minors and Others? Background paper of National Coalition Against Censorship

- Bob Summersgill, History of Sodomy Laws, in Creating Change: Public Policy, Civil Rights and Sexuality (to be provided.)
- 1998 *State of Sodomy Laws*. American Civil Liberties Union, New York
- 2001 (Spring) *Advocates Report: Vol 6, No 2* Sexuality Information and Education Council of the US, New York

D. Di Mauro

Regulation Theory in Geography

1. Introduction and General Argument

The regulation theory aims to understand the economic and extra-economic mechanisms which, relatively, stabilize capitalistic societies despite the fundamental contradictions generating crisis and disintegration. The regulation approach (RA) originated in French academic debates about the economic and social changes of the 1970s. Thirty years after World War II, a first serious structural crisis affected western societies indicating the end of the dream of a never-ending prosperity ('the golden age'). New forms of conflict configuration and new balances of power arose—without undermining the fundamental principles of capitalist societies altogether.

To understand the restructuring of postwar capitalism, neither the equilibrium models of orthodox economies nor the traditional concepts of Marxism seemed to be satisfactory. In this context an integrative approach has been developed—combining the analysis of political economy with that of civil society and the state, sometimes looking like a kind of institutional economies. But the term 'regulation' within RA may not be understood as an authoritative direction in order to govern the economy. Instead a process of self-regulating is addressed to conceptualize the evolution and temporary stability of a crisis-driven capitalist society. Four core concepts help one to understand processes of regulation (Boyer 1990, Boyer and Sallard 1995, Jessop 1997):

(a) an 'industrial paradigm' as a model governing the technical and social divisions of labor;

(b) an 'accumulation regime' as a complementary pattern of production and consumption which is reproducible over a long period;

(c) a 'mode of regulation' as an emergent ensemble of rules, norms, conventions, patterns of conduct, social networks, organizational forms and institutions which can stabilize an accumulation regime; and

(d) a 'model of development' as the result when an industrial paradigm, an accumulation regime and a mode of regulation complement each other sufficiently.

These four aspects of capitalist societies are differently configured in successive stages. As the most

discussed and inquired, the term 'Fordism' refers to a new production concept, introduced by Henry Ford between 1908 and 1925. The Fordist accumulation regime is characterized by a specific combination of mass production and mass consumption. Purchasing power has to be sustained by appropriate wages to ensure demand. This process depends on continuous productivity growth. The corresponding type of state is an intervening welfare state ('Keynesian welfare state') that aims to adjust demand to the supply-driven needs of mass production.

Fordism entered into structural crisis in the 1970s. The regulationists discussed several arguments to understand this phenomenon: differentiation of consumer behavior; rising wages, caused by the bargaining power of strong trade unions, but without a corresponding increase of productivity; the internationalization of economy, especially an increasing low cost production in developing countries; and a lack of a corresponding internationalization of politics and policy.

The stage after Fordism is characterized by productivity gains, ideologies and constraints of 'flexibilization.' Fordist mass production seems to disappear whereas flexible networks of decentralized production units, partly hierarchically controlled, are significant for the new stage using new technologies and new forms of production organization which enable 'flexible specialization.' At the same time there are strong attempts to deregulate the labor market while collective wage bargaining, rigid work relations, and intervening legal forms as institutionalized compromise between capital and labor have been serious constraints for the flexibilization of economy and society. Further, objectives of state acting have changed in order to strengthen the international competitiveness of the national economy. At the same time redistributive welfare rights have been subordinated to a productionist reordering of social policy.

This rather rough sketch of 'post-Fordism' is differentiated in the regulationist debate. Thus, it is not possible to identify a definite post-Fordist model of development. Some authors have distinguished two main directions of post-Fordist development (Leborgne and Lipietz 1994): 'neo-Taylorism' (deregulation and flexibilization without major constraints; for example, in the UK and USA) and 'Toyotism' (decentralization and flexibilization while keeping a high degree of social cohesion; for example, in Japan and Northern Europe).

Spatial patterns are varying between the different stages. The Fordist stage is characterized by strong spatial divisions of labor, homogenization, and core-periphery differentials on all spatial scales (for example, city-suburb, north-south divide). Post-Fordist spatial development is characterized by heterogeneity on all scales: different types of spatial structures (industrial districts, global cities, high-tech regions,

regions with functional mixtures, etc.) built up 'patchworks.' Definite trends of spatial structures are hard to recognize.

2. *Origins and Development of the Approach*

Today there are several RA directions. In the beginning at least two schools of regulationist thinking evolved in France; the 'Groupe de Recherche sur la Régulation de l'Economie Capitaliste' (GRREC) based in Grenoble, and the 'Centre d'étude Prospective d'Economie Mathématique Appliquées à la Planification' (CEPREMAP) in Paris. An important input to both approaches has been the work of François Perroux. Whereas the work of the GRREC is characterized by an elaboration within a more traditional Marxism, the Parisian RA, especially Michel Aglietta, Robert Boyer, and Alain Lipietz, have been basically involved in a renewal of Marxist thought and Althusserian structuralism with respect to evolutionary and institutional economics. Their contributions mark the beginning of the globalization of the RA (see reviews by Jessop 1997, MacLeod 1997). Since the mid-1970s and especially in the 1980s, their ideas spread, leading to new schools of innovation within the RA. A closer look at these discussion groups reveals a broad heterogeneity, bound together by common understandings of basic elements of radical political economy and the ability of social groups to act in order to transform society and economy.

Up to now, three periods of the RA with changing topics are observable: the first one may be labeled as the formative period that is characterized by theoretical and empirical work by the French regulationists. Among others, Aglietta's outline of the theory of regulation based on an analysis of the evolution and crisis of US capitalism in the twentieth century is usually seen as a well recognized starting point (published in French in 1976; in English in 1979, reprinted 2000). In general the formative period delivered the basic concepts of the RA and a deep understanding of Fordism including its crisis in the 1960s and 1970s.

A second period may be characterized by an intensive search and controversial debate around the term 'post Fordism.' Though often rejected as a misleading concept, post-Fordism served as a focus for those who looked after new economic and social principles of capitalist development in the late 1980s and 1990s. Examples are:

- (a) the search for the microfoundation of regulation using convention theory or Bourdieu's concept of habitus (e.g. Painter 1997, Jenson 1993);
- (b) debates about new organizational patterns of enterprises and the internationalization of production (e.g. Waringo 1998);
- (c) attempts to overcome structuralist shortcomings and a reflection of the structure-agency debate;

(d) comparative perspectives on modes of regulation in order to understand crisis and changes of the welfare state more profoundly (e.g., the concept of the workfare state by Jessop (1995) or the 'Wettbewerbsstaat' (competition state) by Hirsch 1995); and

(e) the question of scale related to the diminishing power of the nation state and the growing importance of substates (regional, urban) and supranational levels (international regimes) (e.g., Boyer and Hollingsworth 1997).

As outlined above, these attempts to understand the new (flexible) regime of accumulation and new (deconcentrated) mode of regulation indicate a thematic and regional expansion of the original French RA—a globalization of a theoretical program. However, the critical notes about the term post-Fordism—whether it is the guiding concept of a new development model as temporal spatial fix of capitalism evolved at the end of the twentieth century or it is an overloaded catchword demonstrating the inability of the RA to deliver a consistent explanation of changes—illustrate severe difficulties to open and use the RA for new topics and problems.

A regional perspective within RA has been introduced by different authors like Leborgne and Lipietz (1988), Esser and Hirsch (1994), or (Moulaert 1996). More recently Jessop and colleagues have announced what they call a 'third generation' of regulationists. They emphasize the ability of RA especially to answer questions of the changing form of the state and the relations between scale, space, and time. In their recent reviews (see Jessop 1997, MacLeod 1997), the RA is located between political science, geography, and sociology leaving the traditional macroeconomic perspectives of RA behind. However, in the new edition of Aglietta, the start-up analysis indicates an ongoing interest in macroeconomic regulation (Aglietta 2000).

We may conclude that the RA has experienced a remarkable differentiation in the last two decades of the twentieth century. This diffusion process did not favor a consistent body of thinking—one theory or school—but has led to a range of different concepts which today form the components of the RA. Though these concepts are far away from being parts of a whole, they are proofs of an ongoing interest of academics to organize their thinking around this approach. But there is also a tendency of the spatial and thematic diffusion of RA to a dissolution of the specialties of this approach into the broad stream of critical social science.

3. *Examples of Special Topics*

The diffusion process of the RA makes it difficult to focus on special topics. The following selection tries to outline those we think are probable pathways for future work, especially in geography. The common

question of the first body of literature that we would like to emphasize here is addressing problems and changes of scale-dependant institutional arrangements. One of the major changes between the period of Fordism and today is the dissolution of the former regulatory fix that is grounded on the temporal compromise between capital and labor (based on rising wages related to productivity growth), and on macroeconomic regulation (based on the defined territoriality and relatively autonomous power of the nation state). Owing to the loss of importance of the concept of the welfare state as the democratic form of the 'golden age' of Fordism, a debate about new forms of institutional arrangements has evolved. Deregulation, privatization, and the evolution of new political regimes at local as well as at the international level have resulted in the debate about the concept of governance, which includes new public, private interrelations beyond and within the old concept of nation state.

Within this context, the work of Jessop is of particular interest. The shift from—what he calls—'the Keynesian welfare national state' to an emergent 'Schumpeterian workfare postnational regime' forms the center of his analysis. Three changes are emphasized:

(a) New state activities evolve which should strengthen the competitiveness of state spaces by stimulating and promoting permanent innovation processes. Besides direct incentives to create national systems of technological innovation, new institutional forms, resources, and values are invented in order to create a 'Schumpeterian' mode of regulation.

(b) System competitiveness and labor market flexibility are accompanied by a subordination of social policy to imperatives of the new economic orientation. The term workfare is used to underline a general tendency of post-Fordism to cut back welfare expenditures and to redirect social assistance in order to promote individualistic responsibilities and risk acceptance.

(c) The state itself changes its form substantially. The 'hollowing out' of the national state apparatus favors subnational as well as supranational levels. Though the nation state remains important, new state activities emerge on different scales. This deconcentration of state power is accompanied by a shift 'from government to governance' which reflects a growing importance of semi-public and private actors, illustrated by new urban and international regimes. At the same time, trends towards de-statization or decentralization are counteracted by new state functions (metagovernance), addressing activities such as self-organization of partnerships, networks, and governance regimes.

The concept of the 'Schumpeterian workfare postnational regime' is a promising and far reaching concept of regulation under post-Fordist conditions. Yet, though well elaborated in relation towards RA

and state theory (especially Poulantzas position), the concept has to be grounded in empirical work. In particular, the reconfiguration of scales and state activities need further analysis in order to improve our understanding of new organizational principles beyond the nation state.

A second promising body of literature has drawn attention to local and regional modes of social regulation (Tickell and Peck 1992). There are approaches which address urban management and politics (Mayer 1994), and urban social movements (Hitz 1995), local labor markets (Peck 1996) and regional development in a postsocialist environment (Krätke et al. 1997). In particular, in reaction to challenges of urban regime theory, a new theory has set in (a) to conceptualize the relationships between accumulation and regulation on urban and regional scale; (b) to find adequate solutions for structure–agency problems which are inherent in the RA; and (c) to direct concrete research (Goodwin and Painter 1997). According to Goodwin and Painter, the concept of regulation should not be viewed as a static and stable, but as a process which is constituted by material and discursive practices. Because these practices are temporally and spatially embedded, there are many possible patterns of regulation as contingent combinations that might operate to support accumulation. This variety of modes of regulation is enlarged by the possibility that quality and intensity of regulation might differ due to spatial differentiation and uneven development. Therefore, the process of regulation is constituted geographically as result ‘of the contingent interaction of disparate practices’ (Goodwin and Painter 1997, p. 21).

To view regulation as a process, rather than as a fixed mode, and to conceptualize the process as contingent practices is an important contribution to reduce missing links in RA. But new problems arise when these suggestions are followed. For example, in applying the concept of Goodwin and Painter, it becomes difficult to decide what kind of social practices might be relevant in order to reveal processes of regulation. If it is assumed that regulation might be a result of intended and unintended consequences of social practices, we are trapped in the universe of possible practices. Though the introduction of the habitus concept of Bourdieu may serve as a micro-foundation of the RA and the identification of structures beyond the variety of practices (Painter 1997), the relationships between habitus and fields, on the one hand, and the concept of regulation as process, on the other, remain undefined. Therefore, if we give up the idea that the mode of regulation is like a development model beyond Bourdieu’s fields of political contest, e.g., as a spatially and temporally defined class compromises, which the subjects are reflecting, the term regulation receives a technical connotation of a self-organized system that creates order out of nowhere.

4. Future Directions

There are several missing links and open questions within RA often partly addressed as a general critique. Some of the important deficits are as follows.

The theoretical status of the terms Fordism and especially post-Fordism has become unclear. Some of the regulationist thinkers seem to give up the idea of development models serving as a background necessary for the evolution of structural coherence in time and space.

The ecological question is still neglected in the majority of contributions to the RA and a further concept is needed within RA, addressing an economic–ecological compromise. The evidence that the Western mode of production is self-neglecting in terms of sustainability of ecological resources and is partly elaborated in the work of Alain Lipietz, who has become one of the major political leaders of the red–green movement in France and Europe (Lipietz 1995).

The immanent functionalism of the RA has to be overcome in order to refuse reductionist explanations of modes of regulation for stabilizing capitalist development.

Though the listed deficits are already challenges for future work within RA, some directions are addressed here because they should receive major attention. The first perspective for future research reflects the growing integration of RA, evolutionary economics and institutional approaches in the social sciences. Within this broad field, the term regulation as a mode or as a process has to be worked out in a more detailed way in relation to state theory and the growing body of literature dealing with ‘governance’ approaches. Whereas RA usually puts more emphasis on structure, macroeconomic forces and large-scale institutional mediations, the post-Fordist variety requires differentiated concepts (a) of the enabling and restricting qualities of institutionalized systems of norms and rules, (b) of forms of leaderships and regimes on different scales, and (c) of forms of social network integration through conventions, coercion, and strategic behavior. As we have shown, there are some promising starting points in RA, but a comprehensive approach on regulation is missing. Further, it is rather unclear what a geographical perspective might look like. Interpretations like reterritorialization and politics of scale (Brenner 1999) as a dominant feature of current restructuring of capitalist society, as well as the notion that ‘all regulation is generated in temporally and spatially embedded social contexts’ (Goodwin and Painter 1997, p. 13), are still beyond a clear understanding of the role of space and scale.

A second promising field of future research evolves within the debate about poststructuralism and the societal paradigm of the mode of regulation (e.g., Scherrer 1995, Jenson 1993). Though the more recent literature is very aware of the problem of economic

determinism and the latent functionalism of the RA, it seems to be promising to review the concepts of the RA using discourse analysis (e.g., Laclau and Mouffe 1985). This methodological inquiry may not only serve as a reformulation of basic approaches within RA (like the refusal of a mode of regulation in favor of a process, or a more cautious usage of the term Fordism), but also may lead to (new) answers of the question of political practices deriving from RA analysis.

See also: Capitalism; Finance, Geography of; Globalization, Anthropology of; Globalization: Legal Aspects; Globalization: Political Aspects; International Business; International Law and Treaties; Regulation, Economic Theory of; Regulatory Agencies; World Trade Organization

Bibliography

- Aglietta M 2000 *A Theory of Capitalist Regulation*. Verso, London [first French edn. 1976; first English edn. 1979]
- Boyer R 1990 The regulation school: a critical introduction. Columbia University Press, New York
- Boyer R, Saillard Y 1995 *Théorie de la régulation, état des savoirs*. La Découverte, Paris
- Boyer R, Hollingsworth J R 1997 From national embeddedness to spatial and institutional nestedness. In: Boyer R, Hollingsworth J R (eds.) *Contemporary Capitalism. The Embeddedness of Institutions*. Cambridge University Press, Cambridge, UK, pp. 433–84
- Brenner N 1999 Globalisation as reterritorialisation: the re-scaling of urban governance in the European Union. *Urban Studies* 36: 431–51
- Esser J, Hirsch J 1994 The crisis of Fordism and the dimension of a 'post-Fordist' regional and urban structure. In: Amin A (ed.) *Post-Fordism. A Reader*. Blackwell, Oxford, UK
- Goodwin M, Painter J 1997 Concrete research, urban regimes, and regulation theory. In: Lauria M (ed.) *Reconstructing Urban Politics in a Global Economy*. Sage, London, pp. 13–29
- Hirsch J 1995 *Der nationale Wettbewerbsstaat: Staat, Demokratie, und Politik im globalen Kapitalismus*. ID-Archiv, Berlin
- Hitz H 1995 *Capitales Fatales: Urbanisierung und Politik in den Finanzmetropolen Frankfurt und Zürich*. Rotpunktverlag, Zürich
- Jenson J 1993 Ideas, spaces, and times in Canadian political economy. In: Jenson J, Mahon R, Brenfeld M A (eds.) *Production, Space, Identity: Political Economy Faces the 21st Century*. Canadian Scholar's Press, Toronto, ON, pp. 143–69
- Jessop B 1995 The regulation approach, governance, and post-Fordism. Alternative perspectives on economic and political change? *Economy and Society* 24: 307–33
- Jessop B 1997 Survey article: the regulation approach. *Journal of Political Philosophy* 5: 287–326
- Krätker S, Heeg S, Stein R 1997 *Regionen im Umbruch. Probleme der Regionalentwicklung an den Grenzen zwischen 'Ost' und 'West'*. Campus, Frankfurt/Main, Germany
- Laclau E, Mouffe C 1985 *Hegemony and Socialist Strategy: Towards a Radical Democratic Politics*. Verso, London
- Le Galès P 1998 Regulation and governance in European cities. *International Journal of Urban and Regional Research* 22: 482–506
- Leborgne D, Lipietz A 1988 New technologies, new modes of regulations: some spatial implications. *Society and Space* 6: 263–80
- Leborgne D, Lipietz A 1994 *Nach dem Fordismus*. In: Noller P, Prigge W, Ronneberger K (eds.) *Stadt-Welt. Über die Globalisierung städtischer Milieus*. Campus, Frankfurt am Main, Germany, New York, pp. 94–111
- Lipietz A 1995 *Green Hopes: The Future of Political Ecology*. Polity Press, Cambridge, UK
- MacLeod G 1997 Globalizing Parisian thought-waves: recent advances in the study of social regulation, politics, discourse, and space. *Progress in Human Geography* 21: 530–53
- Mayer M 1994 Post-Fordist city politics. In: Amin A (ed.) *Post-Fordism. A Reader*. Blackwell, Oxford, UK
- Moulaert F 1996 Rediscovering spatial inequality in Europe: Building blocks for an appropriate 'regulationist' analytical framework. *Environment and Planning D: Society and Space* 14: 155–80
- Painter J 1997 Regulation, regime, and practice in urban politics. In: Lauria M (ed.) *Reconstructing Urban Politics in a Global Economy*. Sage, London, pp. 122–43
- Peck J 1996 *Work-place. The Social Regulation of Labor Markets*. Guilford, London
- Scherrer C 1995 Eine diskursanalytische Kritik der Regulationstheorie. *Prokla* 25: S457–82
- Stoker G 1995 Regime theory and urban politics. In: Judge D, Stoker G, Wolman H (eds.) *Theories of Urban Politics*. Sage, London, pp. 54–71
- Tickell A, Peck J 1992 Accumulation regimes and the geographies of post-Fordism: missing links in regulationist theory. *Progress in Human Geography* 16: 190–218
- Waringo K 1998 *Die Internationalisierung der Produktion in der französischen Regulationstheorie*. Campus, Frankfurt/Main, Germany

R. Danielczyk and J. Ossenbrügge

Regulation: Working Conditions

Working conditions encompass the aspects of work ranging from the terms under which workers enter the employment relationship to the environment in which they perform their tasks. Historically, working conditions have varied widely across countries and across time within each country in terms of their quality—from the workers' vantage point. Working conditions have been subject to regulation to varied degrees, a situation which accounts for variability in their quality.

Working conditions are part of employment systems that have been shaped by actions, rules, and regulation emanating from several spheres of society. These include the legal, judiciary, administrative, and collective bargaining spheres as well as personnel policies and informal social norms. All of these spheres shape

the 'way of doing business' (Piore 1990) as well as working conditions for workers. This article reviews how working conditions are shaped by regulation. In industrialized societies, the regulation of working conditions has primarily affected people in waged or salaried employment, that is, those who are in a relationship, recognized legally as economically dependent, with an employer who usually controls the means of production.

1. Working Conditions, Definition

Working conditions consist of the following elements:

(a) *The terms of the employment relationship.* They include whether the employment relationship is durable or short-term, formal or informal. The terms of employment may be explicit or implicit.

(b) *Access to employment.* This access may be equitable or discriminatory. In this latter case, groups of workers may be excluded from employment or have differential access based on their race/ethnicity, gender, age, or other characteristic.

(c) *The workplace itself.* Work may be performed in a single site, multiple sites, or from the worker's home.

(d) *The wage level.*

(e) *Time at work and away from work.* This includes the daily, weekly, and yearly number of work hours as well as their scheduling. It also includes provisions for time away from work due to illness, vacation, or leave for family or medical reasons or for training.

(f) *The supervisory structure and mechanisms.* This structure may or may not include procedures for due process in disciplining, discharges, and layoffs. It may entail varied degrees of protection from supervisory arbitrariness.

(g) *Health and safety conditions in the workplace.* These include occupational health, that is, the ways in which the performance of work tasks affects worker physical health and well-being, and environmental health, that is, whether the environment in which the work is performed has negative effects on physical health.

(h) *Voice at work.* This encompasses access to worker representation and the shape of representation structures. These structures may include representative organizations such as labor/trade unions or worker associations as well as participatory structures such as workplace councils.

(i) *Social protection/social insurance benefits for the individual and their dependents that have been tied to waged employment and are administered through payroll systems.* These include, for example, health insurance and pension plans.

Some of the elements of working conditions listed above have a direct relationship to employment and the performance of work, e.g., health and safety conditions at work. Others have been historically tied to employment with a particular employer in the USA

and to labor market participation in Western European countries but do not have to be. For example, health insurance coverage can also be provided as a universal benefit regardless of work activity.

2. Systems of Employment and Working Conditions

National systems of employment, of which the US system is one example, can be viewed as a whole which provides degrees of regulation and of protection to workers and employers in a variety of ways (Jacoby 1982, Boyer 1986, Piore 1986). Systems of employment can be seen as a conjunction of 'spheres' which contribute to varied degrees toward shaping working conditions over time. These 'spheres' are legal, executive/administrative, judiciary, collective bargaining, and employer personnel policies to the extent that they embody 'social protection' norms, as well as informal social norms. Different countries have a varied mix of reliance upon these spheres to affect working conditions. Some rely on collective bargaining approaches whereas others favor the action of the legislative and executive spheres combined. The regulation of working conditions can therefore be achieved through these spheres in a different manner across countries as well as across time in one country.

3. The Purpose of Regulation on Working Conditions

The regulation of working conditions has several purposes. Most commonly, it establishes labor standards, that is, set minima for wage, hours, or health and safety conditions that employers are mandated to respect and that are enforced through prohibition and penalties. This corresponds to the narrowest role of regulation.

Regulation can permit the formation of structures for negotiation, mediation, or problem solving that contribute to shaping working conditions in varied ways and with diverse outcomes across industrial sectors and occupations (Weiler 1990). One example of such regulation is the federal and state statutes governing worker representation and collective bargaining (e.g., National Labor Relations Act of 1935).

4. The Regulation of Working Conditions

Most narrowly and conventionally, regulation is understood as emanating from government (national, state, local) through laws and statutes (legislative branch), as well as administrative rules (executive). In the USA, separate agencies have been established over

time to develop enforcement mechanisms. For example, the Wages and Hours Division in the US Department of Labor enforces the Fair Labor Standards Act of 1938 and related legislation on minimum wages and overtime pay for hourly workers. Through inspectors and administrative courts, violations are verified and penalties issued for responsible employers. Enforcement mechanisms and their effectiveness play a key role in the shaping of working conditions. Conversely, the work setting affects the effectiveness of enforcement. For example, if workplaces are small, scattered, and numerous, it is difficult for an enforcement agency that relies on visits by inspectors to monitor work conditions and enforce a specific law.

The court system is another feature of the regulation of working conditions. Courts contribute to enforcing existing laws and statutes. For example, an employer responsible for hazardous working conditions having resulted in worker death may be sued. Sometimes, in response to cases initiated by workers, courts issue decisions in areas of working conditions where government legislation and executive enforcement have not acted yet. In this case, court actions precede legislative action.

Working conditions are also affected by the relative bargaining power of workers and their organizations. In turn, this bargaining power is shaped partly by the provisions made in legislation for worker voice and representation. For example, the National Labor Relations Act of 1935 established a mechanism for representation for many US workers in the post-World War II years and also shaped the form that labor unions took. The relative bargaining power of labor unions and worker organizations affects the ability of workers to bargain for improvements, the thoroughness of enforcement of existing laws, and the ability to obtain new laws.

Lastly, social norms and customs along with worker relative bargaining power are reflected in personnel policies and employer practices; these also affect working conditions. In the USA, many aspects of working conditions are shaped by personnel policy rather than law or collective bargaining and thus vary greatly across employers.

In sum, working conditions are shaped by the totality of action in these respective spheres and by the interaction of actions from these spheres. For example, in the USA, provisions in collective bargaining agreements or in innovative personnel policies that affect an aspect of working conditions that is not the subject of government regulation may serve as a model for state or federal legislation.

In other national systems, such as France, national sectoral bargaining in lead industrial sectors on a particular aspect of working conditions often precedes national legislation. National sectoral bargaining is a negotiation of large employers and their federations with national unions that is sometimes mediated by government and whose outcome is extended to an

entire industry. In some countries, national sectoral bargaining on a particular issue (e.g., implementing shift work) is used sometimes as an alternative to government regulation.

5. The Long-standing Debate on the Regulation of Working Conditions

The role and benefits of the regulation of working conditions have been the subject of long-standing debate. One view sees regulation, particularly regulation on labor standards, as an externally imposed cost and/or as having a distorting effect on the operation of the labor market. This view stems from the neoclassical school of economics whose core perspective—presented here in stylized fashion—is that markets are self-regulating and that prices must be allowed to vary freely in order to be efficient allocators of human and capital resources. This view also tends to assume there is no significant conflict or power differential between workers as a group and employers as a group. Rather, workers and employers are individual decision makers who make choices unhampered by power differentials. Critics of federal and state minimum wage laws have argued that the laws push wages above the ‘market clearing’ level and thus cause unemployment. Other analysts have taken a more benign view of their impacts on firm decisions and employment levels. The research literature has centered on the empirical measurement of the relative magnitude of effects on employer costs and employment levels (Card and Krueger 1995, Houseman 1998).

Another view in this debate sees the regulation of working conditions and its effects as a special case in economic analysis because labor markets are deemed the least self-regulating among all markets. In this view, the regulation of labor markets cannot easily be analyzed with the conventional tools of economic analysis.

Consistent with this view, another perspective sees regulation as a reflection of customs and social norms and of power relations between employers and workers—all of which vary across countries. In this view, regulation may permit, hinder, add costs to, or facilitate greater efficiency in production. In other words, the effects of the regulation of working conditions are not conceived as unidirectional. They are expected to vary widely depending on the contents of regulation, the structure of particular industries, and their production organization. For example, minimal labor standard on wages, hours, or health and safety have been argued to discourage ‘sweatshops’ and favor other business strategies that are less onerous to workers (Piore 1990). In sweatshops, workers are unprotected because employers are insulated from the economic impact of variations in productivity among workers and over time thanks to the piece-rate system (workers are paid for measurable units of output). A

minimum wage rate which is high enough to create employer concern with productivity may compel a change in a firm's practice and working conditions; it may deter the use of low-productivity workers such as young children.

In the USA, recent debates on the regulation of working conditions have been colored by the general policy trend toward deregulation of markets. Laws on hours and scheduling, and those on worker representation structures have been questioned. Advocates of deregulation have argued that existing regulations fit an old order of production organization, and market structure and must be removed. Others have called for their replacement with different standards and representation systems that provide equivalent protections for workers but through new means that are more appropriate for contemporary forms of production and work organization.

6. Future Trends and Research Issues

6.1 New Forms of Employment

In the post-World War II period, the contents and enforcement of the regulation of working conditions in late industrialized countries have been constructed, albeit in different ways, around a certain notion of the employment relationship as a relatively long-term, stable attachment to a single employer with the work tasks being performed on the employer's work site. While there were always exceptions to this notion—for example, farm labor, construction, expert consulting, or domestic labor—the bulk of post-war regulatory attention to working conditions has been structured around this employment arrangement.

In the future, most observers expect less employer commitment to long-term employment relationships, less worker attachment to particular employers, and greater lateral mobility (across firms, industries, and possibly occupation). This is due to changes in the organization of production and employment structures within firms; increasingly, firms recruit workers from the outside labor market rather than rely on internal training and promotion. Career paths for workers will become less predictable and may span several companies and industries (Arthur and Rousseau 1996, Cappelli et al. 1997).

The manifestations of these transformations include the emergence and establishment of varied forms of employment which are explicitly short term (limited-duration hiring, temporary assignment), may be brokered by an intermediary (a for-profit or nonprofit temporary/staffing service), or may be the result of contracting arrangements between two businesses (e.g., workforce of a janitorial subcontractor), thus creating a triangulated employment relationship. Independent contracting has also

appeared—a situation where individual workers are treated as self-employed under the law, that is, not wage/salary employees, while they have a sole purchaser of their service and are economically dependent on the relationship.

Short-term employment relationships may render existing regulations and their enforcement ineffective; workers may not stay long enough in a particular job to bring a complaint about a violation of a labor standard. Triangulated employment relationships entail ambiguities as to employer responsibility for compliance with particular laws. Independent contracting means that there is no employment relationship and therefore no option for regulation under most labor and employment law. In practice, the financing and implementation of social protection benefits such as health insurance and pension through employment relationships are weakened if not rendered ineffectual.

As employment relationships change in these ways, the contents and means of the regulation of working conditions will likely change. The regulation of working conditions will require a reconceptualization and analysis.

Although employment relationships change in these ways, they still remain the primary means through which core social protection benefits continue to be provided. In the USA, which has relied on social protection benefits being provided by employers either voluntarily (fostered by tax deduction incentives) or through rare mandates, regulatory attention has focused on mechanisms to enable workers better to meet their responsibilities to family. For example, family leave—time off for caring for a child or sick relative—has been the subject of US legislation (in 1993), of continued activism at the state and federal level, and of innovative benefit policies by some large employers. These developments reflect the increased presence of members of dual earning couples as well as that of single parents (mostly female) in the workforce (Blau and Ehrenberg 1997).

6.2 The Regulation of Working Conditions and Globalization of Production

International economic integration has accelerated in recent years and is expected to continue. Its main features include the reduction or removal of trade tariffs and of restrictions on capital flows either by policy change or simply through the use of new communication technologies (which put distant production areas directly in competition). It has resulted in greater international capital mobility, the freer movement of goods, and the development of production processes ('production chains') that span national boundaries. This pattern holds particularly for manufacturing and, increasingly, for service activities (Whitman 1999).

In this context, the issue of the regulation of working conditions will take on new facets requiring further consideration and research. First, the adaptation and strategic responses of firms in industrialized countries to minimal labor standards were developed during the post-World War II period when mass production was in expansion, and domestic markets were relatively sheltered from foreign competition. Growing product competition from countries with dramatically lower wages and fewer regulations of working conditions, as well as the increased opportunities for firms to expand their production processes cross-nationally, are changing the parameters that affect working conditions for workers in both industrialized and developing countries. These developments challenge the ability of national and local governments to regulate working conditions in ways that provide for effective enforcement.

Second, in developing countries, the development of activities linked to the production processes of cross-national firms has given rise to forms of employment (e.g., electronic parts assembly at home) which are not subject to existing national regulation of working conditions, or which take place in settings and on terms in which monitoring and enforcement are difficult (see debates on the 'informal economy').

Additionally, the ability of countries enforce the regulation of working conditions through labor standards in those developing countries with low wages and low-quality working conditions is weakened by the ability of cross-national firms to shift production to another low-wage country with a workforce of similar skill level.

Three groups of issues will need to be considered in future research. First, the parameters of firms' choices will need to be documented in the context of particular countries and for specific activities, contrasting capital-intensive and labor-intensive manufacturing as well as comparing manufacturing with service activities.

Second, future research on regulation will increasingly examine the sphere of international organizations (e.g., World Trade Organization, Organization for Economic Cooperation and Development, International Labor Organization), the process of international negotiation of conventions on working conditions, and the impacts of these conventions, first, on national regulation efforts and, second, on firm strategies and practices regarding working conditions.

Lastly, for policy analysis and discussion, the core question will be what *should* be the complex of regulation of working conditions that is most compatible with new forms of cross-national production, the productive use of new production and communication technologies, and the well-being of workers in industrialized and developing countries?

See also: Environmental Health and Safety: Social Aspects; Industrial Policy; Industrial Relations and

Collective Bargaining; Industrial Sociology; Injuries and Accidents: Psychosocial Aspects; Safety, Economics of; Social Insurance: Legal Aspects; Work, History of; Work, Sociology of; Workplace Safety and Health

Bibliography

- Arthur M B, Rousseau D M (eds.) 1996 *The Boundaryless Career: A New Employment Principle for a New Organizational Era*. Oxford University Press, New York
- Blau F D, Ehrenberg R G 1997 *Gender and Family Issues in the Workplace*. Russell Sage Foundation, New York
- Boyer R 1986 *La Flexibilité du Travail en Europe*. Editions La Découverte, Paris
- Cappelli P, Bassi L, Katz H C, Knoke D, Osterman P, Useem M 1997 *Change at Work*. Oxford University Press, New York
- Card D, Krueger A B 1995 *Myth and Measurement: The New Economics of the Minimum Wage*. Princeton University Press, Princeton, NJ
- Doeringer P B, Piore M J 1971 *Internal Labor Markets and Manpower Analysis*. D C Heath, Lexington, MA
- Houseman S 1998 The effects of employer mandates. In: Freeman R B, Gottschalk P (eds.) *Generating Jobs: How to Increase Demand for Less-skilled Workers*. Russell Sage Foundation, New York
- Jacoby S M 1982 The duration of indefinite employment contracts in the United States and England: An historical approach. *Comparative Labor Law* 5: 85–128
- Jacoby S M 1997 *Modern Manors: Welfare Capitalism Since the New Deal*. Princeton University Press, Princeton, NJ
- Piore M J 1986 Perspectives on labor market flexibility. *Industrial Relations* 25: 146–66
- Piore M J 1990 Labor standards and business strategies. In: Herzberg S, Perez-Lopez J F (eds.) *Labor Standards and Development in the Global Economy*. Government Printing Office, Washington, DC pp. 35–49
- Standing G 1999 *Global Labor Flexibility: Seeking Distributive Justice*. St. Martin's Press, New York
- Weiler P C 1990 *Governing the Workplace: The Future of Labor and Employment Law*. Harvard University Press, Cambridge, MA
- Whitman M N 1999 *New World, New Rules: The Changing World of the American Corporation*. Harvard Business School Press, Boston, MA

F. J. Carré

Regulatory Agencies

Broadly defined, regulation includes any governmental effort to control behavior by other entities, including business firms, subordinate levels of government, or individuals. Sovereign governments may cooperate under regimes of mutually-recognized restraint, the rules or decisions of which might also be characterized as regulations. Campaigns for elective office may likewise be subject to controls on spending

and advertising. However, in most public discourse and social science inquiry, the term refers more narrowly to formal government programs (or to specific implementing decisions) intended to constrain economic activity in accord with some public-interest justification.

Four primary concerns of regulation in this sense may be distinguished (Breyer 1982, Stone 1982). First, regulation may control (as in the case of occupational licensing) market entry. Second, it may (as in the case of electricity utilities and other 'natural monopolies') establish directly the prices that firms can charge customers. Third, it may limit the competitive practices available to firms, especially as regards price-setting (e.g., the law of antitrust), advertising, and disclosure. Fourth, regulation may aim to improve the quality of specific work practices or products, often in the interest of protecting human health and safety, or the environment. These activities are sometimes subsumed under two more general headings—'economic' (i.e., price, entry, and competitive practices), and 'social' (meaning generally health, safety, and environmental) regulation.

Social scientists have posed three primary empirical questions about regulatory programs. First, what forces lead to the creation of such programs? Second, and most commonly, they have asked what constellation of factors influences program behavior? Third, under what conditions may regulatory programs be significantly scaled-back or terminated? Normative questions also proliferate. Some varieties of regulation, especially in the United States, have been prone to intense controversy. Both business representatives and policy-focused academics have pursued long-running debates about the merits of various regulatory programs, the quality of their specific decisions, and possible institutional alternatives to the status quo.

1. *Empirical Inquiry*

Students of regulation have long found the formally-stated public goals of regulation inadequate to explain program creation and maintenance. From the 1950s to the 1970s, working within disciplines as diverse as economics, political science, and historical analysis, commentators viewed regulatory programs frequently as being prone to 'capture' by regulated interests (Huntington 1952, Bernstein 1955, Kolko 1965, Stigler 1971). By some accounts, regulation was said to have arisen because a diffuse majority favoring regulation (and motivated by a 'public interest' rationale, such as the elimination of predatory pricing or consumer fraud) succeeded, at least briefly, in obtaining legislative consensus, often in the wake of well-publicized scandal. Over time, however, according to this 'life cycle theory' of regulation posited by Bernstein (1955), waning public attention and media scrutiny was said

to have allowed resourceful regulated interests, whose attention to regulatory program activity did not wane, to exercise a strong (even dominant) influence over the program, thus thwarting its original intent. A different version of capture analysis suggested that such programs were sometimes created, at least partially, at the behest of economic interests. Agencies might thus, in effect, be 'born captured' due to business incentives to seek stable and reliably profitable operating environments. The capture model of regulation is in some respects akin to a traditional political science metaphor, widely employed during the same period, that described the politics of narrowly distributive public policies in terms of mutually-reinforcing 'iron triangles' or subgovernments composed of interest groups, legislative committees, and agencies (Freeman 1965). The similarity is striking, even though an analytic scheme widely appropriated among political scientists (Lowi 1964) highlighted a distinction between 'distributive' and 'regulatory' politics.

During the 1970s and thereafter, however, scholars challenged increasingly the descriptive accuracy of both the capture theory and the 'iron triangle' metaphor. In the United States especially, this challenge stemmed importantly from the creation of regulatory programs (e.g., in environmental, consumer, and worker protection) where lack of media scrutiny, public indifference, and business dominance plainly did not apply over the long run (Wilson 1980). Indeed, the creation of these programs often represented sustained triumphs by idea-driven policy entrepreneurs over business interests that had to adjust to the presence of long-term political competitors. In one prominent instance, the regulation of nuclear power generation, a policy milieu changed dramatically from one to which the term 'capture' could arguably have applied in the 1950s. By 1980, the politics of nuclear power in the United States had evolved into a far more complicated set of power relations, in which environmental advocates and citizen groups competed effectively with the nuclear industry (Temples 1980).

Throughout the 1980s, political science commented increasingly about the inadequacy of broad generalizations regarding regulatory politics as a whole. The concentration of costs and benefits prevailing in regulation were found to differ importantly among agencies, and across issues within a single agency (Wilson 1980). The politics of broadcast regulation, pursued within a markedly vague standard of 'public interest, convenience, and necessity,' clearly varied between technical industry structure matters and issues of program content (Krasnow et al. 1982). Within public utility regulation, not only were issues determined to differ in complexity and conflict, with significant implications for the power of different kinds of advocates (Gormley 1983), but the characteristics of commission personnel were also found to vary in ways important to generating distinctive regulatory outputs (Berry 1984).

Regulatory policy also became the staging ground for scholarly argument over the relative influence of various political institutions. A conventional position among journalists and political scientists had for many years held that Congress does little to oversee bureaucratic behavior, largely because of the modest political incentives said to prevail (Scher 1963). In arguing against what he called 'interest-group liberalism,' Lowi (1969) bemoaned the vagueness that had crept into legislative delegations for regulatory and other programs.

A later principal-agent perspective on the legislative-executive relationship argued for the existence of a more robust oversight mechanism grounded in a rational legislative approach to structure and procedure that appeared only superficially to be weak (McCubbins and Schwartz 1984). Careful sifting of agency case histories determined that widely-trumpeted accusations of 'runaway bureaucracy' often directed at regulation were misplaced (Foreman 1988).

Because regulation is politically contentious and generates substantial numbers of comparable implementation and enforcement decisions over time, political scientists (often working within the tradition of rational-choice models and principal-agent theory) have gravitated to regulation as a window on relative institutional influence more generally. Some analysts have emphasized the power of legislative principals (activated by the material stakes of affected parties) over agency behavior (Weingast and Moran 1983, Weingast 1984). Meanwhile, Moe (1982), examining time series data drawn from three US regulatory commissions, concluded that presidential power deserved emphasis. On the other hand, some analysts have looked inside regulatory bureaucracies, to the professional worldviews and incentives of staff members, as key drivers of agency behavior. In antitrust enforcement, economists and attorneys are among those professionals whose views have been found to have a profound impact on agencies, both in case studies (Katzmann 1980, Weaver 1977) and in quantitative analysis (Eisner and Meier 1990). On the other hand, the theory of overhead democracy embraced by Wood and Anderson (1993) ascribed variation in both the level and substance of antitrust regulation to a complex set of institutional forces; changing presidential administrations, congressional committee preferences, and court decisions all helped to mold antitrust activity.

Deregulation is rarer, and politically more problematic, than regulatory initiatives. As regards social regulation, such goals as a clean environment, and safe workplaces and pharmaceutical products, are popular with the mass of the public. In economic regulation, organized stakeholders would be expected to resist unfavorable change (i.e., alteration of profitable regulated environments). Given such forces, it is worth emphasizing that deregulation has occurred on occasion, and asking how this has come about. Political

scientists Derthick and Quirk (1985) point to the power of both ideas and effective leadership. They observe that, by the late 1970s, economists had mounted a sustained, coherent, and nearly unanimous attack on the underlying justifications for price and entry regulation in the transportation sector. This pro-competition critique gained added political impetus from the increasing visibility of economists in national policy debates, from the infiltration of economists into leadership of the regulatory apparatus, and from the transformation of deregulation into an anti-inflation and consumerist cause that both liberals and conservatives could support. Strong presidential and legislative leadership also proved to be pivotal.

2. Normative Assessments

Nearly all regulatory programs will find supporters and detractors in the social sciences. Unlike price and entry regulation, which regularly has found advocates of abolition in the economics profession, other types of regulation enjoy strong academic support on principle (e.g., recognition of 'market failure') though often not in specific detail.

Leading the charge for varieties of regulatory reform have been economists and economically-oriented analysts associated with policy research institutions or 'think tanks.' Criticisms of environmental policy stress varieties of putative inefficiency or irrationality. Costs are said to exceed benefits, to be higher than necessary to achieve a given level of environmental benefit, and perhaps to be masked altogether by inadequate economic analysis (Kraft 1996, pp. 165-6, Litan and Nordhaus 1983, Lave 1981). Inefficiency is sometimes argued to have been caused by the logic of political compromise (Ackerman and Hassler 1981), or by unwillingness to impose overtly visible losses on large numbers of voters (Nivola and Crandall 1995). Inadequate monitoring has long been said to bedevil assessment of pollution control efficacy (Crandall 1983, Davies and Mazurek 1998).

A broader, efficiency-based critique encompasses additional concerns, including the discordant, meandering, and arguably even panic-stricken character of regulatory politics and policy (Breyer 1993). Critics point to a chronic inability to set risk priorities, to confront risk-benefit and risk-risk tradeoffs (Graham and Wiener 1995), and to reconsider regulatory statutes in the light of such concerns.

3. International Perspectives

All developed nations regulate, but they vary substantially in both regulatory structure and regulatory political culture. The United States is often portrayed as having a markedly more adversarial and legalistic system than other countries. One study concluded that

while 'regulatory power has been directed at very similar ends [i.e., control of carcinogens]' in Britain, France, Germany, and the United States, 'the openness, complexity and divisiveness of American decision making have no parallel in Europe' (Brickman et al. 1985, pp. 301–2). It is not clear, however, that the American style of regulation is necessarily more stringent than that of other nations. Conceding the more open regulatory style of the United States in comparison with Canada, Harrison and Hoberg (1994) nevertheless do not find the latter consistently less stringent than the former. Although institutions matter, their effects are 'contingent on a number of other variables that influence policy, such as the balance of power between competing interest groups and the interests of the party in power' (p. 181). Nevertheless there can be little doubt of the generally greater levels of hierarchy, and the more tightly-woven informal networks of personal relations, that characterize regulatory regimes in other advanced democracies in comparison with the United States (Badaracco 1985).

See also: Business and Society: Social Accounting; Policy History: State and Economy; Policy Process: Business Participation; Public Interest; Regulation and Administration; Regulation, Economic Theory of; Regulation: Empirical Analysis; Regulation: Working Conditions

Bibliography

- Ackerman B A, Hassler W T 1981 *Clean Coal/Dirty Air: Or How the Clean Air Act Became a Multibillion Dollar Bail-out for High-sulfur Coal Producers and What Should be Done About It*. Yale University Press, New Haven, CT
- Badaracco J L 1985 *Loading the Dice: A Five-country Study of Vinyl Chloride Regulation*. Harvard Business School Press, Boston, MA
- Bernstein M H 1955 *Regulating Business by Independent Commission*. Princeton University Press, Princeton, NJ
- Berry W D 1984 An alternative to the capture theory of regulation: The case of state public utility commissions. *American Journal of Political Science* **28**: 524–58
- Breyer S 1982 *Regulation and its Reform*. Harvard University Press, Cambridge, MA
- Breyer S 1993 *Breaking the Vicious Circle: Toward Effective Risk Regulation*. Harvard University Press, Cambridge, MA
- Brickman R, Jasanoff S, Ilgen T 1985 *Controlling Chemicals: The Politics of Regulation in Europe and the United States*. Cornell University Press, Ithaca, NY
- Crandall R W 1983 *Controlling Industrial Pollution: The Economics and Politics of Clean Air*. Brookings Institution, Washington, DC
- Davies J C, Mazurek J 1998 *Pollution Control in the United States: Evaluating the System*. Resources for the future, Washington, DC
- Derthick M, Quirk P J 1985 *The Politics of Deregulation*. Brookings Institution, Washington, DC
- Eisner M A, Meier K J 1990 Presidential control versus bureaucratic power: Explaining the Reagan revolution in anti-trust. *American Journal of Political Science* **34**: 269–87
- Foreman C H 1988 *Signals from the Hill: Congressional Oversight and the Challenge of Social Regulation*. Yale University Press, New Haven, CT
- Freeman J L 1965 *The Political Process: Executive Bureau—Legislative Committee Relations* (rev. edn.). Random House, New York
- Gormley W T 1983 Policy, politics, and public utility regulation. *American Journal of Political Science* **27**: 86–105
- Graham J G, Wiener J B (eds.) 1995 *Risk vs. Risk: Tradeoffs in Protecting Health and the Environment*. Harvard University Press, Cambridge, MA
- Harrison K, Hoberg G 1994 *Risk, Science, and Politics: Regulating Toxic Substances in Canada and the United States*. McGill-Queen's University Press, Montreal, Canada
- Huntington S P 1952 The Marasmus of the ICC. *Yale Law Journal* **61**: 467–509
- Katzmann R A 1980 *Regulatory Bureaucracy: The FTC and Antitrust Policy*. MIT Press, Cambridge, MA
- Kolko G 1965 *Railroads and Regulation, 1877–1918*. Norton, New York
- Kraft M E 1996 *Environmental Policy and Politics: Toward the Twenty-first Century*. HarperCollins, New York
- Krasnow E, Longley L, Terry H 1982 *The Politics of Broadcast Regulation*. St Martin's Press, New York
- Lave L B 1981 *The Strategy of Social Regulation: Decision Frameworks for Policy*. Brookings Institution, Washington, DC
- Litan R E, Nordhaus W D 1983 *Reforming Federal Regulation*. Yale University Press, New Haven, CT
- Lowi T J 1964 American business, public policy, case studies and political theory. *World Politics* **16**: 677–715
- Lowi T J 1969 *The End of Liberalism: Ideology, Policy and the Crisis of Public Authority*. Norton, New York
- McCubbins M, Schwartz T 1984 Congressional oversight overlooked: Police patrols vs. fire alarms. *American Journal of Political Science* **28**: 165–79
- Moe T M 1982 Regulatory performance and presidential administration. *American Journal of Political Science* **26**: 197–224
- Nivola P S, Crandall R W 1995 *The Extra Mile: Rethinking Energy Policy for Automotive Transportation*. Brookings Institution, Washington, DC
- Scher S 1963 Conditions for legislative control. *Journal of Politics* **25**: 526–51
- Stigler G 1971 The theory of economic regulation. *Bell Journal of Economics and Management Science* **2**: 3–21
- Stone A 1982 *Regulation and Its Alternatives*. Congressional Quarterly Press, Washington, DC
- Temples J R 1980 The politics of nuclear power: A sub-government in transition. *Political Science Quarterly* **95**: 239–60
- Weaver S 1977 *Decision to Prosecute: Organization and Public Policy in the Antitrust Division*. MIT Press, Cambridge, MA
- Weingast B R 1984 The congressional bureaucratic system: A principal-agent perspective (with applications to the SEC). *Public Choice* **44**: 147–91
- Weingast B R, Moran M 1983 Bureaucratic discretion of congressional control: Regulatory agency policy making at the FTC. *Journal of Political Economy* **91**: 765–800
- Wilson J Q 1980 (ed.) *The Politics of Regulation*. Basic Books, New York
- Wood B D, Anderson J E 1993 The politics of US antitrust regulation. *American Journal of Political Science* **37**: 1–39

C. H. Foreman Jr.

Copyright © 2001 Elsevier Science Ltd. 12985
All rights reserved.

Rehabilitation (Cardiac): Psychosocial Aspects

The primary aim of rehabilitation of patients with physical conditions such as skeletal, spinal, or brain injuries, or following a cardiac event such as a myocardial infarction, is to enable the patient to return to as near their level of pre-morbid functioning as their physical condition allows. Since much heart disease is a product of an unhealthy life-style, an important secondary aim in cardiac rehabilitation is to enable the patient to change their behavior to reduce the risk of further arterial damage and recurrent cardiac events.

Cardiovascular disease is a major problem in most industrialized countries. Currently in the UK approximately 300,000 people per year have myocardial infarction, of whom 50 percent survive for at least 28 days. There are also large numbers with the disabling symptoms of coronary disease such as the 1.4 million individuals in Britain with angina. While there has been progress in the prevention of coronary artery disease in most countries with a high incidence of heart disease, an aging population and advances in medical treatments mean the number of people with coronary disease and its consequences is in fact increasing. A sizeable proportion of these people would benefit from cardiac rehabilitation. While prolonged inpatient rehabilitation—often in sanatoriums in country environments thought to be healthy—has been common in some countries for many years, the development of intensive outpatient programs is a product of the last quarter of the twentieth century. Founded with great enthusiasm, the early results from cardiac rehabilitation programs were disappointing but developments in therapeutic and data analytic techniques have led to a much more positive picture emerging since the mid-1990s. Many now hold that that comprehensive cardiac rehabilitation can improve the quality of life and physical health of patients with coronary heart disease and most sophisticated systems of healthcare aim to offer rehabilitation to patients suffering the effects of cardiovascular disease (see *Chronic Illness: Quality of Life*).

1. Main Features of Cardiac Rehabilitation

The main group of patients who receive cardiac rehabilitation comprises survivors of acute myocardial infarction, although patients who have undergone coronary bypass surgery or angioplasty may also receive it. Since the latter are often elective procedures, rather than the unpredicted disaster of a heart attack, the patients' problems may differ and less is known about the most appropriate forms of rehabilitation for them. The first cardiac rehabilitation programs were primarily based around exercise regimes and this remains the core of many programs. However with the

increasing understanding of the psychological effects of myocardial infarction and the psychological and behavioral determinants of coronary heart disease (Krantz and Lundgren 1998), programs have widened to include many psychosocial components. There is no agreed standard program of cardiac rehabilitation, and programs vary widely in practice, but a modal program would be offered in an outpatient setting, starting six to eight weeks after discharge from hospital, would run for two or three months and would include regular exercise, an educational component about life-style change, and some procedures—such as relaxation training—to reduce distress. Exercise is usually conducted in specialist hospital gymnasiums, often with careful monitoring of the heart rate to diminish the risk to the patient. Rehabilitation programs are usually offered to middle-aged patients but might not be offered to the elderly, although they are the main sufferers from heart disease. Uptake of rehabilitation can be poor, especially in older patients. Such rehabilitation programs are expensive and difficult to deliver to the infirm, the socially disadvantaged, or those who live far from a specialist centre. Fortunately, effective versions of some aspects of rehabilitation can also be delivered by manual and supplementary telephone contact (Lewin et al. 1992), or by brief educational or counselling interventions while the patient is still in hospital (Johnston et al. 1999).

2. The Effects of Cardiac Rehabilitation

2.1 Psychosocial Outcomes

It would appear self-evident that patients will be distressed immediately after an unexpected, painful, and life-threatening event such as an acute myocardial infarction, and that this distress will continue for some time. While this is true of perhaps 15 to 30 percent of patients, a number of studies have shown that anxiety and depression are only slightly raised in most patients while still in hospital, but rise several weeks later and may remain elevated for at least a year (Johnston et al. 1999). Perhaps as significant is the finding that anxiety and depression are very high in the patients' partners while the patients are in hospital and these remain high—higher than that of the patients—for at least the subsequent year. There are many studies using rehabilitative techniques designed to reduce the distress associated with an infarction. Issues covered in such interventions will typically include an explanation of the attack and the healing process, including medical treatments and investigations, information on resumption of activities, and material on the emotional effects of a heart attack and the management of stress. Many recent studies of such interventions are positive and until recently there was a clear consensus, which had been confirmed by meta-analysis (Linden et al.

1996), that anxiety and distress could be reduced by appropriate rehabilitation programs, including programs that focused primarily on exercise (Kugler et al. 1994). However, an analysis by Dusseldorp et al. (1999) failed to confirm this finding. This is almost certainly because of the inclusion of two major studies with substantial samples that failed to find a reliable effect of psychosocial rehabilitation on distress (Jones and West 1996, Frasure-Smith et al. 1997). These studies are aberrant, and while their impressive size (which came about because they were primarily designed to detect effects on reinfarction rate) means that they have a substantial effect on summary analyses, few in the field accept that distress cannot be reduced after a myocardial infarction. Other studies have shown highly significant effects of simple interventions (Lewin et al. 1992, Johnston et al. 1999) on distress and on the resumption of the normal activities of daily life. Johnston et al. also showed that anxiety and depression was greatly reduced in the partners of patients if they were included in the rehabilitation program.

2.2 Risk Factor Reduction

A substantial proportion of heart disease is predicted by a few well-recognized and modifiable risk factors, all of which relate to some extent to behavior. These are elevated blood pressure, some lipids (including cholesterol), and cigarette smoking. Dusseldorp et al. (1999) have provided the most complete account of the effects of psychoeducational programs on the reduction of these conventional risk factors. All were reliably reduced by the interventions.

As well as these well-accepted risk factors there are a number of psychological risk factors that, although not necessarily accepted by the whole medical community, have substantial epidemiological support. Currently the most prominent of these are Type A behavior and depression (see Krantz and Lundgren 1998), and there is increasing interest in social support. Type A behavior, a pattern of behavior characterized by competitiveness, time pressure, and hostility, has been recognized as a risk factor for myocardial infarction for some time. However, possibly because of continuing controversy over which component of this complex behavior pattern is actually harmful, there have been few studies that both modified the behavior and were large enough to detect any effect on physical health. By far the largest of such studies is the Recurrent Coronary Prevention Project (Friedman et al. 1986) in which nearly 900 patients were randomized to either a prolonged program of Type A modification using cognitive behavioral techniques (see *Cognitive Behavioral Therapy*) or an alternative treatment emphasizing cardiological care and information. The behavioral intervention led to substantial reductions in Type A behavior and in recurrent myocardial infarction. There have been a number of successful

small-scale replications of this study, but it does not seem to have been incorporated into clinical practice to a significant extent. Future studies may concentrate more on hostility than the complete range of Type A behavior, since it is generally held that hostility is the component of Type A most reliably related to cardiovascular disease. Depression has emerged much more recently than Type A as a risk factor for heart disease (and indeed other physical conditions) and appears to be finding widespread acceptance. While studies are ongoing, there are no completed studies of the effects of treating depression on recurrent heart disease. Frasure-Smith, one of the leading advocates of depression as a risk factor in recurrent heart disease, has described interventions for related states with disappointingly mixed effects. In an early study she found that distress (which could include both anxiety and depression) could be successfully treated in patients after a myocardial infarction and that this was associated with reductions in morbidity and mortality over the subsequent seven years. However in a substantial replication (Frasure-Smith et al. 1997) she was unable to either reduce distress or affect the recurrence rate.

3. Morbidity and Mortality

Few studies of rehabilitation have large enough samples to demonstrate reliable changes in morbidity or mortality, but meta-analytic techniques enable independent studies to be combined. Oldridge et al. (1988), in a highly influential analysis of exercise and life-style change programs, showed a reduction in mortality, but not reinfarction rate, following rehabilitation. Subsequent meta-analyses incorporating more studies of exercise have largely confirmed these findings and also shown effects on reinfarction, as have the meta-analyses of life-style change and stress management programs (Dusseldorp et al. 1999). They show that such programs were associated with a 34 percent reduction in mortality and a 29 percent reduction in reinfarction rate. The findings were most persuasive in studies in which there was evidence of change in life-style-related risk factors such as blood pressure, cholesterol, and cigarette smoking.

4. The Future

The outstanding issues for the future are effectiveness, efficiency, and access. While it is clear that effective programs exist, the results from ineffective trials are still reported and the critical components have yet to be isolated. In addition we can anticipate increasing efforts to incorporate modifications of the newer psychological risk factors such as depression, hostility and, very probably, social support. Many programs are expensive to implement, they require gymnasiums, physiotherapists and other health professionals, and

time—all scarce resources. Increasingly complex programs will have to be delivered in ever more efficient ways. This will surely involve close targeting of interventions to patients' particular problems and risks as well as more efficient methods of delivery such as patient manuals, both printed and computerized, and the use of the internet. Perhaps the greatest advances need to be made in increasing access. In many countries with otherwise sophisticated healthcare systems, too few people are offered cardiac rehabilitation, and many find it difficult to accept because of factors such as other commitments, poverty, and, no doubt, a failure to realize or accept the possible benefits. In England and Wales 40 percent of centers restrict rehabilitation to those aged below 60, and when it is offered to the elderly it is not geared to their special needs. The situation may be a little different in the USA, where Richardson et al. (2000) report an increase in the uptake of rehabilitation in those aged 65 and over, from 28.3 percent in 1986 to 52.1 percent in 1996 for one particular service. Cardiac rehabilitation in the elderly appears to be safe and of value (Foreman and Farquhar 2000), but Oldridge (1998) has pointed to the lack of good, controlled studies. Most studies, like most rehabilitation programs, have focused on the middle-aged. We can expect to see in the near future a change in both access and a strengthening of the scientific basis of cardiac rehabilitation for the elderly.

See also: Cardiovascular Conditioning; Neural Substrates; Chronic Illness: Quality of Life; Coronary Heart Disease (CHD), Coping with; Coronary Heart Disease (CHD): Psychosocial Aspects; Coronary-prone Behavior, Type A; Gender and Cardiovascular Health; Stress and Cardiac Response; Stress and Health Research

Bibliography

- Dusseldorp E, van Elderen T, Maes S, Meulman J, Kraaij V 1999 A meta-analysis of psychoeducational programs for coronary heart disease patients. *Health Psychology* **18**: 506–19
- Foreman D E, Farquhar W 2000 Cardiac rehabilitation and secondary prevention programs for elderly cardiac patients. *Clinics in Geriatric Medicine* **16**: 619–29
- Frasure-Smith N, Lesperance F, Prince R, Verrier P, Garber R A, Juneau M, Wolfson C, Bourassa M G 1997 Randomised trial of home-based psychosocial nursing intervention for patients recovering from myocardial infarction. *Lancet* **360**: 473–9
- Friedman M, Thoresen E, Gill J J, Ulmer O, Powell L H, Price V A, Brown B, Thompson L, Rabin D D, Breall W S, Bourg E, Levy E, Dixon T 1986 Alterations of Type A behavior and its effect on cardiac recurrences in post myocardial infarction patients: summary results of the recurrent coronary prevention project. *American Heart Journal* **112**: 653–65
- Jones D A, West R R 1996 Psychological rehabilitation after myocardial infarction: multicentre randomised controlled trial. *British Medical Journal* **313**: 1517–21
- Johnston M, Foulkes J, Johnston D W, Pollard B, Gudmundsdottir H 1999 Impact of patients and partners of inpatient and extended cardiac counseling and rehabilitation: a controlled trial. *Psychosomatic Medicine* **61**: 225–33
- Krantz D S, Lundgren N R 1998 Cardiovascular disorders. In: Johnston D W, Johnston M (eds.) *Health Psychology, Volume 8 of Comprehensive Clinical Psychology*. Pergamon, Oxford, UK, pp. 217–36
- Kugler J, Seelbach H, Krueskemper G M 1994 Effects of rehabilitation exercise programmes on anxiety and depression in coronary patients. *British Journal of Clinical Psychology* **33**: 401–10
- Lewin B, Robertson I H, Cay E L, Irving J B, Campbell M 1992 Effects of self-help postmyocardial infarction training program on psychological adjustment and use of health services. *Lancet* **339**: 1036–40
- Linden W, Stossel C, Maurice J 1996 Psychosocial interventions for patients with coronary artery disease—a meta-analysis. *Archives of Internal Medicine* **156**: 745–52
- Oldridge N 1998 Cardiac rehabilitation in the elderly. *Aging* **10**: 273–83
- Oldridge N B, Guyatt G H, Fischer M E, Rimm A A 1988 Cardiac rehabilitation after myocardial infarction: combined experience of randomized clinical trials. *Journal of the American Medical Association* **260**: 945–50
- Richardson L A, Buckenmeyer P J, Bauman B D, Rosneck J S, Newman I, J 2000 Contemporary cardiac rehabilitation: patient characteristics and temporal trends over the past decade. *Journal of Cardiopulmonary Rehabilitation* **20**: 57–64

D. W. Johnston

Rehabilitation of the Offender

The ways in which societies react to those who break criminal laws show major historical changes. In most civilized countries, corporal punishment and the death penalty have been replaced by prison sentences, fines, and a broad range of diversion programs, alternative sanctions, and community treatment (Tonry 1998). The latter include, for example, restitution, offender–victim mediation, social training courses, community service, intensive supervision, electronic monitoring, boot camps, day treatment centers, and other alternatives to institutionalization and routine probation.

Although the scope of criminal justice interventions has broadened, the ‘right way’ forward for crime policy is still controversial in many countries. The respective opinions are related to general political attitudes. They emphasize different aims in reactions to criminal behavior, such as compensation by the guilty, safety of the public, deterrence of other people from similar crimes, and prevention of future law breaking in the offender. The last aim is central to offender rehabilitation and relates to similar concepts such as resocialization or correctional treatment. These concepts refer to educational, psychological, and medical measures that are delivered in a systematic

way to reduce the risk of recidivism and contribute to (re-)integration in society.

During the 1960s and early 1970s, this idea became a centerpiece of crime policy in many countries. However, the concept of rehabilitation was also criticized from different perspectives that emphasized the other aims mentioned above. In the mid-1970s, a comprehensive review of offender rehabilitation programs concluded that the majority of available studies were methodologically too weak to prove treatment success (Lipton et al. 1975). Although this and other reviews revealed various positive effects, the slogan of 'nothing works' in offender rehabilitation became widespread. However, the 1990s saw a re-emergence of the issue of offender rehabilitation. Among other reasons, this was due to systematic research integrations via meta-analyses that led to a more differentiated discussion in research and practice (McGuire 1995, Palmer 1992).

1. General Efficacy

There are now hundreds of evaluations of offender treatment that include an untreated comparison group. The outcomes differ somewhat depending on

the age of offenders, previous criminality, offense types, modes of treatment, criteria of success, length of follow-up, and other issues. For example, various studies of sexual offender treatment revealed sexual reoffending rates between 3 percent and 44 percent with a mean of 18.8 percent (Hall 1995). The respective data for untreated control groups were 6 percent to 75 percent with a mean of 26.8 percent. Therefore, a single study cannot be generalized, and it is necessary to base research syntheses on common measures of effect size (ES). Such an indicator is the correlation coefficient *phi* or *r*. Table 1 summarizes the results of a number of recent meta-analyses on offender rehabilitation.

Although effects differ according to study samples and methods of computation, all mean ESs are positive. However, they are generally small. A realistic estimate of overall effectiveness is only approximately 0.10 or a little less. Taking the example of a control group with 50 percent reoffending, such an effect size of 0.10 would indicate a reduction to 40 percent recidivism. This is a 10 percentage points or 20 percent difference in comparison with the untreated group. Such a small effect may well be relevant in terms of cost utility. A realistic evaluation of correctional programs also has to take into account that the

Table 1
Selected meta-analyses on offender rehabilitation

Authors and topics of meta-analyses	Studies	ES (r)
Andrews et al. 1990 Community/institutional programs for juvenile/adult offenders	154	0.10
Gendreau and Goggin 1996 Community/institutional programs for juvenile/adult offenders	215	0.13
Gottschalk et al. 1987 Community programs for juvenile offenders	101	0.12
Hall 1995 Community/institutional programs for sexual offenders	12	0.12
Lipsey 1992 Community/institutional programs for juvenile offenders	440	0.05/0.08
Lipsey and Wilson 1998 Comm./institutional. programs for serious and violent juvenile offenders	200	0.06/0.09
Lösel 1995 German therapeutic prisons for serious adult offenders	18	0.11
Redondo et al. 1999 European comm./inst. programs for juvenile/adult offenders	32	0.12
Pearson et al. 1996 Community/institutional programs for juvenile offenders	293	0.09
Pearson et al. 1996 Community/institutional programs for adult offenders	211	0.07
Petrosino 1997 Randomized studies on programs for juvenile/adult offenders	150	0.02
Polizzi et al. 1999 Prison- and nonprison-based programs for sex offenders	13	0.09

optimum ES is not 1. Depending on the recidivism rate in the untreated group, 0.30 may already represent the maximum possible efficacy.

2. Different Types of Program

Most research reviews reveal significant differences between kinds of programs: the largest effects come from theoretically well-founded, multimodal, cognitive-behavioral, and skill-oriented programs. Also promising are family-oriented programs in ambulatory treatment for serious juvenile offenders and clearly structured therapeutic communities in the institutional treatment of adult offenders. Successful interventions fit the level of offenders' risk, address criminogenic needs precisely (instead of vague personality changes), and take into account the specific learning styles of the offenders (responsivity principle). Such programs are founded on empirical knowledge about the causes and development of the respective criminal behavior and are not just derived unspecifically from psychotherapeutic 'schools.' They address, for example, motivation for change, self-control, crime-related beliefs and attitudes, social skills, interpersonal problem-solving, self-critical thinking, moral reasoning, anger management, victim awareness, and coping with risk situations for relapse (e.g., Ross and Ross 1995). Cognitive-behavioral and relapse prevention approaches are not only most appropriate for the treatment of 'ordinary,' violent, and personality-disordered offenders but also for sexual offenders (Marshall et al. 1998). Here, combinations with hormonal measures seem to be relatively successful as well.

Effect sizes in the most appropriate programs can be twice as high as the average. In contrast, unstructured case work, traditional psychodynamic, and non-directive approaches show smaller effects than the overall mean. Relatively low-structured, self-governing, permissive therapeutic communities and milieu therapy also seem to range at the lower end of effectivity. The same holds for merely formal variations in punishment or probation; diversion without any educational or psychosocial component; deterrence, boot camps, and other measures of smarter punishment or intermediate sanctions. Aversive behavioral methods for sex offenders are also not successful. In some studies, the effects of inappropriate treatment are even negative.

3. Methodological Issues

Methodological features explain a substantial part of outcome variance. For example, studies with random assignment tend to produce smaller effects than quasi-experimental designs. There are also smaller effects in studies with larger samples, higher statistical power, and low rates of attrition between subject assignment

and outcome measurement. Juridically-proven recidivism shows somewhat smaller effects than measures of institutional adaptation or attitude change. Outcome criteria with good reliability and longer follow-up periods also tend to be associated with smaller effects.

Although the number of evaluations with reasonable methodological quality has increased, there are still very few well-controlled studies on specific modes of treatment and specific offender groups. This is particularly the case for the institutional treatment of serious adult offenders, psychopaths, and sex offenders. In these cases, it can be rather difficult to form equivalent control groups that do not receive any kind of treatment. So far, evaluations often address different degrees or modes of treatment and not an all-or-none dichotomy.

4. Program Integrity and Intensity

A certain treatment label may be misleading with respect to what actually happens in a program. However, high integrity (reliability) of program delivery is necessary for good outcomes. Naturally, this holds only for programs that are successful in principle. Low integrity may result from, for example, weakly-structured programs with no treatment manuals, organizational barriers and staff resistance, or insufficient staff training and supervision. Although psychosocial interventions are basically an individual process and not a standardized technique, it is important to assess and monitor program implementation systematically. Doing this may be one reason why studies in which researchers were involved in program delivery show better effects than other evaluations.

Treatment intensity and dosage also relate to outcome. Lipsey and Wilson (1998), for example, found that the duration of treatment correlated positively with effect size, whereas the mean number of hours a week correlated negatively with outcome. The latter result may be due to problems of implementation and some satiation in very compact programs.

5. Offender Characteristics

The efficacy of rehabilitation programs depends not least on the characteristics of the offenders. For example, child molesters seem to benefit more from sex offender treatment than rapists. In juvenile offenders, Lipsey and Wilson (1998) found that samples with mixed prior offense types (including violent offenses) had better outcomes than those containing primarily property offenders. Effects were also larger when all offenders had prior sentences. These and other data suggest that serious and violent offenders may benefit more from treatment than less serious ones. This is plausible when considered in light of the risk principle: offenders with various prior offenses

are more often at risk of life-long persistent antisocial behavior. In contrast, offenders with fewer prior offenses more often belong to the type of adolescence-limited antisocial behavior. In these cases, natural protective factors in everyday life counteract recidivism in the untreated control groups as well.

However, the finding that high-risk groups benefit most from rehabilitation programs cannot be generalized to very high-risk cases such as psychopaths. These offenders lack the interpersonal and motivational characteristics that are essential for successful treatment, and thus rarely show a good outcome (Lösel 1998). This would indicate that a nonlinear, U-shaped function may be an adequate description of the relationship between risk and treatment efficacy. It is in particular the low-risk and very high-risk cases that show the weakest effects.

The need principle and responsivity principle can be applied adequately only when a systematic assessment of the offender is performed. This should be based not only on static variables such as offender age or criminal history but also on dynamic characteristics that are accessible to change. A thorough assessment is necessary for adequate placement decisions, for example, in the case of dangerous offenders. This should also reduce the number of treatment dropouts. Most studies have shown that the latter group is particularly at risk for reoffending.

6. *Setting Characteristics*

Overall, there seem to be larger effects in community programs than in institutional treatment. As a result, custody should be the last resort (for financial reasons as well). However, placement in a closed setting is normally confounded with a higher risk of reoffending. Sweeping assumptions that successful treatment is not possible within prison or other institutional settings are not supported by empirical research. Negative incarceration effects may depend on personal, situational, and organizational moderators that can be addressed at least partially by adequate programs. In cases of drug use and other dangerous lifestyles, institutions may even have some protective effects. The alternative therefore of 'community vs. institution' is too undifferentiated. The systematic assessment of risk can contribute to a stepwise reduction of custody to accompany positive developments in the offender (e.g., Quinsey et al. 1998).

7. *Organizational Climate and Regime*

Perhaps, it is not institutionalization by itself but the quality of regime and social climate that accounts for positive or negative effects. Although institutional features vary widely even under similar formal conditions, they have not yet been investigated to any great extent. A regime that is emotionally and socially responsive as well as structured, norm-oriented, and

controlling is most promising. From process research on general psychotherapy, it is known that the relations between clients and therapists are very important for efficacy. Similarly, correctional programming requires well-selected, specifically trained, highly motivated, and continuously supervised staff. Attitudes and competencies that are not in line with the aims and content of a program may not only lower treatment integrity but even counterbalance potential efficacy.

8. *Aftercare and Relapse Prevention*

Even intensive rehabilitation programs make up only a small part of the factors that impact on the clients' thinking and behavior. It is therefore necessary to deliver services that consolidate the competencies acquired during treatment, to provide control over risks, and to cope with crises that may arise afterwards. Without such measures, treatment efficacy often can be demonstrated only during short follow-up periods. Although systematic programs of through-care, aftercare, and relapse prevention are particularly emphasized for the treatment of drug-addicted and sexual delinquents (Annis 1986, Laws 1999), they are relevant for other violent and personality-disordered offenders as well. Thorough planning and monitoring of relapse prevention refers to type and intensity of services, cooperation routines between the institutions, probation officers, and other staff involved. These multiagency networks are very important for preventing recidivism (Spencer 1999).

9. *Natural Protective Factors*

Even without any rehabilitation program, a subgroup of high-risk offenders desists from criminal behavior. In these cases, natural protective mechanisms make delinquent behavior less attractive. Although our knowledge about such processes is still rudimentary, various personal and social resources have repeatedly shown a protective function against criminality (see *Religiosity, Sociology of; Non-violence: Protective Factors*): for example, bonds to nondeviant family members or other reference persons; cognitive competencies, school achievement, and planning for the future; nondeviant peer groups; work motivation, vocational skills, and employment; or a supportive and prosocial intimate partner. Such protective factors are relevant for positive turning points in criminal careers. Therefore they need to be assessed and integrated into program planning and delivery.

10. *Community Features*

Individual and microsocial risk and protective factors also interact with the broader social context. Therefore community features such as the concentration of

poverty, unemployment, multiproblem families, violence, drug use, and other deficits in 'social capital' must be taken into account in offender rehabilitation. If, for example, a program promotes vocational skills, it may be particularly successful in communities with low unemployment rates (Sherman et al. 1997). In other contexts, the skills acquired will be less useful. Counterproductive influences may result from neighborhoods with high criminality and violence. In a community in which peers and neighbors model and reinforce deviant behavior, it is more difficult to preserve program effects than in a social context with less negative influences. Offender treatment and relapse prevention should therefore be integrated into broader community programs of crime prevention.

11. Conclusions and Perspectives

During the last decade, there has been growing evidence that some programs of offender rehabilitation work and others are promising. However, there is still a lack of methodologically sound evaluation, particularly of complex programs and multiple services for specific adult and serious offender groups and of the service delivery in everyday practice. More process data are needed on the content of program delivery, staff characteristics, and dimensions of the institutional regime. It is also necessary to assess control group conditions in a similar way. The dropout problem must be tackled more effectively. Adequate assessment procedures, program modules on motivation, and flexible procedures of case management are suggested here. Outcome evaluation should not rely primarily on one-shot studies but refer to a theoretically-based chain of various change measures, including longer follow-up periods. Last but not least, practice-oriented evaluations must place offender rehabilitation programs in their wider social context.

Although more systematic research is needed, empirical program evaluations and experiences from practice already suggest guidelines for effective rehabilitation. These include, for example:

- (a) Realistic expectancies regarding possible efficacy.
- (b) Theoretically sound concept (social learning, multimodal approach).
- (c) Systematic assessment of offenders' risk and matching to service level.
- (d) Appropriate targeting of specific criminogenic needs and not unspecific personality characteristics.
- (e) Teaching of self-control, nondeviant thinking, and social skills.
- (f) Consequence in applying reinforcement.
- (g) Matching of program type, offender, and staff.
- (h) Thorough selection, training, and supervision of staff.
- (i) Development of a supportive institutional climate and regime.

- (j) Neutralization of criminogenic social networks.
- (k) Strengthening of 'natural' protective factors.
- (l) Attainment of high program integrity and adequate intensity.
- (m) Dynamic assessment and monitoring of offender change.
- (n) Provision of systematic through-care, aftercare, and relapse prevention.
- (o) Continuous process and outcome evaluation of the program.

Countries such as Canada, England, and Scotland use similar guidelines for program accreditation. Naturally, we cannot expect that one specific type of program will emerge as the gold standard for successful offender rehabilitation. However, the current empirically-based knowledge on 'what works' is promising in securing service quality for both the society and the offender.

See also: Crime and Delinquency, Prevention of; Criminal Defenses; Criminal Justice, Ethics of; Aspects of Human Dignity; Criminal Law and Crime Policy; Deterrence; Deterrence: Legal Perspectives; Deterrence Theory: Crime; Program Evaluation; Relapse Prevention Training; Treatment of the Repetitive Criminal Sex Offender: United States

Bibliography

- Andrews D A, Zinger I, Hoge R D, Bonta J, Gendreau P, Cullen F T 1990 Does correctional treatment work? A clinically relevant and psychologically informed meta-analysis. *Criminology* **28**: 369–404
- Bernfeld G, Farrington D P, Leschied A (eds.) 2001 *Offender Rehabilitation in Action: Implementing and Evaluating Effective Programs*. Wiley, Chichester, UK
- Cullen E, Jones L, Woodward R (eds.) 1997 *Therapeutic Communities for Offenders*. Wiley, Chichester, UK
- Gendreau P 1995 The principles of effective intervention with offenders. In: Harland A J (ed.) *Choosing Correctional Options that Work: Defining the Demand and Evaluating the Supply*. Sage, Thousand Oaks, CA
- Gendreau P, Goggin C, Smith P 1999 The forgotten issue in effective correctional treatment: Program implementation. *International Journal of Offender Therapy and Comparative Criminology* **43**: 180–187
- Hall G C N 1995 Sexual offender recidivism revisited: A meta-analysis of recent treatment studies. *Journal of Consulting and Clinical Psychology* **63**: 802–809
- Hollin C (ed.) 2000 *Handbook of Offender Treatment and Assessment*. Wiley, Chichester, UK
- Laws D R 1999 Relapse prevention: The state of the art. *Journal of Interpersonal Violence* **14**: 285–302
- Lipsey M W 1992 Juvenile delinquency treatment: A meta-analytic inquiry into variability of effects. In: Cook T D, Cooper H, Cordray D S, Hartmann H, Hedges L V, Light L R, Louis T A, Mosteller F (eds.) *Meta-analysis for Explanation*. Russell Sage Foundation, New York

- Lipsey M W, Wilson D B 1998 Effective intervention for serious juvenile offenders. In: Loeber R, Farrington D P (eds.) *Serious and Violent Juvenile Offenders*. Sage, Thousand Oaks, CA
- Lipton D, Martinson R, Wilks J 1975 *The Effectiveness of Correctional Treatment*. Praeger, New York
- Lösel F 1995 The efficacy of correctional treatment: A review and synthesis of meta-evaluations. In: McGuire J (ed.) *What Works: Reducing Reoffending*. Wiley, Chichester, UK
- Lösel F 1998 Treatment and management of psychopaths. In: Cooke D J, Forth A E, Hare R D (eds.) *Psychopathy: Theory, Research and Implications for Society*. Kluwer, Dordrecht, NL
- Marshall W L, Fernandez Y M, Hudson S M, Ward T (eds.) 1998 *Sourcebook of Treatment Programs for Sexual Offenders*. Plenum Press, New York
- McGuire J (ed.) 1995 *What Works: Reducing Reoffending. Guidelines From Research and Practice*. Wiley, Chichester, UK
- Palmer T 1992 *The Re-emergence of Correctional Intervention*. Sage, Newbury Park, CA
- Quinsey V L, Harris G T, Rice M E, Cormier C A 1998 *Violent Offenders: Appraising and Managing Risk*. American Psychological Association, Washington, DC
- Redondo S, Sánchez-Meca J, Garrido V 1999 The influence of treatment programmes on the recidivism of juvenile and adult offenders: An European meta-analytic review. *Psychology, Crime and Law* 5: 251–278
- Ross R R, Ross B (eds.) 1995 *Thinking Straight*. Cognitive Centre, Ottawa, Canada
- Sherman L, Gottfredson D, MacKenzie D, Eck J, Reuter P, Bushway S 1997 *Preventing Crime: What Works, What Doesn't, What's Promising*. Report to the United States Congress: University of Maryland, US Department of Justice, Washington, DC
- Spencer A P 1999 *Working With Sex Offenders in Prisons and Through Release to the Community*. Jessica Kingsley, London
- Tonry M (ed.) 1998 *The Handbook of Crime and Punishment*. Oxford University Press, New York

F. Lösel

Reification: History of the Concept

Notwithstanding its esoteric overtones and philosophical intricacies, reification (*Verdinglichung*) is a central concept of social theory in general and of critical theory in particular. Although reification has received the greatest attention in Western Marxism, and above all in Lukács' *History and Class Consciousness* (1923), it is important not to restrict the use of the concept to that tradition but to see that it can also and already be found in the work of Kant, Hegel, Nietzsche, Dilthey, Husserl, Heidegger, Simmel, and Max Weber to criticize the dehumanizing, rationalizing, and alienating tendencies of modernity.

As a technical term, the term reification emerged in the English language in the 1860s out of the contraction of the verb *facere* (to make) and the substantive *res* (thing), which can refer both to concrete and empirically observable things (*ens*) and to ab-

stract, indeterminate things (*aliquid*). As a synonym of 'thingification,' the inverse of personification, reification metaphorically refers to the transformation of human properties, relations, processes, actions, concepts, etc. into *res*, into things that act as pseudopersons, endowed with a life of their own. This material transmutation of persons, relations, processes, concepts, etc. into thing-like entities that act like pseudopersons can operate both on a methodological and a social level. In both cases, however, the concept is usually used as a *Kampfwort* to denounce the 'violence of abstractions,' either of conceptual abstractions (Sohn-Rethel's *Denkabstraktionen*) that suppress the reflexive embeddedness of concepts into their social context, treat social facts as things, and transform metasubjects into megasubjects, or of real abstractions (Marx's *Realabstraktionen*) that strip individuals of their autonomy and reduce them to cogs of an abstract social machinery. This reference to the Kantian notion of autonomy is important. Outside of the normative context of an 'enlightened critique of the Enlightenment,' the notion of reification hardly makes sense, as can be gathered from the fact that conservative, skeptical, and ultraliberal theorists such as Gehlen, Luhmann, and Hayek have a theory of the autonomization of social structures but no theory of reification to connect it to the alienation of individuals in modern capitalist industrial societies.

1. Methodological Reification—or the Critique of Conceptual Abstractions

In the philosophy of the social sciences, the concept of reification is used (a) to denounce the hypostasis or substantialization of concepts (nominalist critique of reism), (b) the naturalization of the subject and the life-world (humanist critique of naturalism), and (c) the ideological justification of the status quo (dialectical critique of fetishism).

1.1 Nominalist Critique of Reism

In the case of the critique of reism (or naive conceptual realism), the notion of reification of concepts is used to denounce, from a nominalist, vitalist, or criticist perspective, the categorical error of transforming abstractions (notions, representations, concepts) into a material reality, in a concrete object 'out there.' Reification is here understood as a synonym of the 'fallacy of misplaced concreteness' (Whitehead). What is criticized is the hypostasis of concepts, analytical constructs, and ideal types, the sliding from the substantive to the substance, which involves a subreption of the categorical thing with the 'thing in itself.' This is, for instance, the case with macrosociologists who transform their own conceptual constructs or those of

the actors ('the State,' the *Bourgeoisie*, the 'Proletariat') into historical subjects capable of agency and of determining their own ends ('the State decides,' 'the Anglican Church resists,' 'the glorious Proletariat triumphs,' etc.). It should, however, be noted that due to the absence of a consensus on the ultimate referents of reality and the fact that one can always submit the concepts of the scientist to a neo-Kantian critique of ontology, the charge of reification is almost inevitable. Given that one's typification is another's reification, the critique of 'false conceptual realism' (Weber) is endemic in sociology.

1.2 Humanist Critique of Naturalism

The critique of positivist naturalism in terms of reification of the subject and the life-world is linked to the series of methodological disputes (*Methodenstreit*) which, since the double foundation of sociology by Comte and Dilthey in the nineteenth century, have opposed the partisans of the method of causal explanation (*Erklären*) to the partisans of the interpretative methods (*Verstehen*). Drawing on Vico's principle of the *verum factum* (*verum et factum convertuntur*), according to which we can understand the sociohistorical reality because it is a human product, but not nature which is a divine product, humanists claim that the appropriate method of sociology is interpretative in that it aims to understand, by means of a phenomenological and hermeneutic reconstitution of the meaning of action, the social-historical world (Hegel's objective spirit) as an objectivation of subjective actions. Social facts thus have a meaning and cannot be treated 'as if they were things' (Durkheim). The naturalistic elimination of the meaningfulness of action through statistical observation is reifying in that it transforms psychic acts into pseudophysical facts and reduces culture to (second) nature. Against Durkheim and his fellow 'factists' who 'change the subject' of the human sciences by substituting factors for actors, humanists thus argue that social facts are not things but that things are social facts whose meaning can be understood and which can be interpreted as an 'ongoing accomplishment of the concerted activities of daily life' (Garfinkel).

1.3 Dialectical Critique of Fetishism

The dialectical critique of fetishism offers a meta-theoretical critique of the ideological implications of *bourgeois* theories and methodologies of the social that, due to a lack of reflexivity on their context of genesis and application, legitimize the status quo. Dialecticians accept the limits of 'hermeneutic idealism.' When social relations have crystallized into a

'second nature' and social subsystems follow their own pseudonatural laws, 'dehumanizing' theories (e.g., functionalism, structuralism, systems theory) and methods (e.g., linear modeling, statistical regression) can and have to be applied. But if one does not want to fall prey to a 'reification of the second order' and give a 'reified perception of the reifying' (Adorno) that willy-nilly endorses the reality it registers, the observed facts have to be 'mediated by the totality' (Lukács 1923) and defetishized in such a way that the tension between the real and the possible, between what is and what could or should be, becomes perceptible within the facts themselves.

2. Social Reification—or the Critique of Real Abstractions

In the tradition of Western Marxism, Marx's theories of alienation and the fetishism of commodities are combined with Hegel's dialectical phenomenology of spirit, Simmel's theory of the tragedy of culture, and Max Weber's theory of formal rationalization to form a critical theory of society. The concept of reification is used to refer to the relatively autonomous, alienating and alienated functioning of the social and cultural (sub-) systems of modern capitalist societies that impose their constraints from without on individuals, limit their freedom and tend to reduce them to powerless 'carriers' or passive 'executioners' of the system. As products of praxis, institutions and organizations are human objectivations, but in the course of their development, the cultural and social (sub)systems have been complexified, formalized, rationalized, and depersonalized to such an extent that eventually they have been transmuted into self-referentially closed systems that function independently of the will and the consciousness of individuals, thwart their plans, threaten their autonomy, and perhaps even their existence. The critique of reification is dialectical and thus somewhat paradoxical: the insistence on the alienating autonomy of the system, which is an objectivation of action, aims to reactivate the autonomy of the individuals and to overcome their alienation.

Although the concept of reification (*Verdinglichung*) can already be found in Hegel's *Philosophy of Right*, the real history of the concept begins with Marx and with Lukács's Hegelian interpretation of Marx. The origins of the theory of reification are usually found just where the word itself is absent, namely in the famous section on the fetishism of commodities (Marx 1869, Chap. 1, Sect. 4). Analyzing capitalism as a system of generalized exchange, Marx notes that the commodity has become the universal form of the product of labor, with the result that the exchange value of the commodity supplants the use value. Consequently, the exchange value appears to those

who exchange goods as a property of the commodity itself, whereas in reality it is the result of the labor that is incorporated in the commodity and that expresses itself as a quantitative relationship between the exchanged goods. 'It is nothing but the definite social relation between men themselves which assumes here, for them, the phantasmagoric form of a relation between things' (Marx 1869, pp. 23, 86). This inversion of humans and things is not simply an illusion, however, but an expression of the real nature of social relations in a competitive market environment. In the absence of a central organism that regulates both the production and the distribution of the products of labor, the social integration of humans is imposed from without by the systemic interconnection of things.

In 'Reification and the Consciousness of the Proletariat,' the central chapter of *History and Class Consciousness*, Lukács, a Hegelian Marxist who was once a student of Simmel and Max Weber, presents the classic formulation of the theory of reification. Synthesizing Weber's theory of formal rationalization with Marx's theory of commodity fetishism, Lukács generalizes the theory of commodity fetishism beyond the sphere of circulation. In the problem of fetishism, which he immediately identifies with the phenomenon of reification, he discovers the 'central, structural problem of capitalist societies in all its aspects' (Lukács 1923/1968, p. 257). The universality of the commodity form, conceived as the prototype of all the forms of objectivity that follow their own laws and dissimulate the traces of human relations that subtend them, affects the life of everybody, both in its objective and subjective manifestations. Objectively, individuals are confronted with a second nature of pseudo-things against which they are powerless; subjectively, they are estranged from their own activity, apprehending the products of their own activity in an alienated mode—'as if they were something else than human products' (Berger and Luckmann). Moving from the sphere of circulation to the sphere of production, Lukács rediscovers the theory of the alienation of labor which the young Marx had developed but not published in the *Parisian Manuscripts* of 1844 (see *Alienation, Sociology of*). In the sphere of material production, reification expresses itself most clearly in the reduction of labor power to a commodity and of the laborer to an appendix of the machine. In capitalism, reification is generalized and the fate of the worker becomes paradigmatic of the fate of everyone. Expressing the Messianism of the oppressed, Lukács eventually reintroduces the Proletariat as the 'identical subject-object' of history whose revolutionary actions overcome alienation and reification and thus realize the Hegelian dream of the restoration of the 'beautiful totality.'

The development of the so-called Critical Theory of the Frankfurt School can best be understood as the result of a progressive disillusion with revolutionary

expectations. Eliminating the Hegelian-Marxist dialectic of consciousness, which they replace by a Freudian account of sublimation and repression, Horkheimer, Marcuse, and especially Adorno, who also gave a Nietzschean twist to the concept of reification, radicalize the Weberian-Marxist strand in Lukács' theory. Universalizing and totalizing reification to the point that it appears as an ontological feature of human civilisation, they almost end up indicting Reason as such. Indeed, to explain totalitarianism, Horkheimer and Adorno develop a negative philosophy of history which uncovers in the first protohistorical attempts to dominate nature the origin of the fatal unfolding of a diabolic logic of increasing reification that will find its culmination (but not its endpoint) in the death camps. In his *Theory of Communicative Action*, Jürgen Habermas (1981), the main representative of the second generation of critical theory, reformulates the theory of reification in terms of the paradigm of language. In this perspective, reification is no longer associated with rationalization as such, as was the case with Max Weber and the Frankfurt School, but reconceptualized in terms of the 'colonization of the life-world' by the subsystems of the economy and the administration. When the mechanisms of systemic integration (money and power) force back the forms of social integration from those domains that can only be integrated through language, a reification ensues that leads to a pathological deformation of the life-world.

Bibliography

- Arato A 1974 The neo-idealist defense of subjectivity. *Telos* 21: 108–61
- Arato A, Breines P 1979 *The Young Lukács and the Origins of Western Marxism*. Seabury Press, New York
- Berger P, Pullberg S 1965 Reification and the sociological critique of consciousness. *History and Theory* 4(2): 196–211
- Gabel J 1962 *La fausse conscience. Essai sur la réification*. Editions de Minuit, Paris
- Habermas J 1981 *Theorie des kommunikativen Handelns*, Vols. 1 and 2. Suhrkamp, Frankfurt, Germany (1984 *Theory of Communicative Action*, Vol. 1. Heinemann, London, 1987, Vol. 2. Polity Press, Cambridge, UK)
- Lukács G 1923/1968 Geschichte und Klassenbewusstsein. In: (eds.) *Werke*, Band 2. Neuwied, Luchterhand (1971 *History and Class Consciousness*. Merlin, London)
- Marx K 1869/1972 *Das Kapital*, Vol. I. In: (eds.) *Marx Engels Werke*, Vol. 23. Dietz Verlag, Berlin (1974 *Capital*, Vol. 1, Lawrence & Wishart, London)
- Pitkin H 1987 Rethinking reification. *Theory and Society* 16(2): 263–93
- Rose G 1978 *The Melancholy Science: An Introduction to the Thought of Theodor W. Adorno*. Macmillan, London
- Thomson B C 1982 *Making Sense of Reification: Alfred Schütz and Constructionist Theory*. Humanities Press, Atlantic Highlands, NJ

Vandenbergh F 1997–1998 *Une histoire critique de la sociologie allemande. Aliénation et réification*. 2 vols. Editions la Découverte, Paris (2001 *Critique of Reification*. University of Minnesota Press, Minneapolis, MN)

F. Vandenbergh

Reinforcement: Neurochemical Substrates

1. The Operant–Respondent Distinction

The neurochemical mechanisms that mediate reinforced or operant behavior may differ in a fundamental way from those underlying reflexes or respondent behavior. This is because environmental stimuli appear to control the two classes of behavior in fundamentally different ways. In reflexes, whether conditioned or unconditioned, the controlling stimulus precedes the response and elicits it. In operant conditioning, the controlling stimulus follows the response and elevates its subsequent probability. When the controlling stimulus precedes the response, information flow in the brain is afferent to efferent, as in the conventional reflex arc. On the other hand, when the controlling stimulus follows the response, as in reinforced behavior, the underlying brain organization seems to require an unconventional circuitry in which efferents are activated before afferents. However, the mechanisms for reinforced behavior do not require circuits that directly link efferent to afferent elements. This is because operant behaviors do not directly activate the goal-detecting afferent systems. Rather, the correct response operates on the environment to produce the goal object and it is this environmental change that activates the goal-detecting systems. Thus, although the reinforcement mechanism does not require efferent-to-afferent circuitry, it must recognize efferent–afferent contingencies and must be operated selectively by them, i.e., it must cause behavioral reinforcement only when the neuronal substrates of the correct response and goal object, in that order, are activated sequentially.

2. Characterizing the Brain Reinforcement Mechanism

The problem of characterizing the brain reinforcement mechanism has two main parts: the first is to identify the neuronal substrate that performs the reinforcing function (reinforcing substrate) and the second is to identify the neuronal substrate that is modified by the reinforcement process (target substrate). Since any goal object can reinforce any behavior in an animal's repertoire, it seems likely that there is (a) convergence of goal-object input to the reinforcing substrate, and (b) divergence of reinforcing output from this sub-

strate to the target substrate. Here we briefly summarize the substantial published literature that has been devoted to the identification of the reinforcing substrate. Until very recently, little consideration has been given to the question of the target substrate, but it may be useful to review some initial findings and to consider the directions that research on this problem may take.

2.1 Neurochemical Specialization

The idea that reinforcing functions are specialized neurochemically has guided research in this field for more than 30 years. The hypothesis that certain dopamine and opioid peptide brain cells may serve as reinforcing neurons is supported most directly by evidence from brain self-stimulation and drug self-administration experiments.

2.2 Electrical Self-stimulation of the Brain

In self-stimulation experiments, animals work to deliver electrical stimulation to their own brains through permanently indwelling electrodes. In the absence of other sources of reward, the reinforcement for self-stimulation behavior must arise from the neuronal activity that is excited by the electrical stimulus. If so, it would be logical to assume that some of the neurons under the electrode tip actually are the reinforcing neurons that mediate the effects of natural reinforcers or at least are neurons that directly excite them.

High self-stimulation rates are observed when electrodes are implanted in regions containing dopamine (or opioid peptide) cell bodies or pathways. In particular, self-stimulation tightly overlaps the distribution of dopamine cells in the ventral tegmentum and substantia nigra. Self-stimulation closely follows the anteriorly projecting dopamine fibers through the hypothalamus, but it correlates somewhat less closely with the dopamine terminal fields in the forebrain. The involvement of norepinephrine neurons in self-stimulation is more controversial. Although many laboratories report self-stimulation from sites in the vicinity of the locus coeruleus, it has not been possible to establish convincingly that the noradrenergic neurons that make up this nucleus are responsible for the reinforcing effect. Mapping of opioid peptide sites for self-stimulation is consistent with the idea that certain beta-endorphin and dynorphin neurons are involved in self-stimulation, but these studies are still in an early stage. The dopamine-opioid peptide reinforcement hypothesis also is supported by pharmacologic experiments. Antagonists of dopamine and opioid peptides, such as haloperidol and naloxone respectively, should block chemical transmission of reinforcement messages. In support of the model, there are many

reports that these drugs selectively block self-stimulation.

Interestingly, the pharmacology of avoidance behavior (a prototypical example of negative reinforcement) closely overlaps that of self-stimulation (exemplifying positive reinforcement). In particular, avoidance and self-stimulation behaviors are selectively facilitated by dopamine agonists, such as amphetamine, and both are selectively inhibited by dopamine antagonists, such as chlorpromazine. Such evidence suggests that a common dopamine substrate may underlie both positive and negative reinforcement processes.

2.3 Intravenous Drug Self-administration

In self-administration experiments, behavior is reinforced by systemic or central injections of neurotransmitters or drugs. Although thousands of chemical substances are available, animals and humans avidly self-administer only a few. These self-administered substances may properly be termed pharmacologic reinforcers and it is interesting to ask why these chemicals are selectively associated with behavioral reinforcement. It is no coincidence that most powerful pharmacologic reinforcers have the ability to mimic or release the hypothesized natural reinforcement transmitters. Thus, many dopamine and opioid receptor activators are known to support self-administration behavior. In particular, selective agonists acting at μ and δ opioid receptors function as positive reinforcers, as do the naturally occurring opioid peptides, β -endorphin and dynorphin. Similarly, agonists acting at dopamine D_1 -like and D_2 -like receptors, such as apomorphine (D_1/D_2), piribedil (D_2), and SKF 82958 (D_1) are self-administered. The patterns of self-administration for many of these receptor agonists (particularly, D_1 and μ agonists) resemble the patterns observed when highly addictive stimulant drugs (e.g., cocaine and amphetamine) or opiate drugs (e.g., morphine and heroin) are self-administered. Hence it is reasonable to suppose that addiction to stimulant or opiate drugs may depend largely on their ability to activate or to enhance the activity of natural dopamine or opioid peptide reinforcement systems respectively. Finally, dopamine and opioid peptide antagonists, such as chlorpromazine and naloxone, not only appropriately block the reinforcing effects of agents that activate dopamine or endorphin receptors, but also appropriately block the reinforcing effects of stimulant and opiate drugs. Serotonergic agents generally are not self-administered but serotonin agonists and antagonists have been shown to alter self-administration of psychostimulants such as cocaine.

The self-administration method also has been used in attempts to identify the sites of action of stimulants and opiates. In some studies, animals are trained to self-inject cocaine or morphine directly into the brain

in an attempt to find preferred sites for reinforcing effects. In other studies, dopamine or opioid receptor antagonists are injected into different brain regions in attempts to block intravenous self-administration of drugs of abuse. Results from several laboratories suggest that important sites of action for both stimulant and opiate drugs are concentrated in the 'extended amygdala' including brain regions on an axis from the nucleus accumbens to the central amygdala. Reinforcing actions also are reported after direct injections of cocaine in the prefrontal cortex, morphine in the ventral tegmental area, and dynorphin in the hippocampus. Evidently, target sites for the reinforcing actions of drugs are widely distributed in the brain.

A closer analysis of brain transmitter involvement in reinforcement has been accomplished using *in vivo* microdialysis to monitor neurotransmitter levels in various brain areas during drug self-administration. Each cocaine self-injection caused a large increase in dopamine levels in the nucleus accumbens and these dopamine levels usually fell substantially to a 'trigger' concentration before the animal delivered the next self-injection. The addition of heroin to the cocaine solution caused synergistic increases in nucleus accumbens dopamine levels that are consistent with the higher abuse potential of such 'speed ball' combinations.

3. Methodological Considerations

Although these anatomical and pharmacological findings are generally consistent with the idea that reinforcing functions are performed by specialized systems of dopamine and opioid peptide neurons, some important problems remain. Administration of pure substances in the self-administration experiments strictly controls the chemical nature of the reinforcing injections but because the distribution of injected transmitters to active sites cannot exactly duplicate that of naturally released transmitters, the ensuing pattern of receptor activation could be artifactual or misleading. On the other hand, while electrical activation of transmitter pathways during self-stimulation presumably releases the reinforcing chemical messenger in a relatively natural distribution at appropriate postsynaptic sites, the electrical stimulus also must cause the simultaneous release of many irrelevant transmitters and neurohormones, including some that are still unknown. Thus, in self-stimulation, identification of which cells and which transmitters are relevant to the reinforcement process is largely a matter of inference. Solutions based on mapping studies demonstrate self-stimulation in anatomically coherent systems and those based on pharmacological studies implicate specific neurotransmitters. However, the anatomical mapping data from most self-stimulation experiments are insufficiently detailed. At

the same time, many of the pharmacological reports can be criticized because they often fail to distinguish specific drug effects on reinforcement from nonspecific effects on motor performance or attention. In addition, some self-stimulation findings seem to contradict the dopamine theory of reinforcement. For example, lesions of the nigrostriatal bundle do not permanently eliminate substantia nigra self-stimulation and, if only motorically simple responses are required, hypothalamic self-stimulation survives extirpation of all forebrain dopamine terminals.

4. The Dopamine-opioid Peptide Theory of Reinforcement

Despite such difficulties, the dopamine-opioid peptide theory of reinforcement currently enjoys a widespread acceptance. First, many investigators are impressed by the powerful and reliable reinforcing actions that highly localized electrical brain stimulation and specific chemicals can produce. Second, investigators are impressed by the consistent picture of dopamine-opioid peptide involvement in reinforcement processes that has emerged from the self-stimulation and self-administration data. Other neurochemical systems, in addition to dopamine and opioid peptides, also are likely to participate in reinforcement functions. For example, brain cannabinoid and nicotine receptors are probably involved in the reinforcing actions of marijuana and nicotine respectively. As additional receptors are identified as candidates for reinforcing actions, questions arise regarding their primacy and specific roles in reinforcement. Evidence now exists, for example, that several transmitter systems (e.g., cannabinoid, nicotine) have receptors localized on dopaminergic neurons that in turn may represent the final common pathway for reward. Another indication of a final common pathway utilizing dopamine is that agents that directly stimulate dopamine systems (e.g., amphetamine or cocaine) are more robustly self-administered than are agents that indirectly stimulate dopamine neurons (e.g., marijuana or nicotine).

Although a final common dopamine reward pathway remains a possibility, the crucial importance of reinforcement for survival makes it more likely that redundant neurotransmitter systems are involved. Most research on reinforcement has focused on the familiar transmitter systems mentioned above but complete understanding of reinforcement neurochemistry must await the likely discovery of novel transmitters. For example, several hundred G-protein coupled receptors have been cloned but their endogenous transmitters have not yet been identified. Some of these 'orphan' receptors may well play a role in reinforcement and their endogenous ligands (when elucidated) will serve as novel reinforcement transmitters. Such discoveries will lead to a more subtle and comprehensive understanding of reinforcement.

5. Identification of the Target Substrate

As noted, little theoretical or experimental effort has been devoted to the identification of the substrate that is modified by the reinforcement process. What are the neuronal targets of the reinforcing neurons? Because it is operant behavior that is reinforced, it is plausible to assume that behavioral substrates are major targets. It is commonly believed that operant conditioning involves the strengthening or reorganization of complex (but yet to be identified) neuronal circuits, which presumably form the substrates of reinforceable responses. According to this conventional view, the whole response is the unit of reinforcement. Where in the rat's brain, for example, would one find the circuit for a lever-press response? How might this circuitry be reorganized when lever pressing is reinforced? These anatomical and physiological considerations, as well as difficulties at the behavioral level associated with response definition, have led some scientists to question the idea that elaborate neuronal circuits are the functional units for reinforcement. According to B. F. Skinner, a more useful conceptual scheme recognizes the ultimately continuous nature of operant behavior. In this scheme, behavior is construed—not as a series of discrete acts—but rather as a continuous flow of activity and interaction with the environment, which arises from the largely spontaneous activation of infinitesimal behavioral elements. It is these elements, or 'behavioral atoms,' and not whole responses, that are the units strengthened by reinforcement. If so, and if atoms of behavior can be represented by the activity of individual neurons, then it could be argued that it is the behavior of individual neurons, and not of neuronal circuits, that is directly modified by reinforcing signals.

The cellular reinforcement hypothesis is currently under test in our laboratories by use of a novel '*in vitro* reinforcement' paradigm. This method employs training procedures closely analogous to those of behavioral operant conditioning to answer the question: Can the activity of relatively isolated brain cells in hippocampal slices be reinforced by local micro-pressure injections of dopamine or other reward-related transmitters or drugs? It was found that the bursting responses of individual pyramidal neurons in the CA I layer were in fact progressively increased by burst-contingent applications of dopaminergic and cannabinoid receptor agonists. The bursting responses of CA3 units were similarly increased by opioid receptor agonists. The same microinjections, administered independently of cellular activity, failed to facilitate and frequently suppressed cellular bursting. This observation, and the fact that glutamate (an excitatory transmitter that is not associated with the behaviorally-reinforcing effects of drugs) failed to increase bursting, indicated that general stimulation of cellular activity is an unlikely explanation of the facilitatory action of the burst-contingent injections.

Reinforcement delays exceeding 200 ms largely eliminated the reinforcing action of dopaminergic agonists. These findings support the conclusion that the behaviorally-reinforcing effects of dopaminergic, opioid, and cannabinoid drugs can be modeled *in vitro* at the cellular level.

See also: Amygdala (Amygdaloid Complex); Anxiety and Fear, Neural Basis of; Autonomic Classical and Operant Conditioning; Avoidance Learning and Escape Learning; Brain Stimulation Reward; Classical Conditioning, Neural Basis of; Conditioning and Habit Formation, Psychology of; Drug Addiction; Drug Addiction: Sociological Aspects; Drugs and Behavior, Psychiatry of; Memory: Synaptic Mechanisms; Motivation, Neural Basis of; Reinforcement, Principle of

Bibliography

- Gardner E L 1997 Brain Reward Mechanisms. In: Lowinson J H, Ruiz P, Millman R B, Langrod J G (eds.) *Substance Abuse: A Comprehensive Textbook*. Williams & Wilkins, Baltimore, MD, pp. 51–85
- Klopf A H 1982 *The Hedonistic Neuron: A Theory of Memory, Learning, and Intelligence*. Hemisphere Press, New York
- Self D W, Stein L 1992 Receptor subtypes in opioid and stimulant reward. *Pharmacology Toxicology* **70**: 87–94
- Skinner B F 1953 *Science and Human Behavior*. The Free Press, New York
- Stein L 1978 Reward transmitters: Catecholamines and opioid peptides. In: Lipton M A, DiMascio A, Killam K I (eds.) *Psychopharmacology: A Generation of Progress*. Raven Press, New York, pp. 569–81
- Stein L, Xue B C, Belluzzi J D 1994 *In vitro* reinforcement of hippocampal bursting: A search for Skinner's atoms of behavior. *Journal of the Experimental Analysis of Behavior* **61**: 155–68
- Wise R A 1996 Addictive drugs and brain stimulation reward. *Annual Review of Neuroscience* **19**: 319–40

J. D. Belluzzi and L. Stein

Reinforcement, Principle of

Reinforcement refers to an increase in the probability of a behavior when the behavior results in a positive consequence. That behavior is determined by its consequences is one of the most important principles within behavior therapy (see *Behavior Therapy: Psychological Perspectives*) for understanding the origins, maintenance, and change of clinically-relevant activities. 'Reinforcement' can denote (a) a principle of learning or behavior regulation; (b) a procedure designed to modify the frequency of an approved

behavior (the verbal form 'to reinforce' referring to the contingent delivery of a positive stimulus); or (c) the consequence itself (also 'a reinforcer'). The term connotes 'strengthening' and what is strengthened is the likelihood of a behavior reoccurring in a specific context.

There is circularity in such a statement, as a stimulus event is judged to be a positive reinforcement to the extent that it actually increases the likelihood of the behavior that produced it. Thus, reinforcement is not quite the same as reward, since a reinforcer is defined in terms of what 'works' or functions to strengthen behavior, whereas a reward is identified by what an individual expresses a desire for, or enjoys. In practice, however, one selects reinforcements on the basis of prior experience, observation, and intuition that the planned consequence will be desired or enjoyable. When an organism will engage in a behavior that produces a consequence (food, praise, stimulation, or even apparently negative outcomes such as fatigue or annoying someone else) it is referred to as positive reinforcement; when an organism works actively to reduce or eliminate events currently in existence (such as cold, teacher demands, loneliness, boring activities), it is known as negative reinforcement.

Any previously neutral stimulus that signals the onset or further probability of a reinforcer comes to have reinforcing properties of its own, and is described as a secondary or conditioned reinforcer. Socially, money is the classic example. In applied contexts, a stimulus that has no value other than a designated rate of exchange with something desired is called a token reinforcement—often these are symbols such as poker chips or points that can later be exchanged for privileges.

1. History of the Concept

Reinforcement is closely linked to evolutionary principles, in that it represents adaptive functioning with obvious survival implications, and applies to all species capable of benefitting from experience. Its origin as a formal principle of modern psychology derives from Thorndike's (1927) Law of Effect: the occurrence of a 'satisfying' event 'stamps in' the behavior that was instrumental in producing it. Hull's (1943) influential learning theory proposed that the increase is in the strength of the association between a prior stimulus (cue or context) and the response, with the mechanism underlying reinforcement being the reduction in a basic biological drive. However, many stimuli can serve as reinforcers when they have only incentive value (such as a sweet but non-nutritive substance).

Much past research has been devoted to the question of whether reinforcement is necessary for learning to occur; the present consensus is that it is not. The question, however, may be a spurious one, since it is difficult to imagine a behavior that has no conse-

quences, and the positive nature of the outcome might be as simple as reducing constraints on the individual, or predicting future pleasures.

In the classical conditioning paradigm (see *Classical Conditioning and Clinical Psychology*), Pavlov also used the term to refer to the 'unconditioned stimulus' (UCS: the stimulus that evokes the reflexive response). When this stimulus is regularly paired with (predicted by) a neutral stimulus, the latter comes to evoke a somewhat similar response. The UCS is thus considered reinforcement, since its occurrence increases the tendency to emit a response in the presence of a previously neutral stimulus. Of considerable theoretical and practical importance is the phenomenon that when the UCS is removed and the previously neutral stimulus presented repeatedly, the conditioned response eventually ceases to occur. This phenomenon, as well as the procedure of removing the UCS, is known as extinction. It has an exact parallel in instrumental conditioning: if the overt behavior no longer produces the positive outcome, its probability decreases. This process is not simply the opposite of reinforcement; it is not unlearning or forgetting, but learning to do something else.

Although central to both conventional instrumental and classical learning situations, the concept of reinforcement has acquired much of its practical, applied implication from Skinner's (1969) paradigm of operant conditioning (see *Operant Conditioning and Clinical Psychology*). The typical laboratory procedure is for a specific 'operant' response, such as the pressing of a lever or push button, to result in the automatic delivery of a small amount of food for a hungry animal. The operants studied have been discrete, simple responses that can be repeated easily by the subject. A vast body of experimental research has demonstrated that the form, intensity, and temporal pattern (rate) of responding is a function of the rule specifying the relationship between the operant response, its consequence, and the context in which the consequence occurs (referred to as the 'three-term contingency').

2. Contingency Principles

The effectiveness of this procedure for developing or maintaining a specified operant is influenced by the immediacy and predictability with which the reinforcement is delivered. Although it will slow the acquisition of new responses, the reinforcement need not be delivered after every response, an arrangement called a 'partial schedule of reinforcement.' The most common are ratio schedules, in which reinforcement is delivered after a specified number of responses; and interval schedules, in which reinforcement is delivered to the first response occurring after a predetermined period of time has elapsed. In the laboratory, these

different schedules of reinforcement produce characteristic patterns of response rate. Analogies are often made with workplace contingencies; for example, if one is paid on a piecework basis it is a ratio schedule, whereas being paid every month is an interval schedule of reinforcement. Reinforcement can also be delivered after a variable number of responses or periods of time—gambling on a slot machine represents a variable ratio schedule.

Of greater importance for application is that partial schedules of reinforcement result in a behavior being highly resistant to extinction, such that the subject will persist in the behavior for long periods of time, even when the reinforcement has been withdrawn (extinction). In clinical situations one might think that a behavior is maladaptive as it is apparently not being reinforced and should therefore extinguish. If the individual's history of reinforcement were known, however, it would be recognized that this behavior may have been reinforced sporadically in the past.

The precise form of any behavior can be determined by differential reinforcement, in which the criteria for exactly what response will result in reinforcement are carefully specified. By changing or narrowing these criteria gradually, a behavior can be 'shaped' into a specified configuration. Similarly, it is quite common in practice to reinforce differentially decreases in an undesirable behavior, or reinforce differentially alternative, more appropriate behaviors.

3. Limitations in Generalizing from the Laboratory

Despite the substantial body of knowledge derived from studies of animal behavior, the analogy between the operant laboratory and real life has limitations. Laboratory contingencies are contrived—unlike foraging in the wild, there is no intrinsic relationship between a lever press and obtaining food. There are many comparable artificial contingencies operating in everyday life, particularly work settings, in which a person is reinforced with money, food, or favors, for activities desired by the individual controlling these reinforcers. But more commonly our activities are intrinsically as well as socially rewarding, for instance, reading a book is reinforced by enjoyment of the story as well as by teacher approval, or climbing a mountain may be reinforced by a sense of accomplishment in addition to the admiration of friends.

It has long been debated that if an activity which is usually reinforced intrinsically is placed under an artificial external contingency (such as a monetary payment), this will later interfere with the intrinsic reward, such that once the artificial reward is withdrawn, the frequency of the behavior will drop below its previously (intrinsically) reinforced rate (see Carton and Nowicki 1998 for a discussion). There is also

considerable overlap in practice between the concept of reinforcement and that of feedback—providing people with information regarding their task performance. Feedback too can impair performance, regardless of whether it provides information about success or failure, by directing attention to self rather than the task at hand.

Another major difference between reinforcement studied in the laboratory and the principles affecting human social behavior is that the verbal specification of a contingency can function much like an actual contingency. Thus, for example, people can be promised a reward for an activity and this will change their behavior. Skinner referred to this as ‘rule governed behavior,’ whereas more cognitively oriented theorists prefer to talk of expectancies. Bandura (1977), for example, argued that performance on a task depends partly on the expectation that a certain reward will be forthcoming (outcome expectancy). Another way an individual can learn about a potential contingency between a behavior and a reinforcer is to observe someone else being reinforced for an activity; this is known as vicarious reinforcement.

4. Motivation and Individual Differences

Most general behavior theories argue there is a link between reinforcement and emotion, for example, Staats’ (1996) A-R-D model, which proposes that stimuli serving as reinforcers (R) have simultaneous attitudinal (A) value (can elicit emotion) and discriminative stimulus (D) functions (can serve as a cue for approach). As active approach behavior and reward systems are linked in terms of neural processes, there are plausible biological mechanism for personality differences in the degree to which a given individual is responsive to positive outcomes. Relative sensitivity to reward over punishment may also be a characteristic of individuals exhibiting certain clinical syndromes, such as Attention Deficit Hyperactivity Disorder (ADHD), conduct disorder, and recent onset schizophrenia. There are also individual differences in preference for a small but immediate reinforcer as opposed to a larger, but more delayed reinforcer, a pattern that is generally considered to be maladaptive and detrimental to individual success. Humans are more influenced by reinforcers they have chosen than those arbitrarily selected by someone else. When environmental conditions deprive an organism of the opportunity to perform a particular behavior that would be engaged in a free choice situation, it will perform other activities in order to gain access to that behavior, a phenomenon sometimes referred to as the Premack (1965) principle.

Where reinforcement is artificial, such as a socially determined prize for a job well done, humans learn from prior experience the standard of performance

that deserves to be rewarded. Even young children are unlikely to appreciate a reward given for performance that is below their own standard. Thus, although self-reinforcement is often included as a treatment component for enhancing self-control, it is unlikely to be influential over and above the intrinsic satisfaction of accomplishing one’s own goals. Furthermore, if the outcomes achieved after much effortful behavior are less positive than expected, people tend to reevaluate the outcome in an upward direction (Lawrence and Festinger 1962).

5. Reinforcement as Behavioral Regulation

Much of current reinforcement theory in the operant tradition is concerned not with understanding the motivational features of reinforcement, but with predicting the effect on the distribution of available activities of different conditions of reinforcement. Neobehavioral theories that relate reinforcement to motivation (e.g., need reduction), have given way to economic-type theories that consider the sum total of potential behaviors in a situation as well as the sum total of reinforcements. One of the most important observations underlying this approach is that if there is more than one operant being reinforced (concurrent operant schedules), animals will distribute their responses more or less in proportion to the amount of reinforcement available to each one. This is known as Herrnstein’s matching law (see Noll 1995 for a discussion).

In answer to the apparent anomalies discussed in Section 3, reinforcement is depicted as a homeostatic principle, referring to how an organism must adjust its actions to meet the demands of the environment. An organism will normally distribute its activities so as to minimize the demands (e.g., effortfulness of the response) and maximize opportunities, and this optimum—when there is unconstrained choice—is known as the behavioral ‘bliss point.’ Once reinforcement is seen in this light, it gives credence to applied approaches in which there is an emphasis on enhancing choices and opportunities for individuals in otherwise deprived environments, or those with limited responses in their repertoire.

Theories emphasizing behavioral regulation propose that contingencies serve to constrain the organism’s free flow of behavior. Given the wide range of behavioral choices available to individuals in natural situations, it is logical that removing a reinforcement for one behavior will not be successful in reducing this behavior unless another, more socially desirable, behavior is able to be reinforced. Clinical research has repeatedly demonstrated the value of reinforcing more appropriate alternatives. Since the systems or economic model emphasizes that increases in one behavior must inevitably be accompanied by decreases in

others, extinguishing undesirable behavior and reinforcing appropriate responses may be two sides of the same coin.

6. Conclusions

While context/behavior/outcome contingencies are essential elements in the shaping and development of new skills, the reinforcing event itself has complex influences on human behavior and often serves as a cue initiating other activities. The effects of reinforcement are relative—related to the levels of reinforcement for other possible behaviors—which has important implications for designing environments that will support socially desirable activities. That there are motivational and emotional properties of reinforcement should not be underestimated, such as experiencing success, the relief of reduction in anxiety, or the satisfaction of a craving. There are also cognitive properties, with the history of availability being judged by the individual and influencing expectancies regarding the relative outcomes of different actions. Past experience with certain types of contingencies and deprivation experiences influence the way an individual will respond to a given reinforcement contingency. The relatively simple idea that all behavior has a function—achieves some outcome for an organism—and that function regulates or controls the behavior, is a critical insight for analyzing behaviors that one wishes to change. However, the reverse of this principle—that we can artificially manipulate behavior by managing its consequences—has proved more complex than it appears, partly perhaps because humans at least, resist the experience that they are being deliberately controlled.

See also: Classical Conditioning and Clinical Psychology; Classical Conditioning, Neural Basis of; Conditioning and Habit Formation, Psychology of; Operant Conditioning and Clinical Psychology; Reinforcement: Neurochemical Substrates

Bibliography

- Bandura A 1977 Self-efficacy: Toward a unifying theory of behavior change. *Psychology Review* **84**: 191–215
- Carton J S, Nowicki Jr S 1998 Should behavior therapists stop using reinforcement? A reexamination of the undermining effect of reinforcement on intrinsic motivation. *Behavior Therapy* **29**: 65–86
- Hull C L 1943 *Principles of Behavior: An Introduction to Behavior Theory*. Appleton-Century-Crofts, New York
- Lawrence D H, Festinger L 1962 *Deterrents and Reinforcement: The Psychology of Insufficient Reward*. Stanford University Press, Stanford, CA
- Noll J P 1995 The matching law as a theory of choice in behavior therapy. In: O'Donohue W, Krasner L (eds.) *Theories of Behavior Therapy: Exploring Behavior Change*. American Psychological Association, Washington, DC

- Premack D 1965 Reinforcement theory. In: Levine D (ed.) *Nebraska Symposium on Motivation*. University of Nebraska Press, Lincoln, NE, Vol. 13, pp. 123–80
- Skinner B F 1969 *Contingencies of Reinforcement: A Theoretical Analysis*. Appleton-Century-Crofts, New York
- Staats A W 1996 *Behavior and Personality: Psychological Behaviorism*. Springer, New York
- Thorndike E L 1927 The law of effect. *American Journal of Psychology* **39**: 212–22

I. M. Evans

Relapse Prevention Training

Many psychological and psychiatric problems are characterized by a temporal pattern in which symptoms are either chronic or vary in severity over a long period of time, with oscillations between remissions and return of symptoms. Substance-related disorders such as alcohol dependence are often relatively enduring, and mood disorders typically follow the second pattern of alternating symptom presence and remission. The empirical demonstration of the efficacy of psychotherapeutic interventions requires substantiation of the long-term benefits of therapies for such disorders (see *Psychological Treatments, Empirically Supported*). Accordingly, there has been a growth of interest in the development of therapeutic approaches that are designed to minimize relapse (or, conversely, maximize the maintenance of therapeutic gains) over extended periods of time. Relapse prevention refers to a set of procedures that designed to obtain long-lasting benefits. Marlatt and Gordon (1985) published the seminal work in the area of relapse prevention, in which they described a set of relapse prevention strategies for excessive alcohol consumption and other addictive disorders. This approach applies a cognitive-behavioral perspective to the problem of relapse and involves the employment of a specific set of treatment techniques. Other researchers have focused on relapse prevention for a diverse range of problems, adapting the original approach accordingly. Empirical support for the approach has been mixed, but there are some encouraging findings in relation to some specific problems. Methodological issues in the research place limitations on the conclusions that can be reached at this stage.

1. Conceptual Basis of Relapse Prevention

The term Relapse Prevention (RP) has been associated with a particular set of techniques as described by Marlatt and Gordon (1985). It can be argued that any approach which has as its clear goal the minimization of relapse (or enhancement of maintenance) might be construed as RP. Few approaches other than Marlatt and Gordon's RP, however, have been embedded in a

coherent theoretical framework that is intended to be generalizable across a variety of psychological or psychiatric problems. Cognitive and behavioral theories provide both the framework for the development of RP techniques and the conceptual basis for understanding why a particular person relapses and how best the problem might be ameliorated (see *Behavior Therapy: Psychological Perspectives; Cognitive Therapy*). Specifically, problem behaviors are viewed as arising from the influence of both learning principles (classical and operant conditioning) and cognitive appraisal processes (attributions, predictions, and other judgments). For example, a person's alcohol consumption is regarded as being (a) elicited by environmental cues (presence of stimuli associated with drinking such as a glass, bottle, certain companions, time of day) and (b) reinforced by its consequences (stress reduction or mood enhancement). Apart from environmental cues, a considerable amount of attention is given the role of cognitive factors in RP. Cognitive factors may include, amongst others, causal attributions or outcome expectations. Marlatt and Gordon's RP approach draws heavily upon the concept of self-efficacy (Bandura 1977) in its theoretical foundation (see *Self-efficacy*). Thus, methods which enhance a person's confidence that they will be able to perform the necessary behaviors and the belief that their performance will lead to an effective outcome, are thought to be the ones which are most likely to prevent relapse. A good deal of the research literature on predictors of relapse has focused on self-efficacy. For example, low self-efficacy about the performance of particular behaviors has been found to be predictive of relapse across several problem areas such as alcohol use, obesity, depression, and smoking. The underlying philosophy of RP emphasizes the notion that the maintenance period is an opportunity for the practice of newly acquired self-control strategies. Any 'lapses' that occur are viewed as being useful to provide information about the ways in which people identify, contribute to the occurrence of, or respond to, high risk situations. A 'lapse' may involve a return to the problem behavior on a single, specific occasion (or a small number of occasions), but a 'relapse' involves a repeated accumulation of lapses that approximate the level of severity of the original problem. From this standpoint, 'lapses' are seen as providing positive experience from which to draw important information that might assist in preventing future lapses, and, by implication, future relapse.

2. Relapse Prevention Strategies

In principle, RP procedures may be implemented at one of three different points in time, either as: (a) an integral part of the initial treatment program; (b) a discrete set of sessions that occur towards the end of treatment; or (c) a set of sessions that are spaced across

a period of time after the end of initial treatment. In the case of the latter approach, sessions may sometimes be more properly regarded as 'booster sessions,' especially if the predominant approach involves a continuation of the basic components of the original treatment or the use of prompts for the person to employ the techniques used in the original treatment. Most commonly, Marlatt and Gordon's RP is viewed as an integrated treatment which focuses on the concept of prevention of 'lapses' from the outset, and continues this emphasis beyond the normal period of treatment into a 'maintenance phase.' In practice, studies of the efficacy of RP differ in the precise details of the timing of the delivery of the specific components. These components include (but are not limited to) the identification of high-risk situations, preparation for any high-risk situations that have been identified, explicit rehearsal of coping skills, relapse-crisis debriefing, education (e.g., about the effects of drugs), and lifestyle modification.

The identification of high-risk situations represents an essential component of the RP approach. For example, it might be found that resumption of alcohol consumption occurs when the person has had a clash of opinion with a colleague or manager at work. A list of idiosyncratic high-risk situations can be obtained, based either on knowledge of the person's previous history or on the responses to self-report inventories of typical situations in which the person is asked to indicate (a) how often these situations occur, and (b) how difficult it is to engage in the desired behavior or not engage in the undesired behavior when such a situation arises. In the case of cigarette smoking, typical high-risk situations could be (a) having a cup of coffee; (b) sitting down to study at the desk; or (c) seeing someone else smoke. Learning to recognize that a high-risk situation is occurring or is imminent is a crucial part of the RP approach. People are taught to recognize the importance of 'apparently irrelevant behaviors,' exemplified by the gambler who chooses to drive home from work via a more scenic route (justified by the better view) when the route just happens to pass the casino. On the other hand, other high-risk situations may be more obvious, even to the individual, such as the person who travels business class by plane and is, therefore, likely to be offered free alcoholic drinks. RP is also connected with the concept of the 'abstinence violation effect.' This effect is thought to underlie the occurrence of a relapse in a person who has merely lapsed by breaking a pre-planned limit on daily alcohol or cigarette consumption. It is argued that once a person has broken the limit, they may fail to exert any further control over their behavior. RP decreases this emphasis on 'rule violation' by focusing on self-control and deriving useful information from the occurrence of any lapses.

RP involves teaching people not only to identify such situations as risky, but also to prepare themselves for reducing the likelihood of a lapse. Simulations of

high-risk situations can be used to provide opportunities to practice behavioral responses under supervision of the therapist. Many relapses are thought to occur following the occurrence of negative emotional states such as anger, depressive mood, and anxiety. Thus, RP involves the identification of such emotion-behavior linkages and the use of cognitive-behavioral strategies for the management of these negative states. Other lapses may be triggered by interpersonal conflict and may necessitate a heavier therapeutic focus on helping people to develop better communication and interpersonal problem-solving skills. In the case of addictive behaviors, there may be social pressure to engage in smoking, drinking or other drug taking. Preparation for dealing with such social pressure through assertiveness training may be incorporated into the treatment. Lifestyle modification is also employed in order to increase the frequency of engagement in pleasant, relaxing activities that may also be incompatible with the addictive behaviors.

3. Efficacy of RP

RP has been the subject of a number of experimental evaluations since it was first proposed as a treatment for alcohol and other addictive problems. Wilson (1992) reviewed the existing RP literature in the areas of alcohol problems, smoking, and obesity. Carroll (1996) also published a review of RP to a wide range of problems. Twenty-six studies, on substance use only (covering the years 1978 to 1995), were the subject of a meta-analysis conducted by Irvin et al. (1999). Even before considering the results of the empirical literature, the point needs to be made that there is surprisingly little research published on RP despite its presence as a concept in the treatment-outcome literature for about two decades. The drawing of firm conclusions is hampered by this relatively small body of literature, especially given the diverse types of psychological problems to which RP has been applied and the variety of procedural modifications that have been made in the published studies. A second cause for concern is the relatively short length of follow-up periods, especially given the explicit focus on relapse as an empirical issue. While it is likely that maintenance diminishes as time increases, and this appears to be the case (Irvin et al. 1999), it is not clear what type and level of resources need to be put into RP-based therapeutic endeavors to bring about meaningful results in the public health sense. With these, and other limitations in mind, the following tentative conclusions can be drawn.

The earliest controlled studies of RP were conducted on alcohol problems and it continues to be the main focus of attention for RP researchers. Varying degrees of success have been reported for RP in these studies. A number of methodological explanations may be offered for some failures to find differences between RP

and control groups in some studies, including low statistical power, ceiling effects associated with the success of the initial intervention, the brevity of the follow-up, and the sharing of common elements in the treatments. Irvin et al. (1999) conclude that, amongst the substance use disorders, the strongest positive effects for RP have been found for problem drinking, although even within this group of studies, the outcome variable that is associated with the best response is psychosocial adjustment rather than drinking behavior *per se*. Like problem drinking, relapse rates following smoking cessation programs are notoriously high. While some success has been achieved using booster sessions, these effects frequently are lost after the eventual termination of the booster sessions. Studies of relapse prevention strategies for smoking cessation have also yielded inconsistent results. In their meta-analysis, Irvin et al. (1999) draw attention to the relatively poor outcomes with RP in the area of smoking cessation. It should be noted that a few studies have been conducted on RP for illegal drug use. Cocaine use has been the subject of a small amount of research. As with smoking, evaluations of RP for these addictions have not been noted for their successful outcomes. Irvin et al. (1999) also conclude that the effects for these problems are weaker than those for alcohol problems.

The maintenance of weight loss following cognitive-behavioral treatment was the subject of a series of studies reported by Perri and co-workers in the 1980s (see Perri et al. 1992). In this research, behavior therapy plus relapse prevention training and contact after treatment produced the largest effects. Despite the sometimes positive outcome of longer-term treatment for obesity using adaptations of RP approaches, it is clear that the overall gains associated with treatments for obesity are often modest and that many people regain the weight loss during an initial treatment period.

Other psychological problems for which an attempt has been made to develop and evaluate RP approaches include depression (Wilson 1992) and sexual offending (Laws 1989). Although it is well established that relapse following an episode of depression is as high as 50 percent over a two-year period, there has been little systematic exploration of specific, psychologically informed RP methods for this problem. It is possible that focus on RP for some types of problems such as depression is misdirected because the major impact on maintenance may be derived from the potency of the initial cognitive-behavioral intervention.

4. Methodological Issues

There are several methodological problems in the study of RP and its effectiveness. RP clearly is concerned with long-term maintenance, but the relatively brief length of follow-up periods in many studies

places some limitations on the conclusions that can be drawn. The RP approach contains a number of specific components, but these components are not necessarily employed in all studies and it is not clear which components are necessary for the best outcome. In many studies, RP procedures have been mixed into a broader treatment, making it difficult to separate the effects of RP from those of other components. With some problems, such as depression, it would be preferable for research on RP to be conducted using samples of people who are at high risk for a subsequent episode rather than for all people who complete an initial intervention. In some studies, especially when the RP is being delivered following the completion of an initial broad-based treatment approach (e.g., cognitive-behavior therapy for depression or anxiety), there is a need to separate initial treatment responders from nonresponders. There is little point in examining the impact of a post-treatment RP on nonresponders to an initial treatment program because they have yet to demonstrate responsiveness to treatment. These methodological issues pose particularly difficult problems for which the solutions are generally costly and time consuming to implement.

5. Conclusion

Despite nearly two decades of research on RP, the research produced has been quite meager and many questions remain to be addressed or resolved. The studies of RP for any given problem are so few in number that it is difficult to draw any firm conclusions at this stage. Scattered through the literature, there is some encouraging support for RP in relation to alcohol-related problems, but strong conclusions are limited due to methodological shortcomings and the relative scarcity of studies. There is a need for a large-scale evaluation of RP with addictive disorders. There has also been some progress on the development of techniques for other problems such as obesity, depression, sexual offending, and health problems.

The field of RP research may benefit from greater attention to improving our understanding of the specific ways in which RP affects basic psychological processes (i.e., investigations of the mechanism of change). Further research is also needed on both general and problem-specific predictors of relapse for particular types of symptoms so that the RP approaches can be tailored more specifically to suit the features of specific disorders. The timing of the implementation of RP procedures also needs to be considered in relation to knowledge about the temporal patterns of symptom presence and remission that are characteristic of different disorders. Continued efforts to refine and develop novel variation on RP approaches to chronic mental health and physical problems should remain a high priority for research.

See also: Behavior Therapy: Psychiatric Aspects; Behavior Therapy: Psychological Perspectives; Cognitive Therapy; Psychological Treatment, Effectiveness of; Psychotherapy Integration; Psychotherapy Process Research

Bibliography

- Bandura A 1977 Self-efficacy: Toward a unifying theory of behavioral change. *Psychological Review* **84**: 191–215
- Carroll K M 1996 Relapse prevention as a psychosocial treatment: A review of controlled clinical trials. *Experimental and Clinical Psychopharmacology* **4**: 46–54
- Irvin J E, Bowers C A, Dunn M E, Wang M C 1999 Efficacy of relapse prevention: A meta-analytic review. *Journal of Consulting and Clinical Psychology* **67**: 563–70
- Laws D R (ed.) 1989 *Relapse Prevention with Sex Offenders*. Guilford Press, New York
- Marlatt G A, Gordon J R 1985 *Relapse Prevention: Maintenance Strategies in the Treatment of Addictive Behaviors*. Guilford Press, New York
- Perri M G, Nezu A M, Viegner B J 1992 *Improving the Long-term Management of Obesity: Theory, Research, and Clinical Guidelines*. Wiley, New York
- Wilson P H (ed.) 1992 *Principles and Practice of Relapse Prevention*. Guilford Press, New York

P. H. Wilson

Relative Clauses

A relative clause (RC) can be defined as a subordinate clause which ‘limits the potential reference’ of a noun phrase (NP) in its containing clause, by describing a situation in which the referent of the NP participates. In the following examples, the RCs are enclosed in square brackets:

- (1) (a) The woman [who we met yesterday] owns the local supermarket.
- (b) We saw a lizard [which had a bright blue tongue].
- (c) The developer sent an update to every user [who registered the program].

The phrase ‘limits the potential reference’ needs some further discussion. In (a), the subject of the sentence is a definite NP which is presumed to refer to a single individual, who is being said to own the supermarket. The RC provides a substantial portion of the information needed to identify that individual (that we met this person yesterday), the ‘head’ noun *woman* providing the rest (that the person is an adult female). In (b) on the other hand there is no presumption of unique identification, but the content of the relative clause, ‘having a bright blue tongue,’ must be true of the object NP, so is limiting its potential reference. Example (c) is the most complex. Here, the NP that the

RC is delimiting doesn't refer to any specific individual, but we can still say that it is delimiting the reference of the NP, because the quantifier *every* indicates that if an entity satisfies the conditions indicated by the RC and the noun *user* (is a potential referent of the NP), then the developer sent that entity an update (see *Quantifiers, in Linguistics*).

A problem for defining RCs is provided by so-called 'Nonrestrictive relatives' (NRCs) such as this:

- (2) Lois was introduced to Clark, [who said that he had read all her books].

Here the bracketed clause is not really limiting the reference of the NP of *Clark*, since this is a proper name and presumed to identify an individual on its own. In some languages, such as English, NRCs show significant differences from ordinary RCs, while in others, such as Japanese, they don't (Kuno 1973, p. 233, Andrews 1985, pp. 28–9). It is not entirely clear whether NRCs should be regarded as a kind of RC or as a closely related construction.

1. Essential Structure

Since an RC delimits the reference of an NP by specifying the role taken by the referent of the NP in a situation, it needs to contain an internal position corresponding to the role of that NP, which can be called the NP_{rel} position. Many but not all contemporary syntactic theories postulate that the NP_{rel} position is occupied by an NP, which may be unpronounced, or moved from the normal location for an NP with that semantic role. In the RC (1a), for example, the NP_{rel} would be described as having the direct object role, and identified as the word *who* at the beginning of the RC, preposed from its usual post-verbal location.

From the viewpoint of formal semantics (see *Montague Grammar*), one can propose that RCs have the logical form of λ -abstracts on a variable in the clause, occupying the NP_{rel} position. For example, the logical form of *who we met yesterday* would be along the lines of:

$(\lambda_x) (\text{we met } x \text{ yesterday}).$

Such an expression is taken to denote the set of all entities such that if x refers to that entity, the sentence 'we met x yesterday' is true. If RCs are defined as clauses with this logical form, then NRCs can be treated comfortably as a type of RC.

2. Typology

The typology of RCs across languages can be organized in terms of two major parameters involving the position of the relative clause, and the treatment of

position of NP_{rel} within it, a variety of minor parameters involving other aspects of the form of RCs, and the extremely important topic of Accessibility, which will be deferred to the section on theoretical significance below.

2.1 Major Parameters

The first major parameter is the relationship of the RC to the NP whose reference it delimits. There appear to be five possibilities, falling into two major groups:

- (3) (a) The clause forms a constituent with the NP it modifies:
- (i) positioned after the head N of that NP,
 - (ii) positioned before the head N of that NP, and
 - (iii) appearing in an NP with no head N (alternatively formulated by saying that the head appears internally to the RC).
- (b) The clause doesn't form a constituent with the modified NP, but is a separate constituent:
- (i) at the front of the clause containing that NP, and
 - (ii) at the end of the clause containing that NP.

Types 3(a)(i) and 3(b)(i) can be illustrated clearly from English; all the examples above are 3(a)(i), while 3(b)(i) is illustrated by so-called 'extraposed' relatives like this:

- (4) A person was standing on the corner [who had a cockatoo on his shoulder].

Japanese provides perhaps the best-known examples of 3(a)(ii), which is widely found in verb-final languages, and sometimes in verb-medial languages such as Chinese

- (5) (a) Japanese (Andrews 1985, p. 27):
 Kore wa [ano hito ga kai-ta]
 this THEME that person NOM write-PAST
 hon desu.
 book is
 This is a/the book which that person has written.
- (b) Chinese (Li and Thomson 1981, p. 116):
 [Zhāngsān mǎi de] qìchē hěn guì.
 Zhangsan buy NOM car very expensive
 The car that Zhangsan bought was very expensive.

Example 3(c)(iii) is more exotic, but is commonly found in Native American languages. 6(a) is an example from Ancash Quechua (Cole 1987 p. 279), while 6(b) below is the corresponding 3(a)(ii) construction, also available in Ancash:

- (6) (a) [Nuna bestya-ta ranti-shqa-n]
 man horse-ACC buy-PERFECT-3
 alli bestya-m ka-rqo-n.
 good horse-EVID be-PAST-3

- (b) [Nuna ranti-shqa-n] bestya
man buy-PERFECT-3 horse (NOM)
alli bestya-m ka-rqo-n.
good horse-EVID be-PAST-3

The horse the man bought was a good horse.

In 6(a), *bestya-ta* 'horse' is playing the same basic function as the head N in an English RC, or in the 6(b) example, but both its position and its accusative case-marking are those appropriate to its function as NP_{rel}. Such constructions are often called 'head internal' RCs.

Finally, type 3(b)(i) is especially characteristic of (Indo Aryan) South Asian languages such as Hindi and Marathi (Srivastav 1991), although it is also found elsewhere, such as in Aboriginal Australia (Hale 1976):

- (7) (a) [Jo laRkii khaRii hai] vo lambii hai.
REL girl standing is DEM tall is
The girl who is standing is tall.

- (b) [Nyuntulu-rlu kutja-npa wawirri
you-ERG RC-you(SUBJ) kangaroo
pantu-rnu], ngatyulu-rlu
spear-PAST, I-ERG

kapi-rna purra-mi.
FUT-I(SUBJ) cook-NONPAST

I will cook the kangaroo that you speared.

These structures are often called 'correlative clauses,' and one can question whether they should be seen as a type of RC or a closely related construction.

Srivastav, for example, argues that they are in fact quantifying expressions which include reference delimitation in their function, rather than mere reference delimiters.

The second major parameter of variation in RCs lies in the treatment of NP_{rel}. There are three important and partially independent parameters of variation:

- (8) (a) Reduction of NP_{rel}:
(i) no reduction (NP_{rel} remains a full NP),
(ii) pronominalization (NP_{rel} is expressed as a pronoun), and
(iii) omission (NP_{rel} is not expressed).
(b) Marking of NP_{rel}:
(i) no special marking (NP_{rel} looks like an ordinary NP of a form that also appears in main clauses), or
(ii) a special 'relative' form is used for pronoun or determine of NP_{rel}.
(c) Change of position of NP_{rel}:
(i) remain *in situ*, or
(ii) move (usually to the front).

In both the internally headed RC of 6(a) and the correlatives of (7), NP_{rel} is unreduced. In standard English, NP_{rel} is either deleted or expressed as a special pronoun, which is furthermore moved to the front, perhaps accompanied by some surrounding material

- (9) (a) The people [we talked with *e*] were interesting.
(deletion from position marked with *e*)
(b) The people [who(m) we talked with *e*] were interesting.

(movement of relative pronoun *who(m)* from *e*-position), and

- (c) the people [with whom we talked *e*] were interesting.

(movement of relative pronoun together with containing PP).

In modern Hebrew (Borer 1984, p. 220), NP_{rel} can be omitted or reduced to an ordinary pronoun, which can either remain *in situ* or move to the front:

- (10) (a) Ra?iti ?et ha-yeled she- rina ?ohevet (?oto).
saw-I ACC the-boy that Rina loves (him)

- (b) Ra?iti ?et ha-yeled (she-) ?oto rina
saw-I ACC the-boy that him Rina

?ohevet.

loves

I saw the boy that Rina loves.

Pronominal NP_{rel} that remain *in situ* are often called 'resumptive pronouns.'

2.2 Minor Parameters

These are numerous; only two will be discussed here, clause introductory particles, and nominalization.

There may or may not be a clause introductory particle, which can appear instead of or in addition to a preposed NP_{rel}. In Modern Hebrew as illustrated in (10) above, the introductory particle is either *she*- (cliticized onto the next word) or *?asher*, and it is optional if NP_{rel} is preposed. In modern English, there is an RC-introducing word *that* which is optional, but can't co-occur with proposed NP_{rel}:

- (11) (a) The person (that) I saw was tall.
(b) The person who I saw was tall.
(c) *The person who that I saw was tall.
(d) *The person that who I saw was tall.

A somewhat more subtle form of variation is in clause-reduction. In the examples we have seen so far, the RCs are full-fledged finite clauses, but it is possible for them to become reduced. In Turkish, for example, when NP_{rel} is neither the subject nor contained in the subject (this is an approximation; see Underhill 1972 for further discussion), the subject of the relative clause is put into the genitive case, and the verb of the relative clause takes a nominalizing affix followed by cross-referencing as is normally applied to possessed nouns. The relative clause of 12(a) thus bears a considerable resemblance to the possessed noun phrase of 12(b):

- (12) (a) [oğlan-in git-tiğ-i] mektep
boy-GEN go-NOM-3SG school
The school that the boy goes to

- (b) Ahmed-in şapka-si
Ahmed-GEN hat-3SG
Ahmed's hat.

Lehmann (1986) observes that such reduction is especially characteristic of structures of type 3(a)(ii) (RC preceding head with NP), and is apparently absent for correlatives (type 3(b)(i)).

Many further potential features of relative clauses are discussed in typological surveys such as Downing (1978), Keenan (1985), Andrews (1985), and Lehmann (1986), and in the papers in Peranteau et al. (1972).

3. Theoretical Significance

The theoretical significance of RCs has derived largely from the study of 'accessibility,' the possibilities for the role of 'underlying position' of NP_{rel} within the RC. Historically, this seems to have been divided into two largely independent investigations. Ross (1967) (following on from earlier work by Chomsky) studied systematically what might be called the 'long distance' aspect of the function of NP_{rel}, the capacity of NP_{rel} to be located in a subordinate clause structure within the relative clause itself. Later, Keenan and Comrie (1977) investigated the question of what kinds of functions NP_{rel} can bear in a monoclausal RC.

The long-distance aspect of accessibility revolves around the ability of NP_{rel} to occur, and be moved, specially marked, or deleted in a wide range of environments, including from arbitrarily deep sub-clauses of the RC (NP_{rel} position marked with *e*):

- (13) (a) the book (that) Mary bought *e*,
 (b) the book (that) Mary thinks Jane discussed *e*,
 (c) the person (that/who) Mary thinks Jane spoke with *e*,
 (d) the book that Mary thinks that Susan says that Sam borrowed *e*.

This carries implications about the kind of grammatical theory required to describe languages; simple phrase-structure rules in particular are clearly not adequate, some kind of augmentation being required (in early generative grammar, it was assumed that grammatical transformations were required; subsequently, a wide range of alternatives was found).

Most important is the fact that in spite of the long distance possibilities for NP_{rel}, its positioning is, nonetheless, limited by principles which tend to be similar across languages. In most languages that allow movement or deletion out of subordinate clauses, for example, this cannot happen from clauses inside NPs (the complex NP constraint):

- (14) *the book that Mary believes the claim that Jane discussed *e*.

These constraints have been especially important for linguistic theory because:

- (15) (a) They apply to more than just relativization, but rather a wide range of apparent deletion and movement phenomena.
 (b) They appear in similar forms in many different languages (in both Irish and Hebrew, for example, the Complex NP constraint is observed when NP_{rel} is moved or deleted, but not when it is expressed as a resumptive pronoun).

- (c) It is not clear how they could be learned on the basis of direct evidence of the kind presumably available to children learning their native language.

They therefore raise fundamental questions about universal grammar and what it is that makes human languages learnable by children (see *Movement Theory and Constraints in Syntax*).

One of the more striking discoveries is that long-distance relationships appear to be composed of shorter ones; this was first proposed on theoretical grounds by Chomsky (1973), direct empirical evidence appearing in McCloskey (1977), where it was shown that in Modern Irish, clause introductory particles assumed a special form if they were in an RC, and contained that RC's NP_{rel} (McCloskey 1977, p.17):

- (16) (a) Meas mé gur thuig mé an t-úrscéal.
 thought I that understood I the novel
 I thought I understood the novel.
 (b) an t-úrscéal a mheas mé a
 the novel REL thought I REL
 thuig mé
 understood I
 the novel I thought I understood.

In 16(a) we see the normal form *gur* of the particle introducing a complement of the verb *think*, but in 16(b), when the object of this complement is a (deleted) NP_{rel}, the form *a* is used instead. Similar 'intermediate position' effects were discovered in other languages and collected in Zaenen (1983).

The other aspect of accessibility is the function of NP_{rel} in monoclausal RCs. Keenan and Comrie (1977) proposed that there was an accessibility hierarchy:

- (17) Accessibility hierarchy
 SU > DO > IO > OBL > GEN > OCOMP

such that any RC strategy would apply to NP_{rel} whose functions were a continuous segment of the hierarchy, and languages could always relativize subjects. In Modern Hebrew, for example, SU (subject) and DO (direct object) can be relativized by the deletion strategy, while pronouns (*in situ* or moved) can only be used when NP_{rel} is DO or below.

The validity of the hierarchy corroborates the universal significance of the grammatical relation concepts used in formulating it (see *Grammatical Relations*), and raises significant issues about their nature. For example, Fox (1987) observed that, in spoken discourse, it was relatively rare for transitive subjects to be NP_{rel} (10 percent), most NP_{rel} being either objects (45 percent) or intransitive subjects (45 percent). The high frequency of relativizing objects seems discrepant with a tendency for relativization to be restricted to subjects.

A solution is afforded by a reinterpretation of grammatical relations, in which the traditional 'subject' concept is split into one concept, associated with the semantic role of 'Agent,' and another with the semantic role of 'Topic.' This idea goes back to early work by Schachter (1997), in which relative clauses in

Tagalog constitute some of the evidence, and has been incorporated into many contemporary syntactic frameworks (see Manning 1996, Van Valin and LaPolla 1997 for recent discussion). Fox's observations then indicate that the grammatical relation at the top of the accessibility hierarchy should not be identified with the traditional concept of subject, but with the topic-oriented subject of this more recent work.

Bibliography

- Andrews A D 1985 *Studies in the Syntax of Relative and Comparative Clauses*. Garland Press, New York
- Borer H 1984 Restrictive relatives in Modern Hebrew. *Natural Language and Linguistic Theory* 2: 219–60
- Chomsky N 1973 Conditions on transformations. In: Anderson S R, Kiparsky P (eds.) *A Festschrift for Morris Halle*. Holt, Rinehart and Winston, New York, pp. 232–86
- Cole P 1987 The structure of internally headed relative clauses. *Natural Language and Linguistic Theory* 277: 302
- Downing B T 1978 Some universals of relative clause structure. In: Greenberg J H (ed.) *Universals of Human Language*. Stanford University Press, Stanford, CA, Vol. 4, pp. 357–418
- Fox B A 1987 The noun phrase accessibility hierarchy reinterpreted: Subject primary or the absolutive hypothesis. *Learning* 63: 856–70
- Hale K L 1976 The adjusted relative clause in Australia. In: Dixon R M W (ed.) *Grammatical Categories in Australian Languages*. Australian Institute of Aboriginal Studies, Canberra, pp. 78–105
- Keenan E L 1985 Relative clauses. In: Shopen T (ed.) *Language Typology and Syntactic Description, Vol. II: Complex Constructions*. Cambridge University Press, Cambridge, UK, pp. 141–70
- Keenan E L, Comrie B 1977 Noun phrase accessibility and universal grammar. *Linguistic Inquiry* 8: 63–99
- Kuno S 1973 *The Structure of the Japanese Language*. MIT Press, Cambridge, MA
- Lehmann C 1986 On the typology of relative clauses. *Linguistics* 24: 663–81
- Li C N, Thompson S A 1981 *Mandarin Chinese: A Functional Reference Grammar*. University of California Press, Berkeley, CA
- Manning C D 1996 *Ergativity: Argument Structure and Grammatical Relations*. CSLI Publications, Stanford, CA
- McCloskey J 1977 *Transformational Syntax and Model Theoretic Semantics*. Reidel, Dordrecht
- Peranteau P M, Levi J N, Phares G C (eds.) 1972 *The Chicago Witch Hunt: Papers from the Relative Clause Festival*. Chicago Linguistic Society, Chicago
- Ross J R 1967 *Constraints on Variables in Syntax*. Ph.D. thesis, MIT
- Schachter P 1977 The subject in Philippine languages: Topic, actor, actor-topic, or none of the above. In: Li C N (ed.) *Subject and Topic*. Academic Press, New York, pp. 491–518
- Srivastav V 1991 The syntax and semantics of correlatives. *Natural Language and Linguistic Theory* 9: 637–86
- Underhill R 1972 Turkish particles. *Linguistic Inquiry* 2: 87–100

- Van Valin R D, LaPolla R J 1997 *Syntax: Structure, Meaning and Function*. Cambridge University Press, Cambridge, UK
- Zaenen A 1983 On syntactic binding. *Linguistic Inquiry* 14: 469–504

A. D. Andrews

Relativism: Cognitive

Cognitive (or epistemological) relativism is the view that every belief is bound to some individual, group, place, or time—whence there are no objective, cross-cultural, and universal truths. According to it, what is true for us may be false for them. Hence, all beliefs would be equivalent. Whence there would be no point in searching for objective truths. Consequently, there would be no such thing as objective advancement of knowledge. Under cover of the search for truth there would only be struggle for power. Therefore, the task of the relativist would be to unmask ('deconstruct') scientific discourse and expose the powers lurking behind it. How this task can be accomplished without finding out what is really, truly the case, is not being explained.

1. Relations to Skepticism, Constructivism, and Postmodernism

1.1 Relativism is an Aspect of Skepticism

Cognitive relativism is an aspect of systematic (as opposed to methodological) skepticism. This is the view that truth is unattainable, whence 'anything goes.' It is part of anthropological relativism, and a frequent concomitant of axiological and moral relativism ('All value systems and moral codes are equivalent'). Relativism opposes absolutism, the view that all verities are both universal and timeless, hence uncontroversial. It also opposes scientific realism, which holds that there are universal truths, such as 'Agency and structure are mutually dependent.' But realism contends that, unlike mathematical theorems, propositions about matters of fact may be only approximately true, yet corrigible in the light of further discussion and new evidence. Thus, scientific realism is antidogmatic and moderate, whereas relativism is neither. In small doses, skepticism whets the appetite for research; in large doses, it extinguishes such appetite.

1.2 Constructivism Entails Relativism

Constructivism can be cognitive or ontological. The former holds that concepts and hypotheses are human constructions rather than either innate ideas or the product of revelation, perception, or intuition. On-

tological constructivism claims that the knower makes the world. Cognitive psychology has confirmed the hypothesis that all concepts and hypotheses are constructed rather than given. By contrast, ontological constructivism strains credibility. Moreover, it is destructive, because it dispenses with any tests aiming at checking whether a hypothesis matches the facts it refers to. In particular, it denies the difference between reality and appearance, which may be deceiving: it bans recurrent questions of the forms ‘What is really the case?’ and ‘Which of the two hypotheses is the truer?’

Whether ontological or epistemological, constructivism comes in two varieties: individualist and social. Whereas according to the former the individual is the source of everything, according to the latter the social group, in particular the scientific community, constructs whatever it studies, even natural things such as stars. Whether individualist or collectivist, ontological constructivism implies cognitive relativism. However, cognitive relativism can be held, and is usually held, independently of ontological constructivism. After all, it is logically possible to admit the reality of the external world while claiming that it is impossible to know it objectively, hence intersubjectively. Therefore, relativism must be evaluated separately from constructivism.

1.3 Relativism is a Component of Postmodernism

What is usually called ‘postmodernism’ in social studies and in philosophy is neither a worldview nor a method: it just boils down to the rejection of the ideals of rationality and objectivity. Cognitive relativism fits in with postmodernism because it denies the superiority of rigorous reasoning over ‘weak thought,’ as well as the possibility and desirability of attaining objective knowledge. It is unfazed by hard social statistics, or even by the findings of natural science. It ignores the difference between ‘I feel cold’ and ‘The room temperature is below the freezing point,’ and *a fortiori* between ‘That government is generally perceived as corrupt,’ and ‘The audits found that the government officials are regularly bribed.’

2. The Alleged Accomplishments of Cognitive Relativism

2.1 Relativistic Epistemology

Cognitive relativism is suggested by the multiplicity of simultaneous or successive rival views about many a domain of facts. Actually, such variety only goes to prove that scientific research does not guarantee instant, complete, and definitive truth, let alone instant worldwide diffusion. Experience shows abundantly that we often obtain approximately true data and

hypotheses, which can usually be improved upon. For instance, water was first regarded as simple and continuous, later as constituted by HO molecules, and nowadays as made up of H₂O molecules—and moreover as a mixture of light and heavy water. Scientific realism accounts for such successive approximations to the exact objective truth.

The denial of the possibility of objective truth has several consequences. One is that it discourages the very search for truth. Another is that it renders the idea of error unintelligible since, by definition, error is discrepancy from truth. A third consequence of relativism is that it regards all controversies as unending unless stopped by nonscientific means, such as authority or negotiation—which invites censorship. A fourth is that, by placing magical thinking on the same footing with critical thinking, it debases a key accomplishment of modernity. By the same token, it deprives people of a key tool to diagnose deceit and resist arbitrary authority.

2.2 Relativistic Sociology of Knowledge

Constructivist–relativist sociologists of science have been known to claim to study the sociology of peptides, rather than that of peptide-research laboratories. Moreover, they contend that scientific research is a kind of struggle for power, not truth. They miss the two main generally acknowledged motivations of basic scientists, namely curiosity and peer recognition. This explains why they have not helped explain the occurrence of any of the scientific breakthroughs in recent decades, such as the birth of tectonic plate geology, molecular biology, the neurobiological study of the mind, or socioeconomics. Understanding the search for truth requires realizing that this is what scientific research is all about.

Worse yet, although the constructivist–relativists talk much about power, they overlook the main loci of social power: politics, business, and organized religion. This explains why they have not studied such important facts as the campaigns against evolutionary biology; the budget cuts in pure mathematics, sociology, and the humanities under conservative governments; the persistent overlooking of naturalist (materialist) philosophies in the philosophy curricula; or the current toleration of charlatanism in the faculties of arts alongside the continuing demand for rigor in the faculties of science, engineering, medicine, and law.

2.3 Relativistic Pedagogy

If there is no objective truth, but only opinion, the teacher’s mission is not to teach but to promote the confrontation of rival views. In this view, the teacher is not a master craftsman but a moderator, and the student is not an apprentice but a discussant. This is in

fact how many university classrooms have functioned in the US and in Europe since the mid-1960s: not as learning centers but as debating clubs. Such transformation has had two effects, one positive and the other negative. The former consists in a lessening of the dogmatism, authoritarianism, and stodginess of traditional education. However, this liberation has also deprived students of the motivation and discipline required to learn difficult ideas and procedures—among them the art of informed rational discussion. Consequently, the graduates of relativistic pedagogy have few skills to sell. The most they can hope for is a university post where they can turn students away from serious learning.

3. Intellectual and Social Roots

Skepticism has waxed and waned in different periods since antiquity. It is likely that it prospers in times of social crisis, when the established order of things is challenged, and a number of individuals lose their old faith but are not yet ready to embrace any new belief system. In this connection, it has been claimed that relativism is a perverse effect of egalitarianism and liberalism. However, the Enlightenment and its heirs endorsed reason and science, which they regarded as universal. Their targets were despotism and religious dogma. Much the same holds for the Renaissance skeptics, such as Francisco Sanches and the early modern skeptics, such as Bayle, Montaigne, and Hume. But this does not hold for the ancient skeptics, such as Sextus Empiricus, who only criticized the ignorance and petulance of the professors of their time.

The postmodern skeptics, from the constructivist-relativists to the academic feminists and the 'deep ecologists,' are of a different color. Unlike the skeptics of old, they seldom argue cogently: typically, they just pile assertion upon assertion. Some of them attack reason and science out of horror of the nuclear bomb and other misuses of science. Others seem just alienated from the more rigorous, demanding, and rewarding disciplines. Some are suspicious of science because they confuse it with positivism. Others, because they mistake physical relativity (dependence of the values of certain physical magnitudes upon the reference frame) for cognitive relativism. Still others invoke either the Marxist thesis that ideas are produced by social classes, or the subjectivism inherent in phenomenology and hermeneutics.

The current vogue of cognitive relativism is further congruent with two contemporary social and intellectual currents and attitudes: individualism and uncritical nonconformism. The radical individualists, such as the members of the so-called 'me-generation,' claim that their own views are not inferior to any others. And the emotional non-conformists do not use social science to diagnose or cure social ills: they just revolt against all the powers that be—political, econ-

omic, and cultural. They mistrust or even reject the entire culture that comes with the social order they reject. In particular, they reject mathematics, science, and technology—some, for regarding them as tools of 'late capitalism.'

4. Conclusion

Cognitive relativism, like skepticism in general, has a grain of truth, namely, the thesis that all propositions about matters of fact are in principle fallible. But, to be realistic and fertile, fallibilism must be combined with meliorism, or the thesis that error can be corrected. What counts as truth in science and technology, even as partial truth, is universalizable. If a view is only acceptable to some people, then it is not scientific but aesthetic or ideological. Beauty may be in the eye of the beholder: truth is not. Unless this be admitted, it becomes impossible to distinguish science from pseudoscience and ideology, to tell science-based technology from quackery, and humanity from tribe. Worse yet, relativism undermines democracy and the rule of law, because both require rational debate and agreement about the truth or falsity of certain claims, such as, for example, whether the family is in decline, and whether 'wars on crime' work. In short, relativism paralyzes.

See also: Collective Beliefs: Sociological Explanation; Constructivism/Constructionism: Methodology; Knowledge, Sociology of; Relativism: Philosophical Aspects; Sociology, Epistemology of; Theory: Sociological

Bibliography

- Archer M 1987 Resisting the revival of relativism. *International Sociology* 2: 235–50
- Bottomore T 1956 Some reflections on the sociology of knowledge. *British Journal of Sociology* 7: 52–8
- Boudon R 1995 *Le Juste et le Vrai*. Fayard, Paris
- Boudon R, Clavelin M (eds.) 1994 *Le Relativisme est-il Résistible?* Presses Universitaires de France, Paris
- Bunge M 1996 *Finding Philosophy in Social Science*. Yale University Press, New Haven, CT
- Bunge M 1999 *The Sociology–Philosophy Connection*. Transaction Publishers, New Brunswick, NJ
- Collins R 1998 *The Sociology of Philosophies: A Global Theory of Intellectual Change*. Harvard University Press, Cambridge, MA
- Feyerabend P K 1975 *Against Method*. Verso, London
- Gellner E 1985 *Relativism and the Social Sciences*. Cambridge University Press, Cambridge, UK
- Gross P R, Levitt N 1994 *Higher Superstition: The Academic Left and its Quarrels with Science*. Johns Hopkins University Press, Baltimore, MD
- Knorr-Cetina K, Mulkay M (eds.) 1983 *Science Observed: Perspectives on the Social Study of Science*. Sage, London
- Kuhn T S 1962 *The Structure of Scientific Revolutions*. University of Chicago Press, Chicago

- Latour B, Woolgar S 1986 *Laboratory Life: The Construction of Scientific Facts*. Princeton University Press, Princeton, NJ
- Laudan H 1990 *Science and Relativism*. University of Chicago Press, Chicago
- Ross A 1992 *Strange Weather: Culture, Science, and Technology in the Age of Limits*. Verso, London
- Siegel H 1987 *Relativism Refuted*. Reidel, Dordrecht, Germany
- Sokal A, Bricmont J 1998 *Fashionable Nonsense: Postmodern Intellectuals' Abuse of Science*. Picador, New York
- Trigger B G 1998 *Sociocultural Evolution: Calculation and Contingency*. Blackwell, Oxford

M. Bunge

Relativism: Philosophical Aspects

1. Definition

Relativism is the view that cognitive, moral, or aesthetic norms are dependent on the social or conceptual systems that underpin them and consequently a neutral standpoint is not available to us. Relativists claim that standards of truth, rationality, and ethical right and wrong vary greatly between cultures and historical epochs and that there are no universal criteria for adjudicating between them. Relativism hence stands in opposition to universalism and absolutism. Although the first expression of relativism as a philosophical thesis is found in ancient Greece, it was only in the second half of the twentieth century that relativism became a prominent feature of various intellectual fields, including philosophy, social sciences, and literary theory. In recent years, post-modernism, in particular, has become identified with relativism.

2. Motivations

The appeal of relativism arises out of certain features of our experiences of the world as well as a number of intellectual currents.

2.1 The Indexicality of our Judgements

Many of our judgements and assertions refer to events that happen at a particular time, place, and are experienced by a particular individual. Such assertions seem to be true only at the time, place, or context of their utterance. For instance, the sentence 'It's raining today' is true at the time of writing it but may be false when uttered at a different time or place. This feature of our judgements has led some to argue that the truth of all our judgements depends on, and in that sense is relative to, their time and place, and context.

This argument for relativism, despite its popularity, is not very convincing. Sentences containing indexical

expressions such as 'this,' 'here,' 'now,' 'I,' etc., are abbreviations of statements or sentences containing references to specific objects, dates, geographic locations, and named persons. The content of an indexical statement can be made explicit in such a way that its truth or falsity would arise from specifiable events occurring at a particular time, place, or context. For instance, 'It's raining here today' is equivalent to 'There is rain in Dublin on September 12, at 5:45 local time, 1999' and that sentence is true or false irrespective of when or where or by whom it was uttered.

2.2 Perspectivalism

The perspectivalism of all our reports and experiences, the impossibility of having a view from nowhere, has been used to support relativism. All our judgements, beliefs, or assertions, it is claimed, are made from a point of view, a perspective. A God's eye view or a view from nowhere is not available to us. Perspectivalism goes beyond the mere indexicality of our judgements in that even a nonindexical judgement, for instance, 'there are nine planets in the Solar system,' is a statement made from a human perspective as we cannot step out of our skin, so to speak, and experience reality as it is. In addition, relativists assume, explicitly or implicitly, that different perspectives could not converge and therefore what is true or right from one perspective may not be so from a different one and consequently the truth of a judgement is relative to a perspective or point of view we have or we adopt.

2.3 Cultural Diversity

Increasing awareness of the extent of diversity of beliefs, practices, and customs of different cultures and different historical epochs has cast doubt on the universality and objectivity of our judgements. Most significantly, the reports by anthropologists about the thinking habits of remote peoples have led to the suggestion that all normative judgements, whether cognitive or ethical, may have only a local validity. Opponents of relativism argue that the extent and scope of diversity between different cultures and individuals is often exaggerated. Beyond the apparent dissimilarities, they argue, there are many core similarities which unify all human cultures and systems of beliefs. Relativists, on the other hand, believe that Western ethnocentrism and the intellectual legacy of the Enlightenment prevent us from appreciating or even seeing the uniqueness of different cultures and modes of thought.

2.4 The Disappearance of Old Certainties

The disappearance of old certainties, both in the religious and scientific arena, has been instrumental in the popularity of relativism in recent years. The collapse of a religiously motivated cosmology which

fixed the position of individual human beings within a larger and immutable framework and provided firm foundations for their ethical outlook helped to bolster a climate that was conducive to relativistic outlooks.

In science, the discovery of the possibility of non-Euclidean geometries followed by the scientific developments at the beginning of the twentieth century eroded the confidence placed in classical geometry and physics. Einstein's theory of relativity cast doubt on Newtonian physics and deprived it from the unassailable status it had enjoyed. The new theories in the area of quantum mechanics and particle physics, Heisenberg's Principle of Uncertainty in particular, often misrepresented and misunderstood, added to the belief that universal, absolute laws are rarer than it was once imagined.

2.5 Influence of Early Philosophers

In addition to the above intellectual and socio-historical influences, contemporary relativism is heir to the legacies left by earlier philosophers. The earliest documented source of relativism in the Western intellectual tradition is Plato's account of the Sophist philosopher Protagoras and his famous dictum, 'Man is the measure of all things; of the things that are, that they are; and of the things that are not, that they are not' (Plato 1961). As in twentieth century, the relativistic tendencies in ancient Greece were fostered by the increasing contact between the Greeks and other people; the realization that other societies may arrange their social and ethical beliefs and customs in radically different ways had a great impact on Greek thinkers. Relativism was subsumed under scepticism by the Pyrrhonian sceptics who in turn influenced Michel De Montaigne (1533–92) one of the most important figures at the outset of modern philosophy. Giambattista Vico (1668–1744) and Johann Gottfried von Herder (1744–1803) paved the way for historicism, which in turn has shaped twentieth century relativism. Contemporary relativists are influenced directly by the perspectivism of Friedrich Nietzsche (1844–1900) and the historicism of Wilhelm Dilthey (1833–1911).

According to Nietzsche, all claims of knowledge are representations of reality only from a certain standpoint. Reports of so-called 'facts' are statements of interpretation which could always be supplemented or replaced by other interpretations. Language is not a means of describing what there is, Nietzsche claims, rather it imposes its own interpretation on our thinking. Thus, the very idea of truth and reality is called into question by Nietzsche. Whether Nietzsche's perspectivism should be equated with relativism is open to debate. However, his writings have become a source of inspiration for postmodernist and relativist thinkers.

Dilthey also has influenced a relativistic understanding of social and historical conditions in con-

temporary hermeneutics and social theory through his argument that each nation can be seen as a self-contained unit with its own conception of reality and system of values. Different historical periods, Dilthey maintains, produce different norms and values, each of which is taken to be unconditional or universal. We cannot gain a correct understanding of social and intellectual conditions if we do not understand or take into account the inevitable historicity of all thought.

3. Varieties of Relativism

Relativism takes many shapes and forms. One way to distinguish between different types of relativism is to consider the questions:

(a) What is it that is being relativized?

and

(b) What is it being relativized to?

A reply to the latter question allows us to distinguish between subjective relativism (or subjectivism) on the one hand and social (or historical) relativism on the other. Subjectivists or subjective relativists argue that truth and falsity of all judgements and the right and wrong of actions are relative to the beliefs and opinions of the individual thinkers and actors. The view entails that the truth and falsity of our judgements come down only to what we believe or like. This type of relativism, although quite popular, is too naive and unsophisticated to be worthy of serious philosophical consideration. Subjective relativism rules out the possibility of disagreement and the very distinction between correct and incorrect judgements, for it turns all our judgements, as long as we believe in them, into correct or true ones.

Social (historical) relativism is the claim that the truth and falsity of all judgements and the right or wrong of actions are relative to their social, historical, and cultural contexts. According to this view, we are indeed in a position to distinguish between true and false beliefs and right and wrong actions and judgements, but only within the parameters of socially given norms and beliefs.

The most influential form of social relativism in twentieth century is known as 'cultural relativism.' Cultural relativism was inspired by the work of social anthropologists who conducted their fieldwork among tribal people. Ethnographic evidence seems to suggest that various groups of people hold systems of belief which according to Western standards are false, irrational, and even self-contradictory. The Nuer's belief that twins are birds or the Azande belief in witchcraft and oracles are some of the frequently cited examples.

Anthropologists distinguish between the methodological and the normative components of cultural relativism. The methodological principle enjoins the practicing anthropologist to refrain from taking judgmental positions on the people under study in

order to attain a greater degree of objectivity. The normative consequence of cultural relativism is respect for other cultures and value systems—which can manifest itself as an adherence to multiculturalism and avoidance of ‘ethnocentrism’ in contemporary Western societies.

One important stimulus for the adoption of cultural relativism has come from the study of languages of non-Western people. The work of Benjamin Whorf—known as ‘linguistic relativity’—is the locus classicus of this approach. Whorf’s study of American Indian languages, such as that of the Hopi, led him to claim that ‘we dissect nature along lines laid down by our native languages. The categories and types that we isolate from the world of phenomena we do not find there because they stare every observer in the face; on the contrary, the world is presented in a kaleidoscopic flux of impressions which has to be organised by our minds—and this means largely by the linguistic systems in our minds. We cut nature up, organise it into concepts, and ascribe significance as we do, largely because we are parties to an agreement to organise in this way—an agreement that holds throughout our speech community and is codified in the patterns of our language’ (Whorf 1956). According to Whorf, the Hopi articulate and hence perceive the world in terms of events rather than objects, and see time not in terms of duration as English speakers do, but as relations between events. Whorf’s principle of relativity states that ‘all observers are not led by the same physical evidence to the same picture of the universe, unless their linguistic backgrounds are similar, or can in some way be calibrated’ (p. 214). We see and hear and otherwise experience very largely as we do because the language habits of our community predispose certain choices of interpretation. In this sense, the worlds in which different societies live, he argues, are distinct worlds and not the same world with different labels attached to them.

Further ethnographic and linguistic studies have not supported Whorf’s hypothesis of linguistic relativism. Nonetheless his views have had great impact on the resurgence of interest in relativism in recent decades.

The attempt to answer the question ‘what is it that is being relativized?’ gives us a different set of subdivisions of relativism. Depending on whether cognitive or ethical norms are being relativized, philosophers distinguish between cognitive and moral relativism.

3.1 Cognitive Relativism

Cognitive relativism is the view that what is true or false, rational or valid can vary from society to society and culture to culture and we have no transcultural or ahistorical criterion for adjudicating between conflicting cognitive norms. Some main subdivisions of

cognitive relativism are relativism about truth, relativism about rationality and logic, and conceptual relativism.

3.1.1 Relativism about truth. Relativism about truth is the claim that what is true for one individual, or one social group, may not be true for another and that every effort to adjudicate between context-dependent standards of truth and falsity is bound to be futile. The truth of an assertion, so the claim goes, is relative to the beliefs, attitudes, and other psychological idiosyncrasies of the individuals, or more generally and commonly to the conceptual, or cultural, background of the persons uttering them.

Relativism about truth is the core of all relativistic positions since other subdivisions of cognitive relativism, and even moral relativism, are reducible to it. For instance, relativism about rationality can be expressed in terms of whether there are true (universal) standards of rationality. Relativism about logic may be restated as the question whether logical truths are relative to specific cultures or cognitive schemes or whether they are in any sense universal. Ethical relativism can be seen as the claim that the truth of ethical judgements is relative to the context or the culture in which these particular judgements are being made. In this sense relativism about truth is the most general and strongest form of relativism and the target of most criticisms of relativism (see Sect. 5).

3.1.2 Relativism about rationality. The relativist about rationality argues that various societies or cultures have different standards and criteria of rationality and we are not in a position to evaluate, or choose, between these divergent standards. For instance, Peter Winch has argued that since standards of rationality in different societies do not always coincide, we should use only contextually and internally given criteria of rationality in our assessment of the systems of belief of an alien culture. Winch, who based his arguments on Evans-Pritchard’s account of Azande’s beliefs in witchcraft and magic and who was greatly influenced by the work of the philosopher Ludwig Wittgenstein, argues that we can only use contextually and internally given criteria of reasoning and validity in our assessment of the systems of belief of other cultures, and hence we cannot impute irrationality or illogicality to the Azande. It does not make sense to speak of a universal standard of rationality because rationality means conforming to internal norms given by the language and society, i.e., the form of life in which the language is being used.

In more recent years, the discussion of the Azande material, coupled with a strongly relativistic interpretation of Wittgenstein’s views have led Barry Barnes and David Bloor to argue that not only standards of

rationality but also laws of logic, often seen as the necessary conditions for rational thought, are defined by and hence are relative to their social context. According to them the very distinction between valid and invalid modes of argument is relative to the practices prevalent in a given linguistic community or society and cannot in any sense be seen as universal. Bloor argues that there are no culture-transcendent rules of validity and inference. Different societies may have incompatible, but internally coherent, systems of logic and the norms of logical reasoning are relative to a background of cultural conditioning. According to Bloor, 'The Azande have the same psychology as us but radically different institutions. If we relate logic to the psychology of reasoning we shall be inclined to say that they have the same logic; if we relate logic more closely to the institutional framework of thought then we shall incline to the view that the two cultures have different logics' (Bloor 1976, pp. 129–130).

3.1.3 Conceptual relativism. Conceptual relativism is motivated by philosophical considerations about the relationship between human thought and the world rather than mere observation of cultural and historical differences, and for that reason it is both quite distinct from and much more sophisticated than other varieties of cognitive relativism. The roots of conceptual relativism rest with Immanuel Kant's (1724–1804) distinction between the data of our sense experiences and the principles of organization or categories we use to organize them. Kant believed that all thinkers necessarily apply the same categories or conceptual schemes to the undifferentiated flow of their experiences. However, once the distinction between a conceptual scheme and the content of that scheme was introduced, then it was easy to accept that there may be more than one system or scheme of organization and the idea of conceptual relativism was born.

One of the most influential versions of conceptual relativism in twentieth century was proposed by W. V. O. Quine (1908–2000). According to Quine, to be able to talk about the world and to cope with our stream of sensory experience, we must impose upon the world a conceptual or linguistic scheme or theory. However, it is possible to envisage a plurality of conceptual or theoretical frameworks all of which explain and predict our experiences of the world equally well. There are no determinate facts to enable us to choose between these different frameworks, hence what Quine has called 'ontological relativity' ensues.

The influence of Quine's views is widespread. Philosophers working in differing areas have explicitly, or implicitly, made use of his arguments to support their relativistic positions. We can see the impact of Quine's ontological relativism on the writings of Nelson Goodman and Hilary Putnam who have proposed very sophisticated forms of conceptual

relativism. Putnam, for instance, has denied that there is a fact of the matter as to even what counts as an object outside the standards set by a conceptual scheme. One and the same situation, according to Putnam can be described 'as involving entirely different numbers and kinds of objects' (1992, p.120). Different conceptual schemes will contain bits that are 'true' (or 'right') and bits that will turn out to be 'wrong' by the standards appropriate to the scheme itself—but we do not have an account of truth which would be external to, and independent of, all conceptual schemes.

It is important to note that most conceptual relativists, Quine and Putnam in particular, wish to distance themselves from cultural relativism which they think is fundamentally incoherent (see Sect. 5). Their brand of relativism is based on their theories of meaning and truth rather than the purported differences between cultural traditions.

Conceptual relativism has also influenced the philosophers of science, Thomas Kuhn and Paul Feyerabend, who have emphasized the role of different modes of reasoning at various periods in the history of science and have argued that scientific theories belonging to these divergent styles of reasoning are incommensurable. Kuhn has claimed that scientists working with different paradigms live in different worlds, Feyerabend has argued that a change of universal principles brings about a change of the entire world. Speaking in this manner we no longer assume that there is an objective world that remains unaffected by our epistemic activities, except when moving within the confines of a particular point of view. Kuhn's thesis of incommensurability rested on his assumption that scientific theories cannot be compared or assessed independently of the paradigms in which they are embedded. For Kuhn, each paradigm acts as a unique and untranslatable language or conceptual framework and since rival paradigms do not have access to a common criterion for adjudication, they cannot be compared and hence remain self-contained and beyond the critical reach of other paradigms. According to this understanding, incommensurability implies that evaluation of different conceptual schemes, cultures, or ways of life is impossible.

Subsequent to the first publication of *The Structure of Scientific Revolutions* in 1962, Kuhn, unlike Feyerabend, attempted to distance himself from relativism but his views remain an important source of inspiration for relativistically inclined thinkers in many fields.

3.2 Moral Relativism

Moral inquiry is concerned with how life should be lived and how human beings should behave towards each other and their world. Moral relativism is the claim that there exist diverse, incompatible moral

systems and there are no overarching criteria to decide between them. Moral relativists start from the premise that moral values are grounded on societal conventions, historical conditions, metaphysical beliefs, etc., which vary from one society or social grouping to another and argue that there are no neutral standards available to us for adjudicating between such competing claims. The truth or falsity, the appropriateness or inappropriateness of ethical beliefs, are relative to their sociohistorical backgrounds and moral beliefs cannot be assessed independently of their social framework.

Moral relativism is usually subdivided into descriptive and ethical (or meta-ethical) relativism. Descriptive relativism purports to be an empirical doctrine based on observation of diversity in moral judgements and moral values in different societies and at different times. There is widespread disagreement among individuals and cultures on issues ranging from capital punishment to abortion, from human rights to animal rights, and so on. Descriptive relativists go so far as accepting that there may be broad similarities in the core beliefs of different societies, for instance, injunctions against the killing of innocents, but argue that such shared values are given very different and even incompatible interpretations by different cultures.

Ethical relativists claim that descriptive relativism entails that all ethical judgements are relative to the culture or context in which they are made and that there is no unique criterion for adjudicating between them. Of course moral diversity is compatible with the possibility that there are right and wrong answers to ethical questions. But this would depend on having a dependable path towards the discovery of ethical truths or the establishment of a universal framework for arbitrating between incompatible moral claims. Ethical relativists deny this possibility because they claim that any rational decision making procedure will have deeply seated normative historical and cultural presuppositions embedded in it. Reason and rationality are themselves evaluative concepts, constrained in turn by cultural and historical influences and hence are incapable of leading us to a set of neutral universally applicable standards of evaluation. Thus, ethical relativists deny the possibility of convergence between conflicting ethical frameworks.

Ethical relativism can be embraced independently of cognitive relativism. Ethical relativists may agree that convergence between diverse world views and theoretical frameworks is possible in the natural sciences but deny that this is the case in the realm of ethics. According to this view, moral precepts and judgements are not part of and do not pertain to the natural furniture of the universe. They are man-made and would not exist independently of human actions, beliefs, and customs, hence, there exists a fundamental difference between scientific investigations and moral enquiry.

Ethical relativism is often associated with tolerance and open-mindedness. It is sometimes argued that ethical relativism is itself a normative doctrine insofar as it prohibits interference with the ways of life and moral practices of other cultures. But this degree of tolerance can also lead to moral indifference and even nihilism. A relativist does not have the certitude to condemn genocide, torture, child abuse, etc., so long as it is argued that such actions are part of the fabric of the social life of some cultures and peoples. Furthermore, if ethical relativism is true then even tolerance and noninterference could not be seen as a universal ethical value and could not be recommended to those who do not already embrace it.

4. New Developments

Relativism has gained support from two recent philosophical developments. A new variety of relativism has been introduced through the work of the neo-pragmatist philosophers, the most prominent among whom is Richard Rorty. Although Rorty denies the charge of relativism he does believe that truth, rationality, and objectivity are all defined only in terms of the practices of a community. Furthermore, he argues that no external questions can be asked about the relationship between the practices of a community and a reality that exists independently of it. 'Objective truth,' then, for Rorty, is simply the best idea we currently have about how to explain things. The argument supports relativism insofar as it identifies truth with (best) views current in a society at a given time.

The work of postmodernist cultural and literary theorists has been hugely influential in establishing relativism as one of the strongest intellectual currencies of recent times. Jacques Derrida's proposed deconstruction of Western ideas of logic, truth, and rationality has been seen as a harbinger of a new intellectual era free of the logocentrism and phallogocentrism of Western rational intellectual traditions. Derrida, like Rorty, has denied that he is a relativist but his own and his fellow postmodernists' writings lend themselves easily to relativistic interpretations.

5. Arguments Against Relativism

Relativism has been challenged on several grounds.

5.1 The Self-refutation Argument

The most common criticism leveled against relativism is the charge of inconsistency. The argument goes back to Plato and the main claim is that relativism, at least in its stronger form of relativism about truth, is either

inconsistent or incoherent because of the dubious status of the claim that 'truth is relative.' It is suggested that if 'truth is relative' is itself true unconditionally, then there is at least one truth which is not relative, and hence relativism is not true. If, on the other hand, 'truth is relative' is true only relative to the relativist's framework, then the claim is devoid of significance. To avoid inconsistency a relativist should accept that 'truth is relative' is itself only relatively true, i.e. true only for people who suggest it, but then she or he would not be in a position to try and convince a non-relativist of the veracity of the claim. The relativist is condemned to silence.

5.2 *Argument from Translation*

A new type of criticism of relativism has surfaced in recent years directed at the intelligibility of the claim that one and the same statements may be true for one linguistic community and false for another. Donald Davidson, one of the main advocates of this position, argues that all types of relativism rely on the assumption that there can be radically divergent conceptual schemes or systems of belief. In order to ascertain that a system of belief or cognitive scheme diverges from ours in a significant manner, he argues, we first need to identify such a scheme. In order to identify something as a belief we must be able to interpret it (or translate the sentence expressing that belief into our language). But, to translate a sentence or to assign meaning to it, we will have to regard that sentence, or the belief system underpinning it, as true or in agreement with us. A correct understanding of the beliefs and other propositional attitudes of a person, Davidson maintains, leads to the conclusion that most of her beliefs must be true. If this argument is right, then relativism cannot even get started, because arguments for relativism are built on the premise that there are, or can be, fundamental cultural and conceptual differences between human beings. If this assumption is untenable, then so is relativism.

5.3 *Evolutionary Arguments*

The origins of this approach can be traced to Quine and his infamous dictum that natural selection guarantees that organisms would know many truths and abide by the laws of logic and rational cognitive procedures or become extinct. The philosopher of mind, Daniel Dennett, following Quine, has argued that natural selections guarantee that most of an organism's beliefs will be true and most of its strategies rational. 'True' and 'rational' have to be interpreted univocally, therefore we have empirical grounds for rejecting the possibility of relativism about truth and rationality.

Relativism has survived 2,000 years of attempts to bury it. Its perennial appeal and periodic fashionability are due to its ability to capture and express an essential insight, be it in an exaggerated fashion, about the human condition. The insight that we live in a world of pluralities and do not have access to univocal answers to our diverse problems.

See also: Ethics and Values; Evans-Pritchard, Sir Edward E (1902–73); Fieldwork in Social and Cultural Anthropology; Intentionality and Rationality: An Analytic Perspective; Irrationality: Philosophical Aspects; Kant, Immanuel (1724–1804); Kuhn, Thomas S (1922–96); Nietzsche, Friedrich (1844–1900); Paradigms in the Social Sciences; Pluralism; Realisms and their Opponents: Philosophical Aspects; Truth, Verification, Verisimilitude, and Evidence: Philosophical Aspects; Wittgenstein, Ludwig (1889–1951)

Bibliography

- Bernstein R J 1983 *Beyond Objectivism and Relativism: Science, Hermeneutics, and Praxis*. University of Pennsylvania Press, Philadelphia
- Bloor D 1976 *Knowledge and Social Imagery*. Routledge & Kegan Paul, London
- Davidson D 1984 *Inquiries into Truth and Interpretation*. Oxford University Press, Oxford, UK
- Feyerabend P 1987 *Farewell to Reason*. Verso, London
- Harman G, Thomson J J 1996 *Moral Relativism and Moral Objectivity*. Blackwell, Oxford, UK
- Harre R, Krausz M 1996 *Varieties of Relativism*. Blackwell, Oxford, UK
- Hollis M, Lukes S (eds.) 1982 *Rationality and Relativism*. MIT Press, Cambridge, MA
- Jarvis I C 1984 *Rationality and Relativism: In Search of a Philosophy and History of Anthropology*. Routledge, London
- Krausz M (ed.) 1989 *Relativism: Interpretation and Confrontation*. University of Notre Dame Press, Notre Dame, IN
- Krausz M, Meiland J W (eds.) 1982 *Relativism: Cognitive and Moral*. University of Notre Dame Press, Notre Dame, IN
- Kuhn T S 1962 *The Structure of Scientific Revolutions*. University of Chicago Press, Chicago, (2nd edn., enlarged, 1970)
- Laudan L 1990 *Science and Relativism: Some Key Controversies in the Philosophy of Science*. University of Chicago Press, Chicago
- Margolis J 1991 *The Truth About Relativism*. Blackwell, Oxford, UK
- Norris C 1997 *Against Relativism: Philosophy of Science, Deconstruction and Critical Theory*. Blackwell, Oxford, UK
- Plato 1961 *Theaetetus* [trans. Cornford F M]. In: Hamilton E, Cairns H (eds.) *The Collected Dialogues of Plato*, Bollingen Series. Pantheon Books, New York, pp. 845–919
- Putnam H 1981 *Reason, Truth and History*. Cambridge University Press, Cambridge, UK
- Putnam H 1992 *Renewing Philosophy*. Harvard University Press, Cambridge, MA
- Quine W V O 1984 Relativism and Absolutism. *Monist* 67: 293–6
- Rorty R 1991 *Objectivity, Relativism, and Truth: Philosophical Papers*. Cambridge University Press, Cambridge, UK, Vol. 1

- Siegel H 1987 *Relativism Refuted: A Critique of Contemporary Epistemological Relativism*. D. Reidel Publishing Company, Dordrecht, The Netherlands
- Whorf B L 1956 *Language, Thought and Reality*. MIT Press, Cambridge, MA
- Williams B 1985 *Ethics and the Limits of Philosophy*. Harvard University Press, Cambridge, MA
- Winch P 1964 Understanding a primitive society. *American Philosophical Quarterly* 1: 307–24
- Winch P 1958 *The Idea of a Social Science and its Relation to Philosophy*. Routledge & Kegan Paul, London
- Wong D 1991 Relativism. In: Singer P (ed.) *A Companion to Ethics*. Blackwell, Oxford, UK

M. Baghranian

Relativism, Pluralism, and Skepticism (Philosophical Perspective): Cultural Concerns

Few terms in our critical language are as ambiguous as pluralism and relativism. Under diverse definitions, their relationship could be said to range all the way from twins to deadly enemies. Both have been accused by some monists as furthering rank skepticism or nihilism. Meanwhile, defenders have vigorously denied that either term necessarily entails utter skepticism, or Pyrrhonism, or what Bernard Williams has called vulgar relativism. This article pursues pluralism as the best escape route from naive dogmatism and naive (hyperskeptical) relativism. (For a useful, though somewhat superficial, history of skepticism, especially in the Renaissance, see Popkin 1979.)

Almost every serious critic would agree that some extreme forms of either term can lapse into the kind of skepticism or nihilism that threatens genuine intellectual inquiry or achievement. Genuine inquiry is doomed if no claim to ‘truth’ has any validity except for the claimer, and no one thus has any real reason to listen to the arguments of anyone else. Fortunately, more and more postmodernists, some of whom for a while edged towards such utter skepticism, have attended to diverse cogent attacks on the ‘vulgar’ or ‘total’ forms of doubt. They have acknowledged that all arguments for total nihilism are paradoxically self-destructive; no one can offer a coherent defense of utter skepticism without violating that defense by claiming to offer, unskeptically, this one coherent defense. What is more, the would-be utter skeptics usually reveal many other deep commitments that contradict their skeptical claims. Meanwhile, many have argued that the term relativism need not be identified with nihilism (Pepper 1942, Williams 1972, Shweder 1989, Goodman 1978, Rorty 1989, Krausz 1989).

Cultural relativism—‘though truths exist, they are always relative to some cultural placement’—has been

much more successful in resisting attacks from defenders of hard truth of various stripes. Its defenders have ranged over a vast diversity of arguments. Some have moved dangerously close to utter skepticism by implying, but never stating, that whatever any given culture claims is true must be true, even if it entails mass genocide or the ritual torture and murder of the innocent. At the opposite extreme, some have labeled our ‘scientific,’ ‘rational,’ ‘antisuperstitious,’ ‘intellectual’ culture as inherently, self-evidently superior to all the naive, primitive, ‘believers’ in claims that violate scientific convictions. As Shweder describes the way too many rationalists, including professed cultural relativists, have viewed the world:

Nature is mindless, objective, and visible. It is empty of such unseen, unobservable, meta-physical things as god, sin, obligation, value, morality, and so on. If such things exist at all, they exist only as reality-positives in the mind. And if they exist only in the mind, they are not objective and thus ought not be (sic) allowed to be constraining. If human beings feel constrained by such things, it is only because, not yet realizing their essential nature (self-determination), they do not distinguish between truth and poetry, confusing, quite irrationally, external reality with what exists nowhere else but in their minds (Shweder 1989, p. 121).

Even when we follow Shweder and rightly reject such extremes and decide that truths do exist, though in multiple forms, we are still faced with the complex problem of how to deal with that multiplicity. For cultural pluralists who refuse to dodge the problem, the best possible method is to search for common ground that underlies surface differences. (Sometimes they—we—stumble upon another route, after the probing: ‘By studying my opponent carefully I have learned that I was wrong and he/she was right; now my pluralism includes memory of my previous error.’) When for example we discover that every culture, including the most rationalistic versions of our own, exhibits some form of embrace of an unprovable ‘transcendent’ ‘world’ or ‘reality’ or God, we no longer have good reason to consider religious commitments absurd: they are shared by all cultures. Again to quote Shweder 1989:

In that (our) post-Nietzschean world God is not dead; only positivism and monotheism are dead. Polytheism is alive and well. Its doctrine is the relativistic idea of multiple objective worlds [including the world of scientific rationalists], and its commandment is participation in the never-ending process of overcoming partial views (pp. 132–3).

1. Plurality of Thought-modes

Once we embrace that form of cultural pluralism, value-terms like ‘dead’ and ‘partial’ turn our problem in the direction of philosophical inquiry: how to deal with plurality not just of culture systems but of

thought systems, including value judgments. Even those 'rationalist' philosophers who are totally anti-relativist, antiskeptical, exhibit a plurality of conflicting thought-modes. Too many have faced that conflict with either passionate commitment to 'the one true system' (such as the positivism that most of us, including most professional philosophers, by now question) or the philosophical skepticism that Richard Rorty and Stanley Fish and many others have at times edged toward (even while claiming not to).

What has been too often ignored by those who justifiably attack the first of these paths, dogmatic monism, is the problem of how we are to deal with that plurality of 'contexts' or 'belief systems,' or 'structures of thought' or 'thought-modes' to which all of our assertions are 'related.' Every thinker has to admit that all thinking takes place in a conceptual context in which all assertions are interrelated, interdependent for their full understanding; even claims to factuality have different meanings in different thought-modes. Yet only careless extremists can deny that some 'contexts' are better than others: truer, or 'fuller,' or less vulnerable to refutation.

The pluralistic problem can thus be summarized like this: a devoted, honest Nazi who asserts that the commitment to the one superior race with its superior values justifies genocide, employing as evidence some 'facts' about past misbehavior of Jews and amazing achievements of Aryans like Beethoven, is asserting facts that though in one sense are context relevant, cannot be claimed by an anti-Nazi to be just plain nonfactual. But how, then, can one assert, as we all would, that the entire context of such Nazi thinking is just plain wrong, without falling into dogmatic assertions that are finally totally dependent on an equally unprovable cultural context? How finally do we avoid a form of relativism that grants the Nazis equal intellectual status?

2. Truth Exists—in Multiple Forms

The only possible answer is to develop some form of 'systems pluralism' that allows for distinctions among a plurality of defensible systems and a plurality of inferior, less defensible systems. While there is a genuine plurality of 'truths,' genuine truths that are of equal validity and systems that are universally true in the sense of being irrefutable, no one of them is universally true in the sense of being easily either translatable to other 'systems' or immune to seeming refutation from other 'truths.' But that collection of irrefutable context-systems or thought-modes does not mean that other systems or modes cannot be shown to be inferior. The Nazi illustration reveals that one particular belief system can be finally rejected, first because it will be condemned by every other major mode of thought, and second because its disastrous consequences are revealed by history, as they finally

were to Heidegger when he faced the consequences of his early Fascist thinking.

The history of philosophy, like the history of various cultural relativisms, can be seen as an implicit demonstration both that genuine truth exists, in a variety of defensible forms, and that the diverse forms are never easily intertranslated in intelligible form. Indeed, in most versions, 'believers' believe that they can, in using their own version of truth, refute the other versions (see Russell 1948, as an example of careless refutation of all modes but his own).

The clearest support for a plurality of legitimate thought-modes (along with a larger plurality of illegitimate ones) is found in the ways in which the major, great, surviving philosophies conflict, at least on the surface, and yet reveal immense capacity both to seem to refute other systems and to genuinely resist refutation from the others. In a short discussion, it will be impossible to demonstrate this case fully (for extended treatment, see Booth 1979, Garver and Buchanan 2000, Booth 2000, McKeon 1951).

3. McKeon, the Greatest Pluralist

To me by far the most penetrating of the many efforts to face plurality is that of the much-neglected Richard McKeon. As McKeon probed major thinkers throughout history, he more and more aggressively asserted that there are a small number of absolutely irrefutable (yet seemingly conflicting) thought modes, while there are many totally refutable modes and many other half-defensible but truncated modes. Having pursued every major thinker to the depths, he developed the notion of what his best defender, Walter Watson, has called 'reciprocal priority'—what might be called 'reciprocal irrefutability' (see Watson 1993, the best single introduction to reciprocal seeming refutation and actual nonrefutation; it is based on the lifetime work of Richard McKeon).

McKeon's classification of defensible modes became increasingly complex throughout his life, and it cannot be summarized here, as is evident in his final, most complicated chart, which was not published until after his death (see Fig. 1).

Such complexities may well obscure the validity of his basic project, which was to reveal how first-class philosophers, irrefutable philosophers, have approached multiple truths with valid thought-modes that differ in many conflicting ways, especially in four major matters: (a) what is the proper subject matter, or definition of subject matter ('Modes of fact') of what we should think about; (b) what 'unprovable' first principles, or notions of the 'modes of being,' should one proceed from, taken as unquestionable; (c) what methods are the proper methods ('modes of thinking') for solving all problems; and (d) what is the point of it all, what are the true goals of human thinking. In my view his philosophical multiplicities could well serve as

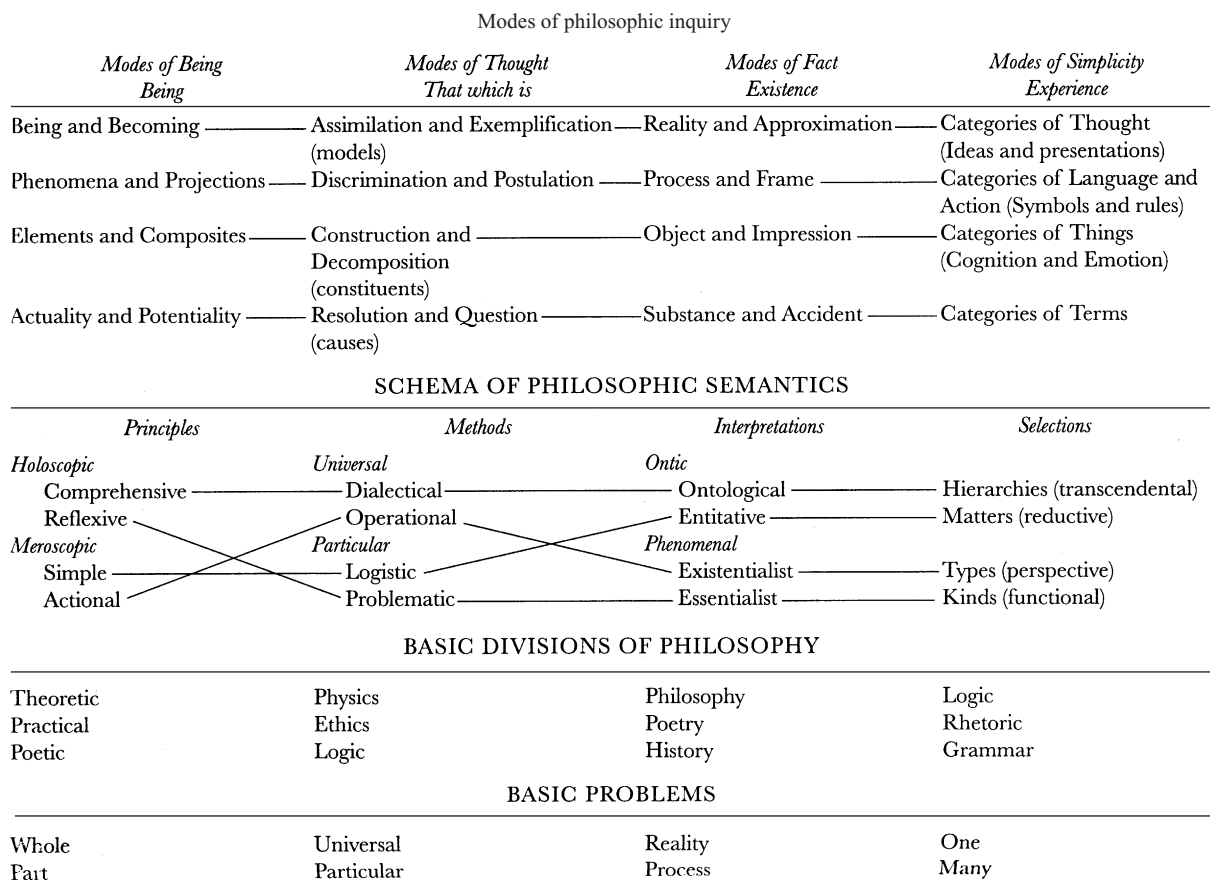


Figure 1

McKeon's final classification of defensible modes (Printed posthumously in *Selected Writings of Richard McKeon*, Vol 1, 1998, p. 218, University of Chicago Press)

a rescuer of those who fear that cultural relativism will inherently destroy such terms as truth or objectivity or reality.

3.1 Definition of the Subject Matter: 'Things,' 'Thoughts,' 'Words,' 'Actions'

3.1.1 Things. The most famous of the classical philosophers agreed that what we should pursue is reality, 'what really is,' substance, being, actuality: whether air, water, fire, atoms, or ideas. Plato always operates with a strong conviction that there are some general truths, or principles, or ideas awaiting our clumsy efforts to discover them. His definition of the subject matter (the world, the universe, the nature of things, the One) is always implicitly a world in which Truth, Goodness, and Beauty are interrelated, harmonizable: a One that can, in theory, finally harmonize it all, even though no one thinker can capture it. Many modern cosmologists begin with a surprisingly simi-

lar pursuit of a supreme truth that is 'really there'—nature and its laws: this or that final theory will explain it all. For all of these, we should begin with some picture of the real, and then move from it to whatever questions arise about the rest of life.

3.1.2 Thoughts. Again and again, as McKeon traces brilliantly, philosophers have concluded that you cannot really begin there, with Being, struggling with ontological questions, because the picture of real reality always depends on how we think about it, and that leads to a take-over by one or another form of epistemology. Follow Kant and begin your thinking by thinking about thought, not about things or substance.

3.1.3 Words. That shift from subject matter to mind has again and again broken down as philosophers concluded that how we think depends on the

language we think with, so we should begin not with thought but with language. Especially in the twentieth century, more and more thinkers took their beginnings in language. For just about everybody thought had to begin with language. So-called analytical philosophy took over, as it had done for some 'skeptics' and 'relativists' throughout the ages, for a while, and later the deconstructionists produced a seeming revolution by claiming that language itself has no real substance.

3.1.4 Actions. Finally, the linguistic revolution has repeatedly been revolted against, by those who see action as the proper subject, the place to begin. Pragmatism in various forms has been seen as the true subject of true thinkers; inevitably they work back up from language to questions about abstract reality, ontology, epistemology, and linguistic analysis, but they always place action—with success and failure of action—at the top of the definition of subject.

3.2 Principles about 'Reality'

3.2.1 Comprehensive whole. Many philosophers, like Plato and some modern cosmologists, put as top priority a conception that there is some whole, some totality, some supreme collection or unity of all truths. Reality consists of that comprehensive unity or totality.

3.2.2 Multiple problems produced by the incomprehensible whole. Others like Aristotle while still in one sense pursuing the whole, see the philosopher's basic principle as to 'pursue truth about this or that particular aspect, or problem,' concentrating on solving the problems. Many such pluralists see no hoped-for 'top' but a collection of diverse subjects and problems; there may be a 'one' uniting them, but for us the world is multiple, created by and consisting of many causes and possible solutions.

3.2.3 Least, parts: 'atoms,' 'sensations,' etc. In contrast, philosophers like David Hume and many modern scientists put as the 'top' of their reality a kind of bottom: elements or particles or epistemological certainties from which they can reason upward to larger truths. Like some particle physicists, they hope to discover, by probing the depths, truths never even dreamed of by those who pursue the 'heights,' and those truths—whether about our most minute sensations or the most minute physical particles, will be totally irrefutable from the other perspectives.

3.2.4 Operational, pragmatic problems: truth is what works. Finally, philosophers who, like Cicero and other rhetoricians, put as top priority the successful operation of the practical world, beginning with questions of just what our pragmatic goals are, can build from their pursuit truths about life that may conflict irrefutably with many of the truths that the other three pursuers have proclaimed.

To say all this does not mean that the four choices of first principles will not reveal, down the line, agreed-upon truths. Many of the undeniable truths that each mode uncovers will be considered undeniable by each of the others. For example, it is always wrong, in principle, in all of the defensible systems, to violate the moral principles that you think your system has revealed, and many of them overlap those revealed in other systems. It is always wrong, in all systems, to deliberately, dishonestly, fabricate evidence to support your system. And there are many other overlapping moral principles, such as the relation of Cicero's 'Civility' to Aristotle's 'Ethics' and Plato's 'goodness' and Hume's moral standards, and the moral commands issued by John Stuart Mill's utilitarianism.

3.3 Methods

Major thinkers have always disagreed about just what methods are effective in yielding proof. Depending in part on their first principles and their definitions of the subject matter, they have ranged from the absolutely deductive synthesizers to the absolutely inductive analysers, with some pragmatists and linguists urging some middle ground of cooperative thinking and testing, not induction or deduction but what could be called co-duction. Plato always falls back on dialectic, even when exploring rhetoric and logic. Aristotle insists on a plurality of methods to meet a plurality of subjects and problems. Consider the contrast between that and the method of Hume, or of other thinkers who begin with the assumption that the true method of inquiry is to work analytically with the indivisible parts of the universe, not some ideological whole, and then to reason analytically 'upward' from those elementary parts. Whether or not such methods lead ultimately to this or that ontology or metaphysics, it is the right way to work.

Thus, without troubling too much about whether one will ultimately find a total view of everything, analytical reductionists know that they are dealing with reality, tough hard reality of the multiple parts. The more defensible among them generally have accorded a place for diverse methods, but some partial or flawed thought-modes have tried to reduce it all to one grand method, as modern scientists too often reduce everything to empirical demonstration of only one of Aristotle's four causes: the efficient, or mechanistic, or 'horizontal' bump-and-grind reality.

3.4 Goals

The diversity of goals is to a large degree implicit in all these other varieties. The goal of thought is to improve truth, for the sake of truth. No, the goal of thought is to improve conditions in the world, for the sake of what Plato called goodness. No, the goal of thought is to pursue beauty, proper kinds of eros, religious ecstasy, some transcendent escape from shitty reality, salvation.

For the monist or dogmatist, the goal is to demonstrate that there is only one right view of principles, subject choice, methods, and goals. For the pluralist who rejects vulgar relativism or skepticism, the goal is both to pursue some one or another of the major, irrefutable modes or, even better, to pursue what is the common ground among the defensible modes, thus demonstrating their defensibility.

The suggestion, in this cursory summary of just one brand of philosophical pluralism, is that all thinkers in all subjects—philosophical, religious, anthropological, political, literary—could reduce pointless debates about truth and objectivity if they thought harder about just which definitions, principles, methods, and goals they are depending on.

4. The Plurality of Pluralisms

One unavoidable problem for anyone who chooses the pluralist route is the immense complexity of modes and the rivalry of pluralisms in dealing with the modes. The rivalries often lead many of us pluralists to retreat, as we sometimes observed McKeon doing, to some version of 3.2.4, a loose-jointed ‘operationalist,’ ‘rhetorical,’ or ‘practical’ (pragmatic) corner of the chart, with an emphasis on how to improve the world by improving thought about it. But even as we do so we have to recognize that other thinkers, in contrasting modes, will be discovering truths that we ignore.

We thus end with a plurality of pluralisms, some more defensible than others, just as cultural relativists end with a plurality, some more defensible than others. All of the pluralisms are ‘relativistic,’ in the sense that within each system every claim, even a claim of fact, is relatable and to some degree dependent on the total thought-mode. Conclusions within a given mode are testable within that mode, and will often even prove to be testable and acceptable within other modes. But no one system is totally refutable by any other system, even when all other systems agree that one system is flawed.

Thus, serious explorers of the issues ‘relevant’ here have two main choices, whenever they find that no one controversy seems resolvable: retreat to the study of some one cultural system or thought-mode, ignoring the others, or pursue the mission of finding common ground among the conflicting systems. Only this second choice can avoid dogmatism, on the one hand,

and on the other skepticism, reductionism, and nihilism, both directions producing absurd destruction of meaningful inquiry.

See also: Cultural Psychology; Cultural Relativism, Anthropology of; Pluralism; Pluralism and Toleration; Relativism: Cognitive; Relativism: Philosophical Aspects; Value Pluralism; Wisdom, Psychology of

Bibliography

- Bakhtin M M 1981 *The Dialogic Imagination: Four Essays* (ed. Holquist M. Trans. Holquist M, Emerson C). University of Texas Press, Austin, TX
- Booth W C 1979 *Critical Understanding: The Powers and Limits of Pluralism*. University of Chicago Press, Chicago
- Booth W C 2000 Richard McKeon's pluralism: The path between dogmatism and relativism. In: Garver E, Buchanan R (eds.) *Pluralism in Theory and Practice*, 1st edn. Vanderbilt University Press, Nashville, TN, pp. 212–30
- Dewey J 1929 *The Quest for Certainty* (Gifford Lectures, 1928–29). Minton, Balch, & Co., New York
- Fish S 1999 *The Trouble with Principle*. Harvard University Press, Cambridge, MA
- Garver E, Buchanan R (eds) 2000 *Pluralism in Theory and Practice*, 1st edn. Vanderbilt University Press, Nashville, TN
- Goodman N 1978 *Ways of World Making*. Hackett Publishing, Indianapolis, IN
- Herskovits M J 1972 *Cultural Relativism: Perspectives in Cultural Pluralism*, 1st edn. Random House, New York
- James W 1907 *Pragmatism: A New Name for Some Old Ways of Thinking*. Longmans Green, New York
- Krausz M (ed.) 1989 *Relativism: Interpretation and Confrontation*. University of Notre Dame Press, Notre Dame, IN
- McKeon R 1943–44 The philosophic bases of art and criticism. *Modern Philology* 41: 65–87, 129–71
- McKeon R 1951 Philosophy and method. *Journal of Philosophy* 48: 653–82
- McKeon R 1998 *Selected Writings of Richard McKeon* (McKeon Z and Swenson W eds.). University of Chicago Press, Chicago, Vol. 1
- Paul E F, Miller F D Jr, Paul J (eds.) 1994 *Cultural Pluralism and Moral Knowledge*. Cambridge University Press, Cambridge, UK
- Pepper S C 1942 *World Hypotheses: A Study in Evidence*. University of California Press, Berkeley, CA
- Popkin R H 1979 *The History of Scepticism, from Erasmus to Spinoza*. University of California Press, Berkeley, CA
- Rorty A O 1989 Relativism, persons, and practices. In: Krausz M (ed.) *Relativism: Interpretation and Confrontation*. University of Notre Dame Press, Notre Dame, IN, pp. 418–40
- Rorty R 1991 *Objectivity, Relativism, and Truth: Philosophical Papers*. Cambridge University Press, New York
- Russell B 1948 *The History of Western Philosophy, and its Connections with Political and Social Circumstances, from the Earliest Times to the Present Day*. Allen & Unwin, London
- Shweder R 1989 Post-Nietzschean anthropology: The idea of multiple objective worlds. In: Krausz M (ed.) *Relativism: Interpretation and Confrontation*. University of Notre Dame Press, Notre Dame, IN, pp. 99–139

- Tracy D 1975 *Blessed Rage for Order: The New Pluralism in Theology*. Seabury Press, New York
- Wahl J A 1920 *Les Philosophies Pluralistes d'Angleterre et d'Amerique*. Alcon, Paris
- Watson W 1993 *The Architectonics of Meaning: Foundations of the New Pluralism*, 2nd edn. University of Chicago Press, Chicago
- Williams B 1972 *Morality: An Introduction to Ethics*. Harper & Row, New York

W. C. Booth

Reliability: Measurement

Whenever an aspect of human behavior is measured, the question always arises: how good is the measurement process? In the behavioral sciences, conceptual and statistical models for answering that question are generally divided into two major categories: reliability and validity. As a first approximation, one may think of validity as concerned with whether the right construct is being measured. Reliability as concerned with the issue of how well we are measuring whatever it is that is being measured (regardless of whether or not it is the right quantity to measure). When choosing a measurement method, one must consider both of these issues; the total error of measurement comes from both the unreliability and invalidity of a measure. While this article primarily concentrates on reliability, it also discusses the measurement process as a whole, and the relationship between reliability and validity issues.

As with many terms, the scientific and psychometric uses of the word 'reliability' derive from, but are different than, the everyday use of the term. In everyday life, a car would be considered reliable if it started every time. If the word were used in the technical sense, one would also say that a car was reliable if it never started; it would only be unreliable if it sometimes started, and sometimes did not. The technical term, in other words, is more concerned with consistency of behavior, regardless of where the behavior is on a scale. The main question is whether one gets the same answer every time a measurement of an individual is made. Whether that measurement is then useful is a question of validity; a measurement procedure could be perfectly reliable but useless.

Reliability theory generally begins with a conceptual framework, which is then translated more formally into a statistical model. Most reliability theory is concerned with a particular type of measurement instrument, multiple-item tests, and some theory is applied in an even more limited setting of items that are scored dichotomously as true or false. The three most commonly used models are (a) classical test theory, based on factor analysis; (b) generalizability theory, based on analysis of variance (ANOVA); and

(c) item response theory, based on 'latent trait models.' Other statistical methods are usually needed for dealing with measurements that are either not tests, such as observations of behavior, or are not multiple-item tests. While much of measurement theory (and terminology) originally derived from consideration of ability tests, the theory applies equally well to many other areas of behavior, such as personality assessment.

1. Classical Test Theory

Classical test theory (CTT) is based on the assumption that a single underlying dimension (a trait or skill) is being measured, and that every person has a single true score, T , on that dimension. The underlying dimension or test domain may be more or less 'pure'; for example one might think of skill in (a) doing multiplication of 3-digit numbers, (b) solving algebra word problems, or (c) solving college mathematics problems. The important consideration is that all items on the test are measuring this skill or trait, and that the skill or trait is well defined. Each person's observed test score, X , is assumed to be determined by his or her true score T plus some error of measurement, e , which would either raise or lower the observed score relative to the person's true score. The basic equation of classical test theory is therefore:

$$X = T + e \quad (1)$$

in which the quantities are as defined above, and the errors are assumed to have a mean of zero and to be uncorrelated with (unrelated to) T .

What exactly does T represent? Conceptually, the definition requires a bit of imagination, and tolerance for ambiguity: imagine a person being tested over and over again, using different sets of items selected from the collection of all items designed to measure the underlying construct. (A test may consist of any number of items, including the special case of a single item, such as an essay.) The average of all these scores (more technically, the expected value) is the person's true score T .

The first major consequence of this model is that the variation among observed test scores (σ_x^2) is the sum of the variation in true scores (σ_T^2) and the variation in the errors of measurement (σ_e^2); that is,

$$\sigma_x^2 = \sigma_T^2 + \sigma_e^2 \quad (2)$$

The optimal case is when the error for each person is zero; then there is no variation in errors, so $\sigma_e^2 = 0$. Consideration of this equation leads to the definition of reliability as the proportion of total variance in

scores that is attributable to true variation rather than error variation; that is,

$$\text{Reliability} = \sigma_T^2 / (\sigma_T^2 + \sigma_e^2) = \sigma_T^2 / \sigma_X^2 \quad (3)$$

For example, if the true score variance were 8, and the error variance 2, then the total variance would be 10, and the reliability would be $8/10 = 0.8$. One useful interpretation of this number is that it is the square of the correlation between true scores (T) and observed scores (X).

Among the interesting implications of the CTT model is that a test can be made more reliable by making it longer; adding items increases the true score variance at a faster rate than the error variance. One aspect of test design is determining how long a test must be to have a reliability great enough for the intended uses of the test.

Another implication of the model is that the reliability of a test is not a single number, but depends on how individuals were sampled to take the test. If a broad range of people are sampled, so that the true score variance is large, then the reliability will be greater than if people with a more limited range of true scores are sampled. (This is because the error variance would be the same in each case.) For example, if a test were given to all high school seniors, it would have a higher reliability estimate than if it were only given to seniors who planned to attend a highly selective college or university.

1.1 Estimation of Reliability

Unfortunately, since true score and error variation are not directly observable, the definitional equation (3) above cannot be applied directly. However, this conceptualization and the accompanying statistical model lead to several possible methods for estimating reliability. First, one could randomly divide any multiple-item test into two halves, and see how the scores on one half related to scores on the other half. If a test were perfectly reliable, then each person would get the same score on both halves of the test. Of course, the correlation is never perfect; the observed correlation allows the calculation of an estimate of the reliability of the whole test. Other methods use the correlation among the items to estimate reliability. These methods, including Cronbach's alpha (developed from an idea of Louis Guttman), are based on the same conceptual and statistical principles as the split-half coefficient.

Because each person is measured at one time point, any variability due to changes in that person across time is ignored by these methods. Here is where the definition of true score becomes critical. One can think of 'true' changes in true score (due to learning or

forgetting in between testing times); one can also think of more temporary changes, due to illness, noise or distractions in the testing room, and so on. In addition, some characteristics are expected to be variable over time; these include measures of mood. In personality psychology the more labile characteristics are called *states*, while the more stable characteristics are called *traits*. In classical test theory, the temporary effects are included in errors of measurement, while changes in traits are generally not. In spite of this conceptual uncertainty, this suggests another commonly used strategy for assessing reliability: Give the same test to a group of people at two different times, and see whether their scores are closely related. For obvious reasons this is called test-retest reliability. Of course, one assumes that the first measurement does not affect the second measurement, which may or may not be the case, and which must be considered. To minimize the effects of 'true' changes in true scores over time, the time period between testings must be kept relatively short.

A related method uses two or more forms of a test, with each form being constructed in an identical manner, so that the forms will be as nearly alike as possible. If in fact they have the same means, variances, and reliabilities, then they are said to be *parallel tests*. If we administer two such forms to the same group of people (e.g., on two consecutive days), the correlation between scores on the forms is called a parallel test estimate of the reliability.

In addition to measuring the overall performance of the test, we are also interested in how well we have estimated the score of each individual. The *standard error* of the test score is such a number; if small, then the score is a more accurate estimate of ability than if the standard error is large. As an example, suppose a person has a score of 120 on an IQ test, and the test has a standard error of 5 points. Then we can construct a 95 percent *confidence interval* around the estimate by adding and subtracting two standard errors (10 points, in this case); here the interval is from 110 to 130. This means, roughly speaking that there is a 95 percent chance that the interval will contain the person's true score.

Besides considering the reliability of the test as a whole, one must also consider how good each item is in the test. The most commonly used method to assess this in CTT is the item-total correlation, which measures how well the item discriminates among people of varying skill or ability levels. To calculate this, one correlates scores on the total test (usually with the particular item omitted) with scores on the item (with 0 used for an incorrect response and 1 for a correct response). In other words, an item is good if people who score high on the test tend to get the item right more often than do people who score at lower levels on the test. An item would be considered bad if people with low scores on the test were just as likely to get the item right as people with high scores on the test.

In the test construction process, items that have low item-total correlation are generally omitted or re-written.

Another common item statistic is the proportion of test-takers that get the item right; this is usually called the *difficulty* of the item (though *facility* would be more appropriate). While it is commonly reported, it is of little relevance for reliability; item difficulties are more important when trying to tailor a test for different ability levels, or to compare the characteristics of tests that are supposed to be equivalent (parallel, in technical terms).

1.2 Weaknesses of Classical Test Theory

CTT has several weaknesses that have led to the development of other models for test scores. First, the concept of reliability is dependent on the group used to develop the test. If the group has a wide range of skill or abilities, then the reliability will be higher than if the group has a narrow range of skill or abilities. Thus reliability is not invariant with respect to the sample of test-takers, and is therefore not a characteristic of the test itself; in addition, neither are the common measures of item discrimination (such as the item-total correlation) or item difficulty (percent getting the item correct). As if this were not bad enough, the usual assumption that the standard error is the same for test-takers at all ability levels is usually incorrect. (In some extensions of CTT, this assumption is dropped, but these extensions are not well-known or widely used.) CTT also does not adequately account for observed test score distributions that have floor and ceiling effects, where a large proportion of test-takers score at either the low or high end of the test score range.

CTT has difficulties in handling some typical test development problems, horizontal and vertical *equating*. The problem of *horizontal equating* arises when one wishes to develop another test with the same properties as (or at least with a known relationship to) an existing test. For example, students who take college admissions tests such as the ACT or SAT should get the same score regardless of which version of the test they take. *Vertical equating* involves developing a series of tests that measure a wide range of ability or skill. For example, while we could develop completely independent tests of arithmetic for each elementary school grade, we might instead want to link them to have one continuous scale of mathematical skill for grades 1 through 6. While vertical and horizontal equating is not impossible within CTT, they are much more straightforward with item response theory (IRT).

These problems of CTT are partly due to some fuzziness in the theory (the population to be sampled is usually not considered in any detail in the theory). But they are also due to the failure of most data

collection to be carried out on a random sample from any population, let alone a population considered appropriate for the test being investigated. In practice, convenience samples are used; these generally fit some criteria that the investigator specifies, but are otherwise taken as they are available to the researcher. Finally, CTT as originally conceptualized was never intended to address some of the practical testing problems described above.

2. Generalizability Theory

While classical test theory provided a framework that has proved useful in a number of situations, psychometricians realized many of its limitations even as they were developing the theory. Most importantly, they realized that a wide variety of factors could influence observed test scores, including how the items were generated, tests were administered, and items were scored (especially if they were free responses scored by raters). In classical test theory, these factors were all included in the (one and only) error term in the statistical model. To include these factors in their statistical models, psychometricians developed what is known as generalizability theory (GT), which included more than one source of error in test scores.

GT is based on the statistical methods of analysis of variance. Factors thought to have a major influence on test scores are included in a design, and the analysis allows the researcher to determine how much influence each factor has on variation in test scores. GT allows the computation of different estimates of reliability, depending on which factors are considered important for a particular use of the test.

As an example, consider a test of writing, in which each student writes three brief essays, each of which is scored by two raters. GT allows the estimation of how much variation in test scores is due to variation in essay topics, and how much is due to variation in raters.

Like classical test theory, GT is primarily concerned with the behavior of the test as a whole, rather than the performance of components, such as subscores or items. For further details on GT, see *Generalizability Theory*. The classic source on GT is the book by Cronbach et al. (1972). Shavelson et al. (1989) provides a less technical overview.

3. Item Response Theory

Since 1980 there has been a steady rise in the use of item response theory, or IRT, to solve some of the problems that CTT could not adequately deal with. IRT, like CTT, is based on an assumption that a single underlying dimension is being measured. But the basic conceptual idea behind IRT is the relationship be-

tween the score (usually right/wrong) on an item and the underlying dimension, whereas CTT is based on the relationship between total test score and the underlying dimension. This relationship represented by an item characteristic curve, or ICC, shows how the probability of getting the item right (for dichotomous items) changes with the underlying trait. As the probability cannot be below zero or greater than one, the ICC generally has an elongated S shape. In three-parameter models, the items may differ in whether the curve starts at zero for people of very low ability (no guessing) or above zero (if guessing correct is likely); how fast the curve rises (discrimination); and whether the curve is to the left (easy), in the middle, or to the right (hard). In two-parameter models, only discrimination and difficulty are used; guessing is fixed at zero. And for one-parameter (Rasch) models, only difficulty is used; guessing is fixed at zero and discrimination at one for all items.

IRT models solve many problems that were difficult for CTT, including the ability to construct a different test for each test-taker, so that students at all levels of ability can be assessed with items appropriate to their status. Such 'adaptive tests,' generally administered by computers, are becoming standard in many large-scale testing situations in some countries, including the Graduate Record Examination (GRE) and licensing examinations in many fields.

In IRT, the traditional concept of reliability has no place. Instead, the emphasis is on the more basic concept of standard error of measurement; that is, how accurately has a test-taker's true score been estimated? For a fixed test given to a group of examinees, the test characteristic curve shows the standard error for each level of ability. Unlike the simplifying assumption of CTT, the standard error can vary with ability. For example, if a test is relatively easy, the standard error will be smaller for students of low ability levels, because the test is better matched to their ability; higher ability students will not be well-measured by an easy test.

4. Reliability in Criterion-referenced Measurement

In some measurement situations the primary interest is in determining whether or not a person meets a predetermined standard or criterion. These judgments are sometimes made for individuals (e.g., is my student Karol reading at grade level?), and at other times for groups (e.g., what proportion of eighth-graders are reading at or above grade level?) In everyday life, drivers' license test provides a familiar example of a criterion-referenced test. In education, the theory of criterion-referenced measurement was developed to provide suitable technical scaffolding for making such judgments.

Again using the concept of consistency as a basis for conceptualization, reliability of criterion-referenced measures is concerned with issues such as whether an individual would have received the same classification if tested by different raters, on a different form of the test, or on different days. It is clear that the consistency will depend on how close the individual's true score is to the cut-point determining mastery. If a person is well above or below the cut-point, decisions will tend to be more consistent than for people near the cut-point. In the extreme, if a person's true score is exactly at the cut-point, then that person will pass half the time and fail the other half, no matter how well the test is constructed.

Because of the 'yes/no' nature of scoring, criterion-referenced test theory is somewhat different than the usual classical test theory, but only because some different statistical methods are appropriate for this type of data. Criterion-referenced measurement is also more obviously concerned with decisions made on the basis of a test; other test theory models are only implicitly concerned with such decisions. For more information about criterion-referenced measurement, see Hambleton et al. (1978), Harris et al. (1974), and Popham (1978).

5. Issues for Observational Measures

While many characteristics can be assessed through tests and questionnaires, others are better assessed through observation of behavior. Behavior can be assessed through direct observation, or it can be recorded for later coding into categories. The recording can be more or less direct, ranging from video and audio devices to the memory of observers about past behavior (where the recording is in the mind of the observer).

Behavioral observations tend to take place in 'real world' settings, and generally (but not always) under uncontrolled conditions. Unlike tests, where those taking them know and expect that their behavior will be observed, in natural settings one can imagine a range of possibilities for reactivity to the act of being observed. The more unobtrusive the observation, the more confidence one has in the quality of the measurement (though this generally would be classified as affecting validity, rather than reliability). The classic reference on unobtrusive measurement is Webb et al. (1966).

Some issues tend to arise much more frequently in the case of observational measures than with tests and questionnaires. One such issue is the decision of whether to consider a characteristic being assessed as categorical or continuous. For example, one might want to say that a person is either depressed or not, rather than being given a rating on a (nearly) continuous scale from not at all depressed all the way

through to extremely depressed. In such a case, the underlying trait is categorical, and different statistical methods are required. The most commonly applied method in this case is latent class analysis, which has been used in settings as diverse as determining stage of development of children, psychiatric diagnosis, and medical diagnosis.

One also tends to see different types of observed measures with observational data. Sometimes a variable has unordered categories, as would be the case if one classified a student in a class as sitting quietly, sitting but making noise, or being out of his or her seat. Other times a variable may be a count or a rate, such as the number of times a student is out of his or her seat, or the number of such times divided by the amount of time during which he or she was observed.

Typically there are multiple observers of behavior, or multiple coders of recorded behavior, or both. (This can, of course, occur for certain types of tests, not just in behavioral observations. Essay tests are a common example, because multiple raters are often used. Similarly, portfolio assessments of students' work are often based on the judgment of multiple raters.)

These considerations present additional problems that affect the reliability of observational measures. While consistency remains the primary concern, and the reliability (and validity) of behavioral observations depends on many of the same factors as for tests, additional factors for observational methods include length of the period of observation, the conditions under which observations are made (including irrelevant stimuli), and the observers, raters, and coders used.

As one might infer from the discussion of generalizability theory (GT), observational measures usually have more than one possible source of error, unlike what is assumed in classical test theory. In addition to inter-observer agreement problems, observers can make their observations at different times, in different settings, and under different conditions; all of these contribute to the error of measurement. GT allows researchers to estimate how much variability in the observations is due to each of these potential sources of error.

These differences from typical test data have additional implications for the statistical theory and data analysis methods generally used with observational data. When data come from observers, the issue of reliability is primarily a matter of whether two or more observers would come to the same conclusion (or, if a rating, the degree to which they agree), and whether the conclusion depends on when and for how long the observation occurs. For variables that fall into categories, one apparently simple method is to calculate the proportion of instances in which two raters or observers agree with each other about the category into which a particular observation should be placed. Unfortunately, the simplicity is only apparent, because for behaviors that are either rare or very common, a

high percentage of agreement can occur by chance. For example, if a student is classified as out of seat on 10 percent of occasions by each observer (and in seat the other 90 percent), then the observers would agree 82 percent of the time even if they were responding independently of the actual behavior. Therefore, such measures are typically corrected for chance agreement. Additional complications arise when there are three or more observers, and when observers do not rate or observe for the same period of time.

Details on statistical methods suitable for analyzing observational data are found in Suen and Ary (1989).

6. *Relationship of Reliability to Validity*

In classical test theory, one cannot have a validity of a test that is higher than the square root of the reliability. This has led some to conclude that reliability is more important than validity of a measure, because a test cannot be valid if it is not reliable. While a test cannot be valid if its reliability is zero or extremely low, it is possible for one test to have a higher validity coefficient than another, even though the first is less reliable. Suppose that two tests (or other measures) are not actually measuring the same true score, because they are measuring different traits. Then the one with a lower reliability can have a higher validity if the trait it measures is more closely related to the outcome that the tests are meant to predict. For example, a paper-and-pencil test of rock-climbing skill might have greater reliability than a performance test done on a climbing wall in a gym, but the latter might still have greater validity as a predictor of performance might in an actual rock-climbing situation.

The (invalid) conclusion about the relationship of reliability and validity has often caused a strain between quantitative researchers and qualitative researchers. Qualitative researchers often use measurement procedures that have lower reliability than the procedures devised by quantitative researchers, but the qualitative researchers prefer these less-reliable measures because they are measuring what they want to measure. It is quite possible that in many cases they are right, and their less reliable measures have greater validity than certain measures that are more reliable.

7. *Suggestions for Further Reading*

Martin and Bateson (1986) is a short introductory work on measuring behavior. The two most popular undergraduate texts on measurement are those of Anastasi (1988) and Cronbach (1990). At a higher level are texts by Crocker and Algina (1986) and Allen and Yen (1979), and the general reference work edited by Linn (1989). McDonald (1999) is a very new text that provides a unified treatment of many issues in test theory. At a very advanced level, the classic text on

both CTT and IRT is Lord and Novick (1968). A more modern text on IRT is Hambleton and Swaminathan (1984); a simpler version is Hambleton et al. (1991).

See also: Factor Analysis and Latent Structure; IRT and Rasch Models; Generalizability Theory; Measurement Error Models; Measurement Theory: Probabilistic; Scaling and Classification in Social Measurement; Unobtrusive Measures.

Bibliography

- Allen M, Yen W 1979 *Introduction to Measurement Theory*. Brooks/Cole, Monterey, CA
- Anastasi A 1988 *Psychological Testing*, 6th edn. Macmillan, New York
- Cronbach L J 1990 *Essentials of Psychological Testing*, 5th edn. Harper and Row, New York
- Crocker L M, Algina J 1986 *Introduction to Classical and Modern Test Theory*. Holt, Rinehart, and Winston, New York
- Cronbach L J, Gleser G C, Nanda H, Rajaratnam N 1972 *The Dependability of Behavioral Measurements*. Wiley, New York
- Hambleton R K, Swaminathan H 1984 *Item Response Theory: Principles and Applications*. Kluwer-Nijhoff, Boston, MA
- Hambleton R K, Swaminathan H, Rogers H J 1991 *Fundamentals of Item Response Theory*. Sage, Newbury Park, CA
- Hambleton R K, Swaminathan H, Algina J, Coulson D B 1978 Criterion-referenced testing and measurement: a review of technical issues and developments. *Review of Educational Research* 48: 1–47
- Harris C W, Alkin M C, Popham W J (eds.) 1974 *Problems in Criterion-referenced Measurement*. Center for the Study of Evaluation, UCLA, Los Angeles, CA
- Linn R L (ed.) 1989 *Educational Measurement*, 3rd edn. Macmillan, New York
- Lord F M, Novick M R 1968 *Statistical Theories of Mental Test Scores*. Addison-Wesley, Reading, MA
- Martin P, Bateson P 1986 *Measuring Behaviour: An Introductory Guide*. Cambridge University Press, Cambridge, UK
- McDonald R P 1999 *Test Theory: a Unified Treatment*. Lawrence Erlbaum, Mahwah, NJ
- Popham W J 1978 *Criterion-referenced Measurement*. Prentice-Hall, Englewood Cliffs, NJ
- Shavelson R J, Webb N M, Rowley G L 1989 Generalizability theory. *American Psychologist* 44: 922–32
- Suen H K, Ary D 1989 *Analyzing Quantitative Behavioral Observation Data*. Lawrence Erlbaum, Mahwah, NJ
- Webb E J, Campbell D T, Schwartz R D, Sechrest L 1966 *Unobtrusive Measures: Nonreactive Research in the Social Sciences*. Rand McNally, Chicago

D. Rindskopf

Religion and Economic Life

Investigation of the relation between religion and economic life has long been a central concern of sociological reflection, but the topic has been largely defined through debates on the relationship of religion

and capitalism, the latter being the specific and dominant expression of economic life in modernity. The work of Karl Marx on the origins and world historical significance of capitalism, and of Max Weber on the Protestant ethic and the ‘spirit of capitalism’ and the economic ethics of the world religions, inevitably overshadow the discussion of both the history and the contemporary forms of the ambiguous and changing relation between religion and economic life. These classical discussions do not, however, exhaust the topic, which can be divided as follows. In Sect. 1, precapitalist configurations of religion and economic life are reviewed both as regards the sociological, or socioeconomic, study of Western history and, very briefly, the anthropological study of primarily non-Western cultures. In Sect. 2, the emergence of modern capitalism is considered in relation to political economy, Marx and Marxism, and Weber’s account of the elective affinities between religion and economic life. Following this, in Sect. 3, some contrasting reflective religious responses to nineteenth and twentieth century capitalism are outlined. In Sect. 4, contemporary capitalism, now unfettered by ideological confrontation with socialism, is represented as the driving force informing globalization, commodification, and managerialism, and as possessing such comprehensive power as to provoke and sustain a wide range of contrasting religious responses. In the concluding section, possible futures and research options are considered in relation to the capacity of economic life to destroy and recreate ‘nature,’ to reforge the parameters of human self-understanding, and to virtualize the condition of humankind. Economic life would seem increasingly to have become ‘life’ itself.

1. Precapitalist Configurations of Religion and Economic Life

The relation of religion to economic life is a function of the totality of any given society and its cultures; besides this, the way in which the interfusion of intrinsic, use, and exchange values may be understood depends on the theoretical assumptions of the inquirer. The most relevant founding fathers of sociology, Marx, Weber, and Simmel, sought to develop grand theory. In general terms, sociologists of religion have devised and reinforced relatively homogenous accounts of the modernization process and attendant phenomena of secularization, and they have used particular economic theories in order to inform and justify their approaches. By contrast, under the influence of Marcel Mauss’ account of the gift and using exchange theory (Gregory 1994), anthropologists have generally paid more attention than sociologists to specific human ecologies and reciprocal community economies as originating in ‘stone age economics’ (Sahlins 1974). Some anthropologists have corre-

spondingly defended distinctive and reciprocal human ecologies and cultures against the ravages of economic modernization in ways inadmissible to orthodox sociologists.

Archeological and textual records indicate that ancient Egypt, Assyria, Babylon, and Israel all had contextually relevant, practical financial institutions, but there is little or no trace of analytical effort. The history of Western economic thought is normally traced back to the primarily agrarian world of the fourth and fifth century BC which knew the small trading firm and had a well-defined merchant class. The representative unit of the economy was the household and commodities were exchanged in free markets. In the Graeco-Roman world early 'economic' reflection was based on the '*nomos*' or law and rule of the household '*oikos*,' and the practical wisdom of household management (*oekonomikos*) was the source from which early economic conceptions were derived. In the setting of the city-state (*polis*), the universal existence of slavery ensured that labor was relatively fixed and immobile. Religion and economic life as part of the cultural superstructure were conditioned by the dependence of ancient Greek and Roman life upon a slave economy; correspondingly, manual labor was regarded as menial. Later, the Classical tradition provided some of the core values of Western culture for two millennia.

Central to Greek thought were the good life and reflection on the nature of virtue: seen thus, economic activity amounted to a mere external factor. Plato and Aristotle were, however, aware of such concepts as the distinction between use value and exchange value, the role of money, and the division of labor. In *The Republic*, the latter corresponded with the human types of ruler, soldier, and worker, and each class pursued its proper, specialized goals. The ideal rulers, the philosopher kings, would ostensibly be free from vulgarity of wealth. Aristotle, as did Plato, assumed the existence of slavery, but he envisaged a wider participation of citizens and reflected upon the distinction between use and exchange value, favoring value based upon use. The arts of commerce and the practice of usury were condemned, above all the latter, which makes gain out of money itself and not from the natural object. Money was to be understood as a medium of exchange, a unit of account, and a store of value; but in itself money was barren, and usury the oppressive exploitation of misfortune. The classical Christian doctrine of the Trinity, couched in terms of the 'economy' of God's self-revelation of his 'essential' being, employed a metaphor with an enduring importance. Later, the evolving theology of the atonement repeatedly exploited certain economic metaphors, not least in the conceptuality of payment of debt and redemption from sin.

Little new was added to Greek economics by Roman thought. The denigration of manual labor was continued by such writers as Cato and Cicero. Pragmatic

management of ever larger estates (*latifundia*) and slaves was required, but the system based upon slavery was first questioned in principle by early Christianity. As with other aspects of life in the Roman Empire (for example, military service), the early Christian Church introduced subversive ideas which underwent rapid accommodation. While debates took place in the early church concerned (e.g.) alms-giving, according to Schumpeter there was little substantial interest in what was to become the dismal science of economics, for 'the Christian Church did not aim at social reform in any sense other than that of moral reform of individual behavior,' nor, moreover, 'did the Church attempt a frontal attack on the existing social system or on any of its more important institutions.' Indeed, 'It never promised economic paradise or, for that matter, any paradise this side of the grave' (Schumpeter 1954, p. 72). This is a view endorsed by Michael Mann, who argues that 'Christianity was a form of ideological power Only by becoming a Christian could one live a truly meaningful life' and that its 'power resides originally in the fit between the Christian message and the motivations and needs of the converted' (Mann 1986, p. 302). The question as to the 'fit' or affinity, elective or otherwise, between the impact of individual Christian belief and spirituality upon individual motivation and economic life has remained a matter of central importance above all in the thought of Max Weber. In a more uncompromising way, Marxist analysis implied that the religious ideology that emerged in the theological debates of the patristic theologians of the first four centuries in the West was but a reflex of the material base; this, however, does less than adequate justice to the complexity of the factors involved.

After the fading of the eschatological expectation of the earliest Christian communities and abandonment of the primitive communism, monasticism did, initially and in principle, preserve the radical exemplar of the economic ethic of the Synoptic Gospels (Troeltsch 1931). The religious orders that emerged from the practice of the Desert Fathers provided both the basis for the continuity of a religious culture and an economic model compatible in key respects with the warrior culture of feudalism. Monastic communities and the religious orders came to be centers of rational economic organization in ways that prefigure the business corporation. Monastic organization created wealth which, in turn, led to inner contradictions tackled by attempts at institutional reform—and which later became the target of acquisitive envy during the Protestant Reformation.

At the high point of scholastic theology and in the context of feudalism, a crucial debate took place which connects the early economic ideas of Aristotle (and of his wide philosophical concerns) with the early modern loosening of the constraints upon economic activity that facilitated the growth of international trade, and the justification of such expansion in

mercantilism. In the Middle Ages, Europe was characterized by a relatively stable agrarian, largely subsistence economy with slow urban growth. Central to urban productive activity was the guild, a form of association which ensured corporate monopolies, the protection of members, standards of quality, and price stability. The presence of the Church in relation to all social institutions ensured an unparalleled degree of cultural interfusion between religion and economic life. Social status and social function implied definite religious duties in a divinely ordered governance that extended throughout the body politic. Nineteenth-century guild socialism attempted to recover something of this integration of meaning into 'menial' work.

St Thomas Aquinas, the leading theologian of scholasticism, defended the just price and the prohibition of usury, the core principles of the attempted control by the Church of economic life. Estimating the just price based on moderate gain was a difficult task: the merchant was justified in trading at a price which ensured the life of his household in its accustomed style. Following Aristotle's account of money as barren, St Thomas condemned usury, the lending of money at interest, arguing that given the true role of money solely as a medium of exchange, it did not exist in the way that a rented house continued to exist when occupied by a tenant. St Thomas' counter-realist account of money applied more to the ethics of the personal loan rather than to large-scale financial transactions. During the course of the fourteenth and fifteenth centuries the increasing complexity of major financial transactions made it increasingly difficult to fit new practices within the constraints of Church law and this was reflected in modifications of St Thomas' rigorist views.

The prohibition of usury was circumvented through the creation of an international system identified with the Jews that was capable of supporting the accumulation of the major loans required for effective warfare. A similar role was also undertaken for a limited period by the Templars. The pariah status of Jews and their indispensable role in facilitating financial transactions and more widely in commerce came to reinforce negative attitudes which provided a long-enduring basis for anti-Semitism (Stern 1977, Chap. 18). The role of such Jewish thinkers as Marx and Sombart in the theorization of capitalism in the nineteenth century and beyond reinforced the impression that Jews and modern capitalism were intimately and fatefully inter-related (Paul Morris in Roberts 1995, Chap. 5). Freudian psychoanalytic accounts of the power of money as supreme 'reality principle' have continued the decoding of economic life. Present-day Islam continues the formal prohibition of the practice of usury, but Islamic banking and investment systems have nonetheless devised ways and means of adapting this restraint to the global economic system (Choudhury and Malik 1992).

In the early modern period, the growth of international trade, the breakdown of comprehensive ideological unity within Christendom following the Reformation, and the destruction of stable communities unified by shared religious values were factors associated with the rise of mercantilism. During the course of the sixteenth, seventeenth, and eighteenth centuries, economic analysis gradually gained increasing autonomy, throwing off the remaining constraints of scholasticism. Mercantilism and then physiocrat economic theory developed as distinct schools, which reflected the conditions of their countries of origin, England and France, where trading and rent of land represented prime sources of value, respectively.

2. *The Emergence of Modern Capitalism and its Critique*

The advent of early industrialization and the development of the factory system brought about unprecedented social disruption, breaking many ancestral ties: this process required both explanation and legitimation. In the *locus classicus* of political economy, *The Wealth of Nations* (1776), Adam Smith articulated a world view informed by the image of the pin factory in which virtue is equated with profit; his analysis of the rational division of labor and its necessary place in the making of profit grounded in the accumulation of surplus value of labor provided the justification for the removal of remaining constraints. The quasi-providential 'Invisible Hand' purportedly provided a spontaneous connection between individual wealth creation and the common good of society. Political economy attained a high status among an ascendent entrepreneurial class as it emancipated economic practice and provided the means for the elimination of embedded social, ethical, and cultural inhibitions, including those traditionally endorsed by religion in pre-industrial agrarian societies.

Marx regarded the intellectual destruction of religion by the eighteenth century Enlightenment as the axiom underpinning the true emancipation of humankind through subsequent social, political, and economic critique: there was, therefore, in principle no basis for any mutual reflective encounter between religion and economic life. Seen in full historical perspective, the theory of alienation adumbrated in the posthumously published *Paris Manuscripts* of 1844, the theory of society as composed of material basis and ideological superstructure in the *German Ideology*, the quasi-Hegelian ontology of capitalism expounded in the *Grundrisse* of 1857–8, the 'fetishism of commodities' in volume I of *Das Kapital*, and the later explorations in *Theories of Value* were all components in a theoretical scheme to which an adequate theological response has never been forthcoming. The representation of the proletariat in the

Communist Manifesto as the class excluded from all classes crystallized the fears of churches already confronted with their loss of influence over a growing urban working class and its anti-clericalism.

The uprooting of agrarian workers from an ancestral status sanctioned by divine order and their regimentation into a disciplined, wage-earning working class differentiated by a new division of labor created rather a dilemma for the churches. On the one hand, the Anglican clergyman Thomas Malthus endorsed population regulation in accordance with economic needs. On the other, the living conditions of the working class were an affront to those sensitive to the historic Christian teaching on wealth and poverty. Such figures as the Scot, Thomas Chalmers, English Christian socialists, and continental religious socialists struggled to find ways to redefine the role of religion under industrialized conditions. An unchurched, brutalized proletariat also, however, represented an inner threat to the role of Christian elites, clerical and intellectual across Europe. In the English religious establishment, serious critical engagement with 'scientific socialism,' or, moreover, with political economy was minimal. Methodism was exceptional in that it successfully harnessed religion and industrial work in ways which, according to the Halévy thesis, spared England a revolution.

Max Weber was fully aware of the threat which Marx's thought and Marxist socialism created for Liberal Protestant European bourgeois culture and society. According to Weber, the *mentalité* associated with the capitalist mode of production had historical connections and elective affinities with 'this-worldly' Protestant asceticism. It is therefore customary (but not undisputed) to link the nascent 'spirit of capitalism' with the much disputed 'Protestant ethic,' above all within the Reformed Christian tradition. Weber explored the affinities between the single-minded pursuit of God and an equally focused desire for profit, but he also applied his analysis to the major world religions, classically, for example, to the religion of India (D. Gellner in Roberts 1995, Chap. 1). Marx and Weber agreed on the centrality of acquisitiveness to the bourgeois mind: political economy both explained and justified the alienative emancipation of single-minded acquisitiveness from contextually embedded human needs. Weber used the Latin scholastic ascription of God as supreme good when he analyses this distinctive mind-set and hinted at its wider implications:

In fact the *summum bonum* of this ethic, the earning of more and more money, combined with strict avoidance of all spontaneous enjoyment of life, is above all completely devoid of any eudaemonistic, not to say hedonistic, admixture. It is thought of so purely as an end in itself, that from the point of view of the happiness of, or utility to, the single individual, it appears entirely transcendental and absolutely irrational. Man is dominated by the making of money, by acquisition as

the ultimate purpose of his life. Economic acquisition is no longer subordinated to man as the means for the satisfaction of his material needs (Weber 1930/1976, p. 53).

In *Religion and the Rise of Capitalism*, R. H. Tawney applied Weber's insights to seventeenth-century England, and provided an indirect critique of the *laissez-faire* liberal capitalism which underwent economic and social crisis during the period between the First and Second World Wars. Later, further British explorations of the processual metaphor of the 'rise' of capitalism were extended by V.A. Demant to its 'decline' and its 'persistence' by R.H. Preston. Christian thinkers such as the Anglican William Temple relied upon the emerging Keynesian standpoint in advocating 'middle axioms' as the basis for socio-ethical intervention. Following the Second World War, the Keynesian program of partly state-managed mixed economy (and declining) capitalism underpinned the ideology and implementation of the 'welfare state' for 30 years until the 'triumph of capitalism' associated with the Reagan-Thatcher era. The re-assertion of neo-classical economics (Friedman 1962) and the ideology of the New Right (Green 1987) stripped away both tacit belief in the 'hidden hand' and all ascetic hesitance with regard to hedonism (Roberts 1992). Mainline religions in Anglo-American Western countries were split between localized variants of the 'option for the poor' and outright endorsement of the 'enterprise culture.' Those who advocated the option for the poor were depicted as naive and ill-informed practitioners of the 'kindness that kills.' Conversely, New Right ideologues proclaimed the virtues of greed, the adequacy of the 'trickle down' theory of welfare, and the necessity of a docile, 'flexible' labor force willing to undergo repeated social deracination.

3. *Reflective Religious Critiques of the Relation of Religion and Economic Life*

The most significant mainline religious response to industrialization and to the growth of capitalism and Marxist socialism is to be found in Roman Catholic social teaching. After lengthy consultation provoked by the changing socioeconomic conditions of the nineteenth century, the papal encyclical *Rerum Novarum* was published in 1891. Aware of the fate of the Catholic Church in revolutionary France (and then later of the Orthodox Church in Soviet Russia), papal teaching represented the historic struggle between capital and labor as conflict between the competing rights and obligations of owners and of workers, in relation to which the Church was to exercise a moderating role. Both parties were to recognize the mutuality of their rights and obligations, and to relativize these through acknowledgment of the

Beatific Vision as the sole and legitimate goal of humankind, thus hopefully resisting the Weberian slippage (under conditions of secularization) from the spiritual worship of God to the practical service of Mammon. For a century the Church developed this doctrine through a series of encyclicals culminating in Pope John-Paul II's *Laborem exercens* (*On Human Work*) (1981). Following the collapse of Marxist socialist states in 1989/90 the changing nature of capitalism was recognized in the encyclical *Centesimus Annus* (1991). The latter contains a cultural rather than a politico-legal analysis and critique of both socialism and capitalism in which the integrity of the human agent is asserted against the contradictions of a destructive, secularizing modernity in need of comprehensive re-evangelization.

Direct intellectual contact between Christians and Marxist socialists did not take place on a significant scale until after the Second Vatican Council (1962–5) and took the form of Christian–Marxist dialogue during the following decade, but even this discussion of the nature of ‘true humanism’ would seem to have had minimal long-term influence. At the same time ‘political theology’ embodying a synthesis between Protestant theology and various strands of revisionist Marxism also developed in Germany. Insights from political theology were transplanted to Central and Latin America, where, after the Second Vatican Council and the Medellín Bishops’ Conference of 1968, ‘liberation theology’ emerged in the context of rapid industrialization and urbanization, a situation similar in important respects to the conditions the young Marx and Engels had observed in the Rhineland in the 1830s. This theology drew critical perspectives from Marxist analysis of society but side-stepped the atheistic aspects of Marxism, as was to be emphasized in the critical response of the Sacred Congregation in Rome. In Central and Latin America liberation theology’s stress upon the ‘option for the poor’ has tended to exacerbate a divergence between Roman Catholicism and Pentecostal Protestantism, where the latter has performed a role analogous to that of Methodism during the European industrial revolution (Martin 1990, pp. 27ff.). The most radical theological critique of capitalism through the representation of the imbalanced distribution of wealth in the global economy was a confessional issue (Duchrow 1987).

The Roman Catholic apologist and critic of liberation theology Michael Novak has written of a ‘magic capitalism.’ Humankind has to abandon the naive and infantile ‘dream’ (the ‘false consciousness’) of socialism, and to recognize the universal necessity of alienation at the heart of a vital economic reality. Every individual must pass through a rite of passage in which each encounters the ‘empty shrine’ of ‘democratic capitalism’: this is the contemporary ‘dark night of the soul’ (Novak 1982/1991, p. 53). Humankind may then draw upon spirituality, theology, and religious values (above all, those taken from Christianity

and Judaism) for the individual strength to survive and compete. Others, like Peter Berger, deny that capitalism has such mythicizing power (Berger 1987). Other accounts of religion and economic life are to be found in Protestant Christian theologies of ‘God’ the Economist. (Meeks 1989).

Recently, the globally influential ‘Third Way’ politics of the center left advocated by such prominent figures as President William Clinton, the British ‘New Labour’ Prime Minister Tony Blair, and the British sociologist Anthony Giddens has been accompanied by a revival in Christian socialism and continuing campaigns to remoralize the individual. This post-Keynesian and postcommunist active capitalism seeks to combine the ideology of entrepreneurial zeal with managerialism and relentless sociocultural ‘modernization.’ The ‘Third Way’ contributes to the increasing salience of the religious factor in economic life, albeit with paradoxical consequences.

4. *Contemporary Capitalism: Globalization, Commodification and Managerialism*

Contemporary capitalism has both diversified into a plurality of capitalisms, each with its contextually relevant elective cultural affinities between religion and economic life, and has consolidated itself through globalization into an infinitely complex singularity. The representation of globalization as a new paradigm for the sociological interpretation of religion (Robertson 1992) has encouraged investigation of the struggle both between and within religions for power and influence in the context of the asymmetrically distributed economic opportunities and outcomes in an ever more competitive core–periphery world system (Beyer 1995). Moreover, the problematic interface between the global and the local has become the focal point of social identity formation (Castells 1996, Vol. II, Chap. 1). Commodification, the ascription of exchange and monetary value to all aspects of the world and human life, has advanced inexorably with the result that the economic function is central, and natural or intrinsic values become ever more marginal and are reduced to mere market values. Global managerialism (Entemann 1994), mediatized marketing techniques, and the application of ‘neo-Fordist’ organization to mind-work (Drucker 1993) have accelerated the uniformitarian ‘production’ of human beings through rational ‘quality’ education and training. These are now processes replicated within some influential strands of both mainline religion and alternative spiritualities.

Information technology has increased the velocity of circulation of all forms of ‘capital’ to the point that space–time compression is having pervasive cultural consequences in new ‘economies of signs and space’ (Lash and Urry 1994). For those at the heart of a

core-periphery economic system, capitalism has become 'soft' and infused with transformative cultural power (Thrift 1997). Under these conditions, religious language and theological vocabulary have become resources for the expression of economico-cultural power. Thus, terms like 'omnipresence,' 'omnipotence,' 'omniscience,' and 'instantaneity' formerly attributed to the 'God' who subsisted at the core of Western consciousness are perhaps the only adequate conceptual categories for such transmutations. Spiritual practices are now recruited into management training in order to assist through the provision of peak experiences the strongest possible integration of individual motivation with corporate goals (Roberts in Roberts 1995, Chap. 9). The religious field may be remapped in relation to these ongoing transformations of economic life. The application of rational-choice theory to religious behavior, a syncretic global marketplace of New Age 'self-religions' and spiritual body practices, the managerialization and marketization of mainline Christianity, Christian fundamentalist 'prosperity theology,' neo-pagan 'nature religion,' and ecological piety are all indications that religion and economic life are in active and sometimes mutually contradictory synergies. Consequently, there is no one 'spirit of capitalism' in advanced modernity—but a multiplicity of affinities and collusions in context marked by growing differentiation between 'religion' and 'spirituality.'

5. Conclusion: Future Prospect—Research Options

In the modern world, the last *residua* of hunter-gatherer societies and subsistence agriculture struggle with globalized economic power: the modern and the postmodern conditions sometimes co-exist in close, yet problematic contiguity (Harvey 1989). From being marginal to the order of things, conscious economic rationality is now for many 'life' itself. 'Economic life' tends to absorb, transmute, and meld 'reality' into seamless economies of virtualized simulacra, a globalized *maya*, outside of which there would appear to be no fulcrum for a critical ideology or politics of resistance (Baudrillard 1981). In a world in which the virtual exerts hegemony over the residual real the cosmos is silent; yet ambiguous opportunities abound, not least for women seeking freedom from both biological and social construction, as cyborg, prosthetic, and virtual cultures are created and colonized. After the so-called 'End of History,' leading sociologists of religion such as Steven Bruce and Bryan Wilson continue to represent religion as subject to terminal attrition by an inevitable process of secularization. Regarded thus, religion can no longer be a resource for the culturally significant critique of total human submission to the dictates of 'economic life.'

Little would seem left for the researcher into the relation of religion and economic life other than to report on the final stages of religious decline.

Given the comprehensive economization of human life and the deteriorating global condition, the putatively normative pursuit of socio-scientific research on the assumption that it is value free and objectively disengaged is ethically untenable. Moreover, the comprehensive absorption by 'economic life' of the physical and biological basis of culture and society—and even of 'nature' itself—is ecologically unsustainable. Under conditions of totalizing inclusion in which the integrity of the ecosystem is reduced to a sociological and economic construction, deviant, marginal, and engaged research is essential. It is perhaps through an interdisciplinary and ecologically informed anthropology and in 'new economics' that the tools will be found with which to develop research into religion(s) and spiritual practices that may provide a vital but neglected resource for a sustainable human ecology and world survival.

See also: Catholicism; Economic Anthropology; Economic Sociology; Marx, Karl (1818–89); Religion and Science; Weber, Max (1864–1920)

Bibliography

- Baudrillard J 1981 *For a Critique of the Political Economy of the Sign*. Telos Press, St. Louis, MO
- Berger P L 1987 *The Capitalist Revolution. Fifty Propositions about Prosperity, Equality, and Liberty*. Wildwood House, London
- Beyer P 1995 *Religion and Globalisation*. Sage, London
- Castells M 1996 *The Power of Identity, The Information Age*. Blackwell, Oxford, UK, Vol. II
- Choudhury M A, Malik U A 1992 *The Foundations of Islamic Political Economy*. Macmillan, London
- Drucker P F 1993 *Post-capitalist Society*. Butterworth-Heinemann, London
- Duchrow U 1987 *Global Economy: A Confessional Issue for the Churches?* WCC, Geneva
- Entemann W F 1994 *Managerialism: The Emergence of a New Ideology*. University of Wisconsin Press, Madison, WI
- Friedman M 1962 *Capitalism and Freedom*. University of Chicago Press, Chicago
- Green D G 1987 *The New Right: The Counterrevolution in Political, Economic and Social Thought*. Harvester Wheatsheaf, London
- Gregory C A 1994 Exchange and reciprocity. In: Ingold T (ed.) *Companion Encyclopaedia of Anthropology*. Routledge, London, pp. 911–37
- Harvey D 1989 *The Condition of Postmodernity*. Blackwell, Oxford, UK
- John Paul II Pope 1991 *Centesimus Annus*. Origins, Washington, DC
- Lash S, Urry J 1994 *Economies of Signs and Space*. Sage, London
- Mann M 1986 *The Sources of Social Power History of Power from the Beginning to AD 1760*. Cambridge University Press, Cambridge, UK, Vol. I

- Martin D 1990 *Tongues of Fire: The Explosion of Protestantism in Latin America*. Blackwell, Oxford, UK
- Meeks M D 1989 *God the Economist: The Doctrine of God and Political Economy*. Fortress Press, Philadelphia, PA
- Novak M 1982/1991 *The Spirit of Democratic Capitalism*. Institute for Economic Affairs, London
- Roberts R H 1992 Religion and the 'Enterprise Culture': the British experience in the Thatcher era (1979–1990). *Social Compass* 39(1): 5–33
- Roberts R H (ed.) 1995 *Religion and the Transformations of Capitalism: Comparative Perspectives*. Routledge, London
- Robertson R 1992 *Globalization: Social Theory and Global Culture*. Sage, London
- Sahlins M 1974 *Stone Age Economics*. Tavistock, London
- Schumpeter J A 1954 *History of Economic Analysis*. Oxford University Press, New York; Allen & Unwin, London
- Simmel G 1978 *The Philosophy of Money*. Routledge and Kegan Paul, London
- Stern F 1977 *Gold and Iron: Bismarck, Bleichröder, and the Building of the German Empire*. Penguin, Harmondsworth, UK
- Thrift N 1997 *Soft Capitalism*. Sage, London
- Troeltsch E 1931 *The Social Teachings of the Christian Churches*. Allen and Unwin, London
- Weber M 1930/1976 *The Protestant Ethic and the Spirit of Capitalism*. Allen and Unwin, London

R. H. Roberts

Religion and Gender

According to Social Sciences religions are social and cultural features, resulting of specific historical constructions. Religious beliefs and symbols as well as church organizations are the result of social processes in which social agents are inscribed. Every religious action is limited and directed by the social situation of religious agents, official leaders, or simple believers. Practices, representations, symbols, and discourses of religious field are accomplished by socially placed believers. Even when they express their faith supposed to be absolutely private, social relations take place. Classical approaches of religions deal with the question of how social class influence religious believers and behaviors. Nevertheless, the ways gender and race influence beliefs and religious institutions have rarely been taken in account.

In the context of feminist critique to the female exclusion of the possibilities of formulating knowledge, the efforts to alter the ways of thinking are an integral part of the historical struggle of women. In the seventeenth century, Sor Juana Inés de la Cruz, a Mexican Catholic nun, invoked the right of women to participate in knowledge in the same manner as men. The development of feminist studies, especially from the 1970s on, led to the formulation of sophisticated feminist criticism of the traditional and modern forms of Western thinking. Thus, practically all areas of

knowledge have been questioned by feminist theoreticians who developed alternative proposals to hegemonic forms of knowledge.

Women have also brought the field of religious studies into question. Feminist theories and methodologies have been applied to the field of religion especially by using the category of gender or social relations with respect to sex, showing how a significant part of traditional studies in this area treat religious agents without distinction of gender and race. This generalization can be found in the use of homogenizing categories of a complex reality. Observe the categories entitled 'clergy,' 'laity,' 'hierarchy,' etc.; they are employed without any reference to the sex or race of their components. The consequences for analysis have been shown by empirical research attesting to the manner in which the field of religion is molded not only by class, but also by gender and race. The approach in terms of gender, therefore, questions the pretentious neutrality of the social sciences in their aim to examine religions.

Among the main critical concerns of these issues is the uncertainty that the sexism inherent in historical religions can be overcome. In the opinion of a few of those who study the issue, these religions are inevitably androcentric. These scholars propose the creation of alternative religious spaces where women can produce new forms of relating with the sacred, which is also thought of in an innovative way. Other scholars believe that the problem of historical religions is that men appropriated them. It is necessary to recuperate the original traditions and foundations for women to find their place in these religions.

Another area of criticism is religious organizations, churches, and others in which women as a whole are excluded. From Judaism to Taoism (Sharma 1987, Gross 1996), religious organizational forms are analyzed as social institutions that guide and control women's lives and bodies. Special reference is made to three aspects that exclude the female population in these institutions: the sphere of decision-making and power, symbology and the predominance of a monotheism in which the divinity is male, and the formulation of theological and moral discourse.

1. Feminist Criticism of the Classics of the Sociology of Religion

The specific field of sociology of religion has also been submitted to the scrutiny of feminist critique. Scholars point to sexism in the analysis of religion found in the work of classical authors in this area, such as Durkheim and Weber (Erickson 1995, Bologh 1990). The first author places a rigid division between the sacred and the profane to explain social life and religion. The power generated by rites and beliefs establishes this separation. Men, holy beings *par*

excellence and having been gifted with a soul, are the protagonists of beliefs and rites whereby new relationships and society itself are created. Women, lacking a soul, are relegated to the sphere of the profane, repetitive day-to-day life and the absence of creative capacity. Only power and violence are able to keep these two spheres separate. In modern, secularized societies, rites and beliefs are meant to help keep the ideal world of men separate from the world of women by force.

Weber's sociology of religion also bases itself on dichotomy: religions founded on both rational and ascetic ethics combat magical and ecstatic religions that incorporate eroticism. The first religions are aimed at action in the world and allow one to attain the greatness of a hero, a charismatic leader or a prophet. The second religions lead to love and only serve to 'comfort the warrior,' and are negative in relation to action in the world. Men are associated with the religions of active asceticism and rationality and are responsible for progress since they have been equipped with heroism. Social order is founded on the empire of reason and greatness, at the cost of the repression of the sentiment, eroticism and magic attributed to women.

The two classical authors defend the dominant social order in relation to the subordinated positioning of women, thus legitimizing the dominance of the male elite both in society and in religion sustained by the female population and the masses. After them, other sociologists addressing the religions institutionalized in churches rarely realize that the nature and functions of religious power are based on the sexual division of the work. The separation of the cleric from the layperson is, at once, a separation of men and women in which the power is attributed exclusively to men. One of the problems with the sociology of social and religious organizations is, as feminist sociologists demonstrate, the abstract treatment of the categories used as supposedly neuter in terms of gender, which prevents the perception of fundamental elements of reality. Even if certain studies include considerations on women, they are generally descriptive and exclude any analysis or explanation of male dominance.

2. *Feminist Contributions to an Understanding of Religion*

Feminist critics envision creative, new directions for classical and current studies of the social studies of religion, challenging their pretentious neutrality and objectivity. For example, Gross (1996) proposes the necessity of academic objectivity and, at the same time, empathy with respect to the subject matter. Empathy would be the most appropriate methodology to deal with polemic themes arousing intense emotions. It allows one to understand the believer's

point of view and creates an important point of contact between academic religious studies and religious practices from a perspective of personal faith. The effect of religious studies therefore is that which radically refutes absolutism.

In the field of theology, scholars bring traditional representations of divinities into question, in an attempt to overthrow the monotheistic, male conception of the Christian God, for example. According to a significant group of feminists, Christian women theologians reject the traditional understanding of 'human nature,' which quarters women into the specific area of maternity by destiny and not by choice. This rejection of the biological resource to explain social and religious order of the sexes leads women to radically criticize the current hierarchical organization of ecclesial institutions, especially the Catholic Church. An affirmation of an individual autonomy that is not opposed to the construction of collective ideas but rather supports them further leads women to claim recognition as moral agents capable of making ethical choices in all fields of life, including those related to the control of sexuality and the capacity to conceive new human beings.

An analysis of the consequences of exclusions that fall on women in the area of religion shows that they are not limited to the fact that women are prevented from being included in the decision-making process in churches and the formulation of authorized knowledge. These consequences also extend to the symbolic field, since they keep women within a closed circle of impurity. The impossibility of presiding over rituals considered fundamental to the development of faith and the delimitation of distinct physical spaces according to sex, with secondary importance being attributed to women, among other examples, add to the religious image of the female population as sinners and polluters of the sacred. In the case of Catholicism, this happens in a radical way. In order for a man to re-establish the relationship of a community of believers with its God, it is not only necessary that he be a man, but also that he be celibate, i.e., that he not touch any women.

Another area in which there are significant feminist contributions is the history of religion. By interrogating existing historiography and the silence with regard to female proaction, and by confronting the difficulty of sources, women scholars have been seeking to write a different history in which facts, figures, and processes are shaped by the social relations established between the sexes. Catholic religious nuns rigorously trained not to have a 'personal or collective history' have been the subject matter of historical and sociological study.

In all fields of religious study, feminist scholars affirm that the fundamental issue is to understand the way in which symbolic activities—beliefs, rites, and religious discourse—that seem to escape sexual differentiation, similar to the ordering of the internal

power structure of religious institutions, truly conform to such differentiation. Feminists are therefore concerned with showing how religions are presented in social and historical reality to be molded and interwoven with gender relations.

3. *Womens' Investment in Religion*

Feminists commonly affirm that religions historically have been spaces for guidance and control of the behavior, thoughts, practices, and representations of women. Nevertheless, many scholars believe that religious space is contradictory and ambiguous for allowing women to have certain gains in concrete historical situations. For this reason, women invest in religion to find possibilities for development and autonomy to counteract the institutional investment in control. Religions can function as mobilizing forces, allowing women to resist disciplining religious and social power.

Feminist scholars of religious history asked historiographical sources demonstrating religious women's protagonism in different religions and diverse societies and historical periods. For instance, some Buddhist traditions became attractive for women because of the large number of famous female practitioners and teachers found in their ranks. In Taoism we can find important women alchemists, monks with their own monasteries and accomplishing high roles as shamanistic mediums. In Christianity, women monks were the chiefs of mixed monasteries, with male and female monks, in medieval Europe. They managed high levels of ecclesiastical power.

Research on nineteenth century Europe and Latin America shows that under certain circumstances religions historically have become unique social spaces for the female population. These spaces further act as a refuge and create conditions for women to escape male control, including by opening up the possibility of not marrying in order to dedicate themselves to spiritual life in convents and similar places. Recent studies have reported alterations in the religious field that are partially favorable to women. This is the case of Christian Ecclesiastical Base Communities, which offered many women from poorer communities the possibility of removing themselves from the domestic environment and unbalancing sex relations as well as an important learning in exercising leadership (Drogus 1997). These communitarian spaces enabled the assimilation of feminist concepts allowing women to criticize the Catholic ecclesiastical institution as male-dominated and patriarchal. It is interesting that apparently opposite fields of discourse—feminism and Christianity—were blended to empower women. New religions were also dealt with in this manner, with research showing their ambiguities in relation to the gains they offered the female population.

Another strategy women use in their favor is the reinterpretation of sacred texts. Like many Christian theologians who developed new readings of the Bible, Islamic female scholars reinterpreted the Koran, showing that male domination of the female population, as well as the idea that women were created for men have no basis on the sacred books. Equally they deny women of being guilty, based on the proposition of the guilty figure of Eve. Buddhist scholars affirm the equalitarian original character of Buddhism. They argue that Buddha taught women as well as men and had both as his followers.

Hence, women's investment in religion can be explained not only by the social control exerted over them through religion, but also by their taking advantage of their religious participation.

4. *Future Directions of Research*

The affirmation of religious experience as a gendered experience has led specialists in religious studies and theologians of various religions to reflect with ever-increasing depth on the different ways in which men and women deal with their religious beliefs and practices. In this vein, many fields have lent themselves to feminist research in order to understand and explain women's relationship with religion and their place in historical religions. One of the most useful fields seems to be that which analyzes the effects of the contemporary re-arrangements occurring in religion both in the West and the East on the lives of women and, as a consequence, men. Another area of research, in terms of claiming sexual and reproductive rights, is that which aims to discover the ways in which different religions appropriate the female body and seek to control its sexuality and reproductive capacity. One area open to empirical research still remains: the conditions for women to produce new religious practices, discourse, and symbols and the effects on religions and the entirety of believers, whether men or women.

In sum, the growth of empirical data prepared taking gender as theoretical tool should offer new clues for understanding and explaining the reality of religions in societies of the third millennium.

See also: Islam and Gender

Bibliography

- Bologh R W 1990 *Love or Greatness. Max Weber and Masculine Thinking. A Feminist Inquiry*. Unwin Hyman, London
- Drogus C A 1997 *Women, Religion, and Social Change in Brazil's Popular Church*. University of Notre Dame Press, Notre Dame, IN
- Erickson V L 1995 *Feminism, Social Theory, and Religion*. Fortress Press, Minneapolis, MN

- Gross R M 1996 *Feminism and Religion. An Introduction*. Beacon Press, Boston
- Sharma A (ed.) 1987 *Women in World Religions*. State University of New York Press, New York

M. J. F. Rosado Nunes

Religion and Health

There are two distinct ways in which religion may be related to health. Religions—complicated systems of symbols, beliefs, practices, and the social institutions associated with them—may be classed with the psychosocial factors with apparent influence on health status. In this view, religiousness may be a protective or preventive factor in reducing mortality or morbidity. A second view reverses this causal order, in that religious beliefs and practices are often turned to by individuals and families faced with serious illness, or at the end of life, and religiousness can be affected by changes in aging or health status.

1. Religion as a Factor Affecting Health

The earliest examples of practices designed to promote the public health are to be found, among primitive peoples, inextricably mingled with the ritual of religion.

These are the opening words of C-E. A. Winslow in the entry for 'Public Health' in the original 1935 *Encyclopedia of the Social Sciences*. Winslow, professor of public health at the Yale School of Medicine, known for his studies of the decline in infectious disease mortality, draws attention to early quarantine efforts in ancient Persia and among the Biblical Hebrews as effective public health practices, motivated and given form by religious beliefs. The burial of the dead, the separation of lepers from the community, or the ritualistic preparation of meat for consumption are but a few of the early examples of practices carried out for religious reasons that had unintended public health consequences. In general, this characteristic of being unintended is typical of the work discussed under this heading—religious practices or beliefs seen to have associations with health are not undertaken for the purpose of improving individual or community health, but rather, are ends in themselves.

The social scientific origins of both theoretical and empirical work on religion and health can be found in Emile Durkheim's (1951) *Suicide*. Durkheim demonstrated, with the use of European national and regional statistical records, that suicide rates varied dramatically and predictably with certain social characteristics, among them the rates of Protestants, Catholics, and Jews in a population. His demonstration that suicide rates were lower among regions with larger Catholic and Jewish populations led him to denote types of suicides (*anomic* and *egoistic*) in modern

societies. Suicides were prevented in religious and other social groups that regulate or constrain the behavior of individual members, and also provide them with integrative and socially supportive relationships.

Despite this foundation in late nineteenth century classical sociological thought, research on religion was largely absent from social scientific work in health until the latter half of the twentieth century. When such work did appear, it was limited to studies of small, peripheral subpopulations, or to suicide as the only cause of death. Beginning in the 1960s, studies finding lower rates of mortality and morbidity (particularly hypertension) for clergy, members of religious orders, and certain sectarian groups including Seventh-Day Adventists, Mormons, and Amish, appeared in US medical and public health literature (Jarvis and Northcott 1987). Work on religiousness and suicide rates has also appeared steadily, with generally consistent support for Durkheim's original findings (Pescosolido and Georgianna 1989). More generalizable findings came in 1979, when the well-known Alameda County (California, USA) analysis of social networks and mortality appeared (Berkman and Syme 1979), quickly followed by similar findings from large population studies in Tecumseh County, Michigan (USA), and Evans County, Georgia (USA). In each of these longitudinal studies, social ties, including ties to religious congregations, were associated with lower mortality risks.

These studies were important for many reasons. Until this time, the social environment, long a member of the agent–host–environment triad in multifactorial models of disease causation, had primarily been thought of as having negative influence on health. Social environments were considered sources of stress; social conditions were frequently operationalized as poverty, urban crowding, time-pressured working conditions, or the stress of modern life. The social network studies opened for consideration the possibility that the social environment also provided positive or protective factors in health. The first social network studies were also important because they included some measure of affiliation with a religious group or attendance at religious services.

In the 1980s and 1990s, studies focusing specifically on religion as a mortality risk factor appeared, primarily from the US and Israel, most showing a moderately large and significant protective effect, even after adjustments for possible selection factors, confounders, and mediators (Kark et al. 1996). For example, Hummer et al. (1999) found that frequent attendance at religious services reduced mortality risk such that seven years was added to life expectancy at age 20. These findings were different from the earlier studies of homogeneous religious groups or clergy, in that they represented the general population, with its natural heterogeneity in religious behaviors and beliefs. Like the social network studies, they were

longitudinal, and thus were able to adjust for the health status of respondents at the start of the study. Data from prospective observational studies with well-designed health status assessment at baseline is critical, particularly when measurement of the religion variable is limited to the indicator of attendance at religious services. The potential selection effects of healthier members of the population attending religious services more frequently, and being at simultaneously lower risk of mortality, are obvious. Additional important selectivity variables included in these analyses include age, sex, socioeconomic status, education, ethnicity, and marital status, as religious participation is generally higher for women, married couples, blacks, older people, and those with higher education and income.

However, a chief criticism of the studies in this emerging field is that most contain only one or two indicators of religiousness, inadequate conceptualization and measurement for such a highly complex set of practices, beliefs, and institutions. Most frequently, attendance at religious services has been the only available measure. When some measure of private religious feelings or practices has been included and compared, attendance is usually the measure with the largest protective effects. Better measurement of the independent variable is clearly necessary before a larger understanding of explanatory mechanisms can be gained. At the present stage of research, however, several mechanisms through which religiousness might affect mortality have been described.

1.1 Reduction of Behavioral Risk Factors

The health economist Victor Fuchs observed in the early 1970s that religious teachings regulate and constrain human behaviors in numerous ways that have known health consequences (Fuchs 1974 pp. 52–5). He noted the enormous differences in infant, child, adolescent, and adult mortality rates in the two adjacent states of Utah and Nevada. Population density, climate, health care, income, and education levels in the two states were rather similar; any advantage went to Nevada. Yet death rates in Nevada for males aged 40–49 were 54 percent higher, and for females 69 percent higher than in Utah. When the causes of death were limited to cirrhosis of the liver and cancer of the respiratory system, Fuchs found that the corresponding figures were 111 percent and 296 percent. He attributed the dramatic differences to the ‘different lifestyles’ of the residents of the two states, Utah being primarily inhabited by Mormons who do not drink alcohol or smoke tobacco and ‘in general lead stable, quiet lives’ with high marriage and fertility rates and low divorce rates. Even within the Mormon church, levels of adherence to church doctrine have been shown to be associated with reductions in smoking and alcohol-related deaths (Gardner and Lyon 1982).

Most religious groups, however, do not have specific, doctrinal beliefs regarding these behavioral health risk factors, and yet rates for cigarette smoking (Koenig et al. 1998a) and substance use (Kendler et al. 1997) are significantly lower among those who attend religious services regularly, even in studies of religiously heterogeneous populations. A US national study of high school seniors found that the more frequent attenders, and those who thought religion was important in their lives, had lower rates of smoking, alcohol and marijuana use, carrying weapons, getting into fights, and driving while drinking. They also had higher rates of positive health practices such as seat belt use, and eating fruits, vegetables, and breakfast (Wallace and Forman 1998).

Religious involvement among adolescents that lowers many behavioral health risks is especially important because it represents a lowering of lifetime risks for behaviors such as smoking, which are most likely to begin in adolescence. Moreover, an analysis of data from the Alameda County study shows that religious attenders who do smoke are more likely than nonattenders to quit smoking during the course of a follow-up study, and hence further reduce their mortality risk (Strawbridge et al. 1997). Thus one of the mechanisms through which religion influences health consists of the constraints placed on behaviors that are sensation-seeking, self-stimulating, and increase risk to the self and others. The reduction in lifestyle risk factors results in lowered rates of lifestyle-related causes of death, and in at least one study a temporal pattern of religious practices preceding risk reduction behaviors has been established.

1.2 Enhancement of Social Support Networks

A second mechanism by which the effects of religious involvement may operate is through the expansion and maintenance of social support networks. The study of religion and health is in some ways derivative of the study of social networks and health, and indeed the latter studies provided the earliest modern findings, in Alameda County and elsewhere, as noted above. This mechanism represents the parallel strain in the Durkheimian theoretical tradition, that religious institutions provide both regulation for the behavior of members, as well as integration, or feelings of belonging in a social group. Religious congregants report larger social networks of both friends and relatives, and also more social support from them, compared with nonattenders (Ellison and George 1994). There are several reasons why religious congregations would serve as centers of effective social support for members of all ages, but particularly for the elderly. Congregations offer a mix of strong and weak ties, and can thus offer a broad range of emotional and instrumental supports.

Congregations of all major faiths share values and beliefs about the importance of helping others who are less fortunate and also provide organizational structure for mounting such efforts. Religious institutions provide both support and opportunities to support others, and reciprocal or bidirectional social support may be the most beneficial. Finally, as a corollary of this, members of religious groups who have served others in the past, and who have seen others receiving help, have reason to anticipate support in the future in the event of need.

Population studies of religiousness and health outcomes, particularly mortality, have generally included some assessment of these two explanatory mechanisms (e.g., Hummer et al. 1999, Strawbridge et al. 1997). Usual findings have been that: (a) religious involvement, particularly public participation in services, is associated both with better health practices and with higher levels of social support, and (b) these known behavioral and psychosocial risk factors explain some proportion (but not all) of the effect of religiousness on later health outcomes. In studies of healthy populations, then, religion (conventionally operationalized as attendance at religious services) appears to be related to a lower risk of mortality after adjustment for known confounders and selectivity variables, and the effect is partly explained by existing explanatory mechanisms and partly independent of them. This is just an emerging field, however; the number of extant studies is relatively few, the measurement of the religion variable is undeveloped, the research has been largely done in the US in regional samples with Judeo-Christian beliefs, and other explanatory mechanisms should be tested.

2. *Illness and the Search for Meaning*

The first part of this article has been concerned the unintended health consequences of religious practices, or measurable changes in health that result from religious practices undertaken for religious and not health-related purposes. Equally important for a social scientific study of religion and health, however, are religious beliefs and practices that are engaged in as a direct result of illness. Religious groups can offer both spiritual and tangible support to the sick, including prayers and comforting rituals, but also meals, visitation, childcare, and even financial help. A US survey of the use of unconventional medicine found that 25 percent of their national sample reported that they had used prayer as a treatment for illness in the past 12 months; only exercise, with 26 percent, was used more often (Eisenberg et al. 1993).

Research on the effects of religious coping (including but not limited to prayer) has frequently taken indicators of mental health and well-being as outcomes; studies of recent widowers, parents who had

lost a child, women with hip fractures, or elderly men with multiple chronic illnesses have shown lower rates of depression among those with religious beliefs or participation compared with those who have none. For example, hospitalized patients diagnosed with major depressive illness were found to have a shorter time to recovery from depression if they felt themselves to be more intrinsically religious (Koenig et al. 1998b). In another study, heart patients who felt that they received strength and comfort from their religious beliefs had a lower risk of death during the six months following surgery than those who did not (Oxman et al. 1995). At the same time, not all forms or styles of religious coping may be benign, for example, when the cause of an illness is attributed to moral failure; research is just beginning on maladaptive religious coping (Pargament 1997).

Classical sociological theories of religion again anticipate these functions of religion in health. Weber's concept of 'theodicy' provides a historical model of the articulation of the problems of individuals with religious language. Religious individuals can draw on the experiences of fellow believers or the stories of their faith tradition to put their own lives into a larger context, to learn lessons from others who have faced similar circumstances, to reduce fear, or to gain hope for the future. Religious ritual practices for mourning or for healing can assist individuals in making transitions to new life stages in which adaptation to loss is required. Religious coping strategies appear prominently in the coping repertoires of aged individuals faced with their own serious illness, or the illness or death of a loved one, situations over which the individual has little direct control, and which may lack clear explanations or challenge beliefs that the world is just (Pargament 1997). In contrast with the population studies described above, in which attendance at worship services was most predictive of later health outcomes, studies of religious coping frequently find that subjective religious beliefs or nonorganizational forms of religious practice (e.g., private prayer) play an important role.

Religious beliefs and practices are especially prominent in the care for and concerns of individuals at the end of life. The modern hospice, begun in England in the 1960s, was modeled after medieval Christian institutions that cared for the dying, and the spiritual concerns of patients are still a central aspect of the new field of palliative medicine. There is also some evidence that the observance of religious holidays may alter the timing of death for elderly individuals; multiple studies in both Jewish and Christian populations have observed short delays before death that permit the celebration of a last sacred ritual (Idler and Kasl 1992).

Despite a background in the Durkheimian and Weberian traditions of the sociology of religion, the field of research in religion and health remained mostly unexplored through the century, but has begun grow-

ing quite rapidly in the 1990s, especially in the US, where levels of interest in religion are generally higher than in Europe. One may anticipate further growth in this field, as the measurement and conceptualization of the relevant domains of religiousness and spirituality improve, and as the population ages and the prevalence of chronic illness increases.

See also: Adolescent Health and Health Behaviors; Aging and Health in Old Age; Bereavement; Chronic Illness, Psychosocial Coping with; Depression, Hopelessness, Optimism, and Health; Emotions and Health; Epidemiology, Social; Health Behavior: Psychosocial Theories; Psychotherapy and Religion; Religion: Definition and Explanation; Religion, Phenomenology of; Religion: Psychiatric Aspects; Religion, Psychology of; Religion, Sociology of; Social Integration, Social Networks, and Health; Social Support and Health

Bibliography

- Berkman L F, Syme S L 1979 Social networks, host resistance, and mortality. *American Journal of Epidemiology* **109**: 186–204
- Durkheim E 1951 *Suicide*. Free Press, Glencoe, IL
- Eisenberg D M, Kessler R C, Foster C, Norlock F E, Calkins D R, DelBanco T L 1993 Unconventional medicine in the United States: Prevalence, costs, and patterns of use. *New England Journal of Medicine* **328**: 246–52
- Ellison C G, George L K 1994 Religious involvement, social ties, and social support in a southeastern community. *Journal of the Scientific Study of Religion* **33**: 46–61
- Fuchs V R 1974 *Who Shall Live? Health, Economics, and Social Choice*. Basic Books, New York
- Gardner J W, Lyon J L 1982 Cancer in Utah Mormon men by lay priesthood level. *American Journal of Epidemiology* **116**: 243–57
- Hummer R A, Rogers R G, Nam C B, Ellison C G 1999 Religious involvement and US adult mortality. *Demography* **36**: 273–85
- Idler E L, Kasl S V 1992 Religion, disability, depression, and the timing of death. *American Journal of Sociology* **97**: 1052–79
- Jarvis G K, Northcott H C 1987 Religion and differences in morbidity and mortality. *Social Science & Medicine* **25**: 813–24
- Kark J D, Shemi G, Friedlander Y, Martin O, Manor O, Blondheim S H 1996 Does religious observance promote health? Mortality in secular vs. religious kibbutzim in Israel. *American Journal of Public Health* **86**: 341–6
- Kendler K S, Gardner C O, Prescott C A 1997 Religion, psychopathology, and substance use and abuse. *American Journal of Psychiatry* **154**: 322–9
- Koenig H G, George L K, Cohen H J, Hays J C, Larson D B, Blazer D G 1998a The relationship between religious activities and cigarette smoking in older adults. *Journal of Gerontology Series A: Biological Sciences and Medical Sciences* **53A**: M426–34
- Koenig H G, George L K, Peterson B L 1998b Religiosity and remission of depression in medically ill older patients. *American Journal of Psychiatry* **155**: 536–42

- Oxman T E, Freeman D H, Manheimer E D 1995 Lack of social participation or religious strength and comfort as risk factors for death after cardiac surgery in the elderly. *Psychosomatic Medicine* **57**: 5–15
- Pargament K I 1997 *The Psychology of Religion and Coping*. Guilford, New York
- Pescosolido B A, Georgianna S 1989 Durkheim, suicide and religion: toward a network theory of suicide. *American Sociological Review* **54**: 33–48
- Strawbridge W J, Cohen R D, Shema S J 1997 Frequent attendance at religious services and mortality over 28 years. *American Journal of Public Health* **87**: 957–61
- Wallace J M, Forman T 1998 Religion's role in promoting health and reducing risk among American youth. *Health Education and Behavior* **25**: 721–41

E. L. Idler

Religion and Politics: United States

Contrasted with other developed Western societies, high levels of religious affiliation and practice persist in the United States. Although the USA has maintained a constitutional separation of church and state, its citizens' religious beliefs and practices have often had political consequences. This article reviews theoretical and empirical literature about the mechanisms that link religion to American politics.

1. Definitions of Central Concepts

According to Durkheim, 'religion' is a form of collective memory for a society that projects order, stability, and predictability in social interactions on to the Divine will. Leege argues that 'People (who) share a sacred community...develop the self-confidence to act in concert. Religion specifies what action to take, and religious beliefs (and religious institutions) create the obligation to act' (Leege and Kellstedt 1993, p. 10). For the adherent, the political consequences of religious beliefs are that if life is not genuine unless lived in a religious way, so politics is not just unless its policies measure up to God's law.

Culture and religion have a deep symbiotic relationship. 'Culture' is a template—in Geertz's term 'a set of control mechanisms' (Geertz 1973, p. 49) for social relationships. When undergirded by religious values, culture establishes a moral order (Wuthnow 1987). Culture performs three central functions for a society: (a) it offers identity, (b) it prescribes norms for behavior, and (c) it maintains boundaries on relationships (Wildavsky 1987).

'Politics' refers to collective action. Politics resolves conflicts over social goals and the means for attaining them through collectively binding decisions (Easton

1953, Lasswell 1958). 'Power,' the ability to get others to do one's bidding (Dahl 1956), is the central relationship in politics. Culture and religion are important to politics because they help to legitimate social hierarchies, set goals, and specify boundaries on political institutions.

In democratic societies, religious institutions are normally located in the civil society. 'Civil society' is that collection of cultural norms, social and economic institutions that compose a way of life but cannot legitimately employ the means of coercion. They range from freely chosen activities (voluntary associations, market transactions) to activities resulting from one's station in society (class obligations). Since 'church' is in civil society and the 'state' potentially possesses the legitimate monopoly over the means of coercion (Max Weber, in Gerth and Mills 1946), Americans often stress the need for constitutional separation. A theocracy merging the two would be dangerous, but a public square devoid of humane religious standards for justice would lack transcendent perspectives and meaning (de Tocqueville 1958).

2. Role of Religion in American Social and Political Life

Max Weber (in Gerth and Mills 1946) argued that to understand a nation's religious world views, one should ask what its people are saved from and saved for. The American social system and its mythology reinforced a deep sense of equality and enterprise; one could become whatever one wanted through individual initiative. But how could some legitimately rule or become wealthy in a nation without a natural aristocracy? After all, the USA was settled heavily by those who had run foul of the law, escaped from conscription, would not submit to ecclesiastical order, or wanted nothing of the poorhouse. Creating a nation of such people, Legee et al. (in press) argue, would be a recipe for anarchy. However, a founding myth joining the Pilgrim covenant of popular sovereignty under God to the righteous life of Puritans would provide a legitimating moral order. Out of the welter of humanity, one could know whom to trust in economic relations or whom to invest with political power by their public righteousness. Evidence of religious commitment afforded social standing. From Thomas Jefferson to William Jefferson Clinton the test of religious commitment has been a test of fitness to rule.

Beyond its cultural function, religion persisted for three reasons, and each gave it a political character. First, contrary to the lazy monopoly that most European churches exercised (see Finke and Stark (1992) for a discussion of this thesis), American churches had to compete for adherents. In addition to dispensing the sacraments, churches developed a wide range of ministries, programs, and social services to

attract and hold the loyalties of their members. Political interests evolved from efforts to maintain and extend such programs. Second, as economic development created an ever more complex society, and individualism became the dominant social philosophy, people needed an institution to provide meaning in everyday life. The church became that solace and gathering place for 'people like us' (see Fowler 1989). Yet urbanization and mass communication thrust people with very different cultural and religious values into proximity with each other (see Engels, in Benson (1979, p. 205) for a realization of how this process undercut a Marxian interpretation of historical stages). Often people felt compelled to seek state action to protect or extend their parochial values (see Hertzke (1993) for two illustrations of cultural populism in recent American politics). Third, a transformationist impulse ran deeply through American piety. Calvinism, the dominant cultural form of American religion, had specified that righteousness would be evident in diligent work, the acquisition of wealth, and charity for the less fortunate (Weber 1946). Nevertheless, righteousness was difficult in an unjust society. Society needed to be transformed from time to time to be a fitting vessel for God's people. The Declaration of Independence, abolition of slavery, Civil War, prohibition, social work and welfare reform, labor rights, anticommunism, the Civil Rights movement, and current antiabortion politics—all were argued in the name of God. The myths of the righteous citizen and 'the city on a hill' not only fueled progressive reforms throughout American political history but also legitimated less-than-righteous foreign policies.

Finally, theorists argue that religion is relevant to American politics because of the permeability of its political structures. Its systems of federalism, division of powers, checks and balances, multitudes of interest groups, etc., all invite citizens to find the appropriate venue for success with their policy or issue (summarized in Wald 1997). For example, if religious groups cannot get an anti-abortion amendment out of Capitol Hill, and if the White House abandons them, they can turn to Supreme Court appointments, or they can argue for a shift in most abortion decisions at the state level (e.g., *Webster vs. Reproductive Health Services* 1989), where it is easier to mobilize a successful coalition locally. Most political careers in the USA depend on support at the micro level, and that is also where organized religion has its strength.

3. Mechanisms that Enhance or Diminish the Linkage between Religion and Politics

Much research in this field has addressed the political consequences of different ways of being religious (see Stark and Glock (1968) and Lenski (1963) for pioneering work on the dimensions of religiosity, and Legee and Kellstedt (1993) for more recent methodological

work on American National Election Studies datasets). The most persistent theme is that the creedal dimension, i.e., what people believe, has powerful effects on politics. Through a complex model, Glock and Stark (1966) showed that the more orthodox the Christian beliefs, the more likely an American is anti-Semitic. The findings remained controversial, however, because of the measures for orthodoxy: The items tap beliefs common to the evangelical Protestant tradition but overlook beliefs that derive from sacramental/incarnational Christianity.

To avoid over-reliance on the beliefs dimension, many scholars, following the lead of Kellstedt (Leege and Kellstedt 1993, pp. 273–304) have developed indices from ‘minimal’ Christian beliefs/practices across several dimensions. Regarding the partisan and ideological consequences of religion, they conclude that the higher the level of religiosity, the greater the Republican loyalty and degree of conservatism—regardless of denomination (Green et al. 1996, especially Chaps. 14 and 15). Leege (1996), however, has shown that among Catholics generational and gender differences override the effects of either beliefs alone or multiple-measure indices of religiosity. Further, higher levels of religiosity among African-American Christians and among Jews are typically associated with some aspects of political liberalism and Democratic loyalty (see recent studies in Smidt and Penning (1997)). Scholars recognize a large unfinished agenda on the measurement of beliefs and practices, especially in the over-reporting of church attendance.

In some respects, greater progress is shown in the political consequences of religion as a group phenomenon. Churches and religious traditions are often reference groups, waxing and waning in salience for their adherents. They develop interests, seek either to protect their values from threat or to evangelize their values, or become the target of persecution. The last is particularly important to political mobilization. Converse’s (1966) study of Catholics in the 1960 election showed that Protestant Democrats not only left their party’s Catholic nominee, but that Catholic Republicans crossed over to Kennedy, and all Catholics were more highly mobilized. Those most likely to show the mobilization effects and to be drawn to Catholic Kennedy were not those who attended mass regularly but those who identified most closely with being Catholic. Social identification trumped social cohesion, i.e., regular interaction with fellow group members (see Turner in Tajfel (1982) for this distinction). Further, reinforcing group loyalties enhanced the attraction of Kennedy, e.g., Irish urban Catholics were drawn more to him than Polish nonurban Catholics. Thus, religious effects partake also of other cultural group identifications.

Nevertheless, frequency of interaction with fellow religionists can have profound political effects. In carefully controlled studies, group norms about political values and candidate selection are shown to be

propagated through behavioral contagion among church attenders (Wald et al. 1988). Where fellow church members are also informal discussion partners, political views are most likely to be reinforced and turned into political actions (Gilbert 1993, Huckfeldt et al. 1993). Even if religious congregations are composed of like-minded people, Jelen (1991) has argued, religiously based political mobilization is unlikely to occur until a negative outgroup, an external enemy, is identified.

Many leaders, both religious and political, attempt to define political agendas, issues, and choices in ways that will mobilize or demobilize religious adherents. This may or may not have political consequences. Earlier, Hadden (1969) warned that liberal activists among the Protestant clergy involved in civil rights and antiwar efforts were driving a wedge between their churches and their congregants. Quinley (1974) argued that the division was not universal, but that there was something of a two-party system emerging among the clergy: Liberals committed to social change addressed largely justice and peace issues, whereas conservatives addressed those moral issues having to do with sexuality, reproduction, family life, and national moral decline; still, the latter were less overtly involved in politics. An important study of the politics of clergy from eight Protestant denominations (excluding the Catholic churches such as Anglican, Lutheran, and Roman Catholic) shows that the two-party division is very evident, the gap in political involvement has closed, and the values of the clergy can be traced directly to theological differences (Guth et al. 1997).

Do the laity follow the beat of these ordained drummers? That is a more complex matter. Most studies are limited to comparing the values and actions of clergy samples with the values and actions of laity samples from their denominations. Already, we observed that most of the action in American religion and politics is local; therefore, an appropriate design for assessing clerical cue-giving effects would use the parish or congregation as the unit of analysis. While the possibilities of controlled studies can be seen (e.g., Roozen et al. (1984), Leege et al. (1984–9), Jelen (1991), Olson (2000)), they are still largely at a typological development stage and offer samples of such small sizes that generalizations are difficult.

Promising hypotheses include the following: (a) the more consummatory the loyalty to a local church, the more influence its pastor has over the religious and political life of its people; (b) the deeper the theological understanding of each member, the more likely that moral values shared by clergy and congregants will guide political choices; (c) the greater the sense of oppression, threat, or status loss felt by church members in modern society—where the subjects of concern are the norms by which we should live—the more the receptivity to cues from religious leaders; (d) the smaller and more homogeneous or selective the religious body, the more dominant its leader’s political

views; (e) the greater the legitimacy of corporate political involvement by the church, the more influential its leadership on a course of political action; (f) the more the issue concerns activities that individuals or families can influence (e.g., consequences of sexuality) rather than the product of impersonal economic forces (e.g. plant closings), the more religious leaders can define the political agenda; and (g) the greater the individual's intrinsic level of interest in politics, the more he/she will hear political cues when given by a religious leader. Impediments to political leadership by clergy abound: (a) separation of church and state illegitimizes the more overt political actions by religious groups; (b) religious groups specialize in salvation, not public policy; (c) religious groups can lose tax exemptions and other privileges when they become too political; (d) within the pluralistic civil society, the church is only one of many groups competing for loyalty and influence; and (e) within the democratic political process, compromise is more characteristic than success on issues defined by churches as zero-sum.

Beyond the local level, churches engage vigorously in politics. Noting a leadership vacuum in many Republican county and state party organizations, Religious Right groups such as the Christian Coalition were able to consolidate dominance over about half of the state party central committees (Rozell and Wilcox 1997). On the policy advocacy front, Wuthnow (1988) has noted a great increase in the number of religious interest groups that are arranged across denominational divides. A number of scholars (see Hertzke (1988) as representative) have examined the organization and tactics of denominational and ecumenical interest groups in Washington. Although their official lobbies are small, they use the full range of inside-the-Beltway tactics. Given the decentralization of policy-making, the larger church bodies or coalitions have similar lobbies at the state level. Outside the Beltway, church interest groups seek to mobilize home district pressure through direct mail and fund-raising, letter writing and contacting, candidate ratings and thinly disguised endorsements, get-out-the-vote drives, and media dramaturgy.

A few studies have shown the impact of the public officials' own religious beliefs and practices on their policy preferences and actions. The most comprehensive of these (Benson and Williams 1982) shows that substantial segments of both political conservatives and political liberals act from religious predispositions. The former affirm an individualism-preserving religion and politics (emphasis on the ways laws create order and protect individuals from their own worst impulses), whereas the latter embrace a community-building religion and politics (emphasis on social transformation, justice, and peace). This study inspired later work by Legee and Welch (1989) where the same types were found in mass samples of Catholic laity, and by Guth et al.'s (1997) clergy study.

Thus, religiously inspired public officials, clergy, and laity appear to have in common certain foundational religious beliefs that divide them from each other both theologically and politically. At the same time it is important to note that (a) a substantial proportion of public officials are only nominally religious, and (b) representational roles (read: the imperative to be re-elected) constrain translating religious beliefs into public policy.

Finally, political elites in the USA need religious groups if they are to achieve their goals of election/re-election. Religious group membership provides a useful basis for political market segmentation. Strategic politicians can mobilize religious group members by advocating policy interests, stressing group affinities (e.g., Kennedy as a modern Catholic, Reagan as an apostate Irish Catholic), or symbolically pillorying the 'enemies' of a religious group. Effective as mobilization is, demobilization of a religious group which composes part of a rival party's majority coalition has the same impact on chances of victory. This is done by stressing how the rival party has deserted the religious group's values (Legee et al. (in press) apply social attribution theory to understand this strategy). Over time, crystallized party images come to include the policy positions of religious groups and even the group itself. For example, the public associated evangelicals and fundamentalists with the Republican Party as early as 1964 and 1980, but they became central to the party image from 1988 to 1996.

Decisions to rely on churches for political mobilization are well advised. Using a resource theory of political participation, Verba et al. (1995) show that churches nurture skills that easily transfer to the political sector. Thus, minorities such as African-Americans associated with black churches participate more than a socioeconomic model would predict, while many Latinos associated with Catholic parishes not developing such skills participate less. Wald (1997) argues that the following skills are germane to politics derive from church participation in the USA: (a) social skills in listening, mediating, and leading; (b) awareness of public issues from a moral perspective; (c) encouragement to join other civic and community betterment activities; (d) a conviction that there is a sacred character to social obligations that transcends self-interest; and (e) the self-esteem that derives from practice with public assignments.

Empirical evidence, then, sustains the argument that religion has measurable effects on American politics, at the same time that certain styles of influence attempts are circumscribed. Basic problems of conceptualization and measurement, underspecified models, endogeneity, failure to anchor puzzles in strong social theory, and inappropriate samples continue to characterize this developing field of inquiry.

See also: Church and State: Political Science Aspects; Religion: Mobilization and Power; Religion, So-

ciology of; Religiosity, Sociology of; Religious Fundamentalism: Cultural Concerns; Religious Nationalism and the Secular State: Cultural Concerns; Social Movements: Resource Mobilization Theory; Weber, Max (1864–1920)

Bibliography

- Benson L 1979 Marx's general and middle-range theory of social conflict. In: Merton R K, Coleman J S, Rossi P H (eds.) *Qualitative and Quantitative Research: Papers in Honor of Paul F. Lazarsfeld*. Free Press, New York
- Benson P L, Williams D L 1982 *Religion on Capitol Hill: Myths and Realities* 1st edn. Harper & Row, San Francisco
- Converse P E 1966 Religion and politics in the 1960 election. In: Campbell A, Converse P E, Miller W E, Stokes D T (eds.) *Elections and the Political Order*. Wiley, New York, pp. 96–124
- Dahl R A 1956 *A Preface to Democratic Theory*. University of Chicago Press, Chicago
- de Tocqueville A 1958 (1835/1840) *Democracy in America* (trans. Bradley P). Vintage, New York
- Easton D 1953 *The Political System: An Inquiry into the State of Political Science* 1st edn. Knopf, New York
- Finke R, Stark R 1992 *The Churching of America, 1776–1990: Winners and Losers in Our Religious Economy*. Rutgers University Press, New Brunswick, NJ
- Fowler R B 1989 *Unconventional Partners: Religion and Liberal Culture in the United States*. W. B Eerdmans, Grand Rapids, MI
- Geertz C 1973 *The Interpretation of Cultures: Selected Essays*. Basic Books, New York
- Gerth H H, Mills C W 1946 *From Max Weber: Essays in Sociology*. Oxford University Press, New York
- Gilbert C P 1993 *The Impact of Churches on Political Behavior: An Empirical Report*. Greenwood, Westport, CT
- Glock C Y, Stark R 1966 *Christian Beliefs and Anti-Semitism* 1st edn. Harper & Row, New York
- Green J C, Guth J L, Smidt C E, Kellstedt L A 1996 *Religion and the Culture Wars: Dispatches from the Front*. Rowman and Littlefield, Lanham, MD
- Guth J L, Green J C, Smidt C E, Kellstedt L A, Paloma M M 1997 *The Bully Pulpit: The Politics of the Protestant Clergy*. University Press of Kansas, Lawrence, KS
- Hadden J C 1969 *The Gathering Storm in the Churches*. 1st edn. Doubleday, New York
- Hertzke A D 1988 *Representing God in Washington*. University of Tennessee Press, Knoxville, TN
- Hertzke A D 1993 *Echoes of Discontent: Jesse Jackson, Pat Robertson, and the Resurgence of Populism*. CQ Press, Washington, DC
- Huckfeldt R, Plutzer E, Sprague J 1993 Alternative contexts of political behavior: Churches, neighborhoods, and individuals. *Journal of Politics* **55**: 365–81
- Jelen T W 1991 *The Political Mobilization of the Religious Beliefs*. Praeger, New York
- Lasswell H D 1958 *Politics: Who Gets What, When, How?* Meridian Books, New York
- Leege D C 1996 The Catholic vote in '96: Can it be found in church? *Commonweal* **27**(September): 11–8
- Leege D C (ed.) 1984–9 *The Notre Dame Study of Catholic Parish Life, Reports 1–15*. Institute for Church Life, Notre Dame, IN

- Leege D C, Kellstedt L A 1993 *Rediscovering the Religious Factor in American Politics*. M E Sharpe, Armonk, NY
- Leege D C, Wald K D, Krueger B R, Mueller P D in press *The Politics of Cultural Differences: Social Change and Voter Mobilization Strategies in the Post-New Deal Period*. Princeton University Press, Princeton, NJ
- Leege D C, Welch M R 1989 Religious roots of political orientations: Variations among American Catholic parishioners. *Journal of Politics* **51**: 137–62
- Lenski G 1963 *The Religious Factor*. Doubleday, Garden City, NY
- Olson L R 2000 *Filled with Spirit and Power: Protestant Clergy in Politics*. State University of New York Press, Albany, NY
- Quinley H E 1974 *The Prophetic Clergy: Social Activism Among Protestant Ministers*. Wiley, New York
- Roozen D A, McKinney W, Carroll J W 1984 *Varieties of Religious Presence: Mission in Public Life*. Pilgrim Press, New York
- Rozell M J, Wilcox C (eds.) 1997 *God at the Grass Roots 1996: The Christian Right in the American Election*. Rowman and Littlefield, Lanham, MD
- Smidt C E, Penning J M (eds.) 1997 *Sojourners in the Wilderness: The Christian Right in Comparative Perspective*. Rowman and Littlefield, Lanham, MD
- Stark R, Glock C Y 1968 *American Piety: The Nature of Religious Commitment*. University of California Press, Berkeley, CA
- Tajfeh H 1982 *Social Identity and Intergroup Relations*. Cambridge University Press, New York
- Verba S, Schlozman K L, Brady H E 1995 *Voice and Equality: Civic Voluntarism in American Politics*. Harvard University Press, Cambridge, MA
- Wald K D 1997 *Religion and Politics in the United States*, 3rd edn. CQ Press, Washington, DC
- Wald K D, Owen D E, Hill S S 1988 Churches as political communities. *American Political Science Review* **82**: 531–48
- Weber M 1946 *The Protestant Ethic and the Spirit of Capitalism* (trans. Parson T). Scribner's, New York
- Wildavsky A 1987 Choosing preferences by constructing institutions: A cultural theory of preference formation. *American Political Science Review* **81**: 3–21
- Wuthnow R 1987 *Meaning and Moral Order: Explorations in Cultural Analysis*. University of California Press, Berkeley, CA
- Wuthnow R 1988 *The Restructuring of American Religion*. Princeton University Press, Princeton, NJ

D. C. Leege

Religion and Science

The intersection of religion and sciences has been both more complex and less conflictive than is frequently depicted by those who cite the trial of Galileo, the dispute over the theological implications of Charles Darwin's theory of natural selection, and the Scopes Trial as evidence of the historical incompatibility of religious beliefs and scientific endeavor. This article will first examine the conflict thesis, then survey the history of the relations between religion and science in the Western tradition from the fifth century BC to the

end of the twentieth century with special reference to Christianity.

1. *The Historiography of Religion and Sciences*

The thesis that Christianity has been historically hostile to science was popularized by two influential works that appeared in the late nineteenth century: John William Draper's (1811–1882) *History of the Conflict Between Religion and Science* (1874) and Andrew Dickson White's (1832–1918) *A History of the Warfare of Science with Theology in Christendom* (1896). White, the first president of Cornell University, argued that Christianity had a long history of opposing scientific progress in the interest of dogmatic theology. Wedding a triumphalist view of science to a dismissive view of theology, his thesis struck a responsive chord in American thought, which was at the turn of the century increasingly committed both to a secular outlook and to recognizing the central role that the sciences played in modern society. The Draper–White thesis became enormously influential in America both among academics and in the popular mind. It came to dominate the historical interpretation of the relations of religion to science during much of the twentieth century. Reflecting a positivist outlook, it viewed science as continually progressing and overcoming the entrenched antagonism of religious opinions, which invariably retreated before its advance.

Some historians regarded the Draper–White thesis as oversimplifying and distorting a complicated relationship. Thus, Burt, in *The Metaphysical Foundations of Modern Physical Science* (1932), suggested that, far from being hindered by religion, science was in fact indebted to theological ideas. Alfred North Whitehead, in *Science and the Modern World* (1926), rejected the view that in controversies between science and religion, religion was wrong, and ... science was always right. The true facts of the case are very much more complex, and refuse to be summarized in these simple terms. A number of important studies, by Robert K. Merton, Herbert Butterfield, and Alexandre Koyré, among others, also challenged the conflict thesis.

A smaller number of historians of science went even further to argue that specific components of Christian theology were responsible for the scientific revolution. Reijer Hooykaas contended that Calvinism's accommodationist understanding of Scripture and its voluntaristic theology encouraged scientific endeavor. Similarly Stanley Jaki, in *Science and Creation* (1974), maintained that Christian theism, the doctrine of Creation, and the belief that the universe is governed by a rational and benevolent deity were necessary preconditions of the development of modern science.

In the last three decades of the twentieth century several scholars have provided a more systematic

reevaluation of the conflict thesis. Two major contributions to the literature came from Lindberg and Numbers (1986) and Brooke (1991). In the introduction to their composite volume, Lindberg and Numbers argued for an historical picture that recognized 'a complex and diverse interaction that defies reduction to simple "conflict" or "harmony"' to replace the largely polemical one drawn by Draper and White. Brooke objected to imparting essentialist understandings to words like 'science' and 'religion,' which had many different meanings over time and could not be reduced to simple static concepts. 'Natural philosophy,' for example, denoted something very different in the seventeenth century from 'science' in the twentieth. The growing acknowledgment that in their relations religion and the sciences have exhibited a multiplicity of attitudes, reflecting local conditions and particular historical circumstances, led Brooke to suggest that a 'complexity thesis' is a more accurate model than the familiar conflict thesis. The views of Brooke, and Lindberg and Numbers have gained increasing acceptance among professional historians of science.

2. *Greece and Rome*

Western science as an explanatory enterprise began with the Greeks. It was in the sixth century BC that the Pre-Socratic philosophers like Thales (fl. 585 BC), Anaxagoras (c. 500–428 BC), and Anaximander (c. 610–546 BC) challenged the mythopoeic picture of the natural world. Most Pre-Socratics understood the world in nonmythological terms, viewing it as the product of natural forces. In the late fifth century BC they experienced some popular opposition in Athens. In c. 437–436 BC Anaxagoras was prosecuted on a charge of impiety and his books publicly burned.

By the end of the fifth century BC natural philosophy (physical science) had come to be recognized by the Greeks as offering the most suitable means of understanding the world of nature and religious issues and hereafter were to play a limited role in either hindering or encouraging scientific endeavor in the classical world. This does not mean that religious ideas were of no account in formulating world views for the Greeks or for the Romans, who took over the philosophical and scientific culture of the Greeks. The Epicureans explicitly repudiated a religious view of the universe while for the Stoics, science (especially cosmology) and religious thought were related intimately. Moreover, the Greek view of the world was predominantly an organicist or vitalist one. Philosophical naturalism, which was espoused by the Pre-Socratic philosophers, the sophists, Aristotle (384–322 BC), and the Epicureans, always remained a minority opinion in Graeco–Roman thought. In general the Greeks conceived of nature as a divinity. Plato (c. 427–347 BC) believed that a world soul (psyche)

animated the universe, in which the sun, the moon, the stars, and the planets were gods. A vitalist world picture, stripped of its pagan elements, was transmitted by the Romans to Western Europeans, who continued to hold it well into the Middle Ages.

3. Late Antiquity and the Middle Ages

Christianity, founded by Jesus of Nazareth (c. 4 BC to 29 AD), spread rapidly throughout the Roman Empire owing to its extensive missionary activity. In the second century AD Christian apologists (theologians who defended their faith philosophically against pagan critics) began the process of harmonizing Christian revelation with Graeco-Roman philosophy. Apologists like Justin Martyr (c. 100–165), Tertullian (c. 160–220), and Irenaeus (c. 130–200) were intellectuals who had received their education in classical culture and were influenced strongly by it. Their opinion of natural philosophy, which was a component of Greek philosophy and inseparable from it, was ambiguous. While they often denounced it for its pagan religious or anti-Christian elements, they employed its methods, thought forms, and vocabulary—and drew on it when it was useful in the rational defense of Christianity. These church Fathers hellenized Christianity by taking over elements of classical culture and incorporating them into the Christian world view. Tertullian, for example, who asked, ‘What indeed has Athens to do with Jerusalem? What harmony is there between the Academy and the church?’ is sometimes cited as an example of early Christian anti-intellectualism. Yet he often borrowed from Greek natural science and medicine to buttress his theological arguments. Not all Christian apologists accepted Greek science, however. A few, like Tatian (second century), who were influenced by ascetic ideals or concerns about potential demonic influences, were either guarded in their acceptance or hostile in their approach to Greek pagan natural philosophy.

The most influential of the church Fathers was Augustine of Hippo in North Africa (354–430). Augustine shared the reservations of some of the Fathers about the value of natural philosophy but he recognized its utility. Like other Fathers he rejected *curiositas*, the desire to pursue knowledge for the sake of knowing, which he regarded as a characteristic feature of pagan learning. But Augustine believed that the study of natural philosophy was legitimated when it served higher purposes, such as aiding in the understanding of Scripture or lending support to Christian theology. By becoming the handmaiden of Christian theology natural philosophy was assured a place in the emerging Christian culture, one that was to last through the Middle Ages and well into the early-modern period.

The early Middle Ages (c. 500–1000) was a period of disintegration that followed the collapse of the Roman

Empire in the West under repeated attacks and occupation by Germanic barbarians. As towns became villages, public services ceased to exist, and roads became impassable, the cultural continuity with the classical world was nearly severed. Educational institutions disappeared and learning remained alive only in Christian monasteries. A few distinguished figures stand out from the intellectual gloom of the period for their interest in natural philosophy, much of it derived from Neoplatonic sources. They include Cassiodorus (c. 485–580), Isidore, bishop of Seville (c. 560–636), the Venerable Bede (c. 673–735), and John Scotus Erigena (c. 810–877). These men could do little more than preserve as much as possible of the accumulated classical heritage, including natural philosophy, but they did so as clerics, within a monastic framework, and with the confidence that scientific knowledge, as a handmaiden to theology (*philosophia ancilla theologiae*), was worthy of being transmitted to future generations.

The eleventh and twelfth centuries experienced the revival of trade between Western Europe and the eastern Mediterranean, together with the concomitant renewal of urban life, the growth of manufacturing, and the beginning of the gradual transformation of Europe from a rural to an urban economy. Cathedral schools sprang up in cities and gradually replaced monastic schools. In the twelfth century natural philosophers avidly studied Plato’s *Timaeus* and attempted to harmonize its cosmology with the Creation account in Genesis. The Arabs had preserved Greek learning, and works by Greek authors on natural philosophy were translated into Latin from Arabic, thus making available much new scientific material. The most important were the extant works of Aristotle, which by the thirteenth century had been translated in their entirety, replacing the influence of Plato and creating a major intellectual transformation of Western European universities. Muslim contributions (for example, in optics) were significant as well.

There was much in Aristotle’s writings that was useful to European Christians and that could be harmonized with Christian theology. But there were also areas of potential conflict, such as Aristotle’s belief that the world is eternal, that the soul is not immortal, and that the universe can be accounted for by naturalistic and even deterministic explanations apart from divine providence. These and other incompatible doctrines caused the teaching of Aristotle’s natural philosophy to be banned from the University of Paris (site of the preeminent theological faculty in Europe) in 1210 and 1215, but by 1255 his writings had gained official sanction and they came to be widely studied. A number of distinguished scholars, including Roger Bacon (1213–1291), Albert the Great (1200–1280), and his student Thomas Aquinas (c. 1224–1274), attempted to create a synthesis of Aristotelian philosophy and Christian theology that would rec-

oncile the Christian faith and classical reason. The existence of a more radical faction at Paris, headed by Siger of Brabant (c. 1240–1284), led the bishop of Paris, Etienne Tempier, to condemn, at the request of traditional theologians, 13 philosophical propositions espoused by Siger. In 1277 he increased the number of propositions condemned to 219. The condemnations represented a conservative backlash that placed limits on speculation in natural philosophy. In the long run, however, they provided arts masters at Paris with incentives to explore other scientific options that lay outside Aristotelian natural philosophy, particularly by way of ‘thought experiments’ that entertained hypothetical possibilities, such as the existence of other worlds or of a vacuum. In fact, freedom of scientific speculation was relatively broad in late medieval universities, which were supported by the church.

4. The Copernican Revolution and Galileo

When joined with Aristotelian physics, the geocentric theory of the solar system, as refined by Ptolemy of Alexandria (second century AD), was almost universally accepted in the West and taken to be compatible with their theologies by Jews, Muslims, and Christians until the seventeenth century. The first major challenge to the Ptolemaic theory came with the publication of Nicolaus Copernicus’s (1473–1543) *De revolutionibus orbium coelestium* (*On the Revolutions of the Heavenly Bodies*), which appeared after the author’s death in 1543. Copernicus advocated a heliocentric model of the solar system. How he came to espouse it is unknown but he might never have published it had it not been for the encouragement of a young Lutheran astronomer, Georg Rheticus (1514–1574) of Wittenberg. Although they initially rejected its heliocentric cosmology, Lutherans at the University of Wittenberg, including the Protestant reformer Philip Melancthon (1497–1560), were attracted to the mathematical convenience of the new theory. Elsewhere, however, European natural philosophers rejected it for a variety of reasons, many of which were scientific rather than theological. In the year 1600 there were probably no more than 10 Copernicans in all Europe.

One scientist who was quick to embrace the new theory was Galileo Galilei (1564–1642), a young professor of mathematics at the University of Padua. Although secretly a Copernican by 1597, Galileo did not publicize his views until 1613, when in *Letters on Sunspots* he declared his acceptance of the heliocentric system. It was, however, his *Letter to the Grand Duchess Christina* (1615), in which Galileo argued that Scripture did not intend to impart scientific truth but accommodated itself to popular language in describing the natural world, that provoked the Catholic Church’s response. Although a layman, Galileo had undertaken to interpret the Bible. In the following

year (1616) he appeared before the Congregation of the Holy Office (the Inquisition). Cardinal Bellarmine stated his willingness to accept the Copernican theory if it could be proven (which it could not be at the time); otherwise Scripture ought to be read literally as teaching geocentricity. The Church declared the theory ‘false and contrary to Holy Scripture’ and recommended that *De revolutionibus* be ‘suspended until corrected.’ Galileo was warned not to teach or defend heliocentrism except as a hypothesis until it could be proven. In 1632, after maintaining silence for 16 years, Galileo published his *Dialogue on the Two Principal World Systems*, an ostensibly even-handed debate between proponents of the Ptolemaic and Copernican theories in which the arguments for the Ptolemaic theory were made to look feeble. Galileo was summoned to Rome, where he stood trial in 1633. Compelled to abjure his belief that the earth moved, he was confined to a country estate near Florence, where he continued writing until his death in 1642.

The traditional picture of Galileo as a martyr to intellectual freedom and a victim of the Catholic Church’s hostility to science is little more than a caricature. The trial of Galileo must be understood against the backdrop of his own times. A vain and ambitious man, he did not suffer fools gladly and he had a penchant for making enemies in high places. Among those he alienated was Pope Urban VIII (1623–1644), who had earlier encouraged Galileo’s researches on the motion of the earth, but came to believe that Galileo had betrayed him in deliberately violating an injunction not to teach Copernicanism. Moreover, strong reaction against Copernicanism (and Galileo’s espousal of it) came from the Aristotelians of the scientific establishment, some of whom were determined to entangle Galileo with the Church. In particular, Galileo aroused the ire of a loosely-organized movement of traditional scientists called the *Liga*, which succeeded in introducing theological issues into scientific discourse and attacked Galileo for his views.

In contrast with the relative freedom of intellectual debate that natural philosophers in European universities had enjoyed in the late Middle Ages, the Protestant Reformation had made the Catholic Church more cautious. In the question of whether Scripture ought to be read as teaching a geocentric universe, the Church was anxious to prove its good faith in interpreting Scripture literally in the face of Protestant attacks on its allegorizing tendencies in biblical exegesis. Aristotelianism became, in its Thomistic form, a rigid dogma and the papal court became, after the Council of Trent (1545–1563), more authoritarian. The Inquisition and the Index of Prohibited Books, introduced into Italy by forces of the Counter Reformation, were employed against Galileo. The prosecution of Galileo represents not the *cause célèbre* in the alleged warfare of science with theology, but a dispute that involved multiple issues, including

powerful personalities and vested professional and ecclesiastical interests, which complicated Galileo's championing of a revolutionary change in astronomical models.

5. *The Mechanical Philosophy and Natural Theology*

By the mid-seventeenth century, the Copernican theory had come to be accepted widely throughout Europe, owing largely to Galileo's promotion of heliocentrism. The seventeenth century saw, too, the formulation of a mechanical model of the universe, which attempted to account for all natural phenomena in terms of matter and motion without reference to action at a distance. The mechanical philosophy had its roots in corpuscular materialism, which explained physical phenomena by the interaction of tiny particles of matter called atoms. While elements of the model had been adopted by several of the leading figures of the scientific revolution, its most influential advocates were René Descartes (1596–1650) and Pierre Gassendi (1592–1655), who sought to devise a new philosophy that would replace Aristotelianism. While Descartes and Gassendi grounded their mechanical philosophies in theological considerations, critics feared that their views would lead to materialism, as they did in the case of Thomas Hobbes (1588–1679), or to deism, as they did in the case of Voltaire (1694–1778).

The mechanical philosophy received its greatest synthesis at the hands of Isaac Newton (1642–1727) in his *Principia Mathematica Naturalis Philosophiae* (*Mathematical Principles of Natural Philosophy* (1687)). By his discovery of the principle of universal gravitation, Newton succeeded in harmonizing celestial and terrestrial mechanics, bringing for the first time the whole of nature into a precise rational interpretation. Newton's picture of a universe governed by uniform laws that could be described without recourse to primary causes resulted in the widespread belief that nature was self-governing and aided the 'natural religion' of deism and its idea of a 'watchmaker' God. It is ironic that Newton, himself a Christian (although an Arian), inadvertently laid the groundwork for the rationalism of the eighteenth-century Enlightenment. On the basis of his synthesis of the mechanical philosophy David Hume (1711–1776) developed a materialistic philosophy in England, while in France philosophers like Julien Offray de La Mettrie (1709–1751) and Baron d'Holbach (1723–1789) embraced atheistic materialism.

The popularity of the mechanical philosophy led in the latter half of the seventeenth century to the creation of a natural theology that continued into the nineteenth century. Natural theology describes the existence and attributes of God by means of knowledge derived from natural reason rather than from special revelation. There were several grounds for its attrac-

tion: the wonders of scientific discoveries, many of which could be seen for the first time with the invention of the microscope; the desire to secure an adequate foundation for a theology of divine providence that would confront the perceived dangers of materialistic atheism; and the enormous influence of Newton's mechanical picture of the world. The argument from design became the cornerstone of a natural theology, in which God was described as a craftsman, a geometer, or an architect. Not all exponents of natural theology were deists; many theists, especially those who were repelled by religious enthusiasm, were involved in the enterprise of demonstrating the majesty of the Creator from the intricacies embedded in his creation. In William Paley's (1743–1805) influential *Natural Theology, or Evidences of the Existence and Attributes of the Deity* (1802) natural theology reached its apex.

6. *Geology and Darwinian Evolution*

The influence of the Enlightenment, with its exaltation of reason and rejection of special revelation, can be observed in the developments that were occurring in science, particularly in the increasing tendency to seek naturalistic explanations and to divorce science altogether from theological considerations. In the seventeenth and eighteenth centuries theories of the earth had sought to weave together geology and theology in what was termed 'physico-theology.' Many of these theories calculated the age of the earth by using biblical chronologies in a way that could be harmonized with physical principles. Invariably they postulated a young earth. While most attributed organic fossils to the Noachian Flood, by the eighteenth century it was becoming increasingly difficult to explain the global distribution of fossils by a single event (the Deluge) that had taken place a few thousand years ago. The French naturalist Georges Louis Leclerc, Comte de Buffon (1707–1788), a deist, proposed a much older earth (publicly of 75,000 years and privately perhaps as old as three million years) and suggested that each of the six biblical days of Creation might correspond to an immense length of time. Georges Cuvier (1769–1832), a devout Protestant, demonstrated the existence of fossil mammals unlike any in his own day and without any apparent intermediate forms. He argued that a series of geologic catastrophes had eliminated species in past ages. Cuvier's catastrophes were not global and in his case did not derive from his religious convictions. But they were later read as such and were appropriated in English natural theology to support catastrophism, which became a popular form of physico-theology that attributed geological changes to great catastrophes of the past, the most recent being the biblical Flood.

The Scottish geologist Charles Lyell (1797–1875) rejected altogether providential interventions in geologic history. In his *Principles of Geology* (3 vols. 1830–1833) he succeeded in severing geology from theology. Lyell rejected diluvialism (catastrophism) for what became known as uniformitarianism: the theory that geological changes on the earth's surface have occurred slowly over long periods of time and that in explaining change we refer to processes that we currently observe. Previous proponents of diluvialism, like William Buckland (1784–1856) of Oxford and Adam Sedgwick (1785–1873) of Cambridge, quickly abandoned catastrophism and adopted the new geology, which Christians harmonized with the Genesis account of Creation in a relatively painless fashion. A group of laymen, who called themselves Mosaic or Scriptural geologists, continued to take their geology from Genesis but they were marginalized since nearly all professional geologists abandoned the Noachian Flood as an organizing idea in accounting for changes in the earth's history.

Charles Darwin (1809–82) had, as a young man, been influenced by Paley's natural theology. During his voyage aboard the H.M.S. *Beagle*, however, he read Lyell's work and began to think that slow changes had produced the geological phenomena that he observed. His theory of evolution by natural selection, published in *Origin of Species* (1859), made use of Lyell's theories that the earth was very old and the fossil record imperfect. Evolutionary ideas had preceded Darwin. Originating in the Enlightenment, they had already been proposed in the eighteenth century by Charles's grandfather, Erasmus Darwin (1731–1802), and by Jean-Baptiste Lamarck (1744–1829), the latter proposing the theory that acquired characteristics were inherited. By the mid-nineteenth century the belief that transmutationism (evolution) represented a divine plan was widespread. In advancing natural selection as a mechanism for evolution, Darwin argued that in the 'struggle for existence' those individuals with even a slight advantage would survive and pass their advantage on to their offspring. Herbert Spencer (1820–1903) referred to it as the principle of the 'survival of the fittest.'

Darwin's theory met with a mixed response that cut across conventional professional and religious boundaries. Some scientists initially rejected it, while a number of clergymen accepted it. Conservative anti-Darwinists, including several leading naturalists like Louis Agassiz of Harvard (1807–1873), geologists like Canadian William Dawson (1820–1899) of McGill, and theologians like Charles Hodge (1797–1878) of Princeton Seminary rejected it, fearing the break with final causes. A second group of Christians accepted evolutionary theory as compatible with a biblical faith. They included the Scot James McCosh (1811–1894), president of Princeton University, and the American botanist Asa Gray (1810–1888) of Harvard. A third group came to believe that evolution had

rendered supernatural religion untenable and opted (most often) for some form of philosophical materialism.

By the 1870s, however, most scientists and a large body of the educated public had accepted some form of evolution (though not necessarily natural selection). Full acceptance of Darwin's views on human nature, as set forth in his *Descent of Man* (1871), was more difficult for Christians. Darwin maintained that moral values were rationalizations of social instincts that had developed by natural selection, thus asserting a full-blown materialism. Alfred Russell Wallace (1823–1913), the codiscoverer of the theory of evolution by natural selection, rejected materialism, maintaining that the creation of the human mind required non-material forces.

In the 1920s evolution was one of several theological issues that divided the major Protestant denominations of North America in the modernist–fundamentalist controversy. While many Protestants (as well as some Roman Catholics and Jews) had come to believe in an old earth and to harmonize belief in evolution with a nonliteral interpretation of the Genesis account of Creation, conservatives emphasized in their rejection the atheistic and materialistic dangers inherent in a naturalistic theory of human origins. But the Scopes trial, held in Dayton, TN in 1925, symbolized the inability of biblical literalists to withstand the force of Darwinism, which was taken for granted by the intellectual elites of Europe and America. Anti-evolutionism was pilloried in the American press as a rearguard reaction to scientific progress and it was associated with nativism and rural and uneducated religious fundamentalism. Although biblical catastrophism was revived in the 1920s, it enjoyed favor only in conservative and sectarian Protestant circles in North America before it exploded into the popular creationist movement (called 'Flood Geology') in the 1960s.

7. Twentieth-Century Physics and Cosmology

While religious issues had been increasingly excluded from scientific discourse in geology and natural history (the study of natural objects, plants, minerals, and animals) in the nineteenth century, they had continued to be discussed in natural philosophy. From about 1895 modern physics began to deal with issues that went beyond those of classical physics, including quantum mechanics and special and general relativity. Several discoveries in modern physics challenged the determinism of the Newtonian mechanical model. Foremost was the indeterminacy principle of Werner Heisenberg (1901–1976), which demonstrated that one could not know perfectly both the position and momentum of a subatomic particle. Heisenberg's principle gave rise to much discussion of its wider

philosophical and theological implications, particularly relating to the possibility of intelligent design of the universe and human free will.

In cosmology too discoveries were made that suggested theological implications. Edwin Hubble's (1889–1953) discovery that the universe was expanding at a tremendous rate led the Roman Catholic priest and physicist Georges Lemaître (1894–1966) to propose what became known as the big-bang theory: that the universe had begun as a dense 'primeval atom' that had exploded in the distant past. It was challenged by the steady-state theory, developed in England and championed by Fred Hoyle (1915–), according to which the universe is eternal but matter is being continuously created in space. In the late 1960s, several observational discoveries led to the almost universal acceptance of the big-bang theory, which was argued by some researchers, like Paul Davies (1946–) and Frank Tipler (1947–), to strengthen belief in God's creation of the universe.

Another cosmological theory that engendered a good deal of theological discussion was the Anthropic Principle (a term coined by Brandon Carter in 1974), which asserts that the emergence of human beings in the universe is not an accidental by product of purposeless cosmic evolution. According to this theory the universe was so finely tuned during its earliest moments that if any one of several physical parameters had been slightly different it would have been unsuitable for the survival of human life. Hence, it is argued that the inhabited universe is anthropocentric, having been designed for human beings. The Anthropic Principle reintroduced into cosmological speculation at the end of the twentieth century issues that were long thought to have been laid to rest, including questions of teleology, intelligent design, and divine providence.

8. Conclusion

By the 1940s the modern synthesis of genetics and Darwin's theory of natural selection, which accounted for random variation by genetic mutation, made it possible for biological scientists to adopt a fully naturalistic position in both method and metaphysics. Naturalistic evolution, it was alleged, was able offer a complete explanation for the origin and development of the universe and for all forms of life, as well as for social factors like language, religion, and ethics. While scientists like George Gaylord Simpson (1902–1984) considered materialism to be the only philosophical position capable of scientific support, it was far from universally accepted within the scientific community. The last decade of the twentieth century, moreover, saw a resurgence of interest in the relationship of religion to the sciences. This interest arose from a variety of sources both within and outside the scientific community, including those of Christian and other

theistic traditions (who most often asserted the complementarity of religion and the sciences), process philosophers, proponents of New Age and pantheistic spiritualism, and non-Western religious traditions. Perhaps it reflected the widespread recognition that, for all its achievements, science does not provide ultimate answers to questions regarding the meaning of the universe or of life itself.

See also: Darwin, Charles Robert (1809–82); Darwinism: Social; Evolution, History of; Evolution, Natural and Social: Philosophical Aspects; Evolutionism, Including Social Darwinism; Historiography and Historical Thought: Christian Tradition; Historiography and Historical Thought: Islamic Tradition; Religion, History of; Religion, Phenomenology of

Bibliography

- Barrow J D, Tipler F J 1986 *The Anthropic Cosmological Principle*. Oxford University Press, Oxford, UK
- Barbour I G 1997 *Religion and Science: Historical and Contemporary Issues*. 1st rev. edn. Harper, San Francisco
- Biagoli M 1993 *Galileo Courtier: The Practice of Science in the Culture of Absolutism*. University of Chicago Press, Chicago
- Brooke J H 1991 *Science and Religion: Some Historical Perspectives*. Cambridge University Press, Cambridge, UK
- Brooke J H, Cantor G 2000 *Reconstructing Nature: The Engagement of Science and Religion*. Oxford University Press, New York
- Burt E A 1932 *The Metaphysical Foundations of Modern Physical Science*, rev. edn. Harcourt Brace, New York
- Durant J (ed.) 1985 *Darwinism and Divinity*. Blackwell, Oxford, UK
- Fantoli A 1996 *Galileo: For Copernicanism and For the Church*. Vatican Observatory Publications, Vatican City, Italy
- Ferngren G (ed.) 2000 *The History of Science and Religion in the Western Tradition: An Encyclopedia*. Garland, New York
- Harrison P 1998 *The Bible, Protestantism, and the Rise of Natural Science*. Cambridge University Press, Cambridge, UK
- Jaki S L 1974 *Science and Creation: From Eternal Cycles to an Oscillating Universe*. Science History Publications, New York
- Koyré A 1957 *From the Closed World to the Infinite Universe*. Johns Hopkins Press, Baltimore, MA
- Larson E J 1997 *Summer for the Gods: The Scopes Trial and America's Continuing Debate over Science and Religion*. Basic Books, New York
- Lindberg D C, Numbers R L (eds.) 1986 *God and Nature: Historical Essays on the Encounter Between Christianity and Science*. University of California Press, Berkeley, CA
- Lindberg D C, Numbers R L 2001 *Science and the Christian Tradition: Twelve Case Histories*. University of Chicago Press, Chicago
- Moore J R 1979 *The Post-Darwinian Controversies: A Study of the Protestant Struggle to Come to Terms with Darwin in Great Britain and America, 1870–1900*. Cambridge University Press, Cambridge, UK

Russell R J, Stoeger W R, Coyne G V (eds.) 1988 *Physics, Philosophy, and Theology: A Common Quest for Understanding*. Vatican Observatory Publications, Vatican City, Italy
 White A D 1896 *A History of the Warfare of Science with Theology in Christendom*. Appleton, New York

G. B. Ferngren

Religion and Youth

Religion has traditionally been an important aspect of personal and social identity. Socially, it provides a 'language'—a set of beliefs, values, and practices—which form an ultimate framework of meaning and order, binding people into one community rather than another. Individually, religion locates a person within a social context, conferring a sense of belonging, a way to understand the self in relation to the world, and a purpose to life. Western society is changing rapidly, however, bringing new uncertainties to the social context of identity formation (Giddens 1991). The last 30 years have seen economic restructuring, expansion of educational provision, new developments in communication and information technology, increased social and geographical mobility, changes in personal relationship and family norms, challenges to gender roles, and the generation of new attitudes and accepted ways of being. This raises the question as to what part religion plays in the formation of young people's identity today?

1. Who are the 'Young People'?

Before considering the relationship between religion and youth identity, it is important to note that the category 'young people' and the concept of 'youth' are problematic. There is no consensus on an upper age for 'young people' and surveys are reported using different age bands for the youngest groups. 'Youth' signifies a period of transition between childhood dependency and full adult responsibility but significant transitional markers, such as the school leaving age and access to employment and financial independence, vary between countries. There can also be age disparities within societies regarding different types of adult responsibility such as marriage and voting rights. Indeed, 'youth' is increasingly being separated from age altogether and becoming a valued personal quality to be retained throughout life. Furthermore, the experience of youth varies greatly according to gender, ethnicity, class, etc. Bearing these conceptual problems in mind, the picture of youth presented here is necessarily a generalized one and the discussion focuses on Christianity which remains the dominant religious tradition in the West (60 percent of

Europeans and 76 percent of North Americans claim a Christian identity, Brierley 1997).

2. Young People's Religious Beliefs and Practices

The data on young people's religious beliefs and practices are complex and sometimes contradictory. Throughout the industrial West young people are less likely than older people to actively engage with institutional religion (Halman and de Moor 1993, Clark et al. 1995). This may reflect a stage in the life cycle, but the evidence suggests long-term change as well.

Table 1 illustrates declining religiosity among youth in Europe on a number of traditional variables and Table 2 indicates how religiosity varies among young people cross-nationally.

From Tables 1 and 2 it can be seen that while Church association is declining, private prayer and belief in God remains quite high. Belief in the soul and life after death is also apparent. Half of European youth and 79 percent of young people in the United States (Inglehart et al. 1998, Vol. 151) describe themselves as 'a religious person.' More locally, there is evidence to suggest that young people take an interest in supernatural phenomenon, nontraditional beliefs, and 'folk' religion. Table 2 indicates the extent of

Table 1

Data from the European Values Survey^a for 18–24 year olds

	1981	1990
Church attendance—percentage who attend weekly or more	43	18
I am a religious person—percentage who agree	53	51
I take moments of prayer—percentage who agree	55	52
Spiritual beliefs—percentage who believe in:		
God	77	67
Life after death	54	50
The devil	26	23
Hell	23	20
Heaven	43	39
Sin	57	51
Reincarnation	27	26
A soul	63	65
I think often about meaning of life—percentage who agree	26	29

Source: Ashford and Timms 1992, pp. 124–8.

^a The European Values Survey samples adults from Great Britain, Northern Ireland, Republic of Ireland, Western Germany, The Netherlands, Belgium, France, Italy, Spain, and Portugal.

Table 2

International comparison of religious beliefs and practices among 16–29 year olds based on the 1990–93 World Values Survey

Country (predominant denomination) ^a	Religion very important in your life (%)	Sometimes or often pray to God outside of religious services (%)	Believes in God (%)	God is very important in your life (%)	Believes in Reincarnation (%)
Ireland (RC)	22	65	95	56	18
Northern Ireland (M)	28	56	96	49	24
Spain (RC)	09	39	78	33	28
East Germany (P)	08	17	24	15	10
Britain (P)	06	24	63	16	30
Italy (RC)	25	58	89	58	28
The Netherlands (M)	09	33	52	20	23
Belgium (RC)	06	23	51	24	20
Austria (RC)	13	44	78	42	36
France (RC)	08	26	51	15	32
Canada (RC)	21	50	85	51	33
United States (P)	46	70	95	68	16
West Germany (M)	05	26	62	25	26
Denmark (P)	03	14	50	08	21
Finland (P)	11	38	57	25	31
Norway (P)	09	28	53	16	22
Sweden (P)	08	20	34	13	27
Switzerland (M)	16	na	na	47	36

Source: Inglehart et al. 1998.

a RC = Roman Catholic; P = Protestant; M = Mixed Protestant and Roman Catholic.

belief in reincarnation. In England and Wales approximately one-third of young teenagers believe in their horoscopes, ghosts, and the possibility of contacting the dead (Francis and Kay 1995, p. 152).

It is also important to recognize that while established mainline church membership is declining, there is some growth among smaller nonestablished churches, especially those which adopt an evangelical and charismatic form of spirituality. In this context the nature of belief and practice tends to be more emotive and personal. For example, young people are more likely to think of God as an intimate friend than as a righteous judge or creator. Various non-Christian new religious movements also attract a minority of youth, although these groups have not counteracted the overall drift away from the Church (Bruce 1996, p. 188).

Young people therefore retain a degree of private belief and practice which could offer an ultimate basis for identity formation. However, Table 2 indicates this is not necessarily the case. While many young people believe in God, God and religion are not very important for a large number of them. Gender studies further reveal that religion is less important to men and boys than it is to women and girls. Women in paid employment, however, tend to be more like men in their religiosity, so as more young women expect to

have working careers, this gender difference is likely to decrease in favor of the perceived irrelevance of religion. What then is the significance of religion for young people's identity?

3. *Young People's Religion and Identity*

The data suggest three broad patterns of relationship between religion and identity among young people in modern Western society.

3.1 *Secular Identity*

Some young people reject religion altogether (according to the 1981 European Values Survey, 27 percent of people under 35 claim to be atheists, Harding et al. 1986, p. 48), but a large number have remnants of traditional belief, echos of an early but incomplete socialization into the Christian framework of their culture. These young people tend not to be hostile to Christianity and may even 'believe' to some extent, but generally they are ignorant and indifferent to its teachings. Christian plausibility has been undermined by increasing contact with the competing truth

claims of other faiths and philosophies, and by the perception that science offers more convincing explanations of life than does religion. Consequently, Christianity has little, if any, significance for these young people's self-understanding. For the most part, they orient their lives according to a utilitarian rational world view gaining their identity through material consumption and through immediate intimate relationships with family and friends. Links between religion and self-understanding are only salient on special occasions (such as the celebration of a birth or a marriage), or at times of crisis (such as during illness or death) when an implicit Christian identity is unearthed. Even then the significance of religion is fleeting. Its main function is to provide a culturally acceptable background for the renegotiation of an essentially secular identity which is largely achieved through the gathered family and friends. Committed religious belief or understanding is unnecessary in this context.

3.2 *Spiritual Identity*

Some young people are more active in their approach to religion and identity. They are aware of a wide range of religious symbols from a number of traditions and approach them as a cultural resource which can be subjectively used in a reflexive process of identity formation. Religious symbols provide a means of self-exploration, a source of empowerment and personal growth, around which a sense of self can be constructed. Without recognizing an external authority, these young people are free to decontextualize, reinterpret, and juxtapose apparently contradictory symbols without any sense of conflict. Religious truth and certainty is judged according to subjective experience and utility—if it feels right, or meets an individual's needs, then it is 'true.' This type of religiosity requires little commitment to any particular tradition and the individual can easily move from one set of beliefs and practices to another.

Religion of this type is heavily bound up with the commercial world, especially as presented in New Age philosophies. It generally appeals to older youth in their late twenties rather than to teenagers who cannot afford the more expensive esoteric practices on offer. Where younger age groups do take an interest in these nontraditional forms of religion, it often overlaps with entertainment, is not taken seriously, and competes with other leisure pursuits. As such it does not form an ultimate basis of identity.

3.3 *Religious Identity*

The remaining young people choose to belong to a particular religious group or church and find their

identity within a community of believers. This may be accomplished through a conversion experience to a new religious tradition, or it may be a reaffirmation of the religious tradition of childhood (sometimes this is conflated with broader issues of ethnic identity by way of 'cultural transition' or 'cultural defence', see Bruce 1996). Either way, the emphasis is on a conscious, autonomous decision to be a part of the group. Consequently, religious identity tends to be strong and salient.

In this respect, the churches which attract young people are often evangelical and charismatic, and are well suited to fostering religious identity. The conservative theology, strong social networks, and high degree of commitment these churches usually require encourages a dominant sense of group belonging around religious themes. A charismatic orientation to the group adds a further dimension in terms of religious experience which helps to authenticate the young person's beliefs, and allows for self-expression and a sense of intimacy with God and other believers. In this context a strong sense of religious identity is established.

4 *Summary*

In conclusion, over the last 30 years established institutional Christianity has generally become less influential in the formation of young people's identity. However, the rate of Church decline is beginning to slow down and religious ideas have not disappeared from society. Consequently, Christianity, independently or as part of a heterodox collection of beliefs and practices, does continue to inform youth identity to a certain extent. It remains to be seen whether the latter two patterns will overturn the general trend towards a secular identity among Western youth in the future.

See also: Adolescent Vulnerability and Psychological Interventions; Education: Cultural and Religious Concepts; Generations, Sociology of; Religion: Family and Kinship; Religion: Morality and Social Control; Religion, Sociology of; Religiosity: Modern; Socialization in Adolescence; Youth Culture, Anthropology of; Youth Culture, Sociology of; Youth Movements

Bibliography

- Ashford S, Timms N 1992 *What Europe Thinks: A Study of Western European Values*. Dartmouth Ltd, Aldershot, UK
- Brierley P 1997 *UK Christian Handbook Religious Trends No. 1*, 1998/99 edn. Christian Research, London
- Bruce S 1996 *Religion in the Modern World: From Cathedral to Cults*. Oxford University Press, Oxford, UK

- Clark W, Carroll J W, Roozen D A (eds.) 1995 *The Post-War Generation & Establishment Religion: Cross-Cultural Perspectives*. Westview Press, Boulder, CO
- Davie G 1998 God and Caesar: religion in a rapidly changing Europe. In: Bailey J (ed.) *Social Europe*, 2nd edn. Longman, London, pp. 231–53
- Francis L J, Kay W K 1995 *Teenage Religion and Values*. Gracewing, Leominster, UK
- Giddens A 1991 *Modernity and Self-Identity*. Polity Press, Cambridge, UK
- Halman L, de Moor R 1993 Religion, churches and moral values. In: Ester P, Halman L, de Moor R (eds.) *The Individualizing Society: Value Change in Europe and North America*. Tilburg University Press, Tilburg, The Netherlands pp. 37–66
- Harding S, Phillips D, Fogarty M 1986 *Contrasting Values in Western Europe: Unity, Diversity and Change*. Macmillan Press, Basingstoke, UK
- Inglehart R, Basañez M, Moreno A 1998 *Human Values and Beliefs: A Cross-Cultural Sourcebook*. University of Michigan Press, Ann Arbor, MI

S. Collins

Religion: Culture Contact

1. A Model of Encounter

In many studies of culture contact there is a perceived dichotomy between the acceptance and the rejection of Christianity. This model of religious change, often advanced by those who have a stake in defending religious boundaries, stems from false assumptions about the nature of the religious encounter. Its premise is of an essential or textual Christianity confronting an essential or ideal traditional (or folk) religion—the religion of the invader confronting the religion of the autochthon. But, Christianity has never been propagated as a monolithic entity encountering a homogeneous local society. A more appropriate model involves a number of modes interaction. First, there was the encounter of a variety of mission Christianities, often mediated by local agents, with a traditional religion which was open, dynamic, and plural in nature. And, this religious encounter cannot be divorced from its political and socioeconomic context. It needs to be situated within a second set of conjunctions between the forces of capitalist imperialism and local institutions. Moreover, those encountering Christianity never adopted it wholesale. Specific ethnic groups, classes, and social categories appropriated particular ideas and practices and new forms of popular or vernacular religion, both traditional and Christian emerged out of processes of gleaning, borrowing, and recoding. This article considers these various modes of interaction with specific reference to Africa but with a Latin American aside.

2. Scholarly and Popular Misconceptions about the Encounter of Religions

2.1 Syncretism

In order to arrive at the above model of religious encounter it is first necessary to deal with a number of scholarly and popular misconceptions about the process. First, there is the issue of syncretism. This is a contentious term, often used to denote inauthenticity, or the pollution of a notionally pure tradition by symbols and meanings taken from alien traditions (Shaw and Stewart 1994). In African studies, missiologists and theologians have used syncretism as ‘a term of disapprobation denoting the confused mixing of religions’ to explain the rise of movements of Christian Independency, such as Zionism in Southern Africa and Aladura in West Africa. In Latin American studies, syncretism has far less of a pejorative connotation. Here scholarship on culture contact has been coloured by the optimism born in the heyday of the American melting pot ideology and portrayed African derived cults such as Vodun or Santera as a celebration of indigenous agency in the face of colonialism and other forms of alien domination (Shaw and Stewart 1994; see also Rostas and Droogers 1993). Recently, anthropologists and social historians have discredited the idea of essential, untainted traditions, exploring instead the processes by which they have been invented or imagined. Where the study of syncretism has most value is when it is understood as a form of cultural politics in which religious practitioners engage in heated polemic about religious hybridity while seeking to claim the moral highground of authenticity.

2.2 African Traditional/Folk Religion

A second set of misconceptions surround the nature of African traditional religion which has been portrayed consistently as ahistorical in contrast to the dynamic religions of the book, Christianity and Islam. If missionaries prior to the twentieth century admitted Africans had a religion at all, most, but not all, saw it as no more than primitive superstition blocking the advance of Western civilization. While early twentieth century anthropologists such as Edward Evans-Pritchard and Meyer Fortes were better informed about African traditional religion they viewed it no less ahistorically. Critical of missionaries whom they argued were destroying ancient and intricate mechanisms of social cohesion, they wrote detailed studies on the function of traditional religion in maintaining the organic stability of African societies. Some of the next generation of missionaries such as Bruno Gutmann or Mabel Shaw (see below), inspired by this functionalist anthropology came to appreciate the supposed communality of African traditional religion and sought to engineer these values in an ‘adapted’ Christianity.

Historical models of African religion also found their way into scholarship beyond social anthropology. Scholars of comparative religion sought to respect all religious traditions and make points of connection between them. Thus, eminent comparative religionists described African traditional religion completely in experiential terms (Parrinder 1954). Leading African theologians have also asserted an historical model of African religion. Mbiti has contended that African religion has no future tense and possesses no founders, prophets, or converts. Africans shared the same underlying philosophy but each 'tribe' had its own religion into which its members were born. While celebrating this integration of religion and society, Mbiti argues that the boundedness of African traditional religion means that it is incapable of adapting to the increasing scale of interaction that colonialism has engendered, and will inevitably decline (Mbiti 1969).

The rise of African history as a discipline from the late 1950s onwards confronted ahistorical notions of African traditional religion. Particularly important has been the emergence of oral history. Some African religions such as the Mbona cult in southern Malawi have shrines maintained by priesthoods who retain extensive traditions referring back to founding figures and cycles of revival, cleansing and prophecy (Schoffeleers 1992). Moreover, the notion of the boundedness of African religious and social identity has been challenged by research on ethnicity. Many of the 'tribal' identities claimed by contemporary Africans are in fact colonial and postcolonial creations. Hence, the concept precolonial Kikuyu or Shona religion makes no historical sense.

African historians have now replaced the old model African religions as a total organic collectivity with a more dynamic and complex one. African cults symbolized and articulated social and economic relationships and thus their adherents were involved in a range of religious interactions that began at the level of family and faction but could extend across their polity and beyond. Guilds of men propitiating hunting spirits united disparate groups of hunters throughout a region. Cults of the land, sometimes operating across a number of polities, enforced ecological rules concerning clearing, planting and reaping. And cults of alien or a-social spirits established the rules of communication between traders and the communities with whom they interacted (Ranger 1987). This more open and pluralist model of African traditional religion helps explain its resilience and adaptivity through the twentieth century.

2.3 Missionaries and Local Agents in Africa (and Latin America)

Neither was African traditional religion 'crushed' by a destructive monolithic mission Christianity acting as

the handmaid of colonialism. Such a wholly negative view of missionaries not only overlooks the contribution of local agents to the foundation of the African church but it also ignores the variety of mission Christianities and complex relationships they had with colonial states and African societies. Missionaries were influenced by different theologies and they were as diverse as the historical contexts and social sources from which their movements arose.

The contrast between the earliest Catholic mission enterprises and those of the first nonconformist missionaries could not be greater. The first wave of Catholic missionaries to sixteenth century Congo were controlled and sponsored by the Portuguese state. Mission took the form of a single religious enterprise conceived of in terms of territory. Co-opting state machinery in a union of faith and politics, Portuguese Catholics sought to create Christian kingdoms through the conversion of elite (Gray 1990). The early nineteenth century nonconformist missions were an extension of the culture of voluntary association and private enterprise that underpinned industrialization. Nonconformists like the London Missionary Society (LMS) in southern Africa were not especially interested in converting kings and aristocrats but rather turning ordinary Africans into sober and industrious Christians (Hastings 1994).

Missionaries also differed on the meaning of conversion. High Church Anglicans from the UMCA, German Lutherans, and Catholics believed that salvation was received through the sacraments and the liturgy, and sought to Christianize entire communities. By contrast, those Protestants influenced by revivalism placed a greater emphasis on salvation acquired in an intense and personal way. Their emphasis on moral conduct and lifestyle led to greater separation of the convert from their former community. While many missionaries were ethnocentric and culturally arrogant others did make a sincere and concerted attempt to come to terms with African culture and religion. As early as 1855 Bishop Colenso of Natal wrote a cogent argument for toleration of polygamy in his *Remarks on the Proper Treatment of Polygamy*. By the twentieth century a whole spectrum of missionary responses to their local settings had emerged. There were imaginative attempts at the translation of Christian ideas into an African idiom. Mabel Shaw of the LMS made extensive use of image of God/Jesus as 'Chief' and the idea of the Christian community as a 'tribe' in her work amongst the Bemba. Another approach was the Christianization of traditional rituals so that they might find force in a Christian community. The most common example of this was the adaptation of traditional rituals of initiation. Here the endeavors to Christianize the traditional rite of passage by UMCA Bishop to Tanzania, Vincent Lucas of Msasi, have been well documented (Ranger 1972).

Indeed, while many missionaries to Africa have been criticized for their cultural imperialism, others

may be taken to task for their conservatism. The German missionary and anthropologist, Bruno Gutmann, believed that the growth of African personality and morality was connected to the essential forms of community life: clan, neighborhood, and age grade. Thus, he developed an 'indirect rule' approach to mission, highly attractive to the Tanzanian colonial state, which sought to create Christian communities through traditional institutions and leadership (Kaplan 1995). When the radical black American educationalist W. E. B. du Bois visited West Africa in 1923, he criticized the fashionable 'emphasis on the preservation of traditional culture.' This du Bois thought, formed part of a white man's plot to maintain the African as a willing but useful servant of the imperial cause.

Perhaps unexpectedly, the last great new wave of mission Christianity, following earlier Catholic and nonconformist waves, has been the most adversarial towards traditional culture and religion. This is the Pentecostal movement that arose in America and northern Europe, 1900–10, and collided with Africa and Latin America almost immediately after its inception, and in subsequent surges throughout this century (Maxwell 1999a, Freston 1995). In part a reaction against the liberalism and modernism of established church theology, Pentecostalism actively believed in the supernatural. But while taking African and Latin American belief systems seriously Pentecostalism nevertheless demonized them (Maxwell 1999b, Lehmann 1996). Traditional spirits, whether good or bad in local cosmological terms, were collapsed into the category of the demonic. Thus, Pentecostals have been characterized by their aggressive stance towards traditional practices, and the destruction of charms and fetishes as polluted objects. Yet, even this Pentecostal attack on things 'traditional' can be interpreted as a form of Africanization: a perpetuation of a tradition of cyclical societal cleansing.

Finally, those scholars who continue to insist on the alien nature of African Christianity fail to grasp the role of local agents in the Christianizing of the continent. Pioneer missionaries were responsible for vast tracts of territory. Few and far between, their priorities were first translation work and then the construction of mission infrastructure. Many simply struggled to survive disease and learn the local language. The work of conversion was left to local agents, freed slaves, and converts from already missionized areas. Because overstretched missionaries were unable to monitor their activities these Catholic catechists and Protestant evangelists had much room for creativity. There were also movements of Christians happening completely independently of missionaries such as the Christian refugees who fled the expansion of the Zulu Kingdom in the 1830s and carried Christianity with them into Lake Malawi region. Another example was labor gangs, such as

groups of Pedi in the Transvaal, South Africa in the 1840s–1860s, who encountered Christianity in their work place and spread it through the process of labor migration. Often missionaries would arrive in a village to find Africans already there teaching the faith.

3. A Latin American Comparison

Latin America did encounter a monolithic mission Christianity. For over three hundred years the subjects of Spanish and Portuguese empires were exposed to state-sponsored Catholicism. Yet, from the very beginning of these Iberian empires there was a debate within Catholicism about the nature of traditional belief. While Spanish imperialists exterminated Indian peoples on the pretext that they were idolaters or devil worshippers Bartholomé de Las Casas described their pagan practices as their own way of approaching the Christian God. Las Casas, a sixteenth century Dominican, also argued for a clearer separation of the church from the imperial project, while the Franciscans and later the Jesuits established local regimes which while patriarchal did offer more humane conditions than their secular coevals.

Latin American Catholicism could also embody regional and national identities. The cult of the dark faced Virgin of Guadalupe, which dates from the mid-sixteenth century, offered a range of protective and curative functions. The Virgin appeared on the banners of those who sought independence from Spain in the early nineteenth century (led by village priests) and was subsequently made universal patron for the whole of Mexico (Brading 1991). Official Catholicism, initially opposed to nationalist movements, took a generation to come to an accommodation with the new nations before becoming embroiled in a century of warfare between conservatives and anticlerical liberals or radicals.

The strength of the sacred canopy of Catholicism varied across the continent. Some of the Amazonian peoples remained 'unreached' until the arrival of twentieth-century Catholic and Protestant missionaries. In other instances the Catholic overlay remained thin. The Maya of Guatemala, who resisted for centuries, retain numerous traditional beliefs, praying to the ancestors as well as the saints, visiting the shaman alongside the priest. In Brazil, Catholicism is often cross-referenced with belief in African Gods imported along with African slaves. In contrast to the literature on African religion, scholarship on Latin America more often equates 'folk' or 'traditional' religion, not with non-Christian religion, but with a 'syncretistic' mix of Catholicism and surrounding religions.

The degree to which Latin American Catholicism drew from other religions rather than retaining a more distinctive identity is explained by combination of two

factors: first, its close relation to the institutions of the state; second, the absence of alternative and competing Christianities. Catholicism was the established religion. Even at local level the church was often a major landowner and the priest involved in local politics. Catholic ideology legitimated social relations of production on the *hacienda*. The local level Catholic representative was usually not a catechist but a foreign priest. In the absence of other Christianities and a strongly independent Catholic dynamic Latin Americans were far more selective in their response to Catholicism and that selectivity came to be accepted by the authorities. Thus, for instance, Ecuadorian Quichua used the cult of the Saints and different visions of the virgin to reach those extrahuman forces they believed dominated them. Prayers, ritual drinking, dancing, and the wearing of lavish costumes took place to influence God's intermediaries. The same process of selection, of course, occurred within African Christianity, but in an atmosphere of Christian pluralism, African Christian identities took on a sharper definition.

Latin American Catholicism, in its different forms, remained monopolistic until the late nineteenth century when liberal elites opened doors to north Atlantic Protestantism and migration. The first wave of missionaries were predominantly older Protestant denominations such as Methodists and Baptists, offering welfare and education, but throughout the twentieth century Pentecostal and conservative evangelical waves came to dominate. The opening up of the religious field was accompanied by the break-up of Catholic sociopolitical hegemony. Agrarian reforms, anticlerical regimes, the rise of the mega-city and the partial liberalization of Catholicism made alternative religious adherence possible.

The persistence of pre-Christian cults is supplemented by their re-invention by middle class intellectuals in search of an authentic religion. One such group is the Association of Concheros in Central Mexico who have fused Catholicism with a blend of notional religious traditions from peoples such as the Aztecs and Chichimecas (Rostas 1993). In the African context such invented religious traditions are not so much sources of spirituality as markers for ethnicities.

Doubtless, if Spain and Portugal had retained and expanded their mercantile empires in Africa then African Catholicism may well have more closely resembled the early Latin American type. But Spain and, to a lesser extent Portugal, did not figure highly in the 'Scramble(s)' and the accompanying Catholic impulse was generally of a non-Iberian variety, involving a plurality of Catholic missionary orders working in competition with Protestantism. Given the open and plural nature of African traditional religion, the varieties of mission Christianity and the contribution of local agents, the African encounter took a profoundly different form from the Latin American one.

4. *Resilience and Innovation in African Traditional/Folk Religion*

Throughout the nineteenth century traditional religion innovated in response to the array of changes that accompanied the slow penetration of first, mercantile, and later capitalist, relations into the African interior. The increasing scale of social interactions that these changes brought led local intellectuals to innovate to make sense of this new world breaking in. In some cases they developed the idea of the High God with a macrocosmic reach to match Africans' expanding horizons. In other contexts, cults of possession became more widespread linking people together over wide regions through their shared affliction by the same amoral spirit. In many places the colonial occupations of the 1870s–1890s coincided with and precipitated ecological crises. In what appeared to many as a millennial moment, great prophetic figures, such as the Shona Chaminuka medium, arose to interpret colonialism and capitalist imperialism in a way that offered their followers conceptual control over what was happening. Thus, the great transformations were not an alien imposition but part of an unfolding local narrative.

In a similar manner, the new religion of Christianity was understood in terms of the old categories of traditional religion. Some missionaries were seen as witchcraft eradicators, whose presence inaugurated movements destroying charms and fetishes (Maxwell 1999b). Other missionary pioneers, with their bags of simple medical cures were seen as traditional healers (Hastings 1994). Others still were welcomed as sources of revelation about the High God.

While specific forms of traditional religion, such as hunting guilds, declined under colonialism, other forms of traditional religion moved to the center of the stage. African religious leaders, such as the famous Kinjikitile in Tanganyika, offered their cults as means of uniting acephalous societies against German occupation. And during the era of decolonization, religious leaders once again helped coordinate resistance against colonial states. In Zimbabwe's war of liberation Shona spirit mediums moved popular understanding of myth and taboo to enable stranger/guerrillas to be legitimated in local terms as hunters and warriors (Lan 1985). Although the labor migration led to decline of some cults of the land, in other instances traditional religion became more ecologically orientated helping Africans to become peasants and articulating their objection to state intervention in their agriculture.

5. *The Foundation of Vernacular Christianities in Africa and Latin America*

Nevertheless, Christianity has made tremendous headway in Africa and Latin America. In Africa great movements of Christianity in Buganda, the Ivory

Coast, Igboland, and Manicaland occurred during the Scramble(s) and in the first decades of the twentieth century. But these grassroots initiatives often happened in a haphazard and unplanned manner far from missionary influence or control, and through this indigenous agency local or vernacular Christianities were founded.

Besides the significant contribution of African agents, vernacular Christianities had two other key defining features. The first was the seizing of the landscape by symbolically sensitive missionaries and African Christians. Mountains, rocks, caves, trees, and pools, which often had traditional religious associations, were resacralized in Christian fashion. Catholics turned them into shrines and grottos—places of pilgrimage. The Anglicans made holy ground through creation of cemeteries.

The second key feature of these vernacular Christianities was the adoption of Christian symbols and powers and Christian literacy because they had utility in helping Africans prosper and relate to the colonial economy. While traditional religion could become more ecological in orientation to help Africans become peasants, Christianity could equally fulfill that role. Catholic prayers, invocations, and rituals were perceived as efficacious in protecting crops and providing soil fertility. American Methodism with its strong emphasis on discipline, an intense personal experience, and gospel of the plough provided the perfect ideology for peasant entrepreneurs (Ranger 1987).

Christian approaches to healing were also important. Pentecostal and revivalist emphasis on the satanic causes of illness resonated with traditional personifications of evil and disease. Likewise, divine healing through the laying-on of hands resembled the practices of traditional healers, though, of course, Pentecostals claimed a different source of empowerment. Miraculous cures through veneration of the saints provided a Catholic counterpart to Pentecostal spiritual healing. St Thérèse of Lisieux was perceived as particularly powerful in Africa. Seminaries of African priests were placed under her protection and she was appointed 'regent' of at least one African country. Finally, African Christian movements were often animated by a grass roots adoption of literacy. For these Africans, literacy, schooling, and church were synonymous in representing progress (Ranger 1987).

Specific social categories were often responsible for the localization of Christianity. In the nineteenth century Africa, kings and chiefs often converted to extend their ideological legitimation and to benefit from the technological and diplomatic advice missionaries had to offer. But these traditional leaders were slow to change their ideas and practices. In both continents younger men and women often converted and changed the very culture these male gerontocrats wished to preserve. The conventional language of Christians: 'coming forward,' 'setting themselves

apart,' and declaring for a 'completely changed life' took on a new meaning when merged with Christianity's individualist conscience (Fields 1985). It offered younger men, returning from the workplace, legitimate reasons for ignoring the burdensome demands of traditional commensality, the Latin American *fiesta* (Earle 1992), or African fertility and harvest rituals. Likewise, it enabled more individualist peasants to separate themselves from the 'wasteful' and time consuming demands of their kin, and in the Latin American context, fictive kin: God children (Goldin and Metz 1991). Faith in Christ gave women recourse to another male authority other than that of the household head, and directly undermined elder male notions of patriarchy by demonizing the ancestor spirits which legitimated it. Likewise, the prayer band and the fellowship meeting offered a vital source of solidarity for women otherwise divided by exogamy, polygamy, and fear of witchcraft accusation, and an engine of resocialization for the wayward male. Pentecostalism offered its own unique source of empowerment. Gifts of prophecy, tongues, healing, and exorcism turned a marginal woman or man into someone of consequence (Gill 1990). Catholicism was frequently an intensely feminine religion. Its Marian emphasis offered women a double-edged model of liberation and resignation.

Material benefits also accompanied conversion. The church offered networks of contacts and solidarity to labor migrants entering the cities, mines, or farms, and it created new alliances of Christian kin within the village. Although Pentecostalism's doctrine of male headship legitimated gender inequalities, its emphasis on industry, sobriety, and sexual propriety diminished aspects of male sex roles that are harmful to women (Gill 1990). Throughout the townships, shantytowns, and villages of Africa and Latin America the puritan ethics advanced by Pentecostals, other non-conformists, and groups like the Watch Tower have pulled families and small communities out of cycles of poverty, or at least stopped them from slipping over the edge.

Where Christianity has not eroded African traditional religion, it has nevertheless often changed it. Christian ideas and practices have found their way into non-Christian religions. Contemporary African genealogies tracing the origins of significant territorial spirits may include references to Hebrew patriarchs such as Noah or Abraham as useful embellishments (Bourdillon 1988). And traditional concepts of God have likewise been sharpened by Judaeo-Christian ideas about the divine.

None of the above should for one moment suggest that missionaries have not been important. It is rather that their importance came in ways they did not expect. Missionary teachings on domesticity shifted gender relations and created new social classes out of missionized elites. And work on languages such as the creation of vernacular alphabets and grammars has

had an immense impact. Not only has it created written languages where none previously existed, but these in turn have helped create new ethnic identities in both continents. In Latin America, the influence of one organization in particular, the Wycliffe Bible Translators, cannot be underestimated. In Africa, literacy and schooling enabled Africans to form their own cultural movements and provided access to political ideas other than those peddled by colonialists.

6. Christianity in Contemporary Africa and Latin America

The process of culture contact of course continues. Despite the fact that the center of Christendom no longer lies in the post-Christian West but in the South, the Western impulse to mission has not abated. It has, however, changed in nature. First, the majority of contemporary missionaries are North American in origin, reflecting the rise of the US as the new imperial power. Second, these missionaries are usually evangelical, charismatic, or Pentecostal and believe that conversion entails an individualized born-again experience. Finally, this new strand of born-again Christianity has a strong affinity with the electronic media.

The arrival of this new type of conservative Protestantism, which exploded across Latin America in the 1960s, and across Africa a decade and a half later, raises the inevitable question of whether it is just another American cultural export (Gifford 1998). While America is the major cultural and missionary dynamo for the born-again movement there are other epicenters in Southern and West Africa and Brazil, each contributing their own 'prophets' and media packages to increase the heterogeneity of the movement. But more importantly, the born-again movement has always been conducive to indigenization. Its lack of tradition, its scorn of formal theological education, its reliance on lay initiative and management, its family and neighborhood cell structure, and its acceptance of the supernatural make it particularly responsive to local agendas. While its stress on releasing converts from the demands of kin, both alive and dead, and its emphasis on building nuclear families may smack of western individualism, these tendencies are useful to African (and Latin American) peasants facing the decline or break-up of the communal system and migrant workers coming to terms with the rise of the mega-city (Martin 1990).

Finally, culture flows do not just move from the West to the South, they are multidirectional. While African and Latin American diaspora have for centuries transported their religions to the West, contemporary Latin American and African Christianities impact on Europe and America in ever more profound ways. In the 1970s and 1980s Latin American

liberation theologians shook up Western theological establishments, while the spirituality of Latin American base communities inspired many Westerners working for justice and peace. Today, mission partnerships bring African Anglicans to British parishes and returning missionaries import Christian song and liturgy from the South into their home churches. Added to this is the growing appeal of African Pentecostal and charismatic leaders in Western preaching and healing circuits. Some of these leaders' churches also finance their own mission initiatives. The arrival of African missionaries in London's Oxford Street, or Brazilians in Portugal, brings mission Christianity full circle.

See also: African Studies: Religion; Christianity Origins: Primitive and 'Western' History; Colonialism, Anthropology of; Cultural Expression and Action; Education: Cultural and Religious Concepts; Globalization, Anthropology of; Religion and Economic Life; Religion: Definition and Explanation; Religion: Evolution and Development; Religion, History of; Religion: Morality and Social Control; Religion: Nationalism and Identity; Religion: Peace, War, and Violence; Religion, Sociology of; Religious Nationalism and the Secular State: Cultural Concerns

Bibliography

- Bourdillon 1988 Gleaning: Shona selections from Biblical myth. In: James W, Johnson D H (eds.) *Vernacular Christianity. Essays in the Social Anthropology of Religion*. Lilian Barber Press, New York
- Brading D A 1991 *The First America. The Spanish Monarchy, Creole Patriots, and the Liberal State, 1492-1867*. Cambridge University Press, Cambridge, UK
- Earle D 1992 Authority, social conflict and the rise of protestantism: religious conversion in a Mayan village. *Social Compass* **39**: 377-88
- Fields K E 1985 *Revival and Rebellion in Colonial Central Africa*. Princeton University, Princeton, NJ
- Freston P 1995 Pentecostalism in Brazil: A brief history. *Religion* **25**: 119
- Gifford P 1998 *African Christianity: Its Public Role*. Indiana University Press, Bloomington, IN
- Gill L 1990 Like a veil to cover them: Women and Pentecostal movements in Le Paz. *American Ethnologist* **17**: 708
- Goldin L R, Metz B 1991 An expression of cultural change: Invisible converts to Protestantism among Highland Guatemala Mayas. *Ethnology* **30**: 325-38
- Gray R 1990 *Black Christians and White Missionaries*. Yale University Press, New Haven, CT
- Hastings A 1994 *The Church in Africa 1450-1950*. Clarendon Press, Oxford, UK
- Kaplan S 1995 The Africanization of missionary Christianity: history and typology. In: Kaplan S (ed.) *Indigenous Responses to Western Christianity*. New York University Press, New York

- Lan D 1985 *Guns and Rain. Guerrillas and Spirit Mediums in Zimbabwe*. James Currey, London
- Lehmann D 1996 *Struggle for the Spirit: Religious Transformation and Popular Culture in Brazil and Latin America*. Polity Press, Oxford, UK
- Martin D 1990 *Tongues of Fire. The Explosion of Protestantism in Latin America*. Blackwell, Oxford, UK
- Maxwell D 1999a Historicising Christian independency: The Southern African Pentecostal movement. *Journal of African History* 40: 2
- Maxwell D 1999b *Christians and Chiefs in Zimbabwe: A Social History of the Hwesa People c. 1870s–1990s*. Praeger, Westport, CT
- Mbiti JS 1969 *African Religions and Philosophy*. Heinemann, London
- Parrinder E G 1954 *African Traditional Religion*. Hutchinson's University Library, London
- Ranger T O 1972 Missionary adaption of African religious institutions. In: Ranger T O, Kimambo I N (eds.) *The Historical Study of African Religion*. Heinemann, London
- Ranger T O 1987 Concluding summary. In: Peterson K H (ed.) *Religion, Development and African Identity*. Scandinavian Institute of African Studies, Uppsala, Sweden
- Rostas S 1993 The Mexica's reformulation of the Conchero's dance: the popular use of autochthonous religion in Mexico City. In: Rostas S, Droogers A (eds.) *The Popular Use of Popular Religion in Latin America*. CEDLA, Amsterdam
- Rostas S, Droogers A 1993 Introduction. In: Rostas S, Droogers A (eds.) *The Popular Use of Popular Religion in Latin America*. CEDLA, Amsterdam
- Schoffeleers J M 1992 *River of Blood. The Genesis of A Martyr Cult in Southern Malawi, c. AD 1600*. The University of Wisconsin Press, Madison, WI
- Shaw R, Stewart C 1994 Introduction. In: Shaw R, Stewart C (eds.) *Syncretism/Anti-Syncretism: The Politics of Religious Synthesis*. Routledge, London

D. Maxwell

Religion: Definition and Explanation

This article sets out the main forms explanations of religion take by reference to their scope and the extent to which they preserve or subvert the self-understanding of religious believers. At the maximum level of scope explanations of religion amount to comprehensive theories of religion. Typically, such theories also seek to subvert the way in which religious individuals understand their own beliefs and actions. The need for comprehensive theories of religion is shown to arise out of the methodological demands of social and behavioral science itself. Theories of religion need to demarcate their subject matter and hence depend on and give rise to definitions of religion. The article sets out the differences between the main styles of definition of religion (substantive, functionalist, experiential, and family resemblance) and explores the possibilities of finding common ground between them.

1. Explaining Religion

1.1 The Nature of Explanation

Explanations are answers to requests for understanding. To explain something may be to: (a) make it known in detail; (b) render it intelligible; or (c) account for it. Disciplines such as the anthropology of religion may explain a religious phenomenon by citing further facts about it, thus explaining it in sense (a). Or they may explain a religious phenomenon by setting out its connections with other, more familiar or easily understood social facts, thus explaining it in sense (b). These forms of explanation need not overturn, but may merely supplement, the motives and reasons the religious themselves acknowledge as behind their beliefs and behavior. However, the social and behavioral scientist may seek to account for a religious phenomenon (type (c) explanation) by giving its cause and in the citing of that cause may set aside the self-understanding of religious actors. For example, the anthropologist may seek to account for ecstatic religious experiences in terms of the social status, particularly social power, such experiences give to their possessors (Lewis 1989). If this factor is a sufficient explanation, then the agents' belief that they are genuinely possessed by gods or spirits is an illusion.

1.2 Radical Explanations in the Science of Religion

It is characteristic of the social sciences in the nineteenth and twentieth centuries that they have gone beyond seeking to explain particular religious phenomena. They have produced explanatory theories seeking to account for the existence of religion as such. Such theories hold that the tendency of human beings to engage in religion cannot be taken as a given; there is something problematic in such engagement. For example, an evolutionary theorist may see religious phenomena as 'survivals,' that is dysfunctional judged against the circumstances of modern society and thus as a remnant of a once adaptive means of organizing social life (Tylor 1903). A psychological theorist may see a striking correlation between religious symbolism and psychic needs, suggesting that religion is motivated by persistent and universal wish-fulfillment (Freud 1927/1978).

The explanations of religion thus produced are radical, comprehensive, and external. They embody the 'hermeneutics of suspicion' (Ricoeur 1970). Religious consciousness is treated as a form of false consciousness. Religious symbolism is a code. Ostensibly it speaks of gods and spirits. In reality its components refer to aspects of the human, social world. A psychological, sociological or anthropological theory of religion enables us to reinterpret and decode religious symbols by correlating them with

mundane forces and facts hidden from the conscious mind of the religious actor.

1.3 Methodological Skepticism

The assumption that religious thought and action cannot be taken at face value but must be explained from the ground up is in part a product of the philosophical skepticism about the gods and the limits of human knowledge engendered by the Enlightenment. Some argue that such skepticism, or naturalism, is required by any sound methodology for the human sciences (Preus 1987). The only explanatory factors the human sciences can know of and thus recognize are mundane ones. They can only produce explanations of religion that deal in natural causes. Hence, they must adopt an attitude of suspicion toward religion's own 'memory' of humankind's contact with supramundane realities, and reinterpret religious symbols accordingly.

1.4 The Cultural-Symbolic Paradigm

The arguments for methodological skepticism are crucial in supporting radical, comprehensive theories of religion in the social and behavioral sciences. Without methodological skepticism, the production of such theories is not mandatory for those unconvinced of the Enlightenment's case against religion. The case against methodological skepticism can be made from within a contrasting approach to the explanation of human, social phenomena: the cultural-symbolic (Byrne 1997). This has its ancestry in the stress on *Verstehen* as the appropriate and unique stance to take toward the human world advocated by German social theorists in the idealist tradition, such as Dilthey. On a cultural-symbolic approach, there is something that marks out the human world as different from the natural. Its objects—human beliefs, actions and institutions—are endowed with meaning independent of any sense social scientists impute to them by their theories. The meaning they have is a result of the fact that they are informed by concepts. A human act, for example, has the sense it has by virtue of it being done for a reason and being directed to an end. That sense is dependent on agent's own beliefs and concepts. A standard way of explaining a human act is to bring out the reason/end the agent had in performing it. Such an explanation will fail unless it uncovers the agent's reason/end. Human action is thus informed by concepts. We explain human acts by bringing out their background in the agent's beliefs and intentions and by making those beliefs and intentions intelligible by reference to the rule-governed forms of social life in which they are embedded. Crudely put: to explain religious phenomena in this way is to account for them in terms of how the world appears to religious actors themselves and in terms of

the forms of individual and social meaning these actors create in the institutions which they fashion (Winch 1958).

The cultural-symbolic approach preaches methodological agnosticism. Religious actors' beliefs relate to suprahuman and supramundane realities. These beliefs need neither be endorsed nor denied by the social scientist in bringing out their sense or relying on their explanatory power. Both their sense and explanatory power relies on the beliefs making action and thought intelligible in terms of how the world appears to these agents. Whether the world is as it appears to these agents is irrelevant. By the same token, the question of whether the referring expressions used in reporting religious actors' beliefs really do refer is irrelevant. It is enough that actors intend to refer to gods and like things by these expressions. Since methodological agnosticism in the explanation of religion is not committed to the truth of actors' beliefs or to the real existence of things corresponding to the referring expressions in reports of those beliefs, it is a thoroughly naturalistic approach to the interpretation of religion (Clarke and Byrne 1993).

The cultural-symbolic approach can admit that there are forms of religious life which function ideologically or are the product of illusion. Investigation may show that individual religious phenomena exhibit forms of human meaning-making which fail of a minimal rationality and are corrupt. In such cases the hermeneutics of suspicion may be appropriately applied (MacIntyre 1971).

2. Defining Religion

2.1 The Purposes of Definitions of Religion

The social scientific study of religion needs a definition of religion in order to demarcate the subject matter of its inquiries. Theories of religion require definitions of religion at this initial stage for a parallel reason. Definitions may also serve the purposes of theories of religion by suggesting an initial interpretation of religion on which the theory may build. A complete theory of religion will also give rise to a definition of religion by way of summing up what it has concluded about the essential nature or origins of religion. Definitions of religion offered at the first stage are operational ones; definitions of religion offered at the end of theorizing can be styled 'theoretical.'

2.2 Types of Definition

Religion may be defined in terms of the typical content of its beliefs. Substantive definitions are illustrated by Tylor's definition of religion as 'belief in spiritual beings' (Tylor 1903). Religion may also be defined by reference to the role religion plays in personal and social life or the structure of religious thought and

action. Such structural–functionalist definitions are illustrated by Durkheim’s definition of religion in terms of a symbols system enforcing a distinction between sacred and profane reality and uniting members of society into a common moral community (Durkheim 1912/1967). Experiential definitions demarcate religion by reference to a putative common, or core, experience religious actors participate in, as illustrated by Müller’s claim that religion consists in the ability to experience the infinite in the finite (Müller 1893). These three modes of defining religion are united in seeking necessary and sufficient conditions for the classification of an institution as religious. Family resemblance (or polythetic) definitions reject this search. Their advocates maintain that the things we call ‘religions’ form a loose class. There is no attribute or set of attributes which unites them all: there is merely a network of overlapping similarities (Southwold 1978).

2.3 Issues Dividing Definitions of Religion

Family resemblance definitions are divided from the rest about the degree of unity to be expected among the class of things the Western academy has chosen to call ‘religions.’ The remaining styles of definition illustrate the conflicting pull of issues of scope and clarity. A substantive definition of religion gives the concept of religion clear boundaries and increases its utility in the demarcation of social phenomena. However, the cost of that clarity is to restrict the scope of the class thus created, threatening thereby to rule out by definition the possibility of nontheistic religion. A structural–functionalist definition gives the concept of religion broad scope, defending the thought that religion can be seen as a universal human phenomenon, but at the cost, its critics will charge, of making the concept too vague.

A means of harmonizing the two dominant approaches (substantive and structural–functionalist) may be found in a moral definition of religion. A religion is a symbolic system (with associated beliefs and practices) which articulates the thought that there is a source of moral order behind the world, that is the thought that the realms of value and fact are ultimately united. There are important social and personal needs met by religion so defined and its symbols have a typical content. They must purport to refer to transcendent realities which function as the ground of the natural order and the source or guarantors of morality. These realities are typically, though not inevitably, modeled on human agents (Byrne 1999).

See also: Belief, Anthropology of; Durkheim, Emile (1858–1917); Freud, Sigmund (1856–1939); Functionalism in Sociology; Religion, History of; Religion: Nationalism and Identity; Religion, Phenomenology of; Religion, Psychology of; Religion, Sociology of; Religiosity, Sociology of; Weber, Max (1864–1920)

Bibliography

- Byrne P 1997 The study of religion: Neutral? Scientific? Or neither? *Method and Theory in the Study of Religion* 9: 340–51
- Byrne P 1999 The definition of religion: squaring the circle. In: Platvoet J, Molendijk S (eds.) *The Pragmatics of Defining Religion*. Brill, Leiden, The Netherlands
- Clarke P B, Byrne P 1993 *Religion Defined and Explained*. St. Martin’s Press, New York
- Durkheim E 1912 *Les formes élémentaires de la vie religieuse*. Alcan. Paris (1967 *The Elementary Forms of the Religious Life*. Swain J N (trans.). Allen and Unwin, London)
- Freud S 1927 *Die Zukunft einer Illusion*. Internationaler Psychoanalytischer Verlag, Leipzig (1978 *The Future of an Illusion*. Robson-Scott W D (trans.). Hogarth Press and the Institute of Psychoanalysis, London)
- Lewis I M 1989 *Ecstatic Religion*, 2nd edn. Routledge, London
- MacIntyre A 1971 Rationality and the explanation of action. In: MacIntyre A *Against the Self-images of the Age*. Duckworth, London
- Müller F M 1893 *An Introduction to the Science of Religion*. Longman, London
- Preus J S 1987 *Explaining Religion*. Yale University Press, New Haven, CT
- Ricoeur P 1970 *Freud and Philosophy*. Savage D (trans.). Yale University Press, New Haven, CT
- Southwold M 1978 Buddhism and the definition of religion. *Man* 13: 362–79
- Spiro M E 1966 Religion: problems of definition and explanation. In: Banton M (ed.) *Anthropological Approaches to the Study of Religion*. Tavistock Press, London
- Tylor E B 1903 *Primitive Culture*, 4th edn. Murray, London
- Winch P 1958 *The Idea of a Social Science*. Routledge, London

P. Byrne

Religion: Evolution and Development

Both evolution and development are ways of looking at the history of religion. Development implies some kind of directional change in the history of a particular religion, whereas evolution implies comparable directional changes in a number of religious traditions. Since histories usually describe directional change of some sort, the idea of religious development merges into the much larger field of the history of religions. Therefore this article will focus on religious evolution.

1. The Term ‘Religious Evolution’

The idea of religious evolution, like the idea of social evolution, has been much criticized in the twentieth century, although neither idea has been entirely abandoned. If the notion of religious evolution is to be salvaged it must be dissociated from several preconceptions of its nineteenth century proponents. These preconceptions include the idea that all religions must be fitted into a single (unilineal) pattern, that progression through a fixed series of stages is inevitable,

and that more advanced stages are 'higher' (ethically and spiritually better) than earlier ones (Comte 1830–1842, Tylor 1871, Spencer 1876–1997, Marett 1914, etc.). It is the notion of inevitable progress that has been rejected by twentieth century writers; the idea that religious change correlates with social change in comparable ways across traditions can still be defended as 'religious evolution.' Contrary to common belief, it is to the writings of Max Weber (1920–1921, 1921–1922/1978) that a defensible conception of religious evolution owes the most.

2. The Protestant Ethic

Weber began his study of religion with his famous *The Protestant Ethic and the Spirit of Capitalism* (1904–1905/1930). At first glance, it might seem that this book is simply a developmental history focusing on a dramatic change within a single religious tradition. However, it is clear from the place of the Protestant ethic argument in *Economy and Society* (1921–1922/1978) and in the *Gesammelte Aufsätze zur Religionssoziologie* (1920–1921) that Weber believed that ascetic Protestantism was an indispensable catalyst for the emergence of a new form of society, which he called modern capitalism. Never before the Reformation in the West was the hold of agrarian society, organized through patrimonial (including patrimonial bureaucratic) and feudal political forms, broken through. Once established, however, capitalism became a worldwide phenomenon, even though taking different forms in different civilizational areas. Thus Protestantism, though occurring in only one tradition, played, indirectly but crucially, a developmental role cross culturally, and thus can legitimately be considered an instance of religious evolution.

Since Weber's time there has been a search for functional equivalents of Protestantism in other traditions that prepared some societies better than others and some groups in some societies better than others for the transition to capitalism, or, perhaps better, to modernity (Bellah 1957, Eisenstadt 1968). Ernest Gellner made the interesting suggestion that there were elements in all the great traditions, Islam in particular, that might have led to a Protestant-like breakthrough, but that a combination of fortuitous circumstances in Western Europe made it possible for it to occur there (1988). This raises the interesting question of what gave the great religious traditions a rationalizing potentiality even if it was only fully realized in one case.

3. The Axial Age

In his comparative studies in the sociology of religion (1920–1921) and in the sociology of religion section of *Economy and Society* (1921–1922/1978) Weber

pointed out that in each of the world religions, sometime in the first millennium BC, there emerged prophets or saviors who radically rationalized previous forms of what he tended to call magical religion. In each case the emergent figure (Confucius, the Buddha, the Hebrew prophets, Socrates, Jesus) preached a systematic form of ethical conduct quite different from the diffuse ritual and sacramental practices that preceded them. By calling these new symbolic forms 'rationalized' Weber was pointing to the fact that they were more coherent, more cognitively and ethically universalizing, more potentially self-critical (reflexive), and more disengaged from the existing society than what preceded them.

Weber's friend, Karl Jaspers, spoke of this period as the Axial Age (1948), and S. N. Eisenstadt has contributed significantly to the further study of what he calls Axial civilizations (1986 and elsewhere). Following Weber one can characterize each of the Axial breakthroughs (first millennium BC China, India, Israel, Greece) as involving, to different degrees and in different ways, an element of world rejection (Weber 1920/1946). Religious renouncers (Dumont 1980) in each case called the empirical world into question and, by invoking a transcendental realm, provided a significant degree of tension between worldly interests and religious demands that enhanced the possibility of social change. Although Weber himself contributed to the stereotype of static traditional societies, in his studies of particular traditions he discerned significant possibilities of social transformation (Weber 1920–1921, 1921–1922/1978).

Subsequent scholars have pointed out the dynamic potentialities in all the great traditions. What scholars such as Gellner have emphasized is that a radically rationalized ethic of conduct, such as that found in Protestantism, was not in itself enough. As long as the existing military, political, and religious hierarchies of agrarian societies could contain these rationalizing potentialities within monastic orders or private conventicles, a new type of society could not emerge. What made Protestantism successful was not just its unique ethic but the possibility of gaining sufficient political influence (in Geneva, the Netherlands, England, and New England) to institutionalize that ethic in the major spheres of social life.

4. The Shape of Religious Evolution

If Weber had at least two stages of religious evolution, namely an Axial stage correlating with the emergence of historic agrarian societies, and a modern stage for which ascetic Protestantism was an indispensable precondition, did he have any more stages? In looking at his historical sociology as a whole it is clear that he did. The world religions were preceded by societies organized primarily in terms of kinship and neighborhood and in which religion was primarily magical, by

which Weber meant, one in which ritual means were used to attain utilitarian ends such as rain, good harvests, human fertility, etc. While such a three-stage typology is for many purposes adequate, it does seem possible to refine it further.

Bellah, in his essay *Religious Evolution* (1964), suggested a further differentiation of the early stages. He distinguished between a 'primitive,' now better, 'tribal,' phase, and an archaic phase. Myth and ritual characterize tribal religion—indeed they may be co-terminous with hominization itself—but they are myth and ritual of a particular type. The spiritual beings whose activities are narrated in the myths are not gods and are not worshipped. These half-human, half-animal, often ancestral, beings express in archetypal fashion the social and ecological life of the people. Ritual involves the reenactment of the primordial events such that mythical time becomes present, not the worship of distanced deities. Tribal religion is most characteristic of hunter-gatherers, but is also common in horticultural societies. With agriculture and the emergence of complex chiefdoms, and particularly with the early state, archaic religion comes into being. While in tribal societies every clan can be traced to an ancestral being, in archaic societies only the aristocratic lineages claim divine descent. A hierarchic pantheon of gods appears, quite different from the mythical beings of tribal religion. The gods require worship, and, significantly, sacrifice, in many cases human sacrifice, and a differentiated class of priests who supervise the worship. In archaic societies the ruler often claims descent from the highest god, even the claim to be a living god himself. In general, archaic religion is characterized by a fusion of political and religious hierarchy largely missing in tribal society. Examples of archaic societies, often identified as Bronze Age societies, would include third and second millennium BC Mesopotamia, Egypt, the Indus Valley, and North China, but also early first millennium AD Japan, and first millennium AD Meso-America and Peru.

It was in the midst of or on the peripheries of such archaic societies that the axial religions of the first millennium BC emerged. They took the hierarchical tendencies of archaic religion in a radically transcendental direction. All value was seen to reside in a realm beyond this world and thus this world and all the worldly hierarchies within it were relativized and potentially delegitimized. But in most cases the new Axial religions either remained marginal to the centers of power or, if they were recognized by the ruling center, they made compromises with it so that they largely legitimated that center rather than challenged it, thus allowing a quasi-archaic situation to continue. This is not to minimize, however, the importance of Axial religions in the elite education and culture of each of the great traditions. Also, the critical potential of the Axial religions was continually reasserted in the form of heresies, uprisings, and rebellions (Eisenstadt

1986, 1999), but these were normally either destroyed, marginalized or, if successful, coopted by the existing structures of power.

A simplified way of looking at the shape of religious evolution was proposed by Bellah (1964). He argued that whereas tribal and archaic religions were primarily this-worldly in orientation, which is what Weber, perhaps unwisely, meant by the word magical, the Axial religions were world-rejecting (and thus, for Weber, very importantly, magic rejecting), although they differed as to whether this rejection was to be worked out within the world (ethically), or as far as possible outside the world (mystically). It should be remembered that in its radical consistency Axial religion was never more than the religion of a minority; the majority continued to entertain beliefs and practices continuous with archaic or even tribal religion, which is what Weber meant by the return to the garden of magic (1920–1921).

With the Protestant Reformation, however, the belief in a radically transcendent God had dramatic this-worldly consequences: the consistent demands of an Axial ethic were to characterize every sphere of daily life. But in the subsequent development of modernity, though the this-worldly dimension remained dominant, its transcendental basis became transformed into immanentism, thus returning the modern world in a much different way to this-worldly immanentism of pre-Axial times. Weber clearly observed this transition, but viewed it almost entirely negatively. The modern world of rationalization would run on its own bureaucratic and economic energies without any transcendental sanction, would become an iron cage (1904–1905/1930). Charles Taylor has described the same transformation in less somber hues as 'the affirmation of ordinary life' and has stressed growing individuation (1989). In any case, though the world of the Protestant Reformation and the Western world since the eighteenth century (and, more recently, much of the rest of the world, as well) are clearly different, it is doubtful whether the Protestant phase needs to be designated as a separate 'early modern stage' as Bellah did in 1964, or whether it should be seen as merely the opening transition toward modern religion, the contours of which, though it has been developing for several centuries, are still far from clear.

The shape of religious evolution should not obscure its many irregularities, its lack of unilineal consistency. Regression was always a possibility: Axial Greek civilization did not emerge from the womb of archaic Minoan and Mycenaean societies, but only after the breakdown of those societies and the return to tribalism. Stages were sometimes skipped, especially on the peripheries of advanced societies: the Israelite Axial breakthrough began in a tribal confederation on the boundaries of archaic empires; Muhammad grew up in the tribal city of Mecca but founded an Axial civilization, to be sure under Jewish and Christian

influence. Not only was every religion from tribal to modern full of internal tensions and divisions but all of them were open to continuous external influence.

5. *The Causes of Religious Evolution*

To move beyond description and consider causation is always difficult in historical sociology; nevertheless, how one answers the question of causation, even tentatively, will determine how one understands the entire process. Religious evolution is certainly embedded in social evolution and cannot be understood apart from it. Long-term tendencies, such as the increasing division of labor, the increasing capacity of states to mobilize social power, the development of media such as money and writing, the differentiation of elites of various types as carriers of distinct cultural traditions, the tendency of pre-modern societies to oscillate between political centralization (patrimonialism) and decentralization (feudalism) all influenced and were influenced by religious beliefs and practices.

There has been a tendency among supporters of the idea of social evolution, Marxist and non-Marxist alike, to emphasize economic factors as causal, and religious developments as epiphenomenal. It has been common to believe that Max Weber reversed this understanding and held that religion was primary in social causation. A reading of Weber's entire work, which puts *The Protestant Ethic and the Spirit of Capitalism* in its proper perspective, makes it clear that such was not his belief. Rather, Weber argued for the relative independence of the religious dimension as a factor in social life, but one that always interacts with other factors, notably economic, political, and legal. S. N. Eisenstadt rephrases Weber by showing that it is not just a matter of emphasizing multi-causality: while never sufficient as a causal factor in itself, religion provides the indispensable element of *cultural codes*, not only for the institutionalization of new social forms but, when internalized in personality systems, for new forms of individual motivation as well (Eisenstadt 1999).

Religious elements have played a critical role in all the great social transitions. Terrence Deacon has even argued that the origin of language among the early hominids was probably facilitated by what he specifically calls ritual (1997). Some primates today can learn words, but only under the most intense pressure of human teaching. Deacon hypothesizes that the bodily and repetitive nature of ritual may have been a necessary precondition for the counterintuitive, from an early primate point of view, shift to symbolic representation that is the basis of language. In any case, Paleolithic evidence, difficult to interpret though it is, suggests that ritual played a central role in creating the social solidarity and respect for the

normative order of kinship relations that differentiate humans from other primates.

The transition to archaic society is associated almost everywhere with monumental temple architecture, large-scale sacrifice, and divine kingship. It would seem that the move from the basic egalitarianism of tribal society to the intensely hierarchical and centralized order of archaic societies was only possible with the institutionalization of new religious codes, in association with an unprecedented concentration of social power.

While causal linkages are far from clear, the first millennium BC did see a number of developments that might have prepared the ground for the rise of the Axial religions. These include the beginnings of a market economy, an increase in bureaucratic and military power, and increasing differentiations of social classes, all of which could have given rise to Weber's 'problems of meaning' (1921–1922/1978). But an indispensable precondition for these radical new religious developments was the existence of cultural elites at least to some degree independent of central political power. These include the Hebrew prophets, the Greek philosophers, the Chinese sages, and the Indian ascetics and mystics. It was among these groups that archaic symbol systems were radically reinterpreted in a transcendental ethical and spiritual direction.

The influence of the Protestant Reformation on the emergence of modern society has been the subject of a vast literature. A significant recent development, however, has been the recognition of the importance of a 'disciplinary revolution,' particularly associated with the Calvinist wing of the Reformation (Gorski 1993). The disciplinary revolution operated at a number of levels: a new code of systematic ethical action expected from believers, new institutions capable of enforcing the code, and new solidary organizations strong enough to influence political and economic developments. The transformation of these new codes under a variety of influences and pressures into such tendencies as utilitarianism, liberalism, Jacobinism, and socialism, and more recently, individualism and fundamentalism, has significantly affected the shape of and the tensions within modern society.

Thus, viewing the process as a whole, it appears that religious evolution is not only a subject of great interest in itself, but is a critical dimension of social evolution.

See also: Civil Religion; Comte, Auguste (1798–1857); Reformation and Confessionalization; Religion: Definition and Explanation; Religion: Nationalism and Identity; Religion, Phenomenology of; Religion, Sociology of; Religious Fundamentalism: Cultural Concerns; Secularization; Social Evolution: Overview; Spencer, Herbert (1820–1903); Weber, Max (1864–1920)

Bibliography

- Bellah R N 1957 *Tokugawa Religion*. Free Press, Glencoe IL
- Bellah R N 1964 Religious evolution. *American Sociological Review* 29: 358–74
- Comte A 1830–42 *Cours de Philosophie Positive*. Bachelier, Paris
- Deacon T W 1997 *The Symbolic Species: The Co-evolution of Language and the Brain*, 1st edn. Norton, New York
- Dumont L 1980 *Homo Hierarchicus*. University of Chicago Press, Chicago
- Eisenstadt S N 1968 *The Protestant Ethic and Modernization: A Comparative View*. Basic Book, New York
- Eisenstadt S N (ed.) 1986 *The Origins and Diversity of Axial Age Civilizations*. State University of New York Press, Albany, NY
- Eisenstadt S N 1999 Axial age sectarianism and the antinomies of modernity. In: McKnight S A, Pierce C I (eds.) *Politics, Order and History*. Sheffield Academic Press, Sheffield, UK
- Gellner E 1988 *Plough, Sword and Book*. University of Chicago Press, Chicago
- Gorski P S 1993 The Protestant ethic revisited: Disciplinary revolution and state formation in Holland and Prussia. *American Journal of Sociology* 99: 265–316
- Jaspers K 1948 *The Origin and Goal of History*. Routledge and Kegan Paul, London
- Marett R R 1914 *The Threshold of Religion*, 2nd edn. Methuen, London
- Spencer H 1876–1897 *The Principles of Sociology*. D. Appleton, New York
- Taylor C 1989 *Sources of the Self*. Harvard University Press, Cambridge, MA
- Tylor E B 1871/1974 *Primitive Culture*. Gordon, New York
- Weber M 1904–1905/1930 *The Protestant Ethic and the Spirit of Capitalism* [trans. Talcott Parsons]. Scribner, New York
- Weber M 1920–1921 *Gesammelte Aufsätze zur Religionssoziologie*. J. C. B. Mohr (Paul Siebeck), 3 Vols. Tübingen, Germany
- Weber M 1920/1946 Religious rejections of the world and their directions. In: *From Max Weber: Essays in Sociology* [trans. Gerth H H, Wright Mills C]. Oxford University Press, New York
- Weber M 1921–1922/1978 The sociology of religion. In: Roth G, Wittich C M (eds.) *Economy and Society*. University of California Press, Berkeley, CA

R. N. Bellah

Religion: Family and Kinship

This article will consider how these institutions, with their associated ideas and practices, intersect with one another. Family and kinship are closely related concepts, which together incorporate a number of issues, such as marriage, reproduction, family roles—particularly mother, father, children, and grandparents—descent and alliance, inheritance and taboo. Religion refers to the sum of beliefs and practices, ideas and attitudes, movements and institutions, symbols and experiences by which individuals and groups give meaning and order to their world, often with

reference to a supernatural being or power. How has religion been involved in the social organization of kinship and the family, and how have issues surrounding kinship and the family affected religious beliefs, attitudes, and practices? In what ways have kinship arrangements and the family served the interests of religions? A range of different religions, family types, and kinship arrangements will be referred to, and contemporary debates about the relevance and ethics of the family, to which religious voices have contributed, will be discussed.

Most of those who have written about the institution of the family comment upon its resilience. Although a variety of family types have been in evidence in different historical periods and societies, families have always existed in order to create and strengthen alliances between groups and to provide a suitable location for reproduction and the nurture of children. A small family unit may incorporate a single parent and child, with larger units incorporating several generations of adults including siblings, and their children. In some societies the idea of family has referred to all those who live in a household, including pseudokin and servants. Kinship, though some have limited its definition to ties of blood, is generally held to refer to systems of both descent and alliance, and to both biological and social relationships. Anthropologists who have researched kinship have been interested, for example, in how certain groups have sought to maintain their blood-line through the generations, or have organized the family to pass down wealth through the male line, in how some groups have developed kinship arrangements to ensure rights over children through the control of women, and how others have used marriage as a means of creating powerful new kinship bonds. Such studies have often involved understanding how religious ideas or ritual practices have been used to reinforce or symbolize such arrangements. Arnold van Gennep, for example, in his famous study of rites of passage published in 1960 [1908], showed how a Turko-Mongolian group called the Bashkir used rituals to mark the important economic and social aspects of marriage. He described the extended visiting and eating rituals associated with the payment by the groom's family of the bride price, and showed how a lengthy game of hide and seek on the marriage day signified the loss of a daughter to one family and her gain by another.

Social scientists have also shown how religious ideas and practices have reinforced common conceptions of kinship and family life. Taking the example of the Christian Right in America in the 1980s, they have shown how religious groups were able to contribute to the formation of a popular family ideal. They have shown, however, that this ideal—of a nuclear family comprised of a heterosexual married couple with children—is not the traditional institution its supporters believed it to be so much as a postmodern

fantasy. 'The family is dead, long live our families,' a phrase used by Judith Stacey, a feminist sociologist, points to the redundancy of this popular ideal in the face of the complex reality of historical and living families in different social settings. How other religious groups have contributed, not only to reinforcing, but also to challenging ideas about the family, will be the subject of section 1.5.

1. The Role of Religion in the Organization of Kinship and the Family

1.1 Kinship and Inheritance

Religious ideas, myths, and practices have had a role in many societies in unholding social institutions, and many social scientists have sought to understand this. One example of this is the Indian caste system, which captured the attention of Max Weber and later anthropologists such as Louis Dumont. Weber held that there was a necessary relationship between caste and religion in India, that without caste there was no Hindu. He drew this conclusion because he saw that the religious rites and duties of Hindus were related to their social rank through a system of closed status groups or castes. These were different to tribes and guilds by virtue of their rules regarding marriage and eating. In particular, Weber noted the importance of the dominant group, the Brahmins, a priestly caste, around whom the entire system was organized. He recognized, however, that the modern caste system had grown to incorporate castes comprised of religious groups other than Hindus, and also suggested that it was severely shaken if not undermined by the challenge of modernism.

Later scholars considered aspects of the caste system in more detail. Dumont focused on hierarchy, examining the use of rules and practices concerning purity and pollution in distinguishing and ranking different castes. Others challenged his view, suggesting that those able to wield power in the system were not necessarily those who were most ritually pure. Most scholars have accepted some kind of relationship between religion and caste, however, acknowledging the role of scripture (especially the *Laws of Manu*), the function of the Brahmins in authorizing the system, the role of ritual in purification, and the part played by religious groups in both manipulating caste for advantage and challenging it.

Just as Indian kinship arrangements were legitimized by religious texts and commentators, so were those in other societies. What came to be understood as the Confucian family ideal—*hsiao*, the virtue governing family relationships—was first prescribed in the classical texts of ancient China such as the *Book of Changes* and the *Book of Rites*. Within Islam, the family was privileged above other social institutions in the *Qur'an* and in later Islamic legal writings.

That European Catholicism reinforced ideas of feudal kinship and the family has been shown by the sociologist Bryan Turner among others. Developing the historical materialist ideas of Friedrich Engels on the family, he stated that the teachings of the Catholic Church in the Middle Ages on sexuality were important for reinforcing the norms of monogamous marriage and the control of women's sexual behavior. These norms, along with laws concerning inheritance through the male line, enabled dominant feudal families to ensure the stable transmission of property from one generation to the next and to avoid the dilution of the estate's wealth through illegitimacy. The Church further served this system through its Religious Orders which were often a destination for those excluded from inheritance, such as younger sons and unwanted daughters.

Religious teachings have played a part in social control at many times and in many societies. Women have often been the focus for this because of the importance of their reproductive role. For example, within Islam, polyandry (a wife taking more than one husband at a time) is prohibited, whereas polygyny is accepted, with the proviso that a husband should be able to provide equally for all his wives (a maximum of four) and children. This may be explained biologically because a woman's reproductive capacity is limited by the gestation period of the embryo, whereas a man may father multiple children simultaneously, thus reproducing his blood-line several times over. Socially, it may be explained by the fact that polygyny was the norm in Arabia at the time of Muhammad. Prevailing social practice was endorsed by the *Qur'an*, and, since that time, some Muslim marriages have been polygamous, particularly in African countries where polygyny was prevalent before the arrival of Islam. Many voices, including religious ones, have spoken out against this practice, but its place in Islamic scripture is unalterable given that the *Qur'an* is Allah's revealed word.

1.2 Marriage

The teachings and rituals associated with marriage are indicative of the value attributed by religions to the family as a significant institution. The family is given legal privilege in Islam, and in Judaism, and Sikhism it is of central importance. In Christianity, the family competes with the Church for priority, and, in Buddhism, with monastic alternatives. The presence of the ascetic life—lived alone or in community—in many religions challenges the adequacy of the family as the best place for spiritual progress, liberation, or salvation. In two modern, global sectarian movements, the Church of the Latter Day Saints and the Unification Church, the family is not only prioritized socially, but is of central theological importance in defining the nature of both the cosmos and divine-human relationships.

The social meaning of marriage is the alliance between two individuals and often between the families or clans of which they are a part. To avoid incestuous relationships, all societies have rules about who may marry whom. In many countries marriage alliances are legitimized in law. Religious understandings of marriage differ, although they all build upon these basic social requirements. In Judaism and Islam, marriage is understood as a contract divinely sanctioned by God, while in Christianity it is a sacrament. In Hinduism, it marks the beginning of a stage of life. Rituals of marriage often symbolize the social and religious ideas embodied by tradition. For example, the exchange of rings, the use of a marriage canopy, the act of circumambulation—together around a fire or of one partner around the other—and the sharing of garlands or food are symbolic of the nature of the marriage relationship, the intention of the partners, and, in the Jewish case of the breaking of a glass, of the fragility of the institution.

Religions have gone beyond these definitions and symbols to reinforce prevalent social ideas about marriage in their teachings, as the case of polygyny in Islam showed. Cousin marriage too is endorsed by Shar'iah or Islamic law, and is also accepted in Australian traditional religions, in order to maintain close, stable relationships and to give sanction to the retention of property within the group. Hindus, though expected to marry within their caste, are prohibited from marrying someone with a common grandparent, thus ruling out close cousin marriages. Arranged marriages, where parents or guardians select marriage partners for sons and daughters, are not obligatory in Hinduism or Sikhism, but they have become a social norm recognized by the majority of Hindus and Sikhs if not always adhered to by them.

Sexual partnerships outside marriage, both heterosexual and homosexual, have frequently been the subject of religious injunctions, often constituting a taboo or ritual prohibition. The breaking of marriage bonds or divorce, similarly, has been an issue where religious teachings have contributed to social control. The use of the *talaq* or marriage repudiation in Islam and of the *get* or bill of divorce in Judaism, both practices which endorse the rights of the male party over the female, have caused consternation among activists campaigning either for equal rights or simply for fairness in situations of relationship breakdown. However, both of these religions have made some provision for the property rights of women during marriage and after divorce, or the death of their husbands, in Judaism through the *ketubah* or marriage contract, and Islam through the *mahr* or dowry provided by the husband. In contrast to these two religions which, owing to their contractual view of marriage, have endorsed divorce, Catholic Christianity has eschewed it, seeing it as the breaking of a sacrament. Christian denominations in the West in the twentieth century, in the face of the increasing divorce

rate, sought to reassert the value of marriage and to encourage reconciliation in the case of marriage breakdown. They were often divided on the question of whether second marriages should be sanctioned.

1.3 Roles and Relationships

The biological functions normally associated with the social unit of the family are reproduction, nurture of children, and preservation and control of fertile adults, particularly women. At different times and places, a variety of social arrangements have been in operation to facilitate these. Responsibilities for children, whether the provision of food and other resources, or of nurture, have been variously divided among men and women, parents and grandparents, though women have had a mothering role in most societies. Perhaps the most commonly recognized family roles have been those of mother, father, children, grandparents, siblings, and ancestors. Religious beliefs and practices have played an important part in endorsing such family roles and the duties associated with them. Roles have often been conferred through rituals, affirmed by religious teachings, and sometimes mirrored and authorized by mythic accounts of divine family roles.

The use of ritual in celebrating entry into different life-stages, conferring appropriate roles, and ensuring group membership, continuity, and loyalty has been the subject of anthropological debate. Functionalists, such as Durkheim and Malinowski, and later, Rappaport, stressed the importance of ritual for social integration, whereas Edmund Leach held the view that rituals reflected conflicting interpretations and often contained the seeds of social instability while legitimizing a particular power structure. This ambiguity was also acknowledged by Victor Turner who held that the symbolic complexity of rituals made space for competing interpretations, thus enabling solidarity to be achieved without compromise. These scholarly perspectives were significant for understanding many types of ritual, including those related to the family. For example, in his study of the Ndembu of Zambia, Turner described the various symbolic meanings of the milk tree, for girls at the time of initiation, for their mothers, and for the society of which they were a part. It symbolized breast milk and the mother-child relationship, the exclusion of men from women's experience, but also the separation of the growing daughter from her mother. Turner suggested that this tree constituted a dominant symbol which enabled the expression both of dissonant interests and of the solidarity of the group as a whole.

Religious rituals and symbols, then, may sometimes affirm divergent viewpoints about family roles and relationships for the social good. Religious teachings, however, are often less ambiguous, constituting normative statements about family life and the roles and duties of mothers, fathers, and children. Householder

life has been prescribed by religious and moral texts in many societies. A nineteenth century example of this may be found in the schoolbooks produced during the Meiji period in Japan when the imperial ideology of State Shinto was gaining its fullest expression. The emperor was held to be the supreme father or head of house, guiding his family, with each individual family being a branch of the same tree, following the same structure. This patriarchal ideal, which came to apply also to corporate families, was authorized by the Shinto myth of the sun goddess, Amaterasu-o-mikami, of whom the emperor was held to be a direct descendent.

A more recent example, which builds on earlier Confucian family ideals, may be witnessed in the teachings of the Unification Church, a movement which began in Korea in the 1950s and was popularized in the West in the 1970s. According to the *Divine Principle*, the movement's founder, Sun Myung Moon, and his wife are the true parents of the world. They seek to heal the broken family, bringing people together regardless of social or religious background. With God as the father of all creation, Jesus as the true spiritual father, and the holy spirit as the true spiritual mother, the movement mirrors this divine family in the human family, with God as the focal point, the male as head of the family, and the mother subject to her husband. With the essential fourth component, their children, they are expected to prepare, through active missionary service, for the second advent. The Unification Church is an example of a religious group for whom the family ideal is central in its theological, ethical, and social teachings, and for whom family relationships are the model for all other relationships.

1.4 The Significance of the Family for Religions

Having seen how religious ideas, practices, and institutions have contributed to forming and legitimizing kinship arrangements, marriage practices, and family roles and relationships, the significance of the family for the maintenance and continuity of religions will now be considered. As was seen in the case of the Unification Church, the family motif has, at times, been an organizing principle for ordering and giving meaning to both divine-human relationships and social life. Other late-modern religious movements, taking their inspiration from Christian rather than Confucian models, have likewise focused on the family as the central expression of community life. The Church of the Latter Day Saints, or Mormons, has seen the eternal family unit (including ancestors) as the primary institution for guiding the individual in his or her spiritual journey. Parents are believed to have duties, not only to care for children, but also to ensure their religious nurture and to confer priesthood upon them. As spiritual progress depends on membership of an eternal family, sex and procreation outside mar-

riage are prohibited, and an active interest is taken in family history and genealogy. The Church's theological and ethical view of the family has led to particular family patterns developing among its members, e.g., a low marital age, large families, support of elders within families, a low divorce rate, and conservative sexual values.

Some Christian denominations and sects depend upon commitment to the family to maintain the group and pass on its beliefs and practices from one generation to the next. In the nineteenth century, Protestant sects in the New World such as the Hutterites, the Mennonites, and the Amish ensured their continuity through the generations primarily by means of the social control of families, marriage within the sect, ethical conservatism, and the religious nurture of children within the family. For some peoples, religious transmission across the generations has been related to ethnicity, itself determined by kinship arrangements, intergroup marriage, and close family ties. Despite having differing myths of origin and notions of identity, Jews, Hindus, Sikhs, and Zoroastrians have been defined by their relationship to a land or a people, and have carried their religious ancestry as a mark of birth. The family has been important for these groups in preserving biological and social ties with others in the group and with ancestors, and in maintaining the group for the future. Current concerns in Jewish and Zoroastrian communities about out-marriage are an example of the importance of family continuity. Additionally, debates within all these groups about the validity of the conversion of outsiders are indicative of the stress they have placed upon ethnic exclusivity. Within Orthodox Judaism, for example, Jewish identity and culture are only transmittable by the mother. If a Jewish woman marries a non-Jew, their offspring may return in the future to the Orthodox Jewish fold; if a Jewish man marries a non-Jew, their offspring may not. This represents an end to the Jewish line. Liberal, Reform, and Conservative Jews have a range of views on this and on the subject of conversion to Judaism.

The family is a major bearer of religious tradition. A new generation of young people may be brought into the life of a faith community by parents and grandparents, thus ensuring the survival of the religion. However, this process of religious reproduction introduces mutation and adaptation. New generations and changing times often require the reinterpretation of traditions, and the introduction of innovation, even if this is within existing religious frameworks. Many scholars of religions have been keen to understand intergenerational religious change, and to examine the methods of religious socialization used within various religions at different times. It can be instructive, for example, to compare the way in which one religious tradition is transmitted differently by families in different locations through variants on the same story, the use of local languages and rituals, and the

involvement of different family members. Traditions surrounding the birth of Jesus in Christianity, for example, are celebrated and passed on in quite different ways by Christian families across the globe.

1.5 Contemporary Religious Views of the Family

Nineteen ninety four was the United Nations International Year of the Family, and the institution of the family and the values associated with it were high on the political agenda in many countries. There was little agreement, however, about these issues. In America and Europe, some supported the nuclear family ideal outlined earlier, of two married parents with several children; others hailed what they called the 'post-modern family,' which was variously constituted of one or more adults of either gender and sexual orientation in a parental role with children. Realistically, in global terms, family types were still more open, including joint and multigenerational families, war-torn and other divided families, families with foster children, and childless families. Religious institutions were often asked for their views on the nature of the family and family values. What were their views on divorce, broken families, remarriage, and the postmodern family? What about birth control within the family, including fertility treatment? Did they have a view on gay and lesbian parenting? What did they think was the effect of women working on the family unit, especially on the socialization of children? A variety of views were offered, some traditional, some liberal, depending on the religious institution or movement and the denomination or group within it. The Catholic Church, for example, under the leadership of Pope John Paul II, continued to articulate a conservative position on marriage, the nature of the family, and sexual ethics. However, in practice, many Catholics contravened Church teachings. Some Reform and Progressive Jewish and liberal Christian groups made provisions to accept cohabitation, second marriages, and gay and lesbian unions, stressing the need for mutual support and care within partnerships and family groupings rather than traditional prescriptions about the inviolability of marriage and sexual ethics.

From the 1970s, the major collective voice for change regarding the family within religious institutions worldwide was that of women. Informed by the research and critiques of feminists, religious women from all over the world called for family-friendly teachings and practices, and respect for women's gifts and roles. In many countries, religious women challenged the law on family issues. In India, feminists fought for women's right to work outside the family, women's property and inheritance rights, and campaigned against dowry and female infanticide, all issues influenced by traditional Hindu and Muslim norms regarding women's roles and expectations.

Women were also able to challenge traditional views of the family by forging new religious roles for themselves or founding alternative religious movements. In Africa, in some traditional religions and new independent churches, women, through the acquisition of spiritual power, were able to circumvent the normal social requirements of wife and mother. Susan Starr Sered, in her research on religions dominated by women, gave examples of independent women who had broken with tradition and established new religious families and communities. She also showed that it was not only contemporary Western spiritual feminists who cried, 'No Father in heaven,' but also Thai, Burmese, Japanese, Brazilian, and Korean women who turned equally to female and male spirits for answers to problems, often seeing them as extensions to their own human families.

See also: Families and Households, Behavioral Demography of; Family and Gender; Family, Anthropology of; Family as Institution; Family Theory; Role of Changing Values; Kinship in Anthropology; Marriage; Motherhood: Social and Cultural Aspects; Religion and Economic Life; Religion and Gender; Religion and Youth; Religion: Morality and Social Control; Religion, Sociology of

Bibliography

- Baker A 1993 *The Jewish Woman in Contemporary Society: Transitions and Traditions*. Macmillan, London
- Burguière A, Klapisch-Zuber C, Segalen M, Zonabend F 1996 *A History of the Family [Histoire de la famille]*. Polity Press, Cambridge, UK, 2 Vols [trans. S H Tenison]
- D'Antonio W V, Aldous J (eds.) 1983 *Families and Religions: Conflict and Change in Modern Society*. Sage, Beverly Hills, CA
- Eriksen T H 1995 *Small Places, Large Issues: An Introduction to Social and Cultural Anthropology*. Pluto Press, London
- Esposito J L 1982 *Women in Muslim Family Law*. 1st edn. Syracuse University Press, Syracuse, NY
- van Gennep A 1960 [1908] *The Rites of Passage*. University of Chicago Press, Chicago
- Gerth H H, Wright Mills C 1970 [1948] *From Max Weber: Essays in Sociology*. Routledge and Kegan Paul, London
- Jackson R, Killingley D 1991 *Moral Issues in the Hindu Tradition*. Trentham Books, Stoke-on-Trent, UK
- Kayongo-Male D, Onyango P 1984 *The Sociology of the African Family*. Longman, London
- Pyper H S (ed.) 1996 *The Christian Family: A Concept in Crisis*. The Canterbury Press, Norwich, UK
- Rabuzzi K A 1982 Family. In: Eliade M (ed.) *Encyclopedia of Religion*. Macmillan, New York
- Sered S S 1994 *Priestess, Mother, Sacred Sister: Religions Dominated by Women*. Oxford University Press, New York
- Stacey J 1996 *In the Name of the Family: Rethinking Family Values in the Postmodern Age*. Beacon Press, Boston

- Turner B S 1983 *Religion and Social Theory*. Heinemann Educational Books, London
- Young S (ed.) 1993 *An Anthology of Sacred Texts By and About Women*. Pandora, London

K. Knott

Religion, History of

1. Historical Background to the Term 'History of Religion'

The Latin term *historia religionis* first appeared in the sixteenth century and was employed alongside related terms such as *historia ecclesiastica* to entitle chronicles of important events in church history, and the life histories of influential theologians and witnesses of the Christian faith. The term was also applied to accounts on the origin and development of Judaism and portrayals of foreign religions. In his *Historia Religionum* (1697), Gebhard Theodor Meier used the concept in the sense of a comparative study, aimed at portraying a history of the three monotheistic religions. Compared with the more descriptive treatises on religion, comparative *historiae religionis* drew on theological distinctions, such as that between God's chosen people and heathens, to the end of justifying the exclusive truth claims of Christianity. In his *Abris einer Geschichte der Religionspartheyen* (1755), the Protestant theologian and student of Christian Wolff, Sigmund Jacob Baumgarten (1706–1757), examined these different methods and aims of religious historiography, explicitly distinguishing between a *usus relativus* and a *usus polemicus*.

From the 1880s onwards we find the introduction of new compound terms of religious historiography which competed with *history of religion(s)*, *Religionsgeschichte* or *histoire des religions* and reflected a heightened interest in scientifically recording and interpreting religious phenomena. The appearance of the terms 'philosophy of religion', 'science of religion', 'religious studies,' and—later—'sociology of religion', 'psychology of religion,' and 'phenomenology of religion' demonstrated the increasing differentiation within the scholarship dedicated to scientifically interpreting religious culture. However, none of these new academic subjects succeeded in establishing itself as a dominant or authoritative discipline for investigating and interpreting religious systems of belief. Academic research into religion is characterized, rather, by a high plurality of divergent and competitive disciplines that have been institutionalized quite differently, mainly in the countries of Europe and in the USA. This rivalry between disciplines also applies to the historical inquiry into religions: besides professional historians, who sought to render their discipline the new leading human science of the nineteenth century,

other cultural scientists and theologians also participated in the historical study of religion and contributed significantly to its advances. As a result, most of the 'classical' works in nineteenth century religious historiography were not compiled by historians but by theologians (or former theologians), philosophers, ethnologists, philologists, sociologists, and professors in the nascent university disciplines, Sinology, Indology, and Egyptology.

In the course of these developments, the term 'history of religion' remained controversial. Friedrich Max Müller (1823–1900), who is regarded as the founder of the 'comparative science of religion' (*vergleichende Religionswissenschaft*), abandoned the term 'history of religion' in favor of the 'science of religion' in order to emphasize the scientific aims of an inquiry directed towards all historically given religious phenomena and traditions, and which strives to emancipate itself from the prejudiced approach of both theology and philosophy. Other researchers of religion, such as Cornelius Petrus Tiele (1830–1902), preferred the term 'comparative religion'. Nevertheless, common to all these different terms and their respective definitions of approach was the motivation to dispense with theologically orientated normative constructions in the study of religion. In Anglo-Saxon countries, this movement found its most important source of support from the so-called 'Lectures', an institution backed by the private endowments of wealthy patrons (in 1878 Friedrich Max Müller inaugurated the 'Hibbert lectures' and, in 1888, the first 'Gifford lectures' were given). The first professorship for the 'General History of Religion' was established in Geneva, in 1873; in 1877 the universities of Leiden and Amsterdam followed suit. In 1879 France's first chair in *Histoire des religions* was founded at the *Collège de France* and in 1886 the Graduate School of Sorbonne, *Ecole Pratique des Hautes Etudes*, instituted its *Section des sciences religieuses*. However, this appearance of several chairs in the 'history of religion' and 'religious studies' did not achieve the full emancipation repeatedly postulated by Christian theologians. In Germany, for instance, most religious-historical chairs were set up within evangelical-theological faculties; and even where this was not the case and provision made for institutional independence, as in The Netherlands, the majority of scholars appointed to such chairs were theologians.

2. History of Religious Historiography

The intellectual reflection on religion(s) can be traced back to scholars in Greek and Roman antiquity. Here we find the origination of several important topoi that continue to influence the study of religion (such as the critique of religion, theories on the origin of religion and its social function, the comparison between polytheism and monotheism). Notwithstanding this

degree of continuity, the ancient world's approach to the subject of religion cannot be termed 'historical'. Beyond the context of Graeco-Roman traditions, we can also find evidence of the examination of religious life in early Arabic, Persian, and South-East Asian scholarship. However, the influence of these traditions on European historiography of religion has been minimal. One central impetus for broadening Europe's religio-historical horizon was provided by the overseas conquests of the fifteenth and sixteenth centuries. What conquerors and travelers had to report on the foreign religious ideas and practices of the far-off lands they had visited led scholars to question the cultural self-evidence of Christianity, particularly after its confessional pluralization following the Reformation. In an attempt to intellectually overcome religious pluralism as well as the denominational conflicts within Christianity, E. Herbert of Cherbury (1581–1648) focused on a 'natural religion of reason' which he believed to be the true essence of all the different guises of positive religion. In doing so, this British deist introduced a distinction that would later be appropriated by Enlightenment theologians.

However, strictly speaking, it was the seventeenth century that gave birth to the historiography of religion proper. In the wake of the ecclesiastical schism and resultant religious wars, the Christian 'culture of unity' (Troeltsch) began to crumble and was soon replaced by an absolutist state which succeeded in politically neutralizing the vying religious parties by way of an absolute leader ruling over all religious parties. At the same time, the religious conflicts of the seventeenth century gave rise to the interest in a history of religion which was not conceived theologically, but had recourse to an impartial 'third' standpoint beyond confessional polemics. This was found in the 'historico-philological method' based on the principles formulated by Jean Mabillon (1632–1707) in his *De re diplomatica*, which was one of the first comprehensive treatises on methods of documentation and the study of hand-written manuscripts. The publication of this work in 1681 can be seen as the birth of textual criticism that played such a central role in the development of religious historiography. At about the same time, Richard Simon (1638–1712) applied literary-critical methods to the study of the Pentateuch in his *Histoire critique de l'Ancien Testament* (1678) and, in doing so, was one of the first to expose the religious history of the Judaeo-Christian tradition to the process of historicization. This rigorous distinction between the theological and historical historiography of religion also formed the guiding precept of Johann von Gro(e)ning's *Historie der heutigen Religionen* which appeared in Hamburg in 1702. In dismissing every form of *theologia polemica*, this author declared his commitment to a historical approach orientated towards the facts.

That the historian had not irrevocably displaced Christian apologetics, however, is evidenced by

Jacques-Bénigne Bossuet's (1627–1704) work *Discours sur l'Histoire universelle* (1681) which appeared in the same year as Mabillon's important treatise and adhered to an interpretation of world history stressing God's saving grace. In opposition to Bossuet, Montesquieu (1689–1755) banished the guiding hand of divine providence from history and declared mankind to be history's focal point. The human being, expounded Montesquieu in *Esprit des lois* (1748), is determined by lots of factors—climate, laws, customs, and traditions as well as religion—the combination of which gives rise to the particular intellectual characteristic of a nation and its underlying 'spirit'. The work of the Göttingen exegete Johann David Michaelis (1717–1791) illustrates how Montesquieu's point of departure could be used to widen the religio-historical perspective. According to his own statements, Michaelis sought to view the Old and New Testament 'through the eyes of Montesquieu'. To this end he reconstructed in his six-volume *magnum opus*, *Mosaïches Recht* (1770–1775), the structure and contents of the Mosaic codex from the climatic, economical, political, and socio-cultural preconditions of its natural environment, and thereby offered an interpretation of this codex in the spirit of Montesquieu and in reply to the demands of his day.

Alongside this *historico-philological* examination of religion, we also find the appearance of the *psychological* approach to religion, as represented by David Hume (1711–1776) in his treatise *The Natural History of Religion*, published in 1757. Hume was not primarily concerned with explaining religion historically, but rather with accounting for religious ideas and practices as phenomena of the human psyche under the dictates of irrational emotions such as fear and hope. Because Hume saw reason as having gradually overcome these irrational forces in the course of history, however, this Scottish philosopher had also furnished the historiography of religion with an evolutionist perspective that would then feature as a key concept in the second half of the nineteenth century.

With the conception of world history as a developmental process from lower to higher, that is to say, from uncivilized to rational forms of life, religion came to represent not an exception to, but a manifestation of these historical dynamics. The evaluation of religion's status within this process of advance was divided: either one adopted a religio-critical stance, and assumed that reason would eventually replace religion (Voltaire, 1694–1778), or one took up a more positive, religio-historical position, and integrated religion into historical progress. One leading exponent of the latter position was Charles de Brosses (1707–1777). In his study *Du culte des dieux fétiches* (published in 1760) he employed a comparison between ancient and contemporary African cult ritual to support his claim that fetishism—understood as the worship of arbitrarily chosen natural objects or living creatures—is to be identified as the original form of all

religions. However, in the course of mankind's historical advance this has been successively superseded by the worship of celestial bodies, by polytheism, and then finally by monotheistic religions, he contended.

In his *Grundriß der Geschichte aller Religionen* (1787), Christoph Meiners (1747–1810) also construed the history of religion as part of a general historical process of cultural advance. As a result of this pronounced tendency in the eighteenth century to assimilate the portrayal of religion into an all-encompassing cultural historiography, the traditional anchor of religious history in the theological field of church history began to corrode, causing a telling shift of emphasis. Johann Christoph Gatterer (1727–1799), for instance, dealt with church history within the framework of a universal history of religion. In a similar vein, Christian Wilhelm Flügge approached church history as an aspect of religious history that he conceived as a branch of the general cultural history of the human spirit.

Resistance against this linear and optimistic idea of historical progress had already begun to emerge in the eighteenth century. For Jean-Jacques Rousseau (1712–1778), scientific and technical progress amounted to a moral step backwards behind the original and good 'natural state.' Even religion, he argued, has not escaped this cultural decline, but degenerated from an honest 'affair of the heart' into an institutionally depraved ecclesiasticism no longer capable of integrating the community. Nonetheless, because Rousseau considered religious feeling to be the strongest social bond, in *Contrat Social* (1762) he spoke up for a religious 'occupation' of the values which hold together the political community, viz. the establishment of a 'civil religion.'

Similar ideas were to be voiced by Alexis de Tocqueville (1805–1859) almost a century later who, in his famous work *De la démocratie en Amérique* (1834/40), prognosticated that the politico-moral consensus within democratic societies would adopt the socially integrative role which religion had played in the social system of the Middle Ages. In eighteenth-century Germany, Johann Gottfried Herder (1744–1803) likewise objected to the belief in progress of Enlightenment thinkers, supporting the Romantic conception of history as a process of organic growth. In doing so, Herder influenced the organo-historical ideas of classical German historicism. This school of thought also rejected the Enlightenment's linear concept of history and belief in universal-historical progress, viewing history as the unfolding of individual cultural units.

The turn of the century into the 1900s was also a transitional period for European historical consciousness which had begun to move away from universal generality and attach greater importance to that which is particular and individual. This change of perspective not only enhanced the status of the historical religions, which had been neglected by the

concepts of historical progress of the eighteenth century, but also paved the way for the empirical research into the history of religion, especially the intensive analysis of ancient religious traditions.

Among others in Germany, Otto Jahn (1813–1869) and Karl Otfried Müller (1797–1840) devoted their scholarly activities to investigating ancient mythology. In his influential study *La cité antique* (1864), the French historian N. D. Fustel de Coulanges (1830–1889) examined the stabilizing function of religion for the Greek polis and Roman system of rule. Furthermore, this new interest in empirical research into religion led to what Edgar Quinet described in his work *Génie des religions* (1842) as an 'Oriental Renaissance': the first translations of India's holy scriptures had already begun to appear at the end of the eighteenth century.

In 1808, Friedrich Schlegel published his writings on the language and wisdom of India (*Über die Sprache und Weisheit der Inder*) whose conception of religious history diametrically opposed that of the Enlightenment. In India's ancient wisdom, Schlegel saw the realization of a unity between nature and spirit (Geist) which 'enlightened Europe' had forfeited and India itself submerged, and which he strove to reconstruct from the oldest of India's religio-historical sources. Drawing on Schlegel, Georg Friedrich Creuzer (1771–1858) and Joseph Görres (1776–1848) called into question the origins of antiquity's mythical tradition, claiming that Greek myths originally derive from the wisdom of Indian priests. With this Romantic enthusiasm for oriental languages and literature, the availability of translations of the latter and, not least, the advances in the field of 'historical and comparative linguistics since the turn of the nineteenth century,' the historiography of religion in the nineteenth century was afforded a wide range of new stimuli.

The endeavors of Franz Bopp (1791–1867) to distill a common Indo-European proto-language from a grammatical comparison between Sanskrit, Greek, Latin, Persian, and German, for instance, gave the comparative study of religions new impetus, temporarily pushing universal interpretations of history into the background. Aside from the work of Eugène Burnouf (1801–1852), a prominent figure in the historiography of Iranian religion, it was above all the services of his, and Bopp's student, Friedrich Max Müller, which rendered the philological approach beneficial to religious historiography.

Analogous to the Romantics' interest in the origins of religious history, Müller linguistically classified the Indo-European languages, particularly with reference to their respective names for God, to the end of reconstructing a primal religion common to all these peoples, which he saw as having once flourished in the advanced civilization of ancient India. Hermann Usener's (1834–1905) investigations into ancient religious history were also based on a combination of philologico-comparative studies and theoretical-

explanatory ambitions. With his programmatic work on an interdisciplinary ancient religious history—which his students Eduard Norden (1868–1941) and Ulrich von Wilamowitz-Moellendorf (1848–1931) advanced—Usener not only influenced the study of ancient religion, but also classical philology, art history and oriental studies, not least cultural theory until well into the twentieth century.

Besides the philological exploration of religious traditions, which chiefly focused on antiquity and the Orient, the second half of the nineteenth century also witnessed the proliferation of ‘ethnological’ research into contemporary, non-European religions which had not been taken into consideration by the romantically inspired religious historiography of such scholars as F. M. Müller. That these religions had become the focal point of attention within nineteenth-century religio-historiographical discourse was due to a revival of the evolutionist theory, aimed at assimilating the individual, historical, and divergent forms of religion into a universally applicable model of cultural advance.

In his major work *Primitive Culture* (1871), Edward Burnett Tylor (1832–1917) held the view that the religions of contemporary tribal societies represented the beginnings of religious history and therefore attached great importance to the investigation of such communities. However, Tylor’s claim that animism (the attribution of a living soul to plants, inanimate objects, and natural phenomena) constituted the primal form of religion was soon superseded by other theories on the origins of religion, the most influential of which was William Robertson Smith’s (1846–1894) totemism theory tracing religion back to the worship of ancestors symbolized by totemic emblems of plants and animals. Emile Durkheim’s (1858–1917) writings on the elementary forms of religious life (*Les formes élémentaires de la vie religieuse*, 1912) can be seen as the most prominent adaptation of Smith’s propositions. Religio-historical evolutionism proved to be an extremely resilient paradigm, providing the working hypothesis for James George Frazer’s (1854–1941) reconstruction of religious history on the basis of contemporary ethnological material in his famous twelve-volume opus *The Golden Bough* (1890–1915).

In the latter half of the nineteenth century, ethnological and philological investigations into the history of religion found enormous support within academia and aroused ever greater interest among the general public. This trend was evidenced in the UK by the positive response to lectures in this field of study (cf. Sect. 1) and the huge popularity of Ernest Renan’s (1823–1892) *La vie de Jésus* (1863) in France and David Friedrich Strauss’ (1808–1974) works *Das Leben Jesu* (1835/36) and *Der alte und der neue Glaube* (1872), as well as Hermann Oldenberg’s book *Buddha* (1890), in Germany. *Historical* scientists, by contrast, showed very little interest in religio-historical phenomena; the majority of professional historians adhered to a narrow politico-national approach to their subject in

which religio-historical questions remained peripheral.

The few exceptions, held in high esteem today, were then decisive outsiders, such as Jules Michelet (1798–1874), who in his 17 volumes on the history of France (*Histoire de France*, 1833–1867) drew attention to the religious folklore of the Middle Ages and, in doing so, became a predecessor of the *histoire des mentalités* (cf. Sect. 3), or Jacob Burckhardt (1818–1897) who in *Weltgeschichtliche Betrachtungen* (1868–73) and *Die Kultur der Renaissance in Italien* reflected on religion as a third historical ‘power’ (*Potenz*) alongside culture and politics.

The cultural-historical revival at the turn of the twentieth century effectively placed religion at the forefront of scholarly discourse throughout Europe and the USA. The rapid erosion of time-honored cultural traditions, the growing insight into the social costs of capitalist modernization, and the new disintegrative tendencies in many Western societies provoked heated debates on the relationship between the religious history of European countries (as well as that of the USA) and their individual formation processes as modern societies. In defiance of traditional disciplinary borders, the historical research into religion became a central medium for the self-reflection of Western societies. From a systematic perspective, three closely connected issues came to play a decisive role in this intensive exploration of religious history around 1900. These issues included: the participation of religious traditions in the genesis of modern occidental rationalism; the socially integrative potential of religion in functionally differentiated and politically fragmented societies; and the cultural relevance of differences between Christianity’s denominations on the one hand, and non-Christian religions, especially those of the East, on the other.

First and foremost, this discourse applied to Protestant theology which had largely viewed itself as a historico-cultural science of Christianity since the mid eighteenth century and, since the 1880s, became increasingly aware of the interrelations between the Christian religion and its non-Christian environment. Above all, it was the proponents of the so-called ‘History-of-Religions School’ (*Religionsgeschichtliche Schule*) within German Protestant theology who, in classical exegetic and religio-historical treatises, analyzed the interconnections between Jewish antiquity and the religious worlds of the ancient Orient, and the syncretistic character of Christianity. Although religio-historically minded scholars of the Old and New Testament, such as Hermann Gunkel (1862–1932) and Wilhelm Bousset (1865–1920), as well as the systematic theologian and cultural philosopher Ernst Troeltsch (1865–1923)—named the ‘systematist of the History-of-Religions School’—viewed ‘religion’ as a relatively autonomous cultural power *sui generis*, they nevertheless gave greater priority to investigating the interactions between religious and other spheres of

culture. Ernst Troeltsch's 'The Social Teaching of the Christian Churches' (*Die Soziallehren der christlichen Kirchen und Gruppen*, 1912) is regarded as the most important work of this religio-historical group. In this publication, Troeltsch embarked on a cultural-historical reconstruction of the formative influence of Christianity's denominational churches and sects on social ideas and the effects of social structures on the formation of religious ideas and communities. Appropriating the typological distinction between 'church' and 'sects', developed by his friend Max Weber (1864–1920), Troeltsch differentiated between 'ecclesiasticism', 'sectarianism,' and 'mysticism' as specific forms of religious socialization. With Weber, Troeltsch shared the ambition to demonstrate the significance of religious ideas and practices for the emergence of the modern world. In *Die Bedeutung des Protestantismus für die Entstehung der modernen Welt* (1906), he focused on Protestantism and Calvinism, in particular.

While Troeltsch transformed church history into the cultural and social history of Christianity, Weber expanded his pioneering study 'The Protestant Ethic and the Spirit of Capitalism' (*Die protestantische Ethik und der Geist des Kapitalismus*, 1904/05) into an extremely elaborate enterprise of categorically comparing the economic ethics of the world religions (*Die Wirtschaftsethik der Weltreligionen*, 1915). In 'The Protestant Ethic', Weber had inquired into the effects of a religiously shaped *habitus* in areas of life and culture far removed from those commonly associated with religion in the narrower sense of the term—above all, in the economic sphere. In his comparative analyses of the economic ethos of the world religions, however, he primarily concentrated on occidental rationalism and sought to grasp its particularity and formative power on cultural life. Today, scholars are more skeptical of the details of Weber's ideas. Nonetheless, it cannot be denied that the worldwide reception of his theories and derivations in the study of religion has scarcely been matched by any other oeuvre.

3. Overview of Recent Research

Since the 1960s, the renaissance of religious movements in many societies, the upward trend of 'fundamentalism' and Protestant sectarianism, the spread of charismatic Christianity, and the manifold processes of a rapidly changing religious landscape (conversions, new syncretism) have attracted the attention of historians and led to a critical revision of the 'secularization theory', which equated social modernization with a general decline in religion. Religion has been rediscovered as a relatively autonomous system of communication and symbolism which shapes the *habitus* of individuals or identity constructions of social groups, and exerts a positive or

negative influence on the development of societies. As a result, the old debates on the definition of religion and master narratives on modernization de-Christianization or secularization are being steadily replaced by micro-studies into the diverse religious milieus of the modern age. While religion has become a central topic in the historical discourse of all countries generally, several national-cultural particularities deserve mention.

In France, the work of Lucien Febvre (1878–1956) and Marc Bloch (1886–1944), together with the journal *Annales, Histoire, Sciences Sociales* (originally entitled *Annales d'histoire économique et sociale*) they founded in 1929, supplied the historiography of religion with an innovative approach which broke with the politico-historiographical tradition of the nineteenth century. Their objective was to uncover a temporal stratum of *longue durée* (Fernand Braudel, 1902–1985), that is to say, extensive periods of time which formatively shape not only political and socio-economical history, but also the concrete experiential world of the general population. Concepts stemming from *Annales*-historiography, such as 'structure' and '*longue durée*,' were then applied heuristically to the broader discussion of collective 'mentalities,' from which the historiography of religion profited in particular.

Although the term *histoire des mentalités* soon became a key concept of *Annales*-historiography, no agreement had been reached on the question as to what 'mentality' exactly means. These definitional problems notwithstanding, *histoire des mentalités* can be generally described as an historical discourse directed towards the collective consciousness of an epoch, i.e. its cognitive, ethical, and affective dispositions, which form the unquestioned cultural basis for the daily life of individuals and social groups, thus constituting their identity. In its application to the history of religion, this approach initiated a change in perspective away from the narrow argumentation of ecclesiastical historiography: religious history came to be seen as a history of religious consciousness whose carriers remain for the most part anonymous. Following this, historians of religion in France began to attach less importance to major events, personalities, and institutions, and focus on 'structural-historical' and, in particular, 'anthropological' lines of questioning.

Here their investigations have been concerned with historically varying attitudes to birth and childhood, mortality and death, corporeality and sexuality, nature and the environment, God and the Church, heaven and hell (cf. Ph. Ariès, M. Vovelle, J. Le Goff, J. Delumeau). To date, however, exponents of *histoire des mentalités* have scarcely researched the religious culture of more recent history, concentrating rather on the Middle Ages and early Modern Age, and—in co-operation with experts in ancient history such as J.-P. Vernant, M. Detienne and P. Vidal-Naquet since the

1970s—on the religious history of antiquity. Further drawbacks of applying this historical approach to religious historiography arise from its emphasis on the alleged durability of mental dispositions, its disregard of the processes of change within religious mentalities, and its neglect of the conflicts caused by differing religious backgrounds (resulting from the acculturation of foreign ideas and practices or due to inner-social differences).

In Germany, the 'socio-historical turn' of the 1960s and 1970s as well as the institutional separation of church history (pursued within theological faculties) and general history encouraged scholars to view religion and confession as nothing other than marginal factors in the history of the modern age. Since the 1970s, however, several attempts have been made to integrate religio-historical perspectives into social historiography (cf. W. Schieder, R. van Dülmen, R. von Thadden) and to open up historical research in church history and theology to the methodological debates of cultural and social scientists (cf. M. Greschat, K. Nowak, F. W. Graf). Numerous younger historians are now conducting research into a range of religious milieus, such as ecclesiastical organizations and festivities, theological controversies and quarrels over religious symbolism as well as the tense co-existence of Catholics, Protestants, and Jews. The collapse of East Germany, moreover, has generated a new sensibility for the political role of churches and their partly antagonistic, partly affirmative relationship to the 'political religions' of the twentieth century. And last, but by no means least, the holocaust initiated worldwide intensive research into the cultural climate of German Judaism and the conflict-ridden history that finally led to the destruction of the 'German-Jewish synthesis.'

In the UK, discussions continue to be determined by the weight of social anthropology (cf. B. Malinowski, A. R. Radcliffe-Brown). Here, from the outset—unlike in Germany—'society' and 'culture' were seen as one, and consideration given to the class-specific religion of the workforce and the processes of religious change in this country's urban centers (cf. H. McLeod). Particularly pioneering were the studies on the connection between religious radicalism and revolution (cf. Ch. Hill) and between Methodism and the emerging labor movement (cf. E. P. Thomson). Research into the processes of religio-theological differentiation in the Anglican churches has also drawn on the interpretative approach of *histoire des mentalités*.

In the USA, historians of religion since the 1980s have been strongly influenced by the 'interpretative' ethnology (stressing the concept of 'understanding' rather than that of 'explanation') and cultural anthropology of C. Geertz. Beyond the classical self-reflective discourse on the religion of native Americans, the faith of the Pilgrim Fathers, pluralism within Protestant denominationalism, the emergence of a specifically American Catholic self-understanding in the late

nineteenth century and a 'civil religion' (N. Bellah) capable of integrating religious pluralism, in the most recent American historiography of religion themes such as gender, race, language, and the religious forms of minority groups have played an important role. In addition, the ethnological debate on 'Writing Culture' (viz. the constructive character inherent in self-portrayal and the description of that which is foreign; cf. F. Clifford/G. E. Marcus), the comparative analysis of fundamentalist movements in the world religions (M. E. Marty), the interpretation of Latin America's religious developments (D. Martin) and the interrelation between law and religion (W. Fallers Sullivan) have featured as central concerns of recent research. Current North American religious historiography has also profited greatly from the application of economico-theoretical approaches (R. Stark, Bainbridge W S (1985), Stark R, Finke R (2000)).

4. Outlook

Religion has gained in importance worldwide and this situation has accordingly afforded the research into religion greater significance. At the forefront of this discourse we find the questions concerning the interrelatedness of religion and ethos, the political relevance of religion, the construction of religious identities of minority and the conflicts between members of different religions and confessions. The debates on migration, religious conversion and new syncretism have shown, however, that historians of religion find it difficult to precisely grasp the long-term changes of religious mentalities. The correlation of conflict-ridden modernization, on the one hand, and the renaissance of strictly binding religion on the other, promises to become an issue of particular concern in the future: for the rapidly advancing globalization of capitalism, together with its constraints on heterogeneity, will only fuel those religious movements intent on communicating a more fulfilled form of human life which transcends an existence solely defined by the utilitarianism of economic rationality.

See also: African Studies: Religion; American Studies: Religion; Christianity Origins: Primitive and 'Western' History; East Asia, Religions of; Historiography and Historical Thought: Jewish Tradition; Islam: Asia; Islam: Middle East; Islam: Sub-Saharan Africa; Judaism; Latin American Studies: Religion; Near Middle East/North African Studies: Religion; Protestantism and Gender; Religion: Evolution and Development; Southeast Asian Studies: Religion; Western European Studies: Religion

Bibliography

Despland M 1979 *La Religion en Occident. Evolution des Idées et du Vécu*. Fides, Paris

- Dubuisson D 1998 *L'Occident et la Religion. Mythes, Science et Idéologie*. Editions Complexe Brussels, Belgium
- Feil E 1986 1997 *Religio*, 2 Vols. Vandenhoeck & Ruprecht, Göttingen, Germany
- Hjelde S 1994 *Die Religionswissenschaft und das Christentum. Eine historische Untersuchung über das Verhältnis von Religionswissenschaft und Theologie*. Brill, Leiden, Germany
- Idinopulos T A, Wilson B C (eds.) 1998 *What is Religion? Origins, Definitions, and Explanations*. Brill, Leiden, Germany
- Jastrow M 1981 *The Study of Religion*. Reprint. Scholars Press, Atlanta, GA
- Jordan L H 1986 *Comparative Religion. Its Genesis and Growth*. Reprint. Scholars Press, Atlanta, GA
- Kemper D 1992 Religionsgeschichte I. In: Ritter J, Gründer K (eds.) *Historisches Wörterbuch der Philosophie*. Schwabe, Basel
- Kippenberg H G 1997 *Die Entdeckung der Religionsgeschichte. Religionswissenschaft und Moderne*. Beck, Munich
- Lehmann E 1925 Zur Geschichte der Religionsgeschichte. In: Bertholet A, Lehmann E (eds.) *Lehrbuch der Religionsgeschichte*, 4th edn. Mohr (Siebeck), Tübingen [1887–1889], Germany
- McCutcheon R T 1997 *Manufacturing Religion. The Discourse on Sui Generis Religion and the Politics of Nostalgia*. Oxford University Press, Oxford, New York
- Molendijk A L, Peter P (eds.) 1998 *Religion in the Making: The Emergence of the Sciences of Religion*. Brill, Leiden, Germany
- Pailin D A 1984 *Attitudes to Other Religions. Comparative Religion in 17th and 18th Century Britain*. Manchester University Press, Manchester, UK
- Pinard de la Boullaye H 1922–1925 *L'Étude Comparée des Religions. Essai Critique*. G. Beauchesne, Paris
- Platvoet J G, Molendijk A L (eds.) 1999 *The Pragmatics of Defining Religion. Contexts, Concepts and Contests*. Brill, Leiden, Germany
- Puech H C (ed.) 1970 1972 1976 *Histoire des Religions*, 3 Vols. Encyclopédie de la Pléiade, Gallimard, Paris
- Rudolph K 1992 *Geschichte und Probleme der Religionswissenschaft*. Brill, Leiden, Germany
- Saler B 1993 *Conceptualizing Religion. Immanent Anthropologists, Transcendent Natives, and Unbounded Categories*. Brill, Leiden, Germany
- Schlesier R 1994 *Kulte, Mythen und Gelehrte. Anthropologie der Antike seit 1800*. Fischer, Frankfurt, Germany
- Sharpe E J 1997 *Comparative Religion. A History*, 5th edn. Open Court, Chicago and La Salle, IL [1975]
- Smart N, Clayton J, Sherry P, Katz S T 1985 *Nineteenth Century Religious Thought in the West*, 3 Vols. Cambridge University Press, Cambridge, London, New York
- Stark R, Bainbridge W S 1985 *The Future of Religion: Secularization, Revival, and Cult Formation*. University of California Press, Berkeley, CA
- Stark R, Finke R 2000 *Acts of Faith. Explaining the Human Side of Religion*. University of California Press, Berkeley, CA

F. W. Graf and A. Reuter

Religion: Mobilization and Power

In what ways does religion mobilize believers? How are they empowered by their beliefs, and to what end? Religion, as a phenomenon, has always contributed to both social conflict and social cohesion. It has the

power to mobilize individuals, communities, and nations. Within religious institutions, power is as inevitable and ubiquitous as anywhere else. It can be the power of good, or morally ambiguous potency. Equally, it can come vested or masked as unction or inspiration. To speak of religion is always to speak of some kind of notion of power, whether it be attributed to divine, human, elemental, or cosmic sources. Relations of power are amongst the most significant aspects of any religion, especially when considering its context and cultural intercourse.

The study of power and mobilization in relation to religion must be rooted in theory, history, and praxis. For example, any consideration of church–state relations inevitably is concerned with control, legitimacy, and authority. It is impossible to conceive of a discussion concerning the origins of the Church of England that does not involve power–relations: those between the Pope and King Henry VIII; between the Reformers and Roman Catholicism; between competing powers at court; between competing ideological and political convictions on the nature of individual salvation and the centrality of the church; vernacular liturgies and scriptures, and their overall hermeneutical control and its centralization; between individual states and their subjects, and their right to any kind of religious choice, including that of dissent. The Church of England—‘by law established’—is a form and exercise of power in nearly every era of its history and at every level of society, from government to parish, from individual to state. To study the church is to study power, in periodic, dispositional, episodic, and latent forms.

Globally, religion operates as a form of power and mobilization almost everywhere. In the USA, although no religion is established (the Constitution separates the churches and State), the outcome of increasing pluralism and secularisation has seen, amongst other phenomena, the mobilization of the New Right, and the movement of Roman Catholicism from a private to ‘public’ denomination, both on a select range of moral issues. As a number of scholars (Haynes, Casanova, etc.), have shown, this is a form of counterpower: the secularization theories (so prevalent in the 1960s and 1970s) have had to capitulate to the rise of a new religious consciousness in industrialized nations. They see modernity not as the harbinger of secularization, but rather of privatization, a by-product of late capitalism. Religion has fought against being squeezed into the private and subjective realm by recovering and reconstituting a public role. It is now a potent force for mobilization on a range of public issues in the political and social spheres. The example of the USA shows that the forces of pluralism, industrialization, and secularization can be combated by religious power at many levels.

In the last quarter of the twentieth century, the competition and association of powers between religion and society (e.g., Bosnia), church, and state

(e.g., Ulster) has operated differently but no less dynamically in Europe. In Eastern Europe, religion (mostly Christianity and Islam) played an important role in the diffusion of the communist monopoly. The vociferous activity of some Muslims in Western Europe has questioned the limits of liberalism in politics and education; in Northern Ireland, religious polarization has attempted to hold out against political reconciliation; the conflicts in Bosnia suggest that the purity of religious and ethnic enclaves is prized above any peaceable settlement. Added to this, there is a discernible rise in new religions, sectarian movements, and a patchy but popular inculcation of 'ordinary' spiritual ideology, indicating that religion remains a perennial public force.

In Africa, the Middle East, South America, Central Asia, the Indian subcontinent and South East Asia, religion has found itself harnessed to many nationalist causes. The symbiosis has often been so intricate as to make the political power play indistinguishable from that of the religious: Lebanon, Israel, Sikh nationalism in India, and Buddhism in Burma are perhaps the most notable examples. The origins of Pakistan lie in religious conflict and differentiation, as does civil war in Rwanda. Equally, it is also important to note that the resurgence of religious identity threatens existing religious powers, especially those allied to the State. Islamic movements in Turkey, Algeria, and Egypt demonstrate that religious fundamentalism is a powerful and popular force for social mobilization. In South Africa, religion has played a vital role in securing peace and democracy following the end of Apartheid; the 'Truth and Reconciliation Commission' was birthed in Christian theology, and chaired by an Archbishop (Desmond Tutu).

1. Religion and Power

Sociologically, distinctions between authority and force, and between violence and resistance, will and desire often cluster around the study of power. In political science, power usually is configured in terms of the ability of its holders to carry out their will, exact compliance, exert force, and compel obedience. To touch on authority, as part of power, is to study of relations, rules, and regulations. To focus on force, as another arm of power, is to consider compulsion, threats, and violence. In any consideration of politics, religion, or society, the subject of power (and its legitimacy) is never far from the surface.

The concept of power in social sciences interfaced with any confessional theology has no unity of discourse. Social scientists assess the 'power of God' through its social and observable outcomes: constructions of reality, rhetoric, sociality, and activity. On the other hand, believers claim to know and experience this same power personally and communally, in ways that appear to be either inaccessible

to or immeasurable for social scientists. Part of the problem is this: 'power' is what we call 'something': it is primarily a noun, not a verb (Morris 1987).

Although the verb does exist in the sense that we may speak of empowering, the fact that it is mainly a noun has implications for the way in which 'power' is described and studied. All too commonly, power is run together with verbs that are deemed to be its associates. For example, from a sociological perspective, a theoretician like Dahl (1957) describes power by contrasting it with near-synonyms like 'influence,' so that power becomes like influence. This becomes more obvious in May's *Power and Innocence: A Search for the Sources of Violence*, in which the noun again assumes the characters of the nearest relevant verbs. For May, power is exploitative, manipulative, competitive, nutrient, or integrative (1976). Each of these types of power represents a different level of threat, force, or coercion, ranging from the mutually empowering (integrative) to the potentially violent (exploitative). Thus, 'power'—at least so far as religious believers are concerned—loses its reality in some sense here, precisely because the verbs are allowed to project on to a passive noun. Perhaps this is why deconstructors of power such as Foucault (1984) have looked at power and seen 'nothing'—only reciprocal relations. On the other hand, 'realists' such as Weber (1947) have seen the concept of power as a 'highly comprehensive' schema within sociology for the analysis of circumstances and situations.

Van der Leeuw's (1938) phenomenology of power also recognises that one cannot prove there is such a 'thing.' He suggests that sensitivity to power has been largely forgotten in 'modern' society, but in primitive society, power developed into religious monism, which reached its climax in monotheism. In contemporary society, power is not encountered in nature or personality, but in the dispersed forms of social organization. However, the essence of religion is power, driven towards monism, which has a tendency to conflate the psychological with the cosmological, the sociological with the theological: religion is a description of concepts of power and their use.

More recently, and from an anthropological viewpoint, Lewis (1996) has also argued that although religious power assumes many different forms (which may appear to be unrelated or mutually exclusive), it is an appropriate 'theme' through which to approach religion holistically. For Lewis, religious power is best understood as a negative force that opposes malign spiritual powers. Thus, the shaman assumes the charismatic role par excellence—being able to exorcise the evil and prepare the way for blessing. In effect, the shaman stands between negative and positive power as a broker, a fixed point of passage and determinacy that both creates and obviates power. Like Van der Leeuw, Lewis cannot 'see' power, except in the way that it is related to, and in the social and ritual structures that are based on its assumption.

Thus, and sociologically, the conceptions of power can be divided into two very broad categories. On the one hand, there are those which are asymmetrical, stressing conflict, will, resistance, and the like: social relations are assumed to be competitive, conflictual, and dialectical. A *Jihad* ('fighting against unbelievers') may be understood in this way (Peters 1996). On the other hand, there are conceptions that imply that all may make some gain through power. Here, power is a collective capacity or achievement, born not out of conflict, but communal welfare (Lukes 1978). As McGuire (1992), many forms of Christian nationalism offer 'social cohesion.'

Although it is helpful to be mindful of these two broad categories, it should be noted that they are somewhat inadequate in the analysis of religion. To an extent, all religious movements have elements of both categories within them. This must be because the religious-social system itself is one of exchange alongside being one of power, which therefore allows for conflictual and communal action together. Beyond this, however, there is no account of subliminal power, which is so critical for an understanding of religious behavior. Power is not just attached to structures and positional personality, but also to nonstructural interaction and roles. Parsons therefore eschews 'canonically correct' definitions of power: it actually covers more than it can define. It must take account of magnitude, distribution, scope, domain, and outcomes. Moreover, it must consider resources, skills, and motivations. In considering power and mobilization in religion, the investigator is always close to the analyses of causality, which in turn refers to power; its measurement, exercise, distribution, and impact.

Morriss (1987) suggests that careful attention to the etymology of power reveals its own diversity, prior to any linkage to verbs. Power is 'ability to do or effect something or anything ... energy, force, effect ... an active property, capacity of producing ... possession of control or command over others ... dominion, rule, legal ability, commission, faculty ... political or national strength.' That there is an overlap with verbs about power is clear: but that there is not complete synonymy is also clear. Morriss argues for a dispositional understanding of power. That is to say, power is seen as a relatively enduring capacity within any given object. Dispositional views of power are held in contrast (but not necessarily in opposition) to episodic ideas, that see occurrences or events as specific exercises of power. Morriss (1987) does point out that this power is not something that can be observed. Power is known only by its reification, which is ultimately an indirect but material way of inferring what power might be at work. This conclusion underlines his plea for power not to be studied in isolation: the subjects and objects of power require a methodological tolerance in the spheres of theory and evidence.

Clegg (1989) syncretizes the work of the previous

theorists discussed, and effectively constructs an 'open' theory of power based on the concept of circuits, nodal points, and agency: it has political, organizational, and theological resonance. Clegg begins his work with an overview of earlier theoreticians such as Hobbes and Machiavelli. This is followed by discussion and critiques of Dahl, Russell, Bachrach, Baratz, Weber, and Wrong, revolving around the 'two faces' of power, namely the relationship between 'intention' and 'structure,' and which enjoys supremacy in society. According to Clegg, these debates, naturally enough, were eclipsed by more substantial models. Lukes proposed a three-dimensional model of power by identifying the centrality of 'interests'; Habermas focused on ideal speech situations, ideology, and hegemony; Giddens on the relationship between agency and structure in terms of power. Each of these speaks to the religion-mobilization axis.

But Clegg moves beyond each of these explications, working in Parsons, Foucault, and Mann (among others) into an overall integration that respects the dispositional and facilitative approaches to power, besides stressing the vital role of strategic agency. As he states, 'a theory of power must examine how the field of force in which power is arranged has been fixed, coupled and constituted in such a way that, intentionally or not, certain "nodal points" of practice are privileged in this unstable and shifting terrain' (1989). This is a critical observation in the field of mobilization, religion, and power, for change is neither predictable or necessary, yet its occurrence is beyond question.

Any discussion about the true nature of power and where it lies will be at the heart of all religious belief. Liberation theology, feminist theology, and other new 'disciplines' within religion could be read by some as being mainly in terms of reconstituting the balance of power in ecclesiastical and social structures. The Nation of Islam has done the same within a North American context, although this is still developing. These theologies are effectively a form of revolution and resistance. The effect of the stress on power has led to more self-conscious questioning by some writers, who are beginning to question the nature of the power, that would generally be uncritically appealed to.

With respect to Clegg and Morriss, one of the difficulties of studying 'pure' power is that it is not one 'thing' to be observed. For example, a nationalist uprising in India may have a religious dimension functioning at many different levels: provision of identity, ideology, legitimization, coupled with notions of sacrifice, taboo, and salvation. Equally, an examination on the origins of Methodism would need to consider the power of mass meetings, the appeal of hymns, testimonies, and religious experience, not to mention the mobilization of displaced rural and agrarian populations in their search for meaning and community. Discerning power in such scenarios is a complex task.

Given the myriad of studies on power and attendant methods, the most practical way forward, especially given the multifarious nature of 'power' that occurs within religious movements, is to return to the agents or 'nodal points' of power identified by Clegg, and focus on their relationality. The advantage of Clegg's conceptualizing of power in terms of circuits, flow, and agency is that one can often identify the delimited 'field' of power (such as an organization), the directionality of power within the structure ('flow' or hegemony), and the specific nodal points of power such as a guru, infallible text or law, or episode (e.g., 'flashpoint'), as well as the latent forms of power embedded in culture, ethnic identity, and political situations.

This theory of circuits of power does not attempt to marry the 'nonrealist' tendencies of Foucault with the 'realist' or 'identification' trends of someone like Weber or Wright-Mills (although a focus on relations steers a middle course between realist and nonrealist accounts of power). Instead, it describes power in terms of circuits, which are organized through agencies. However, it does allow that the agencies cannot be seen as 'effortlessly rational or powerful,' since their 'carrying capacity is itself opened up for scrutiny in power terms.' This view of power is more contingent than monolithic. It has space for dispositional, facilitative, and episodic power within the same or overlapping frameworks. This has implications for the study of mobilization.

Dispositional power, as its description suggests, is the tendency or habit of an individual or group: in many religious movements, this is reflected in the 'grammar of power as a concept' that helps form the social and religious community, possibly even the State (its identity and constitution). Appeals to an 'almighty' God, or the 'monarche' of God, or to a 'Lord,' or to more specified concepts of omnipotence, have direct social and political consequences. Kings, Queens, dictators, and rulers in many faiths, past and present, have depended on such conflation. Any present sociopolitical governance can be a divine instrument and referent that is used in the furtherance of subjugation and legitimization.

Facilitative power describes the points of access through which power can be reached, reified, and exchanged. Episodic power can be used to describe the 'surges' or 'events' of power that may alter the shape, perceptions, or behavior of individuals and groups. Central to these three 'circuits' is organization, and many a religious movement is able to demonstrate systematic configuration in each field. For example, the ideology present in postmodern charismatic worship could be said to be dispositional; the charismatic leaders are facilitative; the 'invocation' of the Holy Spirit a 'cue' for episodic manifestations of power to be unleashed (Percy 1996, 1998).

Episodic power is less easy to identify as a factor within mobilization. Any seminal moment can turn an

ordinary crowd into a passionate army. A martyrdom, miracle, or the raising up of a messiah can provide a single moment from which power flows. The declaration of *Jihad*, the proclamation of a fast, or the threat of heresy to authority can do the same. The very genesis of religious mobilization can be traced back by believers to an event that revealed, defined, and motivated a new community to hold, channel, and disperse power in new and particular forms.

Charisma is also a factor within mobilization. One way of understanding the place of agencies in circuits is to see them as process of power-exchange, a mechanism where power is given up or received, or raw material transformed into a type of 'power.' In other words, John Wesley may have been a charismatic religious leader, but he also functioned within a charismatic situation, which he could not entirely control, although he might have directed.

An account for the resurrection of the charismatic religious leader in postmodernity lies, in part, in Weber's own writings (which in turn is indebted to Nietzsche's notion of 'will to power'). Weber differentiated between 'religion' and 'magic' in the way that he differentiated between ancient and modern. Religion was rational, organized, and functional. Magic was primitive, a legitimate form of domination, and placed power outside ordinary temporal spheres (1947). Within magic, there were prophets, magicians, and shamans. Weber saw the prophet as the most significant bearer of charisma, since their claims were based on personal revelations that ultimately developed a personal following (1965). Yet these distinctions are to be understood as fluid: priests and prophets practiced magic, and there was routinization in the prophetic as much as in the religious. As nearly every preacher knows, propaganda is the highest form of political authority: aims and objectives are seldom met by organization or negotiation, but rather by power and mobilization.

2. Summary: Power and Mobilization

Recent civil strife in Northern Ireland illustrates religion's capacity to promote both social cohesion (e.g., Corrymeela Community) and conflict (e.g., Ian Paisley's anti-Roman Catholic rhetoric). Behind this aspect of sociality lies a complex history of the struggle for power and factors in mobilization—both for war and peace. In terms of power, the divisions between Protestant and Catholic are simultaneously elementary and complex. The 'two traditions' do not just represent different religions, but rather different politics, histories, and ideologies. In terms of mobilization, the preservation and flourishing of Protestant and Catholic communities has often been cultivated through episodic and dispositional conflict, which has seen many casualties and fatalities. Power and mobilization are deeply intertwined through armed

hostility, marches of protest or ritual (Orange Day Parades), sectarian division, and political aspiration. Two 'civil religions' effectively have promoted civil division.

Yet the hopelessness of this situation, in which two forms of 'civil religion' (in terms of their power and mobilization) have failed to cancel each other out, has given rise to a third option: a peace process. Here, religion has been no less involved, although the forces for power and mobilization have recognized that a political solution should bring a degree of peaceable co-existence between competing religious-sectarian convictions. The emerging and fragile détente serves to remind that the goal for many who first engage in hostilities is actually to end conflict; whether they succeed or not, power and mobilization often end in some kind of settlement.

At the beginning of the twenty-first century, it still remains the case that some of the bloodiest conflicts known to humanity have been inspired by or utilized powerful religious symbols and ideas. Keeping the faith 'pure' and protecting the sacredness of the *Koran* has given rise to *Jihad* and *fatwah*. Promised lands and sacred sites continue to call communities and nations to arm, from Judaism to Hinduism, and from Ulster to Amritsar. Equally, faith and fear in the shadow of the Millennium remain a source of violence and violation, from Jonestown to Waco, from Aum Shinrikyo to the Manson murders, and from the Solar Temple and Cargo Cults through to Heaven's Gate. In the midst of this, 'established' religions continue to war in regions such as the Balkans, at great human cost. The history of power and mobilization in religion is, all too often, a litany of tragic events, a requiem for the human race.

Yet to see religion like this is only to view it from the dark side. Religion remains a vital force in much of the world. Millions if not billions of the world's population have a 'religious impulse'; this involves a quest for meaning 'that goes beyond the restricted empirical existence of the here and now' (Haynes 1998). Religion inspires; it binds together, creating communities of affinity and affectivity. It offers revival, revolution, and surprise; and then again respect for the past, remembrance and ritual. Gandhi, Pope John Paul II, and Martin Luther King have been instrumental in turning back tides of nationalism, communism, and racism (even though they have all used a reconfigured form of nationalism and religion to achieve political and social ends). Religion can be an irresistible force that can move the apparently unmoveable. For many, religion remains a viable form of liberation; some hope in the midst of despair.

In spite of the attempts of some sociologists of religion to retain substantial elements of a secularisation thesis (e.g., Wilson, Bruce), religion stands, in many parts of the world, as a protest against and an alternative to 'secular' life. Will American Christian fundamentalists, Egyptian and Algerian Islamists, or

Hindu nationalists someday become nonreligious? It seems unlikely: religion retains its power, and its capacity to mobilize, evangelize, and scandalize.

Any discussion of power and mobilization in religion must take account of the following dynamics. First, religious groups and communities will perceive themselves as vessels or bodies that somehow hold, manifest, or reify divine power in a particular way, distinguishing themselves from 'society,' or perhaps their enemies, or other groups deemed to pose a threat or competition. Second, the structure of the group or community will most likely 'mirror' a proposed divine order; there will invariably be a conflation between the offices of leadership and the concepts of deity. Leaders of religious communities are supposed to resemble their gods. Many social communities attempt to correspond to a higher order that has theistic resonance. And yet, no democracy is immune from theocratic comparison; no theocracy can escape the process of routinization, as well as a degree of secularization. Third, there will be some sort of theology of power that will account for how divine power is to be reified and passed on: this is a key determinant in mobilization. Here, leadership, sacred texts, rituals, symbols, or stories play a vital part in moving believers from passivity to passion, and then to power. Every form of transcendence, no matter how 'other,' is received individually and communally, and worked out in social and political spheres. Fourth, the power that is witnessed to will be antisocial or suprasocial, requiring believers to be mobilized for change and social transformation whenever they perceive their own sense of divine order to be under threat, or require a new order to be imposed. It is against this background that an Islamic nationalist uprising, or Catholic and Evangelical picketing of abortion clinics, or the work of reconciliation and peace in South Africa and Northern Ireland, are to be ultimately understood.

See also: Religion: Definition and Explanation; Religion: Evolution and Development; Religion: Morality and Social Control; Religion: Nationalism and Identity; Religion: Peace, War, and Violence; Religion, Psychology of; Religion, Sociology of; Religious Nationalism and the Secular State: Cultural Concerns

Bibliography

- Beckford J 1987 The restoration of power and the sociology of religion. In: Robbins T, Robertson R (eds.) *Church-State Relations*. Transaction Books, New Brunswick, NJ
- Bettis J (ed.) 1969 *The Phenomenology of Religion*. SCM, London
- Casanova J 1994 *Public Religions in the Modern World*. Chicago University Press, Chicago
- Clegg S 1989 *Frameworks of Power*. Sage, London, pp. 210-15
- Dahl R 1957 The concept of power. *Behavioral Science* 2

- Davis K 1942 A conceptual analysis of stratification. *American Sociology Review* 7(3): 309
- Dowding K M 1996 *Power: Concepts in the Social Sciences*. University of Minnesota Press, Minneapolis, MN
- Foucault M 1984 Space, knowledge and power. In: Rabinow P (ed.) *The Foucault Reader*. Pantheon Books, New York
- Haynes J 1998 *Religion in Global Politics*. Longman, London
- Hemmerle K 1975 Power. In: Rahner K (ed.) *Encyclopaedia of Theology: A Concise Sacramentum Mundi*. Burns and Oates, London
- Lewis I M 1996 *Religion in Context: Cults and Charisma*. Cambridge University Press, Cambridge, UK
- Lukes S 1978 Power and authority. In: Bottomore T, Nisbet R (eds.) *A History of Sociological Analysis*. Basic Books, New York
- May R 1976 *Power and Innocence*. Fontana/Collins, Glasgow, UK
- McGuire M B 1992 *Religion: The Social Context*. Wadsworth, Belmont, CA
- Morris P 1987 *Power: A Philosophical Analysis*. St. Martin's Press, New York
- Percy M 1996 *Words, Wonders and Power: Understanding Contemporary Christian Fundamentalism and Revivalism*. SPCK, London
- Percy M 1998 *Power and the Church: Ecclesiology in an Age of Transition*. Cassell, London
- Peters R 1996 *Jihad in Classical and Modern Islam*. Marcus Wiener, Princeton, NJ
- Rahner K 1966 The theology of power. In: Rahner K (ed.) *Theological Investigations*. DLT, London, p. 391
- Parsons T 1960 *Structure and Process in Modern Societies*. Free Press, Glencoe, IL
- Van der Leeuw G 1938 *Religion in Essence and Manifestation*. Allen & Unwin, London
- Weber M 1946 The social psychology of the World religions. In: Gerth H H, Wright-Mills C (eds.) *From Max Weber*. Oxford University Press, New York
- Weber M 1947 *The Theory of Social and Economic Organisation*. Oxford University Press, New York
- Weber M 1963 *The Sociology of Religion*. Beacon Press, Boston
- Weber M 1965 *The Sociology of Religion*. Methuen, London
- Weber M 1968 *Economy and Society*. Bedminster, New York, Vol. 1
- Wright-Mills C 1956 *The Power Elite*. Oxford University Press, New York
- Wright-Mills C 1958 The structure of power in American society. *British Journal of Sociology* 9(1)

M. Percy

Religion: Morality and Social Control

In most societies at most times religion has been the key institution devoted to answering three of the fundamental questions human beings have posed, namely, 'How and why were the universe, life and humankind created?,' 'What moral rules should we follow in order to lead good lives?,' and 'What becomes of us after we die?.' Most of the great world religions

have answered the three questions together as part of a single coherent narrative. With some religions a just and righteous God was said to have created the world and its inhabitants, communicated moral decrees to them by revelation, and would after their deaths sit in judgment on them and assign them to paradise or to hell according to whether or not they had obeyed these precepts. Other religions held that the sum of an individual's good and bad deeds in this and previous lives would determine the nature of a reincarnation after their death according to the law of karma (Davies 1999). Clearly, beliefs and institutions of such fundamental significance are of great importance for social control. As Professor A G N Flew (1987, p. 71), the distinguished philosopher of religion, has noted:

Can we not understand the hopes of the warriors of Allah who expect if they die in Holy Wars to go straight to the arms of the black-eyed houris in paradise? Can we not understand the fears of the slum mother kept from the contraceptive clinic by her priest's warning of penalties for those who die in mortal sin? Of course we can: they both expect—and what could be more intelligible than this?—that, if they do certain things, then they will in consequence enjoy or suffer in the future certain rewards or punishments.

The nature of the social control thus exercised varies greatly from one society to another, as does the degree to which emphasis is placed (a) on social conformity as part of the natural order of things, (b) on establishing, monitoring, and controlling the nature and content of an individual's conscience, or (c) on claiming to be able to affect or at least to predict a person's fate after death.

Sometimes these forms of social control have simply reinforced other forms of power available to the secular rulers of a society, as when these rulers have also been the key religious authorities or the two groups have been closely allied. Likewise, resistance to the ruler might well be reinforced by religious dissent (Martin 1978) but with the caveat that the degree of social control that the dissenting group, whether the seventeenth century English upholders of the Cromwellian Commonwealth, the Roman Catholic exorcisers of a moral monopoly in Ireland (Inglis 1998), or the Sikhs of the Punjab, exerts over its members may be exceptionally high.

Religion can, of course, be an independent source of power and social control, sometimes exercised over matters that are of little interest to the secular authorities. The medieval Roman Catholic Church's forbidding of marriages between cousins up to the sixth degree or the refusal of the Greek Orthodox Church to permit a person to marry a fourth time after the deaths of three spouses were forms of social control that were at times a nuisance to the rulers of the relevant societies, who were concerned mainly with forming dynastic alliances and producing male heirs. They were part of a set of independent social controls

Table 1
Trends over time in England and Wales

	Reforming Britain 1860–1905	Respectable Britain 1905–60	Permissive Britain 1960–2000
Religious adherence and membership	Rising	Falling slowly and unevenly	Falling rapidly
Crimes of violence and dishonesty	Falling	Rising slowly and unevenly	Rising rapidly
Illegitimacy	Falling	Steady at a low level except in wartime	Rising rapidly
Drug and alcohol abuse	Falling unevenly	Falling to a low level	Rising

exercised through these churches' monopoly over the creation and dissolution of marriages, and hence over sexuality, reproduction, legitimacy, and inheritance, i.e., over crucial aspects of not only personal but also material and political life.

This form of social control may be seen as what Polanyi (1951) has termed deliberate or corporate order comparable to, say, a centrally planned economy, statute law, or the internal workings of a private bureaucracy. In each case there are, of course, divisions of interest, contested areas of authority, and competition between individuals and sectors using the exploitation and interpretation of rules to enhance power, but in principle there is a hierarchical system that may not be contested beyond a certain point. Rome has spoken, the matter is closed and the Brahmins alone can determine a caste's pedigree and thus influence its position in a hierarchy of purity of which they are the apex (Dumont 1970). Yet in all cases in traditional societies, most social control would have been local. Violence, theft, or sex with other men's wives would be kept in check by local opinion in one's village, *mir*, *mura*, *jati*, or kinship group rather than the direct use of priestly authority, the threat of divine wrath, or the dread of hellfire or of reincarnation as a scavenging dog. A mixture of respect for and fear of local chiefs, landowners, or state officials will also have played a part. Religion was just one part, though an important part, of local social control. Beneath the designed order lay a myriad local communities held together by an alternative form of order, the interactive order of face-to-face communities, the essence of *Gemeinschaft*.

Polanyi (1951) contrasts deliberate or planned order with spontaneous order which can be observed wherever stable arrangements evolve as a result of a large number of independent but interconnected decisions, as with the operation of markets through the price mechanism, systems of common law based on cumu-

lative precedence, or scientific progress through conjecture, experiment, refutation, and provisional confirmation. No one directs or designs these large, widely dispersed, flexible, and responsive instances of spontaneous order, yet they are much superior in their efficacy to most attempts to create a centrally controlled designed order. Their equivalent in religious terms and one with a crucial importance for social control may be seen in the flourishing of Protestant sects and denominations, and even of factions within the established churches in Britain during the nineteenth century, as the world's first urban, commercial, and industrial society was established and consolidated. This religious spontaneous order enabled Britain to shift from being a society of small, rural villages and towns based on local social controls to one of large, modern, urban conurbations while experiencing not crime and chaos but steadily declining rates of crime (Gattrell and Hadden 1980), illegitimacy (Hartley 1975), and drink and drug excesses (Wilson 1940). Britain was a more crime free and less deviant society in 1900 than it is today and than it had been 50 years before (Davies 1994).

Over time the pattern of crime and deviance in Britain follows a U-curve, falling markedly in the later half of the nineteenth century (Gattrell and Hadden 1980), reasonably stable in the first half of the twentieth, and rising very rapidly from the mid-1950s (Davies 1994). These changes are inversely shadowed by the growth and decline in church adherence in England and Wales, which grew rapidly in relation to population in the latter half of the nineteenth century and peaked in about 1904–5 (Brown 1992). There then followed an irregular decline with occasional periods of recovery until the late 1950s, after which there has been a steep and unreversed decline. There was a J-curve in piety which was an exact inverse of the U-curve in crime and deviance. No other variable fits this pattern with any such degree of exactitude. It is very

Table 2
Sunday School Enrollment

Year	Percentage of population aged less than 15 years enrolled in Sunday Schools
1818	12
1851	38
1891	52
1901	53
1911	51
1931	46
1961	20
1989	14

strong proof indeed of the effectiveness of spontaneous religious order as a means of social control. The relationship is summed up in Table 1 (Davies 1994, Hartley 1975, Lynn and Hampson 1977) showing the main trends between 1860 and 2000, trends that divide the period clearly into three phases 'Reforming Britain,' 'Respectable Britain,' and 'Permissive Britain' (Davies 1975).

Religious adherence and membership, although at all times involving a minority of the adult population, were a good index of the strength of the values of respectability, and thus of spontaneous self- and social control in British society. The incidence of deviant behavior clearly fell and then rose in Britain as this form of social control rooted in religion first became stronger and then later went into decline. The churches and chapels were the crucial institution promoting respectability in the nineteenth and early twentieth centuries, and membership and attendance were one way in which respectability could be acquired and proclaimed. Rising levels of membership and attendance gave those who were committed to their religion the confidence to promote and foster the virtues of self-control and respectability among the non-attending majority whose attachment to a church or denomination was at best nominal. To a substantial extent, though, they were receptive to the morality of respectability, and in particular they wanted it instilled in their children. The rise and the later fall of Sunday School enrollment and attendance shown in Table 2 (Gill 1992, Laqueur 1976, Martin 1967, Wilkinson 1978) is perhaps an even better measure of the changing importance over time of religion as a means of social control. It has been suggested reliably that between two-thirds and three-quarters of those enrolled were generally present in school on a Sunday (Gill 1992, p. 96).

The importance of popular Protestantism as a source of spontaneous social control and moral order in Britain during a time of massive internal migration to the new industrial and commercial towns and conurbations (Munson 1991) can be seen at work today in contemporary Latin America, a continent

currently undergoing a similar transformation (Martin 1990). Migrants leaving small, agricultural communities to work in the industrial and services sectors in societies as diverse as Chile and Guatemala have, like their British predecessors, sought to create for themselves through religion a spontaneous moral order, a social control based on respectability (Martin 1998, Martin 1990).

Protestantism has 'once again created a critical mass of persons for whom self-discipline, a work ethic and the values of domestic rectitude have become second nature' (Martin 1998, p. 28. See also Weber 1930). The truth has set them free from what otherwise would have been a morally degrading existence. Whether they will be able to sustain these virtues in the long run remains to be seen.

The importance of religion as an autonomous and potent means of social control has often been neglected by sociologists and political scientists, who are blinkered by their own obsession with economic and political power, and unable to appreciate the importance of religion. When they do turn their attention to religion as a means of social control, it is by choosing selectively examples such as the Inquisition, the Ayatollah Khomeini's regime in Iran, or human sacrifice by the Aztecs, which enable them to caricature religion-based social control as backward, oppressive, and irrational. On the contrary, social control rooted in religion is as varied as any other form of social control, and the distinctive spontaneous forms of social control that religion has provided in some countries during the difficult transition from a traditional to an urban and industrial society have been remarkable for their peaceable nature and civilizing influence.

See also: Religion and Economic Life; Religion and Gender; Religion and Politics: United States; Religion and Youth; Religion: Family and Kinship; Religion, Sociology of; Stigma, Social Psychology of

Bibliography

- Brown C G 1992 A revisionist approach to religious change. In: Bruce S (ed.) *Religion and Modernization*. Clarendon Press, Oxford, UK, pp. 31–58
- Bruce S (ed.) 1992 *Religion and Modernization*. Clarendon Press, Oxford, UK
- Davies C 1975 *Permissive Britain*. Pitman, London
- Davies C 1994 Does religion prevent crime? *Informationes Theologiae Europae, Internationales ökumenisches Jahrbuch für Theologie*, Peter Lang, Cologne, Germany, pp. 76–93
- Davies C 1999 The fragmentation of the religious tradition of the creation, after-life and morality: Modernity not post-modernity. *Journal of Contemporary Religion* 14(3): 339–60
- Dumont L 1970 *Homo Hierarchicus: The Caste System and Its Implications*. Weidenfeld and Nicolson, London

- Flew A G N 1987 The logic of mortality. In: Badham P, Badham L (eds.) *Death and Mortality in the Religions of the World*, Paragon House, New York, pp. 171–87
- Gattrell V A C, Hadden T B 1980 The decline of theft and violence in Victorian and Edwardian England. In: Gattrell V A C (ed.) *Crime and the Law, The Social History of Crime in Western Europe since 1500*, Europa, London
- Gill R 1992 Secularization and census data. In: Bruce S (ed.) *Religion and Modernization*. Clarendon Press, Oxford, UK, pp. 90–117
- Hartley S F 1975 *Illegitimacy*. University of California Press, Berkeley, CA
- Inglis T 1998 *Moral Monopoly, The Rise and Fall of the Catholic Church in Modern Ireland*. University College Dublin Press, Dublin, Ireland
- Laqueur T W 1976 *Religion and Respectability, Sunday Schools and Working Class Culture*. Yale University Press, New Haven, CT
- Lynn R, Hampson S L 1977 Fluctuations in national levels of neuroticism and extroversion 1935–1970. *British Journal of Social and Clinical Psychology* 16: 131–8
- Martin B 1998 From pre-modernity to post-modernity in Latin America: The case of Pentecostalism. In: Heelas P (ed.) *Religion, Modernity and Post-modernity*. Blackwell, Oxford, UK, pp. 101–46
- Martin D 1967 *A Sociology of English Religion*. Heinemann, London
- Martin D 1978 *A General Theory of Secularization*. Blackwell, Oxford, UK
- Martin D 1990 *Tongues of Fire, The Explosion of Protestantism in Latin America*. Blackwell, Oxford, UK
- Munson J 1991 *The Nonconformists, In Search of a Lost Culture*. SPCK, London
- Polanyi M 1951 *The Logic of Liberty*. Routledge & Kegan Paul, London
- Weber M 1930 *The Protestant Ethic and the Spirit of Capitalism*. Allen & Unwin, London
- Wilkinson A 1978 *The Church of England and the First World War*. SPCK, London
- Wilson G B 1940 *Alcohol and the Nation*. Nicholson and Watson, London

C. Davies

Religion: Nationalism and Identity

Nationalism and religion share the ambiguous status of being the most potent and inspiring sources of collective identity, but also the most destructive forces for mobilizing large numbers of people and generating social conflict and change. The conventional view of their relationship is conflictual and evolutionary: the nation is opposed to the religious community, as nationalism is to religion, and national identities must, and will, in time supersede obsolete religious allegiances.

For ‘modernists’ like Ernest Gellner, Tom Nairn and John Breuilly, nationalism is a purely modern, secular ideology; it appears only when the age of tradition and religion gives way to that of modernity

and the nation-state. This was very much the view of Elie Kedourie, for whom nationalism was a doctrine invented in Europe at the beginning of the nineteenth century. As a secular ideology of collective Will and as a fruitless quest for unattainable perfection on earth, nationalism embodies a godless, antinomian chiliasm which ultimately can be traced back to medieval Christian millennial doctrines. Its effect is to subvert duly constituted authority and destroy the normal barriers between the public and private domains in the interests of a pure, acosmic brotherly love—in Africa and Asia, as well as in its European seedbeds. With equal regret and ire, but from the other end of the ideological spectrum, Marxists like Eric Hobsbawm can only see in nationalism, particularly of the ‘divisive’ post-1870 ethnolinguistic variety, a new set of ‘invented traditions’ in the hands of ruling capitalist elites and a despairing popular reaction, like fundamentalist religion, to the massive forces of social change and the failure to realize the project of socialist internationalism. Nationalism, in short, like religion, is increasingly a diversion and obstruction to the great globalizing ‘movement of history.’

This view has not gone unchallenged. Impressed by the evidence of ‘religious nationalisms’ in Asia, Africa, the Americas, and Europe, scholars like Mark Juergensmeyer have argued that ‘nationalism’ and ‘religion’ are often natural allies, and that religious movements like Shi’ism are seeking to reclaim the nation from the embrace of the secular state. For others such as Conor Cruise O’Brien, this affinity of religion and nationalism, or ‘holy nationalism,’ can be traced back to the Old Testament. It is also possible to argue that such well-known cases of secular nationalism as Jacobin France, Kemalist Turkey, Stalinist Russia, and Maoist China did not attack religion as such, and were in any cases transient phases of revolutionary politics which always sees the return of a society’s ‘central values.’ From the standpoint of collective identity, religion and nationalism can often be mutually reinforcing, and hence classic ‘weapons of the bourgeoisie,’ in Leninist parlance.

A third view sees in nationalism a species of religion in its own right, competing with traditional world or local religions for mass allegiance. Already in the 1960s, functionalists like David Apter had shown how nationalism performed the functions of a mass-mobilizing ‘political religion,’ a characterization recently developed by George Mosse in his penetrating studies of the liturgies of the civic religion of European, mainly German, nationalisms, and by Bruce Kapferer who, in an illuminating comparative study of Sri Lankan and Australian ideological politics, has demonstrated the ways in which a nationalist religion copes with problems of ontological insecurity and ethnic conflict. Benedict Anderson, too, though he fails to develop the point, sees nationalism more closely aligned to religion than political ideology, and the imagined community of the nation as emerging

out of, and against, the earlier world religious communities.

Each of these views has much to commend it, and can point to considerable supporting evidence. Equally, it is not difficult to find counterexamples. For every 'secular' nationalism drawing on ancient classical or pagan motifs and myths such as the Finnish Kalevala or the Homeric Iliad, we can find equally fervent religious nationalisms inspired by their respective religious ideals and appropriating their ethno-religious heritages, as the modern Hindu Indian movement has drawn on the Ramayana and Mexican nationalism on the cult of the Virgin of Gaudeloupe. The influence of traditional religions may prove equally ambivalent. If Buddhism underpins and is reinforced by Burman and Sinhalese national identities, the conflict between Protestant and Catholic Christianities only served to undermine the growing sense of German or Swiss national identities. As for Islam, it can buttress an all-Arab nationalism, but can equally undermine the Arab circle of allegiance in favor of the wider Muslim umma, or, in the hands of radical populist regimes like the Ba'ath Party, support the narrower circle of the Syrian or Iraqi nation-state.

This suggests that there can be no simple characterization of the relationship between the two modes of culture, identity, and belief. We have only to recall the contradictions that bedeviled the French revolution, often regarded as the first full-blown case of secular nationalism. From the very first oath sworn by the deputies on the tennis court in June 1789 to the last Jacobin journée, the *Fête de l'Être Suprême* in June 1794, classically-inspired secular and Christian religious motifs and rituals were combined, inspiring in the multitudes a powerful emotional blend of patriotic, fraternal political activism. Thus, as Simon Schama points out, at the *Fête de la Fédération* held on the Champs de Mars on July 14, 1790, Lafayette administered the oath to country and king and Talleyrand provided the mass and benediction, intoning: 'Sing and weep tears of joy, for on this day France has been made anew'; while over at Notre Dame, they performed a half sacred, half profane cantata entitled *Prise de la Bastille*, set to passages from the apocryphal, patriotic Book of Judith.

Similar religious motifs pervaded other nationalist celebrations; for example, the patriotic German festivals of the Wartburg in 1817 and at Hambach in 1834, described by Mosse, as well as the six-hundredth anniversary of the Swiss Confederation and the Oath of the Rütli, held in Schwyz and on the Rütli meadow in 1891. In the twentieth century, nationalist celebrations continued to employ religious imagery, most notably in the cult of 'fallen soldiers' and patriotic sacrifice, whether in the two world wars or the many more regional wars fought in the name of the nation, or the nation-to-be.

These considerations point to a more complex

interplay and layering of religious and national identities and experience. Much of the confusion surrounding this problem stems from divergent definitions of both 'religion' and 'nationalism.' The evolutionary modernist 'secular replacement' perspective employs a traditional definition of 'religion' in which our everyday world is believed to be controlled by a supraempirical order (and/or beings) which is the source of salvation—a theocentric and cosmic belief-system; while its definition of nationalism as the aspiration of a people and culture to be autonomous and united in its homeland emphasizes its secular, linear, and anthropocentric character. It is easy to see how, on this schema, nationalism and religion become opposed: the nation replaces the deity, the nationalist movement the church, and posterity the after-life. For the alternative view which sees nationalism and religion as 'natural allies,' or even conflated as 'holy nationalisms,' nationalisms can take as many forms as there are peoples and nations; similarly with religions. So nationalism and religion are really continuous and mutually reinforcing, or superimposed; nationalisms are expressions of belief, or 'faith-communities,' just as churches or religious communities are really 'nations.' As John Armstrong suggests, religious communities easily slide into ethnic groups, and vice versa. But this is, once again, to muddy the waters, and preclude the possibility of a more discriminating analysis. After all, are all nationalisms equally 'holy'? Does religion always reinforce national identities? Isn't the 'faith' of nationalists quite different to that of a traditional believer in a local or world religion? That, at least, is how it has often appeared to many participants who faced difficult choices about the claims of God and state, religion and nation, particularly where the religion was transterritorial, even universal in its claims. Hence also the vital differences between secularizing nationalisms like the republicans in France or the Westernizers in Tsarist Russia, and the religio-monarchical nationalists in France and the religious narodniks in Russia.

This is where the third, or 'political religion,' perspective may help. We do not have to subscribe to its belief in the substitution of 'religion' by 'nationalism,' but we can with profit employ its wider, Durkheimian approach to religion—as a system of beliefs and practices relative to sacred things—to understand how and why modern nationalisms, defined as ideological movements for the attainment and maintenance of autonomy, unity, and identity for populations deemed to constitute 'nations,' can fulfill many of the identity-creating functions of traditional religions. This makes nationalism both a competitor to, and an ally of, traditional religions, depending on historical and social circumstances.

In practice, this tension and symbiosis between 'religion,' so defined, and 'nationalism,' produces a series of familiar mass phenomena. These include:

(a) The politicization of religion, in which tra-

ditional religious motifs are transmuted into secular nationalist features. Thus prophets, saints, and religious lawgivers like Moses, Muhammad, and Confucius are metamorphosed into national heroes, religious revelations are turned into national shrines such as Yasna Gora and Guadeloupe, scriptures like the Bible and Bhagavad-Gita are re-read as national epics, and miracles like the eight-day lamp of Chanukah become national feats of the Maccabees. Even commemorations like the death of Hussein at Karbala, or the Exodus across the Red Sea, are re-enacted as nationalist festivals in Iran and Israel.

(b) The messianization of politics, whereby the nation-state is sanctified and exalted, and its leaders are endowed with religious charisma. This was how their followers came to regard nationalist leaders like Nkrumah, Nasser, Herzl, and Soekarno, investing them with superhuman qualities as messianic deliverers and saviors. But the messianization of politics and the sanctification of the nation-state does not stop with the leaders; it includes the sacralization of ethnohistory, such as the Gaelic history of Ireland, the Christian mission of Poland, and the Hinduization of ancient India; the belief in millennial destiny such as we find in the United States or communist Russia; and the 'invention of traditions' such as the rites and ceremonies of Bastille Day, Independence Day, the Rutli Oath, and Armistice Day at the British Cenotaph.

(c) Religious populism, or the exaltation, even deification, of the 'people,' the ordinary people, as opposed to the elites. This may include the quest for purification of the people and their culture, in the tradition of Herder's cultural populism, sought by so many Central and East European nationalists, or by Indian, Arab, and West African nationalists like Edward Blyden; the return to 'authentic' folk traditions and their elevation to national status, as with the Finnish regard for the Kalevala and the Estonian for the Kalevipoeg, the Gaelic revival in Ireland, and the Mexican and Peruvian return to pre-Columbian Indian culture; and the glorification of the simple or, common people,' a political tradition that can be traced back to St. Joan, through the narodniks and Mussourgsky's depiction of the Simpleton-Prophet in Boris Godunov and Levin's discovery of truth among the peasants in Tolstoy's *Anna Karenina*, to Peronism, Poujadism, and the modern Russian nationalists.

Each of these mass phenomena mixes sacred and secular elements, or more accurately, sanctifies the worldly in unexpected, yet, given the background, quite logical ways. On this reading, nationalism is, as Kedourie affirmed, a secular, this-worldly, ideological movement, with clear terrestrial and anthropocentric goals. It does not seek a world beyond this earth, nor does it posit an unseen world guiding and making sense of the world of our senses. At the same time, nationalisms draw, in varying degrees, on the religious background and tradition of the populations whom

they deem to constitute nations. This is not, or not usually, a case of manipulation of 'atavistic mass emotion,' as Kedourie claims, nor of a need to control and channel the unleashed energies of the newly mobilized and enfranchised masses, as Hobsbawm argues. It is quite simply that most nationalists are steeped in the religious traditions of their forbears, even where they have rejected them, and that to make sense of their new political message, they have no alternative but to couch it in language and images that are familiar and comforting to their audience. Rarely do nationalists seek to destroy an old, religious identity, in order to build up a totally new national one, although they may attack the guardians of religious tradition who refuse to reinterpret the tenets of the faith in a 'politically relevant,' i.e., nationalist, manner. Hence, of any nationalism we must ask, not whether it is 'religious' or 'secular,' but how and to what extent it mingles sacred and worldly elements, and with what consequences. Similarly, we should enquire not whether the nation in question is seen as an 'ethnic' or 'civic' unit, and a religious or secular community, but how and to what degree its national identity is infused with sacred motifs and religious images, and with what kind of effects.

It may be useful to grasp the complexity of this relationship by distinguishing three 'levels of analysis.' The first is phenomenal: it describes and analyzes the kind of nationalism that is in the public domain, the official version proposed by the elites. Thus, we may use a shorthand to describe French nationalism during the Revolution, or Turkish nationalism under Ataturk, as avowedly 'secular,' even irreligious. And we may range official nationalisms and their political symbolism along a continuum of religious to secular poles. Here, of course, we are using the traditional definition of 'religion' as a faith that looks beyond this world for salvation.

A second level of analysis seeks to determine the content of unofficial and more popular kinds of nationalism and national identity. Here we can expect to find considerable variations in the extent and kind of religious motifs and sacred imagery and symbolism in different national identities and nationalisms, and in different religious traditions. On the one hand, are cases where classical, secular myths and imagery predominate; this was the case in France and the United States during and after their Revolutions, and in Italy from the Risorgimento to fascism. On the other hand, we find examples where religious images and motifs pervade the sense of national identity; this was the case among the Greek peasantry in 1821, the Russian peasantry under Tsarism, or the Ethiopian peasantry under the Emperors, as it is in Pakistan and Sri Lanka to this day. In all these cases, the respective traditional religions had played a large part in molding the sense of ethnic identity prior to the modern epoch; and each of these traditional world religions contained a large repertoire of sacred elements—scriptures,

liturgies, symbols, rituals, scripts, ceremonies, and the like. Some religions, indeed, like Christianity and Judaism, as Adrian Hastings argues, have provided particularly fertile terrain for the reproduction and proliferation of vernacular nations based on the Biblical prototype of the ancient Jews in their land.

Moreover, even in more advanced industrial societies, sacred imagery and religious symbolism may persist and underpin a sense of national identity, even where the majority of the population are, by conventional measurements, no longer practicing believers. In the United States, for example, for all its official civic secularism, large numbers of citizens espouse religious forms, images and rituals of the American nation, and continue to envisage its destiny in providential terms. The same is true, though in quite different ways, in France and Israel, despite the avowed secularism of the majority of their populations; the historic rationale and destiny of the nation remains part of the sacred realm, and the various religious symbols, images and motifs have done their work of forging the *raison d'être* of the French and Israeli nations.

This brings us to the third level of analysis. It requires us to adopt a wider, Durkheimian approach to and definition of religion, and a different perspective on the significance of the nation. On this general level, the nation can be considered as a 'sacred communion' that unites the dead, the living, and the yet unborn. This is not just a matter of imagination, as Benedict Anderson claims; it is even more a matter of will and feeling. The nation is, in the first place, a willed and a felt community, a moral community of collective purpose whose members are bound by powerful sentiments arising from a sense of presumptive ancestral affinity. (Presumptive, because myths of origin rarely correspond to the historical record of origins, let alone actual biological descent.)

This points to the dual origin of the nation: on the one hand, as a community based on presumed ethnic origins and descent, tracing their genealogy back to a specific place, time, and ancestor; and on the other hand, as a union of all those espousing common moral purposes and assumptions, a community of destiny who trace their ideological and moral lineage to the sacred properties of beliefs and practices of a community of the faithful. The fusion of these two kinds of community over the *longue durée* creates the basis for the modern nation.

What are the sacred properties of the beliefs and practices of the faithful? And how are they transformed into the national 'religion of the people'? Four kinds of sacred property form the main 'pillars' of the nation: belief in the chosen people; attachments to a sacred territory; memories of the golden age; and rituals of commemoration of the glorious dead.

(a) By the concept of the chosen people, I mean not simply any community that exalts itself and carries ethnocentric sentiments of superiority and exclus-

iveness, such as we find throughout history. The meaning here is more specific. On the one hand, there is a broader sense in which elites of nations feel that they are entrusted with a sacred mission by the deity—to convert the heathen, protect the church, or defeat the infidel. By this divine trust and mission, they are sanctified and, by extension, their people become the elect; they derive their rationale as a community from this trust; and they feel exalted by the sense of their divine calling and the success of their leaders. It is not always clear what the effects of failure are, or indeed what constitutes failure; one has the sense that this kind of chosenness is flexible and capable of considerable reinterpretation.

So indeed is the second, more precise 'covenantal' meaning. Here chosenness is conditional and popular. The whole community is chosen by the deity for a specific purpose, but only for so long as it fulfills its task and mission. Failure to do so leads to the loss of its status as the elect. This is the essence of the 'covenant.' Its purpose is to sanctify the community and to separate its members from their profane surroundings. In the Old Testament model, the Israelites were set apart in order that they might fulfill God's commandments and hasten his universal salvation. They were to become a 'holy people' and a 'nation of priests.' In their footsteps followed Armenians and Orthodox Greeks, Amharic Ethiopians and Orthodox Russians, Afrikaners and Ulster-Scots, and many others in the monotheistic traditions.

The sense of ethnic election has been widespread in all ages and continents, particularly in the looser, missionary form. Its origins are obscure, and it takes many forms, from the conquering zeal of Muslim Arab nomads and Catalan or Hungarian knights to the Celtic-Christian monasticism of the Catholic Irish and the pioneering providential faith of Protestant settlers in America. In each case, chosenness crystallizes the emergent sense of community and confirms its sense of special destiny. As much as any rule of endogamy, it helps to keep the community intact and guard its borders. And, more than any birth rate or territory or conquest, it ensures the long-term survival and persistence of the ethnic community that espouses it. This is especially true of the stricter, more popular, covenantal forms of ethnic election.

The sense of chosenness has carried into the modern world, albeit often in more secular versions. De Gaulle's vision of France's cultural and political mission, the Great Society proclaimed by Johnson for the United States, even the 'New Britain' of Blair and New Labour, echo the missionary kind of election myth. Communist ideals, pared down to 'socialism in one country,' embodied the old myths more shrilly, as did various radical nationalisms in Asia and Africa, from the Meiji reformers in Japan to U Nu's Burma and Nkrumah's African Christianity in Ghana, drawing as they did on Islamic, Hindu, Buddhist, or Christian imagery, but ultimately attempting to fuse

the genealogy of ethnic community with the sacred properties of a moral union, into the sacred communion of the nation.

(b) The origins of ideas of sacred territory are equally obscure, but we may attempt a sociological reconstruction. Territory has been valued not only for security and, of course, as 'hearth and home,' but also as the arena in which the members of a community live, work, and die. And are buried. Reverence for, and protection of, the last resting places of our forefathers and foremothers is one of the main ways in which stretches of territory are invested with special feelings and meaning.

Allied to this is the special collective reverence felt for the tombs and the sites of miracles and exploits of holy men, sages and heroes or heroines. Pilgrimages to shrines like Yasna Gora, Jerusalem, Qom, or Mecca endow these locations with reiterated, and so more lasting, veneration and awe. By extension, any site where heroes or saints displayed the qualities and values central to the traditions of the community is invested with significance and helps to bind the community to its sacred land.

Once again, we are dealing with the confluence of two processes. On the one hand, the sacralization of territory, sketched above; on the other hand, the territorialization of memory. The latter implies two subprocesses: on the one hand, the naturalization of history, turning the community's history into an integral part of its environment and treating it as 'natural,' part of the landscape which it inhabits; on the other hand, the historicization of nature, transforming the community's habitat and landscape into an essential component of its historical development. The effect of these two subprocesses is to bind together community and land into what we may term an 'ethnoscape'—in which the *ethnie* (ethnic community) belongs to, and is an intrinsic part of its landscape, and the landscape belongs to, and becomes an integral part of, its community.

Once ethnoscaping develop, binding together *ethnies* and their poetic landscapes, it is a short step to attaching special sacred qualities to the land. The land is not only 'ours,' it is sanctified as the place of 'our' providential history and the chosen arena for the unfolding of the divine plan through the exploits of His, and our, saints, warriors, and sages. That, of course, is how the Old Testament concept transformed the 'promised' into the 'holy' land of Israel, and how the Swiss shepherds and herdsmen began to interpret the reasons for their remarkable victories over superior Habsburg oppressors, as if the purity of their Alpine mountain fastnesses reflected their simple virtues and steadfast faith in God. Today, ethnoscaping and sanctified nature have often become flashpoints of bitter conflict—in Ulster, Kosovo, the West Bank (Judea and Samaria), Amritsar, and Tamilnad.

(c) The concept of a golden age can only be grasped in the context of the broader 'ethno-history' of the

community. This is the set of traditional, often sacred, histories recounted by the members of the ethnic community, as opposed to the more dispassionate attempts by professional historians to reconstruct the history of the community. The line between ethno-history and ethnic mythology is often blurred, if we interpret 'myth' not as fiction, but simply as a widely-believed tale told about the heroic or sacred past to serve present needs and goals. This is particularly true of those periods and episodes which later generations of the community have come to remember as high or canonical epochs in the unfolding story of the community. These we may call the 'golden ages,' because the memory of these periods is invested with a specific nostalgia and reverence, because they come to serve as models or guides to future action.

There are various kinds of golden age—religious, political, economic, military, artistic—and there may be several in the shared memories of those ethno-histories that are rich in events and relics and are amply documented. Modern Greeks, for example, can look back to the golden age of Athens or to that of imperial Byzantium, Indians to the epoch of the classical post-Vedic states or of the Maurya or Gupta empires, and the English to the kingdom of Alfred and the Anglo-Saxons or to the Elisabethan age. Each of the ages forms part of a wider ethnohistory and its memory can act as the point of reference for a salvation drama of the nation. In such cases, the epoch in question becomes a model, an exemplification of 'our true nature' and a repository of 'our virtues,' spurring the community to emulation. Shared memories of golden ages create a sense of the extraordinary, exalt the community, and confer on their members a sense of collective dignity and the promise of national regeneration.

(d) Among the shrines of the sacred territory, the Tombs of Unknown Soldiers and the monuments to the fallen soldiers and individual warrior-heroes occupy a special place. These are the memorials to the glorious dead. In all ages, from the fallen at Thermopylae or Avarayr to the mass war graves of the two World Wars, the exploits and self-sacrifice of fallen patriot-soldiers and heroes have always evoked a special sombre reverence, and elicited a special determination not to forget their heroism and live up to their example.

In the modern age, the commemoration of the glorious dead has become central to the celebration of the nation. In this respect, the French Revolution with its cult of martyrs like Le Peletier, Bara, and Marat and its commemoration of famous patriots in the Pantheon, *Aux Grands Hommes*, *La Patrie Reconnaissante*, set the tone; while Napoleon's Arc de Triomphe collectivised the memory of the triumphant heroism of La Grande Armée. As George Mosse has documented so vividly, this idea became central to the emergent German nation, with its memorials to ancient heroes like Arminius and to modern heroes

like the Völkerschlachtendenkmal commemorating the fallen at the battle of Leipzig in 1813. In Britain, Lutyens' Cenotaph in Whitehall embodied the grief and pride of the nation; the annual Armistice Day ceremonies there and across the land, to the strains of Elgar and Chopin and the reveille bugle calls, evoke the terrible sacrifices in two World Wars and the collective resolve of the British in an atmosphere of sombre reverence.

There are few modern nations that do not have their memorials to the glorious dead, their ceremonies of commemoration, and their shrines, icons, and symbols of martyrdom and heroism. In this respect, as in the other sacred 'pillars,' nations and their artistic representations have taken over so much of the symbolism and emotion of earlier religious traditions, and infused their this-worldly beliefs in posterity and the judgement of history with earlier sacred longings for collective immortality.

In these ways, nations and their nationalisms have emerged on a religious foundation and are infused with many of their sacred motifs and properties—'religion' being understood here in its wider Durkheimian sense as a collective bond of obligation in respect of sacred things, continually rehearsed in ceremonies and rituals. At the same time, individual historical expressions of a nation's identity may cleave closer to this religious base, or depart from it to embrace a more secular ideal and formulation. However, even in these latter cases, the nation continues to draw upon some of the sacred properties of earlier beliefs and practices and so remains fundamentally a form of sacred communion, uniting its members in a moral and ideological community of intimate belonging and collective moral destiny.

See also: Ethnic Cleansing, History of; Ethnic Conflicts; Ethnicity, Sociology of; Identity in Anthropology; Identity: Social; National Character; Nationalism and Expressive Forms; Nationalism: Contemporary Issues; Nationalism: General; Nationalism, Historical Aspects of: Africa; Nationalism, Historical Aspects of: Arab World; Nationalism, Historical Aspects of: East Asia; Nationalism, Historical Aspects of: South Asia; Nationalism, Historical Aspects of: The West; Nationalism, Sociology of; Reformation and Confessionalization; Religious Nationalism and the Secular State: Cultural Concerns; Rokkan, Stein (1921–79)

Bibliography

- Akenson D H 1992 *God's Peoples: Covenant and Land in South Africa, Israel and Ulster*. Cornell University Press, Ithaca
- Anderson B 1991 *Imagined Communities: Reflections on the Origin and Spread of Nationalism*. Verso, London
- Armstrong J A 1982 *Nations before Nationalism*. University of North Carolina Press, Chapel Hill, NC

- Breuille J 1993 *Nationalism and the State*, 2nd edn. Manchester University Press, Manchester, UK
- Connor W 1994 *Ethnonationalism: The Quest for Understanding*. Princeton University Press, Princeton, NJ
- Gellner E 1983 *Nations and Nationalism*. Blackwell, Oxford, UK
- Hobsbawm E 1990 *Nations and Nationalism since 1780*. Cambridge University Press, Cambridge, UK
- Hobsbawm E, Ranger T (eds.) 1983 *The Invention of Tradition*. Cornell University Press, Ithaca, NY
- Juergensmeyer M 1993 *The New Cold War: Religious Nationalism Confronts the Secular State*. University of California Press, Berkeley, CA
- Kapferer B 1988 *Legends of People, Myths of State*. Smithsonian Institution Press, Washington, DC
- Kedourie E 1960 *Nationalism*. Hutchinson, London
- Kedourie E (ed.) 1971 *Nationalism in Asia and Africa*. World Publishing Company, New York
- Mosse G L 1975 *The Nationalization of the Masses*. Fertig, New York
- Nairn T 1977 *The Break-up of Britain*. New Left Books, London
- O'Brien C C 1988 *God-land*. Harvard University Press, Cambridge, MA
- Schama S 1989 *Citizens*, 1st edn. Knopf, New York
- Smith A D 1986 *The Ethnic Origins of Nations*. Blackwell, Oxford, UK
- Smith A D 1991 *National Identity*. University of Nevada Press, Reno, NV
- Smith A D 1998 *Nationalism and Modernism*. Routledge, London

A. D. Smith

Religion: Peace, War, and Violence

The relation of religion to peace, war, and violence (or nonviolence) is bound up in the various visions of right relationships and harmony within human beings and between collectivites, including nation-states and ultimately global society. Within Eastern traditions, for example, a Buddhist espousal of nonattachment and rejection of anger implies a peaceable attitude, and a Confucian or Taoist espousal of harmony implies reluctance to engage in depredation and offensive war. Within Hinduism a major strain, at least as interpreted by Gandhians, recommends a nonviolent approach and the kind of exemplary spiritual suasion embodied in *satyagraha*.

In all the Abrahamic traditions there is a positive evaluation of peace, so that 'Peace be with you' and 'Grant us thy peace' are written into the most sacred parts of the Christian liturgy, while *shalom* and *salaam* are peaceable greetings in Judaism and Islam respectively. The Hebrew Scriptures themselves are contradictory. There is a tradition of the trickster deploying deception, such as befits ethnic underdogs, and an ideology of legitimate subjugation. At the same time there is a moral reserve about the taking of blood and a positive evaluation of Israel's helplessness (and

dependence on God). This mutates into Isaiah's vision of universal peace and of the righteous rule of God so central to the teaching of Jesus.

In the teaching of Jesus there is a radical reversal of values which includes a rejection of the negative reciprocity embodied in 'an eye for an eye' and the feud. The Kingdom of God inaugurates a peaceful world empire of love in a harmonious universal city, New Jerusalem, which has its foundation in the self-offering of the Savior as he confronts the powers of violence and injustice. Versions of this approach characterize Christianity in the earliest centuries, resulting in conscientious objection to military action, reinforced by refusal to take an idolatrous oath and expectation of a divine foreclosure of human history.

1. Christianity and War After Constantine

However, with Constantine's embrace of Christianity as guarantee of victory, the 'City of God' has to be firmly distinguished from the 'City of Man,' though many Christians came to see the empire as the divinely ordained vehicle of the Christian religion and its triumph. Christianity was now the faith of the majority, including the power holders and senators, not the hope of a minority mostly without access to power. It was implicated in the dynamics of power, even though monks and small groups maintained a devotion to peaceable ideals. Among these small groups were evangelical sects like the Waldensians and dualistic heretics like the Cathars. The church itself iconographically still embodied its original tension with 'the world' of power and violence, set over against icons which celebrate the proximity of the powers that be to the mandates of the Almighty. The church also attempted to adjudicate disputes, and even restrict weaponry, more particularly in the feudal period. It sought to enforce a 'truce of God,' in part to further its own geopolitical ends in confronting the Holy Roman Emperor and reversing the advance of Islam.

The movements of the late Middle Ages, eventually erupting in the Reformation, disseminated the Christian repertoire of themes beyond effective ecclesiastical supervision. There were intimations of collective peaceability in the Lutheran and Calvinist reformations, though these were rapidly checked by the fusion of reformed faith with the autonomy of monarchs and with national sovereignty. There were similar intimations of peace within the natural law traditions of Catholicism. However, the espousal of peace was inevitably a minority option, for example, among Christian humanists such as the Swiss Conrad Grebel, and the anarchopacifist wing of the Radical Reformation, for example, among Anabaptists and Quakers. These minorities came to represent a sectarian encapsulation of Christian radicalism over a wide range

of issues, including corporal and capital punishment, which leaked into the wider society over the next two or three centuries. It was, for example, chiefly among such groups that supporters emerged for the nascent peace societies of the UK and the USA in the early nineteenth century.

2. Proposals for Universal Peace

Ideas of universal peace burgeoned from the fourteenth century on, partly on the model of an imperial Pax Romana, centered on this or that universal monarchy, but also based on projects for an international order. These latter picked up the Christian humanism of Erasmus in his *Complaint of Peace* (1517) and the notions of natural law articulated by Grotius in his *De jure belli ac pacis* (1625). One of the most complex of these projects was the Abbé Saint-Pierre's *Projet pour rendre la Paix perpetuelle* (1713). The last two great peace plans of this period were Kant's *Perpetual Peace* (1795) and Bentham's *Plan for a Universal and Perpetual Peace* (1789).

Here one needs to distinguish types of approach and indicate a partial secularization of the religious vision of peaceable social and international order. A notable secularization was embodied in Machiavelli's statement of Renaissance political realism in *The Prince*, while from rather different viewpoints, Kant fused rational with Christian motifs and Bentham appealed to criteria of human happiness. This mixture of contradictory religious and secular discourses developed into the liberal internationalism of the nineteenth century. According to this tradition, especially as developed in the USA and the UK, there was a potential harmony of interests through the reciprocities established by free trade.

A parallel fusion of religious and political discourses also occurred in the democratic socialist tradition, by contrast with the Marxist variant predominant on the European continent which consigned religion to a realm of fantastic hopes devoid of scientific understanding. For Marxist-inspired movements, as also for anarchosyndicalism, violence was a midwife of a new revolutionary order. Out of the currents of liberalism and socialism emerged hopes of universal human and/or working-class solidarity which were articulated in theories of the social sources of war and the structural alterations needed for peace. However, the wars of 1914–18 and 1939–45 suggested these hopes were chimerical and the theories at best flawed, though capitalism was still widely blamed for international disorder and imperialism.

Out of this iron period grew a new realism about the role of national interest in the genesis of war, which was articulated both in religious and in secular form. The major realist theologian was Reinhold Niebuhr in his *Moral Man and Immoral Society* (1934) and the major exponent of secular realism was the American

Hans Morgenthau. At the same time there also emerged fully fledged empirical enquiries into conflict, including 'peace research.'

3. *Forms of the Peace Sentiment*

Thus over the last two centuries or so the 'peace sentiment' has been part of overlapping causes of a humanitarian and internationalist kind, simultaneously fed by religious and enlightenment sources. There causes have conjoined the gospels to reason, commonsense, and the quest for human rights in a radical critique of war and militarism. The nineteenth century sociologist Herbert Spencer, for example, had roots both in rational and evangelical nonconformity, and identified the alliance of state, priest, and military as a major nexus of conflict. A leftward leaning liberalism attacked the nation state and established religion in the name of a pacific internationalism. This was often referred to as 'pacifism,' meaning it presumed the onus of proof was on the proposers of war and held that the interests of those who stood to gain, for example, the armaments manufacturers, ran contrary to the interests of the majority. Liberal internationalists were not strict 'pacifists' however, because there were times when one had to crusade for liberal principles and for liberation—including that other liberal principle, national liberation.

Other humane causes loosely associated with the peace sentiment were reforms of penal practice, above all corporal and capital punishment and a new attitude towards the animal world, which might include vegetarianism, with its objection to meat, machismo, and blood. To these causes might be conjoined a preference for wholesome food to nourish wholesome bodies and lifestyles. All of these had religious roots, even though one cause need not strictly imply another, so that cornflakes were an outcrop of Adventist teaching without requiring pacifism, whereas the Quaker production of chocolate was clearly linked to it by association. Humanitarians opposed all kinds of brutalization, including male violence in the home and the army. A feminization of the psyche occurred, bred in the Sunday School and in progressive education alongside movements for factory reform and the provision of life-enhancing leisure pursuits, preferably in the context of 'nature.'

The concept of individual conscience took off from religious prescriptions like personal conversion and witness and then mutated into a new sensitivity, including sensitivity towards the effects on other peoples of conquest, imperialism, and exploitation. Thus alongside the nationalistic fervor infiltrating and informing much mass religiosity there ran a counterpoint of moral and political conscientiousness. There were movements for democratic scrutiny of foreign policy and the sale of armaments, for the arbitration of international disputes, the enforcement of interna-

tional law, and eventually the League of Nations and the United Nations. Alongside these ran mostly nonviolent agitations for civil rights and/or national liberation such as those led by Ghandi, Martin Luther King, Nelson Mandela, and Archbishop Tutu.

Of course, given the schooling of populations in patriotic sentiment, dissenting attitudes and conscientiousness usually characterized only a minority in the churches and even more so among people at large. This was above all true in wartime, when dissent from all war or from the claimed objectives of a particular war, could be defined as cowardice or disobedience to lawful authority. Yet conscientious objection still made headway and, at least in the countries of the North Atlantic, Protestant world governments reluctantly accepted that national claims to loyalty were not absolute. This partial tolerance applied to the 'historic peace churches,' such as the Quakers and Mennonites, but also included sizeable groups within the larger Protestant denominations, and it just about extended to secular objections rooted in socialism or humanitarian ideals such as were promoted by Leo Tolstoy.

The social ecology of the 'peace sentiment' over the last quarter millennium is clear enough. It began in the transatlantic world through a combination of religious and rational liberal motifs—Manchester allied to Boston—above all among Protestants, and by degrees affected established churches and the Roman Catholic Church. Given that the Roman Catholic Church was rooted in more organic and communal forms of society, especially in Latin Europe, its defence of the rights of individual conscience was originally weaker. Yet its transnational character always included a sense of the law of nations as nourished by the traditional concept of a just war, such as due proportion of means to ends.

With the development of weapons of mass destruction most mainstream churches expressed reservations about the doctrine of deterrence as it related to the 'just war.' On the one hand there were religious groups allied to secular middle-class radicalism in campaigns against nuclear armament and colonial involvements such as Vietnam, and on the other hand there were specialist commissions into the contemporary nature of war and peacemaking. Thus the Church of England produced a report *The Church and the Bomb* (1982) and the Roman Catholic Church in the USA published an extended critique in 1983.

In summary one may say that there has always been a dialectic relation between the idea of an all-embracing empire (the Roman Pax, Chinese imperial civilization) and the vision of a transnational unity, such as the Islamic 'Ummah' or the Christian *Oekumene*. In times of religious establishment, whether imperial, monarchical, or national, the former will inevitably tend to colonize the latter, so that the wider vision is maintained underground in liturgy, iconography, and movements of protest. However, as

religious groups became separate from the state the visionary repertoire gains more purchase, and is increasingly reinforced by the interdependence of global society. The resources on which the search for peace may draw become increasingly international, so that the message of the gospels, of Isaiah, and Micah mingles with Eastern and especially Gandhian, notions of nonviolent protest and moral suasion. Indeed, as Western protest has drawn on Eastern resources, so have Eastern movements been cross-fertilized by Western ones, not only Christianity but the works of Tolstoy and Thoreau.

See also: Appeasement: Political; Conscription, Policy of; Ethnic Cleansing, History of; Ethnic Conflicts; Genocide: Anthropological Aspects; Genocide: Historical Aspects; Internal Warfare: Civil War, Insurgency, and Regional Conflict; Liberalism and War; Peace; Peace and Nonviolence: Anthropological Aspects; Peace Movements; Political Protest and Civil Disobedience; Religion, History of; Religion: Nationalism and Identity; Religious Fundamentalism: Cultural Concerns; Religious Nationalism and the Secular State: Cultural Concerns; War: Causes and Patterns

Bibliography

- Bainton R H 1960 *Christian Attitudes Toward War and Peace*. Hodder and Stoughton, London
- Beales A C F 1931 *The History of Peace*. G. Bell, London
- Brock P, Socknat T P (eds.) 1999 *Challenge to Mars: Essays on Pacifism from 1918 to 1945*. University of Toronto Press, Toronto
- Burns J P (ed.) 1996 *War and its Discontents: Pacifism and Quietism in the Abrahamic Traditions*. Georgetown University Press, Washington, DC
- Ceadel M 1987 *Thinking About Peace and War*. Oxford University Press, Oxford, UK
- Haleem H A (ed.) 1999 *The Crescent and the Cross*. Macmillan, London
- Martin D 1997 *Does Christianity Cause War?* Oxford University Press, Oxford, UK
- Mayer P (ed.) 1966 *The Pacifist Conscience*. Rupert Hart-Davis, London
- Niditch S 1993 *War in the Hebrew Bible*. Oxford University Press, Oxford, UK
- Sibley M Q, Jacob P E 1952 *Conscription of Conscience: The American State and the Conscientious Objector, 1940–1947*. Cornell University Press, Ithaca, NY

D. Martin

Religion, Phenomenology of

Phenomenology of religion is an academic approach to analyze religion predominantly within religious studies. It is documented in a series of monographs

and handbooks, but it is also defined by a 'scholarly method' that utilizes principles of phenomenological philosophy. For the purpose of this paper, we shall distinguish: (a) the descriptive phenomenology of religion which refers to the classification and systematization of religious phenomena and the creation of typologies which account for different types of religion; and (b) the analytical phenomenology of religion which, in addition to the goals of descriptive phenomenology, is based on some explicit understanding of the philosophical background and methods of phenomenology and hermeneutics developed in the tradition of Edmund Husserl, Wilhelm Dilthey, Martin Heidegger, and Paul Ricoeur.

1. Historical Development, Institutions, and Representatives

Whereas the notion of phenomenology had been proposed by philosophers also preoccupied with religion, such as J. H. Lambert (1721–71), I. Kant (1724–1804), and G. W. F. Hegel (1770–1831), the contemporary notion of phenomenology of religion refers to a specialized field within the academic study of religion. The origins of this field go back to the Belgian historian of religion P. D. Chantepie de La Saussaye (1848–1920). Referring to Hegel's work (which includes the *Phenomenology of Mind*), he divided the science of religion into the history of religion and the philosophy of religion. With respect to the latter, he wanted the phenomenology of religion to not only avoid dogmatics, but to systematically order the main groups of religious phenomena according to the feature of the historical material. The most decisive breakthrough in the field has been due to the work of the Dutch professor of religious history, Gerardus Van der Leeuw who, in 1924, suggested phenomenology of religion to be a special method for the history of religion. Since then, the phenomenology of religion has produced, on the one hand, a series of 'monumental' contributions which provide systematic overviews on a broad variety of religions, or, on the other hand, monographs on typically basic features of single religious phenomena. Therefore, topics of the studies include general issues, such as the distinction between the 'sacred,' and the 'profane,' the 'numinous' or the experience of power, and on the other hand comparative studies on more specific religious phenomena, such as prayer, piety, or religious ceremonies. In the 1950s, phenomenology of religion became increasingly established in the academic studies of the history of religions, so that it became institutionalized as a specialized branch within *Religionswissenschaft*, Comparative Religion or Religious Studies. Although one should not ignore Britain (James 1938) and France (the Rumanian scholar's Eliade's (1949) first important book has been published in French), the main centers of development have been The Netherlands,

Germany, Scandinavia, Italy, and North America (cf. Pettersson and Åkerberg 1981).

2. Approach and Method

Phenomenology of religion differs from theology as well as from other specialized disciplines preoccupied with religion (philosophy, sociology, psychology, etc.) by its history and its method. Historically, phenomenology of religion opposed positivist and evolutionist approaches to religion by integrating historical knowledge of the facts with phenomenological methods. Whereas the sociology of religion, which developed almost simultaneously, focused on the external, public aspects of religion, the phenomenology of religion turned towards the essential aspects and started systematically to classify religious phenomena.

2.1 Descriptive Phenomenology of Religion

Phenomenology of religion draws on the impressive amount of data collected by the history of religion, but also on data collected by sociology, anthropology or the psychology of religion. Therefore, it can be said to be empirical. The basic method of descriptive phenomenology is the comparison of religious phenomena. It assumes that religious phenomena are comparable across different cultures and religions, and that the comparison does not only yield generalizable features but allows us to grasp the general features of religion. In analyzing the empirical forms of historical religions, phenomenology of religion tries to abstract from specific religions, and to then develop ideal types of religious phenomena, such as 'sacrifice,' 'fetish' or 'mysticism,' 'prophecy,' 'priest,' or 'cult.' Therefore, it can be said to be systematical. In order to pursue this task, phenomenology is considered to provide the method for finding the common denominators covering the diversity of empirical data. Religious phenomena include manifestations of the sacred, such as sacred places, mountains, buildings, active realizations of the sacred, such as rituals, prayers, holy words, religious beliefs, such as revelation, redemption, transcendence, and experiences, such as vision, ecstasy, or miracles.

The data on these phenomena are not to be observed from an external point of view, as in most sciences of religion. On the contrary, phenomenology of religion is nonreductionistic in identifying the specifically religious features of religious data. This is the reason for turning to philosophical phenomenology that is expected to analyze the meaning of religion to the person. At the basis of phenomenology of religion lies the analysis of religious experience rather than religious ideas, the beliefs, or morals. The accentuation of the lived experience may be one of the most important contributions of phenomenology to the study of religion.

Basically, religion is the opposite of phenomena: religion posits what it believes to be as existant, whereas phenomenology brackets exactly this belief in order to investigate how it may be constituted by consciousness. Thinking of others, in religion (or parapsychology), gets close to telepathy (think of the dead), in phenomenology, the others are but constructions. Almost simultaneously, J. W. Hauer developed a phenomenology of religion mainly on the basis of data on illiterate cultures. Among the major contributors to phenomenology of religion one has to mention Allen, Bleeker, Bollnow, Hultkrantz, Kristensen, Lanszkowski, Pettazzoni, Wach, Waardenburg, Widengren, etc.

2.2 Phenomenological Methods

Whereas the descriptive and typological branch of the phenomenology of religion remains silent about the methods by which it arrives at results, the phenomenological branch is characterized by additional methodological considerations that are based in philosophical phenomenology.

From the perspective of phenomenological philosophy, experiences are the processes by which consciousness is constituted. As consciousness is always intentional, i.e., consciousness of something, experiences, too, are characterized by intentionality, i.e., the fact that they are related to something else. Also religious experience is experience of something. To give an example: as Chantepie de La Saussaye has already stressed, it is only the meaning of the act that allows us to distinguish between the sacrificial priest and the butcher.

In order to arrive at the meaning of religious experience, Husserl recommended to perform what he called '*epoché*': the bracketing or suspension of all our unexamined assumptions, preconceptions, and prejudices about the phenomena under consideration. In performing the *epoché*, analysts are expected to turn 'to the things themselves,' i.e., to the phenomena as they are experienced. That is to say that phenomena, such as visions, auditions, or miracles, are to be understood according to the ('religious') meanings they have about the experiencing consciousness and not according to, e.g., a neutral scientist's observations. As an additional method, religious phenomena, such as 'snake symbolism,' can be made subject to 'free variation': by mentally varying aspects of this symbol, one attempts to arrive at the essential features of the '*Wesensschau*.' On the basis of these methodological operations, descriptions of the phenomena are considered to yield typical structures, essential meanings or even the '*eidos*' of religious phenomenon.

Attempts to understand the essential meaning of religious phenomena, finally, presuppose hermeneutics, i.e., the method of understanding ('*Verstehen*')

the subjects of religious experience. According to Van der Leeuw (1948), understanding is achieved by empathy: the investigator tries almost to re-experience the religious phenomenon in order to describe these experiences. Understanding, moreover, includes reflecting the context of the interpreters, accounting for the interpretive framework, and explicating their pre-interpretations. To some representants of the phenomenology of religion, such as Eliade, hermeneutics has become a leading principle of their method that allows us to unveil the deeper structure of the experiences of mankind. On these grounds, religious studies have also been equated to a hermeneutic anthropology.

Basically, one can distinguish three related approaches: a descriptive school which is preoccupied with the systematization of religious phenomena; the typological school which tries to distinguish different types of religion, and the phenomenological school in the specific sense of the word that applies methods borrowed from philosophical phenomenology.

3. Criticism

From its beginnings there has been a series of critical objections against the phenomenology of religion.

(a) Especially the phenomenological approach of Eliade is the target of Wagner's (1986) criticism who objects that his concept of the *homo religiosus* is guided by a notion of 'natural religion.' He argues that it presupposes some unwarranted knowledge about the religious situation in what Eliade considers to be 'archaic cultures.'

(b) With respect to religion, phenomenology of religion takes a decidedly substantialist position (Luckmann 1983). Moreover, critics charge that this substantialism is based on nonempirical, extra-phenomenological, and theological assumptions and intentions enter into the analyses of such classic representants of phenomenology of religion such as Kristensen, Eliade, or Van der Leeuw.

(c) To many critiques, this is due to the fact that methods are rarely unveiled. Despite the reference to phenomenological methods, the basic presuppositions often seem arbitrary, and also theoretical reflections are criticized to fall short of the diligent collection and classification of data.

(d) Since phenomenology provides the basic methodology for phenomenology of religion, it is particularly consequential that phenomenology of religion lost contact with the developments in philosophical phenomenology from about the 1950s. As a result, phenomenology of religion rarely is considered of importance in the analysis of religion within the tradition of phenomenological philosophy (Guerrière 1990).

(e) The phenomenology of religion has been criticized for ignoring the social and cultural contexts of

religious phenomena. Moreover, as phenomenology in general has been criticized for its naive attitude towards language and cultural perspectivism, also the phenomenology of religion is subject to criticism as to the linguistic and cultural bias implicit in the analysis of 'phenomena' and 'symbols.' Thus, the results of free variations depend on the phenomenologists' cultural background (Allen 1987).

4. Recent Developments

As a reaction to these criticisms, there has developed what Waardenburg calls the new style of the phenomenology of religion (Waardenburg 1973–1974). According to Waardenburg, this new style of investigation shares a series of assumptions: (a) religious phenomena are characterised by a variety of meanings and intentionalities which are the result of different, often materially based perspectives on the world; (b) religions are coherent, but essentially open systems of signs and symbols, the elements of which are selected according to the corresponding relevance systems; (c) religious phenomena are considered to be human constructions which are subject to social processes (as, e.g., traditionalization, institutionalization, or canonization). At this point, the phenomenology of religion overlaps with positions which have been developed within those branches of the sociology of religion rooted in the phenomenology of Edmund Husserl and Alfred Schutz (Berger and Luckmann 1963). This 'phenomenologically orientated sociology' regards religious experiences as, first, based on the profane transcendences of everyday life, and second, subject to processes of the social construction of multiple realities. By focusing on the communicative processes by which this social construction is accomplished, this approach tries to bridge the gap between empirical research and methodological reflections on this research. Moreover, in admitting that linguistic and cultural background knowledge enters into the phenomenologist's background knowledge, several of the new approaches to the phenomenology of religion accept an orientation towards a 'reflexive phenomenology.'

See also: Hermeneutics, History of; Phenomenology in Sociology; Phenomenology: Philosophical Aspects; Religion: Definition and Explanation; Religion, Sociology of; Religiosity, Sociology of; Schtz, Alfred (1899–1959); Semiotics

Bibliography

- Allen D 1978 *Structure and Creativity in Religion. Hermeneutics in Mircea Eliade's Phenomenology and New Directions*. Mouton, The Hague, Paris, New York

- Allen D 1987 Phenomenology of religion. In: Eliade M (ed.) *The Encyclopedia of Religion*. Macmillan, New York, Vol. 11, pp. 272–85
- Berger P L, Luckmann T 1963 Sociology of religion and sociology of knowledge. *Sociology and Social Research* 47: 61–73
- Bianchi U, Bleeker C, Bausani J A 1972 *Problems and Methods of the History of Religions*. Brill Leiden, Germany
- Bleeker C J 1972 The contribution of the phenomenology of religion to the study of the history of religions. In: Bianchi U, Bleeker C J (eds.) *Problems and Methods in the History of Religions*, Brill, Leiden, pp. 39–41, 51
- Colpe C 1988 Zur Neubegründung einer Phänomenologie der Religionen und der Religion. In: Zinser H (ed.) *Religion-swissenschaft. Eine Einführung*. Reimer, Berlin, pp. 131–54
- Eliade M 1949 *Traité d'histoire des religions*. Paris
- Eliade M 1958 *Patterns in Comparative Religion*. Sheed & Ward, New York
- Guerrière D (ed.) 1990 *Phenomenology of the Truth Proper to Religion*. State University of New York Press, Albany, NY
- Hauer J W 1923 *Die Religionen, ihr Sinn, ihre Wahrheit*. Berlin, Stuttgart, Leipzig, Vol. 1
- Heiler F 1932 *Prayer: A Study in the History and Psychology of Religion*. Oxford University Press, London
- Heiler F 1961 Erscheinungsformen und wesen der religion. Stuttgart 1961. Religionsphänomenologie. In: Müller G et al. (eds.) *Theologische Realenzyklopädie*. De Gruyter, Berlin and New York, Vol. XXVIII
- James E O 1938 *Comparative Religion: An Introductory and Historical Study*. Methuen, London
- Honko L (ed.) 1979 *Science of Religion. Studies in Methodology*. Mouton, The Hague
- Kristensen W B 1960 *The Meaning of Religion: Lectures in the Phenomenology of Religion*. Mouton, The Hague
- Luckmann T 1983 Theories of religion and social change. In: Luckmann T (ed.) *Life-World and Social Realities*. Heinemann Educational Books, London, pp. 133–61
- Pettersson O, Åkerberg H 1981 *Interpreting Religious Phenomena. Studies with Reference to the Phenomenology of Religion*. Almqvist & Wiksell International, Stockholm, Sweden
- Saussaye de la P D 1887 *Chantepie, Lehrbuch der Religionsgeschichte*. Muhr, Tübingen, Germany
- Van der Leeuw G 1948 *Inleiding tot de Phaenomenologie van den Godsdienst*, 2nd edn. Srven F. Bohn, Haarlem
- Waardenburg J J 1973–1974 *Classical Approaches to the Study of Religion: Aims, Methods and Theories of Research*, 2 vols. Mouton, The Hague/Paris
- Wagner F 1986 *Was ist Religion? Studien zu ihrem Begriff und Thema in Geschichte und Gegenwart*. Mohn, Gütersloh

H. Knoblauch

Religion: Psychiatric Aspects

1. Terms

According to *Webster's New World Dictionary*, *psychiatry* is defined as 'the branch of medicine concerned with the study, treatment, and prevention of disorders

of the mind, including psychoses and neuroses, emotional maladjustments, etc.' (p. 1147). In contrast, *psychology* is defined as 'the science dealing with the mind and emotional processes' and 'the sum of the actions, traits, attitudes, thoughts, mental states, etc. of a person or group' (p. 1147). From these definitions, it is clear that the realms of both these fields have considerable overlap in the questions they address about human behavior, and in many cases, knowledge that is garnered in one field is frequently shared and absorbed into the ensuing theoretical development of the other discipline. Hence, for the purposes of this article, which addresses the theoretical underpinnings rather than the actual practices of each profession, the ideas from psychology and psychiatry will be used interchangeably, except where otherwise denoted as specific to their field.

2. The Question

The question about the need for relationship between psychiatry and religion has emerged as a predominantly twentieth-century concern. The movement to inter-relate the two disciplines evolved as participants who believed strongly in the importance of religion in people's lives were concerned by the lack of religious involvement in modern medicine's approach to healing. Many adherents to this movement to intertwine the two fields reject the dualism of body and soul, believing instead that mind and body are inter-related and true healing is best realized when a holistic bio-psycho-social-spiritual approach is utilized.

As such, in recent years a growing number of psychiatrists and other physicians, psychologists, clergy, theologians, and religious philosophers who have not been satisfied with mainstream traditional and institutionalized efforts at healing have sought to combine the best of modern psychiatric medicine with religion/spirituality to develop an alternate method of healing to relieve mental and emotional distress. These practitioners are joined by religiously committed persons who are seeking treatment. At the same time, a number of traditionalists, both psychiatrists and clergy, are critical of incorporating religion and psychiatry into a single treatment mode. They raise the question whether religion and psychiatry should both be considered in aiding individuals, advocating rather that the two remain separate in their approaches to problem solving and actual practice.

3. Defining the Problem

In his book *The Structure of Scientific Revolutions* (1970), Thomas Kuhn defines a paradigm as a nascent model for looking at the universe in a specific manner, and answering questions from within that particular framework's perspective. Furthermore, according to

Kuhn, for growth to occur in science and/or ideology, rather than working toward consensus between paradigms, the premise for philosophical development should be that growth comes from complementary coexistence between paradigms, as each domain converses with the other within its specific applications.

It is arguable that religion and science are two different paradigms for understanding the universe, and humanity's role within it. The difficult and sometimes troublesome relationship between these two fields emerges in part as a result of the two different paradigmatic approaches to answering the same basic questions about humankind. Such questions include, but are not limited to these: what are the fundamental components of human beings? What causes illness? What helps keep individuals healthy?

Furthermore, the more recent developments of specialized disciplines such as psychiatry and psychology in the nineteenth and twentieth centuries also raised alternate perspectives on answering the questions to humanity's interpersonal troubles—answers different from those raised in their antecedent jurisdictions of religion and philosophy (Rechberger 1998).

The question raised is: do these alternate paradigms have communal merit in helping individuals lead happier and healthier lives, or should the division between the two fields of mental health and religion remain distinctly separate?

4. History

Since the 1700s, the period known as the Enlightenment, religion and science have experienced increasing differentiation and tension. Prior to this point in history, religion and science were interchangeable ideologies. In fact, as Murphy (1997) states, 'In the Middle Ages, theology was not only a science but also the queen of the sciences' (p. 155, italics in the original). Indeed, under the church's rule, when sixteenth and seventeenth century scientists such as Giordano Bruno, Tycho Brahe, and Galileo Gallilei offered data for understanding the universe different from the understood Biblical theology of the day, they were declared heretics, punished, and a few were even executed. However, the mathematical evidence presented by scientists such as Copernicus and Galileo was repeatedly validated and supported despite dispute from the church. Hence, for many, science became the primary source of information about the universe and its inhabitants. Furthermore, due to the infamy surrounding the church's declaration of heresy upon scientists such as Bruno and Galileo, religion was at least somewhat discredited as a viable source of answers to humanity's problems (Rechberger 1998).

Nevertheless, it is impossible to discount the effect religion has had on humanity throughout history.

From the earliest times, religious beliefs and rituals have been a central component of the lives of humankind, frequently providing the structural framework from which societal, cultural, and moral norms developed. In the Western hemisphere, the primary religions of Judaism, Christianity, and Islam have as a core belief a universal God who acts as creator, sustainer, and protector. In the Eastern hemisphere, the predominant religions of Buddhism and Hinduism depict God as less actively involved in the day-to-day workings of humanity; nevertheless, almost all world religions have similar beliefs and practices in a supreme being or group of beings that are in some way interested in humanity's development. Furthermore, throughout most of known history, it was the religious leaders—whether priest or shaman—who were considered healers (a tradition which continues today in less Westernized cultures). Indeed, prior to the 1800s, what scant treatment there was for mental illness was provided by the church. Unfortunately, one of the negative outcomes of the overlapping domains of religion and medicine was evidenced in the development of concepts such as demon possession and illness as a result of, if not punishment for, an individual's misdeeds. These explanations were broadly accepted as disease etiology in most Western nations well into the nineteenth century.

In contrast, however, one of the potential benefits of having the same person exercising both religious and healing functions allows for the possibility of positive coordination and integration between the two domains. Reformers such as Philippe Pinel (1745–1826) in France began to see the mentally ill as sick rather than sinful or wicked. Another reformer, William Tuke (1732–1822), opened the York Retreat in England in 1792, and other pious Quakers were early leaders in the care and treatment of the mentally ill.

Beginning in the latter part of the eighteenth century medicine was developing as an independent science dealing with the maintenance of health and the prevention, alleviation, and cure of disease. This was a clear departure from the intertwining of religious practice and medical intervention that had existed from earliest human history. Until this time it was the church that regulated medical practice by the very fact that doctors needed license from the bishop to practice medicine (Foskett, 1996).

Despite this emerging demarcation between religion and medicine as separate disciplines, some religious leaders continued an interest in maintaining the overlapping interests of religion and medical practice. This was particularly notable in the United States during the mid-nineteenth century, during a *zeitgeist* that advanced an increasing emphasis on the mind-body relationship. A strong—albeit minority—view of the positive benefits of intermingling religious belief and medicine began to develop. Many wrote about it, including Mary Baker Eddy (1821–1910), the founder

of Christian Science, and Ellen G. White (1827–1915), one of the founders of the Seventh-day Adventist church.

Eddy wrote, ‘If we understood the control of mind over body, we should put no faith in material means. Science not only reveals the origin of all disease as mental, but it also declares that all disease is cured by divine mind’ (1875, p. 169). Nearly 100 years prior to psychiatry’s recognition of the influence of brain chemistry on disease, Ellen G. White wrote that ‘the mind affects health to a far greater degree than many realize and many diseases from which men suffer are the result of anxiety, remorse, guilt and mental depression’ (1905, p. 241). These religious writers preceded Herbert Benson and Bernie S. Siegel’s twentieth century emphasis on the mind–body connection.

5. The Emergence of Mental Health Disciplines

While the trend of separating science and religion was evidencing itself by the growing separation between medical practice and religious ideology, the emergence of psychology and psychiatry as new scientific disciplines also demonstrated a distinct separation from their religious and philosophical roots. Examining the writings of the earliest primogenitors within these newly emerging disciplines clearly reveals the tension between science and religion as they developed differing paradigms to resolve questions about what the components of ‘good’ mental health comprise.

In the United States, it was pre-Freudian William James (1842–1910) who was at the forefront development of the profession of psychology, along with his Harvard colleague Hugo Munsterberg. Interestingly, James (see article on *James, William* (1842–1910)) refused to treat physiology, psychology, and philosophy as distinct and separate disciplines. Furthermore, his endeavor to maintain a balanced position between these domains is clearly reflected in his 1902 book, *The Varieties of Religious Experience*, which addresses the questions related to religion as it interacts with mental health. As James wrote:

The problem I have set myself is a hard one: *first*, to defend ... ‘experience’ against ‘philosophy’ as being the real backbone of the world’s religious life ... and *second*, to make the hearer or reader believe what I myself invincibly do believe, that although all the special manifestations of religion may have been absurd (I mean its creeds and theories) yet the life of it as a whole is mankind’s most important function’ (1902, p. xix, italics in the original).

Although James acknowledged religion as a social and cultural phenomenon, he was interested primarily in the individual response of a person to religious experience. More specifically, he expressed fascination

with the conversion process, the moral and mystical elements of the religious life, and the interaction between religious belief and illness—particularly healthy vs. pathological mental health responses to religion. James (see article on *James, William* (1842–1910)) espoused a favorable view of religion, based on his perception that ‘healthy-minded’ individuals respond positively to religious beliefs and are the happier for it. Yet this positive view of religion in the scientific community would shortly be denounced by the writings of psychiatrist Sigmund Freud (1856–1939).

The field of psychiatry, although historically a part of medicine, was marginalized as a profession until the appearance of Freud’s writings (Alexander and Selesnick 1966). His ideas about the dynamic structure of personality and the benefits of psychoanalysis revolutionized the field of psychiatry. In his radical new perspectives about the inner processes of individuals, Freud assessed humanity’s desire for God and the function of religion in everyday life as evidence of psychopathology.

In his book *The Future of an Illusion*, Freud (see article on *Freud, Sigmund* (1856–1939)) opined that faith is based on the illusion that an idealized father (God) can replace the lost, real father who has proven fallible to the child and is no longer a comfort and source of security. Freud’s solution to his depiction of this ‘infantile’ response on the part of individuals to create a ‘mythological’ God was for them to develop a realistic view of the reality that there is no God: ‘Man cannot remain a child forever; he must venture at last into the hostile world. This may be called *‘education to reality’*’ (1934, p. 86, italics in the original).

Although the idea that God is a projected illusion did not originate with Freud, his declaration that religious belief is an infantile projective fantasy cast a long shadow over subsequent practitioners’ theories about the relationship between religious belief and mental health. This perspective contributed to many decades in which psychiatry generally viewed religious belief with skepticism and antipathy. What little credit Freud gave to religion for ‘restraining the asocial instincts’ of humankind (p. 65), he reiterated that it had not done so successfully. Instead, Freud advocated that humans should place their trust in the rigors of the scientific method, out of which discoveries of the truth about the nature of individuals and humanity’s relationship to the world around it would emerge.

This review of the origins of psychology and psychiatry illuminates the tension between the positivist assumption—that the scientific method and rigors of research would eventually answer all of humanity’s questions about physical and emotional well being—and the alternate view that religious belief is a constructive addition to solving mental health concerns. Although some theorists continued to typify any religious belief as evidence of naivete or pathology, others dissented from Freud’s strict atheism and

instead posited a more favorable view of religious beliefs in their psychological theories.

Examples of some of the psychologists and psychiatrists who integrated religion with their psychological theories would include such notables as Carl Jung (1875–1961), Erich Fromm (1900–80), and Viktor Frankl (1905–97). The broadest incorporation of the spiritual with the psychiatric is found in Jung, a Swiss psychiatrist and contemporary of Freud. His worldwide travels, interest in the occult, tribal religion, and other non-mainstream mystical beliefs heavily influenced his subsequent psychological theory. In stark contrast to Freud, Jung's liberal premise that all psychic products—including visions, dreams, and hallucinations—are 'facts' and should be considered as having the same basis in reality as other physical facts, opened wide the door to religious belief and behavior in psychic functioning that Freud (see article on *Freud, Sigmund* (1856–1939)) had firmly attempted to close. Furthermore, according to Jung, it is not the business of psychology or psychiatry to prove or disprove the existence of God: it can only address the psychic effects of the God image and its mythic antecedents on human behavior.

Psychoanalyst Erich Fromm recognized the value of religion in providing healthy feelings of stability and security for individuals. In his book *Psychoanalysis and Religion*, Fromm stated, 'Man is not free to choose between having or not having ideals but he is free to choose between different kinds of ideals, between being devoted to the worship of power and destruction and being devoted to reason and love' (1950, p. 25). Fromm further advocated a surrendering of one's individual goals and identity in favor of living selflessly in community with others, a not uncommon ideal of both Judaic and Christian rhetoric.

Finally, Viktor Frankl, the Holocaust survivor and psychiatrist who wrote *Man's Search for Meaning*, articulated in his book the premise that all humans are on an unmitigated quest for meaning—even in the midst of horror and despair. Frankl clearly credited his spiritual belief as sustaining his mental health throughout his imprisonment, and contrasted his experience to other prisoners who lost their faith and subsequently their will to live, succumbed to illness, and died. Of the three highlighted theorists who have acknowledged the positive merits of religious belief on mental health, Frankl's testimony is the most graphic and recent representation of the positive outcome of melding the two.

6. Theology

It is not only within the realm of science that the questions about the potential benefits of intertwining psychiatry and religious belief have been addressed. A number of theologians, religious philosophers, and clergy have also debated the merits of incorporating

religious beliefs into the best scientific practices in mental health.

One of the earliest proponents of such an idea was Anton Boisen (1876–1965), a Protestant clergyman who had once experienced a psychotic illness and subsequently had been institutionalized in a mental hospital. He reported that the hospitalization was a religious renewal experience for him, and he believed that it could be that way for some other patients if the psychiatric community were sufficiently sensitive to the meaning of these experiences.

Another theologian, Paul Tillich (1886–1965), proffered a broader anthropological and philosophical context within which psychiatry could exist and still address its own, more specific concerns. Similar to Jung's placing limits on psychiatry's role in evaluating religious belief, Tillich opined that statements of ultimate concern belong to a domain beyond which psychoanalysis and the scientific method can neither disprove nor verify. Furthermore, Tillich reiterated that when psychiatry makes either negative or positive judgments about the reality of religious ideas and symbols, it has moved out of the realm of science and into the realm of metaphysics.

Current philosopher of religion Paul Ricoeur (1913–) has countered the reductive view of religion espoused by Freud, and the influence Freud's views have had over much of psychiatry and the wider culture shaped by these theories. In contrast to Freud's view that religion is basically a public expression of obsessive compulsive neurosis, Ricoeur espouses belief that humans grow when confronted by sacred figures who challenge them to become more than they are.

Other contemporary writers include Karl Menninger (1893–1990), who has written *Whatever Became of Sin?*, addressing the atheistic stance of psychiatry and the cultural disregard for personal responsibility. Another, Paul Pruyser (1916–87), writes about the need for psychiatry, with its empirically-based findings, and religion with its faith basis, to be utilized separately rather than attempt their integration in understanding and treating mental illness.

7. Practices Which Have Attempted to Merge Religion and Psychology

The concern by some psychiatrists and theologians to interweave the religious with the psychological led to the development of new strains of practice in both disciplines. The most broadly defined psychiatric and religious blend emerged as 'transpersonal psychiatry,' the outgrowth of Jungian mysticism combined with tenets of Eastern religion (such as Buddhism, Hinduism, and Taoism). Proponents of transpersonal psychiatry use drug-free methods familiar in spiritual practices, i.e., meditation, as part of the ritual for maintaining optimal mental health.

A distinctly different approach to the integration of religion and mental health has been seen in the emergence of 'Christian Psychiatry' in the 1970s and 1980s. Led by theologically conservative psychiatrists, this approach was specifically tailored for Christian patients whose treatment mode, psychiatrists believed, could be enhanced by incorporating elements of their Christian faith in their treatment. 'Christian psychiatrists' based their treatment on Biblical principles, including Bible study and prayer.

Also within the theological realm, the development of Clinical Pastoral Education (CPE) emerged as clergy became educated about the nature of psychological problems. Originally developing with the work of Anton Boisen in 1925, CPE emphasizes ministering to the patient by utilizing both theological tenets and psychodynamic theory. CPE subsequently has become an important part of seminary training: a requirement of some, and an elective in even more, to best meet the needs of pastors and chaplains in hospitals and other settings.

8. *Psychiatry Addresses the Issue of Religion*

More than 50 years of writing on the positive interaction between religion and psychological health passed before the field of psychiatry officially re-examined its Freudian views on religious belief. In 1956, R. Finley Gale, Jr., the incoming president of the American Psychiatric Association, devoted his entire presidential address to the 'Conflict and Cooperation between Psychiatry and Religion.' In that speech, Gale called for the development of a committee on psychiatry and religion within the organization, and further suggested that mental health chaplains be invited to join and participate in that venture.

Although the next several decades indicated some gains in psychiatric acceptance of religion as non-pathological, diagnoses as late as those found in the 1987 *Diagnostic and Statistical Manual, Third Edition*, Revised indicated religious belief as a factor indicated in poorer mental health (Larson and Greenwold 1997). However, the 1990 adoption of the *Guidelines Regarding Possible Conflict between a Psychiatrist's Religious Commitments and Psychiatric Practice* indicated that 'patient religious beliefs were being professionally recognized as not only acceptable, but important treatment factors' (Larson and Greenwold 1997, p. 22).

These guidelines highlighted the usefulness of obtaining information about a patient's religious or ideological orientation and beliefs, as well as of addressing the procedure for handling any conflicts between the patient's and the psychiatrist's belief systems. Psychiatrists were reminded that they should not impose their own religious, anti-religious, or other ideological systems or beliefs on vulnerable patients. Furthermore, these guidelines indicated that inter-

pretations concerning a patient's beliefs should be made in a context of empathetic respect for their value and meaning to the patient.

9. *Reflections on the Popular View*

Despite the defamation of religion by the scientific community, demographic studies indicate that those who define themselves as religious/spiritual in the population at large have remained relatively consistent over the past several decades. The first Gallup poll, conducted in 1937, indicated that 73 percent of Americans were members of a church or synagogue. A 1997 Gallup poll (March 24–26, 1997) reported that 96 percent of Americans believe in God or a universal spirit, with 67 percent indicating membership in a church or synagogue. Further, 61 percent indicate religion is 'very important' in their own life (with an additional 27 percent stating it is 'fairly important'), with 43 percent attending religious services weekly or nearly every week.

In addition, one cannot overlook the host of religious-secular self-help support groups in mainstream culture today, particularly Alcoholics Anonymous (AA), the most therapeutically successful of them all. Alcoholics Anonymous (along with its subsequent offspring such as Narcotics Anonymous, etc.) is considered the most useful organization in the United States to assist persons suffering from alcohol abuse and chemical dependence. AA operates much like a religious community, complete with doctrine. The 12 steps proposed as necessary to overcome alcoholism reflect a creed that shapes AA, and has strong religious overtones. A similar group, the Samaritans, operates primarily in the United Kingdom. Its 23,500 volunteers make themselves available by telephone 24 hours a day, by personal visits, and by letters.

10. *Recent Academic Research Supports the Benefits of Religion*

Finally, recent research studies on the interaction between religion and mental health indicate more positive outcomes for believers receiving mental health treatment than for agnostics or atheists. Supporting the anecdotal success stories from Alcoholics Anonymous, a study by Desmond and Maddux (1981) demonstrated that 45 percent of participants in a religious treatment program for opioid dependence were still sober a year later, in stark contrast to the mere 5 percent who had participated in a nonreligious treatment program.

More recently, Koenig and Larson (1998) indicated that in the last 15 years, over 50 epidemiological studies have been published which demonstrate that 'religious persons experience not only greater well-

being and higher life satisfaction, but less anxiety and depression, lower rates of suicide, less alcoholism, less loneliness, and better adaptation to stress than do persons without spiritual resources' (p. 385). Furthermore, a substantive body of research on the acts of religious belief, such as prayer, has demonstrated positive impact on both physical and mental health.

In the preface to his 1993 book *Healing Words*, Larry Dossey, M.D., originally a skeptic, writes about finding 'over one hundred experiments exhibiting the criteria of "good science," many conducted under stringent laboratory conditions, over half of which showed that prayer brings about significant changes in a variety of living beings' (p. xv). Research indicated that patients who became prayer targets experienced altered blood pressure, slowed growth rates of cancerous cells, dissipation of tumors, an increased rate of healing for wounds, and even altered probability for heart attacks.

Finally, in terms of practical and economic outcomes, research indicates that the spiritually committed are less likely to develop serious illness, spend less time in the hospital, and do better upon discharge than those patients who are nonbelievers or aspiritual (Koenig and Larson 1998).

11. Implications for Psychiatry Today

When research indicates that participation in religious activities correlates with a patient's strengthened immune system, a reduced risk (by 60 percent) of arteriosclerotic heart disease and pulmonary emphysema, an increased (between seven and fourteen years) life span (Koenig and Sloan 2000), as well as a decreased likelihood of depression, anxiety, and persistent mental illness, the link between religious belief and improved health both physically and mentally becomes noteworthy. However, some concern remains about the potential negatives of physicians and mental health practitioners encouraging religious belief in their patients.

Richard P. Sloan, Ph.D., director of the Behavioral Medicine program at Columbia Presbyterian Medical Center in New York, opines that when 'physicians depart from their medical expertise and enter the religious realm, there's an element of coercion because patients don't want to offend or disappoint their doctors' (Koenig and Sloan 2000, p. 159). Sloan believes that melding the two realms puts a physician 'in a terrible bind. If he believes that there is strong evidence associating religion with better health outcomes... he would be derelict in his duty as a physician if he didn't advocate religion' (p. 156). He further associates this advocating of religion to prescribing antibiotics: a physician doesn't ask whether the patient is in favor of taking them, but merely prescribes them as being the best course of treatment.

Sloan's concerns about the potential dangers from intertwining religious ideology and mental health are worth taking into account. History has clearly demonstrated problematic responses to the mingling of religion and science, and an uncritical absorption of religious ideology into psychiatric theory would be foolhardy. However, given the recent research that indicates the positive benefits of religious belief in better mental health outcomes, ignoring those data would be very poor science. It would seem that the best approach would acknowledge the inherent tension between the two paradigms of science and religion, with growth in both fields occurring as the result of conversant dialogue between the two, much as Kuhn (1970) advocated in his treatise on scientific revolution.

12. Future Considerations

Given the recent scientific evidence for positive outcomes for religious belief, combined with the statistics about religious believers in the population at large, a move toward educating psychiatrists about the interaction between religious belief and mental health practice appeared necessary. Hence, in 1995 the Accreditation Council for Graduate Medical Education (ACGME) responded with a requirement to include religion and spiritual factors in the human development curriculum. To assist program directors in meeting this requirement, a *Model Curriculum for Psychiatric Residency Training Programs* was developed by David B. Larson, Francis G. Lu, J. Swyers, and Elizabeth Bowman, among others. This model curriculum provided definitions, bibliography, and teaching options to fit individual programs.

Practical considerations aside, the intermingling of religion and science can remind psychiatry that it does not practice in a vacuum or without social, ethical, and public philosophical influences. Religion confronts psychiatry with decisions regarding the proper stance of a secular, therapeutic discipline on determining components of mental health. Because it serves a wide spectrum of humanity, psychiatry should not wed itself to a specific religious or cosmological creed. It should, however, consider a public philosophy that provides 'a broad, intellectual understanding of both its proper focus as well as how it should relate to the wider aspects of social and cultural action' (Browning et al. 1990, p. 20). Such an academic stance allows psychiatrists to take their patients' experiences and beliefs seriously and in a respectful manner.

Psychiatry is on its most solid ground when it is investigating the psychological meaning and motivation of the religious beliefs of any patient. To reiterate the concerns of James and Jung, what psychiatrists must not comment upon is whether there is an ultimate truth justification for religious faith, for 'God' is, logically speaking, an axiom and not an empirically testable scientific hypothesis. Rather, acknowledging

the commensurate questions of the two paradigms (What are the fundamental components of human beings? What causes illness? What helps keep individuals healthy?) and allowing each field to share knowledge with the other as a means of improving practice in both disciplines, appears to be the most prudent course. Complementary coexistence, while maintaining studied attention to the polarizing tendencies of both religion and science, is the most likely path to providing optimal care for patients.

See also: Freud, Sigmund (1856–1939); James, William (1842–1910)

Bibliography

- Alexander F, Selesnick S T 1966 *The History of Psychiatry*. Harper & Row, New York
- Bhugra J (ed.) 1996 *Psychiatry and Religion: Context, Consensus and Controversies*. Routledge, London
- Blazer D 1998 *Freud vs. God: How Psychiatry Lost Its Soul & Christianity Lost Its Mind*. Intervarsity Press, Downers Grove, IL
- Browning D, Jobe T, Evison I S 1990 *Religious and Ethical Factors in Psychiatric Practice*. Nelson Hall, Chicago
- Desmond D P, Maddux J F 1981 Religious programs and careers of chronic heroin users. *American Journal of Drug and Alcohol Abuse* 8(1): 71–83
- Dossey L 1993 *Healing Words: The Power of Prayer and the Practice of Medicine*, 1st edn. Harper, San Francisco
- Eddy M B 1875 *Science and Health with Key to the Scriptures*. Published by the Trustees under the Will of Mary Baker G. Eddy, Authorized Literature of the First Church of Christ, Scientist, Boston, MA
- Foskett J 1996 Christianity and psychiatry. In: Bhugra J (ed.) *Psychiatry and Religion: Context, Consensus, and Controversies*. Routledge, London
- Frankl V E 1987 *Man's Search for Meaning*, 5th edn. Hodder and Stoughton, London
- Freud S 1934 *The Future of an Illusion*. In: Jones E (ed.) *The International Psycho-Analytic Library, No. 15* (Trans. W D Robson-Scott). Hogarth Press, London
- Fromm E 1950 *Psychoanalysis and Religion*. Yale University Press, New Haven, CT
- James W 1902 *The Varieties of Religious Experience*. Longmans, Green and Co., Random House, New York
- Koenig H G, Larson D B 1998 Religion and mental health. In: Friedman H S (ed.) *Encyclopedia of Mental Health*. Academic Press, San Diego, Vol. 3, pp. 381–92
- Koenig H G, Sloan R P 2000 Should doctors prescribe religion? *Medical Economics* 10: 145–59
- Kuhn T S 1970 The structure of scientific revolutions. In: Neurath O, Carnap R, Morris C (eds.) *International Encyclopedia of Unified Science*, 2nd edn. University of Chicago Press, Chicago, Vol. 2(2)
- Larson D B, Greenwold M M 1997 Making the case for spiritual interventions in clinical practice. *Mind/Body Medicine* 2(1): 20–30
- Larson D B, Lu F, Swyers J (eds.) 1996 *Model Curriculum for Psychiatric Residency Training Programs: Religion and Spirituality in Clinical Practice*. National Institute for Health Care Research, Rockville, MD

- Mengh H, Freud E (eds.) 1963 *Psychoanalysis and Faith: The Letters of Sigmund Freud and Oskar Pfister*. Basic Books, New York
- Menninger K A 1973 *Whatever Became of Sin?* Hawthorne, New York
- Murphy N 1997 *Anglo-American Postmodernity: Philosophical Perspectives on Science, Religion, and Ethics*. Westview Press, Boulder, CO
- Rechberger E 1998 *The Impact of Einsteinian Relativity and Quantum Physics Theories on Conceptualizations of the Self in Psychology*. Unpublished dissertation. Under publication review
- Tillich P 1957 *Dynamics of Faith*, 1st edn. Harper and Row, New York
- White E G 1905 *Ministry of Healing*. Pacific Press, Mountain View, CA

G. T. Harding, IV

Religion, Psychology of

The psychology of religion was a central part of psychology at its inception as an independent empirical discipline at the end of the nineteenth century. Psychology is also one of the family of human sciences that has made an important contribution to the study of religion. Paloutzian (1996) provides an accessible introduction. This article will look first at some key historical figures, then at contemporary research contributions, and finally at critical current issues.

1. History

Early work on the psychology of religion included quantitative methods, such as the questionnaire research of Edwin Starbuck on types of conversion. However, the most durable masterpiece of the early period was undoubtedly the qualitative research of William James. The psychology of religion continued to be a vigorous part of the discipline of psychology through the first quarter of the twentieth century, and the most widely read summation of that period was J. B. Pratt's, *The Religious Consciousness* (see Wulff 1997).

From about 1925, there followed a fallow period in the psychology of religion, in part reflecting the behaviorist direction taken by the discipline generally. However, with the growing emancipation of psychology that took place in the 1960s, there has been a revival in the psychology of religion. This recent period has not seen any major psychological theories of religion, indeed many would doubt whether they are appropriate. However, there has been a wealth of investigations in many aspects of religious experience and behavior.

1.1 William James

For his *Varieties of Religious Experience*, written at the turn of the century, James (1985) assembled a large quantity of autobiographical material about religious experience. Several of his conceptualizations of this material have become permanent contributions to the psychology of religion, such as his distinctions between 'once-born' and 'twice-born' conversion types, and between 'sick-soul' and 'healthy-soul' types. His description of mystical experience as having the hallmarks of ineffability, a noetic quality, transiency, and passivity has also stood the test of time.

However, James' approach has been criticized (e.g., Proudfoot 1985) for his overemphasis on the private nature of religious experience as 'the feelings, acts and experiences of individual men in their solitude.' Whereas James saw religious doctrines and church practices as arising out of individual experience, others have emphasized how much individual experiences are shaped by the religious tradition. These criticisms of James have been reinforced in recent years by the growing awareness of the extent to which experience is 'specially constructed.' There has also been concern about James' over-reliance on vivid but unusual experiences; others would want to place the emphasis on the religious experiences of ordinary religious people. James also focused on the special qualities of religious experience.

In contrast, some would want to emphasize that the hallmark of religious experience is not its unusual phenomenal quality, but simply the way in which a religious frame of interpretation is applied to experiences of an ordinary kind.

1.2 Sigmund Freud

Another rather separate strand in the psychology of religion that goes back to the early twentieth century is represented by the contribution of Freud, and the psychoanalytic tradition stemming from him. Freud published several books on religion, but his most enduring contribution has probably been *The Future of an Illusion* (1962). There, he argues that religion is a form of wish fulfillment, and that God is a projection of the human mind. These were not altogether new ideas, and a similar line had been taken by Feuerbach. However, the fact that Freud approached them from a well-articulated general theory of mental functions gave them a new authority.

Though Freud's views have been very influential, there has been a growing recognition that they are highly speculative, and less authoritative than was initially supposed. It has also become clear that an alternative, more sympathetic view of religion can be developed within the framework of psychoanalytic theory, something that Freud conceded might be possible. The main influence on recent work of this kind has been Winnicott's concept of the 'transitional'

world, that was neither wholly internal nor external, but an illusionistic world that plays a constructive role in adaptation (see Meissner 1984). Though Freud's own work on religion was surprisingly weak in case studies, others have now conducted rich case studies documenting the close parallelism that can often be found between relationships to parents and to God. Of course, this need not be interpreted as showing that God is nothing but a projection of the mind, just that concepts of God are often shaped by personal factors.

1.3 C. G. Jung

Jung took an unusual interest in religious symbols and doctrines. Indeed, one of the reasons for his break from Freud was that he was unhappy at what he saw as Freud's disparaging treatment of religion. Jung interpreted religious symbols in relation to the personal journey of 'individuation,' a journey from ego-centeredness towards the 'Self,' involving the integration of the unconscious with the ego. Symbols are important both in representing the goal of this journey, and also in facilitating it. Central to Jung's approach to religious symbols is his theory of the 'collective unconscious' and its archetypes; Jung does not regard religious symbols as just being features of the personal unconscious.

Jung had a strong interest in symbols of the Self, and saw God and Christ as being such symbols. However, he laments the extent to which religious symbols have now lost their psychological power and are no longer felt to be relevant to the journey of individuation. His work on Christianity included extended essays on the symbolism of the Trinity and the Mass (see Jung 1969). He also took an extensive interest in Eastern religion, where he also found symbols of individuation, though he was opposed to Westerners taking up Eastern practices such as yoga.

Jung's view of religion is markedly more sympathetic than Freud's, though his reformulation of religion is at points decidedly heterodox. He tended to broaden religion to include anything 'numinous.' Also he emphasized that as a psychologist he could concern himself, not with religious metaphysics, but only with psychological realities (see Palmer 1997). Though Jung saw 'God' and 'Christ' as psychic images and symbols of individuation, whether there was a 'metaphysical' God was something about which he had no official view. However, it would be a misreading of Jung to imagine that he thought people could work out their own salvation. He thought the ego was quite unable to secure 'individuation,' and emphasized the radical otherness of the 'Self,' of which God is a symbol.

2. The Contemporary Psychology of Religion

Much contemporary psychology of religion employs empirical research methods, and there follows a brief

review of representative research findings. The fullest guide to the empirical psychology of religion is that of Hood et al. (1996). Beit-Hallahmi and Argyle (1997) is briefer, but also helpful.

2.1 Individual Differences

The study of individual differences has long been an important part of the psychology of religion. The search for a particular kind of personality associated with religious allegiance has generally not been successful, largely because of the marked diversity of religious people. However, there is reliable evidence supporting an association between different personality types and different approaches to religion. This emerged first in connection with social prejudice, where 'intrinsic' religious people were found to be less prejudiced on average than extrinsic; the same is true of committed vs. consensual religious people (another set of terms for a related distinction). There has also recently been considerable interest in the relation of Jung's psychological typology, as measured by instruments such as the Myers-Briggs Indicator, to various aspects of religious life such as prayer and Church leadership.

2.2 Mental Health

Mental health (at least as measured by self-report instruments) has been found to show the same relationship as social prejudice to types of religiousness, with intrinsic religious people having good mental health, but extrinsic religious people having relatively poor mental health. Even where religious people are not subdivided, religiousness tends to be associated with better than average scores on dimensions such as Eysenckian 'psychoticism.' When more objective measures are considered, religious people are often found to have lower levels of formal mental illness and abuse of drugs and alcohol. There is also growing evidence that religion has a positive relationship to happiness, coping, and physical health. There are probably a variety of explanations of the relationship between religion and health. These include 'selection' factors, such as alcoholics being less likely to go to church. However, there may be various ways in which religion can improve health and happiness, such as the therapeutic value of prayer, and the social support provided by the churches (see Loewenthal 1995).

2.3 Developmental Psychology

Children's intellectual understanding of religion has been found to follow the usual path, documented by Piaget for many areas of cognitive development, from concrete to abstract modes of understanding.

Early research showed that this applied to the interpretation of Bible stories, but the same general pattern seems to apply to all aspects of religious understanding, such as prayer and sacraments. There have been attempts to extend such 'stages of development' models to other aspects of religion, and to take them into adulthood. The best-known work of this kind is Fowler's theory of 'faith development.' However, many critics have expressed caution about the proper limits of stage models of religious development. They seem to work best for development in the intellectual understanding of religion, less well for other aspects; they also only seem to work better for religious development in children than in adults. The key issue is whether religiousness changes in a sufficiently predictable and universal way for the concept of development to be appropriate.

2.4 Religious Experience

There has been extensive empirical research on mystical and religious experience, though it is beset by the problem of how such experience should be defined. The initial question used in the research of Hay (1987) is characteristic: 'Have you ever been aware of, or influenced by, a presence or power, whether you call it God or not, which is different from your everyday self?' Most empirical surveys find roughly a third of the population respond positively to such questions. Interestingly, the proportion is not much reduced for atheists and agnostics. The experiences often occur at times of stress, and generate feelings of peace and happiness. One of the most debated matters is whether there is a common core to such experiences that is independent of religious tradition and culture. Psychometric research by Hood suggests that there are two groups of descriptors, one focusing on the experience of unity, the other set of descriptors more influenced by religious background.

2.5 Neuropsychology

The study of the brain processes involved in religion promises to become a lively research area. There has long been evidence that meditation alters people's levels of arousal, though probably not in any unique or distinctive way when compared to other procedures such as relaxation. There has been considerable interest in a possible link between the brain processes involved in religious experience and temporal lobe epilepsy. Though there are similarities in the two sets of experiences, there are also marked dissimilarities, and critics have pointed out that evidence that the same brain mechanisms are involved is weak (Jeeves 1997). d'Aquili and Newberg (1998) have proposed another neural theory in which two cognitive oper-

ators, the 'causal operator' and the 'holistic operator' underpin different aspects of religion. As evidence accumulates about the brain processes involved in religion, there is likely to be debate about whether or not they offer a complete explanation of religion.

2.6 Attitudes

The measurement of religiousness is an aspect of the study of attitudes and involves the vexed question of the relationship between self-report and behavioral measures. It has been claimed by some that religious believing currently exceeds belonging. Though, on some measures, this appears to be the case, it depends on exactly what measures of each are used. One of the methodological principles to have emerged in the general literature on attitude measurement is that single-item measures of expressed attitudes are a poor predictor of behavior, whereas fuller self-report measures predict better, and there is reason to think that this applies to religious attitudes. Attitude change can be studied in relation to religious conversion.

2.7 Conversion

Conversion has long been a central topic in the psychology of religion. It is a multifaceted, and by no means uniform phenomenon. There are well-established differences between sudden and gradual conversion types, with the former being more conservative in religion and in other matters. It has also long been noticed that adolescence is the peak time for conversion, but the interpretation of this needs to take account of the fact that it is probably the peak time for apostasy as well. Conversion often does not represent as complete or permanent a change as is supposed. Also, many of those experiencing a conversion are returning to previously held attitudes; conversions among those with no previous religious affiliation are relatively rare. In recent years there has been particular interest in conversions to new religious movements. In fact, these often seem to be associated with improvements in the sense of well being. There is some evidence of mental health problems in those who leave new religious movements, though this no means always occurs.

2.8 Charismatic Religion

Speaking in tongues (glossolalia) is much the best-investigated charismatic phenomenon (Malony and Lovekin 1985). Psycholinguistic research generally indicates that it is to be regarded as some kind of 'ecstatic utterance' rather than the use of an identifiable language. Social learning is apparently required

before people are able to speak in tongues; this is suggested by the way in which people first watch it before joining in, and also get more proficient at it with practice. It has been suggested that it involves some kind of dissociated state of consciousness, perhaps a trance state or a regressive state, but there is as yet no clear support for such hypotheses. It has been difficult to identify personality characteristics of those who speak in tongues. This may perhaps be because it occurs in different ways in different settings, but there is as yet no generally accepted classification. However, there is quite good evidence that it is associated with positive personality changes, though it is not unique in having such beneficial effects.

3. General Issues

The psychology of religion involves a number of general issues about methodology and interpretation, including the following.

3.1 Relation to General Psychology

The psychology of religion currently lacks integrative conceptual frameworks and general theories. A psychology of religion that is no more than a fragmented study of various particular religious phenomena will never look impressive scientifically. The other pitfall to be avoided is, of course, offering a psychological theory of religion that is too simplistic to handle the relevant phenomena adequately. The kind of theory Freud offered, in terms of religion meeting psychological needs, proved neither convincing nor scientifically fruitful, and is not a good precedent to follow. However, the problems faced in studying religion psychologically are not unique. The study of emotion, for example, faces similar problems, in that emotion is as multifaceted a phenomenon as religion. Broadly conceived 'cognitive' approaches to religion may give useful coherence, as they are currently doing in the psychology of emotion. The psychology of religion interfaces with many other areas of general psychology. Indeed, it can be argued that the subject is pursued most fruitfully when its links with general psychology are strong.

3.2 Relation to Sociology

The separation between sociology and psychology is in some ways an artificial one, and affects the study of many social phenomena, including religion. Most religious phenomena have both social and psychological aspects. For example, conversion and speaking in tongues have been studied fruitfully from both perspectives. However, social and psychological

studies often proceed in isolation from each other; it seems unlikely that either alone will be able to offer an adequate explanation of any religious phenomenon though there is a strong case for the psychology of religion having stronger links with the sociology of religion. This tends to raise the question of whether religion is to be understood in social or individualistic terms. Both perspectives are no doubt needed for an adequate understanding of religion, and the psychology of religion does not exclude a social approach. Though the increasing interest in the brain processes in religion are likely to pull the psychology of religion in a biological direction, the psychology of religion, like other aspects of psychology, will need to remain both a social and a biological science, and to hold together the different methodologies involved.

3.3 Relation to Theology

Lastly, what should the relationship be between the psychology of religion and theology? Different countries have different traditions about the relationship between theology and religious studies, some maintaining a sharp distinction between the 'insiders' perspective of theology and the 'outsiders' perspective of religious studies (including the psychology of religion); others have tried to maintain links between them. This issue is affected by the postmodern turn in philosophy, which is casting doubt on the possibility of any wholly detached, objective study of anything, including religion. The psychology of religion has been slow to accommodate itself to these general intellectual developments, though it is hard to see that it can stand aside from them indefinitely. One particular issue is whether there is scope for insiders' and outsiders' explanations to run alongside one another. Charismatic phenomena represent an example. It is not clear that the study of the psychological processes involved in charismatic religion excludes a complementary religious account in terms of the activity of the Holy Spirit.

3.4 The Nature of Explanation

Many of these issues are brought together in questions about the nature of explanation in the psychology of religion. Explanations in psychology are seldom complete explanations, and are certainly unlikely to be such in the psychology of religion. This means that religion will not be reduced to psychology, or 'explained away.' Equally, religion is not a 'natural kind,' and all explanations in the psychology of religion will therefore be culturally contextual, and depend on interpretations of the nature of religion; the psychology of religion can never hope to provide a science of religion that is a 'view from nowhere.'

Nevertheless, there is no sharp boundary between the explanation and interpretation of religion, and it is inherent in the nature of religion that it will adopt a mixed approach.

Bibliography

- Beit-Hallahmi B, Argyle M 1997 *The Psychology of Religious Behaviour, Belief and Experience*. Routledge, New York
- d'Aquili E G, Newberg A B 1998 The neuropsychology of religion. In: Watts F (ed.) *Science Meets Faith*. SPCK, London
- Freud S 1962 *The Future of an Illusion*, Rev. edn. Norton, New York
- Hay D 1987 *Exploring Inner Space: Scientists and Religious Experience*. 2nd edn. Mowbay, London
- Hood R W Jr., Spilka B, Hunsberger B, Gorsuch R 1996 *The Psychology of Religion: An Empirical Approach*, 2nd edn. Guilford Press, New York
- James W 1985 *The Varieties of Religious Experience*. Harvard University Press, Cambridge, MA
- Jeeves M A 1997 *Human Nature at the Millennium*. Baker Books, Grand Rapids, MI
- Jung C G 1969 *Psychology and Religion: West and East*. Princeton University Press, Princeton, NJ
- Loewenthal K M 1995 *Mental Health and Religion*. Chapman and Hall, London
- Malony H N, Lovekin A A 1985 *Glossolalia: Behavioral Science Perspectives on Speaking in Tongues*. Oxford University Press, New York
- Meissner W W 1984 *Psychoanalysis and Religious Experience*. Yale University Press, New Haven, CT
- Palmer M 1997 *Freud and Jung on Religion*. Routledge, London
- Paloutzian R F 1996 *Invitation to the Psychology of Religion*, 2nd edn. Allyn and Bacon, Boston
- Proudford W 1985 *Religious Experience*. University of California Press, Berkeley, CA
- Wulff D M 1997 *Psychology of Religion: Classic and Contemporary Views*, 2nd edn. Wiley, New York

F. Watts

Religion, Sociology of

Though central to the concerns of the founding fathers and to such major mid-twentieth century figures as Talcott Parsons and Robert Merton, religion has, for some time, been regarded as of marginal interest within the discipline of sociology. However, the recent resurgence of ethno-religion and of militant variants of major world religions has incited mainstream reconsideration of the salience of the topic thereby directing attention toward the sociology of religion as a specialized area of research. Meanwhile, the subdiscipline has acquired a significant role in the generation of insights and perspectives among theologians, historians of religion and exponents of *Religionswissenschaft*.

1. *The Role of the Classics*

With roots in the earliest versions of sociological thought, the sociology of religion remains strongly influenced by the classics and Durkheim and Weber still supply the main axes of intellectual tension within subdisciplinary theoretical discourse (O'Toole 1984). The former's concern with religion's role in social cohesion, group stability, and the reproduction of sociocultural forms is complemented by the latter's preoccupation with its part in radical, large-scale social and cultural transformations. Their writings are constantly ransacked for insights while their ideas, often elaborated or even creatively misinterpreted, continue to permeate contemporary discussion of such topics as secularization, sectarianism, civil religion, new religious movements, ecclesiastical organizations, ritual, and religious transformations.

Through the scope of their theories and the depth of their insights, the classics inevitably breed rival readings and multiple interpretations, e.g., while Durkheim and Weber are often viewed as exponents of the secularization thesis, both offer ammunition for its adversaries. The ambiguity of Durkheim's legacy is evident in attempts to explain his fundamental theory of religion. Here, metaphorical parallelism battles symbolic realism as defenders of a materialist Durkheim confront those who proclaim him an idealist thinker.

Though Durkheim and Weber constitute the twin pillars of the subdiscipline's classical legacy, other figures have played significant, though more circumscribed roles. When articulation and development of the Weberian church-sect distinction was a central subdisciplinary obsession, Troeltsch's *The Social Teaching of the Christian Churches* (1923) universally was accorded classic status as the charter document of the study of religious organization. Today, Troeltsch's volumes attract attention, less for insights into church establishment and sectarian voluntarism than for their exploration of 'the secret religion of the educated classes.' With its cultic and mystical components, Troeltsch's treatment of this topic presciently anticipates a contemporary concern with religious individualization and privatization and thus deserves continuing scrutiny.

Rarely granted subdisciplinary recognition until the 1970s, the writings of Marx and Engels may now be said to have achieved minor classical standing. They have influenced theoretical discussion as well as inspiring a minority scholarship whose neo-Marxism is characterized by a sense of the 'relative autonomy' of religion and a belief in continuity rather than conflict between the views of Marx and Weber (O'Toole 1984).

Alexis de Tocqueville has long lurked in the background of sociological discussions of American religion but, until recently, his *Democracy in America* (1835) never achieved the subdisciplinary classic status of Niebuhr's *The Social Sources of Denominationalism*

(1957). Whereas Niebuhr is now largely neglected, however, Tocqueville increasingly is commissioned in the service of a current fascination with the 'vitality' of US religion and the sources of a supposed 'American Exceptionalism' in the religious realm. In its assertion of the significance of religious disestablishment and voluntarism, his work is esteemed as the fountainhead of a perceived 'New Paradigm' in the study of American religion (Warner 1993).

Long invoked without ever approaching the subdisciplinary standing of Durkheim or Weber, Simmel, too, is the subject of revived interest attributable to the uncanny timeliness of his writings (Simmel 1997). Reflecting his resistance to that domination of the objective spirit over the subjective which characterizes modernity, his focus on a diffused religious attitude in preference to institutionalized religious forms has a decidedly contemporary resonance. Careful to distinguish between formal creeds and the subjective experience of pious individuals, Simmel explores religiosity as well as religion, portraying the latter as a co-creation of acting believers which bridges the private and the public. In a subdiscipline currently grappling with issues of religious individualization, autonomy, fragmentation, and *bricolage*, Simmel's potential contribution has perhaps never been greater.

While many sociologists of religion are content to continue a conversation with the classics, others are impatient to break the intellectual spell of what they perceive as the increasingly irrelevant products of a passing industrial era. Those intent on subdisciplinary transcendence of the classics are acutely aware, however, that scholars accustomed to invoking this legacy as new problems arise will not relinquish it lightly.

2. *Definitions*

In most definitional surveys, discussion hinges upon the merits of functional vs. substantive definitions. Within these frameworks, the relative advantages of inclusive or exclusive conceptualizations are assessed while specific formulations are examined for theoretical coherence and empirical integrity. Since its inception, the sociology of religion has been the scene of such a prolonged battle between the proponents of these strategies for defining religion that, in a sense, the debate between Durkheim and Tylor has been interminable. Contemporary substantivist formulations focus on such elusive phenomena as the sacred, the transcendent, the supernatural or the superempirical. Functional definitions portray religion as an indispensable source of social consensus and cohesion which aids human adaptation to the environment and ensures societal survival by explaining the meaning of existence and providing answers to questions of ultimate concern.

The question of how inclusive a definition may be while still retaining analytical utility is one that may be

directed against both functional and substantive formulations. On the substantivist side, the categories of the supernatural and superempirical are notoriously open-ended while, in the most inclusive functional definitions, society itself assumes a religious character. Debate over the relative merits of definitional strategies often hinges on the issue of how closely sociological usage should adhere to societal understandings. On one side, scholars express frustration that sociological usage of the term 'religion' is often remote from and counter-intuitive to anything which ordinary people would recognize as 'religious.' On the other, it is argued that 'religion' is a culturally specific term with no resonance in many societal contexts and that, in any case, no definition of general utility can be expected to conform to the specific demands of ordinary language usage.

Although subdisciplinary practitioners are accustomed to pursuing their research in a continuing context of definitional disagreement, it is an error to believe that lack of consensus in this area has no practical effect. The degree of social and cultural significance which sociologists of religion currently attribute to religion is influenced, to a great extent, by their definitional predilections.

3. Major Themes

In the period of its greatest growth, since the 1960s, the sociology of religion has exhibited both continuity and change. The classics remain a continuous presence although their range of application has expanded. A Judaeo-Christian emphasis is still clearly discernible but it has become self-consciously circumscribed within more globally comparative ambitions. The functionalist tone of earlier theoretical discussions has given way to a polycentric approach that critically assesses the merits and defects of rival perspectives claiming a hearing. Most notably, the variety of topics examined under the subdisciplinary rubric has widened significantly. Whereas the sociology of religion was previously directed largely toward familiar subject matter, it now increasingly ventures into mysterious territory and assumes novel theoretical and methodological challenges. It cannot be said that the sociology of religion has been unresponsive to the unfolding religious processes and transformations of the latter half of the twentieth century. On the contrary, its evolution has been largely propelled by such reactions. The present generation of subdisciplinary specialists is engaged in research on many fronts. While continuing to investigate such phenomena as mainline religion, civil religion, millenarianism, conversion, and the changing role of clergy, its representatives also are exploring energetically a wide array of new topics generated by the turmoil of the last quarter of the twentieth century. E.g., events in Eastern Europe and the former Soviet Union since 1989 have underlined

religion's potency in the forging of ethnonationalism, thereby provoking numerous studies of the contemporary fate of religion in specific former Iron Curtain countries.

Although, in its strict American Protestant sense, it has claimed a longstanding subdisciplinary interest, 'fundamentalism' has also captured scholarly attention in its contemporary comparative meaning, particularly since the Iranian revolution. Adopting a blanket term coined by the mass media, sociologists of religion have participated in a number of subdisciplinary and interdisciplinary projects which examine the currently politically-charged topic of religious extremism (Christian, Jewish, Muslim, Hindu, etc.) under this heading. The effort to discern its underlying characteristics has raised the crucial question of whether the fundamentalist impulse toward differentiation and totality represents an assault on modernity, a manifestation of modernism or an intriguing combination of both.

Subdisciplinary practitioners exhibit an enhanced capacity for scrutiny of specific religious developments within a large-scale comparative context and in terms of more general theoretical debate. The study of Pentecostalism, e.g., is no longer confined to the margins and backwoods of Western societies. In the works of contemporary specialists, it emerges as a movement which offers a serious challenge to mainline religions in the West, is rapidly transforming the religious maps of Latin America and Africa and is making a serious spiritual, social and economic impact on parts of Asia. Sociologists of religion have been forced by world events to situate their subject matter in an increasingly international, comparative, and global framework whether their interest is in fundamentalism, charismatic Christianity, New Religious Movements, or any other major topic of research. It is unsurprising, therefore, that they have led the way in promoting globalization as a concept of central theoretical importance, not just for their subdiscipline but for sociology as a whole.

Born in the political alienation, generational conflict, and crisis of values which characterized the era of the Vietnam War and its aftermath, the study of New Religious Movements (NRMs) revitalized a subdiscipline which appeared to have exhausted the possibilities of the church-sect approach to religious organization. Over the last quarter of the twentieth century, sociologists of religion have continued to study Unificationists, Scientologists, Rajneeshis, Transcendental-Meditationists, Hare Krishnas, and many other movements of religious innovation and experimentation. In the process, they have accumulated a mass of ethnographic material documenting their inner dynamics, recruitment practices, resource mobilization, and ideological development while simultaneously struggling to discern the broader theoretical import of this information. Originating amidst heady notions of an emerging 'new religious con-

sciousness,' another 'Great Awakening' and an imminent 'Return of the Sacred,' NRM research has long subscribed to more sober assessments of its subject matter. Nonetheless, scholars in this field reject suggestions that they are engaged in a mere pursuit of the trivial and exotic. The significance of NRMs is believed to lie in their role as indicators of broader sociocultural trends but, unfortunately, there is wide disagreement concerning their overall significance for contemporary and future religion. For some commentators, they represent a decisive blow to the secularization thesis, their very existence contradicting any notion of a 'disenchantment of the world.' For others, they constitute strong evidence in support of the secularization thesis, confirming by their tiny membership, syncretic ideology, and cultural marginality the inability of religion to restore its fortunes in any significant way.

What might be considered a 'second wave' of NRM research has focused, more recently, on such topics as Satanism, Witchcraft, neo-Paganism, and various manifestations of 'New Age' spirituality. Such studies provide intriguing comparisons and contrasts with first-generation NRMs as well as with occultic movements formed at the end of the nineteenth century. They also examine phenomena such as Wicca and neo-Paganism as indicators of such broad religious trends as individualization, privatization, and *bricolage*, thereby suggesting their essentially post-industrial or postmodern character. A growing fascination with New Age spirituality as a highly individualistic and experiential enterprise has perhaps played a part in the encouragement or revival of subdisciplinary interest in such topics as healing, the natural environment, the human body, and collective memory as they relate to religious issues. As sociologists of religion increasingly have shifted their attention in these directions, the possibility of fruitful interchange between students of NRMs and analysts of so-called 'New Social Movements' from other subdisciplines has been raised. Though the response to this challenge has been rather muted, such cooperation might facilitate a greater infusion of general theoretical ideas into the sociology of religion.

4. Methods and Theories

Within the sociology of religion, methodology and theory typically have been implicit rather than explicit. Perhaps partly due to a certain insulation and isolation, the kinds of self-consciously theoretical disputes which occur in other subdisciplines rarely have been witnessed. Nonetheless, since the unraveling of functionalist orthodoxy in the 1970s, the subdiscipline has experienced significant, if civil, discord which stubbornly resists efforts at securing synthesis.

The subdiscipline employs a wide range of quantitative and qualitative methodological techniques including large-scale social surveys, participant-ob-

servation, and textual analysis of contemporary and historical materials. Here, as in sociology generally, one investigator's sophisticated device is another's blunt instrument, although open debates on this theme are rare. For the most part, a wall of indifference segregates quantitative from qualitative methodologists although it has been breached occasionally. In the broadest methodological sense, the reciprocal relationship between sociological research and religious activity merits attention though it is often ignored on the basis of a glib distinction between religious sociology and sociology of religion formulated in a religious context less bewildering than the current one. To what extent sociologists of religion see their own reflection in the religious realm on which they report is a sensitive issue which surfaces in a number of ways. Thus, in asserting the impact of societal reality upon sociological method, some observers suggest that a purported contemporary transition from institutionalized to individually autonomous forms of religion necessitates a shift from qualitative fieldwork directed at groups to quantitative surveys aimed at individuals. In contrast to this, other commentators ponder the possibility that perception of a rise in privatized religion may itself be partly a product of specific, widely disseminated research techniques directed toward discernment of the dimensions of individual religiosity.

While the complexities and subtleties of the reciprocal relationship between observers and those they observe have come to the fore most notably in discussions of scholarship and partisanship in the NRM field, they are equally evident in other religious settings. Students of Liberation Theology and Base Communities, e.g., are often confronted by sociological ideas in explicit use by those under observation. More generally, widespread use of the term 'secularization' by mainline churches seems to offer an excellent example of social science rebounding upon society. It also provides a cautionary tale of the manner in which societal embrace of a sociological notion may possibly inspire a self-fulfilling prophecy.

In any general assessment of the theoretical situation within the sociology of religion, it is difficult to overcome the impression that scholars are traveling in all directions at once and that every assertion eventually evokes its exact opposite. Empirical-analytic, historical-hermeneutic, and critical conceptions of science are all in evidence and the major tendencies in contemporary sociological theorizing are represented in varying degrees (e.g., functionalism, conflict theory, Marxism, critical theory, symbolic-interactionism, phenomenology, rational-utilitarianism). In such a situation, it is difficult to ascertain the main drift of theoretical innovation but recent important developments may be noted.

A mood of skepticism toward Grand Theories and Grand Narratives appears, in some quarters, to have intensified suspicions of deduction and prediction to a

point where the initial nomothetic impulse of the subdiscipline has been supplanted by an idiographic 'historical turn.' This involves a strategy of methodological individualism, a stress on individual autonomy and creativity, and an overall preoccupation with agency rather than structure. In a parallel development, longstanding inductive and interpretive tendencies have experienced reinvigoration and reinforcement as theorizing has taken a decidedly 'cultural' direction. In their efforts to take seriously the meaning of religion in the lives of believers and practitioners, proponents of this perspective employ the insights of cognitive and phenomenological theory while simultaneously drawing upon anthropological sources and material derived from the field of Cultural Studies. Fundamental to this cultural approach is a fervent antireductionism which rejects the notion of religion's epiphenomenal character and refuses to regard religious phenomena as really something else. Vigorously endorsing methodological individualism and deductive scientific logic, rational-choice theory has had an undeniably profound impact on the sociology of religion during the 1990s. Appointing Adam Smith to the subdisciplinary classical pantheon, its advocates employ economic metaphors to describe, explain, and predict religious behavior. Led by de Tocqueville's identification of a causal link between voluntarism and vitality in the American religious experience, they indicate the paramount importance of choice and competition in US religious 'exceptionalism.' Extending their analysis in a comparative direction, they strongly oppose suggestions of the inevitability of secularization, linking dynamic religious growth to a process of 'deregulation' in the religious realm. While it has won many converts, rational-choice theory has left many unconvinced by either its specific market analyses of religious profit-and-loss or its fundamental assumptions concerning religious behavior. To its most severe critics, its basic propositions appear embedded in ambiguity and tautology while its claims to universal applicability reflect an unfortunate regression into a presociological mode of analysis. Nonetheless, rational-choice analysis now represents such a significant presence that it is regarded by many of its admirers, and perhaps even some of its detractors, as the central constituent of a new subdisciplinary paradigm.

The themes of secularization and the contemporary transformations of religion have permeated subdisciplinary discussion in all its aspects since the 1960s. The concept of secularization has, for many, proven itself to be a plausible, persuasive, and resilient means for interpreting the fortunes of religion in the twentieth century. From another viewpoint it has been judged a misguided myth whose influence has retarded scholarly progress and whose persistence underlines the subdisciplinary need to transcend the classics. Although its classical provenance is tinged with ambiguity, the secularization perspective has long held a

privileged position within the sociology of religion, being described frequently as a dominant 'paradigm.' Logically dependent for its meaning on the definition of 'religion,' the concept has an intrinsic inclination to ambiguity which is often reinforced by vague and careless usage. Despite the best efforts of its proponents to articulate the main elements of the secularization thesis with analytical exactitude, debates over its validity frequently are characterized by argument at cross-purpose. While the validity of empirical evidence is occasionally at issue, facts are not typically in dispute and discussions tend to hinge on the interpretation of various social patterns.

What is new about the current debate is that proponents of the secularization thesis are now decidedly on the defensive. The undermining of the secularization thesis has occurred gradually through the efforts of scholars ranging theoretically from functionalists to phenomenologists but this process has accelerated in the last decade as a result of the arguments of rational-choice theorists and a new subdisciplinary awareness of the historical record. Indeed, some sociologists of religion discern, in the current convergence of these tendencies, the emergence of nothing less than a new subdisciplinary paradigm destined to displace the secularization thesis. Whether this new outlook, which combines deductive theory with social history, may accurately be described as a paradigm is in dispute. There is little doubt, however, that it captures a certain *fin de siècle* subdisciplinary mood of stock-taking and reassessment in which the prospects of religion appear much brighter than a few decades ago. Perhaps the 'new paradigm' is best understood within the context of changing attitudes to modernity. As a by-product of current postmodernist debate, the concept of modernity has experienced revival and refurbishment, a situation with significant consequences for the study of religion. The familiar conception which assumed that as a result of differentiation and relativization, secularization was an inextricable and inevitable component of modernity has given way to a revisionist version of modernity which neither precludes the persistence of religion nor constrains its vitality. Although its significance is challenged by defenders of the secularization thesis, mounting evidence of the vigor of religion on the international scene encourages subdisciplinary embrace of a more encompassing vision of modernity.

Interestingly, this vision is complemented by an increasingly broad conception of religion. Any attempt to portray religion's changing face since the 1970s must stress the extremely wide range of phenomena currently accommodated under the religious rubric before tracing those trends which constitute the common currency of all analyses of the contemporary religious scene. These include the processes of disestablishment, individualization, privatization, fragmentation, and globalization as well as a number of

variants and counter-themes. In the process by which a primarily institutional conception of religion has given way to a far more inclusive and elusive subdisciplinary formulation, the impact of Luckmann's 'invisible religion' thesis cannot be overestimated (Luckmann 1967). Where earlier functionalist definitions had focused on such vast notions as 'ultimate concerns' and 'grounds of meaning,' Luckmann raises the stakes even higher by identifying the transcendence of biological nature by the human organism as a religious phenomenon. He thus describes a general anthropological condition that links the creation of religion inextricably with the formation of the individual and the self. In various guises, this has been a major subdisciplinary theme during the years following the promulgation of his provocative thesis.

Concerned for so long with religion's role in the creation, preservation, and perpetuation of collectivity and community, many sociologists of religion have transferred their attention to the role of the individual and the dimensions of personal religiosity. As with other major subdisciplinary developments, this shift of attention is a product of perceived changes in societal reality as much as theoretical logic. It is not necessary to subscribe to the secularization thesis in order to be aware of the profound crisis which has beset the major institutional forms of religion in Western societies since World War II. Moreover, the vastly diminished role of religion as the 'ideology of community' has become increasingly evident. Thus, the new direction taken by the subdiscipline is part of its effort to cope with the rapid transformation of religion under conditions of postindustrialism and postmodernity.

One aspect of this new reality is starkly evident in the work of Bellah et al. (1985) whose insight into the success and decline of American 'civil religion' has expanded into a concern with the collapse of community. Viewing dwindling involvement in religious organizations within the broader picture of a general decline in participation in political parties, trade unions, voluntary associations, and community groups, Bellah and co-workers perceive religion as increasingly a matter of private concern and personal idiosyncrasy.

This theme is echoed in all current investigations of institutionalized religion, particularly in its 'mainline' forms. Although there is no immediate prospect of the collapse of organized religion, its future arguably depends on its ability to accommodate new attitudes to religious authority and revised forms of religious involvement. Commentary on the contemporary religious climate demonstrates considerable consistency in detecting a number of key related processes at work. Some observers describe individualization and privatization. Others refer, in a US context, to a 'third disestablishment' involving an expansion of personal autonomy. Fragmentation frequently is explored and a growing tendency among believers to 'mix and match' elements from a variety of denominational and

religious traditions is emphasized. The terms *bricolage*, religion *à la carte*, and consumerism are also utilized in portraying the extent to which individuals privately construct personally meaningful syncretic systems of belief for themselves. In examining the dilemmas of organized religion, many scholars discern a disinclination toward total acceptance of any denominational or religious belief system either through inheritance or conversion. This denotes a fundamental alteration in the popular conception of religion. No longer regarded as a source of authoritative meaning, increasingly it is considered instrumentally as a broad, pliant, cultural resource at the disposal of autonomous individuals. Moreover, the accuracy of this assessment is underlined by researchers who document an accelerating incidence of 'belief without belonging' as a mode of religious commitment.

The future of organized religion is precarious and unpredictable but with the waning of the secularization thesis, it is less often discussed in terms of inevitability. Though it seems certain to remain a major player on the religious scene for the foreseeable future, increasingly it will share the subdisciplinary limelight with more 'diffuse' and 'invisible' forms of religion. Often under the general umbrella of 'the sacred,' the attention of sociologists of religion focuses increasingly on uninstitutionalized forms of religion and their involvement in the transformation of the self. In this regard, studies of popular, 'common,' and 'implicit' religion, along with investigations of diffuse versions of spiritual and mystical experience, are now a noteworthy part of the subdisciplinary agenda. They are accompanied by efforts to discern the presence of the sacred in such secular settings as 'self-help' and therapy groups, fitness centers, weight-loss clinics, and healthfood stores which are often depicted as instruments in the pursuit of individual self-perfection.

True to form, however, there is no subdisciplinary consensus on an overall or inevitable passage into more privatized forms of religion in the postindustrial era. On the basis of evidence drawn from Poland, Spain, Brazil, and the US, Casanova (1994) perceives an intersection of religion and politics and a revival of 'public religion' in the modern world. This conception is given credence by globalization theory that regards both privatization and politicization as products of the modern global condition.

Whether religion's privatization or individualization will accelerate or whether its former capacity as a source of authoritative meaning will reassert itself and inspire novel, enduring, and significant forms of collective and public spiritual expression remains to be seen. One thing is certain: reappraisal of the past and adjustment to the present have ensured that, in the sociology of religion, the future is not what it used to be.

See also: Collective Beliefs: Sociological Explanation; Durkheim, Emile (1858–1917); Luhmann, Niklas

(1927–98); Rationalism; Religion: Definition and Explanation; Religion, History of; Religion, Psychology of; Religiosity, Sociology of; Simmel, Georg (1858–1918); Spencer, Herbert (1820–1903); Weber, Max (1864–1920)

Bibliography

- Beckford J A 1990 The sociology of religion 1945–1989. *Social Compass* 37(1): 45–64
- Beckford J A 1989 *Religion and Advanced Industrial Society*. Unwin Hyman, London
- Beckford J A, Luckmann T (eds.) 1989 *The Changing Face of Religion*. Sage, London
- Bellah R N, Madsen R, Sullivan W M, Swidler A, Tipton S M 1985 *Habits of the Heart: Individualism and Commitment in American Life*. University of California Press, Berkeley, CA
- Beyer P 1994 *Religion and Globalization*. Sage, London
- Bibby R W 1987 *Fragmented Gods: The Poverty and Potential of Religion in Canada*. Irwin, Toronto
- Bruce S (ed.) 1992 *Religion and Modernization: Sociologists and Historians Debate the Secularization Thesis*. Clarendon Press, Oxford, UK
- Capps W H 1987 Society and religion. In: Eliade M (ed.) *The Encyclopedia of Religion*. Macmillan, New York, Vol. 13
- Casanova J 1994 *Public Religions in the Modern World*. University of Chicago Press, Chicago
- Clarke P B, Byrne P 1993 *Religion Defined and Explained*. St. Martin's Press, New York
- Davie G 1998 Sociology of religion. In: Swatos W H (ed.) *Encyclopedia of Religion and Society*. AltaMira Press, Walnut Creek, CA
- Davis W 1987 Sociology and religion. In: Eliade M (ed.) *The Encyclopedia of Religion*. Macmillan, New York, Vol. 13
- Dawson L L 1998 *Comprehending Cults: The Sociology of New Religious Movements*. Oxford University Press, Toronto
- Flanagan K, Jupp P C (eds.) 1996 *Postmodernity, Sociology and Religion*. St. Martin's Press, New York
- Glock C Y, Hammond P E (eds.) 1973 *Beyond the Classics? Essays in the Scientific Study of Religion*. Harper & Row, San Francisco
- Hadden J K 1995 Religion and the quest for meaning and order: old paradigms, new realities. *Sociological Focus* 28(1): 83–100
- Hamilton M 1998 *Sociology and the World's Religions*. St. Martin's Press, New York
- Hammond P E (ed.) 1985 *The Sacred in a Secular Age: Toward Revision in the Scientific Study of Religion*. University of California Press, Berkeley, CA
- Hammond P E 1992 *Religion and Personal Autonomy: The Third Disestablishment in America*. University of South Carolina Press, Columbia, SC
- Heelas P 1996 *The New Age Movement: The Celebration of the Self and the Sacralization of Modernity*. Blackwell, Oxford
- Hervieu-Léger D 1987 Faut-il définir la religion? *Archives de Sciences Sociales des Religions* 63(1): 11–30
- Kurtz L R 1995 *Gods in the Global Village: The World's Religions in Sociological Perspective*. Pine Forge Press, Thousand Oaks, CA
- Luckmann T 1967 *The Invisible Religion: The Problem of Religion in Modern Society*. Macmillan, New York
- Luhmann N 1984 *Religious Dogmatics and the Evolution of Societies*. Edwin Mellen Press, New York

- Niebuhr H R 1957 *The Social Sources of Denominationalism*. Meridian Books, New York, (first published 1929)
- O'Toole R 1984 *Religion: Classic Sociological Approaches*. McGraw-Hill Ryerson, Toronto
- Poewe K (ed.) 1994 *Charismatic Christianity as a Global Culture*. University of South Carolina Press, Columbia, SC
- Roof W C 1984 American religion in transition: a review and interpretation of recent trends. *Social Compass* 31(2–3): 273–89
- Simmel G 1997 *Essays on Religion*. Yale University Press, New Haven, CT
- Swatos W H Jr (ed.) 1993 *A Future for Religion? New Paradigms for Social Analysis*. Sage, Newbury Park, CA
- de Tocqueville A C 1835 *De la démocratie en Amérique*, 2 Vols. Librairie de Gosselin [Paris 1946 *Democracy in America*. Alfred A. Knopf, New York]
- Troeltsch E 1923 *Die Soziallehren der christlichen Kirchen und Gruppen*. JCB Mohr, Tübingen, Germany (first published 1911) [1931 *The Social Teaching of the Christian Churches*, 2 Vols. Macmillan, New York]
- Wallis R, Bruce S 1989 Religion: The British contribution. *British Journal of Sociology* 40(3): 493–520
- Warner R S 1993 Work in progress toward a new paradigm in the sociological study of religion in the United States. *American Journal of Sociology* 98(5): 1044–93
- Wuthnow R J 1988 Sociology of religion. In: Smelser N J (ed.) *Handbook of Sociology*. Sage, Newbury Park, CA
- Wuthnow R J 1992 *Rediscovering the Sacred: Perspectives on Religion in Contemporary Society*. Eerdmans, Grand Rapids, MI
- Young L A (ed.) 1997 *Rational Choice Theory and Religion: Summary and Assessment*. Routledge, New York

R. O'Toole

Religiosity: Modern

The *Shorter Oxford English Dictionary* defines 'religiosity' as 'Religiousness, religious feeling or sentiment.' Taking this as a cue, 'modern religiosity' is defined by contrasting it with modern forms of traditional religion. Thus modern religiosity is taken to refer to forms of religion which emphasize experience rather than the doctrines of the religious traditions of modern times; and which are relatively inchoate, indeterminate, or formless rather than being spelled out by way of authoritative belief systems. Religiosity, that is to say, is detraditionalized, the established, spelled-out, regulatory orders of the traditional not being much in evidence, if at all. Instead of living in terms of authoritative orders, religiosity is very much to do with the sphere of consciousness. And the fact that religiosity is so integrally bound up with the life of the person means that it is very much in the hands of the experiencing subject. It largely operates beyond tradition: that is to say, autonomous subjects—not traditions—are authoritative, subjects developing their own religiosities by way of the test of their own life requirements. In sum, rather than

heeding the injunctions of religious tradition, modern religiosity concerns self-based exploration of life's meaning; the autobiographical experience.

1. *Contrasting Modern Religiosity and Modern Religion*

Writing at the beginning of the twentieth century, Emile Durkheim drew a distinction between 'a religion handed down by tradition, formulated for a whole group and which it is obligatory to practice' and 'a free, private, optional religion, fashioned according to one's own needs and understanding' (cited in Pickering 1975, p. 96). With modifications, this contrast provides the basis for distinguishing between 'modern religion' and 'modern religiosity.'

1.1 *Modern Religion*

To begin with the more familiar term, modern religions provide tradition-informed, institutionalized forms of religious activity: the church, chapel, monastery, mosque, or temple. Unlike more insulated premodern religions, however, few modern religions are 'formulated for a whole group' or are 'obligatory to practice' (as Durkheim puts it). For with modernity, information and population flows have greatly increased, this meaning that globalizing and pluralizing processes have provided most people, in most cultures, with a whole range of religions from which to select. This said, however, modern religions strive (to varying degrees and with varying degrees of success) to ensure that participants remain faithful to the tradition on offer. Participants, that is to say, are instructed in a particular, clearly articulated, path to salvation. Examples include long-established forms of tradition (such as Catholicism) as well as more recent developments (such as those new religious movements dating from the nineteenth century, the Mormons providing an illustration, or those organizations that have developed during the later twentieth century, the Japanese 'new new religions' here serving to illustrate).

1.2 *Modern Religiosity*

Generally speaking, modern religiosity flourishes in realms beyond the institutionalized and traditionalized territories occupied by modern religions. Rather than going to church, chapel, or temple, at least to any significant degree, those involved draw on widely available cultural resources—books and magazines, films and videos—to develop their own religiosities. Or those involved resource themselves by way of 'new spiritual outlets,' turning to practitioners who provide spirituality (healings, meditations, empowerments) in

a relatively informal, weakly institutionalized and nondirective fashion. Then again, there are those who exercise their imaginations to conclude that there must be 'something' more to life than the merely mundane.

Reliance on cultural resources and detraditionalized new spiritual outlets goes together with the fact that a key characteristic of modern religiosity is that those involved think of themselves as exercising their own authority to develop their own religiosities (or spiritualities). As Robert Wuthnow (1989) makes the point with regard to North America, 'the religion practiced by an increasing number of Americans may be entirely of their own manufacture' (p. 164). In order to resource their lives, individuals might draw on magical, occult, or esoteric teachings and practices; New Age themes to do with the sacralization of the Self or the importance of connecting to the spirituality of nature; the mysterious forces of 'X-file' culture; notions such as 'spirituality,' 'vital force,' 'higher power,' 'life force' or 'soul,' and even 'angels'; or the 'inner wisdom' of religious traditions. In addition, individual religiosities can include the not-so-obviously religious in the quest for finding meaning in life, self-fulfillment or well-being: a point also made by Wuthnow when he writes of the growing popularity of meaning systems which involve, for example, an 'eclectic synthesis of Christianity, popular psychology, *Reader's Digest* folklore, and personal superstitions, all wrapped in the anecdotes of the individual's biography' (Wuthnow 1989, p. 164).

Although the term 'modern religiosity' can be taken to include vague 'there must be something more' beliefs, which have little direct bearing on the lives of those concerned, modern religiosity primarily dwells on what it is to be 'alive.' Durkheim ([1915] 1971) thought that the future lay with 'a religion which will consist entirely in internal and subjective states, and which would be constructed freely by each of us' (p. 47, my emphasis). What Durkheim is referring to is clearly very much to do with personal life: all those subjectivities which comprise 'being alive.' Now in the words of Wuthnow (1991), who is explaining what he means by the 'subjective' or 'lived realities' dimension of the religious: 'by "subjective", I mean an understanding of religion that privileges its connection to the individual, its location within the inner part of the individual (beliefs, outlooks, felt needs, emotions, inner experiences), and its role within the individual's personality (providing meaning, wholeness, comfort, psychological compensations, even a sense of attachment or belonging)' (p. 272).

Religiosities (or spiritualities) of life are most obviously in evidence among those who are not involved in strong—obeying the Other—traditionalized modern religion. For whereas traditionalized religious institutions serve to liberate people from their 'merely' human lives—as 'fallen,' or, in many Eastern traditions, as contaminated by 'attachments'—in favor of what lies beyond, modern re-

ligiosity concerns the autonomous subject, liberated from the demands of tradition and intent on exploring and developing what it is to be alive. This said, however, modern religiosity is by no means absent within less strongly traditionalized forms of modern religion. Peter Berger (1969), for example, in his discussion of liberal Christian teachings, notes that 'religious "realities" are increasingly "translated" from a frame of reference of facticities external to the individual consciousness to a frame of reference that locates them *within* consciousness' (p. 166). And to the extent that subjectivization or internalization is taking place within modern religions, so too does the divide between modern religion and modern religiosity become undermined.

2. Modern Religiosity and Peoples' Lives

Turning to how religiosities can serve peoples' lives, four main life focuses can be briefly summarized. The first concerns the life of the individual person. What matters in this regard lies with finding religiosity/spirituality within oneself. What matters is addressing personal 'growth,' divorce, illness, ageing, by engaging with what one's own spirituality has to offer. The second concerns the life of the relationship. Religiosity is here put to work to enhance personal relationships, or to operate in therapeutic fashion. The third concerns relationships with nature. The emphasis here is on living in terms of the life of the natural; that life which informs nature from trees to species to the cosmic. And the fourth concerns what life has to do with work, increasing attention being paid in countries like the USA to exploring ways of bringing 'life back to work' by way of religiosity or spirituality.

3. The Growth of Modern Religiosity

Not enough research has been done to provide anything approaching a clear picture as to how modern religiosity is faring on a global scale. It is probably true to say, however, that in much of the world, where religious traditions—Islam, Hinduism, 'popular' Buddhism—remain strong, modern religiosity is relatively unimportant. But even in Islamic countries like Bangladesh, its role should not be neglected: many people could well combine components of Islam with folk beliefs and practices to do with improving the quality of personal life; and Tagore, the great figurehead of the nation, taught a 'religion of humanity' which is a powerfully relational form of life spirituality.

The one thing that is reasonably clear is that modern religiosity is well in evidence in countries where various processes of modernity—rationalization, differentiation (between the religious realm and the political, for example), democratization, pluralization, consumer-

ization, and materialization—have weakened the hold of religious traditions. Indeed, in such countries modern religiosity could well be expanding. The reason is simple. More people are ceasing to participate in institutionalized, traditional religion than are becoming atheists or agnostics, it following that the number of those who are 'somehow' religious without being significantly involved with institutionalized religion has increased.

Northern Europe would appear to provide a major site for growth of this kind. In Britain, for example, the number of those regularly attending institutionalized religion dropped by 30 percent between 1950 and 1990, but the numbers of those becoming atheists or agnostics only increased by some 10 percent (entailing that there has been an increase of some 20 percent in those who are somehow religious without attending). It is true that many of these nonattenders might still be involved with modern religion, reading their Bibles at home, for example. But there are distinct signs that religiosity is growing: belief that 'there is some sort of spirit or vital force which controls life' has now overtaken the (traditional) 'There is a personal God' (Bruce 1996, p. 270), and New Age spiritualities of life are expanding (see Heelas, 1996). Or consider Finland: 15 percent are regular church attenders, with the figure in decline, while 63 percent believe that the 'best way to get in touch with God' is by being 'in' nature, such beliefs apparently gaining in numerical significance (Heino, 1997, p. 21).

Turning to the USA, research by Roof and Gesch (1995) provides convincing evidence of growth. Concentrating on the baby-boom generation, namely those 75 million (or so) people born between 1946 and 1964, by the 1990s approximately half were involved with traditional religion. Of the remainder, namely the 'religious individualists,' approximately half have turned to 'a deeply mystical conception of the Deity,' having little if anything to do with 'organized religion' (p. 141). As for elsewhere, it is highly likely that much the same picture of growth applies to countries like Canada and Australia, or to cities like Rio de Janeiro and Tokyo. Finally, on the question of growth, it is important to note that modern religiosity is also developing within modern religion: see, for example, Donald Miller's (1997) discussion of the (rapidly expanding) 'postdenominational churches' of North America with their strong emphasis on being alive in the right kind of way: 'purity of heart is more important than purity of doctrine' (pp. 127–8).

4. Explaining the Turn to Religiosity

Of the many factors that have contributed to the growth of religiosity, attention is paid here to what is called 'the cultural turn to life.' The key to the argument is that modernity has resulted in very high

value indeed being ascribed to life. In the words of Georg Simmel ([1918] 1997), 'this emotional reality—which we can only call *life*—makes itself increasingly felt in its formless strength as the true meaning or value of our existence' (p. 24). And given its value, life has quite naturally come to serve as the 'home' for religiosity for many of those who remain religious without participating in modern religion.

Looking at this argument in somewhat greater detail, Simmel and Max Weber, then Arnold Gehlen, Peter Berger, and Thomas Luckmann, have claimed that modernity has resulted in significant numbers becoming disillusioned with what the 'primary' institutions of society have to offer. From modern religion (loss of faith seen in secularization) to political ideologies, to the nuclear family, to the workplace (Weber's 'iron cage'), the claim is that the institutional order no longer succeeds in providing adequate sources of identity, purpose, or value in life. Accordingly, so the argument goes, people have been 'thrown back' on themselves as the key source of significance. With little to have faith in beyond themselves, what matters becomes a matter of what lies within: one's psychology, one's personal ethicality ('being true to oneself'), the quality of one's emotions, the importance of being authentic, the value of finding out what one truly is and capable of becoming; becoming 'whole' or healing oneself.

It is but a short step to explain the growth (in certain sociocultural contexts) of religiosities or spiritualities of life. Other than atheists and agnostics, those who have come to have 'ultimate' faith in what their interior or personal lives have to offer are going to seek out those forms of the religious which sacralize experiences of what it is to be alive.

See also: New Religious Movements; Religion: Definition and Explanation; Religion, Phenomenology of; Religiosity, Sociology of; Secular Religions; Secularization

Bibliography

- Berger P 1969 *The Social Reality of Religion*. Faber and Faber, London
- Bruce S 1996 Religion in Britain at the close of the 20th century: A challenge to the silver lining perspective. *Journal of Contemporary Religion* 11: 261–75
- Durkheim E [1915] 1971 *The Elementary Forms of the Religious Life*. George Allen & Unwin, London
- Heelas P 1996 *The New Age Movement. The Celebration of the Self and the Sacralization of Modernity*. Blackwell, Oxford, UK
- Heino H 1997 *Mihin Suomi Tänää Uskoo*. WSOY, Helsinki, Finland
- Miller D 1977 *Reinventing American Protestantism. Christianity in the New Millennium*. University of California Press, Berkeley, CA
- Pickering W S F 1975 *Durkheim on Religion*. Routledge & Kegan Paul, London

- Roof W C, Gesch L 1995 Boomers and the culture of choice. In: Ammerman N, Roof W C (eds.) *Work, Family, and Religion in Contemporary Society*. Routledge, New York, pp. 61–79
- Simmel G [1918] 1997 *Essays on Religion*. Yale University Press, New Haven, CT
- Wuthnow R 1989 *The Struggle for America's Soul. Evangelicals, Liberals, and Secularism*. William B. Eerdmans, Grand Rapids, MI
- Wuthnow R 1991 Religion as culture. In: Bromley D G (ed.) *Religion and the Social Order. New Developments in Theory and Research*. JAI Press, Greenwich, CT, Vol. 1 pp. 267–83

P. Heelas

Religiosity, Sociology of

1. Introduction

Religiosity is a broad and sweeping concept that unites much of the subject matter in the psychology and sociology of religion; consequently, the idea includes much more than it excludes.

1.1 Definition

In the most general sense, religiosity is a term that is used to refer to socially relevant manifestations of: (a) individual beliefs in a supernatural or transcendent reality; (b) relations of individuals to 'high-intensity values' (Roof 1979, p. 18); (c) 'beliefs about the nature, meaning, and purpose of reality' (Glock and Stark 1966, p. 3); and (d) all the behaviors that flow directly from them. Synonyms for the term in the literature of the social sciences include—among others—religiousness, religious affiliation, attachment, commitment, conviction, devotion, interest, and involvement.

2. Conceptualization and Measurement

2.1 Dimensions of Religiosity

Beginning in the 1950s, social scientists of religion directed a great deal of attention to elaborating definitions of religiosity that extended understanding of the concept beyond a mere count of people in the pews of churches and synagogues. Advances accumulated quickly and, according to one review (Roof 1979, p. 17), the 1960s became 'a pace-setting decade in this area.' In particular, sociologists had set to work crafting indicators of religiosity that took account of more than one aspect of religious commitment. After mid-century, scholars encountered an empirical religion that was, more than anything else, multi-dimensional.

2.1.1 Associational vs. communal religion. Some of the most basic quantitative studies of the postwar religious scene in the United States were the work of Gerhard E. Lenski (see, e.g., Lenski 1953). Lenski's major analysis of religious belief and behavior in Detroit, *The Religious Factor* (Lenski 1963 [1961]), gauged two fundamental types of religious involvement: associational (established through attendance at church services and meetings of church-related groups); and communal (the development of a religiously homogeneous circle of primary attachments). He additionally recognized two religious orientations among his respondents: one that stressed doctrinal orthodoxy, or right belief; and another, devotionism, that located religiosity in personal contact with God (Lenski 1963, pp. 22–6; see also Burwell 1999b, Deconchy 1987, p. 35, Roof 1979, pp. 19, 20–1).

2.1.2 5-D and its critics. Also among the earliest examples of religiosity conceived theoretically and captured empirically in more than a single questionnaire item were the efforts of Charles Y. Glock and his colleagues to portray religion along up to five dimensions (Glock 1962, Glock and Stark 1966; see also Burwell 1999a, Deconchy 1987, pp. 33–43, Gorsuch 1984, pp. 232–34, Hilty et al. 1984, p. 252, Roof 1979, pp. 20–2). Initially, Glock proposed that there were at least four distinct ways in which a person could be religious. Religious commitment, he postulated, could take the form of religious emotion or 'direct knowledge of ultimate reality' (the experiential dimension); fidelity to a certain body of beliefs (the ideological dimension); participation in divine worship, prayer, and other ceremonies (the ritualistic dimension); and indications of the 'secular effects' of religion on the individual and society (the consequential dimension).

To this array he later added an intellectual dimension (Fukuyama 1961), because the believer would be expected to be 'informed and knowledgeable about the basic tenets of his faith and its sacred scriptures.' Although Glock admitted that this fifth dimension was 'clearly related' to religious ideology, he insisted that the two were separable, for religious belief demanded familiarity with something in which to believe (Glock 1962, p. S99). Indeed, each of the five dimensions represented to Glock a theoretically and empirically distinct aspect of religiosity; he explained at the outset that 'being religious on one dimension does not necessarily imply religiosity on other dimensions' (Glock 1962, p. S100).

Later studies referred to these constructs together as 'Religiosity in 5-D' (Faulkner and DeJong 1966, Roof 1979, pp. 22–9, Weigert and Thomas 1969). The results of several (see, *inter alia*, Hilty et al. 1984, King and Hunt 1990) offered new or augmented models for research, but the most prominent critics (Clayton and Gladden 1974), aided by large-scale surveys and data

reduction through factor analysis, concluded that there was little empirical warrant for most of the presumed dimensions.

2.2 Meaning Systems and the Religious Imagination

Considering the manifold difficulties and relatively modest benefit in measuring religiosity as socially normative behavior within the confines of churches, sociologists—including Charles Glock himself—in the 1970s and 1980s devoted themselves to fashioning research techniques that would incorporate the broader connotations of the concept. A salient example of this approach is the innovative study by Robert Wuthnow (1976) on the meaning systems to which individuals routinely resort to lend structure and order to their experiences. Using data from the San Francisco Bay area, Wuthnow's survey could discern four different ways of making sense of causation in life (see Deconchy 1987, p. 39, Roof 1979, pp. 33–5). Respondents who attributed events to the intervention of divine agents (God or the devil) in human affairs were classified as ratifying theism. In contrast, those who believed that persons created the conditions of their own lives were said to back individualism. Believers in the explanatory power of social science traced the origins of personal situations to their surroundings in society, while people whose meanings came from mysticism relied on direct experience of the sacred and its mediation through nature, the creative arts, and fantasy.

A somewhat different stream of research has employed sets of interview questions to derive scales that purport to measure the symbolic and narrative basis of religion—what Andrew M. Greeley has termed 'the religious imagination.' Although formal theology is organized into series of logical statements, Greeley contends that personal religiosity predates and often overpowers tame instruction in dogma and doctrine because it is rooted in an imaginative appreciation of stories that renew hope in human actors. Religion is thus primarily composed of 'images that are prior to and perhaps more powerful than propositions' (Greeley 1981, p. 1).

To reach this imaginative capacity scientifically, Greeley and his collaborators have devised combinations of survey items that summarize varying images of God (see also Vergote et al. 1969), chart the perceived 'warmth' of different religious images, probe perceptions of religious figures like Jesus and the Virgin Mary, and describe the degree to which respondents regard heaven as a pleasurable place and the relationship between the divine and the human as 'grace-filled' (Greeley 1981, pp. 23–9).

2.3 Civil and Public Religiosity

Another aspect of religiosity that is sometimes overlooked because it lies just outside the institutional

boundaries of formal faith is what Robert N. Bellah (1967) has termed civil religion (see also Deconchy 1987, pp. 38–9, Roof 1979, pp. 29–31). Civil religion may be defined as a religious perspective that places the past, present, and future of the nation against the backdrop of transcendent values. Commenting on the United States, Bellah (1967, pp. 3–4) maintains that the components of civil religion in America ‘provide a religious dimension for the whole fabric of American life, including the political sphere.’ A special category of symbols and rituals that pertain to the identity and purpose of Americans thereby attains a sacred status in collective life.

In such a perspective nations not only have histories, they have destinies. Although a patriotic citizen may also be civilly religious, belief in the civil religion is different from simple patriotism. Despite, too, the frequent mixture of politics and piety (Glock and Stark 1966, pp. 82–6), civil religion is usually more than the casual consecration of secular aims. In the eyes of an adherent to the civil religion, a nation’s official philosophies, public policies, and international statecraft all invite the intervention of Providence—at the same time that their consequences promise to come under divine judgment. A kind of religious significance thus permeates the entirety of the public realm yet is dependent on no single church or denomination for its vitality.

2.4 Types of Religious Orientations

2.4.1 Intrinsic vs. Extrinsic religion. To a larger extent than sociologists, psychologists of religion have concentrated on the questions of how and why people hold the religious beliefs that they do. Central in this tradition of inquiry has been Gordon W. Allport’s distinction between Intrinsic (I) and Extrinsic (E) religious orientations (Allport and Ross 1967; see also Batson and Ventis 1982, pp. 140–9, Deconchy 1987, pp. 37, 40, Hunt and King 1971, Roof 1979, p. 20). ‘No approach to religiousness has had greater impact on the empirical psychology of religion,’ one scholar observes; Allport’s scale to measure religious orientations, furthermore, is ‘one of the most frequently used measures of religiousness’ (Donahue 1985, p. 400). Another pair of commentators (Kirkpatrick and Hood 1990, p. 442) concurs that the I/E comparison ‘currently represents the backbone of empirical research in the psychology of religion.’

As the labels imply, intrinsically religious persons are those for whom religion is an end and not a means. Religion is ‘their master motive’ that provides the guide to understanding all of life. In contrast, extrinsically religious persons engage in religion with the hope for something else, such as peace, protection, or status in the community. For them, a religious faith ‘is lightly held or else selectively shaped to fit more primary

needs’ (Allport and Ross 1967, p. 434). Extrinsic religion, notes one psychologist (Donahue 1985, p. 400), ‘is the religion of comfort and social convention, a self-serving instrumental approach.’

An enormous amount of research on the I and E scales has been conducted in nearly 50 years. Perhaps the most famous set of these studies has linked the Extrinsic religious orientation to expressions of prejudice. But one of the main complaints about the formulation reverts back to its first appearance—as a more value-laden contrast by Allport (1950) between ‘mature’ and ‘immature’ religious faith. The implication is that Intrinsic religiosity is somehow genuine and ‘pure’—unencumbered by ulterior motives, while the Extrinsic brand is phony, selfish, and manipulative. In the descriptions of I and E, the Intrinsic believer is clearly preferred to the Extrinsic; as Michael J. Donahue (1985, p. 416) concedes, the E scale ‘does a good job of measuring the sort of religion that gives religion a bad name.’ James E. Dittes (1971, p. 375) is much more harsh: to him, I and E struggle beneath a ‘heavy contraband load of value judgment that simply will not be sloughed off.’

Add to this problem of bias the repeated statistical finding that I and E are not two poles of the same continuum, but rather two variables with only a limited correlation (Allport and Ross 1967, p. 437, Donahue 1985, pp. 401–4, Kirkpatrick and Hood 1990, pp. 448–50, 454). For these and other reasons, I and E may survive in the scientific study of religion solely in an amended shape.

2.4.2 The ‘Quest’ variable. Dissatisfied with the dual categories of Intrinsic and Extrinsic religiosity, psychologist C. Daniel Batson has developed a third type of religious orientation that he dubbed Quest (Batson and Ventis 1982, pp. 149–68; see also Deconchy 1987, p. 40, Donahue 1985, pp. 411–14, Gorsuch 1984, pp. 234–5, Kirkpatrick and Hood 1990, p. 447). A person who conceives of religion as a quest is one who actually welcomes ‘a readiness to doubt and to be self-critical’ and who acknowledges the ‘incompleteness and tentativeness’ of religious faith at any moment (Batson and Ventis 1982, p. 149).

2.4.3 Other varieties of religious experience. One of the oldest traditions in the social sciences of religion attempts to study, with precision and objectivity, experiences of the sacred. The subjective nature and ordinarily fleeting passage of religious experiences, however, makes this goal extremely difficult to achieve. Yet the pioneering psychologist and philosopher William James tried, and his classic work, *The Varieties of Religious Experience* (James 1902), is the product.

In spite of this distinguished start, studies of mystical and ecstatic experiences have not yet culminated in a secure body of knowledge. It is nevertheless possible to specify several recurring features of religious experiences, as distilled from numerous cases. People who have experienced a mystical episode remember feeling as if it occurred outside the physical constraints of space and time. In this special state, they recall a spirit of peace and even elation enveloping them. Further, they report receiving personal illumination or enlightenment of unquestioned authority, while still assimilating seemingly contradictory or paradoxical perceptions. They apprehend a sacred presence, and commonly recount a sensation that all of reality was merging into oneness. In the single greatest barrier to researchers, though, mystics also contend, after their re-entry into normal consciousness, that the essence of their experience was indescribable (see Stace 1960, Stark 1965).

3. A New Religiosity?

Many of the ideas and approaches that social science devised in the last half-century to handle religiosity are slowly falling into disuse today, as both conceptions of religion in social science and evidence of its effects are changing.

More and more, scholars in disciplines such as psychology and sociology (e.g., Roozen et al. 1993, pp. 45–8) are describing a contemporary style of religious belief and practice that is at heart individualistic. Although people in plural societies have always had available to themselves a range of religious choices, seldom in the past have these decisions been so free of influence from social factors in one's background like family, ethnicity, or language. Rather, the assumption of a religious affiliation in the current circumstance is increasingly voluntary; it amounts to an expression of identity, and often a private or highly personal one at that.

Thus recent years have seen an abundance of religious pursuits, especially among the so-called 'baby boom' generation (cohorts born between 1946 and 1965), that have an eclectic quality. Popular varieties of faith combine elements of meditative practices, environmental activism, 'New Age' consciousness, and self-awareness with a therapeutic flavor. In the twenty-first century, many who no longer consider themselves conventionally religious nevertheless claim to possess a deep 'spirituality.'

4. Conclusion

Current research and theorizing in the scientific study of religion seeks to depart from older approaches to religiosity that anchored the topic in categorical contrasts, whether bipolar or multidimensional. Instead, sociologists and psychologists are busy tailoring

strategies and tools for research on a concept of religion that is personal more than institutional, amorphous and fluid rather than historical and orthodox, and not static but changing.

In an age when religiosity assumes both a ubiquity in common concerns over truth and authenticity, and an idiosyncrasy in multiplying paths of personal development, flexible ideas and methods of interpretation are definitely needed. Yet, now as before, the challenge for social scientists is to pursue analyses of the sacred into the meanings with which humans endow it and through the actions that proceed from these commitments.

See also: Belief, Anthropology of; New Religious Movements; Religion and Health; Religion: Definition and Explanation; Religion: Evolution and Development; Religion, History of; Religion, Psychology of; Religion, Sociology of; Religiosity: Modern; Religious Fundamentalism: Cultural Concerns; Religious Nationalism and the Secular State: Cultural Concerns; Transnational Religious Identities (Islam, Catholicism, and Judaism): Cultural Concerns; Women's Religiosity

Bibliography

- Allport G W 1950 *The Individual and His Religion: A Psychological Interpretation*. Macmillan, New York
- Allport G W, Ross J M 1967 Personal religious orientation and prejudice. *Journal of Personality and Social Psychology* 5: 432–43
- Batson C D, Ventis W L 1982 *The Religious Experience: A Social-Psychological Perspective*. Oxford University Press, New York and Oxford, UK
- Bellah R N 1967 Civil religion in America. *Daedalus* 96: 1–21
- Burwell R 1999a Dimensions of religious commitment (Glock and Stark, 1966). In: Hill P C, Hood Jr R W (eds.) *Measures of Religiosity*. Religious Education Press, Birmingham, AL
- Burwell R 1999b Religious orientation and involvement (Lanski, 1961). In: Hill P C, Hood Jr R W (eds.) *Measures of Religiosity*. Religious Education Press, Birmingham, AL
- Clayton R R, Gladden J W 1974 The five dimensions of religiosity: Toward demythologizing a sacred artifact. *Journal for the Scientific Study of Religion* 13: 135–43
- Deconchy J-P 1987 Les méthodes en psychologie de la religion: Leur évolution récente [Methods in psychology of religion: Current trends]. *Archives de sciences sociales des religions* 63: 31–83
- Dittes J E 1971 Typing the typologies: Some parallels in the career of Church-Sect and Extrinsic-Intrinsic. *Journal for the Scientific Study of Religion* 10: 375–83
- Donahue M J 1985 Intrinsic and Extrinsic religiousness: Review and meta-analysis. *Journal of Personality and Social Psychology* 48: 400–19
- Faulkner J E, DeJong G F 1966 Religiosity in 5-D: An empirical analysis. *Social Forces* 45: 246–54
- Fukuyama Y 1961 The major dimensions of church membership. *Review of Religious Research* 2: 154–61

- Glock C Y 1962 On the study of religious commitment. *Religious Education [Research Suppl.]* 57: S98–110
- Glock C Y, Stark R 1966 *Christian Beliefs and Anti-Semitism*. Harper and Row, New York and London
- Gorsuch R L 1984 Measurement: The boon and bane of investigating religion. *American Psychologist* 39: 228–36
- Greeley A M 1981 *The Religious Imagination*. William H. Sadlier, Inc., New York and Los Angeles, CA
- Hilty D M, Morgan R L, Burns J E 1984 King and Hunt revisited: Dimensions of religious involvement. *Journal for the Scientific Study of Religion* 23: 252–66
- Hunt R A, King M B 1971 The Intrinsic–Extrinsic concept: A review and evaluation. *Journal for the Scientific Study of Religion* 10: 339–56
- James W 1902 *The Varieties of Religious Experience*. Longmans, Green and Company, New York
- King M B, Hunt R A 1990 Measuring the religious variable: Final comment. *Journal for the Scientific Study of Religion* 29: 531–35
- Kirkpatrick L A, Hood Jr R W 1990 Intrinsic–Extrinsic religious orientation: The boon or bane of contemporary psychology of religion? *Journal for the Scientific Study of Religion* 29: 442–62
- Lenski G E 1953 Social correlates of religious interest. *American Sociological Review* 18: 533–44
- Lenski G E 1963 [1961] *The Religious Factor: A Sociological Study of Religion's Impact on Politics, Economics, and Family Life*. Doubleday and Company, Garden City, NY
- Roof W C 1979 Concepts and indicators of religious commitment: A critical review. In: Wuthnow R (ed.) *The Religious Dimension: New Directions in Quantitative Research*. Academic Press, New York
- Roozen D A, Carroll J W, Roof W C 1993 La génération née après-guerre et la religion instituée: Un aperçu sur 50 ans de changement religieux aux États-Unis [The post-war generation and establishment religion: A window to 50 years of religious change in the United States]. *Archives de sciences sociales des religions* 83: 25–52
- Stace W T 1960 *Mysticism and Philosophy*. Lippincott, Philadelphia, PA
- Stark R 1965 A taxonomy of religious experience. *Journal for the Scientific Study of Religion* 5: 97–116
- Vergote A, Tamayo A, Pasquali L, Bonami M, Pattyn A, Custers A 1969 Concept of God and parental images. *Journal for the Scientific Study of Religion* 8: 79–87
- Weigert A J, Thomas D L 1969 Religiosity in 5-D: A critical note. *Social Forces* 48: 260–63
- Wuthnow R 1976 *The Consciousness Reformation*. University of California Press, Berkeley, CA

K. J. Christiano

Religious Fundamentalism: Cultural Concerns

‘Fundamentalism’ is the name applied to characterize a variety of modern religious movements internationally, but its use is contested in some cases. The key feature in these movements is their reactive

character. That is, they are not simply traditional, classic, conservative, or orthodox. Scholars and others usually speak of fundamentalism only in cases that find sets of religious people responding to certain kinds of challenges. The challenges may vary, but they tend to be variations on what the reactors or observers ordinarily would call ‘the modern,’ ‘modernity,’ ‘modernization,’ or ‘modernism.’

1. The Term ‘Fundamentalism’

One of the most controversial features of the movements called fundamentalist is the term most widely used to designate them. There are two main reasons for some participants in the movements, people in statecraft, and scholars to have reservations about the use. First, the name is derived from a reactive movement in American Protestantism, and for some decades after the 1920s remained chiefly related to them. To some, the transfer of this term to other religious contexts looks like one more example of a Western linguistic imperialism that needs to be resisted. Others contend that features peculiar to a specific Protestant conflict differ in so many respects from those in other reactive movements that only confusion could result from the use of the term. The second reason why many are reluctant to accept the term as their self-description or for scholarly purposes has to do with the fact that it can be used pejoratively and stigmatically. In a way, this resistance is ironic, since in the American context and, by analogy, in other religions and cultures, those who formed and originally adhered to such militant movements originally wanted to be stigmatized, and still may want this, but on their own terms.

Thus, in an American controversy in the 1920s, partisans in denominational warfare regularly complained that while the majority of their fellow-believers claimed to be conservative, as conservatives these did not do battle for the divine cause. Since they were too and compromising, it was argued, a new term had to be coined and applied. This term would help provide boundaries around the movement and create distance between it and conventional conservatism—or against anything else that implied a potential for compromise and comfortable assimilation. But because the strategy of setting boundaries and deliberately seeking to offend others went so far that it often ran counter to some of the positive purposes of the movements, many became discontented with the self-chosen name. Some sought moderating terms. In the American case, ‘neo-evangelicalism’ and then ‘evangelicalism’ served this purpose.

The term will not be found in dictionaries or encyclopedias, even encyclopedias of religion in America, before the time of a fundamentalist–modernist controversy in the US, one that came to a

climax around 1925. When the word does make its way into reference books after that period, for example in German or French dictionaries, all the meanings of fundamentalism get reduced to something about a battle over the verbal inerrancy and literal interpretation of the Christian Bible—a reduction abhorrent to members and confusing to observers of Islamic, Jewish, or other religions' factions.

Given these burdens associated with the term, there is good reason to ask about its continued usefulness and appropriateness. Several reasons are cited in defense, among them the following: first, the word is 'there.' So widespread is the use among reporters, scholars, and politicians that it is not likely to disappear because of discontent over its use by many. Better to clarify what is meant, it is said, than to invent new coordinating terms that will meet similar resistance in due course. The second reason has to do with the need for some such coordinating designations. To mention that need is, of course, to open the door to the many controversies over comparative methods. James (1902) typically made a defense of the practice of typifying and classifying. Without classifying, it is hard to understand many features of phenomena. At the same time James warned against the danger of losing the individual and particular character of the objects studied.

In the case of fundamentalism, the interest in coordination is almost entirely formal, not substantive. That is, one notices aspects of certain religious movements by studying characteristic features while being well aware that these movements all differ from each other in respect to content. Indeed, a Muslim fundamentalist will be 'doctrinally' more distant from a Protestant fundamentalist than modernists in both religions would be to each other. So there is no point in using the word 'fundamentalist' to characterize the central teachings of these religious expressions.

Even here, however, the line is a bit fuzzy. Thus, through comparison one can generalize that all the fundamentalisms, apparently paradoxically because they resist that which is modern, and though electronic media result from modern invention, are adept at using radio, television, film, and the Internet to hold their groups together, seek converts, and express themselves to outsiders. That observation grows out of formal comparison. But a second generalization, that all the fundamentalisms are not only patriarchal but that in their talk about God, they stress God's maleness while moderate and modernist movements play down masculine references to deity does show that there can be certain kinds of substantive similarities.

A third reason for using the term, say its defenders, is that no clear alternative is apparent. Some reporters, trying to avoid the word, will use something like 'militants,' appropriate for one dimension of some of their life. But these words do not signal anything of cultural context: one can be militant about revenge,

the search for territory, or any number of themes that have nothing to do with religious reaction.

Further defense of the term comes from those who say that all words travel, and globally people would make little sense of each other if they did not have terms that could be reappropriated in a variety of cultures. Thus words like 'nationalism,' 'liberalism,' 'colonialism,' 'conservatism,' 'radicalism,' 'democracy,' and 'republicanism' are born within Western contexts and are translated for usage elsewhere. The People's Republic of China means something very different with the term 'Republic' than do those who speak of America as a republic. Yet there are reasons for people to use the term in both contexts—and then clarify what they mean and point to differences.

Attempts to replace the term usually are specific and cannot be used for coordinative or comparative purposes. Thus, one might speak of 'radical revolutionary nationalist Islamic reform,' using other borrowed terms, and then find that they do not relate to phenomena in the Catholic or Hindu world. Mention of such instances suggests why many choose to use 'fundamentalism': it has acquired some rather specific meanings, meanings that deserve examination.

2. The Originating Instance in America as a Case Study

American Protestantism, having been divided into many denominations from the beginning, experienced some transdenominational conflicts in the nineteenth century. The first great breach occurred between the North and the South in the Civil War, when churches on each side inspired, monitored, provided morale for, and negatively defined the other side in religious terms. But all through the century another split was taking shape, thanks to latent tensions between theologically more liberal and more conservative parties.

These tensions increased with the arrival in the US after the middle of the nineteenth century of Darwinian evolution, the higher criticism of the Bible, and various forms of political progressivism. Many conservatives massively resisted theories of evolution, in part because these were seen as assaults on the Bible. They all rejected biblical criticism as a threat to the foundations of the faith. And progressivism did not do justice to apocalyptic messages in the Bible. Meanwhile, modernist parties began to develop toward the turn of the twentieth century. These were strongest at the prestigious theological schools such as Harvard, Yale, and Chicago—but rather dramatically not at Princeton University and Princeton Theological Seminary, from which some intellectual impetus for fundamentalism was to come.

Between 1910 and 1915 the conservative parties issued an influential set of booklets called *The Fundamentals*. From then on the accent on funda-

mentals grew. In 1919 a World's Christian Fundamentals Association was born. In 1920 during a battle for power in the Northern Baptist Convention a Baptist editor proudly chose and publicized the term 'fundamentalist' for his faction. The Presbyterians and to a lesser extent other mainstream Protestant denominations entered into the battle. The celebrated Scopes trial in Tennessee in 1925, one that featured conflict over the teaching of evolution in public schools, was the best known event in the formation and publicizing of fundamentalism.

This is not the place to trace in detail the outcomes of the struggles. Fundamentalists lost the battles for control, but inventively went their way developing their own Bible colleges, seminaries, mission boards and, thanks to schisms, denominations. They were a presence but not a vivid one during this period. In 1941 some formed an American Council of Christian Churches, which engaged in rivalry with a more moderate National Association of Evangelicals born in 1942. Still then seen by nonfundamentalists as marginal, eccentric, and withdrawn, fundamentalists remained generally aloof from politics. However, during the 1970s most of them reversed their stance and began to adopt political interventionist stances. They have been a prominent presence ever since.

To detail this case study further might well bias the case of those who charge that global and interreligious use of the term fundamentalism is too reliant on the American experience. But one could study cognate movements elsewhere—typically the Muslim Brotherhood in Egypt—and find similar reasons for Islamic conservative groups there to move into reaction. A description of elements that make up what is called fundamentalism in any religion will be more profitable, however, than would extending the number of case studies.

3. A Phenomenological Approach to Fundamentalist Features

By phenomenological here is meant a pattern of observation in which the analysts 'bracket' their presuppositions—in this case, for example, assumptions about the positive or negative aspects of fundamentalism—and as if naively asks: what stands out in the phenomena that have been called fundamentalist? Another way to put it is: what phenomena developed in the twentieth century that occasioned the need for fresh naming—since words like conservative and traditional had been easily available? These questions enable one to escape the charge that a priori definitions color the observations. These are some of those features. First, while fundamentalisms originally were dismissed by others as 'the old time religion' and while some parade themselves as traditional, they differ in significant respects from conservatism. In the eyes of doctrinal, political, and sometimes even armed mili-

tant partisans, conservatives may be too passive. They are too content to let the world go about its business not reckoning with them. If in any particular instance the cause being challenged is seen as divine—and we are here dealing only with *religious* fundamentalisms—for anyone to be content with tradition is a betrayal.

Fundamentalisms rise when historically conservative religious movements experience something that challenges them to their core. In Iran and elsewhere in the Middle East after the late 1970s, the challenger might be 'the Great Satan' of Western imperialism, specifically in its American form. In hyper-Orthodox Israel the threat to Jewish identity itself appears to parties such as Gush Emunim to come from non-observant, compromising Jews who might even let some of the land of Israel slip away in diplomatic negotiations. In the US the challenge in the 1920s came in more doctrinal ways, through contests over biblical literalism. But in all these cases, the threat had to be seen as total; fundamentalists do not bother much about trivial issues.

Since the challenge was total, the response had to be drastic. Fundamentalisms characteristically are reactive. They rise among communities that have been acted upon, and they act back. Very often they use the very instruments that had presented the challenge. If a nation-state is the ominous force, they make efforts to win power, through the ballot if possible and through weaponry if necessary. (Let it be noted that not all fundamentalisms take on any form of military or terrorist guise. Some even shun the political arena if they feel that they can convert sufficient recruits from that worldly zone or win their way simply through cultural militancy, as with disputes over 'values' and media.) Again, fundamentalists tend to use some of the weapons that have brought change, whenever, for example, they take instruments of mass media, the agents that penetrated their homes and consciousness, and become sophisticated in using these to fight back.

Somehow, in every case, they react. Not to be reactive is to be not fundamentalist. George Marsden, the foremost historian of the American version, describes a fundamentalist as 'an evangelical who got mad.' In reaction, the fundamentalist seeks ammunition for defense and aggression. Here is where the fundamentals come in. Fundamentalists reach back to foundations, some in the Qur'an, some in Torah, some in the New Testament, still others in classic events of the past, and claim them for their own. In every case in their eyes there had to be some sort of golden moment or precious book to which arguing parties can resort in efforts to determine winners and losers.

Acting upon these fundamentals, members of the group then engage, as noted, in setting boundaries. They are often described as bearing a Manichaean cast. This means that the cosmos is divided into God versus anti-God, good vs. evil, our people vs. everyone else. They may not necessarily withdraw into enclaves,

but they will tend to use code-words, slogans, informal testing devices that help them screen out the most dangerous persons and forces: the moderates. The total outsider, the infidel, the Great Satan is useful for mobilizing forces, but is not a threat the way a compromising modernizer is seen to be.

With the setting of boundaries comes the claim that participants are an elect people, God's own instrument. They know something of where history is going; it has been revealed to them by sacred books in which they find the fundamentals. By the way, it is easier to be fundamentalist in 'religions of the book' such as Judaism, Christianity, and Islam, than it is in faiths such as Hindu and Buddhist, where no single divine scriptural revelation is foundational.

The elect people, called to represent God, must keep themselves pure, try to recruit others, and, to use the words from the American struggle in the 1920s, 'do battle for the Lord.'

The battles may vary in character. Thus in most cases in the Islamic world, the enemy has not been the Great Satan of the US. In Algeria, Egypt, and elsewhere, the foe is a regime seen as corrupt, secularizing, and compromising even though all its leaders are themselves practicing Muslims. In Israel, the parties called fundamentalist may engage in warfare against Arab Muslims, but they take most pains to define themselves over against compromising Jews.

Action may be hortatory, prophetic, political, military, or terrorist, though it is often only the latter two that draw the world's attention.

4. Accounting for Fundamentalisms in Culture

No scholar claims that modern fundamentalisms have no counterparts or analogues in the past within the various religions. All of them have times and forces of traditionalist reaction and conservative aggression. It is valuable to study these analogues. But the word fundamentalism has more appropriately been reserved for responses to what has been seen as modern. One way to put it: these are reactive movements that appeared after the modern academy thought there would be no more like it, that the religions that survived would be small, weak, and always compromising.

Psychologists have on occasion assessed persons in the movements or the movements and generalized about their mentality. Participants tend to be people who cannot tolerate ambiguity, paradox, contradiction, or doctrinal and moral 'fuzziness.' Yet it is difficult to reduce fundamentalisms to this psychological cast, given the fact that people with such predilections in one culture turn religious fundamentalist and those in others do not.

Political and social scientists perceiving fundamentalisms in culture often relate the rise of such movements to class conflict. Marxist interpretations,

for example, can see them as futile resistances by people who take resort in other-worldly solace when they can do little to gain in this world. It is true that in many instances it is the poor and apparently powerless that turn to fundamentalism. But it is also true that those who react may do so from the comfortable confines of Princeton Theological Seminary and Princeton University, as the intellectual definers of early American fundamentalism did. And one will find fundamentalists of wealth in many an American city and note that some of these extravagantly fund fundamentalist projects. Class analysis does contribute to understanding the phenomena, but most scholars again resist reductionism of a sort that fundamentalists are 'nothing but' this or that.

Instead, they give evidence that they are complex responses and reactions to many kinds of cultural stimuli and as such manifest features that are best appraised with a variety of perspectives and disciplines. Among these would be humanistic efforts to bring to bear history, philosophy, and religious studies. These are, after all, self-described religious movements, and seen to be such.

One way to approach the study is to listen as anthropologists do for 'agent's description.' Someone who pays attention to, records, and coordinates the testimony of fundamentalists in many religious contexts will find that the issues that evoke response begin very close to the self of the believer. Diplomats, reporters, and scholars often begin at the furthest extensions of fundamentalisms, as when some engage in terrorism or revolutionary military action.

Instead, whoever is alert to fundamentalist witness will be indeed close to the ego's center. The believer begins by defending a world-view, including one that has been recently changed thanks to conversion. The world looks different to someone who is an initiate, 'born again,' in some versions, than it does to someone else who is distant from or quiet about religious impulses. Very close to that egocenter is the question of personal, social, and cultural identity? To whom do I belong? Whom can I trust? With whom can I check signals and from whom do I draw inspiration? Those who ask such questions in the erosive climate and culture of late-modernity find that identity becomes a major question. They also note that it is threatened by forces such as travel, modern mass communications, mass higher education, especially where the challenges of pluralism and relativism are fierce.

As close to the center of the person as are world view and identity is the issue of gender. Fundamentalists preoccupy themselves with what it means to be a man or a woman. The Manichaean differentiation is strong: there is no room for ambiguity in defining gender. The fundamentalisms, for example, all reject homosexual expression. From gender they move to defense of family. In all the cultures studied by fundamentalist scholars, 'family values' and shoring up the family, however defined, is central to the program.

From the orbits of gender and family one moves to other challenges. Thus, one's mind and habits are shaped by mass communications, and fundamentalists tend to react by trying to screen out uncongenial images, to engage in boycotts, or attempt legally to censor and exclude broadcast or published materials that they believe to be evil or subversive of their values.

Education is another zone in which fundamentalists do battle. They start their own schools to protect and nurture their own young and often engage in efforts to subvert the character of others', especially the public's schools, when their teaching and habits are offensive. Similarly, many of them have challenged modern medical practices, and make the clinic their arena of conflict. They organize against abortion, euthanasia, and many forms of treatment that violate their view of the body as God's temple.

In the sphere of economics it is harder to generalize. Some religious texts favored by fundamentalists—notably those detailing *shari'ah*, Islamic law—may be explicit was when they proscribe the taking of interest. But some fundamentalist movements promote communal, one might say communistic economic patterns, while others, notably in North America, often align themselves with *laissez-faire* economic forces. But when they perceive economic circumstances as a threat to them, they may adopt any number of strategies, and find their holy books less defining and definitive.

When it comes to efforts to change politics, for example, it is through the use of the instruments of political propaganda, the ballot, and constitutional revision, that fundamentalists find themselves reckoned with most strenuously. Where they cannot take over a regime, they may in God's name undertake revolutionary activity. It is activity of that sort that elicits headlines about fundamentalist militants around the world.

See also: Fundamentalism (New Christian Right); Islamic Fundamentalism; Religious Fundamentalism: Cultural Concerns

Bibliography

- Ammerman N T 1987 *Bible Believers*. Rutgers University Press, New Brunswick, NJ
- Arjomand S A (ed.) 1984 *From Nationalism to Revolutionary Islam*. State University of New York Press, Albany, NY
- Arjomand S A 1988 *The Turban for the Crown: The Islamic Revolution in Iran*. Oxford University Press, New York
- Caplan L (ed.) 1987 *Studies in Religious Fundamentalism*. State University of New York Press, Albany, NY
- Cohen N J (ed.) 1990 *The Fundamentalist Phenomenon: A View from Within; A Response from Without*. Eerdmans, Grand Rapids, MI
- Dekmejian R H 1985 *Islam in Revolution: Fundamentalism in the Arab World*, 1st edn. Syracuse University Press, Syracuse, NY
- Dessouri A E H (ed.) 1982 *Islamic Resurgence in the Arab World*. Prager, New York

- Entelis J P 1986 *Algeria: The Revolution Institutionalized*. Westview, Boulder, CO
- Esposito J L (ed.) 1983 *Voices of Resurgent Islam*. Oxford University Press, New York
- Friedman M 1992 *The Haredi (Ultra-Orthodox) Society: Sources, Trends and Processes*. Jerusalem Institute for Israel Studies, Jerusalem
- Heilman S 1992 *Defenders of the Faith: Inside Ultra-Orthodox Jewry*. Schocken, New York
- James W 1902 *The Varieties of Religious Experience*. Longmans, Green and Co., New York
- Juergensmeyer M 1993 *The New Cold War? Religious Nationalism Confronts the Secular State*. University of California Press, Berkeley, CA
- Kepel G 1985 *Muslim Extremism in Egypt: The Prophet and the Patriarch*. University of California, Berkeley, CA
- Lawrence B B 1989 *Defenders of God: The Fundamentalist Revolt against the Modern Age*, 1st edn. Harper & Row, San Francisco
- Lustick I S 1988 *For the Land and the Lord: Jewish Fundamentalism in Israel*. Council on Foreign Relations, New York
- Marsden G M 1980 *Fundamentalism and American Culture*. Oxford University Press, New York
- Martin D 1990 *Tongues of Fire: The Explosion of Protestantism in Latin America*. Blackwell, Oxford, UK
- Marty M E, Appleby R S 1991 *Fundamentalisms Observed*. University of Chicago Press, Chicago
- Marty M E, Appleby R S 1993a *Fundamentalisms and the State: Remaking Politics, Economies, and Militance*. University of Chicago Press, Chicago
- Marty M E, Appleby R S 1993b *Fundamentalisms and Society: Reclaiming the Sciences, the Family, and Education*. University of Chicago Press, Chicago
- Marty M E, Appleby R S 1994 *Accounting for Fundamentalisms: The Dynamic Character of Movements*. University of Chicago Press, Chicago
- Marty M E, Appleby R S 1995 *Fundamentalisms Comprehended*. University of Chicago Press, Chicago
- Mitchell R P 1969 *The Society of Muslim Brothers*. Oxford University Press, London
- Piscatori J 1991 *Islamic Fundamentalisms and the Gulf Crisis*. American Academy of Arts and Sciences, Chicago
- Rahman F 1982 *Islam and Modernity: Transformation of an Intellectual Tradition*. University of Chicago Press, Chicago
- Sivan E 1985 *Radical Islam*. Yale University Press, New Haven, CT
- Tibi B 1988 *The Crisis of Modern Islam*. University of Utah Press, Salt Lake City, UT

M. E. Marty

Religious Nationalism and the Secular State: Cultural Concerns

According to a widely held view, 'nationalism' is an aspect of the process of secularization, which limits the role of religious ideas and institutions in public life. Religious nationalism is therefore considered inimical to the 'secular state,' which is uninvolved with or neutral in religious matters. Empirical evidence how-

ever, reveals, that the relationship is complex and context sensitive. This article examines this complexity with special reference to South Asia where the two kinds of nationalism have long been in conflict and where both secular and religious states exist. Comparative perspectives from the West and the Middle East help frame the argument.

1. European Origins of Nationalism and its Diffusion

Although emphases differ, definitions of nationalism generally converge on the idea of a people's identification of themselves as a 'nation' that already is, or is expected to become, an autonomous state with its own territory. Moreover, the individual's emotional identification with the nation, and loyalty to the state, are of crucial importance. They find expression in the ideology of nationalism, which is a blueprint for the capture and use of political power for furthering common interests.

It has been argued that without nationalism there would be no nations or nation-states—that nationalism 'invents nations where they do not exist' (Gellner 1964, p. 169) and holds them together where they do. It is recalled that in Western Europe, the eighteenth century marked not only 'the dawn of the age of nationalism,' but also 'the dusk of religious modes of thought' (Anderson 1983). Ironically, secular nationalism everywhere, since the French Revolution, has displayed the characteristics of a religion both structurally (doctrine, myth, ritual, etc., are present) and functionally (an overarching moral order is provided). This is illustrated by the growth of 'civil religion' in America (Bellah 1970), transforming a historically developed national political system into a moral order.

In the articulation of nationalism, particular dimensions of collective identity (geographical, historical, cultural, or ethnic), or particular combinations of them, are highlighted. These emphases are variable not only from place to place, but also over time in the same place. One of the identity makers everywhere is religion, but religious nationalism is considered an anomaly, and not seen to be the most suitable cultural setting for a modern secular state in which individual identity is clothed in citizenship rights that override religious and ethnic distinctions.

One of the key questions about nationalism concerns its beginnings. Some scholars have argued for a Western European origin, whereas others maintain that England was the first modern nation ever. Still others consider North America to be its true place of efflorescence: the notion of the nation was implicit in that of 'the people' in the Declaration of Independence and is central to the American Constitution of 1789. Its inspiration is said to have been a unique combination of secularization, industrialization, capitalism, print technology, etc., that 'created the possibility

of a new form of imagined community, which in its basic morphology set the stage for the modern nation' (Anderson 1983, p. 49).

The Western origins and secular character of nationalism have been questioned. Kohn (1955) argued for Hebrew and Greek origins of the concept 2000 years ago, but affirmed that, as a fully fledged ideology, it came into play in England in the seventeenth century. Hastings has written that the Bible presented 'in Israel itself a developed model of what it means to be a nation—a unity of people, language, religion, territory, and government' (Hastings 1997, p. 18). For him the nation and its ideology are 'characteristically Christian things' (Hastings 1997, p. 186).

One of the many strands of Hastings' argument is that since Christianity did not have a sacred language, the availability of the Bible for translation and the activities of the local clergy in different parts of Europe resulted in the fostering of a 'sense of shared local, provincial, or national identity' (Hastings 1997, p. 192). He maintains that nations and nationalism outside the Christian world bear witness to the 'imitation of the Christian world,' although the mimesis may be regarded as westernization (Hastings 1997, p. 186). According to Hastings, while 'Christianity has of its nature been a shaper of nations, even of nationalism; Islam has not, being on the contrary quite profoundly anti-national' (Hastings 1997, p. 187). Orthodoxy notwithstanding, Muslims have embraced nationalism in many places in the twentieth century, e.g., in Turkey, where it appeared as a secular ideology in the 1920s, and in India, where religious nationalism became the basis for the partition of the subcontinent.

2. Secular and Religious Nationalism in South Asia

2.1 India, Pakistan and Bangladesh

For 100 years, beginning in the middle of the eighteenth century, British colonial territories in the Indian subcontinent expanded steadily through conquest and annexation. An image of Indians as a congeries of religious communities, castes, and tribes was assiduously cultivated. District gazetteers, compendiums of tribes and castes, and census reports recorded the divisions among the people. Such divisiveness was welcome from the colonial point of view as it helped strengthen the hold of the alien ruling power.

Ironically, however, British rule itself helped regenerate an Indian national consciousness, a prototype of which had been discernible in the heyday of the Mughal empire (seventeenth century). Establishment of law and order under the aegis of a centralized authority structure, in place of the chaotic conditions that characterized the disintegration of that empire in the first half of the eighteenth century, was one such

major factor. Also important were the introduction of railways, postal and telegraph services, education through the medium of English modeled on British schools and colleges, and newspapers in the nineteenth century. Resentment over the denigration of India's cultural heritage as 'an unmixed evil' (Henry Maine's phrase) by civil servants and Christian missionaries helped create a shared sense of common identity vis-à-vis the British. By the 1870s, 'a feeling of nationality,' in the words of a colonial magistrate, 'had sprung up in India' (see Mehrotra 1971, p. 144).

One of the main objectives of the Indian National Congress, established in 1885, was 'the fuller development and consolidation' of 'sentiments of national unity' (Mehrotra 1971, p. 413). The fact that it was attended by delegates from virtually all parts of the country, who included Hindus, Muslims, Parsis, and Christians, seemed credible evidence of (as a newspaper put it) 'the existence of something like a nation in India' (Mehrotra 1971, p. 419). Indeed, a year later, the highest British official in India, viceroy Dufferin, wrote of his sympathy for 'the aspirations of a nation so similarly circumstanced as my own [the Irish]' (Mehrotra 1971, p. 398).

The secular character of the Congress soon came under severe strain owing to the refusal of certain prominent Muslim leaders of northern India to join it. The position of the Congress was that 'the common benefit' of the people of India transcended the interests of any particular community, and that the two sets of interests were not necessarily in conflict. The emergence of secular nationalism had been preceded by the assertion of a sense of common nationality among Hindus, overriding significant regional and sectarian differences among them. Many of them brought residues of Hindu nationalism into the Congress strengthening Muslim fears.

Muslim separatism led to the formation of the Muslim League in 1906 with the aim of furthering the community's interests. In the course of time the viewpoints of the Congress and the League came to be designated as 'nationalism' and 'communalism,' respectively. Whereas the former considered territorial unity the crucial element of nationalism, the latter considered it irrelevant in Indian conditions and focused on religious identity.

The foregoing antagonism eventually led to the 'theory' that Hindus and Muslims were two nations, irreconcilably divided by history and culture, and each legitimately aspiring to be a nation-state. British rule ended in 1947 with the partition of the subcontinent into India and Pakistan. India created for itself a secular state in which citizenship rights stand above ascribed identities. Contrary to the stated secularist views of its founder, M. A. Jinnah (died 1948), Pakistan was proclaimed an Islamic state where traditional Islamic law would have precedence over other laws, and where Sunni Muslims emerged as the privileged community.

Pakistan was a unique state as it comprised two wings separated by 1,000 miles of Indian territory. The fragility of religious nationalism became obvious within a year of its founding in the context of a *lingua franca*. The people of East Pakistan, who accounted for more than half of the country's population, resisted the declaration of Urdu, the mother tongue of the urban Muslims of northern India, as the national language. Bengali was therefore accorded the status of the second language. Culture and language became rallying symbols for discontented East Pakistanis, who increasingly complained of economic exploitation and denial of political rights at the hands of a succession of 'central' governments that had their seat in West Pakistan and were dominated by the natives of that wing.

Eventually, Pakistan split in 1971, and a new nation-state, Bangladesh, came into existence. Bangladesh adopted a constitution that was secular in character. The original considerations that legitimized partition in 1947 survive, however, and are exploited by some political parties. Islamization of Bangladesh's national culture, and the constitutional redefinition of the state as one that derives its legitimacy from Allah, emphasize the obstacles that confront the secular state in South Asia.

It has been argued that the enormous cultural and religious diversities of India necessitate a pluralist ethos and the secular state, and indeed guarantee them. The Constitution (Article 25) grants every citizen the fundamental right to profess, practice, and propagate the religion of her or his choice. It also prohibits discrimination between one citizen and another on the basis of religion. All major political parties that have held power at the state or union (federal) levels uphold the concept of the secular state. One of them, the Bharatiya Janata Party (BJP), is widely considered communal in view of its following, which is virtually exclusively Hindu, and its adherence to the ideology of Hindutva (lit. 'Hinduness'). Originally put forward in 1923 (BJP itself is a post-independence party), Hindutva equates Indian identity with Hindu identity, and defines the latter in terms of a combination of territorial and cultural criteria (fatherland and holy land should be one). The ideology points to political domination through cultural homogenization, and is hostile to pluralism. BJP emerged as the leading partner in the coalition union government that was formed in 1998. The exigencies of coalition politics have resulted in the de-emphasizing of Hindutva but not its abandonment by BJP. The ideology is synonymous with religious nationalism and therefore a threat to the secular state.

2.2 Sri Lanka

The Sri Lankan Constitution also is broadly secular in character. Although a 1972 constitutional amendment

recognized 'the foremost place' of Buddhism as the religion of the majority, there is no state religion, nor discrimination in respect of citizenship rights between the followers of different faiths (Buddhism, Hinduism, Islam, and Christianity). The fact that the ethnic Sinhalese constitute 70% of the population and are Buddhists has been a source of communal strife and has imposed severe pressures on the state.

Beginning in 1953, when the call for replacing the secular state by a Buddhist one was first made, politically organized Sinhalese have carried on campaigns against the Tamils of the northern and north-eastern regions of the country (mainly Hindus and Christians) for their secessionist ambitions (the Tamils in turn complain of Sinhalese domination) and against a succession of governments for not being sufficiently responsive to Sinhalese aspirations. Prime Minister Solomon Bandaranaike, a convert from Christianity to Buddhism and more sympathetic to such aspirations than his predecessors, was nevertheless assassinated in 1959 by a Buddhist monk. Altogether the Sinhalese nationalists won many concessions during the 1960s and 1970s.

A watershed in the ethnic strife was the countrywide anti-Tamil upheaval of 1983 that had its epicenter in Colombo. It followed the killing of a dozen Sinhalese soldiers by Tamil militants. By official count several hundred Tamils were killed; the Tamils claimed that thousands died. Sinhalese militancy represented by Janatha Vimukhti Peramuna (JVP; People's Liberation Front) was eventually brought under control in 1989–90 by President Premadasa (himself a Buddhist) through accommodation of sections of the Buddhist leadership and ruthless suppression of JVP cadres. But he, too, paid with his life for his resistance to the secessionist movement led by the Liberation Tigers of Tamil Eelam (Nation).

3. Comparative Perspectives

The conflict between religious nationalism and the secular state is vividly illustrated by the wide range of events in South Asia in the second half of the twentieth century. Religious nationalisms are present elsewhere too in confrontation with a secular state or in association with a religious state that is modern in some of its objectives (economic development, political stability) and methods (elective government).

The Iranian revolution of 1978 established an Islamic state after overthrowing the modernist regime of the Shah. The earlier ouster of King Farooq of Egypt in 1952 had been followed by the establishment of a secular government under Gamel Nasser, which set out to suppress religious nationalist groups such as the Islamic Brotherhood. Although his successor, Anwar Sadat, made many concessions in their favor,

he died at the hands of Muslim extremists. Religious fundamentalism and nationalism are a considerable combined force, but President Mubarak has so far succeeded in holding it at bay. In Turkey, too, the secular state founded by Kemal Ataturk remains intact but religious revival has made significant inroads into Turkish society and nationalism. Although the patterns vary, religious nationalism today is a force to reckon with among the Muslim nations.

Religious nationalism played a significant role in the disintegration of the former USSR and the emancipation of East European countries from totalitarian socialist regimes. It obviously is not an essentially reactionary force in all circumstances. The comparative perspective brings into question many assumptions about the western origins and secular foundations of nationalism. Religious nationalism is a mixture of diverse elements, not all of which are present everywhere. Some of them are in principle unexceptionable, such as the emphasis on the moral dimensions of political behavior. The overall record of religious nationalists, however, is marked by their opposition to pluralism, which is one of the primary conditions of humane governance, particularly in ethnically plural societies.

See also: Religion: Nationalism and Identity; Secular Religions; Secularization

Bibliography

- Ahmed R (ed.) 1991 *Religion, Nationalism and Politics in Bangladesh*. South Asian Publishers, New Delhi, India
- Anderson B 1983 *Imagined Communities, Reflections on the Origins and Spread of Nationalism*. Verso, London
- Bellah R N 1970 *Beyond Belief: Essays on Religion in a Post-traditional World*. Harper and Row, New York
- Chandra B 1984 *Communalism in Modern India*. Vikas, New Delhi, India
- Gellner E 1964 *Thought and Change*. University of Chicago Press, Chicago
- Gellner E 1983 *Nations and Nationalism*. Cornell University Press, NY
- Hastings A 1997 *The Construction of Nationhood, Ethnicity, Religion and Nationalism*. Cambridge University Press, New York
- Hobsbawm E J 1990 *Nations and Nationalism Since 1780*. Cambridge University Press, Cambridge, UK
- Jaffrelot C 1996 *The Hindu Nationalist Movement in India*. Columbia University Press, New York
- Juergensmeyer M 1993 *The New Cold War? Religious Nationalism Confronts the Secular State*. University of California Press, Berkeley, CA
- Kohn H 1955 *Nationalism: Its Meaning and History*. Van Nostrand, Princeton, NJ
- Mehrotra S R 1971 *The Emergence of the Indian National Congress*. Vikas, New Delhi, India

Tambiah S J 1992 *Buddhism Betrayed? Religion, Politics, and Violence in Sri Lanka*. University of Chicago Press, Chicago, IL

T. N. Madan

Religious Organizations

Religious organizations seek to establish, promote, and regulate relations between human beings and divinities, supernatural orders, or supreme metaphysical principles. This article explains why religious organizations have been distinguished from non-religious organizations and why their distinctiveness lies mainly in their varied patterns of authority and leadership.

1. Diversity

Religious organizations are immensely varied but they typically aim to promote worship, prayer, meditation, teaching, healing, and spiritual well-being in accordance with authoritative revelations, texts, codes, laws, and principles. They range from the groups of 'clients' who regularly consult self-appointed healers, gurus, and shamans to worldwide ecclesiastical bureaucracies such as the Roman Catholic Church. Informal and irregular religious practices can be nested inside highly formal religious organizations.

The distinctive 'products' of religious organizations are equally diverse: sacred knowledge, transcendental experiences, prophecies, worship and meditation, rituals with the power to cleanse or heal, rites of passage, religious ethics, recruiting missions, political, welfare, and educational outreach, local fellowship, and so on. They also have much in common with nonreligious organizations in so far as they face the need to secure adequate resources, to train and control staff, to retain their authority, to cope with dissent and conflict, to interact with other organizations in their environment, to defend their share of the religious 'market,' and to deter 'free riders,' and unwelcome takeovers. Religious organizations have therefore established schools, seminaries, hospitals, welfare agencies, publishing enterprises, and missionary orders in order to extend and protect their boundaries.

2. History

The history of studies of religious organizations can be divided into two streams. The first is concerned primarily with the distinctiveness of religious organizations and with their subtypes. The second, which has flourished since the 1950s, deals with the internal dynamics and shifting fortunes of religious organizations in their social environments.

2.1 Types and Subtypes

The intellectual rationale for the pioneering and highly influential studies written by Max Weber (1864–1920) and Ernst Troeltsch (1865–1923) had its roots in their respective interests in the interplay between modernity and religion. Having established the crucial but unintended contributions of Ancient Judaism and Christianity towards the forms of modern rationalization, Weber analyzed the ironic implications of rationalization for the differentiation of religion from other institutions and value spheres. Secularization, bureaucratization, and the disenchantment of the modern world were all shown to have reduced the cultural and social significance of religious organizations. Weber perceived clearly that church-type organizations (large, territorially based, and inclusive in membership terms) had the potential to exercise power but were constrained by their ties to the secular world from translating their Christian ethics into uncompromising programs of action. By contrast, sect-type organizations (voluntaristic, relatively small, and exclusive in membership terms) were capable of maintaining a distinctive and often radical version of Christian theology and ethics but were prevented from exercising much influence on the world around them because they distanced themselves from it.

Troeltsch was less preoccupied with rationalization as a process and more concerned with the historical evolution of the social forms of 'church,' 'sect,' and 'mysticism' in which the New Testament's ethical message had found expression. Troeltsch was particularly exercised by the question of whether liberal Protestantism was robust enough and sufficiently well adapted to the conditions of modernity to serve as an effective vehicle for Christian ethics in an increasingly secular world.

The concepts of 'church' and 'sect,' plus numerous variants and subtypes such as 'established sect' and 'denomination,' have played an important part in many attempts to explain patterns of change in Christian organizations throughout the twentieth century. Some scholars have tried to apply these concepts to the analysis of religions other than Christianity, but it remains questionable whether this is appropriate in the case of religions lacking strong notions of membership and corporate authority. Yet, refinements of the sect concept have undoubtedly helped to explain the highly varied developments of minority religious movements, not only in Western societies but also in East Asia. Bryan Wilson's (1970) scheme of seven subtypes of sects has proved particularly fruitful and influential. The concept of 'cult' has also undergone extensive refinement, coinciding initially with public interest in the new religious movements which became controversial in many Western countries and Japan in the 1970s.

A further impetus for sociological studies of 'sects' and 'cults' began to develop in the 1980s when

proponents of rational choice theory claimed that previous usage of these concepts has been largely *ad hoc* and untheorized (Iannaccone 1988). On the assumption that people rationally try to attain their ends by the least costly means, in religion as in any other area of life, the theory holds that people incur the costs of participating in religious organizations in order to derive supernatural compensators which purport to give answers to questions of ultimate meaning. In this theoretical perspective, churches are believed to offer answers which conform with prevailing norms, whereas sects and cults allegedly deviate from the norms and incur sanctions or costs which are nevertheless experienced by their followers as selective benefits unavailable to members of 'lax' churches. The historical formation of sects (by schism) and cults (by innovation) can be accounted for in rational choice terms as a never-ending process of competition between religious organizations standing in different degrees of tension with society.

By comparison, the sociological concept of 'church' underwent relatively little change in the twentieth century. Moreover, little interest has been shown in refining or even applying the kind of cognate concepts of 'ecclesiola in ecclesia,' 'fraternitas,' and 'order' that Joachim Wach (1898–1955) and others elaborated. Nevertheless, studies of Christian organizations in Africa, Asia, South America, and the former Soviet Union have brought to light the importance of such organizational forms as 'independent churches,' 'base ecclesial communities,' and 'underground churches.' The continuing significance of relations between religious organizations, nations, and states in some countries has also preserved a place for 'church' in analyses of, for example, civil society (Casanova 1994) or international relations (Johnston and Sampson 1995). And extensive research in North America has thrown light on numerous aspects of 'denominations' as distinctive organizations. Beginning with the embeddedness of denominations in social classes (Niebuhr 1929) and ethnic groups, studies have analyzed the organizational conditions of schism, the ambiguous position of theological seminaries, the differential responsiveness to female clergy, the increasingly complex webs of 'interorganizational fields' in which denominations operate, and tensions between religious hierarchies and functional agencies. Another rich vein of research has investigated the contributions of denominations towards ideologies of pluralism, pluralistic provision of social welfare, social movements, and the 'independent sector.'

The study of religious organizations has been understandably less important in relation to religious traditions which tend to lack the formality, corporate authority structures, and membership principles characteristic of Christianity. Nevertheless, common foci of scholarly interest have emerged in the study of Muslim 'orders,' brotherhoods, and charitable endowments (*waqf*); Buddhist monastic institutions

(*sangha*); Hindu devotional groups (*bhakti*) and nationalist organizations; Jewish and Sikh communal organizations; and Shinto parishes.

2.2 Dynamics of Religious Organizations

The emerging concern with the organizational dynamics of religion in the 1950s was partly a reflection of the growing salience of 'organization theory' in the burgeoning social sciences, and partly a function of Christian theological interests in 'ministry' as a modern profession and in churches as complex organizations responding to rapid social change (Harrison 1959, Winter 1968, Beckford 1975). Yet the impact of these analyses of religious organizations on the development of organizational theories in general was modest. This state of affairs began to change slightly in the mid-1970s when notions such as 'organizational climate' and 'organizational culture' directed attention towards the significance of cultural factors which had always loomed large in analyses of religious organizations but which were emerging as important influences in virtually all kinds of organization. Investigations of communes gave a further impetus to research into the force of strong commitments and normative beliefs among organizational participants. Subsequent discoveries about the importance of rituals, retreats, flexibility, and networking in many organizations have only enhanced the interest that religious organizations now have for some organizational analysts (DiMaggio 1998). This minor rapprochement between the study of formal organizations and the study of religious organizations received a further boost in the 1980s when so-called New Age and quasireligious business began to attract attention for their practice of combining profit seeking with ethical business practices and of contributing towards the 'spiritual development' of their staff and clients. And fresh insights into the tensions between the 'religious' and 'agency' structure of US denominations (Chaves 1993) offer the prospect of explaining comparable tensions between the normative and functional bases for coordination and compliance in other organizations as well.

Two inter-related aspects of religious organizations stand out as the central foci for past and present research: authority and leadership. This does not mean that issues of organizational effectiveness, efficiency, and complexity are irrelevant in the case of religious organizations. It simply means that organizations purveying allegedly timeless truths and the highest normative commitments tend to be relatively more preoccupied than nonreligious organizations with establishing, reproducing, and defending their teachings and practices against corruption, subversion, or apathy.

3. Authority

Like profit-seeking organizations, religious organizations are engaged in competition with purveyors of rival versions of ultimate truths and in conflict with opponents, but they are also obliged to monitor and police their own members or followers. This is why questions of authority are central to religious organizations and why change often meets with strong resistance. In this sense, religious organizations are distinctive, if not unique, for the extent to which the connections between their founding truths and their organizational structure and everyday activities are constantly rehearsed in rituals, symbols, exegesis of sacred texts, preaching, prayers, teaching, meditation, and the training of leaders.

3.1 Sources of Authority

Christianity is virtually unique among the world's religions for the extent to which its practice has been controlled by centralized, formal organizations cutting across ethnic, tribal, and political groupings. By contrast, Hinduism, Judaism, and Shinto crystallized slowly out of predominantly tribal and ethnic cultures, giving rise to temples, synagogues, and shrines as centers of rites and ritual worship rather than as agencies of civil administration. Diasporic Judaism, lacking the authority previously vested in the Temple at Jerusalem and its hereditary priests, regrouped around rabbis as teachers of the law and around synagogues as ritual centers. Communal organizations have played a major role in preserving the ethnic and religious identity of Jews under persecution, but centralized organizations to control Jewish beliefs and rituals are relatively weak.

Buddhism and Islam emerged from more deliberate attempts to cultivate religious and philosophical systems partly at odds with prevailing culture and social arrangements, and this feature of their historical emergence may help to account for the fact that they have been closely allied with political regimes at various times. Thailand and Japan, for example, have had something approaching Buddhist parish systems. Modern Japan has also seen the rise of massive, hierarchical organizations of Buddhist laypeople such as Risshō Kōsei-kai and Sōka Gakkai. But elsewhere in Asia, Buddhism functions like a diffuse folk religion supported by monastic institutions and temples.

The organization of Islam is subtle, multistranded and variable by region of the world. As a conversion faith Islam has been interwoven with political regimes, but it does not have a single, authoritative organization. There are multiple, competing sources of theological, legal, spiritual, and moral authority in Islam, for the authority of learned experts, jurists, hereditary sheikhs, leaders of brotherhoods and clergy (among the Shi'ites) are all part of the complex

organizational tapestry of Islam. This high degree of complexity makes it difficult for any single organization to claim convincingly that it represents all Muslims. Now that tens of millions of Muslims live as religious minorities in the predominantly Christian countries of Europe and North America, pressure may be increasing for would-be representative organizations to seek to represent all Muslims in their collective relations with nation states, nongovernmental organizations, and the European Union.

3.2 Forms of Polity

The variety of organizational forms and processes which translate these concerns with authority into action is bewilderingly wide. They range from loosely structured fellowships to vastly complex and formal organizations which require their members to accept centrally imposed dogmas and rituals mediated by authoritative officials. In practice, most religious organizations fall somewhere on a continuum between these two extreme positions, but their precise location may change in response to factors internal or external to the organizations. Moreover, the time-honored distinction between church- and sect-type organizations cannot be mapped on to the distinction between religious organizations with either strong or weak systems of authority.

The convention of classifying the polity of Christian churches as Episcopal, Presbyterian, or congregational relates to the distribution of authority as laid down in their respective theologies and legal constitutions. In particular, the distinction between ordained clergy and laity is given theological justification, and the differential authority accorded to different grades of religious professionals is explained in the same terms. Yet this classification can only be a starting-point for social scientific analysis, since the readiness of subordinates to comply with authoritative commands from their superiors is variable, and pressure from outside the churches may make it necessary to modify the practical implementation of the ecclesiastical principle. For example, the Episcopal polity of churches in the worldwide Anglican communion conceals a wide diversity of relations between bishops and other church members. Presbyterian and congregational polities have also come under pressure to adapt the exercise of authority to prevailing socio-cultural values and conditions. Thus, critics of the patriarchalism allegedly underpinning many forms of Christian organizational polity have challenged the gendered distribution of authority in churches and tried to infuse them with feminist, antihierarchical values (Wallace 1992, Nesbitt 1997). Moreover, the ideal of congregational autonomy finds itself in tension with political or economic pressures to extend the authority of regional or national officials.

3.3 Fundamentalism

The increasing popularity and influence of so-called fundamentalist forms of religion roughly since the 1950s in many parts of the world and in most of the major faith traditions has prompted social scientists to pay greater attention to religious authority (Marty and Appleby 1991). Fundamentalism centers on faith in the authority of certain principles, usually enshrined in sacred texts, to generate unequivocal guidance for the conduct of personal and public life, and it favors assertive leadership and a strong suspicion of anything outside the group of committed believers. Yet the organizational bases of fundamentalism are diverse and subtle. There is no single, monolithic organization in any faith tradition or country which promotes fundamentalism. Instead, it has spread by personal witness, small group discussion, and the use of computerized mailing lists, terrestrial and satellite television and radio broadcasts, video cassettes, email, and the World Wide Web. Fundamentalists utilize modern technology and organizational forms in order to transform the world, and are not traditionalists, 'antimodern,' or withdrawn from the world. On the contrary, their aim is to make the world conform with their authoritative principles, and this calls for forms of organization which are flexible and responsive to local circumstances.

4. Leadership

4.1 Roles

Leadership roles, structures, and training vary widely between faith traditions, between their different 'schools' or denominations and between different parts of the world. The precise pattern of leadership at any level is shaped mainly by prevailing ideas about the source and nature of sacred power. Leaders of religions which identify sacred power exclusively with scriptures and laws derive their authority mainly from the capacity to read and interpret them. Judaism and Islam exemplify this pattern of leadership by scholars, teachers, and jurists. Priesthood, by comparison, is found in religions such as Hinduism, Shinto, Christianity, and Sikhism which combine reverence for sacred texts with belief in the capacity of priests to perform rituals allowing humans to interact in various ways with divinities and sacred power. The office of a priest enables this power to be effective through properly conducted rituals. Prophets and charismatic religious leaders may reform religion by opposing institutionalized practices and beliefs with radically new or different ones. Ministers of religion in most of the Protestant traditions offer leadership mainly through their preaching, exposition of scripture and, at least in evangelicalism, efforts to persuade individuals of their sinfulness and urgent need to seek

redemption. Monks, nuns and 'holy men,' as the 'virtuosi' in some religious traditions, exercise leadership on the strength of their extraordinary discipline, devotion, and selflessness. The *sangha*, or monastic community, remains at the heart of Buddhism in many parts of the world; monastic institutions were the bulwark of Christianity throughout the European Middle Ages; Catholic religious orders were in the forefront of missionary work in Africa, Asia, and South America in the modern era; and monks continue to occupy the highest leadership positions in Orthodox churches. But religious traditions which locate authority in the spiritual freedom of individuals to cultivate their own, relatively independent path to the sacred tend to resist strong notions of leadership, preferring at best to have spiritual exemplars, teachers, or gurus.

4.2 Local Groups

It is common for religious leaders to be attached to a particular building or social center where they conduct teaching, preaching, meditation, workshop, prayer, or pastoral functions. Some also have responsibility for the spiritual oversight of institutions such as schools, seminaries, orphanages, hospitals, refuges, and pilgrimage sites. Christianity went further than other religious traditions in establishing a comprehensive system for delivering religious and pastoral services to geographically demarcated areas known as parishes, which represented subdivisions of national or international churches. The idea that the residents of a locality would belong to a parish and form a congregation of regular worshipers under the leadership of local clergy was especially effective in European regions where the dominant church was closely aligned with political regimes.

4.3 Congregations

Congregations are the basic unit of most Christian organizations in the sense of being the social group with which their members in particular localities feel most closely associated. The precise status of congregations varies with the theology and authority structure of the larger religious organizations to which they belong. The congregations of hierarchical churches are local outlets of ideas and practices controlled by regional, national, and international authorities such as bishops, cardinals, and the Pope. By contrast, religious organizations which have little or no theological justification for hierarchy or central control tend to regard local congregations as relatively autonomous units. This is common among, for example, independent fellowships of conservative Protestants, the Reform Synagogues of Great Britain, and the Christadelphians. The implications of the balance that Christian organizations achieve between centrali-

zation and local autonomy have been extensively researched because it reveals differences in the extent to which congregations can generate resources, mobilize members in campaigns, respond to internal problems, and influence public life (Hong and Roozen 1979).

As Hinduism, Islam, Judaism, and Sikhism are not structured on the basis of inclusive, unitary organizations at national or international levels they have not generated congregations or systems of parishes. Yet, temples, mosques, synagogues, and gurdwaras do function as ritual and pastoral centers for people living in their vicinity or choosing to visit them. Lacking strong integration into higher level organizations, however, these local centers are accountable either to their private owners or to their management committees. Consequently, their religious leaders are retained as professionals or qualified volunteers by local groups rather than as employees of ecclesiastical organizations at a higher level.

4.4 Recruitment and Training

In advanced industrial societies the attractions of work in positions of religious leadership are weakening, and the rate of applications for training as religious professionals has been declining since the middle of the twentieth century. At the same time, the characteristics of recruits to religious leadership in Christian churches have also changed. They now include a much greater proportion of women, of people over the age of 30, of people retiring early from other occupations, and of people from a wider range of ethnic and social class backgrounds. The 'feminization' of religious leadership is occurring in many churches in so far as the proportion of women entering theological training and Christian ministry has been rising steadily since the 1960s (Wallace 1992). Yet the upper echelons of most churches are still controlled by men, and there is evidence that various mechanisms still prevent women from enjoying genuinely equal opportunities for professional advancement. Women are even more strongly excluded from leadership roles in the other world religions, although they have been admitted to the rabbinate in some traditions of Judaism.

It is only in developing countries that there is still a relatively strong supply of young men and women for theological training and preparation for the ministry or monastic vows. The disproportionately large numbers of trainee priests and ministers originating in developing countries may mean that the organizational 'center of gravity' of some large Christian organizations is in process of shifting away from Europe and North America. In effect, the new international division of labor is affecting religious organizations as well as transnational businesses.

5. Questions About the Future

Three emerging trends in religious organizations call for special comment. The first concern settlers in countries of large-scale immigration who seek to establish their own faith communities or congregations (see Warner and Wittner 1998). How far will these communities conform with prevailing forms of religious organization in their new setting? How far will their forms of organization come under pressure to meet the criteria laid down by stage agencies if they wish to deliver health, welfare, or educational services? And to what extent will their forms of organization influence longer established religious groups?

The second concerns the effect of new information technologies on the capacity of religious organizations to extend their influence to virtually all parts of the world. Will satellite broadcasting, video recordings, and the Internet increase competition between religious organizations? And will they cultivate a globally standardized form of these organizations?

The third emerging trend is for the appeal of 'special purpose' agencies to cut across the boundaries between Christian congregations and between denominations in the USA, thereby making ideological differences *within* religious organizations more challenging than the differences *between* these organizations (see Wuthnow 1988). Is there evidence of this kind of restructuring in other world religions? Does this repositioning of organizational 'fault lines' facilitate or hinder transnational religious developments?

See also: Civil Religion; Education; Cultural and Religious Concepts; Religion: Definition and Explanation; Religion: Evolution and Development; Religion: Morality and Social Control; Religion, Sociology of

Bibliography

- Beckford J A 1975 *Religious Organization*. Mouton, The Hague
- Casanova J 1994 *Public Religions in the Modern World*. University of Chicago Press, Chicago
- Chaves M 1993 Intraorganization power and internal secularization in Protestant denominations. *American Journal of Sociology* 99: 1-48
- DiMaggio P 1998 The relevance of organization theory to the study of religion. In: Demerath N J III, Hall P D, Schmitt T, Williams R H (eds.) *Sacred Companies. Organizational Aspects of Religion and Religious Aspects of Organizations*. Oxford University Press, New York
- Harrison P M 1959 *Authority and Power in the Free Church Tradition*. Princeton University Press, Princeton, NJ
- Hoge D R, Roozen D A (eds.) 1979 *Understanding Church Growth and Decline 1950-1978*. Pilgrim Press, New York
- Iannaccone L R 1988 A formal model of church and sect. *American Journal of Sociology* 94: 241-68
- Johnston D, Sampson C (eds.) 1995 *Religion, the Missing Dimension of Statecraft*. Oxford University Press, New York

- Marty M E, Appleby R S (eds.) 1991 *Fundamentalisms Observed*. University of Chicago Press, Chicago
- Nesbitt P D 1997 *The Feminization of the Clergy in America*. Oxford University Press, New York
- Niebuhr H R 1929 *The Social Sources of Denominationalism*. Holt, New York
- Wallace R A 1992 *They Call Her Pastor*. SUNY Press, Albany, NY
- Warner R S, Wittner J (eds.) 1998 *Gatherings in Diaspora. Religious Communities and the New Immigration*. Temple University Press, Philadelphia, PA
- Wilson B R 1970 *Religious Sects*. Weidenfeld and Nicolson, London
- Winter G 1968 *Religious Identity*. Macmillan, New York
- Wuthnow R 1988 *The Restructuring of American Religion*. Princeton University Press, Princeton, NJ

J. A. Beckford

Religious Specialists

'Religious specialists' is a generic term referring to select individuals or a category of individuals within different world faiths who undertake a particular responsibility for religious functions such as teaching, contemplation, transmission of scriptures, leadership, pastoral care, or ritual, on behalf of the religious community they serve. The nomenclature used to describe religious specialists differs according to the various religious traditions, but the idea of prophets, teachers, a priesthood and/or contemplative monastic orders is common to many world religions. Most religious specialists either will have inherited their role or will have responded to a sense of Divine calling to a particular and often strict way of life and adherence to religious disciplines.

1. The Scope of Different Religious Specialists

A community of people who share a particular ontological understanding of the world have often been inspired by prophets or religious teachers who have acted as intermediaries between the Divine and earthly realms. The prophetic role has often involved receiving revelations, and acting as an exemplar of these revelations. As religious communities grow in size numerically and geographically, and become increasingly complex, a need arises for individuals with particular knowledge and authority to oversee the religious belief and practice of the group. The need for a range of different religious specialists may become apparent, e.g., to regulate religious activities, to preserve and recite scriptures, or to lead worship, and in time, a hierarchy of different specialists with a specific authority and leadership may emerge. Access to particularly significant holy sites and shrines, or sacred texts, may be limited in some traditions to

religious specialists only. In many ways, prophets are the direct opposite of priests, since the former are often a force for change in society, challenging existing practices and beliefs, while the latter serve as representatives of a religious community and seek to maintain existing practices. Charismatic religious specialists in some ways stand between prophets and priests, since they are often the source of new ideas and practices, but within existing frameworks of community and belief.

In the major monotheistic religions of the world today, Judaism, Christianity, and Islam, there is a range of religious functionaries. Despite the common link between these three faiths, there are some significant differences in the various types of religious specialists that have emerged in each. In Judaism, a rabbi does not, strictly speaking, have a priestly role, but rather a responsibility for teaching and spiritual guidance. Historically, alongside the Hebrew prophets, were sages or wise men whose writings comprise, among others, the Book of Proverbs.

In Christianity, there are two principal avenues for professional religious activity, namely pietistic contemplative, and priestly. Amid the diversity of Christian traditions, the Orthodox, Catholic, and Anglican denominations in particular have a history of contemplative monastic orders for both men and women. These orders have often been inspired by the examples of saints such as St. Francis of Assisi, and members of these religious communities aspire to follow the exceptional qualities of their founders. Other religious specialists in Christianity include priests or clergy whose role often combines a range of functions, such as pastoral care, leading worship, and teaching, generally focused around a church or place of worship.

In Islam, there is no such priestly equivalent, but there is nevertheless a range of religious specialists. An imam may be regarded primarily as a ritual specialist with particular responsibility for leading congregational prayers in a mosque. The title also has honorific connotations, perhaps in denoting the leader of a particular community or group. Among Shi'ite Muslims, the term imam has yet another meaning as an intercessor between Allah and humanity. The framework for Islamic life and conduct, individually and communally, is shaped by the Shari'ah, or Islamic law. Scholars with a particular knowledge of Islamic law and teachings, known collectively as ulema (singular alim) assist with the development of Islamic thought, while muftis are legal functionaries empowered to make fatwas or legal rulings. In the Sufi or mystical tradition of Islam, the spiritual leader of a community is often known as a sheikh.

Even between the three major monotheistic traditions, there is a clear diversity in the type and remit of different religious specialists.

Indian religions, especially Hinduism and Sikhism, share the concept of a guru or teacher. Sikhism is

founded upon the teachings of 10 successive Punjabi gurus during the sixteenth and seventeenth centuries, while in Hinduism, the term guru has a more generic connotation as a religious teacher who gives personal instruction to devotees. Brahmins are at the top of the Hindu caste system, and as religious professionals their role principally involves the transmission of sacred Sanskrit texts and the performance of priestly sacrificial rituals.

The monastic life is perhaps the defining feature of religious specialists in the many traditions of Buddhism, whether Tibetan, Zen, Theravadin, or Japanese. Not surprisingly, therefore, many of the titles associated with Buddhist religious functionaries relate to the hierarchy of the monastic order. With Buddhism becoming increasingly popular in the West, different titles such as Ajahn, Bhikkhu, Roshi, Lama, and Rinpoche are becoming more widely known and understood.

In traditional societies and communities, shamans are religious specialists with the supernatural power to contact the spirit world during ecstatic trance states, often induced through dance, intoxicants, drumming, or chanting. They are not priests, nor medicine men, but are said to have the ability to foresee and predict the future, to quell the forces of the spirit world, and undertake extreme privations without feeling pain.

Despite the wide variations among religions in terms of religious functionaries and their various roles, there are some common sociological features of professional or specialist religious activity.

2. *Becoming a Religious Specialist*

Those who dedicate either all or part of their lives to religious work are often regarded as having a vocation rather than a job. Social scientists are not so concerned with the theological or supernatural implications of vocation, but rather with visible and measurable aspects of professional religious life, such as recruitment and training. In his study of American Catholic clergy, Fichter notes that 'the ecclesiastical vocation is not only a supernatural calling, a state of life and social status. It is also a specialised social role through which the manifold functions of the Church are performed. Obviously then, he (the priest) is engaged in a full-time profession or occupation which can be submitted to sociological comparison and analysis' (Fichter 1961).

2.1 *Recruitment and Training*

Personal connections and example often have a bearing on recruitment to professional religious life. This partially explains why vocations often cluster in families and kinship groups. In some world religions, religious roles are inherited through a lineage of teachers or leaders. Entering professional religious service is a lifelong commitment in many instances,

while in others an individual may through tradition or personal choice undertake to become a religious specialist for only a short period of time. A good example of the latter can be found in the Theravadin Buddhist tradition where it is common for young boys of school age to be ordained as monks for a limited period during their youth.

Most religious traditions have well-established and well-known expectations and qualifications for people interested in becoming religious functionaries, either in terms of knowledge or ethical conduct. As a step towards learning to become a religious specialist, it is common to undergo a period of isolation, training, and/or renouncing of worldly attachments. In both Buddhism and Christianity, those wishing to become monks or nuns often undertake a trial period of training as novices prior to ordination. In religious traditions where scholarly knowledge is a particular feature of professional religious activity, new recruits will often undertake an extensive period of specialist training to acquire the necessary knowledge and skills. Therefore, theological colleges or seminaries are common to the monotheistic traditions. Training institutions may be geographically isolated, or may be fully integrated into local communities, but the shift towards the latter as the locus for professional religious development is becoming more evident. This is partly due to the closing of the gap between lay communities and religious professionals, and the concern that communities have for specialists who are relevant and effective in modern societies.

2.2 *Ordination and Other Rites into Professional Religious Life*

Having successfully completed the necessary training period, whatever form that may take, some form of ceremony, initiation, consecration, vow-taking, or ordination into the religious profession is common. A number of symbolic actions or practices may take place, most of which have the function of setting apart the new religious specialist from the lay community. For example, the wearing of religious garb, such as a monastic habit or robe, or the marking of the body in some way, serves both ecological and structural purposes in 'locating' the individual within a community of fellow specialists. At the same time, a new name or religious title may be taken, and in some religious traditions, kinship terminology may be prevalent, such as 'Brother' or 'Sister.' These actions and rituals are part of the process of transformation into a socially recognized religious specialist.

3. *Religious Specialists—Some Sociological Characteristics*

Despite becoming socially recognized religious adepts, in many traditions it is common for religious specialists

to have a spiritual guide or teacher. This is a reflection of the life-long commitment that individuals make to their vocation, and the pursuit of spiritual development and perfection. The balance between, or combination of, self-centered and socially-centered spirituality and action will depend upon the characteristics of the professional role. Often there is no such thing as 'retirement' for religious professionals; they may retire from active functions certainly, but the nature of their calling entails a continuation of the religious role until death.

3.1 Status and Prestige

Most religious professionals share the social status of the group to which they belong. For example, the prestige and social standing of religious specialists in immigrant minority groups tends to rise or fall as a reflection of the upward or downward mobility of the migrant group. However, religious specialists who have subsidiary secular functions, such as social work or healthcare, may acquire recognition and status through these roles, rather than in their primary religious capacity. In some senses, religious specialists are classless individuals since generally they do not acquire status through earning nor, in the case of celibates, through favourable marriages. This does not mean, however, that they lack the possibility of status, since they can acquire other forms of recognition and prestige such as authority and other attributes of stature and power.

3.2 Professional Qualities and Hierarchies

Religious specialists can be considered as 'professionals' on a number of grounds. They are involved in a nonstandard product—religion. They undertake tasks and services that often cannot be repeated exactly, their work involves personal commitment and dedication, and they are required to have a broad knowledge of a specialized field and have authority both inside and outside the religious community. As specialists, they are often aspiring to personal development and improvement, and generally they share solidarity with other religious professionals. Like most professional groups, religious specialists, particularly in community contexts, are part of a hierarchical structure. Age, length of service, or the responsibilities particular to a role or office often determine hierarchy. The specialists who also have a leadership function face particular challenges in securing conformity, maintaining authority, and religious management of internal and external goals. The most senior members of the hierarchy generally have responsibilities that extend beyond the immediate community to a wide geographical area.

3.3 The Proliferation of Religious Specialists

In modern secular societies there has been a proliferation of work specialization. The study of careers in themselves has become a further branch of knowledge and expertise. The religious professional is yet another specialized career, and within it, there are further specialist activities, such as teaching, missionary work, or care of the sick. For example, some clergy within the Christian church feel called to chaplaincy roles often requiring a particular concern for groups of people in some way set apart from the rest of society, e.g., through imprisonment or hospitalization. The world of chaplaincy has in turn become more complex as society has become more fragmented and 'full-time personnel of the Church have been carried along on a wave of extra-ecclesiastical professionalism' (Fichter 1961).

In most cases, religious specialists are regarded as having a positive influence upon religious communities and society, in as much as their example and ministry inspires action and behavior that is socially beneficial. Most religious specialists are associated with having outstanding personal qualities of kindness, patience, tolerance, concern, holiness, devotion, sensitivity, intelligence, and wisdom. However, amid the proliferation of new religious groups and sects—most requiring leadership—some religious specialists have been so idealized, sometimes through their deliberate flouting of conventional norms and the employment of bizarre rituals, that their humanity and fallibility may be denied, and they are followed with unquestioning loyalty by groups of believers. The dangers of such an attitude have been manifest through the ages in collective acts of destructive or unconventional behavior inspired by the religious specialist or leader. There is sometimes a fine line between being a holy fool and a saint.

3.4 Religious Specialists and Gender

In most world religions, authority and power has been the preserve of men and most religious specialists in history have been male. In some world religions, nearly all formally recognized religious professionals are still men, and in these cases the situation does not seem likely to change in the immediate future. However, in other religious traditions and denominations, women increasingly are taking on roles as public religious professionals, through ordination, particularly in the lower strata of the religious hierarchy. Some of the exceptions to this historical bias towards male religious specialists are the monastic traditions of Buddhism and Christianity where separate orders for women have long been in existence and women have gained credit for their virtuoso religiosity. In various religious traditions over the course of history, individual women have been respected for their wisdom and devotion, such as the Muslim saint and mystic

Rabi'a of Basra (d. 801), and in some instances they have, informally, been the teachers of men. But rarely has this led to their admission into the select community of male religious specialists or scholars. Women have through time often done a great deal of unseen religious work, particularly in prayer and in the care of the sick. The increasingly professional and public role of women in religious life is a reflection of changing social attitudes and the senior roles that some women have achieved in other realms of public life.

4. Religious Specialists in the Contemporary World

The role of religious specialists today is carried out in social settings marked by a greater global awareness, increasing religious diversity, and pressures from within and outside the religious community. The impact of these pressures are manifold and have affected different religious traditions and religious professionals, collectively and individually, in various ways. It is beyond the scope of this article to examine these in detail, but some distinctive features of being a religious professional in modern society can be highlighted.

4.1 Religious Leadership in Minority Contexts

Historically, religious communities and faiths were often separated from each other through geographical distance, but the advent of modern communications has brought traditions together on a new and unprecedented scale. Most large cities around the globe will have a community of believers associated with the major religious traditions. In these contexts, religious specialists face the challenge of maintaining authority in a situation where religious professionals from other faith traditions also claim authenticity and authority.

There is evidence to show that over time religious specialists serving diaspora communities are often influenced by the dominant professional roles of the religious majority. For example, many of the rabbis serving Britain's well-established Jewish community have long-since expanded their primary teaching and education role, and they have a far greater involvement in pastoral care than historically and traditionally they once might. The idea of a 'Jewish chaplain' serving on a prison or hospital chaplaincy team is now commonplace. Even amongst religious communities with a much shorter history in Britain, there is evidence that the roles of religious professionals are changing. The largely ritual responsibilities of Muslim imams, based in mosques, is beginning to change as they too become more involved in public religious affairs, community support, and pastoral care in public institutions. In this respect, the gradual combining of pastoral care

with teaching and leadership of rituals indicates the influence of the dominant professional roles of the religious majority, namely the established Church of England.

4.2 Religious vs. Secular Professionals

Through an ongoing process of social differentiation, many of the functions of the religious professional have been usurped by other, nonreligious professionals or by the state. The marginality that religious professionals may face as a consequence, particularly in modern industrial societies, can and does have an impact on future recruitment and vocation. One of the consequences of the decline in the numbers of religious professionals (in some societies and religious traditions), has been an inevitable need for ordinary, lay members of religious communities to become more actively involved in tasks that were formally part of the priestly role.

A good example of how the remit of religious professionals has changed can be seen in the post-Second World War period, particularly in the West, with the dramatic expansion of interest in holistic and alternative therapies with counseling and psychotherapy part of a new industry of self-development. As Davie (1994) notes this has had two effects, namely, 'in usurping some of the traditional functions of a pastor and, secondly, in obliging that pastor—if he wishes to keep on terms—to acquire skills (including an extensive vocabulary) of a novel and essentially unreligious nature.' The curriculum for those undertaking theological training will now almost certainly require study of counseling, bereavement care, and other forms of pastoral care in order to be able to keep abreast of secular professionals working in the fields of therapy and self-development.

See also: Charisma and Charismatic; Charisma: Social Aspects of; Religion and Gender; Religion: Definition and Explanation; Religion: Evolution and Development; Religious Organizations; Ritual; Weberian Social Thought, History Of

Bibliography

- Beckford J A, Gilliat S 1998 *Religion in Prison: Equal Rites in a Multi-Faith Society*. Cambridge University Press, Cambridge, UK
- Davie G 1994 *Religion in Britain since 1945: Believing Without Belonging*. Blackwell, Oxford, UK
- Feuerstein G 1990 *Holy Madness: The Shock Tactics and Radical Teachings of Crazy-Wise Adepts, Holy Fools, and Rascal Gurus*. Arkana, London
- Fichter J H 1961 *Religion as an Occupation: A Study in the Sociology of Professions*. University of Notre Dame Press, London and Notre Dame, IN

Gombrich R F 1988 *Theravada Buddhism: A Social History from Ancient Benares to Modern Colombo*. Routledge and Kegan Paul, London

McGuire M B 1987 *Religion: The Social Context*. Wadsworth Publishing, Belmont, CA

S. Gilliat-Ray

Religious Stratification

1. Introduction: Two Forms of Stratification

We can distinguish usefully between two separate sociological issues, namely the hierarchical stratification of religious roles in terms of their authority, and the stratification of religious communities within a general system of social stratification. The first line of inquiry considers the hierarchical organization of charismatic power within religious traditions; the second involves the historical and sociological study of the stratification of religious communities, especially in terms of social class and ethnicity. Both forms of religious stratification are analyzed from the sociological perspective of Max Weber (1864–1920).

For further treatment of these topics see *Religion, Sociology of*; *Religious Specialists*; *Charisma and Charismatic*; *Religion and Economic Life*.

2. Virtuoso and Mass Religion

The sociology of social stratification is concerned with the distribution of scarce resources and with the formation of social hierarchies by the processes of social closure and exclusion. The most common illustration of social stratification would be the unequal distribution of economic resources, resulting in a simple formation of economic classes. However, stratification also involves cultural distinctions such as education, lifestyle, and cultural consumption. Weber made an original contribution to the sociology of religion by treating charisma (a 'gift of grace') as a religious value that is also characterized by scarcity. Because these religious values are unequally and unevenly distributed in human societies, there is religious stratification. In *The Social Psychology of the World Religions*, Weber (1991, p. 287) asserted 'that men are *differently qualified* in a religious way stands at the beginning of the history of religion... the sacred values that have been most cherished, the ecstatic and visionary capacities of shamans, sorcerers, ascetics, and pneumatics of all sorts, could not be attained by everyone. The possession of such faculties as "charisma," which, to be sure, might be awakened in some but not in all. It follows from this that all intensive religiosity has a tendency towards *status stratification*,

in accordance with differences in the charismatic qualifications.'

Religious gifts (charisma) are in demand because they bring healing and salvation, but they are inevitably in short supply. The resulting hierarchy of spiritual gifts creates a stratification system that differentiates between the virtuoso and the mass, where the latter are, in Weber's terminology, religiously 'unmusical.' This hierarchy of intense religious values stands in contrast to and frequently in opposition to the formal hierarchy of the ecclesiastical organizations of Christianity that attempted to regulate and to control the flow of charisma.

This model of virtuoso–mass religion is connected with the specific forms of salvational drive in the world religions, namely with their soteriologies. Religious systems of soteriology may be directed either toward inner-worldly asceticism or toward other-worldly mysticism. Now all religions that place a special emphasis on rebirth and spiritual renewal will create a special category of religious aristocracy, because charisma cannot be institutionalized in terms of an egalitarian distribution. This virtuoso pattern emerged as a general characteristic of religious communities, for example, among the Protestant sects, the *perushim* in Judaism, the Sufi dervishes of Islam, and the Buddhist monks (see Weber 1965, pp. 151–65). The existence of distinctive forms of internal religious stratification has often evolved in sharp contrast to the official doctrines of egalitarianism and universalism in, for example, Christianity and Islam.

There is demand from popular religion to enjoy the benefits of virtuoso religion such as healing, and thus there develops a distinctive pattern of exchange between the virtuosi and the mass. In return for acts of charity and direct economic support from their followers, the virtuosi offer healing, guidance, and other charismatic gifts. The layman who is completely immersed in the everyday world needs the blessing of charisma, but in return the virtuosi depend on lay tributes merely to exist. This mutual dependency and reciprocity create the conditions whereby virtuoso religion can always be diluted or manipulated to some extent by the laity who demand tangible evidence of charismatic gifts in the form of magical displays. In fact, with the possible exceptions of Judaism and Protestantism, all religions and religious ethics have had to (re)introduce cults of saints, heroes, or functional gods in order to accommodate themselves to the needs of the masses (Weber 1965, p. 103). In Asia the religious teachers or *gurus* function as living saviors, while in Catholicism the veneration of the saints was the real basis of popular religion. In Hasidic Judaism, despite its popular appeal to the ordinary people through, for example, its use of dance as an expression of religious joy, there was also a clear division between the spiritual elite and their disciples. In these examples, Weber recognized a paradoxical tension between the spiritual quest for perfection in the virtuosi, the

pragmatic needs of the laity for healing and material sustenance, and the mutual reciprocity between elite and mass. Weber's sociology of charismatic roles was an application of philosophical critique of Christian morals by the German philosopher Friedrich Nietzsche (1844–1900), who had condemned Christianity as a system of slave morality and as a product of psychological weakness. Slave morals are created out of resentment by a subservient social class, and hence the Christian doctrine of love was analyzed critically as a subconscious doctrine of a resentful social group. Guilt and bad conscience played a pivotal role in Christian ethics. Against this moral system, Nietzsche argued in favor of a revaluation of values from a position of psychological strength. Nietzsche's genealogy of (Christian) morals in terms of a study of the stratification of values between slave and master moralities was translated directly into Weber's social psychology of the world religions (see Staught and Turner 1988). In Weber's treatment of the psychology of religion, he argued that certain social carriers were crucial in the development of the cultural systems of the various religions. In particular, he claimed, for example, that Islam had been carried by warriors, Buddhism by mendicant monks, Confucianism by court officials and the literati, and Christianity by artisans and craftsmen. These occupational groups were particularly important historically in providing the world religions with a specific soteriology. Thus, the world view of Christianity was shaped by the practical rationalism of artisans and craftsmen, and in the long run this rationalism proved decisive in the connection between the Protestant capitalists and economic rationalism. This thesis has been challenged and various historians have attempted to find similar connections between Jewish culture and the rise of capitalism (see Sombart 1951) or between Islam and capitalism (see Turner 1998). The core of the Weber thesis has survived these criticisms insofar as it attempts to identify important cultural connections between religious discipline, class position, and economic rationalism.

3. Religion and Social Stratification

Religious communities or groups can also be analyzed in relation to the general system of secular stratification in terms of social class or caste. There is an extensive and complex debate about the relationship between caste, occupational groupings, notions of pollution, and Hinduism. The conventional view was that concepts of purity in Hinduism functioned to sustain a social division of labor between manual and nonmanual occupations that over many centuries evolved into a complete hierarchy of caste. Although the Hindu caste system was organized in terms of four principal castes (Brahmin, Kshatriya, Vaisya, and Sudra), there was considerable diversity at the village

level, where these castes were typically subdivided into subcastes or *jati*. Castes were ordered hierarchically in terms of ritual separation around a notion of contagious pollution. This traditional pattern broke down in the twentieth century and there has since been some mobility whereby through the adoption of upper-caste cultural practices lower castes can move up the social scale, and the incorporation of marginal cultural traditions into the Brahmanical system. This process is referred to as 'sanskritization.' There is thus a debate about the differences between Asian and Western systems of stratification and the legitimization of such patterns of inequality by religion (see Dumont 1970). Some sociologists argue that the term caste can be applied to any system of rigid stratification where membership is permanent, hereditary, and guarded or legitimized by religious notions of purity. For example, in the southern states of anti-bellum North America, there is a continuing system of racial orders that is reinforced by the stratification of Christian churches along a division between blacks and whites. The National Baptist Convention, the African Methodist Episcopal, the African Methodist Episcopal Zion, and the Colored Methodist Episcopal churches emerged as exclusively black institutions and illustrate the racial stratification of American denominationalism (see Dollard 1998).

Although religion has played a significant role in the legitimization of social stratification, the relationship between religion and class has to be seen in a dynamic and historical framework. There is a considerable collection of historical and sociological research findings that explore the social mobility of individuals and groups as a consequence of their conversion to evangelical Christian sects. H. Richard Niebuhr's *The Social Sources of Denominationalism* (1957) is the classic illustration of this tradition in the sociology of religion. The basic argument is that 'the churches of the disinherited' have traditionally recruited from the poor and dispossessed through active evangelism. As a consequence of religious asceticism and education, individuals from the lower social classes move up the social scale. The same argument applies to the evangelical sects themselves which, as they become dominated by the lower-middle and middle classes, become more respectable. The process of denominationalization whereby sects acquire the characteristics of middle-class denominations in the United States has been paralleled to the social mobility of their individual members. The history of the Anabaptists, Quakers, Methodists, and Salvation Army exhibit this pattern of social change whereby sects become denominations (see Troeltsch 1931).

The churches of the middle class emphasize personal responsibility and individualism in theology, and restraint and formality in ritual. Their social teaching is concerned with individual behavior rather than with social redemption. The sects of the lower classes appeal to the deprivation of the poor through emotional

preaching, healing practices, and a collective eschatology of redemption. However, because these evangelical sects typically socialize the disinherited into the dominant values of a society, the poor rise through the social scale as they acquire the habits of temperance and personal discipline (Johnson 1961). The education of the laity slowly replaces their conversion. There is some support from the history of the evangelical sects for Max Weber's *The Protestant Ethic and the Spirit of Capitalism* because the personal discipline of the Protestant sects results in the social mobility of the disinherited into the lower and middle classes. Habits of saving, frugality, and temperance produced a reliable workforce that was compatible with the needs of capitalist discipline. In particular, the owners of industrial enterprises recognized that factory discipline could be maintained effectively through the employment of sober Christian workers (see Pope 1942).

4. Religion and Class Conflict

There is a close relationship between class membership and religious affiliation in which religious participation functions as a cultural ladder of social mobility. This association between evangelical Christianity and the working class suggested to Karl Marx (1818–83) that the promise of spiritual salvation compensated for the material deprivation of the disinherited. In the famous argument about human alienation in capitalism, Marx (1974, Vol. 1, p. 83) argued that for a society based on private property, 'Christianity with its cultus of abstract man, more especially in its bourgeois developments, Protestantism, Deism etc., is the most fitting form of religion.' Within the system of capitalist exploitation, Christianity was the opium of the people, because its emotional discharge served to obscure the real nature of class inequalities. While Marx treated Christianity in a cursory fashion, Friedrich Engels (1820–95) offered a more extensive treatment of the relationship between class, religion, and political consciousness. For example, he regarded Christianity in its origins as a religion of a class of slaves and oppressed that became eventually the dominant religion of the Roman Empire. There were close historical parallels between the rise of Christianity and the development of the European working class (see Marx and Engels 1955). Engels recognized that in the Reformation the peasant revolts were an expression of real material interests against oppression, and that Christian belief provided the peasantry with a powerful ideology of protest (see F. Engels, *Peasant War in Germany*, 1965). These peasant protests which drew upon the Bible as a source of social and political criticism were important because they provided a model of the millenarian struggles of the seventeenth century and the secular conflicts of the nineteenth century, when working class movements continued to absorb and transform the biblical language of protest

against capitalism. Methodism often provided the training ground for political protest, because the tradition of lay preaching provided a pedagogy of popular opposition. Popular leaders who emerged within the ranks of the Methodist chapels typically were expelled from the church because their secular radicalism was incompatible with the conservative teaching of the Methodist chapels (see Thompson 1963).

In the seventeenth century, the religious sects of the peasantry and yeomanry were important engines of social change in the English civil war. In the eighteenth century, Methodism had through its emotional preaching provided a compensation for this worldly deprivation. In the nineteenth century, there was an unstable alliance between working-class protest and the evangelical sects, whereby the idiom of protest depended more on the Bible than on Marxist ideology. In the English case, these causal connections between politics and religion were summarized in the Halévy thesis, which suggested that Wesleyan Methodism functioned as a social ladder between the nonconformist dissent of the artisan and the Anglican establishment, thereby contributing to social stability and evolutionary political change (see Halévy 1962).

5. Conclusion: Religion, Class and Egalitarianism

Whereas religions like Confucianism and Buddhism have not systematically provided or embraced an ideology of radical egalitarianism, the Abrahamic faiths of Judaism, Christianity, and Islam are regarded as prophetic religions of the book that generate a revolutionary utopia of salvation. Because they create a radical tension between this world and a sacred realm of justice, they have been seed-beds of revolutionary movements. The actual existence of religious stratification is in this sense an offence to the theology of justice in biblical religion. Logically speaking, this tension between existing social injustice and the prophetic promise of the restoration of the faithful and the punishment of their enemies can take two forms—either a theodicy of other-worldly justice in a future kingdom or radical millenarianism in this world (see Mannheim 1997). These images of revenge and renewal provided slaves and the disinherited classes with a profound ideology of political change. However, in the Abrahamic religions the connections between social class, religious ideology, and political protest have remained highly unstable. Periods of revolutionary fervor have been followed by periods of mystical withdrawal, and the pendulum swing between protest and quietism is well illustrated by the complex history of messianism in Judaism, Christianity, and Islam (see Lantieri 1963). In Islam, the tension between the Koranic promise of a radical egalitarianism and earthly justice has proved to be a resilient source of religious protest against the dominant classes (see Marlow 1997). Although Marx and Engels as-

sumed that religious ideologies of class protest would be replaced by secular or scientific visions of political change, religious inspiration of class protest in the twentieth century has been an important aspect of liberation theology in Latin American Catholicism and of Islamic protests against colonialism. Traditional Shi'ite notions of the Hidden Imam and justice within the Islamic community provided a powerful tool in the emergence of the Iranian revolution which was produced by the alienation of the peasantry and the decline of traditional classes such as the merchants (see Kamali 1998). Religious stratification therefore continues to be an important feature of social and political institutions into the twenty-first century, and religious ideologies provide significant cultural components of both social stability and radical change.

See also: Voting: Class

Bibliography

- Dollard J 1998 *Caste and Class in a Southern Town*. Routledge/Thoemmes Press, London
- Dumont L 1970 *Homo Hierarchicus. The Caste System and its Implications*. University of Chicago Press, Chicago
- Halévy E 1962 *A History of the English People in the Nineteenth Century*. Benn, London, 2 Vols.
- Johnson B 1961 Do holiness sects socialize in dominant values? *Social Forces* 39: 309–16
- Kamali M 1998 *Revolutionary Iran Civil Society and State in the Modernization Process*. Ashgate, Aldershot, UK
- Lantenari V 1963 *The Religions of the Oppressed. A Study of Modern Messianic Cults*. MacGibbon & Kee, London
- Mannheim K 1997 *Ideology and Utopia*. Routledge, London
- Marlow L 1997 *Hierarchy and Egalitarianism in Islamic Thought*. Cambridge University Press, Cambridge, UK
- Marx K 1974 *Capital*. Lawrence and Wishart, London, 3 Vols.
- Marx K, Engels F 1955 *On Religion*. Foreign Languages Publishing House, Moscow
- Niebuhr H R 1957 *The Social Sources of Denominationalism*. Meridian Books, New York
- Pope L 1942 *Millhands and Preachers*. Yale University Press, New Haven, CT
- Sombart W 1951 *The Jews and Modern Capitalism*. Free Press, Glencoe, IL
- Stauth G, Turner B S 1988 *Nietzsche's Dance Resentment Reciprocity and Resistance in Social Life*. Basil Blackwell, Oxford, UK
- Thompson E P 1963 *The Making of the English Working Class*. Penguin, London
- Troeltsch E 1931 *The Social Teaching of the Christian Churches*. George Allen & Unwin, London, 2 Vols.
- Turner B S 1998 *Weber and Islam*. Routledge, London
- Weber M 1930 *The Protestant Ethic and the Spirit of Capitalism*. George Allen & Unwin, London
- Weber M 1965 *The Sociology of Religion*. Methuen, London
- Weber M 1991 The social psychology of the world religions. In: Gerth H H, Wright Mills C (eds.) *From Max Weber Essays in Sociology*. Routledge, London, pp. 267–301

B. S. Turner

Remedies and Damages: Legal Aspects

Remedies can be categorized as monetary or non-monetary. Although much litigation leads to monetary awards, many cases end without a monetary award. Injunctive relief, orders to specifically perform contracts, divorce decrees, probate of wills, and other orders do not necessarily involve damages or monetary payments.

1. Nonmonetary Relief

Several categories of nonmonetary relief exist and their use varies across countries. In contract law, countries divide into three groups with respect to the availability of specific performance: (a) countries, such as Germany and socialist systems, in which specific performance is generally available, subject to exceptions, (b) countries, such as France, which make specific performance available for specific obligations, and (c) common law countries, including England and the United States, in which specific performance is an exceptional remedy, available in limited circumstances.

Nonmonetary relief to prevent injury can be more important than redressing injuries already done. Common law countries address this need by allowing injunctive relief in cases in which irreparable harm can be demonstrated. Other legal systems, such as Germany, Switzerland, Austria, and others influenced by German law, have a general class of actions to prevent threatened injury. Courts in such systems have less discretion to deny preventive remedial relief than do common law courts. Some systems allow remedial relief but lack a coercive component, such as the penalties for violating an injunction, to enforce it (Stoll 1983). Prejudgment nonmonetary remedies, such as attachment and preliminary injunction or its equivalent, are widely available.

2. Damages Rules across Countries

With respect to monetary relief, one cannot completely separate the question of the quantum of damages from the question of liability. Blackstone, in writing about the common law, did not regard damages as a separable ending of the legal process. The right to damages arose the instant the victim was injured. Nevertheless, one can intelligently address many damages issues independently of the question whether any damages should be awarded.

In the last quarter of the twentieth century, damages issues fueled the tort reform movement in the United States and alarmed foreigners doing business in the United States. In many respects, however, damages rules are not strikingly different between the

United States and other countries. All countries seek to compensate victims fully, though most depend more on social compensation systems than does the United States. Most countries allow consideration of the injured person's fault, a consideration similar to the United States doctrine of comparative negligence, and a few (the Netherlands and Sweden) allow for mitigation of damages in view of the parties' financial circumstances and insurance coverage. European and Japanese damages rules governing injury, though less feared than United States rules, are not necessarily simpler (Pfennigstorf 1991). They contain 'a confusing variety of items of non-pecuniary loss and an equally confusing variety of methods for measuring them.'

There are, however, differences in the size of awards. Non-United States systems produce lower and more predictable within-country awards. Courts in several countries observe written and unwritten guidelines with respect to awards. In Japan and Sweden, for example, expert commissions establish consistent guidelines for automobile accidents. Sweden extends this practice to injuries caused by medical mistreatment, by pharmaceutical drugs, and by environmental pollution. Across European countries, however, awards are less consistent. One study estimates that in a wrongful death case involving a 20-year old victim with no dependents, awards across Europe would vary from £210,122 in Italy to £1,250 in England; it also finds that pain and suffering awards vary considerably across countries (McIntosh and Holmes 1990).

3. Social Scientific Study of Compensatory Damages

Media reports, which understandably emphasize extreme awards, promote exaggerated perceptions of large United States damages awards. The media overreport products liability and medical malpractice

cases, exaggerate plaintiff win rates, and distort impressions about the actual distribution of jury awards. Social scientific assessment of damages is critical in helping to ameliorate the effect of this reporting bias.

3.1 General Adequacy of Damages

Despite agreement that awards are higher in the United States, there is little evidence that injury victims are systematically overcompensated in any country. Evidence exists that United States plaintiffs with small losses tend to be overcompensated but that plaintiffs with large losses tend to be undercompensated (Vidmar 1998). Yet it is the large loss cases that tend to be the most controversial. An English study found that serious injuries may not result in substantial recoveries and that minor injuries which do lead to recoveries may have unexpectedly severe consequences (Law Commission 1994). Both findings suggest systematic undercompensation.

3.2 Award Levels and Trends

The most studied system of damages is the tort system in the United States. Concerns about the level of awards, time trends in awards, and the relative competence of judges and juries have generated several studies. Data gathered by the National Center for State Courts from 45 of the largest 75 United States counties, under grants from the Bureau of Justice Statistics (1995, 1999) ('BJS'), provide the least biased sample of awards in litigation. The data allow for meaningful comparisons across case categories. Since the same 45 counties were surveyed in both 1991 and 1996, the data also allow meaningful comparisons over time, discussed below. Table 1 summarizes the

Table 1
Mean and median compensatory damages award, by case category and year, in thousands of 1998 dollars

	1991			1996		
	Mean	Median	n	Mean	Median	n
Auto torts	219	34	1,102	228	21	1,384
Dangerous premises	323	71	394	350	84	441
Products liability	864	319	71	901	363	95
Intentional tort	445	47	91	438	33	112
Medical malpractice	1,424	232	181	1,525	306	172
Professional malpractice	1,303	152	44	392	76	29
Slander/libel	230	23	15	372	19	17
Other torts	447	74	188	11,477	117	186
Contract fraud	774	61	87	488	52	99
Employment relations	412	149	94	537	194	110
Lease	1,963	66	47	239	63	39
Other real estate	205	53	11	154	97	12

pattern of jury compensatory damages awards, in thousands of 1998 dollars, for these 45 counties.

Table 1 shows that damages awards vary substantially by case category. Median compensatory damages awards are highest in cases involving products liability and medical malpractice. Indeed, median awards in these case categories are substantial multiples of the median award in automobile accident cases. The levels of awards are generally modest and are consistent with other comprehensive studies of particular jurisdictions.

3.3 Reliability of Compensatory Awards

In studies that include independent assessments of the defendant's behavior, damages awards correlate with the nature and degree of the wrong. Studies of medical malpractice cases show that those who behaved the worst and caused the most harm are the most likely to pay the most damages. Experimental studies confirm these findings (for example, Wissler et al. 1999). Osborne (1999) uses a different methodology but reaches similar results. His study includes many substantive areas of law and examines the relation of court awards to the pretrial expectations of litigants and their attorneys and to measurable, economically relevant damages. He finds that court awards are highly predictable, as variance in expectations explains much of the variance in awards. The awards also significantly relate to both medical costs and property damage. He rejects the hypothesis of a highly unpredictable court system.

4. Punitive Damages

Punitive or exemplary damages are in addition to the damages awarded to make the victim whole. But normally punishment is left to the state, not to private civil litigation. Thus, many legal systems do not allow damages beyond those necessary to compensate the victim. Little authority exists for punitive damages in Roman law and thus in some of the continental European systems that derived from it. But many countries allow punitive damages or their equivalent (exemplary damages, moral damages) in some classes of cases. And the European Court of Justice has ruled that national courts must be willing to award exemplary damages pursuant to claims or actions founded on Community law, if such damages may be awarded pursuant to similar claims based on domestic law. Nevertheless, punitive damages are substantially less important outside the United States.

The most discussed questions in the recent United States literature on damages relate to punitive damages. Understandably skewed media coverage emphasizing high awards combined with advocacy by lobbying groups has generated enormous pressure on

punitive damages law and helped to shape professional and public impressions. Business lobbying groups have enjoyed some success, with several states enacting provisions raising the burden of proof for punitive damages claims, placing monetary ceilings on awards, and otherwise limiting the likelihood and size of punitive awards.

Social science research on punitive damages expanded as a result of the 1994 award of \$5 billion in punitive damages against Exxon Corporation (now Exxon Mobil) for the 1989 Exxon Valdez oil spill. Exxon funded several leading academics, who produced an impressive body of work that makes important social science contributions. All of this work is critical of punitive damages and supports 'reform.' But the thrust of some of the work seems inconsistent with prior theoretical discussions of punitive damages and some of it has been criticized as methodologically unsound (for example, Eisenberg 1998, Vidmar 1998) and as overstating its implications.

4.1 Rate of Punitive Awards

The social science picture that emerges from systematic studies of actual cases differs in some respects from the implications of the experimental and theoretical results of the Exxon-funded work. Three large studies of punitive damages awards, Moller (1996), Daniels and Martin (1995), and data commissioned by the BJS (1995), cover multiple United States jurisdictions, cover similar time periods, and include case categories that are sufficiently well defined to warrant comparison. Each of the three sources reveal overall punitive damages award rates of less than ten percent in successful jury trials.

In two of the areas of greatest interest to policy-makers, products liability and medical malpractice litigation, punitive awards are especially rare. In medical malpractice cases, the 95 percent confidence intervals for all three sources overlap and show awards in less than five percent of jury trials in which plaintiffs are successful. In products liability cases, all three estimates again overlap at under ten per cent of successful jury trials, and rates under five percent are found in studies of published opinions. Except in case categories dominated by fraud and employment law issues, all punitive award rate levels are likely less than ten percent, and many are less than five percent. These social scientific analyses undermine the belief that punitive damages are frequently awarded.

4.2 Level of Punitive Damages Awards

Large, publicized punitive awards, such as that against Exxon, create the impression that punitive awards are on average much larger than they truly are. BJS data for 1991 and 1996 allow one to assess the level of

Table 2

Mean and median punitive damages award, by year, in thousands of 1998 dollars

	Mean	Mean excluding highest award	Median	n
1991	557	404	58	168
1996	1,887	688	52	119

punitive damages awards. Table 2 summarizes punitive damages awards, in thousands of 1998 dollars, for the years 1991 and 1996. Median punitive damages awards have been more modest than the headline-grabbing awards. In both years they are in the 50,000 dollar range. Mean punitive awards are substantially affected by extremes. Excluding the most extreme punitive award in each year, reduces the mean by over 25 percent in one year and by over 50 percent in the other. In these data, million-dollar awards constitute less than 10 percent of the punitive awards. All studies confirm that the bulk of punitive awards are modest.

4.3 Relation Between Punitive and Compensatory Awards

A criticism of punitive damages is that they bear no relation to the compensatory award—the punishment

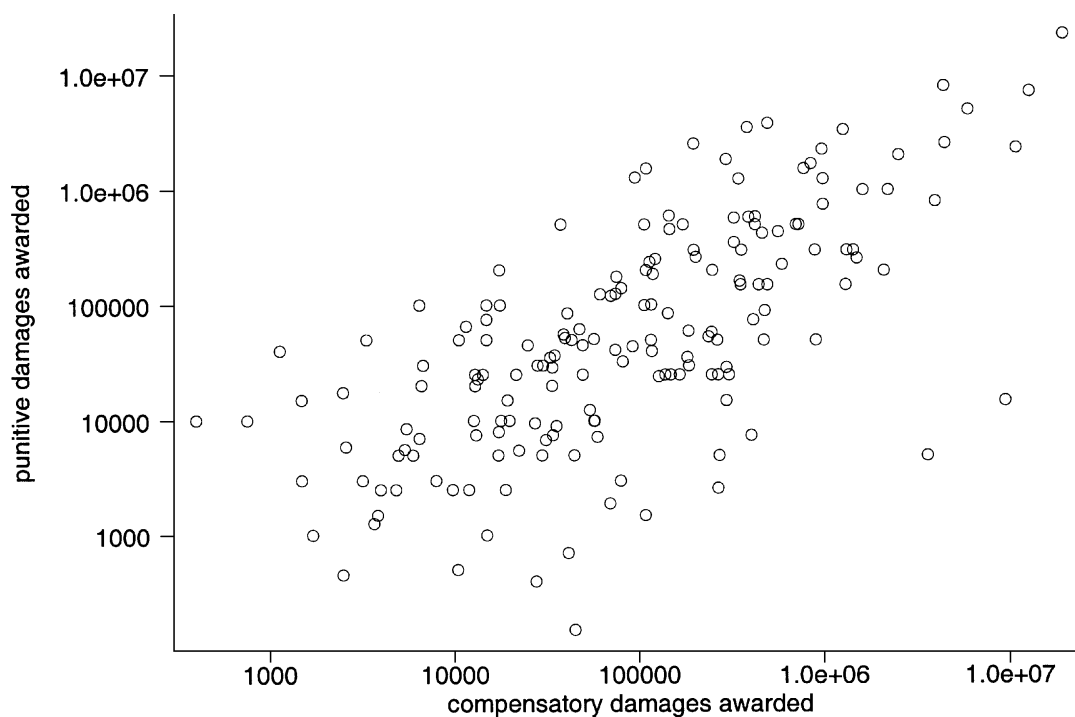
inflicted therefore is not proportional to the harm caused. And the United States Supreme Court has indicated that a punitive damages award normally should bear a reasonable relation to the compensatory award. Social scientific analysis suggests that punitive awards do relate to compensatory awards. Figure 1 shows the relation (in logs) between compensatory and punitive awards in the BJS (1995) study of 45 counties. It shows an unmistakable correlation. Other data sets reveal approximately the same relation (Eisenberg et al. 1997).

Punitive damages awards are thus rarer, smaller, and more closely related to compensatory awards than is widely believed.

5. Time Trends

5.1 Compensatory Damages Awards

Table 1 allows study of award trends over time. Popular perception has it that damages awards are ever-increasing. And some empirical study supports the claim. But these studies have been questioned on methodological grounds (Vidmar 1998). Table 1 shows little systematic evidence of increasing awards. In several case categories, median or mean awards decreased from 1991 to 1996. The only statistically

**Figure 1**

The relation between punitive and compensatory damages awards.

significant change in mean awards was in automobile accident cases and here the 1996 mean (in logs) was statistically significantly *lower* than the 1991 mean ($p < 0.001$). Other changes in award patterns were not statistically significant.

Table 1 is limited to state court cases. Awards in federal court cases tend to be higher, at least partially because of jurisdictional amount requirements in federal court. Median trial awards in federal courts products liability cases did increase from the early 1980s to the mid-1990s. The increase also occurred in judge-tried cases and in contracts cases.

5.2 Punitive Damages Awards

Punitive damages, like compensatory damages, are often said to be ever-increasing in number and size. The two years of BJS jury-trial data allow one to compare the rate and size of punitive awards in the 45 large counties covered by the data. Table 2 summarizes the results, in thousands of 1998 dollars, for the years 1991 and 1996. It shows that median punitive awards were substantially lower in 1996 than 1991. Mean awards were higher in 1996 but that is a function of a few large cases. The means (log) are not statistically significantly different. Regression analysis that controls for locale and case category confirms that punitive damages show no statistically significant movement from 1991 to 1996.

Table 2 also shows that the frequency of punitive damages awards decreased in these 45 counties, from 168 in 1991 to 119 in 1996. Thus, even minimal social scientific analysis provides no support for the belief that punitive damages awards are ever-increasing in size or frequency.

6. The Importance of the Adjudicator

It is widely believed that the choice of adjudicator—judge, jury, arbitrator—substantially affects the damages awarded. Higher and more variable awards in the US are attributed to the greater role of jurors as adjudicators. While real differences exist across adjudicator groups, the social scientific evidence supports skepticism about claims of massive differences across adjudicators. The classic Kalven and Zeisel study used questionnaires to presiding judges in about 4,000 actual civil jury trials in the 1950s (Kalven and Zeisel 1966). Their data showed a 78 percent agreement rate between judge and jury on liability, with no revealing pattern in the cases of disagreement. The rate of judge-jury agreement is high compared to other human endeavors. They also showed that jurors awarded on average 20 percent more than judges said they would have awarded. The modest judge-jury differences reduced interest in attacks on juries.

Renewed attacks on the jury's competence to assess damages, and especially punitive damages, re-emerged in the 1980s and 1990s. Research using interviews and experiments finds that different decisionmakers are quite similar and predictable with respect to their judgments of injury severity (Wissler et al. 1999). Such research also finds that jurors' awards lose consistency in translating severity of injury into dollar awards. Other studies suggest somewhat lesser judge-jury differences. Vidmar and Rice (1993) report experimental evidence about jurors awaiting jury duty and legal professionals bearing on pain and suffering awards. In a medical malpractice burn case, they found that six-person juries would give more variable awards than twelve-person juries but that both groups would be less variable than individual legal professionals. Their second experiment involved an automobile accident case and a medical malpractice case. They found no differences between jurors and legal professionals.

Studies of actual cases, which lack the controls of experimental settings but provide the relevant setting, suggest modest adjudicator differences. Osborne's (1999) study compares the predictability of judge and jury awards based on litigant expectations. He finds that one can reject the hypothesis that jury trials are largely random events. Litigant expectations explain more than a third of the variance in awards. And he finds that the behavior of judges is not significantly more predictable than that of juries. Helland and Tabarrok (2000) conclude that most differences between judge and jury awards can be explained by differences in the sample of cases coming before judges and juries but that some suggestive differences exist. Eisenberg et al. (2000) find little meaningful difference in the pattern of judge and jury punitive damages awards.

Exxon-funded research finds experimental evidence of vast judge-jury differences. For example, Hastie and Viscusi (1998) report mock jurors to be much more likely than judges to impose punitive sanctions in a hypothetical railroad accident case. Vast judge-jury differences on liability in these new studies, which contrast with Kalven and Zeisel's and other findings, calls into question the experimental designs' ability to detect socially meaningful punitive damages differences between judges and juries.

7. Overall Assessments and Comparisons of Damages Awards

Greater injuries lead to greater damages awards in all systems for which data are available. With respect to variability in the United States, where variability is regarded as greatest, from half to almost three-fourths of awards' variation can be accounted for by measures of injury severity, as well as by other predictors, but a

substantial random component also exists. Legal systems that rely on schedules of damages for various injuries have lower variability though not necessarily more complete compensation.

Prospects for definitive social scientific comparison of damages rules' operation are dim. Some inherent limitations apply to comparing performance of legal rules across state and national boundaries. The number of differing legal systems will tend to be small—there are, for example, only 50 states and gathering useful damages information about that many countries is an enormous task. Important differences, such as varying reliance on social compensation systems, different mixes of cases reaching court systems, different methods for compensating attorneys, and varying standards of living and purchasing power need to be considered. For the foreseeable future, reasoned analysis of data available for individual jurisdictions and informal comparison across jurisdictions will have to suffice in assessing performance of the damages component of remedial systems.

See also: Deterrence; Deterrence: Legal Perspectives; Deterrence Theory: Crime; Liability: Legal; Litigation; Parties: Litigants and Claimants; Punishment: Social and Legal Aspects

Bibliography

- Bureau of Justice Statistics Special Report 1995 *Civil Justice Survey of State Courts 1992: Tort Cases in Large Counties*
- Bureau of Justice Statistics Bulletin: Civil Justice Survey of State Courts 1999 *Civil Trial Cases and Verdicts in Large Counties, 1996*
- Daniels S, Martin J 1995 *Civil Juries and the Politics of Reform*. Northwestern University Press, Evanston, IL
- Eisenberg T 1998 Measuring the deterrent effect of punitive damages. *Georgetown Law Journal* **87**: 347–57
- Eisenberg T, Goerd T, Ostrom B, Rottman D, Wells M T 1997 The predictability of punitive damages. *Journal of Legal Studies* **26**: 623–61
- Eisenberg T, LaFountain N, Ostrom B, Rottman D, Wells M T 2002 Judges, juries, and punitive damages: An empirical study. *Cornell Law Review* **87**:
- Hastie R, Viscusi W K 1998 What juries can't do well: The jury's performance as a risk manager. *Arizona Law Review* **40**: 901–21
- Helland E, Tabarrok A 2000 Runaway judges? Selection effects and the jury. *Journal of Law Economics Organization* **16**: 306–33
- Kalven H, Zeisel H 1966 *The American Jury*. University of Chicago Press, Chicago, IL
- Law Commission for England and Wales 1994 *Personal Injury Compensation: How Much Is Enough? A Study of the Compensation Experiences of Victims of Personal Injury*. Law Com. No. 235. HMSO, London
- McIntosh D, Holmes M 1990 *Personal Injury Awards in EC Countries: An Industry Report*. Lloyd's of London Press, London
- Moller E K 1996 *Trends in Civil Jury Verdicts Since 1985*. RAND, Santa Monica, CA

Osborne E 1999 Courts as casinos? An empirical investigation of randomness and efficiency in civil litigation. *Journal of Legal Studies* **28**: 187–203

Pfennigstorf W 1991 *A Comparative Study of Liability Law and Compensation Schemes in Ten Countries and the United States*. Insurance Research Council, Oak Brook, IL

Stoll H 1983 Consequences of liability: remedies. In: Tunc A (ed.) *International Encyclopedia of Comparative Law* **11**, J. C. B. Mohr, Tübingen, Germany

Vidmar N 1998 The performance of the American civil jury: An empirical perspective. *Arizona Law Review* **40**: 849–99

Vidmar N, Rice J 1993 Assessments of noneconomic damage awards in medical negligence—A comparison of jurors with legal professionals. *Iowa Law Review* **78**: 883–911

Wissler R L, Hart A J, Saks M J 1999 Decisionmaking about general damages: A comparison of jurors, judges, and lawyers. *Michigan Law Review* **98**: 751–826

T. Eisenberg

Remote Sensing

Remote sensing is one of a suite of geospatial technologies that are having a growing impact in a wide variety of areas from commerce to science in public policy. While remote sensing's origins and technological advance are an outgrowth of military and intelligence applications, commercial and scientific applications of the technology have been growing. The field of remote sensing evolved from the interpretation of aerial photographs to the analysis of satellite imagery, and from local area studies to global analyses, with advances in sensor system technologies and digital computing. Today remote sensor systems can provide data from energy emitted, reflected, and/or transmitted from all parts of the electromagnetic spectrum. These data can then be transformed into information products either by manual or machine-assisted interpretation and employed in an analytical or management decision-making context. Examples of applications of these data include population and demography studies, study of archaeological sites, energy studies using hydrological models, urban planning, environmental treaty enforcement, and agricultural yields. This article provides a brief overview of the historical development of remote sensing, emphasizing the increasing complexity of platforms, systems, and tasks. The information provided here has been assembled from previously published sources to give an overview of a rapidly evolving field that has had, and will in the future have, an increasing impact on our global society.

The definition of remote sensing used here, provided by the American Society for Photogrammetry and Remote Sensing (ASPRS), is:

In the broadest sense, the measurement or acquisition of information of some property of an object or phenomena, by a recording device that is not in physical or intimate contact

with the object or phenomenon under study; e.g., the utilization at a distance (as from aircraft, spacecraft, or ship) of any device and its attendant display for gathering information pertinent to the environment, such as measurements of force fields, electromagnetic radiation, or acoustic energy. The technique employs such devices as the camera, lasers, and radio frequency receivers, radar systems, sonar, seismographs, magnetometers, and scintillation counters. (Reeves et al. 1975)

What follows is a brief summary of various sensor systems within the context of an historical background of remote sensing, and a short discussion of the future of remote sensing with some reference to social and behavioral sciences.

1. Sensor Systems

Today, a wide variety of systems are used to detect energy patterns, which provide information about the Earth's atmosphere, oceans, and land surfaces. These systems range from cameras in aircraft to advanced sensor systems in Earth-orbiting satellites. The aerial camera systems typically used for remote sensing research and applications today are highly sophisticated. Most satellite remotely sensed data are acquired by sensor systems such as multispectral scanners, linear array devices, and synthetic aperture radar systems that operate in the visible, near infrared, thermal infrared, and microwave portions of the electromagnetic spectrum.

These remote sensor systems can be broadly categorized into *passive systems* and *active systems*. A camera is an example of a passive system as its film records light energy reflected from an existing source: the sun. Add a flash attachment to the camera and it becomes an active system, since it now provides its own source of illumination. Passive systems record solar radiation reflected from the Earth's surface. Data derived from passive systems such as multispectral scanners provide information on, among other things, land use, vegetation types, distribution and condition of landforms, soils, surface waters, and river networks. Active microwave (radar) systems are used commonly in geological and hydrological applications. Because active systems do not depend on reflected energy from the sun they may be able to acquire data during the day or night. In addition, as the atmosphere does not attenuate microwave radiation as much as shorter wavelength energy, radar systems allow data acquisition in cloudy and rainy weather. This ability to operate day or night, and to create images of the Earth's surface in spite of cloud cover, makes radar sensors particularly attractive tools for improving our understanding of both tropical and polar areas.

Thermal infrared data, which represent a record of emitted energy from surfaces, have been particularly useful in monitoring fires and in improving our

understanding of areas of volcanic and geothermal activity. Surface temperature of the ocean is also related to the dynamics of coastal waters and currents. Over land, plant water stress also induces changes in canopy temperatures, which can be detected by thermal sensors. Thermal imagery from satellites has been used to record the nighttime lights of the globe, giving researchers an indication of the spatial distribution of population centers. Other remote sensing systems detect the earth's magnetic and gravitational fields; these tools are used extensively in oil and mineral exploration. Employing imagery from more than one portion of the electromagnetic spectrum can increase the information derived from an analysis, and is typically referred to as multispectral image analysis. Analyses of such multispectral data can, if properly designed, increase both the quantity and quality of information for given applications.

2. Image Interpretation

Interpretation of remotely sensed data is part science and part art. Manual interpretation of remotely sensed data is fundamental to all image interpretation. Even with the ever-increasing sophistication of machine-assisted image interpretation software, it is still the human that ultimately must label the feature or phenomenon being identified. Not all individuals have equal interpretive skills with respect to aerial photographs or images acquired by remote sensor systems. It is axiomatic, however, that the more one knows about the elements, techniques and methods of image analysis the better an interpreter one will be. In addition, the more one knows about the systems which acquire the data one is interpreting and the subject or discipline area most closely related to the interpretation, the better an interpreter one will make. It is seldom that an interpreter uses only the information in the imagery being analyzed as the basis for their interpretation. Ancillary data in the form of interpretation keys, maps, diagrams, textual material, and other forms of site-specific literature often aid in interpretations. Fieldwork is also a key element in image interpretation and the collection of what is termed 'ground truth' should be accomplished whenever possible and practical.

2.1 Resolution

Resolution is at the core of the successful use of remote sensing data. Resolution, as typically applied by a user, normally involves the question 'can I identify what I need to identify on this imagery?' Often this involves the use of the term *ground resolved distance* (GRD). GRD is used to describe the smallest object that can be unambiguously identified on the ground given good object to background contrast. Object to

background contrast is important here. While it may be possible to see white lines on a black road on one set of images, it may not be possible to see the same size white lines on a concrete stretch of the same road. Resolution is thus dependent upon many things.

Indeed, the level of matching between the spatial, spectral, and temporal resolutions of the measurements and needs of the investigation will determine the usefulness of the derived data/information. Spectral range and resolution as used here refer to the portion of the electromagnetic spectrum being used in the measurement. It must be appropriate to the questions being addressed, for example, thermal range for fire detection, visible or infrared for vegetation health and condition. Spatial resolution determines the level of detail that can be extracted concerning objects in a given scene. As a general rule it must correspond to the sizes of typical objects that must be separately identified. A resolution of a meter or less is needed to identify a single tree, a resolution of five meters for a single-family residence, a resolution of 50 meters might be sufficient for identification of most agricultural fields in North America, while a resolution of 500 meters to a kilometer may be enough for studying ocean surface features. A necessary compromise between area to be covered and ground resolution of the measurement must take into account the volume of data which will be generated by, say, a large area coverage—millions of km²—using high resolution data—one to 30 meters.

The frequency of the data acquisition must also match the natural frequency of change in the landscape and the physical, social, or cultural phenomena of interest. For traffic studies, timeframes of minutes can be important. Hourly observations might be of interest for certain meteorological conditions. Daily observations may be needed for an assessment of plant evapotranspiration associated with irrigated agricultural crops, while observation once every year could be sufficient for land cover change assessment.

2.2 Geographic Information Systems

The synergy between geographic information systems (GIS) and remote sensing comes into play here. To be interpreted accurately, remotely sensed data are often supplemented with other data. Often these ancillary geospatial data can be found or included in a GIS for analysis. But to be more valuable in decision-making contexts, GIS data layers should be up-to-date as is practical. Remotely sensed data are a key technology for updating many types of GIS data. Thus when environmental planners, resource managers, and public policy decision-makers want to measure, map, monitor, or model future scenarios in order to facilitate better management decision-making, remote sensing is being employed more and more within the context of a GIS as a decision support system.

3. Historical Development

In the most general sense, the field of remote sensing begins with the invention of photography. Joseph Nicéphore Niepce took the first photograph of nature in 1827. It is a picture of the courtyard of his home in the village of Saint Loup de Varenne, France, with an exposure time of eight hours. In 1829 Niepce and Louis M. Daguerre signed a partnership agreement to work on a practical method of taking photographs. Niepce had been working on what he called Heliography, or sun drawing, and Daguerre on dioramas, which he constructed with the aid of a camera obscura. The process they pioneered was called direct positive image. The result of their collaboration, called the *daguerreotype*, was finalized in 1839. Also in 1839, an Englishman named William H. Fox-Talbot and a member of the Royal Academy invented the *calotype* (from the Greek: *kalos* for beautiful and *typos* for impression). Talbot's positive negative process is essentially the same as that used today, yet at the time the images produced by the two processes were not comparable.

In 1858, Gaspar Felix Tournachon, also known as Nadar a Parisian photographer and balloonist, took the first known aerial photograph from a captive balloon at an altitude of 1,200 feet over Paris. During the late 1870s, German foresters used aerial photography taken from balloons to both map and measure the aerial extent of tree species. This is an example of the earliest known application of both aerial photographic interpretation and photogrammetry (the science of making precise measurements from remotely sensed data).

The use of rockets to launch cameras first occurred in 1906 when Albert Maul launched a rocket propelled by compressed air that was equipped with a camera. The camera took a picture from an altitude of 2,600 feet and the camera was then returned to the ground by parachute. In 1909, Wilbur Wright took one of the first photographs from an airplane in Centocelli, Italy. By 1915, cameras devoted to aerial photography were being produced. World War One saw an explosion in the use of aerial photographs. During the Meuse Argonne offensive some 55,000 aerial photographs were taken and delivered to Allied Expeditionary Forces over a four-day period. During this period thousands of photo interpreters were trained. These interpreters laid the groundwork for the applications of aerial photographic interpretation that began to expand after the end of World War One.

In the early 1920s the first books and articles on the applications of aerial photo interpretation began to be published. Lee's 1922 *The Face of the Earth As Seen From the Air* work 'shows familiar scenes from a new angle,' and contains chapters on mountain features, coastal mud flats, plains, and submerged landforms. Joerg's 1927 work, 'The use of the air plane in city geography,' was published in the *Annals of the*

Association of American Geographers. This work focused on the application of aerial photography for urban studies. In 1927 Bergen published his work on 'Aerial surveys for city planning' in the *Transactions of the American Society of Civil Engineers*, and in 1930 the American Geographical Society published the work by Platt and Johnson, 'Peru from the Air.' While providing a regional overview of the country this work also focuses on archeological sites. Of a more general nature is the work by Fairchild (1927) in the May issue of the *Annals of the American Academy of Political and Social Sciences* entitled: 'Aerial photography, its development and future.' This article discusses potential applications of aerial photography from highway planning to water supply management. During this period, however, by far the foremost applications of aerial photography discussed in the literature are for mapping and geological exploration, particularly on the uses of aerial photography in oil exploration, structural and engineering geological analyses, land feature mapping, and soils mapping.

It is in the late 1920s and early 1930s that the agencies of the US Government begin to make significant use of aerial photography for operational applications. At this time the Agricultural Adjustment Administration began to systematically photograph farm and ranch lands across the entire USA. This use of aerial photography for documenting agriculturally active areas has continued at five to 10-year intervals since this time. This historical record has been very significant in many types of studies, but has been particularly useful in a variety of legal actions where the documentation of land use and/or land cover conditions and changes through time are in question. At the same time, the US Forest Service began to use aerial photography to document much of the nation's timber reserves. Studies such as this began in other nations as well, notably Canada. In this timeframe the US Geological Survey began to photograph many areas of the country in order to make topographical and geological maps based upon photogrammetric techniques and interpretation. The Tennessee Valley Authority and other regional agencies began to use aerial photography for planning purposes, as did state and metropolitan planning agencies. The *Master Plan of Residential Land Use*, published in 1943 by the Chicago Planning Commission, was based largely on data interpreted from aerial photographs.

World War Two led to developments in aerial photography, with a special interest in new and higher resolution film emulsions and thermal infrared imagery. Kodak patented the first false color infrared sensitive film in 1946. The 1950s saw developments in technology that allowed multispectral imagery to be taken. Thermal infrared and side-looking airborne radar systems began to be acquired by the US military. Around this time, research personnel at the US Office of Naval Research coined the term 'remote sensing' to

cover the expanding range of sensor systems and technologies for acquiring images of the world around us. This was the same year that the Soviet Union launched Sputnik, the first Earth-orbiting satellite. Civilian use of remotely sensed data during this period produced some of the first studies on the application of aerial photography for population estimation.

As early as 1946 the US Government began to pursue an operational space-based imaging system when the US Air Force asked the RAND Corporation to explore the possibility of launching an object into orbit. An early experiment in satellite land remote sensing was conducted from the Explorer 6 platform in 1959. The synoptic images of Earth's weather patterns acquired by the Television and Infrared Observational Satellite (TIROS-1), launched in April of 1960, captivated the American public and demonstrated the capability and necessity of earth observation from space. Also in 1960 the first experimental spy satellite, a component of what would later become known as the Corona Program, was launched and, shortly thereafter, classified as secret. These data were withheld from the public until 1995, when the imagery and documentation were released to the public under Executive Order 12951. In 1965 NASA initiated the Earth Resources Survey Program (ERS). This early program sought to develop methods and applications for remote sensing. The Department of Agriculture researched the use of remote sensing for geological, hydrological, geographical, and cartographical applications. In 1966 the Department of Commerce began to participate in the ERS program by establishing the Environmental Sciences Group within the Environmental Sciences Administration (the predecessor to NOAA). Also in 1966, the Department of the Interior initiated the multi-agency Earth Resources Observation Satellite Program, which studied the potential of Earth-observing satellites for natural resources applications.

During the 1960s and into the mid-1970s NASA and a number of other Federal Agencies focused a significant amount of attention on applied remote sensing research. This was, in part, aimed at justifying the need for civil land satellite remote sensing systems and cataloging the variety of fundamental and applied science questions that could be addressed by such systems. It is during this period that efforts began to develop computer-assisted image interpretation techniques. Results of these activities, and the state-of-the-art practice of remote sensing during this time period, are documented in the American Society of Photogrammetry's first edition of the *Manual of Remote Sensing* (1975), as well as other sources.

The first Earth Resources Technology Satellite (ERTS-1) was designed originally as a research tool. When it was launched in July of 1972, however, this satellite ushered in a brief era of emphasis by NASA on the demonstration of operational applications using this series of satellites. Through 1974, ERTS-1

(later renamed Landsat 1) transmitted over 100,000 terrestrial images covering 75 percent of the Earth's land surface. Between 1972 and 1978, major research efforts utilizing this new civil land remote sensing data were conducted in agriculture, forestry, water resources, and geology. Other notable regional-scale applications research in the 1970s included the Large Area Crop Inventory Experiment (LACIE); Agricultural and Resource Inventory Surveys Through Aerospace Remote Sensing (AgRISTARS); and the Western Snow Melt Application System Verification Test (ASVT).

In 1975, Landsat 2 was launched and the Ford Administration approved budget appropriations for Landsat 3, which was launched in 1978. In March of 1979, President Carter announced the Administration's commitment to maintaining the continuity of civilian terrestrial satellite remotely sensed data and recommended the transition of the program to the private sector. This initiated the process of program commercialization that would come to fruition in the mid-1980s. In 1981, the Reagan Administration accelerated the pace of Landsat commercialization by essentially rejecting the Carter Administration's commitment to an operational civilian land remote sensing program and advocating the immediate end of government funding for the Landsat Program.

By the middle of the 1980s, it was clear that satellite remote sensing could provide globally consistent, spatially disaggregated, and temporally repetitive coverage of the land surface of the Earth. Landsat 4, with the first Thematic Mapper instrument, was launched in 1982. Higher spatial resolution multispectral imagery was now publicly available and could be applied by scientists and professionals to the assessment of terrestrial ecosystems and land cover. A number of researchers began to develop techniques to employ the spectral information from Landsat to measure green vegetation status, and phenology, as well as inter-annual variation in these parameters. These studies then led researchers to the development of vegetation measures that provide an index of condition and, in part, describe absorption of sunlight, photosynthetic capacity, and evaporation rates. With this research, the science community began to understand that space-based Earth imaging can and does provide such information concerning the spatial and temporal patterns associated with land processes. During this period of capability demonstration, researchers noted that the possibility of developing fully integrated land-ocean-atmosphere monitoring and modeling capabilities were essentially realized.

In March 1984, Landsat 5 was launched and in July of that year the US Congress provided the legislative authority to transfer the Landsat program to the private sector. Thus began the nearly decade long period of 'commercialization' of the US civil land remote sensing program. Commercialization (and the

subsequent price increases for Landsat data) caused a shift in scientific research emphasis. Researchers began to move away from Landsat imagery and began to design experiments utilizing NOAA polar orbiting meteorological satellites and their Advanced Very High Resolution Radiometer (AVHRR) sensor system as a primary data source for scientific research. This shift was the result of increased costs, greater restrictions on sharing and more limited availability of Landsat data to both the basic and applied land remote sensing community.

The US Government would not regain operational control of the Landsat program until the late 1990s. Essentially, though the law changed in 1992, the governmental control of any orbital assets did not occur until Landsat 7 was launched in April 1999. The commercial pricing and data sharing restrictions prior to this transition back to government pricing had a negative effect on the developing research and application programs at the time which sought to utilize satellite imagery but could not employ it due to the high data cost. It would not be until the twenty-first century that the Landsat program would begin to come under the control of an operational land-oriented federal agency, the United States Geological Survey (USGS).

Despite these difficulties, research using Landsat data during this time included work in many disciplines and application areas. Organizations important in formulating the types of research conducted at this time include the Intergovernmental Panel on Climate Change (IPCC), the International Council of Scientific Unions (ICSU), the International Geosphere Biosphere Program (IGBP), and the Human Dimensions of Global Change (HDGC) project. Of particular significance was the research that validated the use of multitemporal Landsat and other data to monitor and assess tropical deforestation; and also unambiguously demonstrated the value of time series data to provide information on change occurring in every area of the world.

Other countries entered the space imaging arena with the launch of the French *Système Probatoire de la Observation de la Terre* (SPOT) satellite in 1986 and the Indian Remote Sensing Satellite (IRS-1) in 1988. In 1990, the French launched SPOT 2; a year later the European Radar Satellite (ERS-1) was launched, aimed primarily at oceanographic applications, followed shortly after by the Russian radar satellite *Almaz*; 1992 saw the launch by the Japanese of the first Japanese Earth Resources Satellite (JERS-1). In 1995, ERS-2 was launched and Canada launched *Radarsat*, with a primary objective of aiding navigation of shipping in the far north.

The most recent advances, at the end of the twentieth century, involved the launch by US agencies and organizations of Landsat 7 with the Enhanced Thematic Mapper (ETM+) system, the Shuttle Interferometric Synthetic Aperture Radar (IFSAR) mis-

sion, Earth Observing System-1 (now named Terra), and the one-meter commercial IKONOS satellite (all in 1999). The Landsat ETM+ data are already proving useful in a variety of studies from fire fuels mapping to deforestation and urban change monitoring. IFSAR data will be employed to improve the quality of topographical maps between about 60° N and 60° S of the Equator. Terra data is being employed to give continental to global-scale land cover change information with a temporal frequency never before available. IKONOS data give civil researchers the opportunity to see features and objects from space that could only be seen if aerial photography or intelligence satellite data were available to the user. These satellites will add immeasurably to our ability to measure, map, monitor, and model our Earth system as never before.

4. Future Directions

What is even more important is that by 2006, some 20 governments and organizations are expected to launch civilian and commercial high spatial resolution satellites in an attempt to capture a share of the growing market for remote sensing imagery. These satellites will have sensor systems with spatial resolutions that range from several kilometers to less than a meter. It is clear from developments such as these that a number of the long-standing restrictions that the military and intelligence community placed on satellite remote sensing are being put aside as technology advances and becomes more accessible, creating greater transparency among nations of the world. With this increasing transparency the opportunity for both positive and negative uses of remotely sensed data increase. On the one hand it may be easier to detect and identify violations of international agreements, while on the other it may be easier for states to gain tactically important information on their neighboring states.

There is already a growing interest in these data among social, economic, and behavioral users. This interest stems from the realization that whereas remotely sensed data cannot provide all the data/information needed to resolve or answer given questions it may, when combined with other information, allow data to be more reliably interpreted across space and through time. Indeed, an important aspect of remote sensing is its value as a historical record—applications in legal proceedings involving environmental contamination and ownership liability are one example. There is a growing diversity of applications to which remotely sensed data may be applied, in social, economic, or behavioral science, from population estimation, famine early warning and disaster assessments to inventorying industrial capacity, assessing transportation infrastructure, and monitoring of environmental laws and treaties.

There have been a great many advances in the field of remote sensing over the years since the invention of the camera. Yet, from an analysis of the materials covering the field of remote sensing, it is clearly evident that users interested in questions in the physical and biological sciences have been (and continue to be) able to use these data more than users in the social, economic, and behavioral sciences.

The reasons for this disparity are many and varied. They range from the inability of remote sensing to provide needed data/information to practitioners in the social, economic, and behavioral sciences to an unwillingness on the part of users in these areas to take the time to learn both the capabilities and limitations of the information that can be extracted from remotely sensed data. For some applications the spatial and temporal scales of data acquisition are not appropriate; for others the models to translate the data collected into meaningful information specific to a given need have yet to be developed. Remote sensing cannot see what is in the human mind. It can, however, document actions that we as human beings take in response to some stimuli. For example, it cannot tell us what crop a farmer may plant in a given year, but it can, in many instances, given appropriate dates of image acquisition and timely interpretation, tell us what has been planted before the crop is harvested. Indeed, given adequate ancillary data and models it can also give us an indication of how good the farmer's harvest may be.

As our ability to collect data at higher spatial, spectral, and temporal resolutions increases, so too does our ability to derive useful information from remote sensing data. Remote sensing will never replace other, more traditional methods of data collection, such as census and survey instruments. Yet it is a proven, powerful method of collecting data which, when properly interpreted and evaluated, can provide a wide variety of useful information. Information which, when taken alone or when augmented by other data/information collected by more traditional means, can create more useful data sets for social, economic, and behavioral scientists.

5. Organizations that Provide Remotely Sensed Data

- Center for International Earth Science Information Network (CIESIN)
- US National Aeronautics and Space Administration (NASA)
- US Geological Survey (USGS)
- SPOT Image
- GlobeExplorer
- US National Oceanographic and Atmospheric Administration (NOAA)
- Space Imaging
- Earthwatch

See also: Geographic Information Systems; Geographic Information Systems: Critical Approaches; Information Society, Geography of; Landsat Imagery in Geography; Mathematical Models in Geography; Remote Sensing and Geographic Information Systems Analysis; Spatial Analysis in Geography

Bibliography

- Bergen E T 1927 Aerial surveys for city planners. *Transaction of the American Society of Civil Engineers* **90**
- Chicago Planning Commission 1943 *Master Plan of Residential Land Use*. Chicago Planning Commission, Chicago
- Estes J E, Ehlers M, Malingreau J-P, Noble I, Raper J, Sellman A, Star J L, Weber J 1992 *Advanced Data Acquisition and Analysis Technologies for Sustainable Development*. MAB Digest 12, UNESCO, Paris
- Fairchild S M 1927 Aerial photography, its development and future. *Annals of the American Academy of Political and Social Sciences* **131** (May)
- Jensen J R 1996 *Introductory Digital Image Processing: A Remote Sensing Perspective* 2nd edn. Prentice Hall, Upper Saddle River, NJ
- Jensen J R 2000 *Remote Sensing of the Environment: An Earth Resource Perspective*. Prentice Hall, Upper Saddle River, NJ
- Joerg E L E 1923 The use of the air plane in city geography. *Annals of the Association of American Geographers* **13**(4)
- Lee W T 1922 *The Face of the Earth As Seen from the Air: A Study in the Application of Airplane Photography to Geography*. American Geographical Society, New York
- Liverman D, Moran E F, Rindfuss R R, Stern P C (eds.) 1998 *People and Pixels: Linking Remote Sensing and Social Science*. National Academy Press, Washington, DC
- National Research Council 1970 *Remote Sensing with Special Reference to Agriculture and Forestry*. National Academy of Sciences, Washington, DC
- Philipson W (ed.) 1997 *Manual of Photographic Interpretation*. 2nd edn American Society for Photogrammetry and Remote Sensing, Bethesda, MD
- Platt R R, Johnson G R 1930 *Peru from the Air*. American Geographical Society, New York
- Reeves R G, Anson A, Landen D (eds.) 1975 *Manual of Remote Sensing*. American Society of Photogrammetry, Falls Church, VA

J. Estes, K. Kline, and E. Collins

Remote Sensing and Geographic Information Systems Analysis

Remote sensing is the science and art of obtaining information about the Earth's surface through the analysis of data acquired by a device which is at a distance from the surface. Images from Earth observation satellites and aerial photographs are most commonly used. These instruments measure the electromagnetic radiation reflected or emitted by the

surface, at different wavelengths (or frequencies). Geographic information systems (GIS) are computer-based systems that are used to capture, store, analyze, and display geographic information. These two techniques are used widely to assess natural resources and monitor environmental changes. Their main fields of application for terrestrial ecosystems are geology and mineral exploration, land-cover mapping, agricultural statistics, the management of irrigation, forest inventories, and the monitoring of natural disasters such as fires, droughts, floods, landslides, or volcanic activity. Remote sensing and GIS are also used for archeology, land-use planning, urban management, and in the transportation sector. Social scientists can gain insights into fine spatial and temporal dynamics of land-use changes and related social phenomena by analyzing time series of remote sensing data and by linking remote sensing to socioeconomic data using GIS.

1. Remote Sensing Systems

A number of satellite sensors have been acquiring multispectral data over the entire globe since the 1970s, at a spatial resolution of dozens of meters. The most widely used among these sensors are Landsat Multispectral Scanner (MSS) (resolution of 80×80 m), Landsat Thematic Mapper (TM) (resolution of 30×30 m), and SPOT High Resolution Visible (HRV) Imaging Instrument (resolution of 20×20 m for the multispectral product and 10×10 m for the panchromatic product) (Colwell et al. 1983, Lillesand and Kiefer 1994).

Radiance data measured from the Advanced Very High Resolution Radiometer (AVHRR) on the NOAA series of orbiting platforms have been used widely for monitoring changes in the attributes of the Earth's surface. These data cover the period from June 1981 to date. The daily NOAA AVHRR Local Area Coverage (LAC) data have a spatial resolution of 1.1 km at nadir, which is well suited to monitor seasonal vegetation dynamics at the scale of the landscape. Other useful sensors with a 1 km resolution are ERS's Along Track Scanning Radiometer (ATSR), SPOT VEGETATION and the Moderate Resolution Imaging Spectroradiometer (MODIS).

Several active microwave sensors are also available: the Synthetic Aperture Radar (SAR) on the ERS-1 and JERS-1 satellites, and the Advanced Synthetic Aperture Radar (ASAR) sensor as part of the forthcoming ENVISAT mission. The cloud-penetrating capability of microwaves make them an ideal tool to work in areas with a permanent cloud cover. Both active and passive microwave sensing are used. Active systems supply their own source of energy or illumination while passive systems sense energy that is naturally emitted or reflected from the Earth's surface, at extremely low levels however. Note that this

enumeration of sensors used for land applications is not exhaustive (see Colwell et al. 1983, Lillesand and Kiefer 1994).

A new generation of fine spatial resolution sensors is due to be launched around the turn of the twenty-first century. These sensors are capable of generating imagery with spatial resolutions as fine as 1 m in panchromatic mode and 4 m in multispectral mode. These systems are also characterized by high geometric precision, short revisit intervals, and rapid data supply. Such imagery will provide greater spatial details on the land surface and open new applications relevant for social sciences (e.g., for urban studies).

2. Land Surface Attributes Measured by Remote Sensing

The electromagnetic radiation reflected and emitted in different wavelength ranges by the surface is controlled by surface attributes, such as the nature of the material, surface moisture, state of the vegetation, canopy structure, etc. Different features can be identified on an image based on the specific 'spectral signature' of different surfaces. A good understanding of the biophysical processes controlling the radiation of a surface is necessary to interpret the remotely sensed signal in terms of surface attributes.

2.1 In the Spectral Domain

In the visible and near-infrared range of the electromagnetic spectrum, the land surface generally is characterized using vegetation indices, which are arithmetic combinations of spectral bands. The most widely used of these indices is the normalized difference vegetation index (NDVI). Vegetation indices derived from remote sensing data are related to several variables quantifying vegetation activity and state, such as the fraction of photosynthetically active radiation absorbed by the vegetation canopy, canopy attributes (e.g., green biomass or green leaf area index), state of the vegetation (i.e., vegetation vigor or stress) and instantaneous rates associated with the activity of the vegetation. Seasonal variations of vegetation indices are related to vegetation phenology and biome seasonality.

In the thermal range of the spectrum, land surface 'skin' brightness temperature (T_s) can be derived from the thermal channels of satellite sensors. T_s is related, through the surface energy balance equation, to surface moisture availability and evapotranspiration, as a function of latent heat flux. Empirical studies have demonstrated the usefulness of thermal data to locate dense forest boundaries and monitor the growth cycle of crops, including for irrigated systems.

In the microwave range of the spectrum, the intensity of microwave returning from a given surface

depends on the geometric characteristics of the surface (i.e., shape and orientation of objects, as well as their surface roughness) and its electrical characteristics (i.e., reflectivity and conductivity of materials, as measured by their complex dielectric constant). The all-weather capacity of microwave remote sensing combined with the fine spatial resolution of the data opens promising avenues for land cover assessment and monitoring. This technology has however not yet reached the same operational level as optical remote sensing.

2.2 In the Spatial Domain

A major attribute of a landscape is its spatial pattern—i.e., the arrangement in space of its different elements. The concept of landscape spatial pattern covers, for example, the patch size distribution of residual forests, the location of agricultural plots, the shapes of fields, or the number and configuration of landscape elements (i.e., their spatial heterogeneity). Landscape spatial pattern is seldom static due both to natural changes in vegetation and human intervention. Remote sensing offers the possibility to analyze changes in spatial pattern at the scale of landscapes. Textural information of fine spatial resolution remote sensing images can be measured using a variety of indices such as variance, variograms, mathematical morphology, or fractals.

3. Processing and Analysis of Remotely Sensed Data

3.1 Preprocessing

Technical issues of geometric correction and rectification, radiometric correction, geometric and radiometric enhancement, and multispectral transformations of image data are described in Richards (1993). The purpose of these operations is to geo-reference and to transform the image data in a form that can be readily interpreted in terms of Earth surface attributes. The analysis and interpretation of remote sensing data typically is performed through image classification, to generate land-cover maps, or through change detection, to monitor land-cover changes.

3.2 Image Classification

Classification is a method by which labels are attached to spatial sampling units (i.e., the pixels) in view of their character. This character is generally their spectral response in different spectral ranges. It may also include their spatial attributes (i.e., texture) or temporal changes. This labeling is implemented through pattern recognition procedures, the patterns being

pixel vectors. The most commonly used classification methodology is based on the maximum likelihood, a probabilistic classification method (Richards 1993). More advanced classification techniques have been developed. For example, contextual classifiers label a pixel in the context of its neighbors in space. They exploit spatial information of an image. Neural networks are classifier networks, having a decision tree structure, where a collection of simple classifiers is brought together to solve a complex problem.

3.3 Change Detection

Time series of remote sensing data can be created to monitor changes in landscapes at local to regional scales. Classic land-cover change detection techniques are based on the comparison of sequential land cover maps derived from remote sensing data or other sources for the same area. For every sampling unit of the maps, the land cover categories at the two dates are compared. The comparison of successive maps fails to detect subtle changes within broad land-cover classes. Thus, rather than detecting changes on the basis of land-cover categories, change detection is better performed on the basis of the continuous variables defining these categories, whether these are reflectance values measured by a satellite sensor or biophysical attributes derived by model inversion. Coppin and Bauer (1996) review techniques used for this comparison. Recent studies have detected changes in land-cover between successive dates by combining different methods (e.g., Macleod and Congalton 1998).

4. Components of a GIS

A GIS is a powerful tool for handling spatial data. It is designed for the collection, storage, and analysis of objects and phenomena where geographic location is an important characteristic (Aronoff 1989, Maguire et al. 1991). The data input component of a GIS converts data from their existing form to one that can be used by the GIS. Two types of data are entered in a GIS: spatial data such as points, lines and areas that represent the geographic locations of features, and associated nonspatial attribute data that provide descriptive information. The data management component includes those functions needed to store and retrieve data from the database. In a GIS, data are maintained in a digital format. The representation of the spatial component of geographic information follows either the raster model—the homogeneous units are regular cells—or the vector model—the spatial units are the points, lines, and polygons. The data analysis and manipulation functions determine the information that can be generated by the GIS. A GIS automates certain activities and data interpretation procedures. Some of the most commonly used GIS functions are classification, overlay operations, neighborhood operations, spatial interpolations, geo-

statistical analyses, and connectivity functions. The output or reporting functions of a GIS produce maps, tables, or statistical results.

GIS facilitates the interpretation of remote sensing data, by linking biophysical information measured by remote sensing to maps of natural and cultural landscape attributes (e.g., soil type, topography, accessibility as measured by various distances), to field measurements (e.g., crop yield or water quality) and to socioeconomic information derived through household surveys (e.g., indicators of well-being, rate of population growth, agricultural input use). For a discussion of the ideas and social practices that have emerged with the development of new forms of data handling and spatial representation in GIS, see Pickles (1995).

5. The Contribution of Remote Sensing and GIS to Social Science Studies

The scientific literature is becoming increasingly rich in case studies using remote sensing to study social phenomena. Actually, remote sensing techniques are growing in importance for studies aimed at understanding land use dynamics, its driving forces and its impact on society (see Behrens 1994, Liverman et al. 1998). The output of remote sensing and GIS analyses are major inputs for spatial models. Remote sensing data is supporting land-use change analysis in a variety of ways.

5.1 Land-cover Mapping

Remote sensing is used widely to map land cover, land use, and agricultural patterns. The current classification algorithms succeed in identifying from 10–15 land-cover classes, including several succession stages of vegetation. Achieving this level of detail requires however a set of georeferenced field observations on a sample of locations. Current sensors do not identify very well settlements or different cropping systems. Spatially heterogeneous or fragmented classes are identified as mosaics of cover types. The further analysis of land use/land cover maps can lead to the separation of ‘landscape units,’ characterized by different combinations of land use and biophysical attributes.

5.2 Monitoring Land-cover Change

Remote sensing facilitates the introduction of a dynamic perspective in land use studies, by monitoring changes in the spatial patterns of landscapes caused by land use. Sequential remotely sensed data allows to detect changes in land use, such as an expansion of the area under permanent or shifting agriculture, urban growth, or increase in eroded areas. For example, quantitative estimates of rates of deforestation are

derived from time series of remote sensing data. The spatial extent and location of secondary growths on abandoned plots is also well detected by remote sensing. Remote sensing has been used in semi-arid regions to assess dryland degradation and to identify transformations of land use and cultivation. Measuring rates of dryland degradation is a complex challenge, however, given the strong interaction between erratic fluctuations in rainfall due to climatic variability and anthropogenic changes in vegetation cover.

5.3 Identification of Land-cover Change 'Hot Spots'

Land-cover change 'hot spots' are defined as areas where high rates of land-cover changes have been observed. Remote sensing-based methods for identifying land-cover change hot spots at broad spatial scales have been developed. Such methods are based on a series of indicators, such as land-cover fragmentation, occurrence of fires, opening of new roads, etc., which are detectable, by remote sensing. Deforestation, for example, is strongly related to the proximity to forested/deforested edge, the proximity to an access and the degree of fragmentation of forests. Forest fragmentation can be quantified easily from remote sensing data.

5.4 Identification of Land-use Attributes

Remote sensing allows identification and measurement of key socioeconomic and ecological characteristics of land use systems. In the Guinea Highlands, Gilruth and Hutchinson (1990) have been able to discriminate, on the basis of remote sensing techniques, between permanent agriculture (in the form of home gardens) and shifting cultivation. The spatial distribution of shifting cultivation areas can also be mapped from fine to medium resolution satellite data. Guyer and Lambin (1993) succeeded in discriminating and quantifying the total area and proportion of tractor-cleared and hand-cleared fields in a region of Nigeria by using several shape criteria derived from multispectral SPOT data. These authors also computed the crop-fallow cycle and the importance of the land reserve using a remote sensing-based land-cover map.

5.5 Spatial Stratification and Spatial Extrapolation

Remote sensing allows testing of how widely local observations apply to larger areas. Actually, a common problem of in-depth investigations of land-use dynamics is that it requires a large data collection effort in the field, at the household and plot levels. Remote sensing, through the measurement of indicators of land use, helps to define regions characterized by a similar land use, that can lead to a spatial stratification of a study area in terms of land use

patterns. As a result, findings obtained over specific sites within a stratum, that is homogenous in terms of land use, can be extrapolated with more confidence to that stratum. If such a spatial stratification is performed before field surveys, it is a powerful tool to design a sampling strategy and locate villages or households that will be used for in depth data collection and investigation. This leads to an optimization of the field data collection effort.

In a study of land use intensification (Guyer and Lambin 1993), remote sensing analysis provided a means of matching the patterns from a haphazard sample of households to the population, based on a few variables that were estimated both by household survey (on the sample) and by remote sensing (on the entire population). The results tended to give greater confidence to inferences from the sample about ecological dynamics.

5.6 Input into Spatial Models

The spatial patterns, which are detectable at a variety of spatial scales on remotely-sensed data, provide the most obvious link between spatially-explicit land-use change models and remote sensing analysis. With the application of remote sensing technology to the monitoring of land-use changes, it becomes possible to integrate more explicitly spatial relationships into models of land-use change. It is also possible to apply and test these models at a finer spatial resolution and with samples comprising smaller geographic units—e.g., at the scale of individual landscape units.

Spatial, statistical models are born from the combination of remote sensing, GIS, and multivariate mathematical models. Their emphasis is on the spatial distribution of landscape elements and on changes in landscape patterns. The goal of these models is the projection and display, in a cartographic form, of future landscape patterns, which would result from the continuation of current land-management practices or the lack thereof. The approach analyses the location of land-cover changes in relation to maps of natural and cultural landscape variables in a GIS. A model is built to describe the relationship between the dependent variable—e.g., land-cover changes—and the independent landscape variables. Spatial, statistical analyses of tropical deforestation or dryland degradation have already been developed for several tropical regions (see review by Lambin 1997). Spatial models of desertification hazard have also been developed by integrating indicators of water and wind erosion, vegetation degradation, range utilization, and human settlement.

5.7 Validation of Spatially-explicit Models

Dynamic spatial simulation, actor-based models have been developed to predict spatial patterns of land-use

changes at the landscape scale—e.g., for tropical deforestation in roadside areas occupied by shifting cultivators (e.g., Wilkie and Finn 1988) or in regions of official resettlement schemes associated with road development (Southworth et al. 1991). These grid-cell models combine spatially-explicit ecological information with socio-economic factors related to land use decisions by farmers. The modeling is based on rules of behavior determined by locational attributes, land tenure system, and vegetation successions. Remote sensing offers a powerful tool to calibrate and/or validate these models by comparing actual and predicted land-use/cover patterns, and by comparing the level of fragmentation of the landscape.

6. 'Socializing the Pixel'

By 'socializing the pixel' one means to discern information embedded within spatial imagery that is directly relevant to the core themes of the social sciences, and use it to inform the concepts and theories pertinent to those themes (Geoghean et al. 1998). Creating a direct link between spatially-explicit land cover information, as derived by remote sensing, and information on land-use change processes requires the development of new methods and models which are merging landscape data with data on human behavior.

6.1 Linking Household-level Data with Remote Sensing Data

A synthesis of the methods for merging socio-economic household survey data and remote sensing-based information is given in Liverman et al. (1998). The major challenges of this approach are: (a) the definition of the appropriate spatial observation units, i.e., the appropriate level of aggregation of information derived from the domains of social phenomena and natural environment, and (b) the development of the appropriate linkages between household-level and remote sensing datasets. While conceptually straightforward, these links can be difficult to implement operationally. In remote sensing, the unit of observation is the pixel that is not directly associated to any social science unit of observation, e.g., individuals, households, or villages. Establishing the correspondence between biophysical and socioeconomic variables requires taking into account the unit of decision-making, units of landscape transformation, and the spatial scale of ecological processes (e.g., a watershed).

Linking remote sensing observations to socio-economic data at the scale of the administrative units which were used for the collection of socioeconomic data leads to a loss of information as it obscures the variability within the units. More recently, research

efforts have attempted to integrate remote-sensing observations and field surveys at finer levels of aggregation, i.e., at the scale of individuals, households, or villages (e.g., Entwisle et al. 1998). There are difficulties in relating remotely-sensed patterns of land-cover changes with field observations of land-use changes since, in many cases, people live in nucleated villages away from their fields, since households may cultivate multiple non-contiguous plots and since people move—especially pastoralists. For this reason, the integration of the two datasets is generally performed at the village level.

Individual household data allow for a better understanding of the land-use practices within each village, as most land-use decisions are made by individuals and households. Moran and Brondizio (1998) have investigated the linkages between remotely sensed data and traditional field methods in the social and biological sciences. Their results have led them to new research questions, including examination of the role of the developmental cycle of the household in shaping their trajectory of land-use and deforestation.

6.2 Landscape Spatial Patterns

The spatial pattern of a changing landscape have some information content on the processes of land-cover change. Certain categories of land-use changes tend to fragment the landscape (e.g., expansion of smallholder farming, small-scale logging, overgrazing around deep wells). Other land-use changes increase landscape homogeneity (e.g., mechanized cultivation or ranching over large areas). There is a good correlation between remotely-sensed spatial patterns and some important characteristics of farming systems. For example, Lambin (1988) found statistically that, in Burkina Faso, spatial patterns interpreted from Landsat MSS data correlate more closely with ethnic groups than with any other physical or cultural landscape variables. Other authors demonstrated that the analysis of the spatial pattern of forest-nonforest interfaces allows identification of regions affected by different deforestation processes, which are controlled by different driving forces.

The spatial patterns of fires, which are well detected from space, are also good indicators of human activities. In tropical forests, regions of fire concentration indicate frontier areas and variations in the spatial organization of fires are associated strongly with political boundaries and farming practices.

6.3 GIS for Health and the Environment

An emerging area of research concerns the application of remote sensing and GIS techniques to analyze the

relationships between people, their environment and their health (de Savigny and Wijeyaratne 1995). GIS provides information about the spatial distribution of and interaction between disease risk factors, patterns of morbidity and mortality, and the allocation of health resources. Vector-borne diseases (e.g., malaria, trypanosomiasis) have received most emphasis so far, as their occurrence is strongly controlled by environmental attributes which are detectable by remote sensing (Barinaga 1993). However, there is a large potential for GIS application on any environment-related health problem. A classic example is the spatial correlation between cholera deaths and contaminated water supplies.

See also: Geographic Information Systems: Critical Approaches; Geography; Globalization: Geographical Aspects; Spatial Analysis in Geography; Telecommunications and Information Policy

Bibliography

- Aronoff S 1989 *Geographic Information Systems: A Management Perspective*. WDL Publications, Ottawa, Canada
- Barinaga M 1993 Satellite data rocket disease control efforts into orbit. *Nature* **261**: 31–32
- Behrens C A 1994 Recent advances in the regional analysis of indigenous land use and tropical deforestation: Introduction. *Human Ecology* **22**: 243–47
- Colwell R N, Simonet D S, Estes J E 1983 *Manual of Remote Sensing*. American Society of Photogrammetry, Falls Church, VI
- Coppin P R, Bauer M E 1996 Digital change-detection in forest ecosystems with remote sensing imagery. *Remote Sensing Review* **13**: 207–34
- de Savigny D, Wijeyaratne P 1995 *GIS for Health and the Environment*. International Development Research Center, Ottawa, ON
- Entwisle B, Walsh S J, Rindfuss R R, Chamrathirong A 1998 Land-use/land-cover and population dynamics, Nang Rong, Thailand. In: Liverman D, Moran E F, Rindfuss R R, Stern P C (eds.) *People and Pixels: Linking Remote Sensing and Social Science*. National Academy Press, Washington, DC
- Geoghegan J, Pritchard L, Ogneva-Himmelberger Y, Chowdhury R R, Sanderson S, Turner II B L 1998 'Socializing the pixel' and 'pixelizing the social' in land-use and land-cover change. In: Liverman D, Moran E F, Rindfuss R R, Stern P C (eds.) *People and Pixels: Linking Remote Sensing and Social Science*. National Academy Press, Washington, DC
- Gilruth P T, Hutchinson C F 1990 Assessing deforestation in the Guinea highlands of West Africa using remote sensing. *Photogram Engineering and Remote Sensing* **56**: 1375–82
- Guyer J, Lambin E F 1993 Land use in an urban hinterland: Ethnography and remote sensing in the study of African intensification. *American Anthropologist* **95**: 839–59
- Lambin E F 1988 L'apport de la télédétection dans l'étude des systèmes agraires d'Afrique; l'exemple du Burkina Faso. *Africa* **58**: 337–52
- Lambin E F 1997 Modelling and monitoring land-cover change processes in tropical regions. *Progress Physical Geography* **21**: 375–93
- Lillesand T M, Kiefer R W 1994 *Remote Sensing and Image Interpretation*. Wiley, New York
- Liverman D, Moran E F, Rindfuss R R, Stern P C (eds.) 1998 *People and Pixels: Linking Remote Sensing and Social Science*. National Academy Press, Washington, DC
- Macleod R D, Congalton R G 1998 A quantitative comparison of change-detection algorithms for monitoring eelgrass from remotely sensed data. *Photogram Engineering and Remote Sensing* **64**: 207–16
- Maguire D J, Goodchild M F, Rhind D W 1991 *Geographical Information Systems: Principles and Applications*. Longman, Harlow, UK
- Moran E F, Brondizio E 1998 Land-use change after deforestation in Amazonia. In: Liverman D, Moran E F, Rindfuss R R, Stern P C (eds.) *People and Pixels: Linking Remote Sensing and Social Science*. National Academy Press, Washington, DC
- Pickles J 1995 *Ground Truth: The Social Implications of Geographic Information Systems*. The Guildford Press, New York
- Richards J A 1993 *Remote Sensing Digital Image Analysis*. Springer-Verlag, Berlin
- Southworth F, Dale V H, O'Neill R V 1991 Contrasting patterns of land use in Rondonia, Brazil: simulating the effects on carbon release. *International Social Science Journal* **130**: 681–98
- Wilkie D S, Finn J T 1988 A spatial model of land use and forest regeneration in the Ituri forest of Northeastern Zaire. *Ecological Modelling* **41**: 307–23

E. F. Lambin

Renaissance

Two nineteenth-century historians were instrumental in giving the term 'Renaissance' the content that it retains to this day. In his 1855 *Histoire de France*, Jules Michelet (1798–1874) chose the word for the period of the *découverte de l'homme et du monde*, the sixteenth century. But the Swiss Jacob Burckhardt (1818–97) was even more important. His famous book *Die Cultur der Renaissance in Italien. Ein Versuch* (1860) became a milestone of research. Burckhardt sketched a monumental fresco that located the origin of modernity in fifteenth century Italy—in conscious distinction from Hegel's philosophy of history, which posited that the German Reformation marked the beginning of the modern age. Burckhardt referred to the conditions in the Italian ministates as 'masterpieces of statecraft,' in which he saw the preconditions for the 'emergence of the individual' as the most important and far-reaching development of the modern age. He presented the Renaissance, with the many-faceted Humanistic movement as its intellectual substrate, as a period that forced the pace of modernization and secularization and as the beginning of the modern

world in which he himself lived and whose negative sides he analyzed perhaps more acutely than any other thinker of his time.

1. The Development of the Concept of the Renaissance

1.1 Renaissance

Burckhardt's thought-provoking image of the epoch exerted substantial influence, not only on the study of history, but also on literature and philosophy, in particular on Nietzsche's thinking. It remains the more or less explicit background of discussions of the term 'Renaissance,' though it soon roused dissension. For example, in the book *Franz von Assisi und die Anfänge der Renaissance in Italien* (1885), Henry Thode (1857–1920) attempted a romantic reinterpretation, placing the *poverello* and the religious movement he inspired at the beginning of the epoch. Thode and his followers saw the Renaissance more as a reform movement on a medieval-clerical foundation, eliminating its character as an epoch of its own. Thode rejected the idea that classicism and rationality fundamentally constituted the Renaissance and divided it from the Middle Ages. He posited an individualism of piety and sensitivity to beauty, in contrast to Burckhardt's subjects becoming aware of themselves.

But the prospect of another, religious side of the Renaissance had already existed before Thode and did not completely contradict Burckhardt's view. In his influential essay collection *The Renaissance* (1873), Walter Pater (1839–94), an admirer of Ruskin presented several artists and thinkers he considered representative of the Renaissance. The essay on Botticelli became essential to the incomparable cult of this painter, which thrived throughout Victorian society, and not solely among the Pre-Raphaelites. Botticelli's art seemed to correspond to a 'metaphysical need' in the modern, industrialized world of Victorian England.

The inventions of the historians and literary writers stimulated broad segments of the educated bourgeois class. Florence became the preferred destination for an international cultural tourism. People built in the Renaissance style, surrounded themselves with splendid interiors in the style of the fifteenth and sixteenth centuries, and purchased original works of art. Especially in the second half of the nineteenth century, collections of important Renaissance art were amassed from Boston to Berlin. The 'Renaissance' was not only a focus of scholarly interest, but also part of a lifestyle.

Around 1900, the international Renaissanceism reached a zenith. Burckhardt's classic was reprinted, and novels like Heinrich Mann's 'Göttinnen' depicted the sixteenth century as a foil for the contemporary 'decadent' *fin de siècle*. The closing nine-

teenth century saw itself not only in a strange spiritual kinship with the Renaissance: the fiction of a 'lascivious,' 'unscrupulous' epoch paradoxically appeared to offer counterimages to the usual bourgeois existence, taking shape in princes of commerce, patrons, *condottieri*, and courtesans.

At this time, the term 'Renaissance,' was loaded with a plethora of sometimes contradictory meanings, triggering associations of 'gold and purple,' as Johan Huizinga said. The patronage of the Medici, the magnificence of the papal court, and the art of the Renaissance with its heroes Leonardo, Raphael, and Michelangelo provided new cult objects for religious worship in a demystified world. At the same time, the word recalled the art of power and *Realpolitik*, as Machiavelli had analyzed in his trailblazing work. Racist ideas of a victory of hard men and cunning women over decline and enfeeblement were inherent in this image of the epoch. A widely read literature in the succession of Carlyle, Count Gobineau, Nietzsche, and others popularized the cliché of the 'blond beast,' the amoral 'Renaissance man' up to mischief.

The myth of the Renaissance thus offered models for Caesarism and imperialism. It was no coincidence that a nationally-oriented writing of history took up Italy's *Signorie* and identified them with the precursors of the modern military and power-wielding state. But the iridescent construct of the 'Renaissance' also offered fragments for a historical foundation of modern civic societies. As early as the beginning of the nineteenth century, the Swiss Sismonde de Sismondi had described the Italian city republics in this way, and travelers from England and the USA thought they were tracing the ideal roots of their own nations when they came to Florence. It is hardly surprising that the *inglesi* were the largest group among the tourists. The Florence of the Medici was long a primarily Anglo-American myth.

Examples of the ideological interpretation of the Renaissance with an unmistakable relationship to current history are also found in the works of Hans Baron, a German Jew who emigrated to the USA. In the essay he published in the middle of World War II, he saw the crisis of Italian 'freedom' caused by the aggression of the Visconti *signoria* in the early Quattrocento as the precondition for the formation of a modern, pluralistic state system on the Italian peninsula. In Baron's view, the high degree of citizen participation in Florentine politics was a precondition for this most important result of the struggle against the tyrants: He saw it as a kind of Italian America winning the upper hand over Visconti/Hitler (Baron 1955).

1.2 The Revolution of the Medievalists

More recent research on the Renaissance has tried to free itself of mythologizing. Insight has been sought into the clientele systems of the clan that enabled

Cosimo de' Medici and above all Lorenzo the Magnificent to exercise an authoritarian power behind a republican façade (Wackernagel 1938, Thomas 1995), whereby special attention was paid to the patrons commissioning the art (Esch and Frommel 1995). But art history tended to go its own way. Wölfflin and others separated Renaissance art from the cultural-history context and treated it primarily as a problem of style, and thus of aesthetics.

The conception of the art historian Aby Warburg (1866–1929) remained fragmentary and accessible only in a very few essays. It aimed once again at the great synthesis and still provides impetus for new research. He located the epoch in a long-term process of civilization that experienced its climax in the Apollonian beauty of the art of the High Renaissance. In the triumph of the Olympic gods, Warburg saw an important step toward overcoming 'oriental,' 'medieval' irrationality. He was interested in the difficult breakthrough to the Enlightenment, humanity and reason in the history of the West. Warburg saw works of art as documents of the states of the human psyche, and he believed that, through subtle analysis, he could uncover the *Zeitgeist* preserved in them. Warburg's works addressed aspects previously not or only marginally considered. The role of the patron became a theme, along with what Michael Baxandall (1972) calls 'troc': the market of intellectual and material goods where what we call 'Renaissance art' took shape. Warburg sought to illuminate the dark side of the epoch—its interest in divination, astrology, and magic, to which Burckhardt had devoted only a few passages. Equally important was that he conceived the emergence of Renaissance culture as a transfer process and thus examined not only the Orient, but also the culture of the North—the courts of Burgundy, France, and the Netherlands. This underscored what is now the general possession of research: the cultural physiognomy of the Renaissance as an amalgamation of varying influences. Italy remained important, but it was placed in the context of a network cast across the continent (Porter and Teich 1992).

This geographical expansion, the focus on the 'occult Renaissance,' and above all his constant question of the beginnings led Warburg to avoid drawing the boundary between the 'Renaissance' and the 'Middle Ages' too sharply. Although his works had no discernible importance for it, he prepared the ground for what Wallace K. Ferguson has termed a 'revolt of the medievalists' (Ferguson 1948). This debate, characteristic of the 1920s, adopted aspects of the image of the Renaissance as sketched by Thode, Sabatier, and other researchers. It pointed out, for example, that the tradition of antiquity was never completely buried and recalled that various 'renaissances' were palpable between the ninth and thirteenth centuries, in which ancient forms clearly stand out as elements shaping culture. Some researchers, including Etienne Gilson and Johann Nordström saw the

Renaissance as a mere continuation of currents existing in the twelfth century. In this famous book *The Waning of the Middle Ages* (1919), the Dutch writer Johan Huizinga drew an image of the fifteenth century diametrically opposed to Burckhardt's conception. Contemplating the courtly world of Burgundy, France, and the Netherlands, he depicted the epoch as a flowering 'autumn' of burgeoning color, rather than as the spring of a new age.

2. *A Movement of the Happy Few: The Sociohistorical Dimension*

Discussion thus ensued on the boundaries of epochs; it was the hour of the deperiodizers, who made the Renaissance their preferred target. George M. Trevelyan went as far as to locate the turn from the Middle Ages to the Modern Age in the epoch of the Industrial Revolution, the eighteenth century; and for other researchers as well (like Delio Cantimori), the Renaissance of the fourteenth and sixteenth centuries lost its meaning as a significant turning point. A more differentiated sociohistorical research underscored that Renaissance societies were more 'medieval' than widely conceived. Additionally, this culture had been borne by a tiny minority—one of the reasons Ernst H. Gombrich no longer speaks of the Renaissance as an epoch, but as a movement.

It is indeed indisputable that epochs named for styles can describe little more than the cultural stances of an often compoundly interwoven elite. Social history has provided glimpses of the world of the majority of the 'Renaissance people'; these reveal little 'Renaissance' quality (e.g., Herlihy and Klapisch-Zuber 1978, Davis 1982, Squatriti and Merideth 1999). Recently, the theme of gender has also been considered more intensely (see King 1991, with comprehensive bibliography); here, research focuses not only on the 'Renaissance women,' the princesses and courtesans who populate the mythological scenario, but also on other women who lived at the time—witches and saints, nuns, artists, authors, and women in working life in general (Ferguson 1986).

3. *'Modern Individualism' and the Concept of the Renaissance Today*

3.1 *Art Historical Findings*

Despite the still current discussion of the term 'Renaissance,' the concept has prevailed as the name of a cultural epoch (cf., e.g., Burke 1972, Delumeau 1967, Hale 1993). What arguments can be adduced for or against the concept of the Renaissance? The most conspicuous are the findings of art history, and they

were what gave the essential impetus to construe the epoch in the first place. Thus, it is beyond doubt that the painting of Giotto and Massaccio and the sculpture of the Pisani in the fourteenth and fifteenth centuries marked a fundamental transformation. What was decidedly new in painting was that dependence on the Byzantine model was overcome. Here, Flemish influences played a major role, and the Renaissance in general increasingly appears as a complex amalgamation of transfer processes embracing Europe, the Middle East, and finally the New World. The early Quattrocento brought the introduction of a central perspective formulated with geometric precision, a Florentine invention. The goal was to create the illusion of 'real' living beings, objects, and spaces.

One of the most important arguments for Burckhardt's old 'individualism hypothesis' is still that not only did artists' signatures appear even more frequently, but also that, in the fifteenth century, initially in the Netherlands and Italy, portraits were painted that appear closer to nature than anything since classical antiquity. Since the Trecento, sculpture once again clearly approached antiquity's conceptions of the body, and, increasingly frequently, portrait busts clearly strove to convey the individuality of the person portrayed. While it cannot be claimed that the 'modern individual' was invented in the fifteenth century, if we also adduce literary testimony in the broadest sense, for example Ghiberti's *Autobiography* and Giovanni Rucellai's unique *Zibaldone*, we do note that reflection on the self gained decisively in importance (Davis 1986 is highly differentiating; see also Schulze 1996, Porter 1997, Schweikhart 1998).

In architecture, the building forms of classical antiquity reappeared first and initially exclusively on the Italian peninsula. Only at the beginning of the sixteenth century was anything remotely comparable found north of the Alps, starting with the buildings commissioned by the Fuggers in Augsburg. Regional variant forms emerged rapidly. Their common denominator was a more or less clearly articulated reference to antiquity. With Mannerism, the Renaissance flows into a typically European style (cf. Sect. 5).

3.2 The Revolution in Discourse and Renaissance Humanism

The art of the Renaissance developed in a complex content with a theoretical discourse that, on the one hand, used texts from antiquity, for example Vitruvius' theory of architecture. This was new and genuinely Italian. In the fifteenth and sixteenth centuries, no other European country produced a theory and critique of art whose quality and differentiation were even remotely comparable. Italy was also the first place with a palpable group of collectors, art agents,

and patrons who, for aesthetic considerations, acquired real or putative ancient objects. These inspired an almost romantic veneration and learned discussion and correspondence. One of the earliest representatives of this modern phenomenon of collecting was the Florentine Niccolò Niccoli (1364–1437).

Interest in ancient and contemporary art was part of the humanistic culture that was closely tied to the Renaissance. The term *studia humanitatis* is of ancient origin and refers to education in the broadest sense, the acquisition of knowledge as well as of 'good' ways of living. The fifteenth and sixteenth centuries saw the emergence of the social type of the *humanista*, the scholar concerned with the canon of humanistic studies—grammar, rhetoric, poetics, history, and moral philosophy—with recourse to the authors of classical antiquity. Humanistic scholars and humanistically educated elites became the real bearers of the Renaissance. They gave art, architecture and literature essential elements of form and content (see e.g., Wind 1968, Panofsky 1962). Modern research sees the first beginnings in upper Italy in the second half of the thirteenth century. But Francesco Petrarca (1304–74) was the first thinker to unite in himself the essential aspects of the humanistic movement. The older idea that the humanists were a group of traveling scholars staging a revolution against scholastic theology and Church doctrines is no longer tenable. Social history cannot view the host of humanists as homogeneous, even as early as the fifteenth century. The majority of the humanists of Florence, the 'capital' of the Renaissance, were economically well off (Martines, 1968). Monks, prelates, princes, university scholars, urban jurists, and many others spread the *studia humanitatis* in Europe, and in many places humanism established itself as a national movement with national characteristics.

Humanism, and with it the culture of the Renaissance, were in no way pagan. The vast majority of humanists were firmly based in Christianity, and this did not change with the Reformation (Verdon and Henderson 1997). Significantly, pictorial subjects remained mostly religious during the Renaissance. In the sixteenth century, however, we note their gradual dwindling (Burke 1972), a telling indication that Renaissance humanism contained substantial potential for secularization. The great expansion of the spectrum of questions discussed and written about in the fifteenth and sixteenth centuries was of long-term importance. Theology and religion still dominated, but texts on every possible aspect of life increasingly moved to the foreground; finally, and again with recourse to ancient thinkers, a new theory of politics developed (Pocock 1975). This 'differentiation of discourse' did not depend on the Gutenberg Revolution, but began much earlier. In a certain way, the invention of the letterpress with movable type can be viewed as a response to a market ravenous for written material.

4. *The Research Problem of the Genesis of the Renaissance*

A question equally fascinating and controversial is why the beginnings of humanism and the Renaissance lay in Italy and what led to the formation of these movements. As Burckhardt had already underscored, there was the multiplicity of *signoria* and city republics, whose political rivalry sought artistic representation. Ancient monuments, statues, and other relics also provided models. Italy and Flanders, the regions in which the new culture developed, were probably the richest in Europe at the time, so here is where the economic prerequisites for extensive patronage were given. But the connection between art and economy should not be absolutized. There is no necessary parallel between the unfolding of Renaissance culture and a general economic upswing; Roberto S. Lopez (1963) even attempts to interpret the Renaissance as an attempt to compensate for crisis-like developments. The financial position of the patrons was decisive (see e.g., Roover 1963). The *condottieri* wars in fifteenth-century Italy resulted not merely in a destruction of resources, but also in processes of redistribution from the great centers to small seats of courts like Urbino and Bergamo, where successful military commanders spent astronomic sums on books and in patronizing the arts. It is difficult to estimate the significance of the dynamics of economic shrinkage triggered by plague epidemics and failed harvests in the first half of the fourteenth century. Along with the resulting acceleration of urbanization, the consolidation of capital in the hands of the survivors contributed to the financing of early Renaissance art—in part as an ‘investment in saving one’s soul,’ which must have suggested itself and seemed necessary against the backdrop of the crisis of Church hierarchy and the experiences of the plague era. The Renaissance, humanism, and the Reformation thus appear as varying answers to one and the same problem, namely the determination of the human position in a fundamentally changed world. The writings of Giannozzo Manetti and Pico della Mirandola, the political theories of Machiavelli, the primary Reformation writings of Luther and the invention of the realistic portrait and of central perspectives thus all had the same preconditions.

5. *The ‘Mannerist’ Renaissance*

The question of the end of the Renaissance is no less complex and controversial than that of its beginnings. Art history sees indications of a metamorphosis to the ‘Mannerist Renaissance’ or to mannerism as early as the second decade of the sixteenth century (Shearman 1967, Arasse and Tönnemann 1997). Literary history usually locates the end of the epoch around 1600. And music history, too, has its difficulties with the question

of the definition and temporal boundaries of mannerism (important but controversial, see Maniates 1979). National differences in particular must be differentiated. Setting the boundaries of the epoch is also made difficult by the fact such important intellectual currents constitutive of the Renaissance as humanism and such aesthetic ideals as the orientation toward the standards of Greek classicism lived on after 1600 and experienced their own renaissances.

There has been no lack of attempts to set the stylistic transformation of the sixteenth century in direct relationship to political and societal changes (the best known is Arnold Hauser 1979). Structural changes at least in the art market and thus in the relationships between artists and patrons probably played an important role. Europe’s princes, the rulers of the rising early modern states, moved increasingly into the foreground as patrons. They used the arts to buttress their legitimacy and to display their grandeur: now, in the sixteenth century, the art of the power availed itself of the power of art more than hardly ever before. Political competition also manifested itself as competition among patrons of the arts. The artists’ virtuosity, or *maniera*, had to prove itself in this complicated field.

See also: Art History; Burckhardt, Jacob (1818–97); Middle Ages, The; Reformation and Confessionalization

Bibliography

- Arasse D, Tönnemann A 1997 *La Renaissance Maniériste*. Gallimard, Paris
- Baron H 1955 *The Crisis of the Early Italian Renaissance. Civic Humanism and Republican Liberty in an Age of Classicism and Tyranny*. Princeton University Press, Princeton, NJ
- Baxandall M 1972 *Painting and Experience in Fifteenth Century Italy. A Primer in the Social History of Pictorial Style*. Oxford University Press, Oxford, UK
- Brucker G 1977 *The Civic World of Early Renaissance Florence*. Princeton University Press, Princeton, NJ
- Burke P 1972 *Culture and Society in Renaissance Italy*. B. T. Batsford, London
- Burke P 1998 *The European Renaissance*. Blackwell, Oxford, UK
- Davis N Z 1982 Women in the crafts in sixteenth-century Lyon. *Feminist Studies* 8: 47–80
- Davis N Z 1986 Boundaries and the sense of self in early modern France. In: Heller T C et al. (eds.) *Reconstructing Individualism*. Stanford University Press, Stanford, CA
- Delumeau J 1967 *La Civilisation de la Renaissance*. Editions Arnaud, Paris
- Esch A, Frommel C L 1995 (eds.) *Arte, Committenza ed Economia a Rome e nelle corti del Rinascimento*. Giulio Einaudi, Turin, Italy
- Ferguson W K 1948 *The Renaissance in Historical Thought. Five Centuries of Interpretation*. Houghton Mifflin/The Riverside Press, Cambridge, MA

- Ferguson W K 1986 (ed.) *Rewriting the Renaissance. The Discourses of Sexual Difference in Early Modern Europe*. University of Chicago Press, Chicago
- Goldthwaite R A 1993 *Wealth and the Demand for Art in Italy, 1300–1600*. Johns Hopkins University Press, Baltimore, MD
- Hale J 1993 *The Civilization of Europe in the Renaissance*. Harper Collins, London
- Hauser A 1979 *Der Manierismus. Die Krise der Renaissance und der Ursprung der modernen Kunst*. Beck, Munich, Germany
- Herlihy D, Klapisch-Zuber C 1978 *Tuscans and their Families: A Study of the Florentine Catasto of 1427*. Yale University Press, New Haven, CT
- Holmes G 1986 *Florence, Rome and the Origins of the Renaissance*. Clarendon Press, Oxford, UK
- King M L 1991 *Women of the Renaissance*. Chicago University Press, Chicago
- Kristeller P O 1961, 1965 *Renaissance Thought Vols I and II*. Harper Torchbook, New York
- Kristeller P O 1972 *Renaissance Concepts of Man and Other Essays*. Harper Torchbook, New York
- Lopez R S 1963 Hard times and investment in culture. In: Ferguson W K (ed.) *The Renaissance. A Symposium*. Seemann, New York, pp. 29–54
- Maniates M R 1979 *Mannerism in Italian Music and Culture 1530–1630*. Durham, NC
- Martines L 1968 *The Social World of the Florentine Humanists, 1390–1460*. Princeton University Press, Princeton, NJ
- Panofsky E 1960 *Renaissance and Renaissances in Western Art*. Amquist & Wiksell/Gebers Förlag AB, Stockholm
- Panofsky E 1962 *Studies in Iconology*. Harper Torchbook, New York
- Pocock J G A 1975 *The Machiavellian Moment*. Princeton University Press, Princeton, NJ
- Porter R 1997 *Rewriting the Self, Histories from the Renaissance to the Present*. Routledge, London
- Porter R, Teich M 1992 *The Renaissance in National Context*. Cambridge University Press, Cambridge, UK
- Roover R de 1963 *The Rise and Decline of the Medici Bank, 1397–1494*. Harvard University Press, Cambridge, MA
- Rubinstein N 1997 *The Government of Florence Under the Medici (1434–1494)*. Clarendon Press, Oxford, UK
- Schmitt C B, Skinner Q (eds.) 1988 *The Cambridge History of Renaissance Philosophy*. Cambridge University Press, Cambridge, UK
- Schulze W (ed.) 1996 *Ego-Dokumente. Annäherungen an den Menschen in der Geschichte*. Akademie-Verlag, Berlin
- Schweikhart G (ed.) 1998 *Autobiographie und Selbstporträt in der Renaissance*. Walther König, Cologne, Germany
- Shearman J 1967 *Mannerism*. Penguin, London
- Squatriti P, Merideth B 1999 *The Renaissance in the Fields. Family Memoirs of a Fifteenth-century Tuscan Peasant*. Penn State University Press, PA
- Thomas A 1995 *The Painter's Practice in Renaissance Tuscany*. Cambridge University Press, Cambridge, UK
- Verdon T, Henderson J (eds.) 1997 *Christianity and the Renaissance. Image and Religious Imagination in the Quattrocento*. Syracuse University Press, Syracuse, NY
- Wackernagel M 1938 *Der Lebensraum des Künstlers in der florentinischen Renaissance. Aufgaben und Auftraggeber, Werkstatt und Kunstmarkt*. Leipzig, Germany
- Wind E 1968 *Pagan Mysteries in the Renaissance*. Faber & Faber, London

B. Roeck

Repartnering and Stepchildren

1. Introduction

Sexual unions create children and hence, parenthood. Union dissolution creates single parents and absent parents. When these parents enter a new partnership, they may have a new kind of children—stepchildren—children who are not the joint concern of the couple at the start of the union, as at least one member enters the union as a step-parent. This asymmetry in parenthood is likely to shape the partner dynamics that lead to union formation and hence to affect the calculations of those with children (whether coresident or not) and any partners who might join them in a new union. This asymmetry affects their lives as a couple, particularly whether they have additional, joint children, and also affects whether their union survives. This article on children and new partnerships examines three issues: (a) the role of children in the likelihood that their parents repartner (considering different effects for men and women); (b) the role of such children on the fertility of the unions formed, and (c) the effect of stepchildren on union dissolution.

2. Background

The rapid rise in divorce in most industrialized countries together with the rise in cohabiting unions, which tend to be more unstable than marriages, is creating a new demographic situation for men and women. A union disruption places them in the partner market again, which is particularly challenging for those who are the parents of young children. Although historically, high mortality often left young parents widowed, divorce creates twice as many new prospective partners. Further, whatever wisdom had accumulated for those facing this challenge, has likely been lost. The divorce revolution that began in the 1960s and 1970s in the industrialized countries in the west followed a half century or more of rapidly declining mortality, which sharply reduced widowhood for both sexes and concentrated it at ages when children are older, or even adults. In addition, the early years of the increase in divorce rates disproportionately involved couples without children far more than has been the case in the 1980s and 1990s (Andersson 1997). Hence, the effect of existing children on new partnerships is not well understood, either in the community of scholars of the demography of the family or in the communities experiencing the new situation. Little is known about how children affect their parents' repartnering, and not much more is known about those partnerships as they develop over time in terms of their fertility and stability. The research community has been further hampered by a lack of appropriate data to study these issues.

The numbers involved are substantial. In the United States, it is estimated that 40–50 percent of all marriages dissolve and a substantial minority of all children spend some time in a single-parent family (Bumpass and Raley 1995). Half of both previously married cohabiting couples and remarried couples have children in the household and many of the rest have children living elsewhere (Bernhardt and Goldscheider 1998). Although such rates are higher in the US than in other industrialized countries, they are rising rapidly outside the US (see *Demographic Transition, Second*), while US levels appear to have reached a plateau in the last two decades of the twentieth century. At whatever level, these trends mean that a substantial proportion of adults are at risk of entering a new partnership, and many of those who do will have children, either living with them or dependent upon them financially. Those considering forming a union have to take their potential partner's children into account.

A critical question might be whether the presence of children increases the likelihood that the unions formed are cohabiting ones. There is some evidence that this is, in fact, the case. Although the rise in cohabitation has not fully offset the decline in proportions married among the never married, the offset is total among the formerly married, suggesting that remarriage is being replaced by cohabitation (e.g., see *Demographic Transition, Second*). However, most research, and most of the data on which research could be based, has made it difficult to study cohabiting partnerships that involve children. Research on 'single-parent' families has frequently studied only 'unmarried' parents. The most common control for resources in families was 'family income,' which typically excluded the income of 'non-family' members, i.e., cohabiting partners. New surveys, such as the Family and Fertility Surveys (FFS), should overcome these difficulties.

3. Repartnering of Parents

Nearly all research on repartnering when pre-existing children are involved has considered only women with coresident children, focusing almost entirely on remarriage and on the US. These studies have typically found that the presence and sometimes the number of children reduce women's likelihood of marrying, with effects often depending on the ages of the children and of the mother. One of the few studies that considers cohabitation suggests that children reduce the partners' commitment to the union, decreasing the risk of getting married but increasing the risk of cohabiting (Bennett et al. 1995).

There is much less evidence on how children affect men's likelihood of union formation. A major problem for men is distinguishing between coresidential and

non-co-residential children, a distinction that has been less important for women. One study examined widowers, where presumably fewer of the children lived with other relatives (Smith et al. 1991). That analysis showed that having dependent children reduces middle-aged (< 60) men's likelihood of remarriage after widowhood, although the reduction was much less than is the case for women with children. However, more recent research on both the US and Sweden, which has included divorced and never-married men with coresident children, as well as widowed men, has found that such men enter partnerships significantly more quickly than men without children (Goldscheider and Sassler 2000, Bernhardt and Goldscheider 1999). This suggests that the effects of coresidential children on union formation are quite different for men than for women.

There is even less research on how non-coresidential children affect union formation, although this is a particularly important issue for men (and the women who consider entering unions with them). The studies mentioned above also examined this issue, and found that in both the US and Sweden, men with non-coresident children are significantly more likely than otherwise comparable men with no coresident children to enter a union than to remain outside a partnership. In Sweden, such men are also more likely to enter a union with either a woman who has coresident children or one who does not than to remain unpartnered; in the US this effect does not appear for union entry for women without children. The few women who have non-coresident children are also more likely to enter a union than to remain outside a partnership. In this case, however, at least in Sweden, they are much more likely to choose a partner with no coresident children, although this effect may reflect the small numbers of nonpartnered men who live with their children.

Perhaps a more important question, however, is the broader one: what factors affect entry into a union with a partner who has children? Research on the US and Sweden shows that two family-related factors distinguish men by whether the partnerships they enter into include children: their parents' marital history and their own marital history (Bernhardt and Goldscheider 1999, Goldscheider and Sassler 2000). Men who experienced a nontraditional family structure in childhood are significantly more likely to form a marriage that converts them into stepfathers than one that does not, and the same general difference appears for cohabitation.

US men who have been previously married resemble those who are already parents, in being very likely to enter a union (Goldscheider and Sassler 2000). If they marry, there appears to be no difference in the likelihood that their partner has children or not. If the new union is a cohabiting one, however, they are more likely to enter one with children. In fact, divorced men are more likely either to marry or cohabit with children

than to remain unpartnered; and they are significantly less likely to cohabit without children than to remain outside a partnership. Similar results were obtained for Sweden (Bernhardt and Goldscheider 1999).

Men who enter unions with children are not simply different in terms of their family histories, however; the evidence for both Sweden and the US suggests that they are also drawn from among the less educated, but the studies do not show why. In Sweden, those with a post-gymnasium education are the most likely to enter a union with a partner without children but the least likely to enter a union where the new partner has children (Bernhardt and Goldscheider 1999). This pattern is the same for US men; college graduates are more likely to enter a union than to remain outside a partnership if the woman does not have children. They are more likely, however, to remain outside a partnership than to enter a union with a woman who has children (Goldscheider and Sassler 2000).

These results suggest that there are important differences in the likelihood of repartnering for those with and without children, with patterns that differ between men and women. Children, whether co-resident or not, increase repartnering for men, but only non-coresident children do so for women; for them coresident children have the opposite effect. This suggests that surveys need to go beyond providing information in their partnership histories about whether partners brought children into the union and at least also provide information on whether partners have children living elsewhere. Such children clearly affect union formation; it is likely that they would also affect partnership dynamics (fertility and union stability), though this cannot yet be assessed fully, as we shall see in the next section. It is also likely that other characteristics of the children, such as their sex, might affect their parents' repartnering. No studies to date, however, have assessed whether the gender of the child(ren) matters. This question will take on increased importance later in this review.

More fundamentally, it becomes increasingly important to understand the reasons why children reduce the likelihood of repartnering, which we now see is only the case when they are coresident children and are children of the woman and not the man (unless he is older and widowed). How important is the fact that, although in most industrialized countries stepfathers have few legal rights or responsibilities vis-à-vis their stepchildren, most men realize that the children will be claimants on their incomes, at least while they remain in the household? The effect of having coresident children appears to be less for Swedish women than for women in the US (cf. Bernhardt and Goldscheider 1999 and Bennett et al. 1995). It may be that the greater support provided to children of all kinds by the Swedish state offsets some of the perceived liability for prospective stepfathers.

A perhaps equally important problem is ambiguity over the nature of stepfather-stepchildren relation-

ships. Stepfathers are discouraged from assuming parental responsibilities for their stepchildren's lives by the children, their mothers, and by the larger society in most western, industrialized countries. Paternal involvement is an important factor in family dynamics, particularly for union stability, and we will examine this topic below.

A further interesting question is how much of the decline in marriage formation is a result of the increase in potential partners with children? Most studies that have examined whether the quality of the partner market has declined have considered women's decisions and have examined trends in the earnings of unmarried men who would be their potential partners. It might be equally interesting to study men's decisions and examine whether trends in men's marriage likelihood are influenced by the changing composition of potential female partners, as increasing proportions have children living with them.

4. Fertility of Couples with Children from Previous Relationships

There have been more studies on the effects of children from a previous union on childbearing in a new union than on repartnering. Few, however, distinguish between coresident and non-coresident children, and none have taken into account the selectivity of repartnering when children are already present as revealed in the studies reviewed above. Nevertheless, these studies have assembled an impressive set of results for a wide range of industrialized countries in Europe, as well as in the US.

Not surprisingly, the general finding is that couples in later unions have lower fertility within those unions than do couples in otherwise comparable first unions, which indicates clearly that many individuals have childbearing goals they do not want to exceed. Their previous children are considered in their decisions to continue childbearing. This result has appeared in studies in Germany, Czechoslovakia, the US, France, and Sweden (cf. references in Vikat et al. 1999).

How much previous children matter and whether it makes a difference whether they are coresident or not remain open questions. They are much more difficult questions to assess for the moment, as there is rarely information on previous children for both members of a couple, particularly on their non-coresident children. The most complete study is a recent analysis of Swedish data (Vikat et al. 1999), which, although it only had data on non-co-resident children for one member of the couple, produced a number of interesting findings.

First, the study found that the effect of previous children is definitely not total, as couples in Sweden appear to have at least one shared biological child independently of previous childbearing by either

partner. The authors interpret this result as indicating that joint children are an important element of a union. Second, the authors were able to rule out some alternative reasons for childbearing in later unions. There was little evidence that the risk of having a first child in a union was higher when one of the partners was childless, suggesting that Swedes do not have a child simply as an indicator of adult status. Further, there was little evidence the couples have a child to produce a sibling for a pre-union child, although couples did appear to avoid having only one shared child. This was interpreted as suggesting that it is full siblings that are considered important by Swedish couples in new unions, not half-siblings. Research on the US indicates that while stepsiblings remain fairly close to each other in adulthood, they are much less close than full siblings (White and Riedmann 1992), although it is unclear how much this difference reflects their different histories prior to coresidence and how much it reflects problems in their relationships while sharing a home. It was not possible to assess how much fertility in repartnerings is affected by co-resident as opposed to non-coresident stepchildren, as full information on the latter was not available.

Fertility in repartnerings is clearly a complex subject, and one that might depend not only on the numbers of children each of the partners has had previously and where they live, but also the genders of those children. Strong gender effects appear in research on stepfamily relationships, and there is evidence that gender of children affects union dissolution (Brooks-Gunn 1994). No studies to date, however, have assessed this issue, as no research has addressed the selectivity of repartnering.

5. Stepchildren and Union Dissolution

One of the most widely observed characteristics of second partnerships is that they are less stable than first partnerships. Although many have speculated that this pattern was the result of one or more elements reducing commitment (from religious prohibitions to personal orientations toward divorce), a landmark study found that most of the difference in marital satisfaction and in stability could be attributed to the presence of stepchildren. Previously childless persons who repartnered had very little greater risk of union dissolution than those in first partnerships (White and Booth 1985). Studies of union dissolution since then have generally included indicators of the presence of non-joint children as well as joint children, and found that whatever protective effect joint children might have does not characterize stepchildren, who increase the risk of separation. Little research has been done, however, on what types of stepchildren contribute more to this effect. Are the ages of the children a factor? Is their sex? Does the involvement of the step-

parent and/or of the non-coresident parent matter? An even more important element might be the parenting practices of the new partners. Two studies of paternal involvement that was restricted to first partnerships showed that greater involvement of a father with his children increased the wife's marital satisfaction (Harris and Morgan 1991) and their children's well-being (Amato 1994); stepfathers are normally less involved with their stepchildren than are biological fathers (Marsiglio 1991).

Related to this issue is a startling finding—that couples with daughters are more likely to divorce than couples with sons (Morgan et al. 1988). The key to this result appears to be paternal involvement with children. Men tend to spend more time with their sons than their daughters, both in one-on-one conversation and in terms of participation in *family* activities, sharing more meals, leisure time, and home activities with them (Cooksey and Fondell 1996). In a sense, stepchildren may be like daughters for many men, children who they wanted less, who they do not relate to as closely, and who they can leave more easily. This may also account for the commonly observed phenomenon that sons often benefit more by their mother's remarriage than daughters (Brooks-Gunn 1994). Given the increased importance of repartnering in the lives of children and their parents in western, industrialized societies, understanding this phenomenon needs to be a high priority.

See also: Child Abuse; Children and the Law; Divorce and Children's Social Development; Divorce and Gender; Divorce, Sociology of; Family Business; Family Theory: Economics of Marriage and Divorce; Fatherhood; Marriage; Motherhood: Economic Aspects; Motherhood: Social and Cultural Aspects; Parenting: Attitudes and Beliefs

Bibliography

- Amato P R 1994 Father-child relations, mother-child relations, and offspring psychological well-being in early adulthood. *Journal Marriage and the Family* **56**: 1031–42
- Andersson G 1997 The impact of children on divorce risks of Swedish women. *European Journal of Population* **13**: 109–45
- Bennett N G, Bloom D E, Miller C K 1995 The influence of nonmarital childbearing on the formation of first marriages. *Demography* **32**: 47–62
- Bernhardt E, Goldscheider F forthcoming Men, resources and family living: The determinants of union and parental status in the US and Sweden. *Journal of Marriage and the Family*
- Bernhardt E, Goldscheider F 1999 Children and Union Formation in Sweden. Paper presented at the European Population Conference, August, The Hague, The Netherlands
- Brooks-Gunn J 1994 Research and policy issues in the study of stepparenting. In: Booth A, Dunn J (eds.) *Stepfamilies: Who benefits? Who does not?* L. Erlbaum Associates, Hillsdale, NJ, pp. 167–89

- Bumpass L L, Raley R K 1995 Redefining single-parent families: cohabitation and changing family reality. *Demography* **32**: 97–109
- Cooksey E C, Fondell M M 1996 Spending time with his kids: Effects of family structure on fathers', and children's lives. *Journal of Marriage and the Family* **58**: 693–707
- Goldscheider F, Sassler S 2000 Men, children, and union formation. Unpublished paper.
- Harris K M, Morgan S P 1991 Fathers, sons, and daughters: Differential paternal involvement in parenting. *Journal of Marriage and the Family* **53**: 531–44
- Marsiglio W 1991 Paternal engagement activities with minor children. *Journal of Marriage and the Family* **53**: 973–86
- Morgan S P, Lye D, Condran G 1988 Sons, daughters and the risk of marital disruption. *American Journal of Sociology* **94**: 110–29
- Smith K, Zick C D, Duncan G J 1991 Remarriage patterns among recent widows and widowers. *Demography* **28**: 361–74
- Vikat A, Thomson E, Hoem J 1999 Stepfamily fertility in contemporary Sweden: The impact of childbearing before the current union. *Population Studies* **53**: 211–25
- White L K, Riedmann A 1992 When the Brady Bunch grows up: Step/half sibling and full sibling relationships in adulthood. *Journal of Marriage and the Family* **54**: 197–208
- White L K, Booth A 1985 The quality and stability of remarriages: The role of stepchildren. *American Sociological Review* **50**: 689–98

F. K. Goldscheider

Replication

'Replication' refers to the repetition of the same or very similar research study. The term 'replication study' refers to a study conducted specifically to verify the results of an earlier study. The concept of replication of research studies is central to the scientific study of psychology and the social and behavioral sciences more generally, as many authors have pointed out in Neuliep (1991).

1. The Importance of Replication

Scientists of all disciplines have long been aware of the importance of replication to their enterprise. Now, thanks to the efforts of science writers for daily newspapers and reporters for network news, even the general public has become aware of the importance of replication in all the sciences.

The undetected equipment failure, the rare and possibly random human errors of procedure, observation, recording, computation, or report, are known well enough to make scientists wary of the unreplicated experiment. When we add to the possibility of the random 'fluke' common to all sciences, the fact of individual organismic differences and the possibility of systematic experimenter effects, the

importance of replication looms larger still to the behavioral scientist.

What shall we mean by 'replication'? Clearly, the same experiment can never be repeated by a different worker. Indeed, the same experiment can never be repeated even by the same experimenter. At the very least, the participants and the experimenters themselves are different over a series of replications. The participants are usually different individuals, and the experimenter changes over time, if not necessarily dramatically. But to avoid the conclusion that there can be no replication in the social and behavioral sciences, we can speak of relative replications. We can rank order experiments on how close they are to each other in terms of participants, experimenters, tasks, and situations. We can usually agree that this experiment, more than that experiment, is like a given paradigm experiment.

Replications may be crucial, but some replications are more crucial than others. Three of the variables affecting the value, or utility, of any particular replication are:

- (a) when the replication is conducted
- (b) how the replication is conducted
- (c) by whom the replication is conducted.

1.1 When the Replication is Conducted

Replications conducted early in the history of a particular question are usually more useful than replications conducted later in the history of a particular research question. Weighting all replications equally, the first replication doubles our information about the research issue, the fifth replication adds 20 percent to our information level, and the fiftieth replication adds only 2 percent to our information level. Once the number of replications grows to be substantial, subsequent investigators' felt need for further replication is likely to be due not to a real need for further replication but for a real need for the more adequate evaluation and summary of the replications already available.

1.2 How the Replication is Conducted

It has already been noted that replications are possible only in a relative sense. However, there is a distribution of possible replications in which the variance is generated by the degree of similarity to the original study that characterizes each possible replication. If we choose our replications to be as similar as possible to the study being replicated, we may be more true to the original idea of replication but we also pay a price; that price is external validity (see *External Validity*).

If we conduct a series of replications as exactly like

		Result of replication	
		Successful	Unsuccessful
Precision of replication	Fairly precise	Supports the theory	Damages the theory
	Fairly imprecise	Extends the theory	Limits the theory

Figure 1

Effects on theory of the success of replication and the precision of replication

		Result of replication	
		Successful	Unsuccessful
Precision of replication	Fairly precise	Supports the investigator	Impugns the investigator
	Fairly imprecise	Supports the investigator	Impugns the investigator very little

Figure 2

Effects on original investigator of the success of replication and the precision of replication

the original as we can, we have succeeded in ‘replicating’ but not in extending the generality of the underlying relationship investigated in the original study. The more imprecise the replications, the greater the benefit to the external validity of the tested relationship if the results support the relationship. If the results do not support the original finding, however, we cannot tell whether that lack of support stems from the instability of the original result or from the imprecision of the replications. Figure 1 summarizes the consequences for the theory that is derived from the initial study of (a) successful vs. unsuccessful replications, and (b) the precise vs. the imprecise nature of the replications.

Figure 2 summarizes the consequences for the original investigator of (a) successful vs. unsuccessful replications, and (b) the precise vs. imprecise nature of the replications. Compared to the theory tested by the replication, the investigator has much to lose if a fairly precise replication is unsuccessful since such failure is often associated with ascriptions to the original investigator of having been careless, incompetent, and in some cases, even dishonest.

1.3 By Whom the Replication is Conducted

So far in our discussion of replications we have assumed that the replications are independent of one another. But what does independence mean? The usual minimum requirement for independence is that the participants in the replication studies be different persons. But what about the independence of the replicators? Are 10 replications conducted by a single investigator as independent of one another as 10 replications each of which is conducted by a different investigator? This issue of potentially correlated replicators bears additional comment.

1.4 The Problem of Correlated Replicators

To begin with, an investigator who has devoted her life to the study of vision, or of psychological factors in somatic disorders, is less likely to carry out a study of verbal learning than the investigator whose interests have always been in the area of verbal learning. To the extent that (a) experimenters with different research interests are different kinds of people, and to the extent

		<i>First study</i>	
		$p > 0.05^*$	$p < 0.05$
<i>Second study</i>	$p < 0.05^a$	A Failure to replicate	B Successful replication
	$p > 0.05$	C Failure to establish effect	D Failure to replicate

Figure 3

Common model of successful replication: judgment is dichotomous and based on significance testing (*by convention 0.05 but could be any other given level, e.g., 0.01; ^a in the same tail as the results of the first study)

that (b) it has been shown that different kinds of people, experimenters, are likely to obtain different data from their subjects, we are forced to the conclusion that within any area of behavioral research, the experimenters come pre-correlated by virtue of their common interests and any associated characteristics. Immediately, then, there is a limit placed on the degree of independence we may expect from workers or replications in a common vineyard. But for different areas of research interest the degree of correlation or similarity among its workers may be quite different. Certainly we all know of workers in a common area who obtain data quite opposite from that obtained by colleagues. The actual degree of correlation, then, may not be very high. It may, in fact, even be negative, as with investigators holding an area of interest in common, but holding opposite expectancies about the results of any given experiment.

A common situation in which research is conducted nowadays is within the context of a team of researchers. Sometimes these teams consist entirely of colleagues; often they are composed of one or more faculty members and postdoctoral students, and one or more predoctoral students at various stages of progress toward the Ph.D. Experimenters within a single research group may reasonably be assumed to be even more highly intercorrelated than any group of workers in the same area of interest who are not within the same research group. Perhaps students in a research group are more likely than a faculty member in the research group to be more correlated with their major professor. There are two reasons for this likelihood. The first is a selection factor. Students may elect to work in a given area with a given investigator because of their perceived and/or actual similarity of interest and associated characteristics. Colleagues are less likely to select a university, area of interest, and

specific project because of a faculty member at that university. The second reason why students may be more correlated with their professor than another professor is a training factor. Students may have had a large proportion of their research experience under the direction of a single professor. Another professor, though collaborating with colleagues, has most often been trained in research elsewhere by another person. Although there may be exceptions, even frequent ones, it seems reasonable, on the whole, to assume that student researchers are more correlated with their advisor than another advisor might be.

The correlation of replicators that we have been discussing refers directly to a correlation of *attributes* and indirectly to a correlation of *data* these investigators will obtain from their subjects. The issue of correlated experimenters or observers is by no means a new one. Nearly 100 years ago Karl Pearson spoke of 'the high correlation of judgments...[suggesting] an influence of the immediate atmosphere, which may work upon two observers for a time in the same manner' in his 'On the mathematical theory of errors of judgment with special reference to the personal equation' (Pearson 1902, p. 261).

2. The Meaning of Successful Replication

There is a long tradition in psychology of our urging one another to replicate each other's research. But, although we have been very good at calling for replications, we have not been too good at deciding when a replication has been successful. The issue we now address is: when shall a study be deemed successfully replicated?

Successful replication is ordinarily taken to mean that a null hypothesis that has been rejected at time 1

is rejected again, and with the same direction of outcome, on the basis of a new study at time 2. The basic model of this usage can be seen in Fig. 3. The results of the first study are described dichotomously as $p < 0.05$ or $p > 0.05$ (or some other critical level, e.g., 0.01). Each of these two possible outcomes is further dichotomized as to the results of the second study as $p < 0.05$ or $p > 0.05$. Thus cells A and D of Fig. 3 are examples of failure to replicate because one study was significant and the other was not. The problem with this older view of replication is that two studies could show identical magnitudes of the effect investigated yet differ in significance level because of differences in sample sizes.

2.1 Contrasting Views of Replication Success

The traditional, not very useful view of replication modeled in Fig. 3 has two primary characteristics:

(a) It focuses on significance level as the relevant summary statistic of a study.

(b) It makes its evaluation of whether replication has been successful in dichotomous fashion. For example, replications are successful if both or neither $p < 0.05$ (or 0.01, etc.), and they are unsuccessful if one $p < 0.05$ (or 0.01, etc.) and the other $p > 0.05$ (or 0.01, etc.). Psychologists' reliance on a dichotomous decision procedure accompanied by an untenable discontinuity of credibility in results varying in p levels has been well documented, for example, by Nan Nelson, Robert Rosenthal and Ralph L. Rosnow (1986).

The newer, more useful views of replication success have two primary characteristics:

(a) A focus on effect size as the more important summary statistic of a study with only a relatively minor interest in the the statistical significance level.

(b) An evaluation of whether replication has been successful made in a continuous fashion. For example, two studies are not said to be successful or unsuccessful replicates of each other, but rather the degree of failure to replicate is specified.

See also: Experimental Design: Overview; External Validity; Generalization: Conceptions in the Social Sciences; Hypothesis Testing: Methodology and Limitations; Internal Validity

Bibliography

- Campbell D T 1988 Overman E S (ed.) *Methodology and Epistemology for Social Science: Selected Papers*. University of Chicago Press, Chicago
- Cooper H M 1989 *Integrating Research: A Guide For Literature Reviews*, 2nd edn. Sage, Newbury Park, CA
- Hunt M 1997 *How Science Takes Stock*. Russell Sage Foundation, New York

Light R J, Pillemer D B 1984 *Summing Up: The Science of Reviewing Research*. Harvard University Press, Cambridge, MA

Nelson N, Rosenthal R, Rosnow R L 1986 Interpretation of significance levels and effect sizes by psychological researchers. *American Psychologist* **41**: 1299–301

Neuliep J W (ed.) 1991 *Replication Research in the Social Sciences*. Sage, Newbury Park, CA

Pearson K 1902 On the mathematical theory of errors of judgement with special reference to the personal equation. *Philosophical Transactions of the Royal Society of London* **198**: 235–99

Rosenthal R 1976 *Experimenter Effects in Behavioral Research*, enlarged edn. Irvington Publishers, Halsted Press Division of Wiley, New York

Rosenthal R 1991 *Meta-Analytic Procedures For Social Research*, rev. edn. Sage, Newbury Park, CA

Rosenthal R, Rosnow R L 1969 *Artifact in Behavioral Research*. Academic Press, New York

Rosnow R L, Rosenthal R 1997 *People Studying People: Artifacts and Ethics in Behavioral Research*. Freeman, New York

R. Rosenthal

Representation: History of the Problem

Philosophers and social scientists have always been concerned about the relation of their knowledge about the social world to (the rest of) this world. This concern has sometimes been phrased as the question of the representation of the social world in and through the writing of scholars, or as the *adaequatio intellectus et rei*, being the objective of scientific inquiry. From Plato and Aristotle, and then Thomas of Aquinas, this *problématique* stretches into the positivism of the social sciences of the nineteenth century as well as its critical alternative, scientific Marxism. With important qualifications, it was revived in the neopositivism of the twentieth century and the unified science program that tried to (re-) integrate the social sciences with the natural sciences.

Despite this continuity of concern, the very concept of representation of the social world has also been heavily criticized. At several points in intellectual history it appeared as if the whole idea needed to be abandoned, because there is no way to verify the relation between the world and thoughts about the world other than again by thought. Thus, knowledge of *adaequatio* is in principle unattainable. In most recent debate, during the closing decades of the twentieth century, the accumulation of such doubt has often been labelled 'crisis of representation,' implicitly and misleadingly pretending novelty of these doubts. Through all of earlier such crises, however, the *problématique* keeps reviving, often with new qualifications, and those who revive it often argue that it cannot be entirely abandoned either, since otherwise

the status of statements about the (social) world would become entirely unclear.

1. Representation in Philosophical Modernity

Even though this debate has been going on virtually through all of the history of philosophy, the break that came with the radical skepticism characteristic of what can be called the philosophy of modernity is often seen as shifting the terrain. Between the middle of the seventeenth century, Descartes's time of writing, and the end of the eighteenth century, Kant's time, the ground on which representation could be possible was systematically explored. Descartes's *cogito ergo sum* radically subjectivized knowledge claims. Nevertheless, he was able to develop on that basis a rationalism according to which mathematical representation was adequate to the world. Kant's distinction between *noumena* and phenomena has sometimes been interpreted in a similar way, that is, by considering *noumena*, although they cannot be experienced, as theoretical entities the relations between which are governed by laws. A science developing such laws would then be truthfully representing the world. Arguably, however, the more significant impulse from these one and a half centuries of debate was a deep crisis of representation, at least within the philosophical mode of reasoning. This crisis of representation stemmed from the increasing gap between the perceiving and knowing subject and the world.

The combined outcome of the Cartesian and the Kantian moves, namely, was that, on the one hand, philosophical reasoning appeared to have exhausted its possibilities, at least for the time being. At the end-point of this reasoning, on the other hand, a space had been opened for a different form of inquiry. The gap between the subject and the world, or between the *noumena* and the phenomena, permitted, maybe even invited, attempts to fill it by positive, empirical knowledge, elaborated in systematic, conceptually guided ways (see *Positivism, History of*). It allowed the subjectivation and historicization of knowledge claims, which in turn made the social sciences possible, as psychological inquiry into human perception and experience of the world and as sociohistorical collection and systematization of the 'facts' of human social life. Much of the social sciences of the nineteenth century, often devoid of any explicated philosophy of the social sciences, indeed worked on such foundations.

2. Representation in Political Modernity

While philosophical debate opened the possibility for developing a conception of representation of the social world through social science inquiry, roughly parallel political transformations made such a representation appear as an increasingly urgent necessity. These transformations made the concept of representation

increasingly problematic in political thought and practice, a culmination of this process and a turning-point being reached with the American and French revolutions.

A concept of representation had been almost completely absent from political thinking in Europe and the Mediterranean region during antiquity and the Middle Ages. First elements of such a *problématique* gradually emerged with the rivalry between various sources of religious power during the era of the schism, and between religious and political power with the contest between the pope and the emperor. Representation here, most broadly, refers to the relation between a holder of authority to that over which this authority is exercised. Employing the concept of representation for this question presupposes that authority is not held in its own right, but for something or somebody else. The disputes during the late Middle Ages and early modern times were largely cast in terms of the legitimate representation of God, as the uncontested supreme authority, on earth. William of Ockham may have been the first to conclude on the impossibility of deciding in the face of competing claims for representation and, thus, to have opened up a political version of the crisis of representation.

As in the philosophical debate about the knowledge of the world, the period between the middle of the seventeenth and the end of the eighteenth century is of crucial importance for a recasting of the issue of political representation. The first systematic attempt to solve the issue in modern form, that is, by fully acknowledging both the exigency of representation and the undecidability of the question of the sources of authority, was made by Thomas Hobbes. Faced with the religious wars and thus a violent contest over the rightful representation of God, he ruled out any religious source, but insisted on the need for unity in the polity to avoid persistent warfare. The idea of state sovereignty, prepared by Bodin, and of the unitary state represented by the monarch were the key ingredients of his solution to the problem. By making the constitution of the state itself, however, dependent on a contract between individuals, he had opened the breach for individualism, as Carl Schmitt observed. The idea of popular sovereignty, as developed by Locke and Rousseau, should indeed turn Hobbes's reasoning around and prepare the intellectual ground for the political transformations of the late eighteenth century.

Quite analogous to the situation in general philosophy, the development of political philosophy was, on the one hand, stalled with the full explication of an individualist liberal political theory, that is a political theory of modernity (see *Liberalism: Impact on Social Science*). On the other hand, however, insufficiencies of the liberal assumptions, particularly visible in the post-revolutionary situation, lent themselves to the development of the social sciences as means to deal with those insufficiencies by other means, through a

kind of 'empirical political philosophy' namely (Wagner 1998).

Liberal political philosophy was insufficient in the sense that the set of rules on which viable political institutions could be built, starting out from individuals and their potential diversity, needed propositions beyond those that could be derived from the basic assumptions of individualist liberalism. The concept of self-determination needed, for instance, procedural rules of participation in communication and in deliberation. Since such rules could not unequivocally be derived from the principles themselves, the issue of representation emerged much more forcefully than ever before. In response to the French Revolution, Burke (1993) still argued that, while there was a practical issue of representation, best handled in line with grown traditions, there was no specific legitimacy attached to representation. Paine (1993), in contrast, linked most forcefully the idea of the sovereignty of the people to the representation of the people in the governing bodies. The controversy between Burke and Paine 'marks the turning point of two ages of representation' (Sternberger 1971, p. 594). As soon as any position closer to Paine than to Burke was accepted, debates on modes of representation, such as those that had already occurred in both the American and the French constituent assemblies, became inevitable.

And it was in conjunction with this inevitability of disputes over representation that the creation of social effects of the liberal rules moved to the center of attention. The identification and the analysis of these effects called for social science inquiry and conceptualization. The observation of actual, empirical structures of representation also served the interest of enhancing some stability and certainty in political procedures that could appear to be opened to all contingencies by the abolition of any legitimacy of preordained orders (see *Freedom/Liberty: Impact on the Social Sciences*). Two main strategies for rediscovering certainties can be identified: systematic observation and reflective conceptualization. The one was more typical for American thinking, where the structure of the vote was of key interest, whereas European social thinkers focused on what came to be called social structures behind the vote. These two intellectual responses to the political *problématique* of representation inaugurate two modes of social theorizing, the behaviorist and the structural-functional one.

In sum, thus, during the nineteenth century an increasing overlap of concerns between cognitive representation, that is, the truthful representation of social reality, and political representation, that is, the need for a social map to draw up rules for electoral representation can be observed. One may speak here of the gradual transformation of a philosophical conception of representation, based on the political philosophy of the Enlightenment, into a sociological one, based on the sociology of industrial and mass

society emerging at the end of the nineteenth century (d'Arcy and Saez 1985). This rephrasing of the problem of representation was brought again into crisis, however, by the turn of the nineteenth century and, yet again, in the late twentieth century. These crises, although they were also crises of philosophical representation, showed specific features of a crisis of the representation of the social world—and this was again related to political issues of representation.

3. *A Crisis in Modern Representation and its Temporary Solution*

From the late nineteenth century on, a shift in political orientations, which entailed a declining faith in the viability of liberalism, was paralleled in epistemology by a renewed skepticism about the other central tenet of the Enlightenment tradition, the intelligibility of human action and the social world. The period around the turn of the century is now known as an intellectually extremely fruitful, even a classical era in many fields of social science, most notably sociology, psychology and economics. At that time, however, much of the work was driven by a sense of crisis, a feeling that many of the assumptions of earlier social science were inadequate.

In terms of epistemology, social science saw itself forced largely to abandon the idea of representing social reality and accepted the view that conceptual constructions were dependent both on the means and forms of observation and perception and on the interest of the observer in the social world. American pragmatism is possibly the most explicit case of such a reorientation, but similar, often much more tension-ridden, discussions were led in European debates, a prominent example being Max Weber's methodological writings. As a consequence, the turn-of-the-century approaches were less convinced about a determinist course of human history than earlier social science, and also less persuaded of the direct insight into any laws of the social world from empirical observation. Such uncertainty also entailed a crisis in the usability of social knowledge for purposes of action. The earlier call had been for better, 'positive' knowledge that would lead to better action; it relied on the conception of an objective social world outside the mind of the observers, but adequately represented by them in social theory and research. According to such a conception, action based on uncertain knowledge should entail uncertain outcomes.

Through the early decades of the twentieth century, however, novel conceptions of the relation between knowledge and action were proposed that were based not least on a critique of what Dewey (1984, p. 195) called the 'spectator theory of knowing.' With this term, he summarized all views, for him characteristic of most philosophy and science, that saw the world as exposed to the human gaze and assumed that it is

exactly this distance between the knowing subject and the object to be known that allows for certain knowledge. The distancing from the sensations of the world had been considered a precondition for the identification of those assumptions on which claims for valid knowledge could be erected. By the same move, the question of epistemology itself had emerged as the problem of the relation between the 'reality' of the world and the 'representations' that human beings provide of the world in their philosophy and science. The preference in the philosophy of science was for a long time, as Dewey points out, for an epistemological formalism that presupposes this distance between the knower and that which is to be known. Pragmatism precisely rejects this assumption of distance.

Similar turns were made in European philosophy during the same period. Ludwig Wittgenstein had hoped to clarify the problem of representation in the *Tractatus logicus-philosophicus*, in which he employs a picture theory of knowing, whereas in the *Philosophical Investigations* he sees our languages as being part of our forms of life, and thus not looking at life from a distance. In *Holzwege*, Martin Heidegger analyzed the modern theory of knowledge as representation comprehensively as the attempt 'to provide oneself a picture' of the world by making a world that is placed in front of oneself. Rather than making 'a picture of the world,' this move 'grasps the world as picture' (see *Imagination: History of the Concept*). In two intertwined processes, the world becomes a picture and the human being a subject. Rorty tried to summarize this reinterpretation of the history of philosophy and the sciences provided by Dewey, Heidegger, and Wittgenstein as the abdication of the attempt of seeing knowledge as 'the mirror of nature' (Rorty 1979, Friese and Wagner 1997, pp. 208–11).

The critique of representation, in this sense, was linked to the fundamental and critical epistemological presupposition that the social world is in important respects not found and discovered but made and invented. This novel combination, however, provided merely for one strand of debate in the social sciences of the first half of the twentieth century, and it did not become dominant. Partly disregarding epistemological issues, partly referring to the 'neopositivism' of the Vienna Circle, the other major strand tried to deal with the problem of representation by methodological means. At this point, the history of the problem of representation links up closely with the history of statistics and of empirical social research.

In statistics, the period between the end of the nineteenth century and the 1930s witnessed intense debates about a seemingly very specific methodological question. The problem was how to generalize observations collected from the study of a part to statements about the whole. Its solution was a number of statistical techniques summarized under the label of representativity. These techniques were developed simultaneously in debates within the statistical pro-

fession and about the design of social insurance arrangements. An institutional, a cognitive, and a political transformation were closely interlinked, 'the nationalization of social statistics, ... the diffusion of the 'representative method' and of random selection, and ... the beginnings of the development of the Welfare State' (Desrosières 1991, p. 228) (see *Statistics, History of*).

During the interwar period, mathematical techniques spread to other areas of empirical research, not least to the emergent fields of opinion research, of business cycle research and of research into the social structure of society. Special research institutes were created, during those years and particularly after the Second World War, that were to provide the data to which those techniques were to be applied. From the late 1920s onwards, thus, the contours of a new social science, which then should dominate the image of those sciences in the second post war era, are recognizable. This social science liberated itself from the doubts of the 'classical period' by methodologically circumventing the problem of representing society. The doubts about epistemological and conceptual issues could not be entirely removed, but they could be dealt with, it was assumed, by starting from the most secure elements one could find, that is the empirical observation and collection of data on the preferences and behaviors of individual human beings. Conclusions referring to the larger scale of society and politics were arrived at by aggregation of those data; and the questions towards the answering of which data were generated or analyzed were derived from policy needs for 'social control.' Thus, a new connection between cognitive and political representation was created. Or, in other words, a 'soft' behaviorism aligned with a similarly 'soft' pragmatism.

4. The Re-emergence of a Crisis of Representation

However, such solution to the problem of representation remained temporary. During the closing decades of the twentieth century, the 'interpretative turn' or, more broadly, the 'linguistic turn' in the human sciences has had strong, though uneven, effects in the sciences devoted to the study of contemporary society. Emphasis on the linguistic constitution of the social world and on the interpretative openness of social representations brings the social sciences back to a period of epistemological, ontological, and methodological reconsideration that shows many parallels to the 'classical era' at the beginning of the twentieth century.

It has been argued that this recent period of intellectual transformation occurs in parallel to an overall process of sociopolitical transformation. Crises of cognitive and of political representation then again occur in tandem (Wagner 1994). During such periods,

it becomes evident that a social science that tightly couples the concern with social order to the solidity of its own representations of the social world applies at best in periods of relative social stability, that is, only contingently. Conceptually, however, the ideas of unanimity or similarity of behavior being established by common material constraints or well mastered by shared ideas (beliefs, representations) or a diversity of actions being integrated through systematic articulation (division of functions, of roles) works with too restrictive definitions of social life (Thévenot 1993 p. 276). Attention needs instead to be directed towards 'the immense historical labor that is necessary to unite disparate beings around the same system of representation, to constitute the reality of such a heterogeneous ensemble, to inscribe it into devices (*dispositifs*) through an intensive effort of objectification and to endow it with a common interest' (Boltanski 1990 p. 70).

It is important to note that many of the authors who recently have renewed the interest in the *problématique* of representation, such as Michel Foucault and Reinhart Koselleck, have done so by emphasising the linkage between cognitive and political representation. In their own research, they have also underlined that the critical relation between these two forms of representation becomes accentuated during the political and intellectual transformations of the late eighteenth and early nineteenth century. This is at the same time the period of the emergence of the social sciences. They were created to deal with this *problématique*, but they remain nevertheless, or for that very reason, marked by it.

See also: Knowledge Representation; Marxism/Leninism; Modernity; Political Representation; Positivism, History of; Propositional Representations in Psychology; Representations and Rituals of the State; Wittgenstein, Ludwig (1889–1951)

Bibliography

- d'Arcy F, Saez G 1985 De la représentation. In: d'Arcy F (ed.) *La représentation*. Economica, Paris
- Boltanski L 1990 *L'amour et la justice comme compétences. Trois essais de sociologie de l'action*. Métailié, Paris
- Burke E 1993/1790 *Reflections on the Revolution in France*. Oxford University Press, Oxford, UK
- Desrosières A 1991 The part in relation to the whole? How to generalise. A prehistory of representative sampling. In: Bulmer M, Bales K, Sklar K K (eds.) *The Social Survey in Historical Perspective 1880–1940*. Cambridge University Press, Cambridge, UK, pp. 217–44
- Dewey J 1984/1929 *The Quest for Certainty: The Later Works 1925–1953*, Vol. 4. Southern Illinois University Press, Carbondale, IL
- Friese H, Wagner P 1997 Bild und Wissen. Reflexionen zu Erkenntnis und Darstellung in den Sozialwissenschaften. In: Hoffmann U, Joerges B, Severin I (eds.) *LogIcons. Bilder zwischen Theorie und Anschauung*. Edition sigma, Berlin, pp. 207–22
- Hacking I 1983 *Representing and Intervening: Introductory Topics in the Philosophy of Natural Science*. Cambridge University Press, New York
- Lynch M, Woolgar S (eds.) 1990 *Representation in Scientific Practice*. MIT Press, Cambridge, MA
- Paine T 1993/1791–2 *The Rights of Man*. Dent, London
- Rorty R 1979 *Philosophy and the Mirror of Nature*. Princeton University Press, Princeton, NJ
- Sternberger D 1971 A controversy of the late eighteenth century concerning representation. *Social Research* 38
- Thévenot L 1993 Agir avec d'autres. Conventions et objets dans l'action coordonnée. In: Ladrière P, Pharo P, Quéré L (eds.) *La théorie de l'action. Le sujet pratique en débat*. CNRS Editions, Paris, pp. 275–89
- Wagner P 1994 *A Sociology of Modernity. Liberty and Discipline*. Routledge, London
- Wagner P 1998 Certainty and order, liberty and contingency; the birth of social science as empirical political philosophy. In: Heilbron J, Magnusson L, Wittrock B (eds.) *The Rise of the Social Sciences and the Formation of Modernity: Conceptual Change in Context, 1750–1850*. Kluwer, Dordrecht, Germany, pp. 241–63

P. Wagner

Representation of Serial Order, Cognitive Psychology of

Stimuli come and go. The serial order of their appearance is determined by rules in the external world as well as by the actions of the organism experiencing the stimuli. Consequently, successions of stimuli are almost never random but rather comprise some form of structure. Structure allows prediction, and predicting forthcoming events is advantageous because organisms can prepare for what will happen. Thus, sensitivity to serial order can be considered as being one of the basic mechanisms by which humans (and organisms in general) adapt behavior to constraints in the succession of events. That is why the examination of serial learning has a prolonged research history (e.g., Ebbinghaus 1902, Lashley 1951, Miller 1956).

Serial order refers to deviations from randomness in the succession of events. Consider, for instance, the three colored lights green (g), red (r), and yellow (y) as distinct events. The series g–g–r–g–y–g–r–y–y–g–r–... is random, as each of the lights was selected with equal probability in each position. In contrast, the series g–y–r–y–g–y–r–y–g–... presents the highly regular sequence we experience in watching ordinary traffic lights. Knowing the serial order of traffic lights facilitates responding: if we see that the light switches from green to yellow we can expect that it will next switch to red so that we can prepare to stop before we are really forced to do so. Almost all event sequences entail some regularity: clouds are frequently followed

by rain, words follow each other according to grammatical rules, tone sequences constitute melodies, and baking a cake requires an ordered sequence of actions, etc. In this article, first two types of serial order are distinguished, then several effects of serial order on human cognition are sketched, and finally, the main theoretical approaches to account for serial order effects are mentioned.

1. Statistical and Relational Constraints in Serial Order

Statistical constraints refer to the conditional probabilities of events in dependence on preceding ones. Traffic lights are a good example: after 'red' and after 'green,' 'yellow' is certain to follow, whereas after 'yellow' either 'red' or 'green' follows. If one takes two preceding lights into account, however, the light following 'green-yellow' and 'red-yellow' is also safely predictable. An example of more complex statistical constraints is provided by sentences: the next word in a sentence can mostly be predicted with a probability higher than chance but almost never with certainty, regardless of how many preceding words are taken into account. *Relational constraints* refer to regularities in the relations between successive events instead of conditional probabilities. Consider, for instance, an arithmetic series such as 3-8-13-18-23-28. The regularity of the series is based on that between every pair of adjacent numbers the same relation holds. An example of a somewhat less regular relational structure is a digit sequence such as 23461938, which contains only two equally related pairs of numbers, namely 23-46 and 19-38.

2. Serial Order Effects in Human Cognition

2.1 Serial Order Effects in Prediction

In order to examine the adaptation of predictions to serial order, experiments have been performed in which items are successively presented and participants are asked to predict each forthcoming item. Typically, prediction accuracy improves corresponding to the statistical and/or relational constraints in the succession of items. Moreover, the data indicate that constraints in immediate successions are effective before constraints in the succession of items further apart from each other are taken into account. Consider, for instance, a well ordered sequence such as 1-2-1-2-2-3-2-3-6-5-6-5-5-4-5-4. Experimental data show that participants first learn to predict the respective second items of those pairs whose elements are likewise related, (12), (23), (65), and (54). They continue to learn the repetitions of pairs (1212), (2323),

(6565), and (5454). Finally, participants manage to predict the order of the quadruples (Restle and Brown 1970). In other words, serial learning starts by connecting adjacent items into subunits or chunks corresponding to the given relational (as in the present case) or statistical constraints. Following this, constraints in the succession of these elementary chunks results in the formation of higher order chunks, i.e., connections between elementary chunks, which again are subject to chunking, etc. Thus, consecutive chunking may lead to a representation of the whole sequence through a hierarchy of chunks.

2.2 Serial Order Effects in Remembering

Ordered sequences of items are generally remembered better than unordered sequences (see *Memory: Organization and Recall*). For instance, sequences of words are remembered better if they follow the statistical constraints of word order in sentences (Miller and Selfridge 1950), a list of words is remembered better if words belonging to the same category (e.g., names of fishes, flowers, birds, or trees) are presented adjacently (Underwood et al. 1974), and a series of letters will be remembered better if it consists of familiar abbreviations as in FBICIAPHDVIP. In all these cases subsequences of the items to be remembered correspond to familiar knowledge units like word phrases, categories, or abbreviations which are used to reduce memory load. In the latter case, for instance, participants have to keep in mind the four abbreviations FBI, CIA, PHD, and VIP instead of 12 letters. Moreover, the order of the storage units or chunks may provide a useful retrieval plan which additionally helps remembering. In the categorically ordered word list mentioned above, for instance, memory will be further improved if also the categories are presented in a systematic order such as the animals (fishes and birds) first and then the plants (flowers and trees).

However, new and unfamiliar serial constraints also improve memory. Consider, for instance a so-called 'finite state grammar,' which restricts the way in which the letters T, P, V, X, and S are allowed to follow each other in letter strings. For instance, all grammatical strings may start either with the letter T or P, they may finish either with V or S, only the letters X and V may be allowed to be repeated, after letter T only the letters S, X, and V may be allowed to follow, etc. In this way the 'grammar' constitutes a statistical structure which all grammatical strings such as TXS, TSXXTVV, PTVPS, or PVPXVPS have in common. If such letter strings are to be memorized, grammatical strings are remembered increasingly better than random strings and after some training participants are even able to distinguish between grammatical and agrammatical strings fairly well (Reber 1967). The findings clearly indicate that the completely new and arbitrary stat-

istical constraints have been learned and are used to put order into the letter strings to be remembered.

2.3 Serial Order Effects in Motor Performance

In motor performance, similar effects as in predicting or remembering stimulus sequences have been reported: well ordered sequences of actions are executed faster and more correctly than unordered sequences and, even more informative, the timing of sequence execution mirrors its serial order (Rosenbaum et al. 1983). Consider, for instance, the times for pressing a series of eight distinct keys with the index fingers (I, i) and the middle fingers (M, m) of the right and left hand, respectively. The order of the keys to be pressed shall be such that a first pair of keystrokes (e.g., I–m) is to be repeated (I–m–I–m) and then a second pair is to be repeated (e.g., M–i–M–i). In performing such a regular keystroke sequence as fast as possible, execution times are highest at the beginning of the quadruples *ImIm* and *MiMi*, they are somewhat reduced at the beginning of each repeated pair *ImIm* and *MiMi*, and they are lowest for the respective second element in each pair *ImIm* and *MiMi*. The correspondence of this time pattern with the serial order of the keystrokes suggests that the time to execute a motor act in a sequence of acts is determined by hierarchically organized chunks rather in the same way as the prediction of items is: at the beginning of each new chunk execution time increases as prediction accuracy decreases and for items within chunks execution is fastest and prediction is best (see *Motor Skills, Psychology of*).

2.4 Serial Order Effects in Serial Reaction Time (SRT) Tasks

In SRT tasks, participants are required to respond to successively presented stimuli by distinct responses whereby each response triggers the presentation of the next stimulus, which in turn requires the next response, etc. Numerous studies have shown that nearly any kind of serial order in the succession of stimulus–response pairs results in a decrease in response times (Hoffmann and Koch 1998). Serial-order effects in SRT tasks deserve special attention because besides constraints in the succession of stimuli and of responses, also constraints in the succession of the responses and the following stimuli, i.e., in response–effect relations, may contribute to serial learning. Although not much research has been done yet in this respect, the available evidence clearly indicates that consistent response–effect relations do indeed contribute to the acquisition of serial order (Ziessler 1998). This is most noteworthy because consistent response–effect relations are an essential part of any ordinary behavioral sequence.

3. Intentional and Incidental Acquisition of Serial Order

Some of the mentioned cognitive effects surely result from the explicit detection and use of serial order. For instance, if one tries to predict the continuation of a stimulus sequence on the basis of regularities in the preceding series, or if one tries to reduce memory load by parsing a string of letters into abbreviations, the given serial order is explicitly used to improve performance. However, there are other cases in which it does not seem to be necessary that serial order is detected in order to become effective (Stadler and Frensch 1998). For instance, memory for letter strings generated by a ‘finite state grammar’ improves without the grammatical constraints being reportable. Likewise, in SRT tasks participants respond increasingly faster to statistically ordered stimulus sequences without being aware of the statistical constraints, and consistent response–effect relations have also been shown to decrease response times without being detected. Obviously, serial order can become effective in two ways, once via a strategic use of detected constraints and second via more elementary mechanisms by which serial constraints seemingly inevitably affect performance, irrespective of whether they are detected or not.

4. Theoretical Accounts

The mechanisms underlying the acquisition and use of serial order are not yet well understood. However, it is generally acknowledged that serial learning is accomplished by processes that lead to the formation of *chunks*, i.e., to an integration of successive events into coherent subsequences (Miller 1956). The theoretical accounts differ with respect to the processes which are assumed to mediate the formation of chunks as a result of experienced serial constraints. On the one hand, *explicit strategies* for searching for serial order have been analyzed (Simon and Kotovski 1963). From this perspective, the formation of chunks results from the storage and comparison of subsequences explicitly in the search for identical structures which can be generalized to the whole sequence. On the other hand, it has been acknowledged that constraints in the succession of events also become incidentally effective. For instance, an implicit learning mechanism has been assumed that incidentally associates current/preceding and forthcoming events if the former allow the prediction of the latter. Corresponding simulations in so-called *recurrent networks* have proven that adaptations to grammatical constraints may indeed result from such simple associative learning (Elman 1990, Cleeremans 1993). Other accounts have proposed that events are automatically integrated into a chunk if they are frequently and consecutively experienced in the same succession (Servan-Schreiber and Anderson

1990). Finally, the formation of response–effect associations has also been acknowledged as an autonomous process which may contribute to the formation of chunks in the sense of a series of effects which can be reliably predicted as the outcome of a certain sequence of actions (Hoffmann 1993). Thus, one can conclude that the acquisition and representation of serial order certainly is not a result of only a single process but rather the outcome of different mechanisms which all aim at improving the predictability of forthcoming events.

See also: Lashley, Karl Spencer (1890–1958); Learning and Memory, Neural Basis of; Mathematical Learning Theory; Mathematical Learning Theory, History of; Memory Models: Quantitative; Memory Retrieval; Serial Verb Constructions; Short-term Memory, Cognitive Psychology of; Short-term Memory: Psychological and Neural Aspects

Bibliography

- Cleeremans A 1993 *Mechanisms of Implicit Learning*. MIT Press, Cambridge, MA
- Ebbinghaus H 1902 *Grundzüge der Psychologie*. Veit, Leipzig, Germany, Vol. 1
- Elman J L 1990 Finding structure in time. *Cognitive Science* **14**: 179–211
- Hoffmann J 1993 *Vorhersage und Erkenntnis: Die Funktion von Antizipationen in der menschlichen Verhaltenssteuerung und Wahrnehmung*. Hogrefe, Göttingen, Germany
- Hoffmann J, Koch I 1998 Implicit learning of loosely defined structures. In: Stadler M A, Frensch P A (eds.) *Handbook of Implicit Learning*. Sage, Thousand Oaks, CA, pp. 161–99
- Lashley K S 1951 The problem of serial order in behavior. In: Jeffress L A (ed.) *Cerebral Mechanisms in Behavior*. Wiley, New York, pp. 112–36
- Miller G A 1956 The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological Review* **63**: 81–97
- Miller G A, Selfridge J A 1950 Verbal context and the recall of meaningful material. *American Journal of Psychology* **63**: 176–85
- Reber A S 1967 Implicit learning of artificial grammars. *Journal of Verbal Learning and Verbal Behavior* **6**: 855–63
- Restle F, Brown E R 1970 Serial pattern learning. *Journal of Experimental Psychology* **83**: 120–5
- Rosenbaum D A, Kenny S B, Derr M A 1983 Hierarchical control of rapid movement sequences. *Journal of Experimental Psychology: Human Perception and Performance* **9**: 86–102
- Servan-Schreiber E, Anderson J R 1990 Learning artificial grammars with competitive chunking. *Journal of Experimental Psychology: Learning, Memory, and Cognition* **16**: 592–608
- Simon H A, Kotovski K 1963 Human acquisition of concepts for sequential patterns. *Psychological Review* **70**: 534–46
- Stadler M A, Frensch P A (eds.) 1998 *Handbook of Implicit Learning*. Sage, Thousand Oaks, CA
- Underwood B J, Shaughnessy J J, Zimmermann J 1974 The locus of the retention differences associated with the degree of hierarchical conceptual structure. *Journal of Experimental Psychology* **102**: 850–62

- Ziessler M 1998 Response–effect learning as a major component of implicit serial learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition* **24**: 962–78

J. Hoffmann

Representations and Rituals of the State

‘The state,’ wrote Michael Walzer (1967, p. 194), ‘is invisible; it must be personified before it can be seen, symbolized before it can be loved, imagined before it can be conceived.’ Like all human institutions, the state is a symbolic construction. It is not something that can be touched, nor does it exist independently of people’s ability to relate individual actions to a complex and abstract set of ideas. In this article the ways in which the state is constructed through such symbolic processes and how political actors make use of a variety of symbolic means to exert control over it are examined.

1. The Role of Symbolism in Politics

This perspective on the role of symbolic processes in understanding the state is not universally shared. From the time of the Enlightenment, and through to the current day, many scholars and other analysts have identified the real stuff of politics with the play of material forces and the formation of groups by people defending their common interests. In this view, the symbolic dimension of politics—the flags, the ceremonies, the songs and insignias—is of little consequence.

Politics, however, is at its root a symbolic process, for both the creation of political groupings and the hierarchies they spawn depend on symbolic activity. Political reality does not come to us already formed. It must be created out of the symbolic framework that we have previously acquired. The selective perception and radical simplification of potentially relevant stimuli are necessary if we are to make any sense at all out of the infinite number of potential observations that we could theoretically make. The state, like other political institutions, groupings, and roles—political party, ethnic group, president, judge, king—requires symbolic work of this kind. The notion that the world is divisible into a series of mutually exclusive states, corresponding to lines on a map, and that any individual can be assigned to membership in one of them, is part of this symbolic construction of reality (Anderson 1983, Scott 1999).

1.1 Politics as Nonrational Action

Given that people’s views of the state are symbolically constructed and the symbols on which they base their

views are manipulated by various political actors, it is misleading to see the state as the product of rational processes. This is all the more true because one of the most salient characteristics of the symbolic process is that it not only fashions certain views of reality, but also invests these perceptions with heavy emotional weight. 'A fine symbol,' wrote one Italian analyst, 'has greater weight than a mediocre fact' (Camillo Pellizzi, cited in Gentile 1993, p. 84).

1.2 The Nature of Political Representations

If, as beings who attempt to make sense of the world, we relate to politics not directly, but only through symbolic means, just how does this process work? Anthropologists have long been interested in such questions, although the pioneering work in this field suggested that it was peoples in small-scale, nonliterate societies who were particularly at the mercy of such symbolic activities in their political life. Meyer Fortes and E. E. Evans-Pritchard, in their 1940 volume, *African Political Systems*, which created the subfield of political anthropology, pointed to the mystical credentials of the African ruler, who served as the symbol of the people's unity. 'Members of an African society,' they wrote, 'feel their unity and perceive their common interests in symbols, and it is their attachment to these symbols which more than anything else gives their society cohesion and persistence' (1940, p. 17).

More recent anthropological investigation of state-level politics has retained the Durkheimian emphasis on the key role that symbols and rituals play in promoting social order. But it has expanded the paradigm in two ways. First, the Fortes/Evans-Pritchard formulation regarding the political role of symbolism has been generalized to apply to all societies, no matter how large or complex. Second, it is now recognized that the symbolic process is just as crucial in bringing about political change as it is in maintaining the political status quo. Clifford Geertz's work has had a major impact here. 'At the political center of any complexly organized society,' he wrote, 'there is both a governing elite and a set of symbolic forms expressing the fact that it is in truth governing ... they justify their existence and order their actions in terms of a collection of stories, ceremonies, insignia, formalities, and appurtenances that they have either inherited or, in more revolutionary situations, invented' (1985, p. 15).

The views of Pierre Bourdieu, who has directed attention to the symbolic processes through which potential political leaders create followings, have also been influential. Bourdieu writes of 'the alchemy of representation...through which the representative creates the group which creates him.' Hence we have Louis XIV's magical equation '*L'état, c'est moi*,' or what Bourdieu tells us amounts to the same thing: 'the world is my representation' (1991, pp. 105–6). Of great importance is determining the conditions under which

an individual is in a position to create representations that others are likely to follow, a matter dealt with by Bourdieu in terms of 'symbolic capital.' Anthropologist Simon Harrison (1995, p. 255) has picked up on this approach, viewing political conflict as a battle fought through symbolic means over symbolic capital. 'Competition for power, wealth, prestige, legitimacy or other political resources seems always to be accompanied by conflict over important symbols, by struggles to control or manipulate such symbols in some vital way.'

1.3 The Political Power of Ritual

Ritual provides one of the most powerful mechanisms through which symbolism enters the political realm. Here we adopt a broader view of ritual than the traditional equation of rites with religion. In this broader view, ritual is seen as consisting of culturally standardized, repetitive activity, which is primarily symbolic in character, yet ritual need not have any direct reference to the supernatural realm. Durkheim contributed to this perspective by linking ritual to the division of the world into two categories: the sacred and the profane. Ritual, he argued, consists of 'the rules of conduct which prescribe how a man should comport himself in the presence of these sacred objects' (1912 [1947], p. 41). The fashioning of rituals surrounding the state is in good part a matter of the creation and maintenance of sacred objects and the separation of these objects from the profane world.

Rich in symbolism, ritual derives its potency in part from its ability to arouse the emotions and the sensations through physiological stimulation. Rituals are filled with motions, gestures, colors, noises, songs, and odors. Through rituals emotions are joined to certain cognitions; that is, to certain views of the world.

Following Durkheim, scholars previously emphasized the conservative role of political ritual. Rites surrounding royalty took on a paradigmatic status. The rituals of rulers, the rites of court, and the ritualized means by which rulers interacted with their subjects—as in royal processions—were seen as means by which traditional authority was bolstered. More recently, however, a different perspective has emerged, in which ritual is viewed as a potent form of political action that serves not only power holders but also those seeking to overthrow the political status quo. In societies where the legitimacy of power is contested, we find the political battle fought, in part, through ritual. Rites (a) help develop political organization, not least through linking individuals to larger, abstract political entities (whether states or revolutionary organizations aimed at overthrowing them); (b) legitimize political power holders, would-be power holders, and political institutions; (c) create a sense of solidarity, even in the absence of consensus; and (d) foster certain views of political reality (Kertzer 1988).

The forms taken by political rituals are diverse. Individuals are linked to larger political entities through a variety of ceremonial means, from rites of membership in political parties, to rites of swearing allegiance as a new citizen (dubbed, tellingly, 'naturalization'), to participation in mass demonstrations. Leaders are legitimated through rites of inauguration, and the performance of an assortment of rituals that associate them with powerful symbols. Thus, when the president of the United States makes a public appearance, he enters to the stirring tune of 'Hail to the Chief'; all present rise, and he steps behind a podium decorated with the presidential seal, with the American flag waving from a pole at his side. Likewise, stripping a leader of his power outside regular rhythms of replacement necessitates rites of degradation, as in the case of impeachment proceedings. The death of a political leader requires special ritualization, affording a crucial opportunity for defining political reality and political values, as in the case of state funerals (Verdery 1999). These funerals have their own peculiar grammar, with the presence or absence of other political leaders sending powerful messages that can affect interstate relations.

2. Representing the State

Social scientists have long debated the nature of the state. Anthropologists have paid most attention to it as a stage in sociopolitical evolution, while political scientists debate whether there is anything necessarily Western about the modern state. For our purposes we view the state as consisting of 'the institutions of government and its individual representatives which together possess legitimate means of coercion and ability to tax and plan fiscal policy on a national level.' The state is charged with providing for its population's defense, security, and welfare (Biersteker 1987, pp. 50–70).

Having considered the nature and importance of symbolism to the political process, and the role of ritual in politics, we now turn specifically to how states are represented through symbolism, and the role played by ritual in the state.

2.1 Organizing the State

Given that people can only perceive the state through its symbolism, how do symbols do their work? The French historian, Maurice Agulhon (1985), addressing this question, distinguishes four categories of symbols through which modern states are represented. The state, first of all, must have a flag. Second, it is personified, either by a current or past head of state (e.g., a king or a dictator), or, as in the case of the French republican state, by an allegorical figure (Marianne). Third, each state has a variety of other kinds of symbolic representations of a nonanthropo-

morphic type: these include coats of arms, official seals, and the like. Finally, every regime, writes Agulhon, has its own Pantheon of historical notables, and typically places statues of such heroic figures (or, as the case in the Soviet Union, their embalmed bodies) in public squares.

Historians point to the French and American Revolutions as creative cauldrons for inventing new modes of representation of the modern state, with innovations ranging from the national flag to the national anthem. Historians have also shown that many of these forms of representation, while conveying an image of great antiquity, are in fact relatively recent. They are used by state elites both to solidify their own position internally and in international competition to bolster the state's position in the battle for influence (Hobsbawm and Ranger 1983).

The use of a variety of symbols and rites to link the masses of people to the state took on new importance with the rise of nationalism and democratic politics. When the state's legitimacy was vested in the leader by virtue of supernatural sanction (as in the case of divine right), the main requirement was for representations coming from above to reinforce the view that the rulers ruled by such higher authority. But when states came to rely on the notion that they existed at the behest of their citizenries, with no intermediate social body mediating between citizens and the state, the multifarious ways in which individuals identified with the state became much more important.

The heavy use made of such participatory representations of the state by twentieth-century European dictators, and especially by Hitler and Mussolini, led analysts to pay special attention to the importance of symbolism to the Nazi and Italian fascist regimes. George Mosse (1976, p. 39) has written of the ways in which in Nazi Germany the 'myths and symbols of nationalism, its rites, monuments, and festivals,' transformed an inchoate mass of people into a cohesive mass movement of citizens who saw themselves as in some sense acting as one. Mosse portrayed this process in religious terms, as 'the shaping of the crowd into a congregation.' Others have examined Mussolini's similar attempts to craft a system of symbols and rites to alter Italians' relationship with the state. Yet, in fact, such systems of representation and ritual are common to all modern states.

The importance of ritual in establishing the link between citizens and the state means that there are few more effective means of showing opposition to a regime than through contestation of the rites that bind the citizen to it. Postunification Italy offers a typical example. In the decades following the September 20, 1870 conquest of Rome, which brought with it the final construction of the Italian state, patriotic rites were organized each year to commemorate that sacred date. These annual celebrations, however, served only to inflame the passions of Catholic loyalists, who, following papal instructions, refused to recognize the

new state's legitimacy. The result was that every September 20 Catholics organized counter-rites to challenge the state-organized rites, and thereby challenged the legitimacy of the new state itself (Gentile 1996).

2.2 *Legitimizing the State*

If states come to be known through the ways they are symbolized, any threat to the symbols with which they are identified threatens the legitimacy of the state itself. Harrison (1995, p. 258) refers to symbolic conflict between states of this sort as 'proprietary' contests, involving competition over the state's distinguishing symbols. The bitter dispute that erupted in the early 1990s between Greece and the former Yugoslav republic of Macedonia over the use of the name 'Macedonia' is a case in point.

Harrison distinguishes a variety of means through which the political representation of the group is in a process of continuous contestation and flux. This perspective applies well to the representation of states, again seen both in terms of internal competition and the competition that is waged among them. In the 'expansionary' contest, for example, one competitor attempts to displace another's symbols of identity with its own. This is found whenever revolutionary movements take power. In the case of the Russian revolution, as in revolutions everywhere, these battles were waged in part through the symbolism of space (e.g., St. Petersburg became Leningrad).

Although leaders of new states or new regimes try to replace their predecessors' symbols with their own, they rarely attempt to displace all such symbols. The reason is that, in the eyes of the population, the new regime's legitimacy is almost always tied to its identification with pre-existing symbols of state legitimacy. Hence, the Bolshevik leaders kept the Kremlin as the symbolic center of the state. Alternatively, older symbols of political legitimacy of the state are re-introduced, as in Mussolini's attempts to link his regime to ancient Rome through the symbolic elaboration of *romanità*.

Where older symbolic forms are judged unsuitable, new regimes are as likely to transform them for their own ends as they are to abandon them, for the continuing popular influence of the rites has its uses. Hence, the Nazi regime retained most of the old state holidays—from May Day to Mother's Day—but largely altered their symbolism and ritual in order to promulgate the new state ideology (Taylor 1981). In the post-World War II period, as the Polish Communist government attempted to win popular support, leaders similarly attempted to borrow legitimacy from older symbols, while changing their meaning. Accordingly, the new state authorities chose not to replace the crowned White Eagle, which had served as a national symbol of Poland for centuries. Rather, they simply removed the eagle's crown, thought to be

too blatant a symbol of the old aristocracy. Instructively, on various occasions on which workers protested against the state, they displayed an image of the white eagle recrowned (Kubik 1994, pp. 50–1).

This emphasis on the struggle over representations also allows us to examine how political symbols and rituals are employed by combatants for power within states as a means of establishing their own legitimacy and carving out a particular political identity for themselves. A number of observers, in this connection, have cited the example of the Orange Order's annual parade on July 12, when men wear black bowler hats and orange collars, and march for hours holding aloft banners commemorating battles that three centuries earlier helped establish Protestant rule in Ireland. These rites have some parallels in Independence Day celebrations in the United States, but here the political context makes all the difference. In the United States, the legitimacy of the state is largely undisputed, and the July 4th rites are ceremonies that foster solidarity and reinforce the legitimacy of the state. In Northern Ireland, by contrast, the state's authority remains contested, and such rites as those of the Orangemen provide emotionally powerful means of organizing the combatants.

2.3 *Establishing New Regimes*

As the cases of the French and Russian revolutions show, new regimes require symbolic elaboration, and they must deal with the lingering power of the symbols of legitimacy of the old regimes that remain behind. Struggles over symbols and rites play no small part in the battle to establish such regimes and in the attempts of competitors to delegitimize them. Hence the tricolor flag appeared in France in 1789 immediately following the taking of the Bastille, then was replaced by the royal white flag on Napoleon's fall in 1814, before the July Revolution of 1830 again resulted in the banishment of the white flag and the triumph of the tricolor. Agulhon, in reviewing the history of these developments, argues that battles of this sort, fought through and over symbolism, were central to the course of political events. 'The obscure militants,' he writes, 'who, on July 28, 1830, went to raise the tricolor on the towers of the Notre-Dame and over the Hôtel-de-Ville while the outcome of the struggle was as yet uncertain, contributed to throwing the people into battle with a redoubled ardor' (1985, p. 190).

Rituals of allegiance offer particularly effective means of identifying those whose loyalty to the new regime is weak, and at the same time furnishing ammunition for the regime's claim that it enjoys widespread popular support. Because state authorities can require their performance, such rites force individuals who oppose the regime either to be publicly identifiable (and hence risk reprisal) or to undermine their own self-esteem and their sense of being opposed

to the rulers. Such was the case of the arm-stretched 'Heil Hitler' salute required in daily social encounters during the Third Reich. Similarly, in 1931, concerned about rooting out subversive university professors, and at the same time eager to offer evidence of its claim that it enjoyed overwhelming support from all segments of the population, the Italian fascist regime required that all faculty members swear allegiance to it. Only 12 (of over 1200) professors in the entire country refused to take part in the rite. Their symbolic act of resistance did not go unpunished; they were stripped of their posts (Goetz 2000).

New states formed out of former colonies continue to face dramatic problems of representation and legitimation. In sub-Saharan Africa, where many new states emerged in the 1960s, the difficulties were compounded by the fact that their boundaries were those that the colonial powers had drawn. How could such divisions of land be represented as 'natural' bases for political organization, much less objects of popular adulation? What role would continuity have in such transitions, and if the symbols of the colonial past had to be rejected, what would take their place?

Swaziland provides an intriguing example (Kuper 1978, pp. 301–4). Following decades of British colonial rule, Swaziland gained its independence in 1968, using the figure of the Swazi king as the centerpiece of the new order. The ceremonies marking independence took on great importance to the political elite, who saw them as a means both of defining the new state and building public allegiance to it.

Independence was marked by a great public ceremony, held in an open stadium, packed with thousands of enthusiastic citizens. The king made his entry in a huge new black Cadillac limousine, but as he stepped out people saw that he was wearing traditional regal garb, with high-plumed headgear. After the king took his seat, the British ambassador arrived, as the band played 'God Save the Queen.' From there the ceremonies followed Western form until, suddenly, thousands of swaying Swazi warriors began singing a traditional war song, and the king and prime minister left their seats to join them. As the song reached its crescendo, and the king along with the thousands of warriors raised his shield over his head, the crowd roared its approval. The king and prime minister returned to their seats, later getting in their stretch Cadillacs to take them to a diplomatic reception.

The Swazi case illustrates the power of symbols and rites not only to define a new political reality through the combination of symbols derived from various sources, but also their power to foster popular legitimation and a sense of identification between citizens and the new state.

2.4 The Uses of History

A common feature of representations of the state is that they contain an image of history. It is through

symbolic expressions, including rituals, that leaders link the state today to its earlier incarnations, and likewise link both themselves and their enemies to particular predecessors, viewed in certain ways.

New leaders do all they can to identify themselves with popular predecessors. American presidents need not rely solely on verbal reminders, citing Washington, Jefferson, Lincoln, and Roosevelt in their speeches. By performing certain state rites performed by these predecessors—as most notably in the Inauguration ceremonies—they effectively make this connection.

In the relationships between states, the importance of representing history in certain ways, and establishing legitimacy for the present through symbolic links of a certain kind to the past, is likewise crucial. This was evident, for example, in the difficulties the German government experienced in the decades following World War II in dealing with the historical memory of the war. These came to a head in a series of state rituals held in the mid-1980s. In 1984, leaders of the former Allied nations held a commemorative rite on the anniversary of the Normandy landing. This was similar to many such rites that served to keep the memory of the war alive, and the leaders who participated in it no doubt derived some political benefit themselves. But German leaders, representing the enemy nation, were not invited.

Angered at having the memory of the war represented in this way, German Chancellor Helmut Kohl invited US President Ronald Reagan to come to Germany to celebrate with him a commemorative ceremony in honor of the war dead, to be held at the Bitburg military cemetery. By having the American president stand with him there, the German leader hoped to refashion the history of the war. Indeed, the American President cooperated in this effort, referring to the dead German soldiers as 'victims' of the Nazis, 'just as surely as the victims in the concentration camps.' Yet, history is not infinitely malleable; the symbolism of World War II in the United States could not be so rapidly and radically altered, and as a result Reagan came under sharp attack. Attempting to make partial symbolic amends, the President added a visit to a concentration camp to his itinerary that day, yet this then opened him up to the charge that he was thereby symbolically equating the concentration camp victims with their Nazi persecutors. Despite such criticisms, however, the Bitburg rites had their effect. Opinion polls following the ceremony showed a considerable rise in the proportion of Americans who saw the German soldiers as victims of Hitler rather than as responsible for Nazi atrocities (Kertzer 1988, pp. 92–5).

2.5 The Future of Representations and Rituals of the State

As long as there continue to be states, they will continue to require cultural means of representation,

and in this, ritual in its various forms will play an important role. However, as states face competition for sovereignty and for influence both from within and from without, the role and importance of such representations are likely to change, and some of their forms may alter as well. Insofar, for example, as one role of the rituals we have discussed here is linking the individual to the state, any change in the importance of the state as a basis for personal identity will be reflected in changes in participation in and popular perception of the rituals themselves.

These competing claims for people's allegiance, in turn, will require their own forms of representation, including ritual. Regionalist, ethnic, religious, and other movements opposing the state's power presuppose both the delegitimation of the state—a battle fought in good part through efforts at altering existing representations and rituals of the state—and the creation of an alternative identity. In Italy, for example, the Northern League has devoted tremendous attention to the creation of symbols of representation and rituals that link individuals to a previously nonexistent territory—Padania—and to the League itself (Destro 1997). At the same time, ritualized display of the symbols of the Italian state—such as the Italian flag—is studiously avoided.

In addition to such threats from below (in the Balkans, south Asia, Africa, and other areas), states also face threats from above. For decades theorists have been predicting the demise of the state and the emergence of larger, supranational entities. Many of these (e.g., the Arab League) have foundered on the unwillingness of national elites to cede authority. The most notable current effort, the European Union, demonstrates the link between weak popular allegiance and an underdeveloped system of representation and ritualization (although also suggesting that causation goes both ways). French anthropologist Marc Abélès (1996, pp. 125–32) refers to '*le déficit symbolique*' of the EU, writing that the supranational organization sorely lacks both the symbols and the rituals that could furnish it with some power of popular attraction. Without such symbols and rites, it is unable to mobilize popular passions. The reasons for this deficit are many, but include the inability and indeed unwillingness of national elites to agree on a clear representation of what the EU stands for. Indicative is the fact that while the EU has its own anthem (a version of Beethoven's '*Ode to Joy*'), people cannot sing it, for it has no words. The EU's rites—such as the commemoration of the fortieth anniversary of the European Parliament in 1992—have been so devoid of emotionally resonant ritual elements that they do little but reinforce the EU's image as a creature of Euro-bureaucrats.

Future research must take aim at these potential competitors to the state from below and from above, and examine the ways in which these competitors employ representations in general and ritual in par-

ticular to delegitimize the state. In all such studies, it is crucial that people's reactions to these efforts be examined carefully. It is one thing for state elites or would-be revolutionaries to employ symbolic means to try to win the allegiance of those whose 'representatives' they would like to be. It is quite another for the 'represented' to respond to these symbolic efforts in the ways the elites would like. What is needed is a mix of methods—from participant observation to the analysis of texts—and an examination of a range of social settings, from large periodic national demonstrations to quotidian interactions in rural communities. Political scientists and sociologists need to pay more attention to the symbolic dimension, while anthropologists need to take aim more directly at state-level forces.

See also: Collective Identity and Expressive Forms; Hegemony: Cultural; Legitimacy: Legitimacy: Political; Legitimacy, Sociology of; Political Advertising; Political Communication; Political Culture; Political Discourse; Ritual

Bibliography

- Abélès M 1996 *En attente d'Europe*. Hachette, Paris
- Abélès M 1997 La mise en représentation du politique. In: Abélès M, Jéudry H-P (eds.) *Anthropologie du politique*. Armand Colin, Paris
- Agulhon M 1985 Politics, images, and symbols in post-revolutionary France. In: Wilentz S (ed.) *Rites of Power: Symbolism, Ritual, and Politics Since the Middle Ages*. University of Pennsylvania Press, Philadelphia, PA
- Anderson B 1983 *Imagined Communities*. Verso, London
- Biersteker T 1987 *Multinationals, the State, and Control of the Nigerian Economy*. Princeton University Press, Princeton, NJ
- Bourdieu P 1991 *Language and Symbolic Power* [trans. Raymond G. Adamson]. Harvard University Press, Cambridge, MA
- Destro A 1997 A new era and new themes in Italian politics: The case of Padania. *Journal of Modern Italian Studies* 2: 358–77
- Durkheim E 1912 *Les formes élémentaires de la vie religieuse* [1947 *The Elementary Forms of the Religious Life*. Free Press, Glencoe, IL]
- Fortes M, Evans-Pritchard E E 1940 Introduction. In: Fortes M, Evans-Pritchard E E (eds.) *African Political Systems*. Oxford University Press, London
- Geertz C 1985 Centers, kings, and charisma: Reflections on the symbolics of power. In: Wilentz S (ed.) *Rites of Power: Symbolism, Ritual, and Politics Since the Middle Ages*. University of Pennsylvania Press, Philadelphia, PA
- Gentile E 1993 *Il culto del littorio*. Laterza, Rome [1996 *The Sacralization of Politics in Fascist Italy*. Harvard University Press, Cambridge, MA]
- Goetz H 2000 *Il giuramento rifiutato: I docenti universitari e il regime fascista*. La Nuova Italia, Milan
- Harrison S 1995 Four types of symbolic conflict. *Journal of the Royal Anthropological Institute (n.s.)* 1: 255–72
- Hobsbawm E, Ranger T (eds.) 1983 *The Invention of Tradition*. Cambridge University Press, Cambridge, UK
- Kertzer D 1988 *Ritual, Politics, and Power*. Yale University Press, New Haven, CT
- Kertzer D 1996 *Politics and Symbols*. Yale University Press, New Haven, CT

- Kubik J 1994 *The Power of Symbols Against the Symbols of Power: The Rise of Solidarity and the Fall of State Socialism in Poland*. Pennsylvania University Press, State College, PA
- Kuper H 1978 *Sobhuza II: Ngwenyama and King of Swaziland*. Africana, New York
- Lukes S 1975 Political ritual and social integration. *Sociology* 9: 289–308
- Mosse G L 1976 Mass politics and the political liturgy of nationalism. In: Kamenka E (ed.) *Nationalism: The Nature and Evolution of an Idea*. St. Martin's Press, New York
- Scott J C 1999 *Seeing Like a State*. Yale University Press, New Haven, CT
- Taylor S 1981 Symbol and ritual under National Socialism. *British Journal of Sociology* 32: 504–20
- Verdery K 1999 *The Political Lives of Dead Bodies*. Columbia University Press, New York
- Walzer M 1967 On the role of symbolism in political thought. *Political Science Quarterly* 82: 191–205

D. I. Kertzer

Reproductive Medicine: Ethical Aspects

In reproductive medicine a variety of diagnostic methods are used to discover causes of involuntary human infertility. Diseases in this area can be treated by a number of techniques, ranging in complexity from the simple and inexpensive to the highly sophisticated and expensive (Reiss 1998). The diagnostic methods used include both the invasive, such as laparoscopy, and the noninvasive, such as ultrasonography. They may be specific, in the sense that they give the answer to a question posed, or non-specific, providing information that has not been asked for. Some diagnostic tests are quick, while others take longer, since they have to be processed in laboratories.

The development of a variety of more or less effective approaches to involuntary human infertility, as well as the separation of human reproduction from sexuality, raises a number of ethical problems, discussed in books, journals, and the mass media (Shenfield and Sureau 1997).

1. Ethical Controversies

Cloning has recently been the subject of debate in many countries. There are different kinds of cloning, but the World Health Organization (WHO) and the Council of Europe have banned the cloning of human fertilized eggs for reproductive purposes. The objections to cloning, apart from safety issues, concern the identity and individuality of human beings. Nature produces identical twins, but many people argue that cloning of human beings will undermine the respect of the human being by making its identity and individuality subject to human ingenuity.

Current ethical issues include the disposal of frozen embryos (how and under what conditions this may take place), and posthumous treatment, much discussed in the UK in the late 1990s. Other controversies are raised by third party gamete donation and sperm cryopreservation, particularly concerning payment of donors, anonymity of donors, and selection of donors and recipients (discussed in, e.g., Council of Europe 1996). The different methods used in reproductive medicine raise different ethical problems, and hence need to be considered separately. The issues of oocyte cryopreservation, and reproductive tissue freezing (ovaries and testis), also raise issues of safety and consent—particularly for young people affected by cancer.

For example, assisted insemination using the semen of the husband or designated partner raises a somewhat different problem from that of assisted insemination using the semen of a third party. The need for careful counseling is important in both cases, but the status of the child, its right to know its biological father, needs to be considered in detail when a third party is involved. When semen is frozen, safety issues need to be considered, in addition to how long the semen may be stored as well as under what conditions and for what purposes it may be used. When semen banks are established, regulation is needed. If and when the semen is tested or manipulated, further problems arise.

At present, at the beginning of the twenty-first century, preimplantation diagnosis is limited to sorting out fertilized eggs carrying genes associated with severe diseases for which there is currently no cure, like Huntington's chorea and certain X-linked diseases. Concern has been expressed that the practice may be extended to diseases which are less serious. Obviously, there is a sliding scale between more or less serious and also room for subjective differences in this grading of diseases. There is thus a risk of a slippery slope in this area (Shenfield 1997).

A fundamental issue concerns the moral and philosophical status of the fertilized egg and to what sort of protection it is entitled (Steinbock 1992, Hermerén 1993). The status of the embryo is crucial in many debates of the ethical dilemmas of reproductive medicine. Other basic concerns include the partly culture-dependent reasons for reproduction.

2. A Framework for Analysis

The underlying values in the debates on dilemmas in reproductive ethics can be described in different ways, in terms of the rights and interests of the various stakeholders concerned, or in terms of duties towards patients and the future child. These duties may clash with each other as well as with the interests and rights of other parties. To make the underlying values and value conflicts explicit is to take a first step towards a rational discussion of these dilemmas.

An ethical problem arises in a situation where: (a) there are two or more possible courses of action; (b) one course of action satisfies or is compatible with certain interests, rights, or values, and another course of action satisfies others; and (c) all parties or all interests, rights, or values cannot be satisfied at the same time by one possible action. This is particularly the case if (c) is intensified in such a way that both options are incompatible with each other or if the values, which are included within the options, are conflicting with each other. Whatever alternative is chosen, there will be winners and losers. This is clearly illustrated by some of the well-known dilemmas in reproductive medicine, where the stakeholders (or parties involved) include the woman or the couple, present and future children, doctors and clinics, researchers, the pharmaceutical industry, taxpayers, and society at large.

3. Terminological, Empirical, and Value Issues

If a rational discussion of these ethical dilemmas is wanted, it is necessary to separate conceptual, empirical, and normative issues, as illustrated by the controversies over when life begins and the proposed distinction between embryo and pre-embryo (Jones and Schrader 1989).

Evaluation of criteria of success like 'take-home baby rate' also involves issues of all three kinds. Moreover, whether certain methods would harm children and parents or alter people's understanding of themselves or of the act of procreation are empirical issues that people may disagree about, even if they agree on the goals of reproductive medicine. To be able to substantiate these empirical claims it would be necessary to proceed with long-term follow-up studies of children produced with the help of the various reproductive technologies available.

The utilitarian principle of maximizing the utility or achieving the greatest good for the greatest number of people is open to many interpretations, some of which will clash with Kantian principles that each person should be treated as an end in itself and not as a means to an end. They may also clash with rights-based theories of the sort advocated by Dworkin (1977). This is particularly evident in the debate over research on embryos and the conditions for destruction of spare embryos. Here respect for the autonomy of the gametes providers, by requiring their joint consent, may yield different conclusions than a utilitarian starting point would do. According to the latter, the consent of the gametes providers may be disregarded if it can be shown that using the spare embryos for research or for donation will maximize the greatest good in society. Whereas Swedish legislation requires that sperm donors be willing to reveal their identity to the child when he or she is grown up, if the child wants to know, French legislation requires anonymity

(Shenfield 1997). These different requirements may have an effect on the number and kind of people willing to donate sperm. Interestingly, both requirements can be justified by considerations based on the welfare of the future child. Thus, there is a common agreement on the goal but different views about the best way to achieve it.

These differences could clearly be solved in principle by empirical research, involving long-term longitudinal follow-up studies in both countries of children who are the result of sperm donation. But these studies themselves raise further ethical problems. As the children reach a mature age, it becomes increasingly difficult to carry out the follow-up without their informed consent. How is this to be obtained in a country which makes anonymity compulsory? And can it be obtained without risk of harming or stigmatizing the children?

4. Goals

It may be helpful to distinguish between questions concerning the ethical acceptability or legitimacy of certain goals of reproductive medicine, such as the improvement of success rates in *in vitro* fertilization (IVF), and the ethical acceptability of certain methods used to achieve these goals. As indicated above, it is possible to agree on the goals and differ on the methods, and vice versa.

Do the techniques mentioned above help to achieve the goals of reproductive medicine? Do they help to achieve other goals as well?

If the goal of reproductive medicine is to cure diseases, it becomes necessary to clarify what is meant by a disease to understand this goal. If a particular condition is regarded as a disease, it falls within the goals of reproductive medicine to cure it; if not, it does not.

A fundamental issue is whether involuntary infertility is a disease, the result of a disease or a socially constructed need. This is not just a matter of semantic definition, since the different ways of regarding human infertility have social and economic consequences. If it is a disease, those suffering from it are entitled to help from the national or publicly funded health care system, whose resources should be used. Moreover, if infertility is a disease or a response to a disease rather than a social need, a case for financial support of the new reproductive technologies can also be made, as well as for research on such technologies. If, however, it is a socially constructed need, the case for financial support is much weaker.

5. Means

The methods available to achieve the goals of reproductive medicine vary, and new methods are being

tested. Owing to breakthroughs in molecular biology, development is rapid (Fauser 1999, Harris and Holm 1998). Some methods, like surrogacy, have been criticized because they lend themselves to commercial exploitation and to seeing children as commodities, as 'made' rather than 'begotten' (O'Donovan 1984).

If and when the methods of reproductive medicine involve or presuppose preimplantary or prenatal diagnostics, further issues are raised. The case for preimplantary diagnostics in connection with IVF is that it can be used to avoid X-linked diseases such as hemophilia or certain forms of dystrophy. However, concerns are raised over the possibility that, for instance, indications may be expanded, and that they may be misused for eugenic purposes. Hence regulations or guidelines are necessary, and have been proposed in many countries.

Sex selection for social and economic purposes is practiced in many countries, like India and China. If there are normally about 105 girls born to 100 boys, population statistics indicate that perhaps 100 million girls are 'missing' in the world because of sex selection. The status of women in the world is subject to many cultural differences, and the most important arguments against sex selection focus on the equality of the sexes, emphasizing that gender is not a disease (Shenfield 1994). Suppose, however, that there is no social security for parents when they reach old age. Moreover, suppose that parents have to pay an expensive dowry for their girls, and that girls are not able to make a living by working outside of the home. Then the much-criticized practice of sex selection could be viewed as a culturally chosen means to achieve a certain goal (economic and social safety for the parents). This obviously further suggests that in order to abandon this practice of sex selection, alternative means of reaching this goal have to be made available. Since this may take considerable time, compromise positions have been suggested.

Further concerns have been expressed that, because some of the technologies used in reproductive medicine are expensive they will only be available for rich people in the industrialized parts of the world. They will thus help to increase the gap between the health care and quality of life available for the rich and that for the poor or, more generally, the gap between rich and poor countries.

6. Some Important Arguments

As already mentioned, the basic ethical dilemmas of reproductive medicine involve conflicts of interests, rights, or values. The first step in the attempt at a rational discussion of these conflicts is to make the conflicting interests, rights or values explicit, and to try to separate normative and empirical issues. This is sometimes easier said than done, because, for example, in the discussion of the importance of a genetic relation

between parents and children all these issues are intertwined in a complicated way. Various arguments are used in discussing these conflicts. One rarely takes a stand on the basis of one argument or principle only—often several arguments or principles are used. Sometimes those advocating different positions use the same arguments or principles, but rank the importance of the arguments differently. The central arguments seem to be the following ones: the autonomy argument, the non-maleficence argument (an argument based on the principle that one should not harm), and justice arguments of various kinds.

6.1 Arguments of Autonomy

These are derived from a position that attaches great weight to individuality and liberty. The principle of autonomy is the basis of the requirement of informed consent. If the technologies used involve risks, the principle of autonomy requires careful counseling of the candidate couples. If great importance is attached to this principle, it may overrule also the concern over commodification of human beings, their bodies, and their bodily products. If people have a right to do what they want, they also have a right to serve as surrogate mothers, to earn money on their bodies, and to sell their organs or bodily products. The problem of the limits of autonomy then becomes an important issue. This is also an issue of power: who is to set the limits, and in accordance to what other principles?

6.2 Nonmaleficence Arguments

Arguments based on the principle that one should not harm are clearly legitimate and have a long history in medical ethics ('primum est non nocere,' a principle known already in antiquity). It has been claimed that children who grow up in nontraditional families (for example, with two parents of the same sex) or are the result of particular reproduction technologies will suffer, because they lack role models of both sexes or because their intellectual or emotional development will be impaired. At present, though, the evidence supporting this or the opposite position is missing, or at best inconclusive.

6.3 Arguments of Justice

Arguments based on the just distribution of scarce resources are also relevant. Clearly the resources for health care are not unlimited in any country, and there is always the problem of the fair use of these limited resources. Is the unequal access to treatment also unfair? If fewer than half of those who ask for it obtain assisted treatment by IVF, the reasons why some obtain treatment and others not take on an ethical

significance (Shenfield 1997). Is this a case of discrimination on the basis of economic, social, or ethnic grounds or not? Moreover, the practice in different countries as to how many cycles are subsidized is also relevant here, since incomes are unevenly distributed. For some, a fee of EUR.10000 is a small sum, but to others it is an effective barrier.

Fair access is clearly also a legitimate concern. If the available resources are to be used in an equitable way, and used to benefit those who suffer most, rather than to benefit those who are willing and able to pay, it has to be demonstrated that the methods are cost-effective. Though doctors should respect the autonomy of those who come to seek their help, doctors are not under any obligation, legal or moral, to provide patients with any treatment they might request (McCullough and Chervenak 1994). The request has to be compatible with good professional standards. The proposed treatment might be not benefit the patients or it might be incompatible with the interests of the child.

A related question that has been the subject of heated discussions is who should have access to the benefits provided by some of these reproduction technologies. Should IVF be available for single people, or for homosexual or lesbian couples? Should HIV couples or couples where one of the partners is HIV positive be entitled to treatment with artificial reproduction? Are there any age limits? Are any age limits appropriate and ethically justifiable? If the answer is yes, what sort of inquiries are appropriate before proceeding? In short, on what ground, if any, should some people be denied the access to these techniques?

7. Research Ethics

Here one set of problems focuses on the ethical acceptability of research on embryos *in vitro*, and another on the conditions of such research, if acceptable. According to the views of the Roman Catholic Church, neither such research nor IVF is morally acceptable. Embryo research is also explicitly illegal in some countries, such as Norway, and in practice illegal in others, such as Germany.

If, however, research on embryos *in vitro* is morally accepted, a number of more specific questions have to be asked and answered. They include what kind of research should be permitted, questions about the source of the embryos and for how long may embryos be used for research (14 days is a limit in many countries, like Sweden). Moreover, conditions have been laid down concerning the length of time frozen embryos may be saved and how spare embryos may be destroyed. All these issues are controversial, as is the question of whether embryos may be created solely for the purpose of research. The latter is explicitly forbidden in French legislation and in the Council of Europe Convention on Human rights and Bio-

medicine (1996). The fundamental issue here concerns the philosophical and moral status of the embryo.

The human embryo deserves 'respect,' but a much-contested question is what precisely this respect allows and what it forbids. The practice of research on embryos is subjected to strict regulations, as is experimental research on human beings. Informed consent by the male and female providers of gametes is a cornerstone in these regulations. The knowledge sought must be impossible to achieve in other ways, for instance, by research on animals. Further, this knowledge has to be of vital interest to present and future patients. Others regulations concern safety issues, in particular the safety of the applications of research.

Research on new reproduction technologies (diagnostic methods or treatment options) involves the standard issues of providing patients and healthy volunteers with information they understand and which enables them to make autonomous choices. Obtaining consent from research subjects who have deficiencies in understanding poses other difficulties. In this area of medicine, specific problems include that the fact the embryo and the future child are stakeholders, or parties concerned, but for obvious reasons, they cannot be heard. Someone, therefore, has to look after their interests. Often the parents do this, but sometimes they disagree. Who is then to look after the interests of the embryo or the child-to-be? Moreover, the women or couples seeking help may be under emotional or economic pressure, which will not aid rational decision making.

Yet another problem is over the safety of the woman or the embryo. Some of the methods used in research may involve physical, psychological, or economic risks. Protection against possible harm is a requirement for research—there always has to be a reasonable balance between the anticipated benefits of research and foreseeable harm to research subjects. If unexpected harm arises, however, measures have been taken to interrupt the study and to compensate the research subjects.

When sperm or human tissues are taken from tissue banks or sperm banks for research purposes, special problems have to be dealt with. The conditions under which such research may be performed have been regulated in many countries like the US, Sweden, and the UK.

Finally, there is the problem of fair access to research results, as well as controversies over payment or compensation to research subjects.

See also: Assessment and Application of Therapeutic Effectiveness, Ethical Implications of; Bioethics: Philosophical Aspects; Ethical Dilemmas: Research and Treatment Priorities; Ethics for Biomedical Research Involving Humans: International Codes; Family Systems and the Preferred Sex of Children; Fertility and Culture: Anthropological Insights; Fertility Control:

Overview; Fertility: Institutional and Political Approaches; Fertility: Political and Political-Economic Perspectives; Fertility: Proximate Determinants; Gender and Reproductive Health; Reproductive Rights in Affluent Nations; Reproductive Rights in Developing Nations

Bibliography

- Callahan J C (ed.) 1995 *Reproduction, Ethics and the Law: Feminist Perspectives*. Indiana University Press, Bloomington, IN
- Chadwick R F (ed.) 1992 *Ethics, Reproduction and Genetic Control*, rev. edn. Routledge, London
- Council of Europe 1996 *Convention on Human Rights and Biomedicine*. Council of Europe, Strasbourg, France
- Dworkin R M 1977 *Taking Rights Seriously*. Duckworth, London
- Fauser B C J M (ed.) 1999 *Molecular Biology in Reproductive Medicine*. Parthenon, New York
- Harris J, Holm S (eds.) 1998 *The Future of Human Reproduction: Ethics, Choice and Regulation*. Oxford University Press, New York
- Hermerén G 1993 *The Philosophical and Moral Status of the Embryo*. Council of Europe, Strasbourg, France
- Jones H W, Schrader C 1989 And just what is a pre-embryo? *Fertility and Sterility* **52**: 189–91
- McCullough L B, Chervenak F A 1994 *Ethics in Obstetrics and Gynecology*. Oxford University Press, New York
- Meschede D, Horst J 1997 The molecular genetics of male infertility. *Molecular Human Reproduction* **May** **3**(5): 419–30
- Nuffield Council on Bioethics 1993 *Genetic Screening: Ethical Issues*. Nuffield, London
- O'Donovan O 1984 *Begotten or Made?* Oxford University Press, Oxford, UK
- Reiss H E (ed.) 1998 *Reproductive Medicine: From A–Z*. Oxford University Press, Oxford, UK
- Shenfield F 1994 Control of human sex ratios—sex selection—A matter for fancy or for ethical debate. *Human Reproduction* **9**: 569
- Shenfield F 1997 The ethics of embryo research. In: Shenfield F, Sureau C (eds.) *Ethical Dilemmas in Assisted Reproduction*. Parthenon, New York, pp. 15–21
- Shenfield F, Steele S J 1995 Why gamete donors should not be paid—A gift is a gift is a gift, or why gametes donors should not be paid. *Human Reproduction* **10**: 253–5
- Shenfield F, Sureau C (eds.) 1997 *Ethical Dilemmas in Assisted Reproduction*. Parthenon, New York
- Statens medicinsk-etiska råd. 2000. *Om Livets Början*. Etiska vägmärken, Stockholm
- Steinbock B 1992 *Life Before Birth: The Moral and Legal Status of Embryos and Fetuses*. Oxford University Press, New York
- Strong C 1997 *Ethics in Reproductive and Perinatal Medicine: A New Framework*. Yale University Press, New Haven, CT
- Tong R 1997 *Feminist Approaches to Bioethics: Theoretical Reflections and Practical Applications*. Westview, Boulder, CO

G. Hermerén

Reproductive Rights in Affluent Nations

Reproductive rights can be defined as the freedoms and opportunities that allow individuals to choose if, when, and how to have children. Advocates of reproductive rights have argued that this choice involves access to contraceptives, abortion, and reproductive technologies; freedom from forced sterilization; and access to health care that allows women and their partners to participate in decisions regarding prenatal care and childbirth. Many feminists argue that genuine reproductive choice requires a broad set of changes, including gender equality, employment opportunities, and available high-quality daycare.

1. Historical and Intellectual Context

The notion of reproductive rights is very much a product of nineteenth and twentieth century women's movements. Feminist perspectives on reproductive rights include both individual and social components (Petchesky 1984). The individual element, rooted in liberal thought, is the idea that women must control their bodies to be independent persons. Liberal discourse has been critical to both reproductive rights advocates, who discuss reproductive rights as essential to women's equality, and their opponents, who make claims about the competing rights of fetuses, male partners, and parents of minors (O'Connor et al. 1999, pp. 158–9). The social component, which draws on Marxist thought, is the recognition that material conditions and social relations constrain reproductive choices. As Rosalind Petchesky (1984) argues, reproductive freedom is necessarily individual, involving women's bodies, but at the same time social, involving social arrangements and class and racial divisions that affect reproduction.

In the nineteenth century, birth control movements emerged in many affluent nations. In her history of birth control in America, Linda Gordon (1976) argues that feminist efforts to gain political power through suffrage, and power within the family through birth control, were a response to rapid industrialization; by gaining power and control, women might bolster the family and help to solve social problems. Although nineteenth century advocates of 'voluntary motherhood' disapproved of contraceptives as 'unnatural,' they advocated women's right to refuse sex with their husbands. This demand for the right to refusal was part of the fight against unequal marriage laws, including legalized rape, and was considered a condition of women's independence. In the words of one voluntary motherhood activist, it was 'Woman's Natural Right to ownership and control over her own body-self—a right inseparable from Women's intelligent existence...' (Gordon 1976, pp. 103–4), a sentiment that would be echoed by twentieth-century

feminists in their continuing battles for reproductive rights.

Although nineteenth-century American feminists firmly asserted women's right to control their bodies, they never endorsed contraceptives and never rejected motherhood. Gordon explains their continued disapproval of contraception as a result of their fear that contraceptive use would lead to extramarital sex, and that this would lead to greater freedom for men, but not for women. Similarly in England, nineteenth century feminists failed to support birth control, fearing that women would be victimized by men's sexual desires (Banks and Banks 1964). Lacking economic independence, women needed men's commitment to their wives and children in marriage. The voluntary motherhood movement was an attempt to restrict sexual morality, rather than create sexual freedom, which at that time was not seen as beneficial to women. Women's movements in other Western countries were animated by similar ideologies. It was not until additional social changes made it possible for women to broaden their worlds that women's movements broadened their conceptions of reproductive rights.

By the early twentieth century, women were entering the labor force in greater numbers and sexual activity outside of marriage was on the rise. 'Birth control,' a term coined by US activist Margaret Sanger in 1915, was advocated for both economic and sexual reasons. Socialists in Europe and North America pushed for access to birth control to improve the lives of the working class. Neo-Malthusians, who were concerned about population growth, advocated contraception and were generally supportive of sexual equality (Francome 1984). Advocates of sexual liberation, influenced by European sex theorists such as Havelock Ellis, saw birth control as freeing women of their dependence on men. Early twentieth-century feminists, however, were not in complete agreement about the benefits of birth control for women. Some feared sexual promiscuity and the degradation of motherhood as a source of status for women, and some had race and class biases that influenced their views on birth control. Nevertheless, birth control movements in affluent nations expanded greatly and were linked to women's sexual freedom. When women's movements in various countries were 'reborn' in the 1960s, access to birth control was an important demand.

Movements to legalize abortion also emerged in many countries in the twentieth century. Attitudes toward sexual behavior were becoming more liberal, and various women's groups advocated liberalization of abortion laws. By the 1930s, a number of European countries passed laws that provided for some limited access to abortion. The contemporary wave of abortion reform laws in affluent countries began with the passage of the British Abortion Act in 1967, which was pushed by the Abortion Law Reform Association, an organization founded in 1936. Although the British

law gave medical practitioners rather than women control over the abortion decision, it was a significant advance in women's access to abortion (Francome 1984).

Abortion rights activism increased in the 1960s as women's movements flourished in many Western countries. Feminists demanded legal abortion as part of women's right to control their bodies and developed extended analyses of reproductive rights. In North America, collectives like the Montreal Health Press and the Boston Women's Health Book Collective distributed works such as *The Birth Control Handbook* and *Our Bodies, Ourselves*, which were early sources of information to women about contraception, abortion, childbirth, and other women's health issues. Women's health activists criticized male gynecologists, obstetricians, and other practitioners for demeaning women and failing to include them in decisions about their own bodies. They questioned the safety of some contraceptives and drugs and the necessity of some surgical procedures. They criticized high-technology childbirth practices that took control over childbirth away from women. As more black women and other ethnic minorities were drawn into the women's health movement, they raised the issue of forced sterilization of poor and nonwhite women. As alternatives to established medical control over women's health care, feminists developed 'self-help gynecology' and opened feminist health clinics (Ruzek 1978).

In countries where the women's movement was strong, feminists influenced abortion reform movements to advocate abortion as a women's rights issue. In the United States, the Supreme Court legalized abortion in 1973 on the basis of a 'right to privacy,' recognizing the rights of both women and doctors to make decisions about abortion. Other Western countries also legalized abortion in the 1970s, with varying degrees of input from feminists. In Canada, for example, feminists framed abortion as a reproductive right, but the discourse of medicalization was stronger than that of feminism, and the 1969 abortion reform law focused on the concerns of doctors rather than women (Brodie et al. 1992).

Although feminists continued to be concerned with a broad range of reproductive rights, abortion became a major preoccupation after legalization of abortion provoked strong antiabortion countermovements in many countries. In the USA, the single-issue anti-abortion movement expanded the conflict to new battlefields such as electoral politics, and the abortion rights movement was forced to engage in single-issue political tactics of its own. In the late 1970s and early 1980s, some feminist reproductive rights activists tried to influence single-issue abortion rights groups to broaden their approach, but they had little success at a time when the very legality of abortion was threatened (Staggenborg 1991). Nevertheless, abortion rights groups helped to maintain the larger women's movement in the 1980s and 1990s, attracting

many young women concerned about losing abortion rights. During periods when the antiabortion movement ebbs, reproductive rights groups have more freedom to pursue a broader range of concerns beyond abortion rights.

In sum, notions of reproductive rights have shifted with the larger socioeconomic and political environment. Nineteenth century voluntary motherhood activists, who lacked economic independence, demanded rights that would strengthen women's positions in families. As women gained more economic freedom with expanding industrialization, movement activists advocated birth control and later abortion rights. The rebirth of feminism in the 1960s inspired an elaborated notion of reproductive rights, which questioned existing medical authority. As black and other nonwhite feminists added their voices to the movement, consciousness of race and class issues, such as sterilization abuse, expanded. Countermovement threats to abortion rights since the 1970s limited the concerns of some movement organizations, but also helped to bring new activists into the reproductive rights movement.

2. Research on Reproductive Rights

Social scientific research on reproductive rights has examined reproductive rights movements and countermovements, changes in the medical profession, variations in public policies on abortion, court rulings on abortion, voting patterns of elected representatives, public opinion, demographic profiles of contraceptive and abortion users, public discourse and media coverage of reproductive rights campaigns, and a variety of other issues. Three major categories of research are described here.

2.1 Social Movement Research

Analyses of reproductive rights movements have focused on the conditions giving rise to such movements, their evolving organizational structures, strategies and tactics, conflicts with opponents, and outcomes. With regard to the origins of reproductive rights movements, research points to the importance of the political environment as well as available organizational bases and resources. The US abortion rights movement, for example, arose in the context of the protest cycle of the 1960s, in which the successes of the civil rights movement influenced the growth of other movements. Civil rights victories in the courts were certainly one impetus for abortion rights litigation. Other social movements, including both the population movement and the women's movement, provided organizational bases for the abortion rights movement (see Staggenborg 1991).

Beyond explaining the emergence of reproductive rights movements, analysts have examined their on-

going activities and outcomes. Studies of the women's health movement have examined the difficulties involved in establishing and maintaining feminist health clinics, changes in the movement over time, and the extent to which women's health concerns have become institutionalized in current medical practice. In the case of the abortion conflict, analysts have examined interactions between the abortion rights movement and the antiabortion countermovement. Organizational analyses of movement changes over time show that abortion rights movements in various countries have generally grown in strength in response to antiabortion threats.

To explain the intensity of the abortion conflict, many analysts have sought to connect the conflict to twentieth century shifts in values and lifestyles. For those opposed to these changes, abortion symbolizes a whole cluster of threatened values. In *Abortion and the Politics of Motherhood*, Luker (1984) argues that the abortion battle is a clash between two groups of women. Among activists in California, she found that advocates of abortion tended to be educated, career-oriented women who saw motherhood as only one part of their lives, whereas opponents had committed themselves to the traditional female role of wife and mother and felt threatened by its declining status.

Such case studies have been important in revealing the symbolic meanings of abortion for activists and in showing how the conflict has intensified over time. In Toronto, Cuneo (1989) found that the composition of the antiabortion movement changed over time. Early opponents of abortion were liberal Catholics who defended the civil liberties of the fetus and did not oppose changes in gender roles, but they were gradually pushed out of the movement by defenders of traditional family and religious values who regarded abortion as both a symbol and a cause of declining values. In Fargo, North Dakota, Ginsburg (1989) found that early activists on both sides of the abortion conflict actually shared some common ground, including a concern with community and collective responsibility. However, the conflict became more polarized over time, particularly as media coverage encouraged extremism in the antiabortion movement.

2.2 Analyses of Public Discourse on Reproductive Rights

Related to research on social movements are analyses of public discourse on reproductive rights, particularly abortion. Condit (1990) examines shifts in public discourse on abortion in the USA from 1960 to 1985, using sources such as newspapers and television shows. She demonstrates the persuasive force of different types of discourse, including value-laden phrases such as the 'right to choice,' narratives about women's experiences with abortion, and characterizations such as the 'back alley' of illegal abortion. Condit describes

how early 1960s narratives about women's experiences with illegal abortion, which portrayed ordinary women needing abortions, aroused public sympathy and made abortion a public concern. These narratives were limiting in that they focused on the needs of 'good' women with 'good' excuses for abortion rather than on the reproductive rights of all women. Nevertheless, they helped to change the public view of abortion as 'wrong'—and in doing so provoked 'prolife heritage tales,' which countered that abortion is always wrong. Later in the 1960s and in the 1970s, the argument broadened as abortion rights activists demanded equality for women and the 'right to choice' while antiabortionists argued for the 'right to life' of the fetus. Condit analyses in detail the social force of the different images and rhetoric used by each side and the social compromises that have been negotiated in different arenas in response to the debate.

Comparative research on abortion politics is needed to show how discursive processes vary with different political cultures and institutional structures. Myra Marx Ferree, William Gamson, Jurgen Gerhards and Dieter Rucht have been comparing the changing influences on abortion discourse in Germany and the USA. Ferree and Gamson (1999) analyse the impact of the organization and discourse of the supranational women's movement, the national women's movements in each country, and the political context of abortion policy in each country. The International Reproductive Rights Research Action Group has found both an international diffusion of feminist discourse on reproductive rights and remaining silences on the subject of female sexuality (Petchesky 1997).

2.3 Research on Public Opinion

An extensive body of literature examines public attitudes toward abortion. Analyses of public opinion surveys in the USA show that approval of abortion increased sharply between 1965 and 1973, when the US Supreme Court legalized abortion, and has remained relatively stable since then. Opposition to abortion is correlated with religiosity and other conservative positions on issues of personal morality, and is also linked, more weakly, to opposition to women's rights (see Cook et al. 1992). European data on trends in values suggests a long-term process of secularization and acceptance of individual discretion in terms of personal morality in Western industrialized countries, which may also explain trends in US attitudes on abortion (Rossi and Sitaraman 1988).

Comparative analyses of abortion attitudes in the USA and Canada reveal strikingly similar patterns in the two countries (see Jelen and Chandler 1994). In both countries, abortion attitudes are remarkably stable, with majorities of the public falling into the middle ground of continued support for abortion under certain circumstances. Nevitte and his

colleagues find that support for legal abortion is predicted by similar variables in both countries, including religiosity, 'postmaterialist' values, feminism, and education. Such analyses are important in predicting the future of the abortion conflict. With economic prosperity, increases in postmaterialism and the salience of feminist issues are likely to increase support for reproductive rights, although waning materialist concerns may also increase concerns about abortion among religious fundamentalists who oppose abortion (Nevitte et al. 1993, p. 28).

Other affluent countries exhibit similar patterns. Gallup polls taken in 1981–4 and again in 1990–4 in a number of different countries show that majorities of respondents approve of legal abortion for certain reasons such as threat to the mother's health, and that these attitudes are stable over time (Simon 1998). Although opposition to feminist positions on reproductive rights remains, public opinion in affluent countries is generally supportive of women's access to abortion.

See also: Baby Booms and Baby Busts in the Twentieth Century; Children, Value of; Eugenics as the Basis of Population Policy; Family Size Preferences; Family Theory: Feminist–Economist Critique; Feminist Movements; Feminist Theory: Liberal; Fertility Control: Eastern Europe; Fertility Control: Overview; Gender and Health Care; Gender and Physical Health; Gender and Reproductive Health; Gender and Technology; Gender Role Stress and Health; Gynecological Health: Psychosocial Aspects; Human Rights in Intercultural Discourse: Cultural Concerns; Infanticide and Induced Abortion; Men's Health; Motherhood: Economic Aspects; Poverty and Gender in Affluent Nations; Regulation: Family and Gender; Regulation: Sexual Behavior; Reproductive Medicine: Ethical Aspects; Reproductive Rights in Developing Nations; Sex Preferences in Western Societies; Sexual Attitudes and Behavior; Sexual Behavior: Sociological Perspective; Sexuality and Gender; Sexually Transmitted Diseases: Psychosocial Aspects; Teen Sexuality; Teenage Fertility; Women's Health; Women's Suffrage

Bibliography

- Banks J A, Banks O 1964 *Feminism and Family Planning in Victorian England*. Schocken Books, New York
- Condit C M 1990 *Decoding Abortion Rhetoric: Communicating Social Change*. University of Illinois Press, Urbana, IL
- Cook E A, Jelen T G, Wilcox C 1992 *Between Two Absolutes: Public Opinion and the Politics of Abortion*. Westview Press, Boulder, CO
- Cuneo M W 1989 *Catholics Against the Church*. University of Toronto Press, Toronto, Canada
- Ferree M M, Gamson W A 1999 The gendering of abortion discourse: Assessing global feminist influence in the United States and Germany. In: della Porta D, Kriesi H, Rucht D

- (eds.) *Social Movements in a Globalizing World*. St. Martin's Press, New York, pp. 40–56
- Francome C 1984 *Abortion Freedom: A Worldwide Movement*. Allen & Unwin, Boston
- Gavigan S A M, Brodie J, Jenson J 1992 *The Politics of Abortion*. University of Toronto Press, Toronto, Canada
- Ginsburg F D 1989 *Contested Lives: The Abortion Debate in an American Community*. University of California Press, Berkeley, CA
- Gordon L 1976 *Woman's Body, Woman's Right: A Social History of Birth Control in America*. Grossman, New York
- Jelen T G, Chandler M A (eds.) 1994 *Abortion Politics in the United States and Canada*. Praeger, Westport, CT
- Luker K 1984 *Abortion and the Politics of Motherhood*. University of California Press, Berkeley, CA
- Nevitte N, Brandon W P, Davis L 1993 The American abortion controversy: Lessons from cross-national evidence. *Politics and the Life Sciences* 12(1): 19–30
- O'Connor J S, Orloff A S, Shaver S 1999 *States, Markets, Families*. Cambridge University Press, Cambridge, UK
- Petchesky R P 1984 *Abortion and Woman's Choice: The State, Sexuality and Reproductive Freedom*. Longman, New York
- Petchesky R P 1997 Spiraling discourses of reproductive and sexual rights: A post-Beijing assessment of international feminist politics. In: Cohen C J, Jones K B, Tronto J C (eds.) *Transforming Politics*. New York University Press, New York, pp. 569–87
- Rossi A S, Sitaraman B 1988 Abortion in context: Historical trends and future changes. *Family Planning Perspectives* 20(6): 273–301
- Ruzek S B 1978 *The Women's Health Movement: Feminist Alternatives to Medical Control*. Praeger, New York
- Simon R J. 1998 *Abortion Statutes, Policies, and Public Attitudes the World Over*. Praeger, Westport, CT
- Staggenborg S 1991 *The Pro-Choice Movement: Organization and Activism in the Abortion Conflict*. Oxford University Press, New York

S. Staggenborg

Reproductive Rights in Developing Nations

No sharp distinction can be made between affluent and developing nations with respect to the evolution of the concept of reproductive rights. The notion developed through intense dialogues involving feminist communities in Northern and Southern nations. The concept of sexual rights developed concurrently. The two concepts occasioned theoretical controversies within the feminist community and opposition from national and global conservative political forces.

The two concepts are now enshrined in United Nations documents. In 1994, the Program of Action of the International Conference of Population and Development (ICPD Cairo, 1994) defined reproductive rights as:

[Embracing] certain human rights that are already recognized in national laws, international human rights documents and other consensus documents. These rights rest on the rec-

ognition of the basic right of all couples and individuals to decide freely and responsibly the number, spacing and timing of their children and to have the information and means to do so, and the right to attain the highest standard of sexual and reproductive health. It also includes their right to make decisions concerning reproduction free of discrimination, coercion and violence. ... (Paragraph 7.3)

A year later, the Platform of Action of the IV World Women Conference (IV WOW Beijing) reaffirmed the ICPD principle and spelled out the substantive contents of women's sexual rights:

The human rights of women include their right to have control over and decide freely and responsibly on matters related to their sexuality, including sexual and reproductive health, free of coercion, discrimination and violence. Equal relationships between women and men in matters of sexual relations and reproduction, including full respect for the integrity of the person, require mutual respect, consent and shared responsibility for sexual behavior and its consequences.

1. Historical Background

Almost a century of North–South interactions concerning birth control preceded the recent establishment of the concept of reproductive rights. From the early twentieth century, eugenic societies, neo-Malthusians, the socialist movement, and advocates of sexual hygiene had developed connections with colonial territories and Southern independent nations. Beyond these intentional efforts, ideas had been largely spread worldwide in a more incidental way, by labor migrants, colonial functionaries, and progressive political activists.

The expansion of eugenic societies and Malthusian networks has been extensively examined. The biography of Margaret Sanger, the North American birth control advocate, is a privileged subject of analysis. From the late 1920s, Ms. Sanger was involved in international advocacy in association with neo-Malthusian and eugenic networks. Research has revealed the close links between her efforts, colonial administrations, and local ideologies of class, caste, ethnic, and racial domination. These networks would play a major role in the second wave of global advocacy for fertility control in the 1960s and 1970s. By then Ms. Sanger had sponsored research leading to the invention of the contraceptive pill which, together with the first IUDs, were being disseminated in developing nations.

In contrast, the path through which Emma Goldman's socialist views on women's reproductive self-determination traveled south remains largely unexplored. The Cuban experience suggests a resonance far beyond the pioneering initiatives taken by Goldman in the early days of the Soviet Union, as abortion was legalized in 1959, immediately after the revolution. Recent research has also drawn attention

to some of the more progressive ramifications of Ms. Sanger's endeavors. One illustration is provided by Mexico where, in 1916, in Mexico, after a Feminist Congress, a booklet written by Margaret Sanger, *'Birth Regulation,'* was disseminated. In 1925, the same publication was distributed freely by the anti-clerical Calles Administration.

Having originated from contradictory theoretical sources, birth control ideas were given different meanings by diverse interpreters. In India, where the spread of neo-Malthusian ideas coincided with the intensification of the independence struggle, they sparked harsh controversies within the Gandhian movement. In Mexico, contrasting with the climate of the previous decade, a pronatalist General Population Law was approved in 1936. In the same period, mother and child care programs and family benefits were adopted in Brazil, Argentina, Chile, and Uruguay. These initiatives, which were conceived as accompaniments to early industrialization strategies, converged with the Catholic Church's resistance to all variants of birth control advocacy. This climate silenced existing progressive strands and inspired adaptations of conservative eugenic streams. In Brazil, for instance, ideologies advocating racial and ethnic miscegenation to 'whiten' the population gained legitimacy.

This background determined the immediate post-Second World War environment. In Latin America, previous such policies gained leverage from the adoption of import-substitution industrialization strategies. In other regions, the acceleration of decolonization brought the idea that large populations were a positive element in the establishment of new nation states. Rapid economic growth favored a lowering of mortality rates worldwide. However, as industrialized nations entered a second cycle of fertility decline, in developing nations, where demographic transitions were taking place, fertility rates remained high. In the 1950s, Alfred Sauvy, the French demographer, conceived the categories *First, Second, and Third World*. The central parameters for this demarcation were the demographic differentials between the three zones.

In the late 1960s, fertility control premises became a 'grand ideology,' supported by United Nations agencies, public financial institutions, bilateral donors, private foundations, and the International Planned Parenthood Federation and its national associates. Scientific evidence was collected to prove that rapid demographic growth had negative impacts on the economic development of Southern countries. Extreme positions went as far as to argue that population control would have the additional benefit of preventing revolutionary uprisings across the South.

Although pressured from outside, developing nations were not easily or immediately persuaded. Indeed, at the Population Conference of Bucharest of 1974, they counteracted by proposing a policy agenda that combined reform of the international economic

order, domestic economic growth and investments in basic needs. This position became crystallized in the lemma: 'Development is the best contraceptive.' Yet, not all developing nations remained faithful to their Bucharest commitment. By the late 1970s, draconian population control was already in place in Asia, particularly in China and India. In the other regions countries resisted longer.

Feminist initiatives were also blossoming south of the Equator. Feminist groups initially supported the Bucharest critique of population control. Gradually, however, they came to perceive that women's needs and rights were not being addressed by either pronatalist or antinatalist programs. In some settings, women's fertility was targeted. In others, women's ability to freely decide on reproductive matters was curtailed by pronatalist ideologies. In a few countries the two approaches were implemented in combination, the racial and ethnic origin of the client being the determining factor for selecting which approach was used for whom. Overall, the quality of information and care provided by both MCH and family planning programs was extremely poor.

The early feminist critique of population policies was preceded and surrounded by other relevant institutional processes. The 1968 International Human Rights Conference, held in Tehran, approved a principle ensuring couples the '*basic human right to decide freely and responsibly on the number and spacing of their children and a right to adequate education and information in this respect.*' In Bucharest, it was rephrased to extend these rights to individuals (not just couples) and to ensure access to the means of exercising them. Nevertheless, until the 1990s, individual rights in regard to reproductive matters remained subordinated to the primacy of preventing the population crisis.

Bringing gender concerns into family planning frameworks evolved along with other efforts aimed at women's equality. In 1975, the First United Nations sponsored International Women's Conference was held in Mexico City. Coercive measures of population control were debated and the final document proclaimed respect for the bodily integrity of women as '*a fundamental element of human dignity and freedom.*' This was followed by the approval of the Convention on the Elimination of All Forms of Violence against Women (CEDAW) in 1979. CEDAW affirmed women's rights to family planning, their equal right (with men) to decide on the number and spacing of their children, the right to maintain their jobs while pregnant, and to have access to maternity leave and child-care. In 1981, a UN Symposium on Population and Human Rights held in Vienna also declared that '*Both the compulsory use of abortion and its unqualified prohibition would be a serious violation of human rights.*'

Despite this emerging consensus, the climate of debates began to change. In the International Population Conference of Mexico in 1984, developing

countries expressed a broader acceptance of family planning programs and reaffirmed women's right of access to contraception. However, an alliance between the Reagan Administration and the Roman Catholic Pope made the subject of abortion a major controversy. The United States cut funding for international family planning activities, saying that the programs facilitated access to abortion. (This US policy remained in place until 1993.) Some family planning organizations openly contested this US policy and disseminated information about it among their Southern partners. Abortion was openly debated at the NGO Forum of the Women's Conference of 1985 in Nairobi. In 1988, the meeting of the International Federation of Gynecology and Obstetrics in Rio de Janeiro was preceded by the Christopher Tietze International Symposium on Women's Health in the Third World, a critical event in addressing abortion rights internationally. By the end of the decade, the World Health Organization adopted the notion of reproductive health that would be rapidly incorporated by other institutions and advocacy groups becoming an important conceptual adjunct to the idea of reproductive rights.

Other contributions to the conceptualization of reproductive rights came from women's organizations in Southern nations. Though small-scale, these grassroots organizations engaged in constant dialogue with their societies to transform sexist attitudes and practices. Sometimes they influenced legislation and policy to enhance women's health. In all contexts, feminist organizations had to confront the powerful influence of international population policies or the resistance of pronatalist ideologies.

This blossoming of feminist advocacy coincided with the first ripples of globalization. The debates evolving in industrialized countries filtered into Southern societies through various communications streams. Feminist publications such as *'Our Bodies, Ourselves'* were read in English speaking circles. Political exiles brought back home the stories of feminist struggles in relation to reproduction and sexuality and adapted innovative consciousness raising strategies, as in the case of gynecological self-examination. The mainstream media provided some information.

Local groups built connections with each other and with the Committee for Abortion Rights and Against Sterilization Abuse (CARASA) in the United States, and the International Campaign on Abortion, Sterilization and Contraception (ICASC) in Europe, which had spawned the very first formulation of reproductive rights. In 1984, a first Latin American Women's Health Meeting was held in Colombia. The International Campaign on Abortion, Sterilization, and Contraception also hosted, in Amsterdam, the first International Reproductive Rights Conference.

In Amsterdam, the International Campaign was renamed the Women's Global Network for Repro-

ductive Rights (WGNRR) and given a broad mandate to address all issues related to women's reproductive health, not only those involving abortion, sterilization, and contraception. Reproductive rights were defined as *'women's right to decide whether, when and how to have children—regardless of nationality, class, race, age, religion, disability, sexuality, or marital status—in the social, economic and political conditions that make such decisions possible.'*

The next WGNRR conference held in San Jose, Costa Rica (1987) started an international campaign on maternal mortality and inspired the creation of the Latin American Campaign for the Decriminalization of Abortion. In 1988, in a feminist meeting preceding the Tietze Symposium in Rio, the debates concentrated on ethical aspects of population control policies, abortion, and new reproductive technologies. In Manila, in 1990, during another WGNRR Women's Health Conference, the idea was raised that the international reproductive rights movement should conceptualize a feminist population policy. As globalization intensified, a clear acceleration occurred in the circulation of these ideas.

The United Nations Conferences of the 1990s were conceived as an interconnected series of global debates starting with Environment and Development (Rio 1992) and moving forward to address Human Rights (Vienna 1993), Population and Development (Cairo 1994), Social Development (Copenhagen 1995), Women (Beijing, 1995), Human Settlements (Istanbul, 1996), and Food Security (Rome, 1996). Whereas in the 1970s and 1980s, Southern feminist intellectual and political endeavors evolved largely in isolation from international policy arenas, in the 1990s, 'they involved' meant direct engagement in these negotiations. In Rio, access to reproductive health was added to conventional family planning language. Vienna's definitions with respect to women's human rights affirmed that they applied to both the public and private realms, thereby creating the basis for the further developments of Cairo and Beijing regarding reproductive and sexual rights.

2. *The Contents of Reproductive and Sexual Rights in the Southern Feminist Perspective*

The adoption of a reproductive rights framework by advocacy groups within developing nations entailed a broadening of its original focus—abortion, contraception, and sterilization—to include maternal mortality, prenatal and obstetric care, breast and cervical cancer, and adolescent and postmenopause health. The Southern perspective also underlined the need to incorporate attention to women's economically productive roles in addition to their biologically reproductive functions. This required a better understanding of the correlation between sexual and reproductive

health and endemic and chronic diseases, as well as occupational, environmental and mental health hazards, and afflictions.

In the 1980s, the scope and intensity of abortion-related initiatives varied across regions, being more visible in Latin America (particularly Brazil), the Philippines, and South Africa. Abortion was always addressed as part of a broader reproductive health and rights agenda, never as a 'single issue.' Efforts to decrease maternal mortality were prioritized by most countries and within these efforts new venues for addressing abortion have emerged. Both ICPD and IV WCW final documents recognize abortion as a major public health problem, affirming that, in circumstances where it is legal, it should be safe. The Beijing Platform of Action recommends that countries review laws containing punitive measures against women who have undergone abortion.

Contraception has always been a cross-regional concern, often involving United States and European advocacy efforts. Research has been devoted to sterilization, IUDs, injectables and implants, methods privileged by fertility control programs. Long before the feminization of the HIV-AIDS pandemic, Southern feminist organizations emphasized the disease-preventing advantages of barrier methods, such as the diaphragm.

In the 1990s, campaigns were launched against specific technologies: Norplant, the contraceptive vaccine, and Quinacrine (a method of chemical sterilization). Advocates engaged in dialogues with researchers, pressuring for better ethical standards and redefined research priorities. One critical outcome of these efforts was the investments aimed at developing microbicides to protect women against HIV-AIDS and pregnancy. Concern arose regarding the expanding use of prenatal sex-selection techniques in India and China.

In their efforts to address reproduction as a right, not merely a woman's obligation, Southern advocacy groups very early challenged gender inequalities in the sexual domain. Issues such as early marriages, rape, forced prostitution, trafficking, and female genital mutilation have been raised under the umbrella of violence against women. Sexuality, as such, however, gained enormous visibility under the impact of the HIV-AIDS pandemic. The inadequacy of contraceptive methods used in most Southern countries became clear, as they offered no protection against infections. Research in Africa showed the association between HIV and other sexually transmitted infections.

The World Health Organization coined the notion of sexual health, and initiated preventive strategies that required a better understanding of sexual behavior. Most Southern governments resisted recognizing the impact of the pandemic, partly because it was initially portrayed as affecting only certain risk groups (homosexuals, sex workers, and drug addicts). Among Southern feminists, some time elapsed before

the feminization of HIV-AIDS became a priority. An important recent political development is the multiplication of gay and lesbian rights advocacy groups in developing nations. The concept of sexual rights presently encompasses nondiscrimination on the basis of sexual orientation or HIV-AIDS status. Other ideas have diffused that promote responsibility by men in sexual matters and transforming notions of masculinity.

Thus, although the emphases do vary, the content of reproductive and sexual politics within affluent and developing nations are similar. The concepts exemplify an 'ideoscape' which, having broken through national and cultural boundaries, has gradually crystallized in a globally accepted discourse. These ideas are present in Southern national and regional advocacy groups and among Southern academic and research initiatives, particularly in Latin America and South and South East Asia.

3. Conceptual Debates

A number of inter related controversies developed as reproductive and sexual rights were conceptualized in developing nations. A first one can be retraced back to the global debates around abortion in the 1980s. In developing countries, should abortion be addressed as a single issue, as in the United States advocacy strategies, or part of a comprehensive reproductive health and rights agenda? A theoretical and political problem underlay this early controversy. In affluent societies, where women's basic needs had been largely met, restrictions in the access to abortion strikingly expresses gender inequality with respect to the individual right to decide. However, women's exercise of reproductive rights in developing regions (and among marginalized groups in affluent societies) is primarily curtailed by economic disadvantages that affect men and women.

A related controversy arose from the critique of the intrinsic individualistic and contractual nature of the Western concept of rights. Feminist thinkers recognized that any approach to sexual and reproductive politics must take into account political, economic, and cultural constraints on individuals. The full exercise of these rights is dependent upon an enabling environment including democratic conditions that allow citizenship rights, women's empowerment, and material support, such as transportation, childcare, jobs, and education. Examples of the importance of such support are easily found. In all developing nations, women die because they do not have transportation or because basic equipment is lacking in health facilities. In rural India and China, vaginal infection results from precarious sanitation. Poor women die of botched abortions, while rich women can pay for safe procedures. In India, abortion has been legal since the 1970s, but the quality of care

remains poor. Young women are infected by HIV-AIDS by having sex to get money to survive.

The emphasis on enabling conditions does not imply ignoring gender inequality and how it restricts women's ability to protect their bodily integrity and personhood. Research that explored ordinary women's perceptions of sexual and reproductive rights found that few women use these terms, but they do understand that the principles refer to being able to make decisions about procreation, to be free from harassment by men, to protect themselves against domestic violence and preventable illness, and to have access to adequate health care. When these rights are not in place, women cannot exercise civil and political rights. The framework implies both positive rights, such as reproductive health care and employment, and negative rights, such as protection against rape and discrimination, regardless of sexual orientation or marital status.

Some have criticized the emphasis on sexual and reproductive health as being too narrow when compared with the broad women's health agenda defined in the 1980s. To address this limitation requires understanding links between gender, social inequality, and health. One priority is to ensure that health reform policies improve women's access to health services. Another concern is to understand how men's violence against women affects women's health and mount policies to confront this. Exploring this topic may also lead to addressing the adverse effects on women of cultural constructs of masculinity.

4. Policy and Political Challenges

A human rights approach to sexual and reproductive politics still faces major obstacles to effective implementation. The definitions of related rights as enshrined in United Nations documents are not binding. Even if these new principles are translated into binding instruments, international human rights mechanisms have little power to enforce them, especially if violators are private parties rather than states.

Additionally, the effective implementation of women's reproductive and sexual rights must challenge a rigid and persistent 'public'-'private' distinction in international human rights law and national legal systems. This breaking up of boundaries is crucial for women, whose lives in many cultures are still locked within domesticity. To help women, human rights claims must penetrate the 'private sphere' within which marital rape, female genital mutilation, virginity codes, and repudiation of birth control occur. On the other hand, public actions of the state and its agents also very often violate women's sexual and reproductive self-determination.

Greater challenges may lurk in globalization. Poorly understood and poorly regulated, globalization is creating growing inequalities of income within and

across countries. Women are entering labor markets everywhere. Earning more income often raises a woman's value, but can also result in even tighter controls over her life. Earnings sometimes provide women with greater physical mobility, increased personal autonomy, and the possibility of breaking through gender barriers and patriarchal or other mechanisms of male control.

Concurrently—in reaction to the negative effects of globalization—political, national, religion-based, ethnic, or other identities are being strengthened which revive traditional gender systems and controls. There is a proliferation of both antiabortion and 'profamily' groups associated with religious fundamentalists—Catholic, Protestant, Islamic—and of conservative politics that systematically oppose women's reproductive and sexual freedom on behalf of 'fetal rights,' womanly duty, and the defense of hierarchical power within the family. These forces are at play in the United Nations, and other global policy arenas, whenever gender related issues arise. Moreover, racism, nationalism, and ethnic hatreds also pose threats to women's reproductive and sexual health. Chronic violence, genocide, 'ethnic cleansing,' and military occupation have often meant mass rape, abandonment, refugee status, increased vulnerability to forced prostitution, trafficking, unwanted pregnancy, and sexually transmitted diseases.

The promoters of a globalized economy often support the breaking down of traditional gender systems and greater freedom for women. Nevertheless, they remain largely unconcerned with growing economic inequality which, in most cases, lies at the root of fundamentalist trends, patriarchal revivals, and ethnic strikes. For developing countries, globalization has led to large-scale debt and trade inequities. Fiscal austerity, imposed through structural adjustment programs, has negatively affected the availability and quality of public social services.

The fulfillment of the twentieth century promise of women's sexual and reproductive freedom requires these contradictory trends to be challenged and eventually tamed. Redefining development policy priorities to create an enabling environment that will both provide a firm basis for individual self-determination while restraining the appeal of moral conservative forces remains a daunting task for the early years of the new millennium. The historical experience in both the South and the North also suggests that such fulfillment additionally demands sustained intellectual and political efforts aimed at promoting and enhancing the acceptance of all forms of diversity and tolerance within and across societies.

See also: East Asian Studies: Gender; Economic Development and Women; Economic Globalization and Gender; Eugenics as the Basis of Population Policy; Family Planning Programs: Development and Outcomes; Family Planning Programs: Feminist Per-

spectives; Family Planning Programs: Quality of Care; Family Theory: Economics of Intergenerational Relations; Family Theory: Economics of Child-bearing; Family Theory: Economics of Marriage and Divorce; Family Theory: Feminist-Economist Critique; Female Genital Mutilation; Feminist Movements; Fertility and Culture: Anthropological Insights; Fertility Change: Quantum and Tempo; Fertility Control: Overview; Fertility: Political and Political-Economic Perspectives; Fertility Transition: China; Fertility Transition: East Asia; Fertility Transition: Latin America and the Caribbean; Fertility Transition: Middle East and North Africa; Fertility Transition: South Asia; Fertility Transition: Southeast Asia; Fertility Trends in the Formerly Socialist Countries of Europe; Gender and Health Care; Gender and Physical Health; Gender and Reproductive Health; Gender and Technology; Gender Role Stress and Health; Gynecological Health: Psychosocial Aspects; Health in Developing Countries: Cultural Concerns; HIV and Fertility; Infanticide and Induced Abortion; Labor Markets, Labor Movements, and Gender in Developing Nations; Land Rights and Gender; Latin American Studies: Gender; Lone Mothers in Nations of the South; Male Dominance; Masculinities and Femininities; Men's Health; Motherhood: Economic Aspects; Motherhood: Social and Cultural Aspects; Nation-states, Nationalism, and Gender; Near Middle East/North African Studies: Gender; Population Policy: International; Poverty and Gender in Developing Nations; Regulation: Sexual Behavior; Reproductive Medicine: Ethical Aspects; Reproductive Rights in Affluent Nations; Rural Industrialization in Developing Nations and Gender; Science and Industry; Southeast Asian Studies: Gender; Women's Health

Bibliography

- Blatliwala S L 1994 The meaning of women's empowerment: new concepts from action. In: Germain G S A, Chen L C (eds.) *Population Policies Reconsidered: Health, Empowerment and Rights*. Harvard Center for Population and Development Studies, Boston
- Bunch C 1993 *Organizing for Women's Human Rights Globally, in Ours by Right: Women's Rights as Human Rights*. Zed Books, London
- Chesler E 1992 *Woman of Valor: Margaret Sanger and the Birth Control Movement in America*. Simon & Schuster, New York
- Cook R 1995 *Human Rights and Reproductive Self-Determination*. Washington, DC
- Corrêa S, Petchesky R 1994 Reproductive Rights: A Feminist Perspective. In: Sen G, Germain A, Chen L C (eds.) *Population Policies Reconsidered: Health, Empowerment and Rights*. Harvard Center for Population and Development Studies, Boston
- Corrêa S, Reichmann R 1994 *Population and Reproductive Rights: Feminist Perspectives from the South*. Zed Books, London

- Dixon-Mueller R 1993 *Population Policy and Women's Rights: Transforming Reproductive Choice*. Praeger, Westport, CT
- Hartmann B 1995 *Reproductive Rights and Wrongs*. South End Press, Boston
- International Planned Parenthood Federation 1995 *IPPF Charter on Sexual and Reproductive Rights*. International Planned Parenthood Federation, London
- Petchesky R P 1995 From population control to reproductive rights: feminist fault lines. In: *Reproductive Health Matters*
- Petchesky R P, Judd K (eds.) 1998 *Negotiating Reproductive Rights: Women's Perspectives Across Countries and Cultures*. International Reproductive Rights Research Action Group, London and New York
- Peters J, Wolpe A (eds.) 1995 *Women's Rights, Human Rights: International Feminist Perspectives*. Routledge, New York and London
- United Nations 1994 *Programme of Action of the International Conference on Population and Development*. United Nations, New York
- United Nations 1996 *Beijing Declaration and Platform for Action, adopted by the Fourth World Conference on Women*. United Nations, New York

S. Corrêa

Republic

'Republic' comes from the Latin *res publica*, the shared or common thing. It first described the political regime instituted by the Romans after they deposed their kings in 509 BC. The founding of the Republic symbolized the rejection of rulers who treated the political community as their private property. The word *publica* probably derives from the Latin *populus*, 'people.' Cicero in the first century BC calls the *res publica* a *res populi*, a thing that belongs to the people. That the political community belongs to the people rather than those who rule reverberates throughout the long history of the term as it takes on and sheds concepts drawn from the Roman experience.

1. The Roman Legacy

Polybius, in the second century BC, described (not entirely accurately) the Roman Republic as a 'mixed regime' where monarchy, aristocracy, and democracy balanced one another, providing stability and protection from the constant cycling of regimes from one form to another. Sallust, writing during the political chaos of first century BC Rome, turned away from institutional analyses to character and emphasized the foundation of the Republic on political virtue and traditional moral values, qualities whose loss would bring about the demise of the Republic. The civil wars ended with the Principate of Augustus which returned Rome to one-man rule. Tacitus, writing in the early

years of the Principate, recalled the liberty of the Republic in contrast to the tyranny of the increasingly despotic Roman emperors. All these elements, the mixed regime, political and moral virtue, and liberty as the absence of the oppressive rule of a king have followed the word 'republic,' though neither consistently nor necessarily in unison. 'Republic' recalls Rome, but the Greek city-states, different in structure and size, also came to be called republics, and Athens and Sparta, along with Rome, became models of republican regimes for later ages.

During the Middle Ages, *respublica* simply meant the political community in contrast to the church; it was the body of which the prince was the head, the wife to whom the prince was joined in matrimony. It incorporated none of the elements of liberty, institutional balance, or political virtue that connected the idea of republic back to the experiences of Rome. At the close of the medieval period, however, republic recaptured the connotations of liberty understood as hostility to tyrannical rule. The early Italian renaissance witnessed a shift from medieval praise of the Principate as a reflection of God's rule to admiration for the political liberty of the Roman Republic. A number of the Italian city-states of the Renaissance gained political independence and instituted republican regimes of self-rule. Meanwhile, humanist writers of the early Renaissance, recalling republican Rome, appealed to the liberty and political virtues of antiquity against the faith in a God-willed monarchy that dominated the medieval period. Nevertheless, tensions arose between republican virtues which fostered renown for service to the secular community and religious virtues which glorified God. With Machiavelli in the early 1500s, concern about devotion to and participation in a secularized polity disappeared.

2. Machiavellian Republicanism

Machiavelli's begins *The Prince* by rejecting the conventional Aristotelian typology of monarchy, aristocracy, and democracy; he boldly stated that all dominions 'now and in the past' have been 'either republics or principalities.' Machiavelli's dichotomies never persevere throughout his writings, but *The Prince* largely treats principalities and his *Discourses on the First Ten Books of Livy* (1513) discusses republics with special attention to Rome. Other republics like Sparta and Venice surface as well offering variations and qualifications on the Roman model. More nuanced than works by Renaissance writers who simply identified republics with liberty, or sought to find a republican heritage for their cities in the history of the Rome, Machiavelli identifies the tensions inherent in republican regimes, their occasional need for dictatorships, their dependence on conflict between parties as a condition for freedom, and their glory in imperialist expansion. Machiavelli,

along with republican authors of the period, praised the mixed regime, but he saw it as necessary for a regime to order the conflicts that plague any free polity. Further, while praising republican virtues of devotion to and participation in the political regime, and while worrying about the effects of a corrupt self-interested citizenry, he also recognized that republican virtues entailed the conquest of others.

Despite Machiavelli's somewhat ambiguous presentation of a republican Rome that needed occasional dictators and men like Brutus who were willing to sacrifice their sons for the preservation of the regime, his *Discourses*, with its reliance on Polybius' mixed regime, became a primary source for antiroyalist, republican literature of seventeenth century England. James Harrington, for instance, begins his utopian republican tract *Oceana* praising Machiavelli as the sole retriever of ancient prudence concerning governments devoted to the public interest. In addition to praising economic equality as the foundation of a peaceful regime, Harrington continued the republican theme of the mixed government by proposing a bicameral parliament where an elected and rotating aristocratic senate would be checked by a popular assembly. As with the other republican authors of his time, Harrington relied on ancient Sparta and contemporary Venice along with Rome as models. All demonstrated the viability of the republican model with their mixed regimes, and all were frequently paraded as exemplars of popular engagement and liberty in contrast to monarchical oppression.

The participatory expectations meant that seventeenth and eighteenth century English authors often associated republicanism with the aristocratic leisure essential for an engagement in political life. The contemporaneous emergence of liberal thought based on the rational pursuit of self-interest set up a conflict between an elitist republic relying on men of public virtue with resources and leisure for public life and the egalitarian principles of a liberal society focused on rights, consent (a much weaker political act than actual engagement), and the absence of oppression. The liberal perspective did not require self-sacrifice; the liberal individual, unlike the republican, pursuing private interests could be oblivious to the public good. Liberal virtues of frugality and moderation supporting economic success were not the republican virtues that supported the moral and political life of a nation.

3. Enlightenment Republican Thought

The English republicans, drawing on the Polybian mixed regime, articulated a theory that allowed for the presence of a monarch, albeit one limited by assorted checks. Debates arose as to whether the checks were sufficient or too strong, but a monarchical regime that might allow for liberty contrasted starkly with the absolutism of eighteenth century France.

Montesquieu's *Spirit of the Laws* appeared in this latter context. Montesquieu replaced the still dominant Aristotelian typology with despotism, monarchy, and republic; the last he divides into the aristocratic republic, marked by the love of honor, and the democratic republic, characterized by 'political virtue' or the 'love of equality.' Drawing on themes of civic virtue from Sparta, the Roman Republic, and the *Discourses* of the 'great man' Machiavelli, Montesquieu writes that political virtue in a democratic republic means the denial of self for the sake of the whole. According to Montesquieu, the democratic republic would be possible only in a small, homogenous community where there was no attention to individual economic success nor worries about economic inequalities. The simplicity of the democratic republic would allow for none of the conflicts that arise with the emergence of the wealthy few. Montesquieu's democratic republic depended on political virtue, but it did not offer much happiness. In contrast to the democratic republic, Montesquieu wrote about a free and happy people who inhabited a large commercial nation, namely England, that with its 'mixed regime' could incorporate variety, economic activity, and inequality in a way that a democratic republic could not and that could offer liberty through the just execution and impartiality of its laws.

Montesquieu's praise of the English constitution suggested freedom would be possible in large, heterogeneous regimes, if laws were well executed and a proper balance between executive and legislative powers and within the legislative branch existed. Rousseau's *The Social Contract*, in contrast, affirmed that freedom was possible only in small homogenous republics where all citizens, as members of the sovereign body, participated in the life of the city through law making. A large city where citizens devote themselves to commerce, crafts, and comforts rather than to political life undermines attachment to the community. Free regimes, according to Rousseau, depend on the political engagement of all in self-rule; we are slaves if do not legislate for ourselves. Rousseau returned to Rome for his model of participatory republican life where, in addition to popular assemblies, there was a censorial tribunal ensuring the purity of opinions and morals. Rousseau's republic depended on a similar control of citizens' opinions and required a civil religion to instill 'sentiments of sociability.' A decline in morals would bring about political decay and the loss of the self-rule at the heart of free republics. Preservation of morals and attachment to the city for Rousseau rested on maintaining a small state, unperturbed by the bustle of economic activity or the cosmopolitanism of Parisian culture.

4. The Issue of Size

The question of whether republican liberty was possible in large nations or only in small ones was central

to the debates about ratification of the US Constitution. *The Federalist Papers*, arguing for adoption of the constitution, drew upon the equation between a republic and liberty. Using Publius as their common pen name, the authors of *The Federalist Papers* identified with a founder of the Roman Republic who (in a history their audience would have known) helped overthrow the Roman kings. *The Federalist Papers* famously argue for a republic, not a democracy. The republican principle affirmed: 'The people are the only legitimate fountain of power,' but when a multitude of people exercise legislative functions in person, the ambitious intrigues of the executive magistrates lead to tyranny. Freedom in a republic does not require that the people themselves exercise legislative authority directly. The power that rests in the people must be filtered. Liberty may require that all power derive from the people and that those entrusted with power remain dependent on the people, but stability requires the opposite. Admitting that the petty 'republics of Greece and Italy' evoke 'sensations of horror and disgust' at the rapid succession of revolutions, the defenders of the constitution claim that it incorporates newly discovered maxims derived from political science that apply 'republican remedies' to the 'diseases most incident to republican government': namely, representation, legislative checks and balances, and the enlargement of the sphere of government. The new republic does not rely only on the virtue of its citizens, but on institutions that protect against oppression, afford liberty for its citizens, and ensure the excellence of those elected to office.

The Anti-Federalists arguing against the new constitution feared that the large centralized government necessary for a populous and heterogeneous nation would oppress its citizens, undermine any attachment they felt for one another, and stifle self-rule. Consent, critical for a free republic, they argued, cannot be represented; it must come from the individual citizen. A 'representative republic' is a contradiction in terms. Lost would be the community of like individuals; citizens would be left to nurture vices of self-interest rather than cultivate republican public virtue. A centralized government would fall into the hands of the wealthy few transforming the new republic into an aristocracy or even a monarchy. The Anti-Federalists emphasizing a small, homogenous polity of virtuous citizens were the last great defenders of the classical republic.

5. Republic Today

Though *The Federalist Papers* distinguish democracies with their demagogues and upheavals from orderly republics, democracy and republic have over the nineteenth and twentieth centuries converged conceptually. Both identify the body of citizens as the fount

of political authority and the overlap between the two regimes has obscured distinctive political structures that might limit recollections of Roman institutions only to republics. With the absence, indeed impossibility, of direct democracy, both democracies and republics have taken on representative forms of government and most democracies have adopted variations of the Polybian mixed regime. Nevertheless, a return to republican ideals of political virtue and engagement has surfaced in an effort to address what is portrayed as the excessive individualism of citizens in liberal democracies. Among contemporary communitarians, ancient republics are symbols of a lost political world where citizens found fulfillment in political life, not in the pursuit of economic success. This evocation of ancient republics has revived discussion about the role of republican ideals during the founding of the USA with communitarians eager to find in the early years republican rather than liberal roots. The concept of the republic to which they appeal goes well beyond the people as the fount of political authority and the mixed regime to the intangibles of political virtue and a shared life of the ancient republic.

See also: Citizen Participation; Civic Culture; Civil Society/Public Sphere, History of the Concept; Democracy; Democracy, History of; Democratic Theory; Participation: Political; Republicanism: Impact on Social Thought; Republicanism: Philosophical Aspects

Bibliography

- Fink Z S 1945 *The Classical Republicans*. Northwestern University Press, Evanston, IL
 Kramnick I 1990 *Republicanism and Bourgeois Radicalism*. Cornell University Press, Ithaca, NY
 Machiavelli N 1998 *The Prince*. [trans. Mansfield H] Chicago University Press, Chicago
 Pangle T L 1988 *The Spirit of Modern Republicanism*. University of Chicago Press, Chicago, IL
 Pocock J G A 1975 *The Machiavellian Moment*. Princeton University Press, Princeton, NJ
 Rahe P A 1992 *Republics Ancient and Modern*. University of North Carolina Press, Chapel Hill, NC
 Wood G S 1969 *The Creation of the American Republic, 1776–1787*. University of North Carolina Press, Chapel Hill, NC

A. W. Saxonhouse

Republican Party

The Republican Party is one of the two major political parties in the United States that have dominated American politics since the mid-1850s. The Repub-

lican party (also known as the Grand Old Party—GOP) emerged amid the sectional conflicts of the pre-Civil War era. Throughout most of its lifetime the party's electoral strength was centered in northern and western states. It was not until the 1960s with the realignment of southern whites toward the Republicans that the party was able to compete effectively in all sections of the country. The Republicans are the right of center party of American politics. It has a broad, multiclass base of support, which tends to be strongest among white, middle, and upper middle-class citizens. Since the 1980s its activist base has become increasingly conservative. Republican national level organizations have become well financed and professionalized and as a result intraparty influence has shifted toward the national organization.

1. Historical Development

The Republican Party emerged amid the conflicts of the 1850s as the Democratic and Whig Parties were unable to reconcile differences within their ranks over the slavery issue and the growing economic and cultural differences between the North and South. The immediate reason for the emergence of the Republican Party was passage of the Kansas–Nebraska Act of 1854, which overturned the ban on slavery in the Northwest Territory. This action by Congress touched off antislavery protest meetings that led to the creation of the party in 1854. In its first election, the Republicans had unusual success winning approximately 40 seats in the House of Representatives (Jones 1965, Moos 1956). In 1856, the Republicans established themselves as the principal opposition party to the then dominant Democrats when the Republican presidential nominee ran a strong second in the popular vote. Abraham Lincoln, formerly a Whig, became the first Republican president after the election of 1860 as the Democratic and Whig parties splintered along a North–South axis. In the Civil War that followed Lincoln's election, the Republicans were the party of national unity and freedom for slaves.

Despite occasional losses, the Republicans were dominant in national politics from the 1860s until the 1930s. It was, however, a sectionally based party of the North and West as the legacy of the Civil War and Reconstruction made it noncompetitive in southern states. The extent of Republican electoral domination during this period is demonstrated by its control of the presidency for 56 of the 72 years between 1860 and 1932. Despite Democratic control of congressional seats in the South, the Republicans also tended to dominate the Congress—controlling the House in 23 sessions and the Senate in 31 sessions out of a total of 36 (Jones 1965).

The Great Depression of the 1930s ended Republican dominance as the party was blamed for economic hard times and the Democrats achieved electoral

hegemony. The GOP's only presidential victories between 1932 and 1964 came in 1952 and 1956 when it nominated Dwight D. Eisenhower, a World War II hero, whose appeal was above party.

2. *The Republican Party in the Post-New Deal Era*

After a disastrous defeat in 1964, Republican fortunes improved although the party continued to have fewer party identifiers among the voters than the Democrats. Benefiting from deep schisms among Democrats over the Vietnam war, civil rights, and social policy, the GOP regained the presidency with Richard Nixon's 1968 election and 1972 reelection. Evidence of the party's continuing minority status can be found in the fact that in these elections it still was unable to win control of Congress and most state governments.

The resurgence of the Republicans after 1964 was based upon fundamental electoral changes as politics became less party-centered and more candidate-centered. There were also significant electoral realignments (Aldrich 1995). Particularly significant was the movement of southern whites toward the Republicans and away from the Democrats. Beck (1999) estimates that white southerners constituted 30 percent of the GOP's base electorate in 1996. With a base of support in southern and western states, the Republicans won three presidential elections in the 1980s and in the 1994 midterm congressional elections gained control of both houses of the Congress for the first time in 40 years.

The transformation of the Republicans from a party which for most of its history drew its electoral support almost exclusively from northern and western states into one with a southern base is demonstrated by the fact that after the 1994 election, the Republicans held a majority of the seats in the House and Senate from the South. Southerners constituted the largest regional group among both House and Senate Republicans. At the same time, Republican representation from the Northeast, once a bastion of party support, declined substantially so that northeasterners were the smallest regional group among House Republicans in the 1990s.

In addition to the shift in the partisanship of white southerners, other changes in the composition of the Republican party's electoral base have occurred since 1968. Whereas women traditionally outnumbered men, by the 1990s they comprised a minority in the party and as a result the party became less sensitive to the concerns of moderate and liberal females. There was also a change in the religious composition of the party. Support among mainline and nonfundamentalist Protestants has eroded steadily since the mid-1960s when they comprised 60 percent of Republican identifiers to slightly more than one-third of party loyalists in 1996. These shifts in the demographic characteristics

of Republican identifiers have been accompanied by an ideological change as the party's loyalists have become increasingly conservative. Beck's (1999) analysis revealed that whereas a majority of Republican identifiers were self-declared conservatives in 1980, two-thirds carried that designation in 1996.

Changes in the composition of the Republican electorate and the heightened conservative ideological commitment of Republican loyalists have had significant implications for the party. Conservative ideological interests have achieved greater influence within the party organization, among elected officials, most notably in the House of Representatives, and over nominations. There is also substantial internal division within the party between committed conservative ideologues, who tend to dominate the many party organizations as well as the House Republicans, and the more pragmatic and moderate Republicans exemplified by the party's governors.

3. *Republican Party Organization*

The party organization is a layered structure of national, state, and local units rather than a hierarchy controlled from the top down. The Republicans' organizational resources, however, involve more than its formal, legally constituted units. There are the thousands of personal campaign organizations created by Republican candidates, an array of partisan campaign consultants and allied interest groups which provide assistance in voter mobilization, money, staff, technical assistance, and media advertising (Bibby 1999).

3.1 *The Republican National Committee*

The Rules of the Republican Party, which are adopted at each quadrennial national convention, call for the creation of a Republican National Committee (RNC) which is charged with responsibility for the 'general management of the party, subject to direction from the national convention.' This authorization makes the RNC the official governing body of the party between conventions. Since its creation in 1856, the RNC has evolved from an organization in which functions were sketchily differentiated and performed by party notables every 4 years into a continuous operation with a large professional staff and budget (Cotter and Bibby 1980).

The RNC resisted reform efforts in the 1970s and has maintained a confederate organizational mode in which each state, plus the District of Columbia, Puerto Rico, American Samoa, Guam, and Virgin Islands have equal representation through a national committeeman and woman and a state chairman. RNC members are selected through procedures mandated by state law and state party rules. Since the full RNC

meets only twice a year, responsibility for operations of the committee rests primarily with the national chairman and senior staff. When the party controls the presidency, the national chairman is normally designated by the President and works in close conjunction with the White House staff.

Since the 1960s, the RNC has developed the capacity to raise massive amounts of money to support party activities and candidates. Thus, in the 1999–2000 election cycle, the RNC raised \$205.3 million, plus \$156.2 million in ‘soft money’ (money raised outside the restrictions of the Federal Election Campaign Act).

With this level of financial resources, the RNC has achieved a degree of autonomy that could not have been imagined in the 1940s and 1950s when it was dependent upon contributions from state affiliates. Through an extensive program of financial and technical assistance to state and local parties, it has gained intraparty leverage because RNC aid is given with requirements attached that the state units must meet. Through this process, Republican state organizations have been professionalized and gained enhanced candidate-support capabilities, but they have also become increasingly dependent upon the RNC for the resources essential for organizational maintenance and effective performance in state level campaigns.

The RNC has also used its financial resources to gain an unprecedented level of intraparty integration in implementing national campaign strategies. For example, in the 1999–2000 election cycle the RNC transferred \$93.0 million in ‘soft money’ to state and local party organizations for spending on general party overhead, voter registration, get-out-the-vote drives, and state and local campaigns. The RNC allocated these funds among the state parties in a manner designed to implement a national strategy geared to winning critical states in the presidential race and maximizing the party’s seats in the House and Senate. Through its programs of technical assistance and fund transfers the national party organization has been able to nationalize campaigns and integrate national and state party campaign strategy to an unprecedented extent (Bibby 1999).

3.2 The Senatorial and Congressional Campaign Committees

In addition to the RNC, the national organization of the party includes two autonomous units which focus their activities upon electing members of the House and Senate. Like the RNC, the National Republican Congressional Committee (NRCC) and the National Republican Senatorial Committee (NRSC) have evolved into well funded organizations with professional staffs. These campaign committees are headed respectively by an incumbent representative and senator. They engage in programs of candidate recruitment, fund raising, technical assistance to candidates,

and media advertising. They also channel and mobilize nonparty money to candidates from political action committees, wealthy individuals, and incumbent members of the House and Senate. Support is doled out to candidates on the basis of electoral criteria—the competitiveness of the constituency—not on the basis of candidate ideology or party loyalty.

Since 1996, the congressional and senatorial committees have placed increasingly heavy emphasis on ‘issue advocacy’ advertising in support of their candidates. These are media advertisements funded with ‘soft money’ that do not expressly advocate the election or defeat of a candidate, but which effectively convey either favorable or unfavorable images of candidates. The campaign committees also engage in independent expenditures which do involve express advocacy of the election or defeat of candidates. Through the use of these techniques the campaign committees have expanded their influence over the national political agenda, helped shape the issue agendas of individual House and Senate campaigns, and attacked their candidates’ opponents (Herrnson and Dwyre 1999).

3.3 State Parties

State level party units have also gone through a process of institutional strengthening since the 1960s. Republican state organizations tend to be stronger than their Democratic counterparts in their ability to engage in party building and candidate support activities (Cotter et al. 1984). As is true at the national level, there is a division of labor among state level organizations. Republican state central committees concentrate on state-wide races, voter mobilization, and party building, while legislative campaign committees focus their resources upon electing Republican state legislators.

4. The Republican Party Network of Issue-oriented Activists

American parties are increasingly networks of issue-based participatory activists. The sources for this trend can be found in a series of complex and interacting forces: the development of a postindustrial society in which noneconomic social/cultural issues have gained heightened saliency; sociological and economic changes that have created higher levels of educational attainment, reduced blue collar employment, and increased number of white collar workers; a decline in the availability of patronage as an incentive to participate in politics. Issue oriented activists have also been aided by changes in the rules governing presidential nominations that have diminished the role of elected officials and party leaders.

In the Republican Party these changes have had the effect of significantly shifting the locus of power toward ideologically conservative activists. In the highly participatory politics of Republican presidential nominations, candidates normally find it essential to stress their conservative bona fides if they are to have any hope of winning the nomination in a process dominated by presidential primaries and participatory party caucuses. Activists taking a leading role in these events as well as national convention delegates have been shown to be substantially to the right of the party's rank and file voters (Mayer 1996). Similarly, Bruce et al. (1991) found in a study of county level campaign workers that ideological 'true believers' outnumbered 'vote maximizers' by a ratio of 4.3:1 among Republicans.

Issue-based groups have gained an increasingly entrenched position within the Republican Party. The Christian Right, for example, has become an increasingly influential and sometimes dominant force within some state party organizations. The increasing involvement of groups and individuals whose motivation is not maximizing the vote but a commitment to policy and ideological concerns has created a problem for the Republican Party. With a conservative activist base that is needed for its campaign support and often to win party nominations, candidates often find themselves catering to the GOP's most conservative elements. However, such actions tend to alienate the more moderate elements of the party and voters without strong partisan commitments—the voters that are needed to win general elections. The party's future depends in large part on how this intraparty problem is resolved.

See also: Campaigning: Political; Congress: United States; Democratic Party; Legislatures: United States; Party Identification; Party Systems; Political Parties, History of; Presidential Government

Bibliography

- Aldrich J H 1995 *Why Parties? The Origin and Transformation of Party Politics in America*. University of Chicago Press, Chicago
- Beck P 1999 The changing American party coalitions. In: Green J C, Shea D M (eds.) *The State of Parties: The Changing Role of Contemporary American Parties*, 3rd edn. Rowman and Littlefield, Lanham, MD
- Bibby J F 1999 Party networks: National-state integration, allied groups, and issue activists. In: Green J C, Shea D M (eds.) *The State of Parties: The Changing Role of Contemporary American Parties*, 3rd edn. Rowman and Littlefield, Lanham, MD
- Bruce J M, Clark J A, Kessel J H 1991 Advocacy politics in presidential parties. *American Political Science Review* 85: 1089–110

- Cotter C P, Bibby J F 1980 Institutional development of parties and the thesis of party decline. *Political Science Quarterly* 95(Spring): 1–28
- Cotter C P, Gibson J L, Bibby J F, Huckshorn R J 1984 *Party Organizations in American Politics*. Praeger, New York
- Herrnson P S, Dwyre D 1999 Party issue advocacy in congressional election campaigns. In: Green J C, Shea D M (eds.) *The State of Parties: The Changing Role of Contemporary American Parties*. J Rowman and Littlefield, Lanham, MD
- Jones C O 1965 *The Republican Party in American Politics*. Macmillan, New York
- Mayer W G (ed.) 1996 *In Pursuit of the White House: How we Choose our Presidential Nominees*. Chatham House, Chatham, NJ
- Moos M 1956 *The Republicans: A History of their Party*. Random House, New York

J. F. Bibby

Republicanism: Impact on Social Thought

In the modern age, republicanism influences social thought by offering a distinctly political articulation of the general belief in the historical contingency of political and social order (Pocock 1975). Modern republicanism holds that political and social order is legitimate only if it serves the constitution of political freedom as nondomination. The possibility of changing forms of rule is understood as a condition of the institution of these forms (Arendt 1963). The existence of ineradicable and productive social divisions is a fundamental presupposition of the republican idea of politics, according to which social conflicts must be given their political expression and regulation (Lefort 1992).

Originating with Machiavelli's critical reception of the tradition of classical republicanism, pursued in the English and Dutch struggles against monarchic absolutism by the likes of Harrington, Sidney, Locke, and Spinoza, and systematized in the Enlightenment by Montesquieu and Rousseau, modern republican theory is directly put into practice in the American and French Revolutions by figures such as Madison, Jefferson, and Sieyès (Pocock 1975, Bock et al. 1990, Fontana 1994). In this sense, modern republicanism has been justly termed a 'revolutionary republicanism' (Wootton 1994).

1. Political Freedom as Nondomination

Modern republicanism describes political freedom as that arrangement in which 'the rule of law, resting on the power of the people, would put an end to the rule of man over man' (Arendt 1972 p. 139). Politics should be that domain of human affairs where domination

and rule can be questioned, checked, and even possibly abolished (Pettit 1997).

The modern republican ideal of freedom as non-domination or no-rule moves beyond both the negative and positive ideas of freedom (Berlin 1965). Negative freedom is an individual and private freedom from interference that requires a noncontesting relation to the commands of the political sovereign that protect and enable it from the outside. In contrast, the ideal of nondomination is political, not private, and consists of being 'subject to no man's will' (Sidney 1798, III, 16) by having the power to contest both arbitrary and legitimate domination. Freedom as nondomination also differs from ideals of self-rule or positive freedom insofar as these view some forms of rule (e.g., of the community over its members, of reason over passion) as intrinsic to the meaning of freedom, whereas from the perspective of freedom as nondomination, political activity is primarily understood in terms of its capacity to check the dynamics of domination and processes of ruling by empowering those who would otherwise fall into relations of dependence. Political freedom 'is there only when there is no abuse of power.... To prevent this abuse, it is necessary from the very nature of things that power should be a check to power' (Montesquieu 1748, XI, 4).

Modern republicanism sees the political abolition of domination as achievable through a combination of the 'rule of law' and the 'power of the people.' Although both of these elements can be found in classical republicanism as well, their meaning and relative priority undergo important changes in modern republicanism where 'all power is vested in, and consequently derived from, the People' (*Virginia Declaration of Rights* 1776, article 2) because its members are understood as the carriers of the ideal of non-domination: 'all men are by nature equally free and independent, and have certain inherent rights of which, when they enter into a state of society, they cannot, by any compact, deprive or divest their posterity' (*Virginia Declaration of Rights* 1776, article 1). As a consequence, the republican ideal requires that the rule of law find its support in the power of the people.

2. The Power of the People

The connection between the power of the people and the ideal of nondomination can be traced back to Machiavelli in his *Discourses on Livy* where, for the first time, 'the people' is defined as that subject position occupied by whoever acts in political ways that are motivated 'only [by the] desire not to be dominated' (Machiavelli 1517, I, 5).

The 'power' that is vested in 'the people' refers to those political practices and institutions of will and

opinion formation that integrate this political subject as bearer of the desire for nondomination into political life, alongside the established practices and institutions of legitimate domination, of which traditionally the most important elements were the deliberations of the senate and the commands of the monarch (Machiavelli 1517, I, 4). Harrington does away with the monarchical element, and takes into account only the nobles, represented by the senate, and the people, in order to give ultimate say to the latter: 'as the wisdom of the commonwealth is in the aristocracy, so the interest of the commonwealth is in the whole body of the people.... Whatsoever upon debate of the senate is proposed unto the people, and resolved by them, is enacted by the authority of the fathers and the power of the people, which concurring make a law' (Harrington 1656 p. 24). As Rousseau frames the issue, 'it is only the general will which compels individuals and there can be no assurance that an individual will is in conformity with the general will until it has submitted to the free suffrage of the people' (Rousseau 1762, II, 7). It is an axiom of modern republicanism that without this kind of political power vested in the people there can be no political freedom.

Likewise, since the people is conceived as a political subject whose foremost interest is nondomination or no-rule, it follows that it should not, as a people, engage directly in the process of governing or ruling on pain of self-contradiction. For this reason Madison is perfectly consequent when he claims, in the same article of *The Federalist*, that 'the people can never willfully betray their own interests,' and that the distinguishing feature of 'American government lies in the total exclusion of the people in their collective capacity from any share in the latter' (Hamilton et al. 1787, p. 63). Modern republicanism requires the principle of representation because it understands political representation not as a condition for the pursuit of more effective ways of governing a large population, but as the primary way to limit the powers of the state as an agency of legitimate domination.

3. Politics and the Primacy of Social Division

The analysis of society in terms of the conflict between heterogeneous social subjects and its causality in the emergence of political forms is a fundamental contribution of modern republicanism to social thought. Machiavelli shows that the opposition between those who 'desire not to be commanded or oppressed' and those who 'desire to command and oppress' constitutes a basic social division that, depending on how it is politically interpreted, gives rise to 'one of three effects in cities: principality or freedom or license' (Machiavelli 1513, IX). Similarly, Harrington observes that every state finds its basis in the social division of property. For that reason, he argues that an 'agrarian law' must be set at the

'foundation' of the republic. In so doing, Harrington both acknowledges the fundamental role of social conflict and seeks to regulate it by giving it a political form (Harrington 1656 p. 33). Modern republicanism holds that social divisions provide the basis of politics, and that politics has the task of giving or taking away form (laws and orders) in response to such conflicts, thereby expressing and regulating rather than extirpating or denying them (Pocock 1975, Lefort 1992).

After the American and French Revolutions, and well into the twentieth century, this relative autonomy of politics with respect to 'the social' comes under criticism thanks to the emergence of the idea of society as a form of organization that, in and through its divisions, is self-constituting and self-regulating. This idea of society opens the prospect of the abolition of politics and of the state, under the appropriate social conditions. Much of the impetus for the late twentieth century retrieval of the republican tradition comes from the need to recover a positive conception of politics and, along with it, a sense of the productive tension between state and society.

4. Republicanism and Institutionalization

For modern republicanism, politics consists of the practice of giving an institutional expression, or interpretation, to the divisions that constitute social relations. In particular, republican politics sets out to refashion the state so as to be able to offer institutional responses to the demands for nondomination issuing from society. The state, viewed as a form of legitimate domination, can only respond to these demands by internalizing its own negation and becoming internally self-checking. This principle underlies the recurrent attacks that modern republicans wage on the very idea of a 'standing power,' against which they introduce the need for 'motion' or 'the successive revolution of authority' (Nedham 1659 *Mercurius Politicus* 4 December 1651–1 April 1652). It also accounts for the constant thematization of political change and social conflict in the very constitution of political form. The debate between Jefferson and Madison on the uses of constitutional conventions to check the power of the state is exemplary in this sense: the latter believes in designing a political form of government capable of checking itself, while the former holds that a check on government ultimately can only come from what lies outside it (Hamilton et al. 1787, pp. 49–51).

The innovation of modern republicanism with respect to thinking about institutions consists of providing both theories and examples of what may be called institutions of critique, or counterinstitutions, which take the form of inscribing a resistance to domination into the state apparatus that administers the rule of law. The most significant example of such an institution of critique is the separation of state powers

when the latter is understood as a *system of checks and balances* (Hamilton et al. 1787, p. 51) whereby 'one branch of the government may be authorized to exercise some active influence on another in order to resist and counteract its power' (Manin 1994, p. 31) in view of keeping the entire state in check with respect to the people.

Modern republicanism offers other examples of such institutions, whose sole purpose is to serve as a forum for the contestation and resistance of the normal administration of state rule. Among them stands out the Tribune of the people, interpreted by Machiavelli as a 'guard of freedom' (1517, I, 3–4), which, in the words of Rousseau, 'although it can do nothing, it can prevent anything from being done. As the defender of the laws, it is more sacred and more venerated than the prince who executes laws or the sovereign which legislates' (Rousseau 1762, IV, 5).

Another example is the Agrarian Law as discussed in Machiavelli (1517), Harrington (1656), and Hutcheson's *System of Moral Philosophy* (1755), where it serves to check the unregulated distribution of private property. Among such institutions also figures Jefferson's 'republic of wards,' designed to serve as 'a system of fundamental balances and checks for the government' (Kurland and Lerner 1987, p. 142). The same principle animating the ward system is embodied by the 'council systems' (Arendt 1963) that emerged during the republican revolutions of France in 1871, Russia in 1905 and 1917, Germany in 1918, Hungary in 1956, and Poland in 1979.

5. Republicanism and Revolution

Modern republicanism also recognizes the possibility of an external critique of the state-form or revolution. In political revolutions, it is the people that attempt to institute nondomination directly, and it can do so only by suspending or dissolving the state form as such. Although prefigured by Machiavelli with his theory that a free political life can be maintained only if there are periodic 'return to beginnings' in which the social and political orders are renewed (Machiavelli 1517, III, 1), one of the firmest formulations of the state-dissolving power of the people is found in Locke, for whom 'the community may be said in this respect to be always the supreme power, but not as considered under any form of government, because this power of the people can never take place till the government be dissolved' (Locke 1690, XIII, 149). Likewise, Rousseau insists that 'the moment the people is lawfully assembled as a sovereign body all jurisdiction of the government ceases' (Rousseau 1762, III, 14). Hamilton reaffirms in *The Federalist* 'that fundamental principle of republican government which admits the right of the people to alter or abolish the established Constitution whenever they find it in-

consistent with their happiness' (Hamilton et al. 1787, p. 78).

It is important to note that for modern republicanism, unlike other late modern revolutionary traditions, political revolutions are recurring events out of which all forms of legitimate domination emerge and into which they can all be revoked. By definition these revolutionary events cannot be subsumed into or become in turn a political form without denying themselves. Republican moments are in themselves evanescent, having as their fundamental purpose that 'men who live together in any order whatever [be able to] often recognize themselves...and look over the account for the men who are in that body' (Machiavelli 1517, III, 1). The purpose is the mutual recognition of the liberty and equality of all.

6. Political Equality and the Rule of Law

The 'rule of law' is a feature of both classical and modern republicanism. But whereas classical rule of law is substantive and intends to prescribe the common good, modern rule of law is formal and has as its essential purpose 'the restriction of each individual's freedom so that it harmonizes with the freedom of every one else' (Kant 1793, p. 73). As evidenced in the American and French Declarations of rights, the rule of law is always limited by, because it is founded upon, two elements that are absent in classical republicanism: individual natural rights and sovereign power of the people.

Following the trajectory that leads from Machiavelli through Rousseau up to the American and French Declarations, some contemporary republican theorists hold the view that unless the rule of law is grounded on the power of the people, it cannot guarantee a political life free of domination (Arendt 1963, Habermas 1992). Other republican theorists, supporting themselves on Montesquieu's definition of liberty as 'the right of doing whatever the laws permit' (Montesquieu 1748, XI, 3), believe that the relation between the rule of law and the power of the people should be reversed because it is 'the laws of a suitable state...[that] create the freedom enjoyed by citizens' (Pettit 1997 p. 36).

At stake in this discussion is the interpretation of a basic republican principle of equality before the law. On one understanding, freedom as nondomination basically requires 'an empire of laws' (Harrington 1656, p. 20) assuring that no one is above the law and thereby excluding arbitrary rule. On the other understanding, freedom as nondomination also requires equality in the making of the laws before which everyone is formally equal. So, for instance, 'No Man can be imprisoned, unless he has transgressed a Law of his own making' (Trenchard 1697, p. 2). Rousseau simply says that 'a people, since it is subject to laws, ought to be the author of them' (Rousseau 1762, pp. II, 6).

Without the equality to make law, the rule of law harbors the possibility of paternalism and legal domination (Ferry and Renaut 1985, Habermas 1992).

To be politically free is to have the power to 'make' law in the sense of having the power to constitute, and therefore check the exercise of, any constituted legislative power (Sieyès 1789). Such constituent power cannot itself fall under the rule of law but rather grants it its ultimate legitimacy ground, as is made clear not only in Sieyès, but also in Rousseau and Locke. For both, the power to make law belongs to the people originally, which is constituted by a contract that everyone makes (with everyone else or with themselves) (Locke 1690, XI, 134, Rousseau 1762, I, 6). Locke then permits the transfer of the legislative power of the people to the government (X, XI); but, as an element of the state, the rule of law is always checked by the deciding power of the people (XIII, 149; XIV, 168; XIX, 240).

The question of constituent power as it is developed in modern republicanism has an impact on social thought because it highlights the existence of political but prelegal struggles for recognition (Taylor 1994), through which individuals and groups seek the constituent power of citizenship that gives access both to political and individual rights.

7. Representation and Participation

For modern republicanism, political representation is a response to Montesquieu's call for a separation of legislation and execution 'lest the same monarch or senate should enact tyrannical laws, to execute them in a tyrannical manner' (Montesquieu 1748, XI, 6). Such a separation of powers is echoed as much by Rousseau (1762 III, 1, III, 16) as by Kant, who defines republicanism as that 'political principle whereby the executive power (the government) is separated from the legislative power' (Kant 1795, p. 101) vested in the people as 'co-legislators' (Kant 1798, p. 184).

The modern republican discourse on political participation follows, rather than contradicts, Montesquieu's doctrine of the separation of powers. Since a republic entails the division between those who have the power to make laws and those who have the power to apply it, it follows that the activity of governing or the administration of the constituted powers of the state must be given over to representatives, and in that sense all government must be representative in order to be legitimate. But the generation of constituent power has to be kept in the hands of the people, and to that extent calls forth their participation, without which no such power would come into existence in the first place. The debate between Locke and Rousseau on whether the fundamental power of the people to make law can or cannot be represented (Locke 1690, X, 132, XIII, 149,

Rousseau 1762, III, 15) is resolved in the form of the distinction between a constituent legislative power, which is what Rousseau and Kant have in mind when they advocate its inalienability from the people, and a constituted legislative power, which is what Locke, Sieyès, and Madison have in mind when they argue for its representative character. In both cases, the goal is to assure Montesquieu's requirement of contrasting government by setting power (of the people) against power (of the state).

8. Political Virtue and Corruption

The discourse of 'virtue' in modern republicanism (Pocock 1975) for the most part refers to a political conception of virtue, rather than to the classical moral conception of virtue as the pursuit of the common good (Machiavelli 1513, Montesquieu 1748). The debate about political virtue is a place-holder for the necessity of a certain kind of 'government by men' in order to maintain the ideal of nondomination in the face of the potential for 'corruption' (Skinner 1998) harbored in the modern rule of law and other systems of negative liberty, of which the most significant for modern republicans is the system of commerce and money. Since the modern rule of law and modern commerce are not viewed as being morally substantive, but rather as protecting and fostering the negative liberty needed for all nonpolitical forms of self-enactment (Wootton 1994), simple adherence to them cannot replace the political activity needed to sustain both the participatory and the representative requirements of republicanism. Left to their own devices these systems of negative liberty could degenerate into forms of bureaucracy (no-man's rule), as well as deepen social and economic inequalities, thereby threatening freedom as nondomination.

Modern republicans speak of two kinds of political virtues, depending on whether they take virtue to be synonymous with the set of characteristics that make political subjects have 'power' or 'authority,' and depending on whether they are more interested in the 'freedom' or in the 'stability' of the republican form. Harrington and Madison concentrate on the need for an 'aristocratic' virtue, to be present in the representatives, for the sake of the 'wisest' and most 'authoritative' outcome in their deliberations, which is in turn the main guarantee of the stability of the republican form. By contrast, Machiavelli, Jefferson, and other proponents of 'public happiness' offer a distinctly more 'democratic' conception of political virtue, designed to foster in the people that 'power' which comes from being, in Jefferson's felicitous phrase, 'a participator in the government of affairs' (Kurland and Lerner 1987, p. 142). As a result, the rule of law requires virtuous individuals, and there is no virtue without the power to do 'strong' and perhaps

'uncivil' actions (Machiavelli 1517, I, 18, III, 1) of the kind required by engagement in practices of active nondomination, of empowerment and resistance against practices of domination and dependency.

See also: Cleavages: Political; Corruption: Political and Public Aspects; Equality and Inequality: Legal Aspects; Equality: Philosophical Aspects; Freedom/Liberty: Impact on the Social Sciences; Freedom: Political; Institutionalization; Machiavelli, Niccolò (1469–1527); Montesquieu, Charles, the Second Baron of (1689–1755); Political Representation; Power in Society; Power: Political; Republicanism: Philosophical Aspects; Revolutions, History of; Revolutions, Theories of; Rousseau, Jean-Jacques (1712–78); Rule of Law

Bibliography

- Arendt H 1963 *On Revolution*. Penguin Books, New York
- Arendt H 1972 *Crises of the Republic; Lying in Politics, and Disobedience on Violence, Thoughts on Politics, and Revolution*. Harcourt Brace, New York
- Berlin I Sir 1965 *Four Essays on Liberty*. Oxford University Press, Oxford, UK
- Bock G, Viroli M, Skinner Q (eds.) 1990 *Machiavelli and Republicanism*. Cambridge University Press, Cambridge, UK
- Ferry L, Renaut A 1985 *Philosophie Politique, vol. 3: Des Droits de l'homme à l'idée républicaine*. Presses Universitaires de France, Paris [1992 *Political Philosophy 3. From the Rights of Man to the Republican Idea*. University of Chicago Press, Chicago]
- Fontana B (ed.) 1994 *The Invention of the Modern Republic*. Cambridge University Press, Cambridge, UK
- Habermas J 1992 *Faktizität und Geltung: Beiträge zur Diskurstheorie des Rechts und des demokratischen Rechtsstaats*. Suhrkamp Verlag, Frankfurt am Main [1996 *Between Facts and Norms: Contributions to a Discourse Theory of Law and Democracy*. MIT, Cambridge, MA]
- Hamilton A, Jay J, Madison J (eds.) 1987 [1787–8]. *The Federalist*. In: Kramnick I (ed.) *The Federalist Papers*. Penguin, Harmondsworth, UK
- Harrington J 1992 [1656] *The Commonwealth of Oceana*. Pocock J G A (ed.), Cambridge University Press, Cambridge, UK
- Hutcheson F 2000 [1755] *A System of Moral Philosophy*. Thoemmes, Bristol, UK
- Kant I 1793 *Über den Gemeinspruch: 'Das mag in der Theorie richtig sein, taugt aber nicht für die Praxis'* [1991 Reiss H (ed.), Nisbet H B (trans.) *Kant: Political Writings*. Cambridge University Press, Cambridge, UK]
- Kant I 1795 *Zum ewigen Frieden. Ein philosophischer Entwurf* [1991 Reiss, H (ed.) *Kant: Political Writings*. Cambridge University Press, Cambridge, UK]
- Kant I 1798 *Der Streit der Facultäten* [1991 Reiss H (ed.) *Kant: Political Writings*. Cambridge University Press, Cambridge, UK]
- Kurland P B, Lerner R (eds.) 1987 *The Founders' Constitution*. University of Chicago Press, Chicago

- Lefort C 1992 *Écrire. A l'épreuve du politique*. Calmann-Levy, Paris
- Locke J 1980 [1690] *Second Treatise of Government*. Macpherson C B (ed.), Hackett, Indianapolis, IN
- Machiavelli N 1996 [1517] *Discourses on Livy*. University of Chicago Press, Chicago
- Machiavelli N 1998 [1513] *The Prince*. University of Chicago Press, Chicago
- Manin B 1994 Checks, balances and boundaries: the separation of powers in the constitutional debate of 1787. In: Fontana B (ed.) *The Invention of the Modern Republic*. Cambridge University Press, New York
- de Secondat Montesquieu C 1989 [1748] *The Spirit of the Laws*. Cohler A M, Miller B C, Ston H S (trans. and eds.). Cambridge University Press, New York
- Nedham M 1659 *Mercurius Politicus*. Cornmarket Press, London
- Pettit P 1997 *Republicanism: A Theory of Freedom and Government*. Oxford University Press, Oxford, UK
- Pocock J G A 1975 *The Machiavellian Moment. Florentine Political Thought and the Atlantic Republican Tradition*. Princeton University Press, Princeton, NJ
- Rousseau J-J 1968 [1762] *The Social Contract*. Penguin, London
- Sidney A 1990 [1798] *Discourses Concerning Government*. West T G (ed.), Liberty Fund, Indianapolis, IN
- Sieyès E J 1970 [1789] *Qu'est-ce que le tiers état?* Droz, Geneva
- Skinner Q 1998 *Liberty before Liberalism*. Cambridge University Press, Cambridge, UK
- Taylor C (ed.) 1994 *Multiculturalism: Examining the Politics of Recognition*. Princeton University Press, Princeton, NJ
- Trenchard J 1697 *An Argument, Shewing, that a Standing Army is Inconsistent with a Free Government, and Absolutely Destructive to the English Monarchy*. London
- Virginia Declaration of Rights 1987 [1776]. In: Kurland P B, Lerner R (eds.) *The Founders' Constitution*. University of Chicago Press, Chicago
- Wootton D (ed.) 1994 *Republicanism, Liberty, and Commercial Society, 1649–1776*. Stanford University Press, Stanford, CA

M. E. Vatter

Republicanism: Philosophical Aspects

Republicanism is the doctrine that laws and the state should always serve the common good (or '*res publica*') of a nation's people or citizens ('*populus*'). The basic desiderata of republican government, as articulated in the republican tradition derived from Rome, secure government for the common good through the checks and balances of a mixed constitution, comprising a sovereign people, an elected executive, a deliberative senate, and a regulated popular assembly, constrained by an independent judiciary, and subject to the rule of law (Sellers 1994). Some republicans would add representation, the separation of powers, or other new structural devices to protect public liberty ('*libertas*') and avoid Rome's eventual descent into popular tyranny and military despotism (Sellers 1998). Republican liberty signifies

subjection to the law and magistrates, acting for the common good, and never to the private will or domination ('*dominatio*') of any private master (Pettit 1997).

1. The Republican Tradition

'*Res publica*' was the Romans' term for their state, its public business, all public property, and the purposes these served. The word notoriously evades translation, most often appearing in English as 'commonwealth,' or simply (more recently) 'republic,' the use preferred here. Marcus Tullius Cicero (BC 106–43) and Titus Livius (Livy) (BC 59–AD 17) constructed the first and most influential comprehensively republican ideology, to try to explain how and why the Roman republic had failed to serve the public good. Both agreed that republican institutions collapsed when military expansion and party conflict upset traditional checks and balances between the senate, the magistrates, and the people of Rome.

Cicero and Livy inaugurated a republican tradition of 'liberty' that fortified principled resistance to demagogues, emperors, and kings for the next two thousand years. Niccolò Machiavelli (1469–1527) did the most to revive this republican tradition in Italy, in his *Discorsi sopra la prima deca di Tito Livio* (1517). The resistance of the Swiss, various Italian cantons, and the United Provinces of the Netherlands to imperial control added practical models for republican liberty, as did the constitutional and theoretical writings of various English authors, in their efforts to restrain or remove their kings during the Civil War and Commonwealth (1642–60), the Glorious Revolution (1688), and the extended British controversies over American Independence (1763–83).

2. The Republican Revival

Republican political and legal philosophy provides an enduring model for philosophers and lawyers who oppose the domination of arbitrary power. The constitutional controversies of the French and American revolutions left institutional relics, which established republican ideals in the structures of Western politics, with a pervasive commitment to 'liberty' and even (in the United States), a federal 'guarantee' that every state in the union must enjoy 'a republican form of government' (*United States Constitution*, 1787 Art. IV § 4). Similar provisions requiring senates and independent judges give republican theorists a practical foundation in the constitutions of most Western democracies. Lawyers such as Frank Michelman (1936–) and Cass Sunstein (1954–) first revived republican doctrine in law schools, to justify judicial intervention for the common good, against partisan legislation and political corruption (*Symposium* 1988).

Philip Pettit inaugurated a similar revival among philosophers, embracing three central claims of the republican tradition: that government must serve the common good; that certain institutional arrangements will do so; and that liberty will be the result (Pettit 1997). Republicanism since Cicero has offered a political epistemology for finding the common good, or approximating it as closely as possible. The republican question has not been (in the first instance): ‘What is the common good?’, but rather: ‘Which political procedures will best find and protect the common good, given the possibilities and limitations of history and human nature?’ (Sellers 1991).

3. Liberty as Nondomination

Republican liberty consists in not being interfered with by fellow citizens or the state, except as regulated by the common good of the people, established by republican politics. This conception of liberty as ‘nondomination’ contradicts a widely-held (but more recent) understanding of liberty as not being interfered with at all. Public and private actions that constrain or influence a person’s activities do not restrict republican liberty unless they violate the law or republican moral principles established to serve the common good of the people as a whole. Republican liberty cannot exist without a republican state because nonrepublics will always dominate their subjects, as masters dominate their slaves.

Domination marks the end of liberty, in republican theory, because domination subjects one person to the arbitrary will of another, without regard to the common good. The unchecked power of the *dominus* (or master) makes his subordinates slaves, whether the master chooses to exercise power or not. Slaves to good masters remain slaves, and dominated, because they have no security against their master’s will. States dominate citizens when their constitutions give legislatures or public officials arbitrary power to act, without regard to the public welfare. Private individuals dominate others when laws and public officials do not protect citizens against private oppression.

The common good is the measure of domination, and therefore of liberty, in republican political vocabulary, because only the common good distinguishes ‘arbitrary’ interference from justified state action. *Libertas* (liberty) in Rome signified the status of a free citizen in a republican state. The *res publica* (republic) represented every citizens’ common interest in the public good. Restraining citizens equally with just laws in pursuit of the common good enhances their liberty by restraining arbitrary power. Neo-republican doctrine revives liberty as a worthy object of public policy by rescuing the sense in which the word first developed as a political ideal. Liberty is not, as Thomas Hobbes (1588–1679) maliciously redefined it, the unfettered license to do what one wants, but

rather a public shield against uncontrolled license in others. Republican liberty consists in equal subjection to the rule of laws, made for the common good, through the carefully balanced mechanism of a republican form of government. For further references see *Freedom: Political*.

4. The Common Good

The fundamental republican insight that government and laws should serve the common good did not originate in Rome. Most governments have made this claim, which Cicero (Tullius Cicero *Off.* I.xv.85) attributed to Plato (circa BC 427–347). Cicero defined the purpose of the state as seeking to create harmony from the disparate interests of all members of society (*Off.* III.vi.26). This does not require that all public interests always serve every individual’s separate desires, but that laws and government exist to make every citizen’s life worthwhile. Most human goods have a social component, making liberty essentially communal in character.

The first difficulty republicans face in finding the common good lies in defining their community. Ancient republics closely restricted their citizenship, excluding women, immigrants, and slaves. Most republicans would now extend citizenship and membership in the *populus* or people to all inhabitants of a given territory. This leaves open the question of boundaries. Cicero endorsed a universal republic or society of humanity, but also recognized the value of local politics. Cultures and human relationships develop and solidify locally. Therefore, the republican interest in harmony and community implies small homogenous republics, where people will be very much the same (Montesquieu 1748). But the natural diversity of human talents and interests guarantees that small communities will find their own internal minorities to dominate and oppress. Larger republics with greater diversity better protect minorities, by making dominant factions harder to assemble (Hamilton et al. 1787–8, p. X).

The republican solution to republican boundaries has been to encourage different levels of nested republics, for different purposes, with different capacities. Immanuel Kant (1724–1804) advocated a federation of national republics, assembled in perpetual peace (Kant 1795). The formula suggested by the United States and modern European Union has been to entrust the protection of individual human rights and commerce, which local majorities might threaten, to federal suprarepublican authorities, while delegating social and cultural concerns to smaller, more homogenous republics. When regional authorities consistently abuse their powers to oppress local minorities, their oppression implicitly recognizes such minorities as separate people, entitled to separate republican homelands, within the republican federa-

tion (Sellers 1996). For further references see *Public Interest*.

5. Popular Sovereignty

Republicans realized very early in Rome, and have maintained the principle ever since, that the common good will never be found or maintained without the *imperium populi* or popular sovereignty to support it (Tullius Cicero *Cat.* IV.14; *Phil.* IV.iv.8). Macchiavelli credited Cicero with the insight that although the people may be ignorant, they are capable of grasping truth, when good men place the truth before them (Macchiavelli 1517 I.iv.10). The purpose of republican popular sovereignty is not that the people should govern every day, but rather, as Benjamin Rush (1745–1813) suggested in advocating an American ‘republic,’ that the people should select their rulers themselves. The people exercise their power on election days, then defer to magistrates that they themselves have selected (Rush 1777).

This republican reliance on popular sovereignty distinguishes republican government from liberalism, which diverged from the republican tradition at the beginning of the nineteenth century (Sellers 1998). Early liberals rejected popular sovereignty in the wake of the French revolution, while retaining republican commitments to liberty, the rule of law, and certain specific rights against of government (Constant 1819). Liberals divorced liberty from its basis in public deliberation by distinguishing ‘political liberty’ from personal independence. This left liberals dependent on judges and magistrates to verify their assertions of rights, and protect individual liberty against the state. Liberalism started as a flight from politics, but found that this implied a new definition of liberty as the ability to do what one wants, without state regulation (Pettit 1997).

Contemporary liberalism has largely abandoned the fear of democracy that separated liberalism from republicanism in the first place, and most liberals now endorse the widespread political participation of citizens in elections (Rawls 1993). Republican popular sovereignty never signified democracy in any case, but rather a carefully balanced and controlled mechanism for gathering the insights of all members of society, through a process of public deliberation, in pursuit of the common good. As modern liberals have sought a method of democratic deliberation to legitimate their sense of community, and regulate conflicting private interests, they have increasingly returned to republican principles and institutions, embedded in Western constitutional democracy. For further references see *Sovereignty: Political*.

6. Checks and Balances

The primary, process-centered challenge for republicans has been to identify those institutions that

find and secure the common good best. Popular sovereignty provides access to every citizen’s interests and insights, but democracy would become an elective despotism, without checks and balances to restrain it (Pettit 1999). Cicero blamed the Greeks’ misfortunes on the turbulence of their popular assemblies, lacking internal balance or a senate to control them (Tullius Cicero *Flacc.* 1976). Republics since Rome have maintained bicameral legislatures to prevent self-seeking in either assembly (Harrington 1656, p. 22). Democracy is one of several foundation stones of liberty, not an end in itself.

The necessary checks and balances of republican government prevent public officials from making themselves and not the *res publica* the object the state (Paine 1792, p. 168). The dispersion of power through bicameralism, federalism, and the separation of powers makes it harder for any one person or faction, including the majority, to wield arbitrary power over others (Pettit 1997, p. 177). American republicans, such as John Adams (1787–8, p. I.ii–iii) and Alexander Hamilton et al. (1787–8, IX) added representation and the life tenure of judges to bicameralism and the checks on governmental powers as basic desiderata of balanced republican institutions. If the people and judges are outside government, they can better control their government’s mistakes (Hamilton et al. 1787–8, p. LXIII).

Modern republicans such as Philip Pettit differ from democrats in viewing democracy as a derivative value, in service to the broader ideal of balanced or ‘contested’ government (Pettit 1997, p. 187). All government decisions should be subject to challenge by institutions that prevent private inclinations from ruling public interests and ideas. Republicans have proposed bills of rights, public hearings, and many other devices designed to constrain and channel public decision making, so that ordinary citizens may influence their government’s decisions, without diverting the public purposes of the state. For further references see *Constitutionalism*.

7. The Rule of Law

Republican government consists of whatever institutions ‘will compel the formation of good and equal laws, an impartial execution, and faithful interpretation of them, so that citizens may constantly enjoy the benefit of them, and be sure of their continuance’ (Adams 1787–8, p. I.128). Such institutions secure the ‘*imperia legum*’ or ‘empire of laws and not of men’ praised by all republican authors from Livy to Pettit. Laws secure liberty against domination by delimiting and protecting the line across which private or state behavior violates individual autonomy against the common good. Cicero insisted that such laws must serve the public welfare (*populi utilitas*) not the public will (*populi voluntas*) (Tullius Cicero 1906 [*Sulla* 25])

because votes cannot alter the natural laws of justice (Tullius Cicero *Leg. I.xvi. 44*). Republics need institutions that will find justice by securing the common good.

The empire-of-law condition of republican government entails that laws should be publicly promulgated, intelligible, consistent, stable, general, and apply to everyone, including the legislators themselves. This prevents arbitrary decisions by public officials, by subjecting their will to known constraints and purposes. Republics maintain a general presumption that government action, when needed, will operate by law, not by *ad hoc* or *ex post* decisions. All agencies of government must act through principled, regulated structures, maintaining the 'due process' of laws to prevent the abuse of governmental power (Pettit 1997, pp. 174–5).

Codifying the law in every detail will not be possible or desirable. Some discretion must remain, but subject to the dispersion-of-power and popular sovereignty procedures that generate republican laws in the first place. Discretion subject to constraints against arbitrariness may sometimes secure the common good, but unlimited discretion produces mistakes about justice, through the natural partiality and limited viewpoint of any individual decision-maker. The republican rule of law protects liberty by respecting the welfare of every member of society, pooling their insights to secure the common good of the people. Legislation is the public reason of the republic, protecting citizens' liberty against private power (Harrington 1656, pp. 19–20).

8. Neorepublican Ideals

Contemporary philosophical republicanism reasserts the republican conception of liberty, proposing new republican institutions to support republican liberty in the modern multicultural state. Philip Pettit has given the most detailed proposal for neorepublican politics, supplanting the Benthamite liberal opposition of 'negative' and 'positive' liberty with older republican conceptions of liberty as nondomination, and protection against arbitrary power (Pettit 1997). Pettit understands arbitrary power ('*arbitrium*') to include any action against the 'interests and ideas' of those who suffer interference. This surpasses the old republican standard, which measured arbitrary power by its violation of an objective public, common, or collective good.

Neorepublican deference to private interests and ideas reflects the residual liberal commitment to moral pluralism and cultural diversity. This enduring hesitation to overrule anyone's personal ideas makes the revised standard of nondomination very strong. Interference or influence over others' choices becomes domination unless it tracks their private values. This

introduces a subjective element into republican doctrine. Older republican ideals tolerated hierarchy and private influences when toleration served the common good of all citizens.

Neorepublican rhetoric describes the state as seeking to 'promote' freedom as nondomination, where older republicans would have understood republican pursuit of the common good simply to be freedom as nondomination. The difference lies in the pluralistic language of the newer theories. Some suggest that immigrant groups and minority cultures will maintain their separate identities in republics, because republics help to preserve hereditary differences within the larger culture (Pettit 1997, p. 144). This represents a significant departure from traditional doctrine. Republics exist precisely to create and maintain an overarching culture that embraces all citizens. This necessarily influences and discourages dissenting subcultures by promoting an image of national unity, which will tend to assimilate minority perspectives.

9. Neorepublican Institutions

The neorepublican commitment to pluralism encourages a greater emphasis on countermajoritarian aspects of republican popular sovereignty (Pettit 1997, *Symposium* 1988). Neorepublicans embrace the empire-of-law condition of republican government and the dispersion-of-power constraint, but they often seek to displace public decision-making onto independent panels of social scientists or courts (Pettit 1997, Sunstein 1993, *Symposium* 1988). This favors 'contestability' over consensus to limit the weight of unified public opinion.

Neorepublicans do not endorse interest-group pluralism in the classical amoral sense of liberal democracy, but their concept of representation favors cultural minorities. Some believe that public policy should be able to be justified according to the lights of every member of society (Pettit 1997, p. 169). This overstates the value of mistaken perceptions of the good. Republican government differs from liberal democracy precisely in that it offers a deliberative technique for overruling mistaken views, while allowing constructive dissent by those who oppose the consensus.

Traditional republics maintained geographical representation to encourage the development of local communities. Most recent proposals prefer religious, cultural, or sexual diversity. Popular elections now seem too crude to identify administrative competence, or to guarantee minimum statistical representation for every stakeholder grouping (Pettit 1997, p. 192). Neorepublicans often propose deliberation among statistically selected 'representative' officials, in place of elected representatives or public assemblies. This leads to difficulties in identifying which sub-

cultures or stakeholder 'groups' to represent. Most people have many such partial or self-defined sources of identity.

10. The People

The fundamental republican conception of the common good encourages social solidarity by seeking a community of interest among the *populus* or people of any given republic. Cicero described republics as the property of a people (*res populi*), in the most influential and frequently repeated passage in the republican canon. A *populus* is not just any collection of humans, but a partnership about justice, in pursuit of the common good (Tullius Cicero *Rep.* I.xxv.39). There will be no republic without a people and no people (in the fullest sense) without a republic, which is to say the mutual recognition of common citizenship (fraternity) and the equal importance (equality) of all citizens in the eyes of the state.

The concept of peoples represents the most contested element in contemporary republican doctrine, and the most likely source of disagreement between neorepublican philosophers. The United Nations Charter recognized the republican principle of self-determination of peoples (Article I (2)), implying a people's right 'freely [to] determine their political status' (International Convention on Civil and Political Rights 1966, Article 1). This gives the republican identity a practical importance in international politics, implying that recognized status as a 'people' entails some degree of political separation. The salient examples of liberated 'peoples' have been the inhabitants of colonial territories, who constitute peoples in international law for the purpose of pursuing their independence from imperial domination.

Some neorepublicans would separate the national and territorial principles to create politically separate republican peoples on the basis of ethnic, religious, or linguistic affinities within larger multinational empires. Different laws would apply to different subjects of the overarching state, according to their different internal 'citizenship' or status. This violates the republican principle of equal citizenship to achieve the republican ambition of social solidarity among state-maintained subgroups of the population. Traditional republican doctrine would insist that all citizens enjoy the same privileges and immunities in their public capacities, while permitting private initiatives to pursue elective affinities. This protects (but may subtly discourage) private diversity, because the people, as public citizens, will be the same under law.

11. Republican Philosophy

Republican philosophy since Cicero has sought to construct a harmony of interests and common sense of

justice among citizens through the 'empire of laws and not of men.' Republican laws draw the line between liberty and license, in pursuit of the common good. Republican theory seeks to find and establish good laws, by discovering the principles and basic structure that serve the *res publica* best. So although republican philosophy begins, as Thomas Paine (1737–1809) put it, by making the '*res publica*, the public affairs, or the public good' the object of all governments, and 'republican government is no other than government established and conducted for the interest of the public' (Paine 1792, p. 168), the idea of the republic entails a constellation of political structures to secure republican legislation, embedded in two thousand years of republican tradition, derived from Rome.

The republican commitment to political institutions that protect liberty challenges the later 'liberal' flight from politics after the French revolution. Former republicans such as Benjamin Constant (1767–1830) hoped by securing certain rights or liberties to protect their own private liberty against the state. But republican philosophy recognizes the futility of 'rights' without power. Republican doctrine offers a political epistemology to discover and protect public justice and fundamental human rights. Republicans believe that without popular sovereignty, the rule of law, deliberative senators, elected executives, independent judges, a representative popular assembly, and proper checks and balances, the people cannot know or enjoy their rights and duties to each other, or to the state. The basic desiderata of republican government may appear more (or less) important in different contexts, so that even a monarchy could seem nearly 'republican' when limited by two independent branches in the legislature, and subject to the rule of law (Adams 1787–8, pp. I. xxi–xxii). Contemporary republican philosophers adapt republican principles to new circumstances to protect public liberty against improper domination ('*dominium*') by private interests, or the arbitrary government ('*imperium*') of the state (Pettit 1997). What makes such theories republican is their fundamental commitment to the common good, and to the political structures that support the *res publica* best. Republican philosophy is a theory of freedom and government, or rather of freedom through government, to secure a shared sense of justice, in pursuit of the common good. Republicans believe that there can be no justice without community, no liberty without the law.

See also: Aristotle (384–322 BC); Democracy; Justice: Philosophical Aspects; Republic; Republicanism: Impact on Social Thought

Bibliography

Adams J 1787–8 *A Defence of the Constitutions of Government of the United States of America*. 3 Vols. C. Dilly, London

- Adams J 1776 Thoughts on government. 1983 In: Hyneman C S, Lutz D S (eds.) *American Political Writing During the Founding Era 1760–1805*, 2 Vols. Liberty Press, Indianapolis, IN
- Baron H 1966 *The Crisis of the Early Italian Renaissance: Civic Humanism and Republican Liberty in the Age of Classicism and Tyranny*. Princeton University Press, Princeton, NJ
- Bleicken J 1972 *Staatliche Ordnung und Freiheit in der römischen Republik*. Lassleben, Kallmünz, Germany
- Bock G, Skinner Q, Viroli M (eds.) 1990 *Machiavelli and Republicanism*. Cambridge University Press, Cambridge, UK
- Constant B 1988[1819] De la liberté des anciens comparée à celle des modernes. In: Fontana B (ed.) *Political Writings*. Cambridge University Press, Cambridge, UK
- Cornelius Tacitus P 1906 *Ab excessu divi Augusti annalium libri*. In: Fisher C D (ed.) Oxford University Press, Oxford, UK [1974 *The Annals*. Penguin, Harmondsworth, UK]
- Cornelius Tacitus P 1911 *Historiarum libri*. In: Fisher C D (ed.) Oxford University Press, Oxford [1964 *The Histories*. Penguin, Harmondsworth, UK]
- Fink Z S 1962 *The Classical Republicans: An Essay in the Recovery of a Pattern of Thought in Seventeenth-Century England*, 2nd edn. Northwestern University Press, Evanston, IL
- Fisher H A L 1911 *The Republican Tradition in Europe*. G. P. Putnam's Sons, New York
- Fontana B. (ed.) 1994 *The Invention of the Modern Republic*. Cambridge University Press, Cambridge, UK
- Gordon T 1728–31 *The Works of Tacitus with Political Discourses Upon that Author*, 4 Vols. Woodward and Peele, London
- Gordon T 1746 *The Works of Sallust Translated into English with Political Discourses Upon that Author*. Woodward and Peele, London
- Habermas J 1992 *Faktizität und Geltung: Beiträge zur Diskurstheorie des Rechts und des demokratischen Rechtsstaats*. Suhrkamp, Frankfurt, Germany [1996 *Between Facts and Norms: Contributions to a Discourse Theory of Law and Democracy*. MIT Press, Cambridge, MA]
- Hamilton A, Jay J Madison J ['Publius'] 1787–8 *The Federalist*. [1987 Kramnick I. (ed.) *The Federalist Papers*. Penguin, Harmondsworth, UK]
- Harrington J 1656 *The Commonwealth of Oceana*. [1992 Pocock J G A (ed.) Cambridge University Press, Cambridge, UK]
- Hobbes T 1651 *Leviathan, or The Matter, Form, and Power of a Commonwealth Ecclesiasticall and Civil*. [1991 Tuck R (ed.) Cambridge University Press, Cambridge, UK]
- Kant I 1795 *Zum ewigen Frieden*. F. Nicolovius, Königsberg
- Koenigsberger H G (ed.) 1988 *Republiken und Republikanismus im Europa der frühen Neuzeit*. R. Oldenbourg, Munich, Germany
- Livius T *Ab urbe condita I–V*. 1974, Ogilvie R M (ed.) Oxford University Press, Oxford, UK [1971 *The Early History of Rome*. Penguin, Harmondsworth, UK]
- Liviu, T *Ab urbe condita VI–X*. 1919, Walters C F, Conway R S (eds.) Oxford University Press, Oxford, UK [1982 *Rome and Italy*. Penguin, Harmondsworth, UK]
- Macchiavelli N 1517 *Discorsi sopra la prima deca di Tito Livio* [1984 Inglese G (ed.) Rizzoli Editore, Milano] [1970 *The Discourses*. Penguin, Harmondsworth, UK]
- Montesquieu C de Secondat, baron de 1734 *Considérations sur les causes de la grandeur des Romains et de leur décadence*. [1967 Truc G (ed.) Editions Garnier, Paris] [1968 *Considerations on the Causes of the Greatness of the Romans and Their Decline*. Cornell University Press, Ithaca, NY]
- Montesquieu C de Secondat, baron de 1748 *De l'esprit des lois*. [1973 Derathé R (ed.) 2 Vols. Editions Garnier, Paris] [1989 *The Spirit of the Laws*. Cambridge University Press, Cambridge, UK]
- Nicolet C 1982 *L'idée républicaine en France (1789–1924): Essai d'histoire critique*. Gallimard, Paris
- Paine T 1792 *The Rights of Man, Part II*. [1989 In: Kuklick B (ed.) *Thomas Paine: Political Writings*. Cambridge University Press, Cambridge, UK]
- Pettit P 1997 *Republicanism: A Theory of Freedom and Government*. Oxford University Press, Oxford, UK
- Pettit P 1999 *Postscript to the Paperback Edition of Republicanism: A Theory of Freedom and Government*. Oxford University Press, Oxford, UK
- Plutarchos *Bioi paralleloi*. [1949–1971 In: Lindskog C, Ziegler K (eds.) *Plutarchi Vitae Parallelae*. Teubner, Leipzig] [1914–26 *The Parallel Lives*. Harvard University Press, Cambridge, MA]
- Pocock J G A 1975 *The Machiavellian Moment: Florentine Political Thought and the Atlantic Republican Tradition*. Princeton University Press, Princeton, NJ
- Polybios *Historia* [1866–1868 Dindorf L (ed.) 4 Vols. Teubner, Leipzig] [1922–7 *The Histories*. Harvard University Press, Cambridge, MA]
- Powell J G F (ed.) 1995 *Cicero the Philosopher: Twelve Papers*. Clarendon Press, Oxford, UK
- Rahe P A 1992 *Republics Ancient and Modern: Classical Republicanism and the American Revolution*. University of North Carolina Press, Chapel Hill, NC
- Rawls J 1993 *Political Liberalism*. Columbia University Press, New York
- Robbins C 1959 *The Eighteenth-century Commonwealthman: Studies in the Transmission, Development and Circumstance of English Liberal Thought from the Restoration of Charles II until the War with the Thirteen Colonies*. Harvard University Press, Cambridge, MA
- Rousseau J-J 1762 *Du contrat social*. [1973 Guillemin H (ed.) Union Générale d'Éditions, Paris] [1997 *The Social Contract*. Cambridge University Press, Cambridge, MA]
- Rush B 1777 *Observations upon the Present Government of Pennsylvania in Four Letters to the People of Pennsylvania*. Steiner and Cist, Philadelphia
- Sallustius Crispus C *Bellum Catilinae*. [1991 In: Reynolds L D (ed.) *Catilina; Jugurtha; Historiarum Fragmenta Selecta; Appendix Sallustiana*. Oxford University Press, Oxford, UK] [1921 *The War with Catiline*. Harvard University Press, Cambridge, MA]
- Sallustius Crispus C *Bellum Jugurthinum*. [1991 In: Reynolds, L D (ed.) *Catilina; Jugurtha; Historiarum Fragmenta Selecta; Appendix Sallustiana*. Oxford University Press, Oxford, UK] [1921 *The War with Jugurtha*. Harvard University Press, Cambridge, MA]
- Sandel M J 1996 *Democracy's Discontent: America in Search of a Public Philosophy*. Belknap Press of Harvard University Press, Cambridge, MA
- Sellers M N S 1991 Republican impartiality. *Oxford Journal of Legal Studies* 11: 273–82
- Sellers M N S 1994 *American Republicanism: Roman Ideology in the United States Constitution*. Macmillan, Basingstoke, UK
- Sellers M N S 1996 Republican principles in international law. *Connecticut Journal of International Law* 11: 403–32
- Sellers M N S 1998 *The Sacred Fire of Liberty: Republicanism, Liberalism and the Law*. Macmillan, Basingstoke, UK

- Sidney A 1798 *Discourses Concerning Government*. [1990 West T G (ed.) Liberty Classics, Indianapolis, IN]
- Skinner Q 1998 *Liberty Before Liberalism*. Cambridge University Press, Cambridge, UK
- Spitz J-F 1995 *La liberté politique: Essai de généalogie conceptuelle*. Presses Universitaires de France, Paris
- Sunstein C R 1993 *The Partial Constitution*. Harvard University Press, Cambridge, MA
- Symposium: The Republican civic tradition 1988 *The Yale Law Journal* 97(8)
- Trenchard J, Gordon T 1723 *Cato's Letters or Essays on Liberty, Civil and Religious, and Other Important Subjects*, 4 Vols. [1995 Hamowy R (ed.) 2 Vols. Liberty Fund, Indianapolis, IN]
- Tullius Cicero M *De legibus*. [1914 Mueller C F W (ed.) Teubner, Leipzig] [1998 *The Laws*. Oxford University Press, Oxford, UK]
- Tullius Cicero M *De officiis*. [1994 Winterbottom M (ed.) Oxford University Press, Oxford, UK] [1991 *On Duties*. Cambridge University Press, Cambridge, UK]
- Tullius Cicero M *De re publica*. [1969 Ziegler K (ed.) Teubner, Leipzig] [1998 *The Republic*. Oxford University Press, Oxford, UK]
- Tullius Cicero M. *M. Antonium orationes Philippicae*. [1986 Fedeli P (ed.) 2nd edn. Teubner, Leipzig] [1926 *Philippics*. Harvard University Press, Cambridge, MA]
- Tullius Cicero M *Orationes in L. Catilinam*. [1901 Clark A B (ed.) Oxford University Press, Oxford, UK] [1976 In *Catilinam*. Harvard University Press, Cambridge, MA]
- Tullius Cicero M *Pro L. Flacco oratio*. [1906 In: Mueller C F W (ed.) *M. Tulli Ciceronis orationes, Pro P. Sulla, Pro A. Licinio Archia poeta, Pro L. Flacco*. Teubner, Leipzig] [1976 *Pro Flacco*. Harvard University Press, Cambridge, MA]
- Tullius Cicero M *Pro P. Sulla oratio*. [1906 In: Mueller C F W (ed.) *M. Tulli Ciceronis orationes, Pro P. Sulla, Pro A. Licinio Archia poeta, Pro L. Flacco*. Teubner, Leipzig] [1976 *Pro Sulla*. Harvard University Press, Cambridge, MA]
- Washington G 1789 The First Inaugural Speech. [1988 In: Allen W B (ed.) *George Washington: A Collection*. Liberty Classics, Indianapolis, IN]
- Wirszubski C 1960 *Libertas as a Political Idea at Rome during the Late Republic and Early Principate*, 2nd edn. Cambridge University Press, Cambridge, UK
- Wood G S 1969 *The Creation of the American Republic, 1776–1787*. University of North Carolina Press, Chapel Hill, NC

M. N. S. Sellers

Reputation

A reputation exists in the form of an image shared within some public. Applied to a person, it identifies ‘not an intrinsic quality he possesses but rather the opinion other people have of him’ (Bailey 1971). It can also refer to certain collectivities, like a company, occupational category, professional group, organization, or its leadership and, by extension, to just about anything that can somehow be rated or ranked. Thus, we talk about the reputation of a dining establishment, of a vintage year of wine, of an annual event, of a

locality or city. All such reputations, whatever they be, are socially constructed. They may or may not be deserved and are always open to challenge. But, as long as they endure, they are real in their consequences.

Over 300 years ago Locke referred to the ‘law of opinion and reputation’ in order to highlight how much reputations were a part of the small politics of everyday life (1690/1894): ‘People have a strong motive to accommodate themselves to the opinions and rules of those with whom they converse ... and so they do that which keeps them in reputation with their company, with little regard for the laws of God, or of the magistrate ...’ Smith, a century later, argued in similar vein when he located the love of virtue in the desire for ‘being what ought to be approved of,’ a motive he believed to be most strongly felt among persons of middling or inferior status who, possessing none of the privileges associated with noble lineage or inherited wealth, must, in order to distinguish themselves, bring their talents into public view. Besides, they ‘can never be great enough to be above the law ... (Their) success ... almost always depends upon the favour and good opinion of their neighbours and equals; and without a tolerably regular conduct, these can very seldom be obtained’ (1804).

These ideas persist. Economists represented in a volume edited by Klein (1997) depict society as permeated by a flowing network of reputational relationships, with institutional credentials and certification substituting for reputations based on personal acquaintance.

Insofar as one can extract value from it, a reputation can be further conceptualized as a form of symbolic capital (Bourdieu 1992). It can grow or diminish yet still retain some value so long as people remember. This value is real enough for anyone who feels his or her good name to have been unjustly sullied by the careless or malicious actions of others. Redress can be sought through the courts—not always successfully, to be sure. And tax laws in some countries allow the corporate write-off of expenditures for building and protecting the reputation of a company or its brand name.

Reputations vary along three dimensions: best known is the vertical dimension that locates an agent or object on a hierarchical scale. A high reputation signifies the attribution of a level of respect, credibility, authority, or competence denied to those who do not measure up to that standard. Reputations, high as well as low, also vary along a lateral dimension, where placement is determined by the size of the public within which a name has currency. To differentiate between these two dimensions, we refer to a reputation for upright character or some talent shared by neighbors, colleagues, professional peers, and other insiders as *recognition* while reserving the term *renown* for that more universal recognition that extends beyond a particular reputational community (Lang and Lang 1990). A parallel distinction at the low

Table 1

Vertical (Status)	Lateral (Expansion)		Temporal (Duration)	
	Insiders	Public Sphere	Short	Long
High	<i>Recognition</i>	<i>Renown</i>	<i>Celebrity</i>	<i>Eminence</i>
Low	<i>Discredit</i>	<i>Notoriety</i>	<i>Ignominy</i>	<i>Infamy</i>

end of the vertical dimension is between being in *disrepute* among peers and the *notoriety* attained when disreputing information carries to the general public. The terms are diagramed in the Table 1 below.

A third *temporal* dimension has to do with duration in time. Public attention is forever being distracted by new entrants into the limelight. Reputations follow trajectories of different lengths. A celebrity is typically a person who may enjoy moments of fame but thereafter is remembered only by the few with a special interest in preserving the memory. Only the 'truly great' achieve the kind of *eminence* that accounts for their survival in the collective memory as major historical figures (Collins 1998). Alternatively, the ignominy experienced for some heinous deed may be quickly forgotten or be forever remembered in infamy. Most reputations fall somewhere between these extremes. The categories are depicted in the accompanying figure.

Sociologists, and especially historians as well as some literary scholars, have focused mainly on the lateral and temporal dimensions of high status, namely on how particular individuals attain renown and are canonized. Economists and political scientists, many of them employing game theoretical models, as well as some anthropologists have exhibited more interest in how reputations function within particular settings.

1. Effects of Reputation

Effects can be divided into two general categories: those involving the reactions of others to the reputation of an individual and those that stem from an individual's anticipation of reactions to their own conduct. Obviously, there is a good deal of overlap between these two.

The effect of a personal reputation on those with whom one directly interacts has been illustrated famously in the fable about the boy who cried 'Wolf' once too often. When a wolf actually showed up, as we all have learned, no one in the village heeded his warning, having come to expect just another false alarm. So it is that signals from a source deemed untrustworthy are apt to be ignored and dealings with a party of dubious reliability will, wherever possible, be avoided. But we do not always know enough about

the other nor are we at liberty to choose. Where such doubts exist, the willingness to co-operate or enter into an agreement becomes contingent on external enforcement.

How reputations derived from past performances of parties engaged in repetitive play affect both the strategies adopted by competitors and their preferences for partners has been modeled by game theorists. Their highly suggestive work is commendable for its logical rigor. Still, one can question how well their propositions tested in experiments, most of them with only two players interacting within controlled settings, replicate the complexities of real life. For one thing, reputational effects are based not just on personal experience but on word that gets around. Thus, collective attributions are a cardinal factor in the power an American president can wield. Neustadt locates its source, not just in the Constitution but in what others in the Washington political community think and expect of the incumbent. Justified or not, 'it is the reality on which they act, at least until their calculations turn out wrong ... (1990). What the president must demonstrate in the many different games he plays, often all at the same time, is his ability and will to make use of his bargaining advantages. Success breeds success but a pattern of repeated failure is bound to result in a loss of faith in his effectiveness that will reduce his power to persuade.

Similar self-enhancing effects show up in nonpolitical settings. Distinguished scientists get disproportionate credit for their contribution to a collaborative project (Merton 1967). The prestige of a university affects the quality of the faculty it attracts, whose members then, in turn, benefit from their position and raise the prestige of the institution; artists who are already well-known get more exhibitions; the works of writers who have already won prizes are more actively promoted. The epitomizing case is that of a celebrity, described by Boorstin (1961) as someone 'known for his being known.'

Images may be poor substitutes for more direct knowledge; yet in a world where most people are strangers to one another, there may be nothing else to govern our calculations. Images certainly affect what we buy, how we invest, and how much we are willing to pay (Fombrun 1996). The reassurances provided by an established brand name usually exacts some premium. But nowhere are the returns from a good

reputation greater than in the arts, where values are somewhat intangible, subject to fashion, highly dependent on critical and public endorsement (Moulin 1986). The name attached to a work determines not only price but also the attention it gets. The media, very much oriented toward a star system, feed this proclivity, thereby contributing to a pattern in which small differences among a number of top performers, none of them unequivocally superior to the rest, translate into large differences in acclaim (Goode 1978; Frank and Cook 1995).

Even in the corporate world the relation of reputation to measurable performance is not as close as it may seem. High publicity given the Valdez oil spill in Alaska waters was followed by a significant short-term decline in the value of Exxon stock. At least some investors, it has been found, back away from companies not so much because of any loss of faith in their future performance but doubts about their commitment to general community values (Fombrun 1996).

The fear of damage to one's reputation becomes an incentive for behavior against one's immediate self-interest. What economists, following Adam Smith, have called the discipline of repeated dealings acts as a deterrent, not always effective, against business selling shoddy goods to naive customers or against lobbyists providing clearly misleading information to legislators. But this response is not simply a matter of honesty but of strategic impression management as, for example, when a college reduces its acceptance rate as part of a marketing strategy to portray itself as highly selective. On the international scene, where to yield 'on a series of issues when matters at stake are not critical ... may (make) it difficult to communicate to the other (side) just when a vital issue has been reached,' a nation may be moved to act for no other reason than to retain its 'reputation for action' (Schelling 1966). Against such an instrumental view, Chong (1992) has argued that, when a reputation becomes part of their identity, people try to live up to it for moral reasons as well, trying to turn themselves into the kind of person they want others to think they are.

Another set of reputational incentives comes into play when the expansion of a reputation into the public sphere creates what Braudy (1986) calls the 'frenzy of renown' with numerous effects, some of which are potentially deleterious to one's moral commitments, whatever these may be. Fame and the appearance of power it gives are sometimes seductive. For example, Gitlin (1980) describes how media attention undermined the connection of some leading activists of the New Left, a movement with a weak organizational structure, to their original reputational community. Competing for a place in the limelight, these 'leaders' transformed themselves into 'un-attached celebrities,' no longer accountable to the rank-and-file but drawn toward increasingly 'extravagant, "incidental," expressive actions—actions which

made "good copy" because they generated sensational pictures rich in symbolism.' In sports, the advent of a national public for major-league baseball, largely a product of television, has been associated with a decline in the advantage of playing on one's home turf. The present-day public, according to Leifer (1995), is less loyal to home-grown athletes and teams but looks for any winner it can celebrate. It is not averse to abandoning a team that falls short of its expectations. By interpreting what were initially random performance differences between contestants, the public casts competitive equals into stable winner and loser roles. The teams react to reputations established early in the season, not only by the public at the game but also by sportscasters and writers when in fact, Leifer contends, many outcomes might easily have gone the other way.

2. *From Recognition to Renown*

Propositions about the acquisition of fame derive largely from life histories, mostly of successful individuals who by virtue of their background, talent, or associations seemed destined for success. To gain recognition within some domain for some special talent or accomplishment depends on a number of general conditions. Endowment and ability are, without a doubt, prerequisites. But the pool of talent is large and some acclaimed figures have had to overcome what would appear to have been formidable handicaps. Albert Einstein is known to have failed his school examination and Glenn Cunningham, the champion miler for several years, had legs so severely burned as a child that his doctors were not at all sure whether he would even walk again. Regardless of any such handicap, everyone in pursuit of recognition has to do what is necessary to develop the skills that count.

There is, to be sure, a world of difference between measurable improvements of athletic performance or scores in games played according to strict rules and the evolving consensus by which the less tangible abilities prized in domains other than sports come to be validated. Nothing counts unless people take notice. The achievement of a recluse working all alone is more likely to be overlooked than comparable ones by more gregarious contemporaries. Insofar as reputations carry through networks, acceptance by peers and mentors is often the first step toward gaining recognition. Thereafter it becomes a question of how far these networks extend. Persons highly esteemed in one domain may not even be recognized outside its boundaries or can be subject to all sorts of invidious distinctions. Hence, reputations do not easily transfer, even between genotypically similar activities, especially if they are as differentially valued. In most arts, where only some genres are ordained by an arts establishment, an excess of popularity can actually be the kiss of death. Kapsis (1992) traces the career of

Alfred Hitchcock, the Hollywood director of many famous films, who transformed himself deliberately from a merely successful maker of movies into a genuine *auteur* by adopting a new and more acceptable idiom.

The general rule that one is better off not straying too far from the mainstream holds even in the sciences where the highest honors usually go to the innovators. To be working on a problem whose importance has been widely acknowledged increases the likelihood of gaining attention for work that adds only incrementally to existing knowledge. Attaching oneself to important people is a second strategy, equally appropriate to the sciences and to the arts. Yet dissidents denied their place by a complacent establishment can also collectively advance their reputations by banding together, forming their own societies, and using them as a springboard to raise their status. This is how the group of artists denigrated as 'impressionists,' due to their apparent disdain for serious subject matter, drew attention to themselves.

Whether recognition within a group converts into renown is highly contingent on distribution and publicity (Becker 1982). Visual artists depend on exhibition space for display of their accomplishments; writers on publishers and booksellers; playwrights on performers and on a suitable stage, and the performers themselves on recordings and bookings in a major medium to display their accomplishments. To be singled out for special praise by influential critics and other mediators of fame is still more important (Moulin 1986). Especially if their judgments stimulate museum purchases, give a theatrical work a permanent place in the repertory, or generate substantial interest in the writings of an author, they can make a household word out of an obscure name. Most likely to bring instant fame are the feats that receive the greatest publicity, like the breaking of an established record in one of the major sports or ventures into entirely new territories in science or in space. A few individuals are singled out as stars with very little acknowledgment of the collective effort that made their achievement possible. Only a small number of scientists know the names of the many highly competent engineers who have made space flight possible.

3. Measuring Reputations

In a society set on rating and ranking everything, there are essentially two approaches to the measurement of reputation: through responses to survey questions and through aggregations of indicators about the actual reactions of some public.

Judgments expressed by survey respondents are most meaningful in settings where just about everyone knows everyone else. Accordingly, performance can be predicted from measures of certain qualities based

on average ratings by groups of their peers. No less ascertainable is what the residents of a village community think of their neighbors or which professionals are held in the highest regard by their colleagues. However, the content of any such consensus becomes more problematical the farther removed those queried are from whomever they are asked about. For years pollsters have obtained what legitimately can be called a reputational ranking by asking a representative sample of Americans to name the man and also the woman they most admire. Because respondents were free to name anyone alive, their choices have always been widely scattered with the president and the first lady almost always heading the lists. Their rankings, as well as that of others, has have as much to do with their prominent coverage in the news media as with the respect they enjoy within their particular reputational community.

The American president is of course one of a very small number of celebrities or infamous villains about whom just about everyone has some opinion. Thus, standardized questions about presidential performance approximate a vote and can be the basis for comparing a president's reputation to that of other actors prominent on the political scene. But there is as yet no general agreement on a standard for measuring other reputations in the public sphere. Should the reputations of corporations be based on the collective judgment of high-level executives, of customers, or of their employees? That of schools on responses from their alumni or of academics from those working in the same discipline? Who does the ranking is particularly important in the arts, governed as they are by the distinction between real 'quality' ascertained by art experts and what is merely popular.

When it comes to the past, there is no alternative to querying specialists. This method works best when evaluating a narrowly defined category, like American presidents or justices of the Supreme Court. In the first such survey, 55 experts on the presidency were asked to place each occupant of the office into one of five categories ranging from 'great' to 'failure' (Schlesinger 1997). Others have queried larger samples with more 'refined' methods only to find that the same three—Washington, Lincoln, and Franklin D. Roosevelt—consistently placed at the top, though not always in the same order. Choices of the best and worst presidents have also remained relatively stable with more fluctuation among those ranked in between. Most striking has been the steady rise in reputation of Eisenhower as new information about his presidency has come to light.

Citation counts are one of the most widely used indicators of the reputation of individuals and of academic departments. Such information is very accessible. Because scholars are obliged to cite prior publications on which their own work is built or with which they take issue, the number of references to the work of a scientist or scholar in any particular

discipline is assumed to indicate the level of recognition by colleagues. Yet strategic considerations also enter. Citing a particular predecessor can help a neophyte establish a link to one or more of the prestigious figures in their field, thereby drawing the attention of colleagues and enhancing the significance of one's own work (Camic 1987). The more a field is subject to intellectual fashion the more likely it is that credit for an idea that is very much in the air goes to the more famous.

The frequency with which a particular name appears in the more popular media indicates its role in a broader discourse but there are, at least in some fields and for some individuals, serious discrepancies between name counts for introductory texts and those for more esoteric publications aimed at narrower constituencies. In drawing conclusions from any such enumerations, one needs always to take into account the perspective of publications examined. But any set of similar publications can be useful in examining trends. Rosengren (1987) noted that in texts of Swedish literature the same names that turned up in an earlier period were still prominently reviewed four decades later.

A simple way of gauging the level of eminence (or infamy) attained by any individual is through the cumulation of literature in major libraries, most of whose holdings are now on line. After all, writers are the primary purveyors of fame (Cooley 1927). The number of items a search turns up is a measure of how much what the subjects did and the issues they raised have retained their salience as historical markers. The larger the universe searched the more finely do the findings discriminate. A search limited to full-length biographies would omit many minor figures somewhat renowned in their lifetime. Searches covering long time spans are open to another kind of bias. Although material evidence gets lost over time and some of the eminent of long ago may have left few traces, posterity has had many more opportunities to develop a consensus about those whose 'lives' have somehow survived. In terms of number of mentions and pages devoted to each in an authoritative encyclopedia of communication, a few 'greats' from the past outranked even the most distinguished contemporary scholars in this burgeoning field (Beniger 1990). It therefore makes sense to restrict such counts to differentiating between the luminaries of a particular time period, whose ideas remain of great interest, and their contemporaneous lesser lights as Collins (1998) does in his study of philosophers.

Another measure focuses less on citations than on decisions made over the years by the custodians of tradition as to what part of the past to preserve. The results are manifest in the number of memorials, shrines, plaques, exhibits, and commemorative events designed to keep alive the memory of a particular person or persons. To obtain a full record of these tributes is difficult, except perhaps for some of the arts

where one can ascertain the number of times a composer's work is performed by major orchestras (Simonton 1986; Mueller 1958) or a play is performed or revived (Griswold 1987). In the visual arts, the aggregate number of works in major museums and on permanent display is an appropriate indicator of collective curatorial judgment but somewhat confounded by rarity. Because there is not enough work by some of the greatest artists to meet the demand of museums, a single indicator, based on their holdings, has to be supplemented by information about the frequency with which work is circulated in special loan exhibitions, many of which are organized precisely to commemorate the artist.

The market is the ultimate but albeit imperfect arbiter of value. Many variables, like age, rarity, and condition, enter into prices paid for works and memorabilia associated with past eminence. Still, for items comparable on these variables, consistent price differentials represent the consensus of their relative worth. They are, to this extent, valid measures of reputation.

4. *The Survival of Reputation*

Over time, as memories fade, the lives of most people end up in the dust bins of history. There is a continuing sorting out of those deemed memorable. As to likely survival in the collective memory, posthumous reputations are the combined effect of two conditions: the rhetorical force of the accomplishments of the deceased on society or some domain in their lifetime and the extent to which, even if for altogether different reasons, an accomplishment still touches the imagination of posterity. 'Greatness' can also be conferred retrospectively as validation of or correction to earlier valuations.

The renown achieved in a person's lifetime enhances the chance for reputational survival. It facilitates contact with persons who become the mediators of one's reputation across generations (Zilsel 1972). Kings, popes, presidents, generals, and executives may have archivists at their command but those without an official position have to find their place in the right networks. Attention from colleagues attracts students, who will pass their valuations of a scholar on to the next generation. For an artist, critical acclaim brings wealthy collectors, whose purchases have a good chance of ending up in a museum. It is by no means uncommon for the admirers of some creative person to form a society dedicated to preserving and promoting their reputation.

Famous people also leave more traces. Anything connected with them gains value, which provides incentives for preserving mementoes of their lives, especially any pictures, diaries, and correspondence that evidence their thoughts and character. Insofar as their paths have crossed those of other famous people,

many of the letters they have written or received are likely to survive in some archive and to attract scholars prepared to ferret out still more material.

Subjects have some, though limited, control over the record on which their reputational fate will come to depend. Those who take care to keep and organize records strengthen their claim on the memory of posterity. It has become standard for a president to initiate the setting up of a library, even though his papers would be saved in any event. Some famous people are known to have taken the further step of purging from their files items inconsistent with the image they want to project. Writers have burned manuscripts they believed detrimental to their reputation. There are painters who destroyed systematically all pictures produced during a period for which they did not want to be remembered. Theodore Roosevelt dictated letters intended for posterity. Thomas Hardy actually controlled the biography his wife published under her own name following his death (Hamilton 1994). President Richard Nixon's attempt to retain control over his secretly recorded tapes failed. The same impulse shows up in instructions to estate executors to convert the residence of the deceased into a landmark or to use remaining assets to preserve his or her reputation. To be sure, these wishes are not always followed or followed only half-heartedly.

Some factors affecting the survival of reputation remain beyond the control of the individual: first, there are the accidental circumstances of birth and death. Possible candidates for reputational immortality can be born into a time or live in a place where no one acts to preserve the record of their deeds. Thus, the Chinese predecessor of Gutenberg remains unknown and Columbus rather than Leif Ericson is hailed as the discoverer of America. As to the timing or circumstances of death, a long productive life full of honors is an obvious reputational asset but to be struck down in the prime of life by an assassin, a tragic illness, or unfortunate accident has the potential to transform the victim into an icon. No one can possibly know how John F. Kennedy would have been remembered had he served for a full one or two terms. Yet the halo put around him by his untimely death overshadows the modest achievements and all too apparent failures of his administration. A still more striking example is that of Anne Frank. Would her diary, translated into many languages, have had the same impact if she had survived the Holocaust? And had she survived, would she have tolerated the publication by her father of this very private document from which he excised passages that displeased him?

The availability of survivors, like Anne Frank's father, with a personal stake in cultivating remembrance depends, in part, on such demographic factors as marriage, fertility, and longevity. Lang and Lang (1990), studying the revival of etching as an original art form in the late-nineteenth century, found that it

had attracted an unusually large number—at least for that time period—of talented women well recognized during their lifetimes, yet today much less well remembered than their male peers. Two demographic circumstances contributed to this reputational disparity: the extraordinary longevity of the women and the much smaller number ever married or, if married, with children. Hence, they outlived, often by decades, not only their spouses but also their colleagues, close friends, publishers, dealers, admiring collectors, as well as the collapse of the etching craze, so that there was often no one still around to mourn their death and write the obligatory obit. Also, because unmarried women had neither mate nor daughter to take on the tedious task of record keeping, material for an artistic biography or definite catalogue raisonné are often lacking.

All reputational communities have boundaries. The heroes celebrated in one may be largely unknown in another. Nowhere is this more obvious than in the different statesmen, military heroes, literary figures and so forth each nation celebrates. National perspectives even intrude into the allocation of credit for achievement universally recognized as breakthroughs. Obviously, the pioneers of space exploration who have received the greatest attention in America differ from those celebrated in the former Soviet Union. The inventor of the incandescent lamp, as remembered by Americans, is Thomas Edison, who was in fact preceded by Joseph Swan, who first exhibited his successful carbon-filament lamp in England, where he had greater impact. His factory near Newcastle rivaled Edison's in prosperity. But whereas Edison immediately patented everything, Swan was slow to do so. Nor did he have a wealthy sponsor like Henry Ford, who thought him the greatest man alive and started collecting Edison memorabilia as early as 1905, and ultimately erecting a shrine in Dearborn, Michigan by transporting, piece by piece, and reassembling the inventor's New Jersey laboratory. But for a reputation to cross all cultural and geographic boundaries, the achievement behind it must appear unique and be an inspiration for people all over the world.

Time creates its own boundaries. There is only so much room at the top, so that at some point what Collins (1998) calls the law of small numbers comes into play. Secondary figures recede into the background as new interests emerge. The survivors are those whose deeds and accomplishments still speak to the present. Posterity has some leeway in its view of the past. Not only does it select but it reinterprets continuously, particularly when it comes to the work of writers and artists. The qualities that contemporaries most admired in Hawthorne's oeuvre (Tompkins 1984) and in Da Vinci's *Mona Lisa* are not the same ones that touch us today (Boas 1940). Put another way, eminence derives from the persistence of beliefs about persons and their deeds and creations that have changed their domain in some fundamental way. The

larger the number of domains affected, the greater the likelihood that their reputations will live on.

The development, change, preservation, and decline of reputations as well as how they function have to be understood as part of an ongoing social process. Experiments and statistical studies provide useful information, mostly about outcomes, but require supplementation of case material on the process as it unfolds in a particular context. Qualitative evidence, when derived from more than one case and subjected to systematic analysis, is not to be discounted as merely anecdotal.

See also: Celebrity; Charisma and Charismatic; Charisma: Social Aspects of; Potlatch in Anthropology; Sport, Sociology of

Bibliography

- Bailey F G 1971 Gifts and poison. In: *Gifts and Poison: The Politics of Reputation*. Blackwell, Oxford, UK, pp. 1–25
- Becker H 1982 *Art Worlds*. University of California Press, Berkeley, CA
- Beniger J R 1990 Who are the most important theorists of communication. *Communication Research* 17: 698–715
- Boorstin D J 1961 *The Image: Or, What Happened to the American Dream*. Atheneum, New York
- Bourdieu P 1992 *Les règles de l'art*. Editions du Seuil, Paris
- Braudy L 1986 *The Frenzy of Renown*. Oxford University Press, New York
- Camic C 1987 The making of a method: A historical reinterpretation of early Parsons. *American Sociological Review* 52: 421–39
- Chong D 1992 Reputation and cooperative behavior. *Social Science Information* 31: 683–709
- Collins R 1998 *The Sociology of Philosophy: A Global Theory of Intellectual Change*. Belknap Press, Cambridge, UK
- Cooley C H 1927 Fame. In: *Social Process*. Scribner's, New York, pp. 112–24
- Fombrun C J 1996 *Reputation: Realizing the Value from the Corporate Image*. Harvard Business School Press, Boston
- Frank R H, Cook P J 1995 *The Winner-take-all Society: How More and More Americans Compete for Ever Fewer and Bigger Prizes, Encouraging Economic Waste, Income Inequality, and an Impoverished Cultural Life*. Free Press, New York
- Gitlin T 1980 *The Whole World is Watching: Mass Media in the Making & Unmaking of the New Left*. University of California Press, Berkeley, CA
- Goode W J 1978 *The Celebration of Heroes: Prestige as a Social Control System*. University of California Press, Berkeley and Los Angeles, CA
- Griswold W 1986 *Renaissance Revivals: City Comedy and Revenge Tragedy in the London Theatre, 1576–1980*. University of Chicago Press, Chicago
- Hamilton I 1994 *Keepers of the Flame: Literary Estates and the Rise of Biography from Shakespeare to Plath*. Faber & Faber, Boston
- Kapsis R E 1992 *Hitchcock: The Making of a Reputation*. University of Chicago Press, Chicago
- Klein D B (ed.) 1997 *Reputation: Studies in the Voluntary Elicitation of Good Conduct*. University of Michigan Press, Ann Arbor, MI
- Lang G E, Lang K 1990 *Etched in Memory: The Building and Survival of Artistic Reputation*. University of North Carolina Press, Chapel Hill, NC
- Leifer E M 1995 *Making the Majors: The Transformation of Team Sports in America*. Harvard University Press, Cambridge, MA
- Locke J 1690/1894 *An Essay Concerning Human Understanding*. Clarendon Press, Oxford, UK
- Merton R K 1967 The Matthew effect in science. *Science* 159(5): 56–63
- Mueller J H 1958 *The American Symphony Orchestra: A Social History of Musical Taste*. Calder, London
- Neustadt R E 1990 *Presidential Power and the Modern Presidents: The Politics of Leadership from Roosevelt to Reagan*. Free Press, New York
- Rosengren K E 1987 Literary criticism: Future invented. *Poetics* 16: 295–325
- Schelling T C 1966 *Arms and Influence*. Yale University Press, New Haven, CT
- Schlesinger A M Jr 1997 Rating the presidents: Washington to Clinton. *Political Science Quarterly* 112: 179–90
- Smith A 1804 *The Theory of Moral Sentiments*. Cadell & Davis, London
- Tompkins J 1984 Masterpiece theatre: The politics of Hawthorne's literary reputation. *American Quarterly* 36: 617–42
- Zilsel E 1926/1972 *Die Entstehung des Geniebriffes: Ein Beitrag zur Ideengeschichte der Antike und des Frühkapitalismus*. Olms, Hildesheim, Germany

K. Lang and G. E. Lang

Resampling Methods of Estimation

Most scientists occasionally or frequently face problems of data analysis: what data should I collect? Having collected my data, what does it say? Having seen what it says, how far can I trust the conclusions? Statistics is the mathematical science that deals with such questions. Some statistical methods have become so familiar in the scientific literature, especially linear regression, hypothesis testing, standard errors, and confidence intervals, that they seem to date back to biblical times. In fact most of the 'classical' methods were developed between 1920 and 1950, by scientists like R. A. Fisher, Jerzy Neyman and Harold Hotelling who were senior colleagues to statisticians still active today.

The 1980s produced a rising curve of new statistical theory and methods, based on the power of electronic computation. Today's data analyst can afford to expend more computation on a single problem than the world's yearly total of statistical computation in the 1920s. How can such computational wealth be spent wisely, in a way that genuinely adds to the

classical methodology without merely elaborating it? Answering this question has become a dominant theme of modern statistical theory.

1. The Bootstrap

The following fundamental question arises in almost every statistical data analysis: on the basis of a data set $\mathbf{z}$ we calculate a statistic $s(\mathbf{z})$ for the purpose of estimating some quantity of interest. For example, $\mathbf{z}$ could be the nine cholesterol reduction scores shown in Table 1, and $s(\mathbf{z})$ their mean value $\bar{z} = 28.58$, intended as an estimate of the true mean value of the cholesterol reduction scores. (The 'true mean value' is the mean we would obtain if we observed a much larger set of scores.) How accurate is $s(\mathbf{z})$?

This question has a simple answer if $s(\mathbf{z})$ is the mean $\bar{z}$ of numbers $z_1, z_2, \dots, z_n$. Then the *standard error* of $\bar{z}$, its root mean square error, is estimated by a formula made famous in elementary statistics courses.

$$\left[\sum_{i=1}^n (z_i - \bar{z})^2 / (n \cdot (n-1)) \right]^{1/2} \quad (1)$$

For the nine numbers in Table 1, Eqn. (1) gives 10.13. The estimate of the true cholesterol reduction mean usually would be expressed as 28.58 ± 10.13 , or perhaps $28.58 \pm 10.13 \cdot s$, where s is some constant like 1.645 or 1.960 relating to areas under a bell-shaped curve. With $s = 1.645$, interval (2) has approximately 90 percent chance of containing the true mean value.

The bootstrap, (Efron 1979), was introduced primarily as a device for extending Eqn. (1) to estimators other than the mean. For example, suppose $s(\mathbf{z})$ is the '25 percent trimmed mean' $\bar{z}\{.25\}$, defined as the average of the middle 50 percent of the data: we order the observations $z_1, z_2, \dots, z_n$, discard the lower and upper 25 percent of them, and take the mean of the remaining 50 percent. Interpolation is required for cases where $0.25 \cdot n$ is not an integer. For the cholesterol data, $\bar{z}\{.25\} = 27.81$.

There is no neat algebraic formula like Eqn. (1) for the standard error of a trimmed mean, or for that matter almost any estimate other than the mean. That

Table 1

Nine men's cholesterol reduction scores after taking cholestyramine; the nine scores are an ordered random sample from the scores of 164 men (data from the Stanford arm of the LRC-CPPT experiment, courtesy of Drs D. Feldman and J. Farquhar, Stanford University (see Efron and Feldman 1991).

-21.0	3.25	10.75	13.75	32.50	39.50	41.75	56.75	80.0

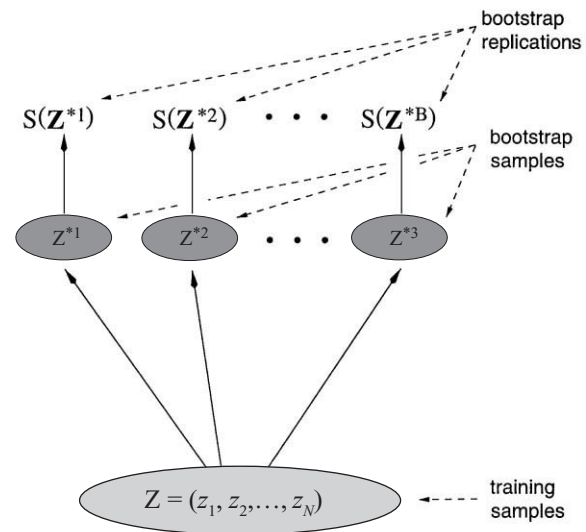


Figure 1

A diagram of the bootstrap algorithm for estimating the standard error of a statistic $s(\mathbf{z})$; each of the B bootstrap samples is a random sample of size n drawn with replacement from the original data set $(z_1, z_2, \dots, z_n)$

is why the mean is so popular in statistics courses. In lieu of a formula, the bootstrap uses computational power to get a numerical estimate of the standard error. The bootstrap algorithm depends on the notion of a *bootstrap sample*, which is a sample of size n drawn *with replacement* from the original data set $\mathbf{z} = (z_1, z_2, \dots, z_n)$. The bootstrap sample is denoted $\mathbf{z}^* = (z_1^*, z_2^*, \dots, z_n^*)$. Each z_i^* is one of the original z values, randomly selected, perhaps $z_1^* = z_7$, $z_2^* = z_5$, $z_3^* = z_5$, $z_4^* = z_9$, $z_5^* = z_7$, etc. The name *bootstrap* refers to the use of the original data set to generate 'new' data sets $\mathbf{z}^*$.

The bootstrap estimate of standard error for $\bar{z}\{0.25\}$ is computed as follows: (a) a large number ' B ' of independent bootstrap samples each of size n , is generated using a random number device; (b) the 25 percent trimmed mean is calculated for each bootstrap sample; (c) the empirical standard deviation of the B bootstrap trimmed means is the bootstrap estimate of standard error for $\bar{z}\{0.25\}$. Figure 1 shows a schematic diagram of the bootstrap algorithm, applied to a general statistic $s(\mathbf{z})$.

Trying different values of B gave these bootstrap estimates of standard error for the 25 percent trimmed mean applied to the cholesterol data

B :	25	50	100	200	400
bootstrap estimate					
of standard error:	12.44	9.71	11.50	10.70	10.48

Ideally we would like B to go to infinity. However, it can be shown that randomness in the bootstrap standard error that comes from using a finite value of B is usually negligible for B greater than 100. 'Negligible' here means small relative to the randomness caused by variations in the original data set $\mathbf{z}$. Even values of B as small as 25 often give satisfactory results. This can be important if the statistic $s(\mathbf{z})$ is difficult to compute, since the bootstrap algorithm requires about B times as much computation as $s(\mathbf{z})$.

The bootstrap algorithm can be applied to almost any statistical estimation problem. (a) The individual data points z_i need not be single numbers; they can be vectors, matrices, or more general quantities like maps or graphs. (b) The statistic $s(\mathbf{z})$ can be anything at all, as long as we can compute $s(\mathbf{z}^*)$ for every bootstrap data set $\mathbf{z}^*$. (c) The data set ' $\mathbf{z}$ ' does not have to be a simple from a single distribution. Other data structures, for example, regression models, time series, or stratified samples, can be accommodated by appropriate changes in the definition of a bootstrap sample. (d) Measures of statistical accuracy other than the standard error, for instance biases, mean absolute value errors, and confidence intervals, can be calculated at the final stage of the algorithm. These points are discussed in Efron and Tibshirani (1993). The more complicated example below illustrates some of them.

There is one statistic $s(\mathbf{z})$ for which we don't need the computer to calculate the bootstrap standard error, namely the mean $\bar{z}$. In this case it can be proved that as B goes to infinity, the bootstrap standard error estimate goes to $\sqrt{(n-1)/n}$ times Eqn. (1). The factor $\sqrt{(n-1)/n}$, which equals 0.943 for $n = 9$, could be removed by redefining the last step of the bootstrap algorithm, but there is no general advantage in doing so. For the statistic $\bar{z}$, using the bootstrap algorithm is about the same as using Eqn. (1).

At a deeper level it can be shown that the logic that makes Eqn. (1) a reasonable assessment of standard error for $\bar{z}$, applies just as well to the bootstrap as an assessment of standard error for a general statistic $s(\mathbf{z})$. In both cases we are assessing the standard error of the statistic of interest by the 'true' standard error that would apply if the unknown probability distribution yielding the data exactly equalled the empirical distribution of the data. The efficacy of this simple estimation principle has been verified by a large amount of theoretical work in the statistics literature of the 1990s (see the discussion and references in Efron and Tibshirani 1993).

Why would we consider using a trimmed mean rather than $\bar{z}$ itself? The theory of *robust* statistics, developed since 1960, shows that if the data $\mathbf{z}$ comes from a long-tailed probability distribution, then the trimmed mean can be substantially more accurate than $\bar{z}$, that is, it can have substantially smaller standard error (see, e.g., Huber 1964). The trouble, of course, is that in practice we don't know *a priori*

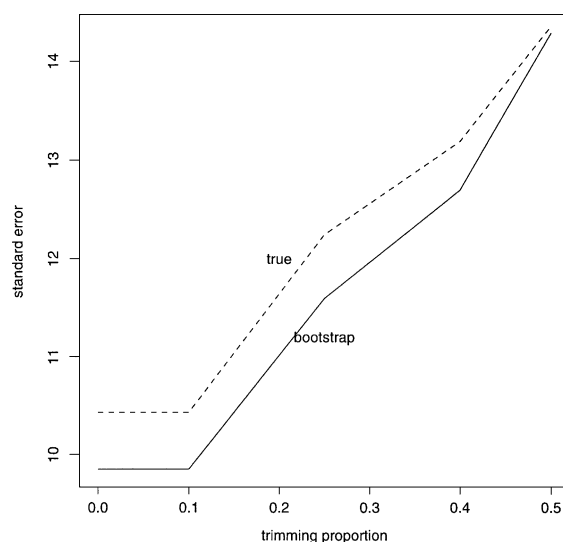


Figure 2

Bootstrap estimates of standard error for five different trimmed means $\bar{z}\{p\}$, $p = 0, 0.10, 0.25, 0.40, 0.5$, applied to the cholesterol data of Table 1. Based on $B = 400$ bootstrap samples. Also shown is the true standard error of $\bar{z}\{p\}$, obtained by taking random samples of size nine from the population of 164 cholesterol reduction scores. In this case the bootstrap correctly indicates that $\bar{z}\{0\}$, the ordinary mean, gives the smallest standard error

whether or not the true probability distribution is long-tailed. The bootstrap can help answer this question.

Figure 2 shows the bootstrap estimates of standard error for 5 different trimmed means $\bar{z}\{p\}$, ' p ' being the proportion of the data trimmed off each end of the sample before the mean is taken. (So $\bar{z}\{0\}$ is $\bar{z}$, the usual mean, while $\bar{z}\{0.5\}$ is the median.) These were computed using the bootstrap algorithm in Fig. 1, $B = 400$, except that at step 2 five different statistics were evaluated for each bootstrap sample $\mathbf{z}^*$, namely $\bar{z}\{0\}$, $\bar{z}\{0.10\}$, $\bar{z}\{0.25\}$, $\bar{z}\{0.40\}$, and $\bar{z}\{0.50\}$.

According to the bootstrap standard errors in Fig. 2, the ordinary mean has the smallest standard error among the five trimmed means. This seems to indicate that there is no advantage to trimming for this particular data set.

In fact the nine cholesterol reduction scores in Table 1 were a random sample from a much bigger data set: 164 scores, corresponding to the 164 men in the Stanford arm of a large clinical trial designed to test the efficiency of the cholesterol-reducing drug called cholestyramine (Efron and Feldman 1991). With all of this extra data available, we can check the bootstrap standard errors. The solid line in Fig. 2 indicates the true standard errors for the five trimmed means, 'true'

meaning the standard error of random samples of size nine taken from the population of 164 scores.

We see that the true standard errors confirm the bootstrap conclusion that the ordinary mean is the estimator of choice in this case. The main point here is that the bootstrap estimates use only the data in Table 1, while the true standard errors require extra data that usually isn't available in a real data analysis problem.

2. Growth Curves

Longitudinal data are often modeled using growth curves. Our example uses educational testing data from students in North Carolina (see Williamson et al. 1991, Rogosa and Saner 1995). The data we use for the example are eight yearly observations on achievement test scores in mathematics Y for 277 females, each followed from grade 1 to grade 8. The straight-line growth curve model for each individual can be written as

$$Y_p(t) = \zeta_p(0) + \theta_p t + \varepsilon_{pt} \quad (2)$$

where the observable scores Y contain error of measurement indicated by ε_{pt} . Here t denotes time and p denotes person. The constant rate of change for each person is indicated by θ_p ; θ_p has a distribution over persons with mean μ_θ and variance σ_θ^2 .

To illustrate bootstrap estimation we pick out two of the questions that are investigated using the growth curve models:

(a) Can individual change be measured reliably? In behavioral and social science research the reliability index for the estimated rate of change, $\rho(\hat{\theta})$, is the parameter of interest.

(b) Is there a relation between initial standing (math achievement in grade 1) and the rate of improvement? The parameter to be estimated is correlation between θ and initial status (over the population of individuals), written as $\rho_{\theta\zeta(1)}$.

Maximum-likelihood estimates of $\rho(\hat{\theta})$ and $\rho_{\theta\zeta(1)}$ can be obtained from standard mixed model analyses. Standard errors and interval estimates for these quantities are not available from widely-used computer packages.

Figure 3 shows histograms for 4000 bootstrap estimates of $\rho(\hat{\theta})$ and $\rho_{\theta\zeta(1)}$. Bootstrap samples were constructed by resampling individuals (where the data for each individual is a vector of the eight observations). The 90 percent confidence intervals were obtained using the so-called 'percentile method': the endpoints are the 0.05 and 0.95 percentiles of the bootstrap histograms. (The percentile method is the simplest of a number of methods for confidence interval construction via the bootstrap: a generally better method is the 'BC_a'; see Efron and Tibshirani

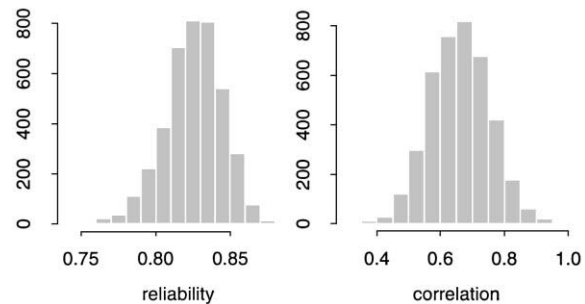


Figure 3

Histograms of 4000 bootstrap estimates of $\rho(\hat{\theta})$ (left) and $\rho_{\theta\zeta(1)}$ (right).

Table 2

Results for growth curve example

Parameter	Estimate	Standard error	90% CI
Reliability $\rho(\hat{\theta})$	0.828	0.019	(0.792, 0.855)
Correlation $\rho_{\theta\zeta(1)}$	0.651	0.093	(0.510, 0.814)

1993, chap. 12–14. The results in Table 2 show good precision for estimation of the reliability, but relatively less precision for the correlation parameter.

3. Related Methods

There are a number of important related methods, with strong similarities to the bootstrap. The *jackknife* systematically deletes one or more observations from the dataset, and applies the statistic to the remaining data. By doing this for each observation or groups of observations, one can obtain estimates of standard errors and bias that roughly approximate to the bootstrap estimates.

Cross-validation is similar to the jackknife but is used to assess the prediction error of a model. It systematically excludes one or more observations from the dataset, fits the model to the remaining data, and then predicts the data that was left out. Application of the bootstrap for estimating prediction error is not straightforward and requires some special modifications. With these, the bootstrap and cross-validation have similar performance for estimating prediction error. Details and references are given in Efron and Tibshirani (1993) and Efron & Tibshirani (1997).

There is a connection of the bootstrap to Bayesian inference. Roughly speaking, the bootstrap distribution approximates a nonparametric posterior distribution, derived using a noninformative prior. Rubin (1981) gives details.

A difficult theoretical question is: when does the bootstrap work? There is much research on this question (see, e.g., Hall 1992). In 'regular' problems where asymptotic normality can be established, first and higher order correctness of the bootstrap (and modifications) have been proven. However, the bootstrap can fail, for example, when it is used to estimate the variance of the maximum or minimum of a sample. More generally, it is difficult to prove theoretically that the bootstrap works in complicated problems: but these are the very settings in which the bootstrap is needed most. Eventually, theory will start to catch up with practice. In the meantime, the bootstrap, like any statistical tool, should be applied with care and in the company of other statistical approaches.

Some popular statistics packages have bootstrap functions, including S-PLUS and Resampling Stats. In most packages that facilitate some customization by the user, it is easy to write one's own bootstrap function.

4. Conclusions

The bootstrap is a powerful tool for data analysis, with many applications in the social sciences. It is being used more and more, as fast computational tools and convenient software become widely available. Good general references on the bootstrap include Efron and Tibshirani (1993), Davison and Hinkley (1997), Hall (1992), Shao and Tu (1995), and Westfall and Young (1993).

Bibliography

- Davison A C, Hinkley D V 1997 *Bootstrap Methods and their Application*. Cambridge University Press, Cambridge, UK
- Efron B 1979 Bootstrap methods: Another look at the jack-knife. *Annals of Statistics* 7: 1–26
- Efron B, Feldman D 1991 Compliance as an explanatory variable in clinical trials. *Journal of the American Statistical Association* 86: 9–26
- Efron B, Tibshirani R 1993 *An Introduction to the Bootstrap*. Chapman and Hall, London
- Efron B, Tibshirani R 1997 Improvements on cross-validation: The 632+ bootstrap method. *Journal of the American Statistical Association* 92: 548–60
- Hall P 1992 *The Bootstrap and Edgeworth Expansion*. Springer-Verlag, Berlin
- Huber P 1964 Robust estimation of a location parameter. *Annals of Mathematics and Statistics* 53: 73–101
- Rogosa D R, Saner H M 1995 Longitudinal data analysis examples with random coefficient models. *Journal of Educational and Behavioral Statistics* 20: 149–70
- Rubin D 1981 The Bayesian bootstrap. *Annals of Statistics*: 130–134
- Shao J, Tu D 1995 *The Jack-knife and Bootstrap*. Springer-Verlag,
- Westfall P H, Young S S 1993 *Resampling-based Multiple Testing: Examples and Methods for p-value Adjustment*. Wiley

Williamson G, Appelbaum M, Epanchin A 1991 Longitudinal analyses of academic achievement. *Journal of Educational Measurement* 1: 61–76

B. Efron, D. Rogosa, and R. Tibshirani

Research and Development in Organizations

1. Increasing Division of Labor in Knowledge Production and Use

The growth of research and development activities (R & D) reflects the increasing division of labor in the production and use of scientific and technological knowledge—a process foreseen by Adam Smith at the beginning of the Industrial Revolution. Since then, professional education, the establishment of laboratories, and improvements in techniques of measurement and experimentation have progressively increased the efficiency of discovery, invention, and innovation (Price 1984, Mowery and Rosenberg 1989). Three complementary forms of specialization have developed in parallel, each contributing to the growth of R & D.

1.1 Specialization by Discipline: The Growth of Science and Engineering

First, progress and specialization in scientific disciplines accelerated in the nineteenth century with the development of more powerful research tools and techniques. Increasing opportunities for application also led to the emergence of the engineering disciplines: hence chemical engineering in addition to applied chemistry, and electrical engineering in addition to applied physics (Rosenberg and Nelson 1994). The differences in timing of the emergence of new opportunities for innovation reflected radical but uneven improvements in the knowledge base underlying technological change, and in particular the emergence of new technological paradigms (Freeman et al. 1982, Dosi 1982).

The mechanical paradigm was the basis of the Industrial Revolution. It did not grow out of contemporaneous scientific advances, but of improvements in the quality of metals and the precision with which they could be formed and machined. These enabled the *design, development, and testing* of prototypes of families of machines with greatly improved performance based on materials of homogeneous and predictable quality, and on shapes with compatible sizes. Similarly, the initial improvements in metals processing technologies depended less on scientific understanding than on *development*

activities, experimenting with different alloys and processing conditions in pilot plant. Even today, such development activities in prototypes and pilot plant typically account for about three quarters of the R & D expenditures of private business firms and for about two-thirds of total R & D.

1.2 Specialization by Corporate Function: The Growth of Business R & D

The science-based chemical paradigm (based increasingly on synthetic chemistry) and the electrical paradigm (based on electromagnetism and radio waves) laid the basis for the industrial R & D laboratory, initially in the chemical and electrical firms in Germany and the USA. Many of the initial applications—like the use of techniques of chemical analysis to control the quality and composition of materials in the increasingly large-scale processing industries—were relatively mundane, but very profitable in reducing costs. Management learned progressively that the science-based technological opportunities were applicable over a range of existing and new product markets, and therefore opened up opportunities for what is now called ‘related’ product diversification. Mowery and Rosenberg (1989) have described the spread of the establishment of R & D laboratories in the USA from the chemical industry to other sectors, and from larger to smaller firms. Similar processes have been at work in Europe, and more recently in East Asia (Caron et al. 1995, Nelson 1993).

Thus, in addition to the benefits of the cognitive division of labor into more specialized fields, the functional division of labor within business firms also augmented the rate of technical change. Corporate R & D laboratories and similar groups devoted full-time to inventive and innovative activities provided an effective method for creating, combining, and coordinating increasingly specialized knowledge. They provided improved and specialized instrumentation, which enabled firms to monitor and benefit more systematically and effectively from advances in specialized academic disciplines. They also created skills in the development and testing of laboratory concepts and prototypes, and the translation into commercialized products. Firms were able to experiment with a wider range of products and processes than had previously been possible when constrained by established products and production lines. In fields rich in technological opportunity, firms have in consequence become multiproduct as well as multitechnological.

Technological paradigms have been associated with the emergence of large dynamic firms that have been successful in exploiting the new opportunities. The largest R & D spenders today are in companies that grew with the emergence of the mechanical (and automobile), chemical, and electrical–

electronic paradigms. Aerospace is a special case, having been technologically force-fed, especially since World War Two, by government R & D subsidies and procurement linked to military requirements. The fastest growing R & D spenders today are those closely associated with ICTs and software technology.

1.3 Specialization by Institution: The Growth of Academic Research

In the nineteenth century, de Tocqueville foresaw that the dynamics of capitalist competition would greatly stimulate the development of innovative activities that showed the prospect of a commercial return. He also argued that public authorities would need to support complementary public research of a more fundamental nature, in order to avoid diminishing returns, to open up new opportunities, and to provide trained researchers. This has happened too. In all advanced countries governments have become the main source of the funding of research activities, and related postgraduate training, in universities and similar institutions. Corporate R & D laboratories have come to depend increasingly on a supply of scientists and engineers aware of the latest research results and trained in the latest research techniques (Salter and Martin 2001).

1.4 R & D and Modernization

This historical pattern of specialization and growth in knowledge production has been broadly repeated in the processes of modernization of late-coming countries, and is observable in countries at different levels of development today: R & D expenditures as a percentage of Gross Domestic Product (GDP) increase with GDP per head, both in specific countries over time, and in cross-sections of countries at any point in time. In successful modernizing countries, R & D is preceded by systematic business investments in improvement-generating activities: in particular, in production engineering, quality control, and design activities, initially making minor modifications and improvements, and later becoming the basis for indigenous innovative activities (Lall 1992, Bell 1984). At this stage, business-funded R & D increases rapidly, accompanied by the growth of university-based research in underlying disciplines, with both reflected in the growth in the numbers of international patents and scientific papers. This pattern can be observed most clearly in both South Korea and Taiwan during the 1980s and 1990s.

The centrally planned Soviet model of modernization, practiced in the former USSR and imposed or adopted elsewhere, gave high priority to R & D. By the late 1950s and early 1960s, R & D expenditures in

Central and Eastern European countries were apparently higher than those in Western Europe and the USA. Certain Western observers, therefore, concluded that the Soviet system was superior in promoting R & D and technical change. However, Soviet economic performance subsequently deteriorated, and it later became clear that a very high proportion of Soviet R & D was oriented towards weapons development, and that the government-established R & D laboratories established for each industry were decoupled from the requirements of producers and consumers (Hansen and Pavitt 1987). The major reductions in R & D activities since their collapse in 1989 can be seen as painful adjustments to make R & D activities—and the underlying activities in production engineering, quality control, and design—become an integral part of a process of economic modernization (Dyker and Radosevic 1999).

The Soviet system also had inadequate linkages with technical advance in the rest of the world economy. One important feature distinguishing today's modernizing countries from those of the nineteenth century is the availability of more productive technologies in more advanced countries. Countries successful in assimilating these more advanced technologies have had two characteristics (Hobday 1995): first, strong linkages with the more advanced countries' technology whether through inward direct investment (e.g., Singapore), inward technology licensing by locally owned firms (e.g., Japan, S. Korea), or subcontracting agreements from advanced countries to local firms (e.g., Taiwan); second, the development of local, change generating activities, culminating in R & D activities, that were essential inputs to effective imitation, as well as to innovation (Cohen and Levinthal 1989).

2. Measurement of R & D Activities

2.1 Patterns of R & D Activity

By the beginning of the 1960s, initiatives by the US National Science Foundation and the Organization of Economic Co-operation and Development (OECD) had established common definitions of R & D activities, and led to the collection by governments of systematic data on R & D activities. These have since been complemented by privately funded surveys comparing the R & D activities of individual business firms. They show the following common and largely invariant features of R & D activities.

(a) In industrially advanced countries, business and government are the main sources of funds for R & D activities. Business funding is in general larger, and spent in-house mainly on *applied research and development* activities. Government funding is divided between *basic research* performed mainly in university-

type institutes, and various types of R & D associated with health, the environment, defense, and the support of industry and agriculture. This R & D is performed mainly in the laboratories of government agencies and business firms.

(b) In the industrially advanced countries, the share of GDP typically spent on R & D varies between 1.5 and 3.0 percent of Gross Domestic Product (GDP), compared to 0.5 and 1.5 percent in the newly industrializing countries, and less than 0.5 percent in the rest. The share of national R & D funded and performed by business firms tends to increase along with GDP per head.

(c) More than 60 percent of all business-funded R & D is typically performed in the chemical, electrical, electronic, and automobile industries. The largest individual corporate spenders on R & D are the world's leading automobile, electrical, and ICT (information and communication technologies) firms.

2.2 The Economic Determinants and Impact of R & D Activities

The improved quality of R & D statistics has enabled economists and other analysts to deepen understanding of both the determinants and the economic impact of R & D activities, at the level of countries, industries, and firms. A number of analysts have been able to show that—amongst the industrially advanced countries—differences in the levels and rates of growth of national R & D activities have a significant influence on differences in national performance in exports and productivity (Fagerberg 1987).

At the industry level, considerable attention has been given to the effects of firm size and industrial structure on R & D activities. Given the dependence of specialization on scale, the proportion of firms performing R & D increases with firm size. However, there is no clear consensus on how R & D intensity varies with firm size amongst large firms: evidence can be found either way. And although R & D is found mainly in concentrated industries, there is compelling evidence that variations amongst industries in both the degree of concentration and R & D intensity are determined jointly by a third factor, namely, inter-industry variations in the extent of technological opportunities (Levin et al. 1985).

Less progress has been made so far in measuring the effects of differences in R & D expenditures on company performance. This is partly because comprehensive and comparable R & D data at the company level are only slowly becoming available. It is also because of the difficulty of defining a proper measure of corporate performance. Those used include the R & D production function, stock market evaluation, and long-term growth (Patel 2000).

2.3 Limitations

Sections 2.1 and 2.2 and show that statistics on R & D activities can be a useful proxy measure for innovative activities. But they have their limitations, the most often mentioned of which is that they measure inputs and not outputs. Equally, if not more important are the following:

(a) In large firms, R & D statistics do not measure the often considerable expenditures on complementary activities in production and marketing, that are necessary to transform the outputs of R & D into commercially successful innovations.

(b) In small firms, they do not measure part-time innovative activities, which are particularly important in the machinery and software sectors.

(c) They capture only very imperfectly the growing volume of innovative activities in software now being performed in firms in service sectors, such as finance and distribution.

Other indicators of knowledge-generating and innovative activities have therefore been developed to complement R & D indicators, the most important of which have been counts of patents, papers, citations, and numbers of innovations (van Raan 1988).

3. Some Contemporary Debates in R & D Management

Many of the characteristics and problems of managing corporate R & D have changed very little since the nineteenth century.

(a) The need to orchestrate and integrate specialized knowledge and skills across disciplines, professions, functions, and divisions.

(b) The progressive improvement in fundamental scientific understanding, accompanying the increasing systemic complexity of innovations—a process that will continue with the more widespread application of the techniques emerging from ICT.

(c) The difficulty (impossibility?) of making reliable predictions about the success or otherwise of specific innovations, especially major ones.

At the same time, new problems and challenges have emerged.

(a) With increasing specialization in both the production of knowledge, and of the components of increasingly complex systems, firms are faced with increasingly difficult choices about what products and knowledge to outsource, and what to retain in-house. An important part of the in-house R & D function is now devoted to the monitoring and co-ordination of innovative and production activities external to the firm.

(b) In addition to the well-established function of supporting foreign production, large firms are performing a growing share of their R & D outside their home country, in order to tap into the increasingly

numerous international sources of leading-edge scientific and technical advance. This poses new challenges for managers in integrating skills and knowledge over long distances (Niosi 1999), and for national policy-makers in assessing the location of the benefits of public investments in academic research.

(c) There is some evidence that academic research is becoming increasingly linked to commercial R & D activities. For some, this is the consequence of unwelcome financial pressure from governments on universities to demonstrate short term 'relevance' in their research. For others, it is the consequence of fundamental changes in the locus of knowledge production. Other evidence suggests that reductions in the costs of technical experimentation through simulation software now make it easier for academic researchers to develop and test experimental prototypes.

See also: Competitive Strategies; Organizational; Corporate Culture; Design and Form; Organizational; Information and Knowledge; Organizational; Innovation and Technological Change, Economics of; Innovation, Theory of; Intelligence; Organizational; Intelligence, Prior Knowledge, and Learning; Learning; Organizational; Organizational Behavior; Psychology of; Organizational Decision Making; Organizations, Metaphors and Paradigms in; Organizations: Unintended Consequences; Strategic Intelligence; Technological Innovation; Technology and Organization; Technology Transfer

Bibliography

- Bell M 1984 Learning and the accumulation of technological capacity in developing countries. In: Fransman M, King K (eds.) *Technological Capability in the Third World*. St Martin's Press, New York
- Caron F, Erker P, Fischer W 1995 *Innovations in the European Economy between the Wars*. W de Gruyter, Berlin
- Cohen W M, Levinthal D A 1989 Innovation and learning: The two faces of R & D. *Economic Journal* 99: 569–96
- Dosi G 1982 Technological paradigms and technological trajectories: A suggested interpretation of the determinants and directions of technical change. *Research Policy* 11: 147–62
- Dyker D, Radosevic S (eds.) 1999 *Innovation and Structural Change in Post-socialist Countries: A Quantitative Approach*. Kluwer, Dordrecht, The Netherlands
- Fagerberg J 1987 A technology gap approach to why growth rates differ. In: Freeman C (ed.) *Output Measurement in Science and Technology: Essays in Honour of Y. Fabian*. North-Holland, Amsterdam
- Freeman C, Clark J, Soete L 1982 *Unemployment and Technical Innovation: A Study of Long Waves and Economic Development*. Greenwood, Westport, CT
- Hansen P, Pavitt K 1987 The Comparative Economics of Research Innovation in East and West: A Survey. In: *Fundamentals of Pure and Applied Economics*. Harwood, Chur, Switzerland, Vol. 25

- Hobday M 1995 *Innovation in East Asia*. Elgar, Brookfield, VT
- Lall S 1992 Technological capabilities and industrialisation. *World Development* 20: 165–86
- Levin R, Cohen W M, Mowery D C 1985 R & D appropriability, opportunity and market structure: New evidence on some Schumpeterian hypotheses. *American Economic Review* 75(2): 20–4
- Mowery D, Rosenberg N 1989 *Technology and the Pursuit of Economic Growth*. Cambridge University Press, Cambridge, UK
- Nelson R (ed.) 1993 *National Systems of Innovation: A Comparative Study*. Oxford University Press, Oxford, UK
- Niosi J 1999 The internationalisation of industrial R & D: from technology transfer to the learning organisation. *Research Policy* 28: 107–17
- Patel P 2000 Technological indicators of performance. In: Tidd J (ed.) *Measuring Strategic Competencies*. Imperial College Press, London
- Rosenberg N, Nelson R 1994 American universities and technical advance in industry. *Research Policy* 23: 323–48
- Salter A, Martin B 2001 The economic benefits of publicly funded basic research: A critical review. *Research Policy* 30: 509–32
- Van Raan A 1988 *Handbook of Quantitative Studies of Science and Technology*. North-Holland, Amsterdam

K. Pavitt

Research Conduct: Ethical Codes

This article will focus on the use of ethical codes in dealing with moral problems concerning research involving humans, nonhuman animals and the environment, that is, moral problems that arise mainly in the experimentally working sciences, especially biomedicine. More general issues such as truth telling, or usage of best scientific methods that if neglected may cause moral problems, will not be considered here.

1. History of Ethical Codes of Research Conduct

The (ab-)use of humans for research purposes has a long tradition reaching back far beyond the beginnings of modern experimental science in the Renaissance to ancient Egypt and Persia. Unfortunately, morally questionable experiments were also performed in the recent past. Well known are the atrocities perpetrated by physicians during the Nazi regime in Germany during the 1930s and 1940s. But even later, contemptible experiments were uncovered elsewhere (Beecher 1966): an inglorious example is the US Public Health Service Syphilis Study (better known as Tuskegee Syphilis Study) in the course of which a large number of black men diagnosed with syphilis had been left untreated in order to observe the ‘natural’ course of the disease.

Although the abuse of human experimental subjects is by no means the only moral issue in research

conduct (see Sect. 3), it gave rise to the first international moral framework regulating research conduct—the Nuremberg Code of 1947. It was followed by the World Medical Association’s (WMA) *Declaration of Helsinki* in 1964 (with several revised versions), and in 1993 by the *International Ethical Guidelines for Biomedical Research Involving Human Subjects*, which was drafted in cooperation by the Council of Medical Organizations of Medical Science (CIOMS) and the World Health Organization (WHO). Besides these widely adopted international guidelines there is a large number of further international and national guidelines. Although regulating research conduct was for a long time equivalent to regulating research on human subjects (for an overview see Veatch 1995), the debate on animal rights as well as on the protection of the environment has also stimulated the issuing of many guidelines covering these areas. (A large selection of relevant texts, including those named above, are reprinted in Spicer 1995.)

2. Ethical Codes—Philosophical Remarks

A moral, or an ethos, is understood as a set of rules of action (norms). Ethics, a scientific subdiscipline of philosophy, is the critical assessment of different morals. There are many morals (e.g., the ethos of physicians, the ethos of lawyers, etc.) but only one ethics; the latter should not, however, hide the fact that there is no agreement between ethicists as to what the correct criteria for this assessment are (see *Bioethics: Examples from the Life Sciences*).

A moral or ethos can exist in unwritten or (partly) written form. An ethical code is a written, i.e., codified, moral or ethos; an ethical code is, therefore, a moral and not, as its name might suggest, a part of ethics. Insofar as oaths (such as the Hippocratic oath), prayers, and creeds present a certain moral, they are open to an ethical assessment too.

An ethical code is an instrument of self-regulation established by a certain profession (e.g., physicians’ organizations) or those groups directly involved in the actions of that profession (e.g., patients or funding bodies). In contrast, national laws and international conventions (e.g., the Convention on Human Rights and Biomedicine adopted by the European Council in 1996) are issued by government bodies. Using the principle of subsidiarity, i.e., the principle that what can be regulated on a lower level should not be regulated on a higher level, ethical codes allow an increased flexibility in view of new moral challenges when compared to statutory regulations.

An ethical code of research conduct is a codified set of rules of action (norms) that should be adopted by every researcher to whom the code is addressed. If a researcher obeys an ethical code he or she is said to

perform her research in a morally acceptable way. 'Performing morally acceptable research' is relative to the justified ethical code applied in testing a certain piece of research. It might be the case, however, that the evaluation of a certain act of research is morally acceptable relative to one ethical code but not to another. For example, a medical experimenter by performing a certain act of research (e.g., on human embryos) might obey the ethical code to which he is subject in his country, but by the very same act might disobey the ethical code obligatory in another country. In those cases where the differing evaluations of acts of research is deemed unacceptable, the philosophical evaluation of ethical codes has its place. The main areas of interest in this context are the application of results from research, the generation of knowledge through research, and the credibility of researchers. Conflicts may arise due to the incompatibility between—as already noted—different ethical codes, certain parts of a certain ethical code, or an ethical code and what might be called public morals.

The ethics of science deals with the reconstruction (and thereby critical assessment) of rules of action, whether codified or not. This reconstruction displays presuppositions concerning the scope of the research in question, the criteria for the 'scientific' generation of knowledge, and the acceptable application of this knowledge. (Motivation of research, be it longing for social acceptance or money, does not play any role in the ethics of science; only prescriptive claims of validity related to certain acts of research are at issue here.) Ethical codes of research conduct can be subjected to such a reconstructive process. The task of ethics is to make distinguishable by way of reconstructing those moral norms that can be universalized, in other words, that are valid for everybody. This assessment should be done by using a method that is, in principle, accessible to everyone. Those moral norms that have proven to be valid can then justifiably be used to test an act of research with regard to its moral acceptability.

The application of moral rules of action in research contexts—the testing of certain acts of research concerning their norm-conforming character—is easily established in small social groups only. For example, working together in a laboratory on a daily basis, the members of the peer group develop a strong intuition of whether certain acts of research do obey the relevant rules of action or not. Acts that do not conform to the rules of action are, therefore, easily identified and the committing actor is kept from doing those acts again. This habitualization of rules of action seems to be, therefore, an appropriate measure of enforcing rules of action in small social groups (Gethmann 1989).

In complex scientific communities with a highly differentiated division of labor the habitualization of moral norms does not work satisfactorily, because the peer review between researchers cannot be sufficiently secured. In this situation it is necessary to design

procedures that achieve the control of obedience to moral norms and their enforcement (institutionalization): though not everyone can evaluate whether a certain experiment is morally acceptable (according to certain standards), because not everyone has the expert knowledge necessary for that evaluation, everyone—at least in principle—should be able to assess whether the procedure used by an institutional review board for evaluating this research (e.g., by a peer-reviewed journal, where the review is done by making use of certain rules of action) is correct. Or to give another example: not everyone can assess whether a certain clinical study obeys the relevant rules of action, but everyone should be able—at least in principle—to assess whether the procedure by which the responsible ethics committee evaluates this study is correct (see also *Ethical Practices, Institutional Oversight, and Enforcement: United States Perspectives; Ethics Committees in Science: European Perspectives*).

The codification of moral norms, that is the generation of ethical codes, runs largely in parallel with the institutionalization of these norms. Codifying norms enhances their ability to be put into operation: codified norms are less open to inter-individual interpretability of their meaning than norms that are subject to an oral tradition. This stability of meaning supports the efficiency of institutionalized procedures of control of obedience to and enforcement of those norms (cf. the legal requirement of writing for certain documents).

Connected to this is the issue of the status of obligation of ethical codes: statutory regulations must be obeyed (with the problematic exception of toleration of illegal acts or illegal acts that are exempt from punishment). In contrast to that, ethical codes can have a status of obligation varying between a mere declaration of intent and binding professional law. Depending on the act at issue, the consequences of non-obedience of an ethical code may vary from a mere moral reproof to a prohibition from practicing a profession. In addition, the status of obligation of ethical codes may vary over time. This can be illustrated by the *Declaration of Helsinki*: issued by a professional organization, the World Medical Association, it was originally seen as an ethical code binding only for the members of the WMA. Today, however, the Declaration has gained a status comparable to the United Nations' Universal Declaration of Human Rights, to the extent that even non-members of the World Medical Association cannot simply depart from the rules of action it lays down. There is a case, therefore, for transferring the ethical code into an international convention or even law issued by the United Nations.

Finally the question arises, by what procedures should ethical codes be developed and changed. Since the development of ethical codes starts from consensus statements, many codes have been blamed for being developed without a wide enough consultation of the relevant stakeholders. It is claimed, for example, that

codes with an allegedly international validity are negotiated mainly by representatives from Western countries.

Taking into account the respective opinions of the relevant stakeholders is an important step in the preparation of an ethical code. But, as said before, the task of a critical assessment of an ethical code or its draft versions is to single out those moral rules of action that can be universalized, i.e., are valid for everybody. This means, however, that the assessment and its results should be independent of the person, and their background, who is actually performing the assessment. Therefore, procedures for developing ethical codes do not make it necessary to involve all stakeholders; even involving only a very few experts can lead to an ethically justified result. But reflecting on the preconditions of a procedural development of justified norms should not lead to a neglect of the difficulties in the practical implementation of these procedures: the debate on changes to international ethical codes like the *Declaration of Helsinki* involves highly political issues such as conflicts of interest, where the involved parties persist in their claims, be they justified or not. Deciding the question of who should participate in the negotiation of ethical codes is, therefore, an important task in addition to the purely ethical assessment of the rules of action laid down in these codes (Schüklenk and Ashcroft 2000).

3. Major Issues of Ethical Codes of Research Conduct

Research bears chances and risks for, among others, patients, society, future generations, nonhuman animals, and the environment (for an introduction see Brody 1998). An important issue of ethical codes is to state criteria relative to which the weighing of these chances and risks can be achieved. An extensively discussed example is the concept of informed consent, which features prominently in many medical codes of research conduct. The lesson drawn from the history of medical research during the twentieth century led to a broad consensus that research without informed consent of the research subject—roughly, a decision made understandingly, intentionally, and without controlling influences—is highly problematic from the moral perspective. Taking this into account the Nuremberg Code demands that in research on humans ‘the voluntary consent of the human subject is absolutely essential,’ virtually making impossible research on those with diminished decision-making capacities. In contrast to that, the *Declaration of Helsinki* decreased the importance of informed consent and introduced ‘best interest’ criteria that are supposed to regulate research on incompetent research subjects (e.g., demented or comatose patients). Though vastly treated in the literature, the issue of

whether informed consent is a necessary precondition of any research—and if yes, what informed consent actually means—is not satisfactorily solved. Especially in the context of international ethical codes that are designed to supersede national codes, codes that often rest on diverging traditions, this issue remains a controversial subject of debate (see *Research Subjects, Informed and Implied Consent of*).

Another criterion used in ethical codes for the sake of weighing chances and risks is that of benefit. But benefit for whom? Basically the question is whether those affected by research may be burdened with the risks connected to research without benefiting from this research. For example, the debate on the acceptability of nontherapeutic research focuses on this problem: the research subjects in nontherapeutic research are burdened with the risks of this research (like drug side-effects) without benefiting from it (see *Medical Experiments: Ethical Aspects*).

Delivering criteria for weighing the risks and chances of research conduct is not the only topic usually covered by ethical codes. Other topics include fraud and misconduct in research, conflicts of interest, and ownership of results of research (see *Intellectual Property Rights: Ethical Aspects; Research Funding: Ethical Aspects*).

An issue of increasing importance for ethical codes is posed by moral problems connected to research involving nonhuman animals and the environment. The former brings up the question as to whether the burden of such research can legitimately be left to those animals. In particular, the so called animal rights movement has done much to raise the level of attention to these problems (see *Animal Rights in Research and Research Application*).

Although the effects of research on the environment have so far not been the main threat that pushed environmental protection forward, research nonetheless can have morally questionable effects on the environment. The release of genetically modified organisms (GMOs), for example, may endanger biodiversity, in that the number of wild-type plants and animals may be reduced by suppression. Although a large number of living species is by itself not necessarily morally preferable, a reduced biodiversity may have (probably only in the future) adverse effects on the ecosystem and thereby on mankind. This would mean, however, that the potential benefits of GMOs are allocated to the living, while the potential burdens due to a damaged environment are left to future generations. Besides postulating rules for actions that demand consideration in dealing with the environment insofar as it is the ‘subject’ of research, ethical codes may also contain the demand that research actively seeks to improve environmental protection.

See also: Litigation: Effects on Medical Research

Bibliography

- Beecher H K 1966 Ethics and clinical research. *New England Journal of Medicine* 274(24): 1354–60
- Brody B A 1998 *The Ethics of Biomedical Research. An International Perspective*. Oxford University Press, New York
- Gethmann C F 1989 Proto-ethics. Towards a formal pragmatics of justificatory discourse. In: Butts R E, Brown J R (eds.) *Constructivism and Science. Essays in Recent German Philosophy*. Kluwer, Dordrecht, The Netherlands, pp. 191–220
- Spicer C M (ed.) 1995 Appendix. Codes, oaths, and directives related to bioethics. In: Reich W T (ed.) *Encyclopedia of Bioethics*, 5 Vols. Macmillan, New York, pp. 2599–842
- Schüklenk U, Ashcroft R 2000 International research ethics. *Bioethics* 14(2): 158–72
- Veatch R M 1995 Medical codes and oaths. History. In: Reich W T (ed.) *Encyclopedia of Bioethics*, 5 Vols. Macmillan, New York, pp. 1419–27

F. Thiele

Research: Ethical Aspects of Long-term Responsibilities

The development of the modern sciences in the twentieth century brought about unexpected consequences for humankind both of a welcome and an unwelcome nature. The ethical issue of long-term responsibility in research takes on particular importance in view of the potential threat to future generations posed by the far-reaching and unforeseeable effects of realizing the possibilities of science. Yet our experience of the unforeseeable nature of innovation processes and the sheer impossibility of foreseeing the life circumstances of future (especially of distant) generations gives us reason for serious doubts about whether it is at all reasonable to speak of long-term responsibility in reference to research. Thus affirmation of this must be founded upon an enlightened ethical analysis of long-term responsibility. If, indeed, such responsibility is to be affirmed, it remains to be clarified who shall bear this responsibility: the researcher, the user, or those who benefit from the results of research. Finally, it may be necessary to determine how long-term responsibility should be implemented within the reality of societal decision-making processes.

1. Scientific Aspects

The discipline of the ethics of science is concerned with the systematic reconstruction of those specific orientations for action that are determined by the understanding, immanent in each field of science in respect of its individual subject area, and the understanding of the scientific procedures for its description and explanation and an appreciation of the possibilities for

action opened up by scientific knowledge. In this, ethics of science shall be understood as a part of philosophical ethics that refers to a specific, societal field of action determined by particular forms of knowledge. Above all, ethics of science, in respect of the task on hand, has to deal with two interrelated areas of phenomena. First, ethics of science draws on the specific ethos of the scientific community in order to reconstruct those orientations to which the scientist is committed, in the interest of seeking out the truth. This involves a smooth, unbroken transition from general rules of action (e.g., the prohibition of falsifying research results) to specific regulations for individual disciplines within the framework of their respective methodology. Second, ethics of science concerns itself with the relationship between general moral orientations and the problems of generating and applying scientific knowledge. Both in generating and applying, it is primarily a question of the practical consequences resulting from technical knowledge.

An example of the moral problems involved in the generation of scientific knowledge can be seen in the field of experiments on human beings in medical research and other medical science disciplines (above all in the field of psychology); more recent discussions target experiments on animals or genetic intervention into the genome of humans, animals, plants and micro-organisms, for the purpose of getting to know more about genetics. The moral problems inherent in the application of scientific knowledge and the responsibility of the scientists concerned are now under discussion in almost every field of modern science, and have been especially since the 'Manhattan Project' (the building of the atom bomb by US physicists).

The beginning of modern science, at its core, has moved away from the concept of knowledge generated by contemplation, since it is now essentially characterized by knowledge that owes its existence to technical intervention into nature under controlled conditions. Not until a fundamental appreciation of the problems of a modern concept of knowledge was undertaken by Kant and other German proponents of idealism did it become clear that this, too, is thanks to moral presuppositions. Whereas initially it was not the technical consequences of the application of such knowledge that was in the foreground but rather the self-commitment of the scientist to the truth, in the interest of generating reliable knowledge, the discussion on the ethics of science in the twentieth century, in comparison, was first and foremost determined by the experience of the technical and practical consequences of scientific knowledge. In this it was predominantly the natural scientists, in the wake of the development of weapons of mass destruction and other large-scale technical applications of knowledge, who posed the question as to the moral principles of their actions. Above all, with the development of the modern biosciences, the question of the moral responsibility involved in the generation of knowledge

has become the focus of attention. The problems arising especially in the medical disciplines and the biosciences are proving to be highly complex issues, involving the fundamental ethical orientation of societies characterized by scientific and technical civilization.

2. Fundamental Problems of Long-term Responsibility

As far as ethics is concerned, the idea that human beings not only bear responsibility for their actions towards those with whom they are faced in the 'here and now' is by no means a new idea ensuing from problems arising from modern science and technology. Nevertheless, the issue of obligation beyond the immediate circumstances of our actions is dramatically brought home to us by modern science and technology in a manner hitherto unknown, in that modern technology generates for successive generations qualitative and quantitative risks—but also chances—unknown to premodern technology, and this not only when actions are undertaken but also when they are omitted.

It is an indisputable fact that present-day actors expect the acceptance of a kind of risk that earlier generations never had to face. In the sense that this is a quantitatively and qualitatively new situation, it may be stated that humankind is now confronted with a new ethical problem. Thus the question of risk is materially the core question when considering the issue of the scientist's responsibility for the future. Taking actions involving risks, however, is not in all cases conducive to conflict and thus is not always of ethical relevance. It is only of interest in those cases where others are expected to accept the taking of actions that involve risks. As far as and as long as the consequences of the actions concerned involve risks that are of a purely private nature, e.g., because (a) the consequences only effect the actors themselves; or (b) the consequences are not recognizable due to insufficient knowledge concerning the relevant causal or conditional nexus; or (c) though the consequences are foreseeable, the persons concerned, as moral subjects, are not deemed equals to the actors, so that as far and as long as this position is held, taking actions involving risks is not an ethical problem; it may be reduced to the problem of acting with predetermined consequences.

The ethical problems of taking actions involving risks are now exacerbated even more by a further dimension, i.e., when it is a matter of the consequences of technical activities for coming generations in the distant future. These are the guiding problems in the issue regarding long-term obligation. Thus, in principle, it should not be necessary to demand a new code of ethics but rather to attempt to come to terms with the 'at any time' concept of our obligation as seen in

existing ethics. The question is to be asked more precisely as to what it means for human beings to be under obligation to other human beings in the distant future in view of the unforeseeable long-term consequences of their actions in applying the means of modern scientifically founded technology.

In issues of applied philosophy, like the question of responsibility for climate consequences, for the problems of the final disposal of nuclear waste, for the consequences of intervention in the human genome or for population growth, the problem of obligation and responsibility mainly concerns generations in the distant future. This entails asking a series of nontrivial questions that have to be answered if one seeks to justify the reason for and nature of such long-term responsibility. In this connection it is of primary importance to answer the following three fundamental questions:

(a) Are actors only under obligation toward those with whom they are concurrently interacting?

Only if the answer to this question is negative can there be any talk at all of obligation that extends beyond the actor's present into the future. This question is to be answered in the affirmative by proponents of such ethical concepts from which obligations arise exclusively from claims uttered by currently existing conscious beings (this primarily implying preference-oriented utilitarian approaches). Within the framework of such concepts the paradox arises that for members of a generation G_0 in relation to members of future generations G_n no obligations will ever be able to exist if the future generations are in reality not in a position actually to raise claims against the members of G_0 , which after sufficient passage of time (approx. $n > 3$) is always the case. In respect of concepts that reduce obligations to the actual voicing of claims of this nature it must be pointed out that there is a distinction between having the right to make claims on the one hand and the ability in reality to assert such claims on the other.

However, if future generations have justified claims against earlier generations (even when they are not yet able to assert them), then, in appropriate cases, corresponding obligations are incumbent upon the earlier toward the later generations.

(b) Does obligation end with generation $k > i$ (e.g. where $i = 3$)?

Everyday intuition would seem to suggest what many economists profess, namely that due to the lack of knowledge concerning the needs of future generations, which increases with the growing distance in time, obligations extending into the future must end with a generation of a certain, relatively large magnitude of k . However, if one attempts to pin down this intuition precisely, it does, in fact, begin to seem implausible: on principle to allow obligations to end at certain generation markers would mean in effect that for the members of G_0 in relation to G_{i+1} , in contrast to the members of generation G_i there would be no

obligations whatsoever, and this for the only reason that the former belong to a grade $k > i$ generation. A restriction of this kind on the scope of obligation would be incompatible with an ethical universalism, in principle presupposed both by obligation ethic as well as more recent approaches to an utilitarian ethics. If one grants recognition of moral justification to moral subjects at all, then it seems simply arbitrary to allow this to end at a particular time (or spatial) limit. To allow obligations to end with however near or distant a generation would be particularism of the present, an 'ego-preference' of those consciously concerned at the time. And in any case, such a form of particularism would be in contradiction to cognitive universalism, which is also inherent, so to speak, as a sort of regulative idea, in the search for scientific truth.

(c) Are actors under the same degree of obligation to members of the k th generation as they are to the first generation succeeding them?

The conclusion is often drawn from an affirmative answer to question (b) that the answer to question (c) must also be affirmative: Either there exists an obligation on the members of generation G_k or there exists no such obligation—but if it does exist, then it does so for the members of generation G_k just as for the members of generation G_{k-1} , G_{k-2} , ... G_1 . This consequence, however, is not inevitable. On the contrary—irrespective of the question of temporal distance between those under obligation and those entitled to make a claim—a distinction must always be made between the existence of an obligation, on the one hand, and the degree of real liability on the other. Correspondingly the gradation of obligations must be sharply distinguished from the concept of discounting liabilities. Whereas, in fact, the idea of discounting (at any rate in most of the variants put forward) presupposes an ego-preference that is not compatible with universalism and which is therefore rightly rejected by most moral philosophers, the idea of a gradation of obligations, on the other hand, is unaffected by the presupposition of ego-preferences of those consciously involved at the time.

3. Social Implementation of Long-term Responsibility in Research

Obligations come into existence through discursively justified claims. In the case of long-term responsibility, however, the very element that is missing is the moral proponent. It can merely be introduced as an entity not yet asserting claims, albeit entitled to make claims. In such cases the process of justifying claims and obligations is therefore not based on real but on fictitious discourse. It must be said, of course, that fictitious discourses are not confined to long-term responsibility. Even in the case of short-term responsibility, moments of fictitious discourse are a

matter of course, e.g., concerning forms of delegated, anticipated and representative obligations to act. The fictitious discourse has to be conducted as though the proponent were actually raising claims (with all the difficulties involved in the growing distance between the generations).

It is just as difficult to reconstruct the discursive arguments of the proponent as it is to reconstruct those of the opponent, i.e., the addressee of the claims introduced. Whereas the real proponents bring themselves to the notice of the real opponents, it can easily happen in the case of long-term responsibility that no one feels addressed by the issue. Thus the question must be posed as to who exactly shall be obliged to bear the long-term responsibility. Also for this question, the starting point is short-term responsibility because here, already, there is the problem of determining the addressee.

One of the first conclusions that can be drawn is that an answer like 'Everyone shall take care of their own claims' cannot constitute a satisfactory solution, since actions likely to result in conflict (distribution of resources, exercising usufruct etc.) are the very ones that do indeed concern other people. In respect of long-term responsibility, however, a special case arises in that 'the others' are not known to the actors and as yet may not even exist. Indeed, however, the 'everyone for themselves' solution already comes up against insurmountable barriers in connection with short-term responsibility.

Individuals who have come to the conclusion that they cannot take responsibility for the application of chlorinated carbohydrates as a propellant gas, that individually this conclusion is being justified, nevertheless fail to achieve the aims of acting because they are not in a position to evade the use of CFCs. Such examples can also be cited from the field of large-scale technical plant.

Examples taken from the field of short-term responsibility show that in addition to individual obligation there is also a responsibility of the kind that devolves upon the actors since 'all' the addressees constitute a plea and thus the claim which arises from it. In short it shall be referred to as 'collective responsibility.' Of course, this only indicates the direction of the solution. Collective responsibility could be regarded as a concept of escapism from obligation. It must therefore be clarified in what manner the validity of collective obligations may be asserted in real discourses.

For the further clarification of this question, it is necessary to make a terminological remark. A distinction must be made between obligation arising from an action and the responsibility for an action. Actors are subject to direct and undivided obligation for their actions, and responsibility is borne by them according to delegation of collective obligation. Responsibility arises from obligation through 'moral division of work.' Thus there exists not only the productive

division of labour (in the economy) and the cognitive division of work (in the field of science). A moral division of work is especially applicable to those moral problems that exist in connection with the weighing of risks and chances in respect of the possibilities to act and the compulsions and constraints pertaining to actions in the context of modern science and technology. Due to the complexity of these problems the obligations of all concerned towards future generations must be delegated to the responsibility of individuals.

The first solution that springs to mind in this respect is the creation of bodies of scientific experts. In this, however, the problem of (non)acceptance arises which manifests itself in mistrust towards a perceived 'expertocracy.' According to the degree in which obligations are delegated to those bearing responsibility there is naturally a growth of uncertainty as to whether those responsible are (in generalizable terms) properly assuming their responsibility.

Thus it constitutes a basic problem of modern scientifically and technically based civilization to organize the social assumption of responsibility in such a way that, on the one hand, the effect of preventing an overburdening is maintained through the process of delegating obligation, while, on the other hand, the moral competence of the decision maker remains transparent. Although not every action is relevant in respect of distant consequences, the question must be asked in those areas that are sensitive to such distant consequences as to which social agencies are in a special way capable of, and therefore should be entrusted with, bearing long-term responsibility.

In this there are many complicated issues that have to be resolved. Above all the question of who should bear the responsibility is particularly explosive at present. Is it primarily a matter for the state and its three main organs of power or is it a matter for society and its subsystems to take over responsibility for the future? The state, through the medium of its legal system, is not a good advocate of long-term responsibility for several reasons. First it is subject to the specific democratic rules governing the lifetime of a parliament and secondly, the functionality of its laws is restrictive. The democratic rules controlling the period in which a parliament stays in power have the effect that the parliamentary system is geared to ultrashort periods of obligation, i.e., the members of parliament, as mandate holders, assume responsibility for four or five years, after which they have to answer to the electorate; the voters, in their turn, have to optimize their decision as to how they should vote with regard to a wide variety of political dimensions. Since the support given by the mandate holders to the idea of long-term responsibility is only one dimension among many, it is likely that for many voters this will be the dominant dimension—as long as voters each have only one unweighted vote.

For a restructuring of the political system, which could be imagined, it would not be sufficient to introduce marginal changes like merely extending the period of parliament or establishing a system of overlapping parliamentary recruitment. In accordance with the principle of diminishing obligation, as adumbrated above, it would be ethically not permissible—and certainly not enforceable either—to abandon the foundation of parliamentary democracy in the (putative) interests of distant generations. For another reason the state is not a good agency to take over long-term responsibility, because its most important medium, the law, is not suitable to regulate subtle moral issues. An essential difference between ethics and law consists in the fact that there is no judge for morally reprehensible actions. Ethics apply rigorously—in contrast to the law; which is why in ethics the saying, 'Where there is no plaintiff, there is no judge' makes no ethical sense at all. Otherwise, there would be no such thing as long-term responsibility since persons who do not yet exist cannot stand up as plaintiffs. Thus there would be no defendants either. A further problem with regard to law in connection with future generations is the difficulty arising from the fact that the relevant occurrence that would be subject to legal punishment has not yet taken place, or may only take place with a certain (possibly very small) degree of likelihood. It can be seen clearly that principles of law are inapplicable here. Furthermore, in respect of the relationship between law and morals the principle of subsidiarity applies, i.e., if something can be accomplished by moral discourse, no attempt should be made to regulate it by law.

Without the necessity of excluding the possibility that in many issues the law may be utilized as an instrument in establishing the long-term responsibility of the members of society, these reflections show that when it comes to answering the question of who should bear responsibility, the focus of attention naturally turns to society with its moral forming and preserving groups and their respective subsystems. In this process a number of subsystems play a prominent part, such as religion, culture (in the narrow sense of the arts), education, and the economy. The findings of the sciences and their respective applications, however, are only of particular importance from the point of view of long-term responsibility.

Their future perspectives are characterized by being particularly concrete and particularly long-term and are marked by having an especially profound effect on human life as we know it. Therefore, it is a paramount task of the scientific community to promote the capability of institutions of assuming long-term responsibility.

See also: Ethical Dilemmas: Research and Treatment Priorities; Objectivity of Research: Ethical Aspects; Research Conduct: Ethical Codes; Research Ethics,

Cross-cultural Dimensions of; Research Ethics: Research; Research Funding: Ethical Aspects

Bibliography

- Apel K O 1987 The problem of a macroethic of a responsibility to the future in the crisis of technological civilization: An attempt to come to terms with Hans Jonas's 'Principle of Responsibility'. *Man and World* 20: 3–40
- Birnbacher D 1988 *Verantwortung für zukünftige Generationen*. Reclam, Stuttgart, Germany
- Gethmann C F 1989 Protoethics. Towards a formal pragmatics of justificatory discourse. In: Butts R E, Brown J R (eds.) *Constructivism and Science. Essays in Recent German Philosophy*. Kluwer, Dordrecht, pp. 191–220
- Gethmann C F 1993 Langzeitverantwortung als ethisches Problem im Umweltstaat. In: Gethmann C F, Kloepper M, Nutzinger H G (eds.) *Langzeitverantwortung im Umweltstaat*. Economica Bonn, Germany, pp. 1–21
- Gethmann C F 1998 Reason and cultures—life-world as the common ground of ethics. In: Lange H L, Löhr A, Steinmann H (eds.) *Working Across Culture*. Kluwer, Dordrecht, Germany, pp. 213–24
- Gethmann C F 2000 Die Krise des Wissenschaftsethos. Wissenschaftsethische Überlegungen. In: *Ethos der Forschung. Ringberg-Symposium Oktober 1999*. Max-Planck-Gesellschaft, Munich, Germany, pp. 25–41
- Gethmann C F, Kamp G 2000 Gradierung und Diskontierung von Verbindlichkeiten bei der Langzeitverpflichtung. In: Mittelstraß J (ed.) *Die Zukunft des Wissens. XVIII. Deutscher Kongreß für Philosophie (Konstanz 4–8 Oktober 1999)*. Vorträge und Kolloquien, Berlin, pp. 281–95
- Golding M 1972 Obligations to future generations. *The Monist* 1: 85–99
- Höffe O 1993 *Moral als Preis der Moderne*. Suhrkamp, Frankfurt, Germany
- Jonas H 1984 *The Imperative of Responsibility. In Search of an Ethics for the Technological Age*. Chicago University Press, Chicago
- Lehrer K (ed.) 1987 *Science and Ethics*. Rodopi, Amsterdam
- Mittelstraß J 1992 Zur Ethik der Naturforschung. In: Mittelstraß J (ed.) *Leonardo-Welt. Über Wissenschaft, Forschung und Verantwortung*. Frankfurt, Germany, pp. 120–54
- Mittelstraß J 1995 Ethics in science—substance or rhetoric. In: Götschl J (ed.) *Revolutionary Changes in Understanding Man and Society. Scopes and Limits*. Kluwer, Dordrecht, Germany, pp. 269–77
- Norton B G 1978 Future generations, obligations to. In: Reich W T (ed.) *Encyclopedia of Bioethics*. Macmillan, London, pp. 892–9
- Norton B G 1982 Environmental ethics and the right of future generations. *Environmental Ethics* 4: 319–37
- Parfitt D 1983 Energy policy and the further future. The identity problem. In: MacLean et al. (eds.) *Energy and the Future*. Rowman and Littlefield, Totowa NJ
- Passmore J 1974 *Man's Responsibility for Nature. Ecological Problems and Western Traditions*. Scribner, New York
- Shea W R, Sitter B (eds.) 1989 *Scientists and their Responsibility*. Watson, Canton, MA
- Sikora R I, Barry B (eds.) 1978 *Obligations to Future Generations*. Temple University Press, Philadelphia, PA

- Weiss E B 1989 *In Fairness to Future Generations: International Law, Common Patrimony, and Intergenerational Equity*. United Nations University, Tokyo
- Wieland W 1999 *Verantwortung—Prinzip der Ethik?* Winter, Heidelberg, Germany

C. F. Gethmann

Research Ethics, Cross-cultural Dimensions of

For the researcher working in a culturally distant setting, for whom cultural and behavioral differences impinge directly on methodological design, data collection, and everyday life in the field, maintaining ethical equilibrium can be a difficult, never-very-satisfying discipline. Anthropologists are most prone to this situation and have reflected the most about it, but to the extent that other practitioners (political scientists, demographers, geographers, etc.) utilize ethnographic approaches, they, too, expose themselves to significant ethical dilemmas. Faraway places are perhaps somewhat extreme in this regard, but studying ethnic minority groups in one's own society can present particular ethical challenges, especially when the observed practices are illegal or are more legitimately at odds with the tastes and standards of the wider community, to which the researcher typically belongs. The difficulty stems from the researcher's need to balance the demands of multiple moral codes, not only those based on personal and professional values—a requirement common to much scientific investigation, whatever the content and venue—but, in addition, those of the host culture and the laws of its land, and those derived from the interests and cultural traditions of persons to whom the researcher has become emotionally attached and obliged during the course of research. Ethical clarity is not helped when, as frequently happens, some of its elements are invisible or incomprehensible to the researcher—from whom, at least early in the project, but to some extent throughout, blunders are to be expected. Charity from the hosts usually redeems the investigator, but not always, and those are the instances of ethical infirmity which rarely survive as classroom and dinner-party anecdotes in the after years.

1. Changing Ethical Circumstances

In the early days of fieldwork among 'native peoples,' prior to about 1920, ethical issues were not so pressing. For British, French, German, and Dutch researchers of that era, the terms of engagement were largely governed by colonial authorities; in North America and South America, comparable oversight was exercised by internal government agencies, such as the

Bureau of Indian Affairs of the US Department of the Interior. Whatever retrospective moral judgment one might pass on those institutions, they were taken for granted at the time, and were the reality within which ethnographic inquiry was conducted. Considerable research was actually carried out by scientists employed or sponsored by colonial or governmental administrations, though with little effect, apparently, on the policies and procedures of rule. Access to field sites was readily achieved by bona fide researchers, who, so long as they disturbed neither village peace nor administrative purpose, were left alone. Host communities had little power to resist being studied, should they have wished to do so. In any case, such communities could not have understood very fully why the equipment-toting sojourner was there, nor could they have foreseen the consequences of the study for themselves and their posterity.

In those days, too, ethnographic research often consisted of participation in regional survey work conducted by teams of scientific specialists and government officers. Theirs was a 'collecting' mentality, with no need to stay long in one place. The expedition would record cultural inventories (word lists, kinship terms, technology surveys, basic institutional forms, various traits that were part of the standard ethnological catalogue, etc.), take anthropometric measurements and photographs, and, particularly if a museum was its sponsor, collect artifacts (see Sect. 2). Gifts and payments would be distributed—items such as steel tools, for example, which were relatively inexpensive for the expedition but highly prized by the recipients—and the party and its carriers would move on to the next group.

Beginning in the 1920s a revolutionary new field-work methodology expanded the ethical implications. Based on a technique loosely known as 'participant observation,' the approach involved prolonged (1–2 years or more) residence in a host community with the aim of acquiring intimate familiarity with the people and their culture. Close familiarity in turn invited the innovation of the 'case method' approach, in which specific events (often disputatious in nature) were contextually examined as indicative of underlying cultural understandings or social organization. Beyond their analytic usefulness, well-executed case studies added vividness and probity to the ethnographic account. Case studies embedded in monographs usually identified individuals by name, along with relevant genealogical histories and information of a potentially embarrassing or inflammatory nature. Risks to informants thereby entered into the research equation, but, initially, only to a slight degree. Missionaries and colonial administrators might have searched out published information and used it to the detriment of individuals or cultural traditions, and yet it appears that this was a rare occurrence. As for the people themselves, information revealing the hidden deeds and motives of neighbors and enemies might

have been very useful, and damaging to the persons named but, under those remote, preliterate conditions, the idea that some explosive item from a book or journal would find its way back to the village and there make mischief would have been very farfetched.

Nowadays, such possibilities are quite thinkable. With the spread of literacy, some monographs are finding their way back to the village, or into the courts, where facts and events previously noted are being used, for example, in litigation concerning land claims. Starting in the late colonial period and continuing until the present day, educated elites in many of the countries traditionally included in cross-cultural studies comprise a new community of critics: some are the children of cultures ranking as 'classics' in the ethnographic literature, some are social and behavioral scientists, some are even anthropologists. The new intelligentsia, many of them educated in the (self-) critical Western tradition, many of them eager to forget the past and move on, are harshly critical of their colonial past and willing to condemn out of hand studies carrying the odor of those times. Ready allies are to be found from within the various critical movements that swept Western academic institutions during the last two decades of the twentieth century. Anti-orientalism, postmodernism, neomarxism, 'science studies,' the idea of ethnography as a constructed text, the distrust of studies conducted under pre-feminist, 'patriarchal' assumptions—these are part of the ethical-cum-intellectual apparatus currently used to discredit the tradition of cross-cultural study.

Although a good deal of this criticism is ill-informed and irrelevant, the best of it is very good indeed, offering novel insights and perspectives, the value of which even the most diehard traditionalist cannot reasonably deny; devising new concepts and new research goals and relationships, such as befit the new parameters of cross-cultural studies in a world very different from what it was as recently as a generation ago. Gone are the reassuring isolations which formerly shielded cultures, and those who studied them, from outside scrutiny—the never-perfect isolations which lent credence to the always dubious doctrines of cultural essentialism and cultural relativism (Gellner 1985, Cook 1999). Globalization engenders recurrent issues, such as local environmental degradation, the AIDS epidemic, poverty, war, and political revolution, which pique not only the researcher's interest and intellect, but his or her deep questionings into the ways of goodness, justice, beauty, and the future of the human species. These are supremely moral and ethical questions, which, while not displacing the urgent quest for objective findings, are adding immense complexity to the task of the researcher.

2. Cultural Property

Reference was made to the routine practice of earlier researchers of collecting information and emblematic

artifacts from peoples they visited. Now, generations later, descendants of some of these peoples, or those speaking for them, have come to demand the return of these objects. Never mind that perishable objects, both tangible and intangible, would never have survived in their home place, their removal is retrospectively construed as a form of cultural banditry. Today, legal claims concerning cultural patrimony and indigenous peoples' exclusive rights to traditional property—material or intellectual, commercial or sentimental—circumscribe an international, cross-cultural arena of high-stakes ethical debate. The cases, to name a few, include the longstanding litigation between Greece and the United Kingdom over rights to the 'Elgin Marbles'—sculptures from the Parthenon and other buildings of the Athenian acropolis which were removed by Lord Elgin in the early nineteenth century and were later sold to the British Museum; the legal contest between scientists and local Native American groups over custody of the remains of a prehistoric individual of obscure genetic provenance known as 'Kennewick Man'; more generally, the repatriation of skeletal material to indigenous descendant groups—a process now mandated by US federal law; efforts by Australian Aboriginal groups to recover legal title to ancestral lands containing sacred sites; the question of potentially huge royalties payable to various Amazonian groups who have guided drug companies to forest products having pharmaceutical properties; and the commercial exploitation of design elements taken from among Native American traditions of the US Southwest.

Social scientists are being ever more frequently drawn into these issues as witnesses, commentators, teachers, analysts, and advocates. Increasingly, sometimes unavoidably, the role of the researcher is shifting from that of detached observer to that of social activist. What to a previous generation of investigators would have seemed an unconscionable abandonment of scientific objectivity is nowadays becoming a common, practical necessity—a commitment demanded by newly sophisticated, politically aware groups as a condition of their acceptance of a research visitor. Other researchers, decrying 'scientific objectivity and detachment' as simply one ideology among others, gladly construe their mission to be that of assisting in the struggle of powerless ('subaltern') groups to improve their lot vis-à-vis dominant state or corporate interests (D'Andrade 1995). Wherever one stands on the detached vs. committed gradient, social movements and the emergent social and political aspirations of postcolonial groups are, themselves, of compelling intellectual interest to the social scientist, further blurring the ethically consequential line between actor and observer. Gone is the age of innocence during which many ethnographic habits were formed, when ethical choices were a good deal simpler.

In less controversial terms, professional ethics now calls for the dissemination and preservation of re-

search findings. Some host countries require, as a condition of granting research permits, that copies of field notes be deposited in their national libraries or archives, or that debriefing talks be given to academic groups. Fearing the continued neglect and eventual loss of much unpublished material, some research communities are establishing specialized archives for the preservation of field notes and other cultural records, and for their systematic repatriation to the host countries, with the aid of microfilm, microfiche, or other inexpensive, readily accessible technologies (Tuzin 1992). Not only do such practices secure priceless cultural information and benefit future generations of researchers, they also contribute immediately to goodwill and a collaborative spirit between scholars belonging to reciprocal sides of the cross-cultural divide. These trends are further assisted by the requirement of many host countries that foreign researchers be sponsored by a local academic institution or engage an assistant or co-investigator from among the local scholarly community. The importance of behaving responsibly toward host peoples and nations and toward the long-term scientific good is an explicit principle of professional ethics. Some funding agencies, such as the Wenner-Gren Foundation for Anthropological Research, are actively encouraging, and committing resources to, an approach that emphasizes custodianship rather than ownership as properly describing the researcher's relationship to cultural treasures he or she is privileged to collect.

The distinction between custodianship and ownership further points to an unresolved issue of looming ethical importance: on moral and legal grounds, what should be the tenurial relationship between a body of cultural ideas and the persons who, by accident of birth, happen to be carrying them? Or, as put by one commentator (Brown 1998), can culture be copyrighted? The alternative would be to say that, just as culture is an attribute of humanity, so all cultures, none of which are formed in perfect isolation, are manifestations of that attribute, and therefore are ultimately the property of all humanity—granting that particular groups have contingent, custodial rights over particular cultural assemblages. The former view accords with notions of cultural distinctiveness and cultural relativism, the latter with ideas of universal ethical standards and a body of interests common to the human species. The former view would seem to revere human rights—the right to their cultural traditions being the only right left to many otherwise dispossessed groups—while the latter would appear to remove the last protection against wholesale cultural trespass.

This issue is a minefield of ethical problems that will be prominent in the landscape of intercultural relations and research in the twenty-first century. The seismic convulsions of that landscape will be traumatic moments in the globalization of social and economic relations; increased human migration, possibly in

response to local population pressures and global-warming effects, bringing different cultural groups increasingly into competition with one another; the collapse of autochthony as an argument for indigenous groups in fending off migrant groups beating on the gate—groups once autochthonous themselves, driven by necessity rather than choice, with human rights of their own. In such a world, the conflation of cultures and peoples, collective memories and current circumstances, could produce dangerous vulnerabilities, generating the kind of xenophobic, stereotyping mentality that justified campaigns of ethnocide and 'ethnic cleansing' in the darker moments of the twentieth century. The better alternative, perhaps, is for universalist ethics to have a rightful place alongside the value of respecting and preserving cultural identities, and the rights that may be attached to them. Whatever shape that future takes, however, it is certain to affect the ethical circumstances of cross-cultural researchers.

3. Regulations and Codes

In recent decades, research ethics, once the privileged business of conscience, personal character, professional integrity, and scientific organizations, has come to the attention of lawmakers. During the 1970s federal government regulators in the USA developed rules for the protection of human subjects (Sieber 1992). With ethical reasoning rooted in the Nuremberg Code developed after World War II, the regulatory impulse was triggered by certain scandals involving biomedical research. Early in the lawmaking process, however, authorities recognized that potential risks to human subjects also existed in the social and behavioral sciences. Social and psychological hazards were recognized, as were the need for confidentiality and the importance of informed consent by subjects—not only at the outset, but through the entire course of the research. Procedures were instituted to disclose risks to peer reviewers and granting agencies, assess whether or not they are outweighed by potential benefits of the research, including to the human subject; and, in the field or laboratory, inform subjects of these risks as part of the consent process. A plan for compliance with these procedures, formally approved by university-based, faculty-run internal review boards, became a prerequisite for receiving government grants. Research centers quickly adopted these federal guidelines as their own, mandating the safeguards regardless of funding source.

These measures address ethical anomalies that are egregious in the sense of being built into the research design. No regulatory guideline can govern the many unforeseeable, quotidian dilemmas that may arise over an extended period of research in another culture—a reality acknowledged in the Code of Ethics promulgated by the American Anthropological Association (1998). Still, these measures have the salutary effect of

heightening researchers' consciousness and accountability concerning the welfare of human subjects, while helping to shield institutions and innocent investigators from liability if things go wrong and unethical conduct is alleged. On the other hand, regulatory experience and input from practicing social and behavioral scientists have shown that applying the 'biomedical model' to research in their fields raises problems, some of which are peculiar to cross-cultural studies. Even if language differences *per se* are not an obstacle, other communication barriers can impede the way of genuine 'informed consent.' If the research subjects are lacking in literacy, this would preclude written consent—a circumstance for which official guidelines permit equivalent oral procedures—but there are more subtle, cultural, difficulties. Even where written consent is feasible, subjects usually would not be accustomed to direct questioning by strangers (except, ominously, by police authorities) nor would they be likely to understand the foreign culture of litigation and the bureaucratic concept of 'risk' which jointly inspired the formal consent procedure in the first place. Depending on the political environment in which such an encounter occurs, a procedure innocuously designed to identify and mitigate risk triggers, instead, reactions of suspicion, alarm, and flight. For the would-be subject, for whom the compensatory benefits of the research may seem ethereal and remote, refusal might well be the easiest course, ending the research before it begins. Where traditional communities are receptive to the researcher, an overzealous concealment of individual identities would deprive the people of the only documentary history they might ever have. And so on. Government overseers and internal review boards, acknowledging such technical difficulties, have revised the regulatory guidelines so as to accommodate the special needs of ethically responsible research in other cultural settings.

The protections just discussed focus on the welfare of individual human subjects. What ethical imperatives pertain when not only individuals but entire cultural traditions are potentially at risk? Anthropology, by virtue of its disciplinary investment in cross-cultural research, celebration of cultural diversity, and preoccupation with the culture concept, has a special affinity for ethical issues that arise when science crosses cultural boundaries. Even if the threat stems not from the research *per se* but from other external factors (missionaries, commercial interests, etc.), is the anthropologist not duty bound to intervene vigorously in defense of the host culture? After all, in its submission to United Nations Commission on Human Rights, in 1947, the executive board of the American Anthropological Association recommended that the UN Declaration should be a statement 'of the rights of men to live in terms of their own traditions' (quoted in Cook 1999, p. 169). Should not anthropologists be willing to intervene, by word or deed, in defense of those traditions?

Critics might condemn such intervention as a veiled form of paternalism verging on racism—a denial of the ability or right of individuals to make their own decisions, to dispose of their cultural legacy in a manner they, and not well-meaning outsiders, choose. From this standpoint anthropologists are sometimes accused (sometimes by designing outsiders, with self-serving plans of their own for the group) of being culturally conservative, of wishing to see traditions preserved, of standing in the way of perceived material development and ‘modernization.’ Whatever their motives, the critics have a point. If the idea of indigenous rights is to mean anything, it must hold sacred the element of choice—even if the choice in question proves to be an unfortunate one in the view of contemporary observers or of future generations who must inherit the consequences. The role of the concerned outsider (anthropologist or other) is to help, if invited, to insure that the choices made are informed choices, based on the advisor’s experience, expertise, and external perspective. Such advice would be offered with knowledge of authority relations occurring in that cultural setting, specifically, the mechanisms through which decisions affecting individuals or the group are made.

4. Conclusion

Social and behavioral science research has ethical implications in any setting. Though we may control the interaction to a maximum extent, we are still humans studying other humans. Lives can be affected. To a greater or lesser degree, expectations, projections, aspirations, and the reflexivities accompanying all of these, are part of the immediate research situation. They influence what we observe and how we observe it; they shape the larger lessons we draw in the long afterthought. The publication of results can further affect the original subjects or, years later, their descendants. All of these ethical implications are greatly compounded when the research engagement is intimate and prolonged, and extends across a cultural boundary separating the investigator from the persons he or she is attempting to understand (Geertz 1968). Obligations accrue. Reminiscent of the methodologically engineered interplay of transference and counter-transference in psychoanalytic practice, as the ethnographic relationship deepens, complex personal motives intervene, and the desire to belong and to assist can become so strong as to compete with the desire to know. The whole point of maintaining ethical equilibrium is to transform this competition into a synergy. This is not an easy task; it requires thoughtful, honest reflection upon the ethical dimensions of what one is doing, together with the courage not to be paralyzed by them. In cross-cultural research the effort is worth the candle, for it is this synergy which stands

the best chance of producing a result that is at once truly humane and truly informed.

Bibliography

- American Anthropological Association 1998 *Code of Ethics*. American Anthropological Association, Washington, DC
- Brown M 1998 Can culture be copyrighted? *Current Anthropology* 39(2): 193–222
- Cook J W 1999 *Morality and Cultural Differences*. Oxford University Press, New York
- D’Andrade R G 1995 Moral models in anthropology. *Current Anthropology* 36(3): 399–408
- Geertz C 1968 Thinking as a moral act: ethical dimensions of anthropological fieldwork in the New States. *Antioch Review* 28(2): 139–58
- Gellner E 1985 Relativism and universals. In: Gellner E (ed.) *Relativism and the Social Sciences*. Cambridge University Press, Cambridge, UK, pp. 83–100
- Sieber J E 1992 *Planning Ethically Responsible Research: A Guide for Students and Internal Review Boards*. Applied Social Science Research Methods Series, no. 31. Sage, Newbury Park, CA
- Tuzin D 1992 The Melanesian archive. In: Silverman S, Parezo N (eds.) *Preserving the Anthropological Record*. Wenner-Gren Foundation for Anthropological Research, New York, pp. 31–42

D. Tuzin

Research Ethics: Research

Research topics, approaches and contexts bring with them ethical issues—some well understood and easily resolved, some unexplored and potentially risky, and some in which one’s moral outlook is inadequate to the problem. Sometimes the social or political context is such that morally wrong actions are rewarded and morally right acts are punished. Thus, ethical issues in research include conflicts between ethical principles, questions of what action will produce the most good, questions that call for a new outlook or approach, pursuit of personal gain at the expense of others (e.g., research fraud, endangerment of research participants) and research in contexts where the scientist is not free to choose his or her own scientific goals or methods.

Until issues are understood, scientists can only guess about the risks involved and how to ameliorate them. Although ethical questions are normative (What is the morally right action to take?), answers must often be grounded in scientific inquiry. For example, when investigators first undertook major field research on illegal behavior (e.g., prostitution, drug dealing, child pornography) and promised confidentiality in order to obtain honest responses from informants, perceptive researchers wondered, e.g., Can I protect

my data from subpoena, and can I gather valid data otherwise? How serious is the threat of subpoena? Would I go to jail if I were to ignore a subpoena of my data? What will happen to research careers and research in criminal behavior if protection from subpoena cannot be assured? Research by criminologists, sociologists, and political scientists has answered these questions, and led to the development of legal protection of data from subpoena. Research administrators, researchers, and students have learned to use these tools.

Empirical and methodological research coupled with social science theory is used to explicate how ethical issues arise, their possible consequences, and how they might be resolved. This article outlines a range of topics and approaches employed in research on research ethics. Other articles herein on ethics provide detailed examples.

1. History

Prior to the 1960s, most scientists considered their work value free or inherently good. Some scientific societies developed codes of ethics that were quite general and largely ignored. Some survey researchers worried about possible breaches of confidentiality; and some applied researchers recognized that knowledge is power that may be used for good or evil. However, most researchers were oblivious to the ethical dimension of their work. The field of research on research ethics was beyond the imagination of almost everyone.

With the rise of concern about ethics of human research, scientists found that social and behavioral research methods and theories could be used to explore and resolve problems in research ethics. Government funding for such research began to be offered. Increasingly, scientific societies encouraged such research, scientific journals published the findings, and new journals devoted to ethics and science came into existence.

2. Approaches

Most research on research ethics uses theory, has methodological components and involves empiricism. However, for simplicity, the discussion in this section focuses on main approaches:

2.1 Empirical Research

The entire range of empirical social/behavioral research methods may be used in pursuit of empirical answers to ethical questions. For example, survey research methods have been used to assess whether psychology departments comply with federal regu-

lations in the operation of their subject pools. Experimental methods have been used to examine the effects of various approaches to informed consent on comprehension and decision making, and entire research funding programs have invited proposals on this topic. Psychometric methods have been used to study the scientific and ethical values of scientists in various disciplines and their relationship to kinds of problems studied and methods employed (e.g., Kimble 1984, Tarnow 1999). Participant observation methods have been used to study the nature of communication concerning informed consent. Econometric measures have examined the costs and benefits of increased institutional monitoring of human research.

2.2 Methodological Research

Methods of designing and organizing research, evoking and measuring responses, and collecting, storing, and sharing data have ethical implications. For example, does a given procedure or method evoke trust, respect privacy, assure confidentiality, or yield valid results? The design and evaluation of methods to ensure that they meet ethical criteria require methodological research. Two major areas of methodological research in the service of ethics are privacy/confidentiality, and the design of field experiments. The development of methods for protecting subject privacy and data confidentiality has focused on such diverse contexts as interviews, sampling, collecting data from internet chat rooms, and inter-file linkage of shared data. Technological developments increasingly permit unobtrusive observation and measurement of behavior which could destroy trust in major institutions in society; discovering and evaluating uses of such technology have become topics of intense interest (e.g., Agre and Rotenberg 1999). The second major area of intensive methodological research has focused on field settings where rigorous experimental methods with well designed controls are needed, but ethical considerations prohibit jeopardizing the well-being of subjects by placing them in conditions not of their choosing. This has led to development of quasi-experimental methods that are responsive to both ethical and scientific requirements.

2.3 Theoretical Research

This approach conceptualizes and synthesizes what is known about a given ethical problem in the service of ethics and good science. For example, Laufer and Wolfe (1977) developed a theory of privacy that accounts for how age, culture, and status determine persons' regulation of the access of others to themselves. Thompson (1992) synthesized what is known about youngsters' age and sensitivity to risks of research. Gunsalus (1998a, 1998b) has summarized

the destructive conflicts and institutional mishandling that often surround whistleblowing, and has synthesized best practices for institutional and individual handling of whistleblowing situations.

2.4 Legal Research

The regulatory environment that is growing around scientific endeavors inevitably brings with it ambiguities and conflicts when applied in areas that the lawmakers did not anticipate. Exact standards of research conduct may leave no room for intelligent interpretation or for resolution of conflicting ethical principles. The effort required to follow a given law may be disproportional to the expected results. As examples, Western standards of informed consent and signed consent forms are inappropriate and even disrespectful in many contexts such as fieldwork in cultural anthropology and research in third world cultures. Oral histories often involve sensitive details about living persons but cannot meaningfully be rendered anonymous. Examining and challenging laws governing research and proposing more effective laws requires ability to research and interpret laws in relation to their effect on research within specific disciplines, methods, and research topics.

3. Topics of Research on Research Ethics

Topics are organized into 17 categories, nested under five major headings:

3.1 Theory, Method and Design Issues. These Background Issues Set the Ethical Context of Research

3.1.1 Values and epistemology. Scientists' beliefs about knowledge and reality determine what are considered valid scientific problems and how data should be collected, organized and disseminated. Whether knowledge is regarded as an unlimited good or as power affects perception of research risks and benefits. Whether reality is regarded as something 'out there' or as a construction composed of interaction between investigator and subject influences which methods (e.g., deception) are acceptable.

3.1.2 Validity. To be ethical, research must be valid; otherwise it disrespects subjects and society by producing false information. Conversely, in pursuit of validity, investigators may use unethical methods (e.g., putting subjects at risk). Much research on research ethics focuses on finding ways to optimize both beneficence to participants and validity of findings.

3.1.3 Equitable treatment of participants. Procedural and distributive justice issues may arise in planning, conducting, and applying research. Who sponsors the research? Who benefits? Will the results be used in a way that may harm the subjects or their culture? How are subjects selected? How is the power of the researcher balanced against that of the subjects? What kinds of control groups are acceptable? Who sponsors and pays for the research? When is it unethical to withhold information about sponsorship from subjects?

3.2 Risk and Benefit

The perspective of scientists is influenced by their research objectives, culture, and experience with the circumstances of those studied. A narrow perspective may negate the promise of scientific research to be beneficial. Research on risk and benefit seeks to broaden this perspective and enhance the value of research to subjects, their community, researchers, their institution, funders, science, and society. Risks and benefits are generic to all ethical considerations and research topics. However, specific risks vary markedly in relation to the research context (e.g., laboratory, community, school, workplace, hospital, prison).

3.2.1 Risk, wrong, and harm. How does one weigh moral wrongs (e.g., deceiving subjects), versus risks of harm? What determines perception of risk and safety? How can this determination be reached accurately and usefully? How can risk, wrong or harm be assessed?

3.2.2 Benefit and promise. Benefit is the moral justification of research, but it is often simply taken on faith that research is, by definition, beneficial. How are benefits identified, estimated, and maximized? Who benefits? Often research has little likelihood of benefiting society; does the researcher therefore have a duty to benefit the subjects or their community directly via some services, products, training, or job opportunities connected with the research?

3.2.3 Risk/benefit assessment. The price of scientific knowledge is risk and financial cost. How is risk justified in relation to benefit? What is the risk or cost to society of research opportunities that are lost due to financial, regulatory, or political pressures? (e.g., see Rosenthal and Rosnow 1984.)

3.3 Agreements and Communication

Research participants have some understanding of what is occurring; how correct and complete is that

understanding? How informed and autonomous is the decision to participate?

3.3.1 Informed consent. There are many possible modes of communication between researcher and subject about the purpose, procedure, and intended outcome of research. What modes are most effective? What information should be included? How can comprehension and decision making ability be enhanced, especially when subject competency (e.g., of children, institutionalized persons, the mentally ill) is at issue? When must guardians and gatekeepers be included in the consent process?

3.3.2 Deception. This may include concealment, mental reservation, lying, self-deception, and distraction of subjects from the main purpose of the study in order to study spontaneous behavior. Deception raises many difficult questions, e.g., how does deception affect persons across their life span? What are methodological alternatives to deception? When is deception acceptable?

3.3.3 Relationships as a source of data. Qualitative research may involve getting to know persons then reporting on some aspect of their lives, often from a critical perspective that differs from the way the persons describe themselves. How is the research problem defined and is this disclosed to the persons? The nature of the relationship, as perceived by the participant and the researcher, may affect how the research problem is defined, how the data are collected and organized, how the research report is framed, how the results are disseminated and used, and whether the participant is wronged or harmed by this incursion into his or her private life. With the advent of the Internet, researchers are now free to observe or participate in chat room discussions of quite personal and intimate topics. What are their responsibilities to those they study? Should they reveal their presence as researchers? Will subjects' knowledge of the researchers' presence have a chilling effect on their chat room participation? What steps should be taken to disguise the identity of the chat room and its participants?

3.4 Acquisition, Presentation and Use of Data

The many facets of privacy, confidentiality, and data use have been topics of much theory, and subsequent empirical and methodological research.

3.4.1 Privacy. The need to establish boundaries between oneself and others is universal, but differs depending on learning, cultural, social, and developmental factors. People may want to disclose to researchers more than the researcher wants to know, or lie to protect privacy. Researchers' views on privacy may be inadequate for judging what privacies others want, expect, or create. How can researchers learn what people consider as private and respect that privacy? What new threats to privacy are emerging? New technology brings with it new threats to privacy, but also more powerful safeguards, and a new generation of privacy activists to establish better policies for deterring invasions of privacy.

3.4.2 Confidentiality. Part of the informed consent process typically includes an agreement about how the researcher will limit access of others to individually identifiable data. Many risks to confidentiality are not immediately obvious, e.g., subpoena of data, theft from computer files, or deductive identification. How can researchers understand sources of risk to confidentiality and prevent their occurrence? A wide range of statistical, procedural, and legal solutions continue to be developed.

3.4.3 Uses of data: data sharing and dissemination. How should scientific data be disseminated and used? Some data such as videotaped data cannot be rendered anonymous as individuals might be recognized; what safeguards are appropriate to protect the privacy of the research participants? Who may be benefited or harmed? Should harmful data be suppressed? How is misunderstanding of findings avoided? What should be included or omitted? What are effects of various modes of dissemination? What is the obligation of scientists to contribute to the scientific literacy of society? How should raw data be documented, archived and shared? Who owns data and knowledge (intellectual property)?

3.5 External Influences on Research

3.5.1 Government regulations. Studies of the regulatory process contribute to the development and improvement of research regulation.

3.5.2 Institutional review boards. Studies of IRBs focus on their processes, and on their effects on science.

3.5.3 Questionable, taboo or controversial topics of research. Some espouse prohibiting research that could cause irreparable social harm, produce misleading findings, or be used in politically harmful ways. Which is more harmful, suppression or publication of such research? Studies of the impact of research that blames the victim or otherwise seems to misdirect society's response to problems seek to answer this difficult question.

3.5.4 Scientific integrity and responsibility. Pressures to obtain grants and to publish may interfere with self-correcting scientific procedures, and motivate scientists to resort to unethical practices or outright fraud e.g., fabrication of data, plagiarism. Research has investigated conditions that inhibit or facilitate scientific responsibility.

3.5.5 Ethics and politics. Scientists who investigate controversial topics may be accused of scientific irresponsibility. Charges of scientific irresponsibility or immorality often mask profound political disagreements. Who sponsors or employs the researcher and how the data are used raise complex ethical and political questions. Because these issues usually arise in highly partisan settings, the ensuing conflict often fails to enlighten either side. Analysis of specific case studies can clarify the issues somewhat and suggest guidelines and safeguards.

3.5.6 Animal rights activism. An extraordinarily high degree of political activism has surrounded animal research. In what ways has this stimulated the advancement of theory and methodology? In what ways has it damaged scientific endeavors? Controversy surrounding specific cases of activism is usually aimed at proving the other side wrong, rather than resolving conflicts. Scholarly examination of activism and of the changes that have occurred as a consequence of conflict holds promise of guiding scientists to methods and principles that are both more acceptable morally and more beneficial scientifically.

3.5.7 Education in research ethics. In recent decades, there has been increasing concern to educate scientists and research administrators about issues of ethics. The goals have been, variously, to impart greater sensitivity to ethical issues, to teach approaches to conducting research ethically, and to examine and interpret current laws and regulations governing research. A variety of educational formats have been employed within Western societies. The rise of worldwide research programs on sensitive

issues such as the epidemiology of AIDS and the intense ethical concerns this has raised, has now resulted in the development of ethics education programs in collaboration with scientists and applied ethicists in Third World countries. Despite significant expenditures on research ethics education, there is little understanding of the effects of these various education programs, of the relative effectiveness of the diverse approaches to teaching, or even of the best ways to measure their effectiveness.

4. Summary

While ethical theory and codes of ethics offer important general guidance, intelligent application requires understanding of how ethical problems may arise and of effective solutions to those problems. By its nature, social/behavioral research continues into new territory, raises new ethical issues, and requires research to explicate those issues and, where possible, to develop better methodology. Such research may be conducted as part of a project that raises an ethical issue, or as a 'stand alone' research activity. It may be the main focus or a sideline of one's research career. Whatever its origins, it deepens and broadens the ability of social and behavioral research to respect and be responsive to human (or animal) nature, to produce more valid results, and to better serve society.

See also: Animal Rights in Research and Research Application; Archiving: Ethical Aspects; Censorship in Research and Scholarship; Criminal Justice, Ethics of: Aspects of Human Dignity; Cultural Property and Ancestral Remains, Repatriation of; Deceptive Methods: Ethical Aspects; Ethical Dilemmas: Research and Treatment Priorities; Ethical Practices, Institutional Oversight, and Enforcement: United States Perspectives; Ethics Committees in Science: European Perspectives; Fieldwork: Ethical Aspects; Intellectual Property Rights: Ethical Aspects; Medical Experiments: Ethical Aspects; Mentoring: Supervision of Young Researchers; Normative Aspects of Social and Behavioral Science; Objectivity of Research: Ethical Aspects; Placebos, Use of: Ethical Aspects; Privacy of Individuals in Social Research: Confidentiality; Research Conduct: Ethical Codes; Research Ethics, Cross-cultural Dimensions of; Research Funding: Ethical Aspects; Research Publication: Ethical Aspects; Research Subjects, Informed and Implied Consent of; Surveys and Polling: Ethical Aspects

Bibliography

Agre P E, Rotenberg M 1999 *Technology and Privacy*. MIT Press, Cambridge, MA

- Gunsalus C 1998a Preventing the need for whistleblowing: Practical advice for university administrators. *Scientific and Engineering Ethics* 4: 75–94
- Gunsalus C 1998b How to blow the whistle and still have a career afterwards. *Scientific and Engineering Ethics* 4: 51–64
- Kimble G 1984 Psychology's two cultures. *American Psychologist* 39: 833–9
- Laufer R, Wolfe M 1977 Privacy as a concept and a social issue: A multidimensional developmental theory. *Journal of Social Issues* 33: 44–87
- Rosenthal R, Rosnow R 1984 Applying Hamlet's question to the ethical conduct of research: A conceptual addendum. *American Psychologist* 39: 561–3
- Stanley B, Sieber J, Melton G 1996 *Research Ethics: A Psychological Approach*. University of Nebraska, Lincoln, NE
- Tarnow E 1999 The authorship list in science. *Scientific and Engineering Ethics* 5: 73–88
- Thompson R 1992 Developmental changes in research risk and benefit: A changing calculus of concerns. In: Stanley B, Sieber J (eds.) *Social Research on Children and Adolescents*. Sage, Newbury Park, CA

J. E. Sieber

Research Funding: Ethical Aspects

The issue of ethics in research funding, a feature of research policy generally, is closely connected with the issue of freedom of research, an ideal of high standing in the science community. Special focus is on relations between providers and recipients of funding (see Orlans 1967). A vitally important principle says that research must be free in the sense of being shielded against any influences that may interfere with the obligation to seek truth. Yet, research must cope with regulation and restrictions, and discussions of these issues are concerned with questions of what freedom in which areas and with what consequences (Alver and Øyen 1997). The quest for freedom of research requires that conflicting and contradictory concerns and objectives be weighed against each other. In the process, the consideration of the policies, principles, procedures, and objectives of research funding often demands ethical reflection. Some principles of ethics seem self-evident. For example, no source of funding may be used to 'buy' a researcher to produce 'right' answers. Other issues require careful consideration—and there may be no clear-cut solution.

1. Sources of Ethical Concern

A sound process of research funding depends on organization and guiding principles. Research must meet standards of quality under conditions that favor innovation and contribute to knowledge accumulation. Research ought not to be done on problems that

have already been solved and where sufficient knowledge exists. But who ought to organize and govern? The science community may favor the view that the researchers themselves are in the best position to face new challenges and play key roles in deciding what research objectives to pursue. Nevertheless, responsibility for research must be shared between several partners: political authorities, research councils, universities and other research agencies, and executors of concrete research tasks. Research is cost demanding and requires highly qualified personnel resources and efficient infrastructure services. Also, research involves risks to people and environment. Priorities have to be set independently of the aspirations of researchers, and particularistic ambitions require tempering in view of universalistic standards. Coordination is needed to secure teamwork or to assure that researchers in different locations contribute their share of a totality. Funding agencies must deal with such issues, often involving rivalry among research institutions, among representatives of disciplines or fields of specialization, and among individual researchers. All of these matters may evoke ethical concern. The main point is to assure sufficient freedom and autonomy, protect the research process against unnecessary control, and guard against interference having undue effect on conclusions.

Thus, a decision to fund research is a result of a process involving several actors and partners whose goals hardly ever are identical. Provider and recipient of funding are main partners, but they cannot relate to each other without considering the interests of the context from which they act. The funding agency cannot defy interests of its hinterland or constituency, whether voters, taxpayers, or stockholders. The recipient of funding must relate to norms of scientific conduct, principles of data access and use, interests of institution of affiliation, and personal career interests and breadwinning needs. Ethical concerns become accentuated when different values and normative systems confront each other, and thus through the process of defining unified direction, seeking alignment of goals, and establishing common ground.

Traditionally, the institution of science asserts moral obligation to seek 'knowledge for its own sake,' without specific concern for applications. The ideal of basic or 'pure' research, driven by intellectual curiosity and awareness of missing links in the body of knowledge, plays a fundamental role in building and updating an integrated knowledge base. The universities have special responsibility for basic research and its integration in higher education. However, the mechanism whereby basic research is funded is going through significant changes as the role of research assumes new dimensions. Scientific discoveries accelerate change in society, but at the same time, scientific research becomes an instrument for solving problems brought about by change. Events during the

twentieth century contributed to widespread distrust in knowledge as intrinsically good, and perceptions of evil consequences of scientific findings brought public opinion to hold that some discoveries ought not to have been made—and thus not funded. These and other circumstances create a demand for more control and more pronounced goal orientation. Some funding agencies routinely demand that even basic research be justified in terms of applicability. Then, as an increasing proportion of research funding goes to applied research, the goals, and subsequent utilization of findings, become matters for ethical evaluation.

Standards of applied research bear heavily upon criteria of assessing the merits of basic research. The science system demands investment priorities, coordination for optimal utilization of funding, efficiency, and a favorable cost–benefit ratio. In academia, the freedom to pursue theories challenging scientists’ imagination is under pressure from funding sources needing quick answers to urgent problems. But at the same time, large sectors of the science community are morally devoted to research on issues of environment, poverty, health, inequality, and political instability, to mention but a few critical themes.

2. Sources of Research Funding

Funding for research comes from a multitude of sources. Governments channel public funds through foundations or research councils to individual researchers and to research establishments. Industrial corporations engage researchers as their employees or commission researchers on contract for particular research jobs. International organizations have special budgets for research. Private endowments and voluntary organizations provide funding for specialized research areas. The pluralism of research funding provides opportunities for research across an extensive array of issues and problems. Yet, there will always be concern that funds are skewed in favor of certain research goals at the expense of others. Not surprisingly, a host of ethically loaded questions concern funding priorities and processes of assuring optimal conditions for research.

2.1 Funding from Public Sources

Political authorities establish guidelines and policy principles for research, usually without entering into details of strategic decision making or interfering in the design of programs or projects. However, when investments in research are perceived as vitally important in view of political or national interests, funds may become earmarked for specific purposes. In rare instances, politicians address questions of whether

certain disciplines are worthy recipients of funding from public sources, as in the recurrent debates in the US Senate about the place of social sciences in the National Science Foundation (Larsen 1992).

Generally, the science community has a strong hand in strategic decision making in research funding, as much authority is delegated by politicians to science foundations, research councils, and academies of sciences, where researchers influence policies, grant funds, and resolve ethical issues. However, their performance may be subject to evaluation by political authorities in terms of criteria of political expediency and objectives. The representatives of the science community may have to accommodate to demands from above, under threats of reduced flows of public funds. Within this contest area, the flow of public funds to basic research is particularly vulnerable, as evaluation tends to follow criteria of applied research.

Increasing internationalization of research requires attention to issues of compatibility across national boundaries of legal and ethical principles bearing on the use of public funds, as for example, privacy protection and confidentiality provisions. Standards of ethical conduct have been adopted by most professional organizations, while modern versions of globalization, internationalization, and interdisciplinarity of research bring up new issues needing continued attention. For example, when research funds transgress national boundaries, sensitive issues of sponsorship and of true or hidden agendas may stir up moral and political controversy, as in the case of ‘Project Camelot’ (see Horowitz and Katz 1975, Homan 1991).

2.2 Funding from Corporate Sources

In corporate funding, a corporation enters into a contract with a researcher or a research agency engaged to produce knowledge applicable for a defined purpose. Normally, a contract is preceded by negotiations to secure mutual rights and duties. It may not be possible to foresee all problems involving ethical aspects that may come up in the course of an investigation, but it is always helpful when potential sources of problems may be dealt with at the time of negotiating a contract. In applied research, the linkage of funding and spending often involves a certain urgency and a limited time horizon. While the funding source has an interest in attaining a useful research result quickly and efficiently, the investigator is eager not to sacrifice researcher rights. The investigator must relate to ethical principles adopted by the professional association and also wishes to clarify questions of ownership to research data and results, respect for rules of confidentiality, freedom to publish—and perhaps, freedom to have second thoughts about the results of an investigation. A funding source may want provisions for secrecy or first use, while the

researcher is on guard against any clause that restricts truthful reporting. Other issues requiring clarification involve the funding agency's participation in planning and execution of research, terms of monitoring a project while in progress, publicizing or utilizing partial information before the conclusion of a project, and terms of application of findings. Both parties in a contract negotiation must be acquainted with relevant legal provisions concerning intellectual property rights. Attention to 'small print' may require qualified advice, and researchers may be well served when the institution with which they are affiliated provides backing and takes part in contract negotiations.

3. Researcher Influence on Funding

Research policy formation as well as decision making in matters of concrete research tasks requires input from the science community. Science policy advice may be solicited or volunteered; in either case, the decision-making process invites attention from lobbyists and special-interest groups seeking influence on policies, budgets, and the allotment of funds. Although the process is regarded as necessary and legitimate, it may generate many ethical issues that concern political authorities, funding agencies, the science community, and the general public. It is assumed that those scientists who possess the highest professional competence within highly specialized fields are best qualified to serve as judges and gatekeepers for research programs and initiatives. However, as research plays an important role in the pursuit of social, economic, and political goals, the role of scientists in research funding, as policy advisors and as participants in the decision-making process, may to some appear too closely interconnected with the political process.

Furthermore, fair disbursement of research funds requires safeguarding against unfair delimitation of turf, undue promotion of special interests, and questionable manipulation of the research opportunities of friends and foes. Yet, measures to secure competence and impartiality ought not to lead to the paradoxical situation that researchers having specialized knowledge of an area have to recuse themselves from participating in evaluation and decision making within that area. In fact, the system needs mechanisms for the protection of the integrity of the science advisor.

4. Freedom, Market, and Control

Research has become closely linked to the solution of pressing problems encountered through rapid social change. At the same time, the possibilities of controlling research—and the desire to control—have become stronger. With increasing control, in a situ-

ation where research functions much like a market commodity, the center of gravity is shifting from basic to applied research. Increasingly, the firm linkage of research to practical utilization determines the criteria by which decisions to fund research are made. For the research community, there may be a price to pay in terms of freedom of research and standards of research. Surely, there will be more attention paid to the ethical issues surrounding decision making and policy formation in the area of research funding.

See also: Censorship in Research and Scholarship; Intellectual Property Rights: Ethical Aspects; Objectivity of Research: Ethical Aspects; Research Ethics: Research; Research Publication: Ethical Aspects; Science Funding: Asia; Science Funding: Europe; Science Funding: United States

Bibliography

- Alver B G, Øyen Ø 1997 *Forskningsetikk i forskerhverdag*. Tano Aschehoug, Oslo, Norway
- Homan R 1991 *The Ethics of Social Research*. Longman, London
- Horowitz I L, Katz J E 1975 *Social Science and Public Policy in the United States*. Praeger, New York
- Larsen O N 1992 *Milestones and Millstones: Social Science at the National Science Foundation, 1945–1991*. Transaction, New Brunswick, NJ
- Orlans H 1967 Ethical problems in the relations of research sponsors and investigators. In: Sjöberg G (ed.) *Ethics, Politics, and Social Research*. Schenkman, Cambridge, MA, Chap. 1, pp. 3–24

Ørjar Øyen

Research Publication: Ethical Aspects

Ethical aspects in research publications are all the moral issues and problems raised by any behavioral misuse or abuse of the system for communicating and disseminating scientific results and information within the international scientific community. It is intended as a social body genuinely co-operating for the cultural and material improvement of mankind regardless of any personal, political, economic, or racial interests. Misuses or abuses include plagiarism, fraudulent and repetitive publication, guest- and ghost-authorship in multi-authored papers, conflict of interests in peer reviewing, censoring, and rejecting publications.

Scientific knowledge is community property unceasingly increased by individuals or groups in mutual co-operation and fair and honest competition. Even in the professional and competitive world of today researchers are intended as primarily engaged in the disinterested pursuit of truth and in the quest of new

insights and discoveries. Honesty towards oneself and towards others is considered as a fundamental condition for co-operation and fair competition since researchers depend on each other and cannot be successful unless they are able to trust each other and their predecessors. Publications are the primary medium through which scientists give an account of their work, co-operate, contribute to the advancement of knowledge, and claim the rights to be recognized for their discoveries. Therefore, dishonesty in publication not merely throws research open to doubt but destroys the very fundamentals of science.

The publication of a fraudulent article (the very outcome of a scientific fraud consisting in the fabrication, falsification or in 'cooking' and manipulation of statistical data) remains the most serious transgression of ethics of scientific publication. It is commonly considered to be an essential part of the scientific misconduct as well as plagiarism, the deliberate presentation of another's ideas or texts as one's own. Compared to these repetitive publications (publishing the same article in different journals), sequential publication (reporting follow-up of the same study with additional subjects but without new results), 'salami slicing' (dividing up different parts of the same study for publication in different journals), and guest and ghost authorship (giving or not giving authorship credit to persons who did or didn't take part in the research or in drafting or revising the article) are all considered as minor transgressions. These result in a huge amount of unwieldy and polluted literature that distorts the direction of science. It has been roughly calculated that flawed papers published on the 5,600 scientific journals indexed by The Science Citation Index amount to about 10,000 each year. This raises the problem of the reliability of the peer-review system designed to assess the quality of scientific literature, and that of the fate of the invalid ideas and data generated by fraudulent publication (Friedman 1990).

Good scientific journals today publish original articles only after they have been examined by competent reviewers for their validity and originality. Antecedents of peer review practices go back to the seventeenth century and the first organizations of scientific societies in Italy, France, and Germany but the modern system was only fully developed after World War II and it was adopted firstly and mainly in the Anglo-Saxon cultural area. The system is credited to have greatly contributed to the efficiency and professional organization of the scientific enterprise but since the eighties it has been severely criticized not only for its enormous costs in time and money but also for the progressive loss of accuracy (Lock 1986). Reviewers have indeed some moral obligation to make every necessary effort to recognize manipulations and falsifications. However, to expect irregularities to be reliably detected would be misleading: the original data are not available to reviewers, and if they were

they would not have time to replicate experiments and observations. This is why, according to many, peer review is so vulnerable to dishonest conduct. Other flaws have been detected in negligence (it resulted from inquiries that the higher the academic status and age of the reviewer, the lower the quality of the review), in vested interests of the referees and in many forms of bias (against particular individuals, subjects, or institution), the worst of which is that against the innovators. Some critics maintain that the peer review system favors mediocre basic science and fails to recognize and promote creativity, originality, and innovation (Horrobin 1990, Campanario 1996). Even the anonymity of the referees has been questioned for its finality as contrary to the moral norms of universalism, communality, disinterestedness, and organized skepticism that are thought to characterize the ethic of science. In fact the peer review system holds some risks for the authors because ideas, research findings, and texts which are unprotected by patents or intellectual property rights are submitted to persons whose identity is unknown to the authors and who may happen to be their direct competitors.

As for the fate of invalid data generated by transgressions to the ethics of scientific communication, the scientific community has yet to agree on the level of inaccuracy required to mandate a retraction vs. an erratum or who may retract an article. Although early in 1987 the International Committee of Medical Journal Editors published guidelines for the format and terminology of retraction notices, articles known to be fraudulent may be not retracted at all, or are retracted with ambiguous retraction. In other cases, only part of the data or conclusions are retracted. Editors of journals have little leeway for action since publishing corrections is fraught with juridical risks unless all authors jointly sign them. However, omitted or ambiguous retractions are not the only reasons of invalid work not being effectively purged from the scientific literature. It resulted from inquiries on US authors (Pfeifer and Snodgrass 1990) that after retraction the invalid information they produced continues to be used since the majority of retractions are not indexed by their own journals and no reliable source exists to locate fraudulent or erroneous work.

Some ethical relevance also has the soft censoring activity by which administrations and companies tend to control and restrict the dissemination of sensitive scientific and technological knowledge on the grounds of national security and economic reasons (Relyea 1994).

Such ethical problems depend on the very structure of the scientific enterprise as shaped in the seventeenth century but they have only grown in relevance and extension in recent times. Cases of plagiarism or fraudulent publication have been noticed in the past history of science, some of which involving such outstanding figures as Ptolemy, Luca Pacioli, Ernst Haeckel and even Galilei and Newton. However, it is

generally admitted that if not the number, the relevance and the social impact of the transgressions to the ethics of scientific publication began to rise in the 1940s. This was when modern science policy was established in view of the expected contributions of the scientific community to national defense and industrial progress. The new reward and funding system enhanced the traditional competition between scientists and it was then, in 1942, that the sociologist L. Wilson apparently coined the expression 'publish or perish.' It could hardly be said however that in the first 30 years competition did in some way sensibly promote fraudulent behaviors or the pollution of scientific literature. The turning point has in fact been stated in the late 1970s when the social structure of the scientific community underwent some crucial changes due mainly to the exponential increase of the number of people involved in research and the proportional (though not absolute) decreasing rate in assigning funds to research activity. The ensuing economic constraint (upon which the resources poured in by private companies had no or little effect) led to a situation of even stronger competition among scientists for academic rewards, position, or research funds. One crucial decision was that of the adoption of publication as the main performance indicator and of measuring productivity as the number of publications per length of time. The additional tool of citation analysis which was introduced in those very years (Garfield 1972, 1979) and based on the key concept of 'impact factor', (as the ratio of the number of citations a journal receives to the number of papers published over a particular time period) and then extended to calculate the impact of the work of individuals, groups, or departments generated a race to produce, in the shortest possible time, the least publishable unit in a journal with the highest impact factor. At the same time the growth of the scientific population implied a statistical decline in ethical standards and in the quality of production. Lotka's law of productivity in fact states that in a certain field, half of the scientific literature is produced by a population of the square root ($\sqrt{N}$) of all N scientists. De Solla Price pointed out that the exponential growth of the scientific population implies a proportional increase in the number of scientists able to author professional papers, but not important and innovative articles. From a statistic compiled by the Institute for Scientific Information (ISI) 55% of the scientific papers published between 1981 and 1985 in journals indexed by the Institute received no citation at all (Hamilton 1990). This led some students to draw the conclusion that more than half, or perhaps more than three-quarters, of the scientific literature are worthless. Others reacted against the notion that most uncited papers are without value. As for the use of citations as a measure of impact and of the quality of a publication, it has been questioned whether this may not lead to abuses such as citation cartels. It was estimated by the

same ISI, that self-citation accounts for between 5 percent and 20 percent of all citation.

It is from, or in any case within, those changes that according to the current views (see, for example, Petersdorf 1986) the pathological phenomenon of scientific misconduct arose to which the major ethical issues in scientific publishing are tightly related. As a consequence the analysis of these issues, the evaluation of the frequency and gravity of the transgressions, and the discussions on how to manage the problems they pose have been conducted as part of a general study on scientific misconduct.

Unfortunately, despite increasing publicity during the past 25 years, attempts to assess the quantitative dimensions of the problem of scientific misconduct did not lead to conclusive results: its magnitude is unknown and its detection largely serendipitous. Guidelines and actions adopted to manage scientific and (as a consequence) publication frauds and other abuses vary according to which of the two main different approaches they are based on. Sociologists and the vast majority of scientists share the opinion that the frequency and incidence of misconduct did not sensibly change during the twentieth century. The impression that fraud has vastly increased may be the result of the heightened self-consciousness in society coupled with strong moral expectation that scientists in the pursuit of reliable knowledge will live up to a higher standard of probity. This view seems to be substantiated both by the reports of the two most important US authorities responsible for dealing with misconduct cases, the Office of Inspector General (OIG) of National Science Foundation (NSF), and the Office of Scientific Integrity (ORI) of the Public Health Service and by that of analogous authorities in Denmark, Norway, Finland, and Australia. The ensuing suggested strategy is firstly, and mainly intended to save the autonomy right and self-regulation privilege of the scientific community from legal or government unnecessary intervention. It also recommends that inquiries primarily be made by local academic bodies, at the same time stressing the crucial role of education to improve the ethical standards of researchers.

Opposite is the view according to which, though there are no official data on the scale of the problem, the reports suggest that only a comparatively small proportion of cases of misconduct is actually detected and that what emerges is just the tip of an iceberg (Lock 1993). According to this view, self-regulation has proven to be ineffective and unsuitable. Scientists are not trained in conflict and conflict of interests resolution. Their intuitive response is usually wrong and they tend to set up shaky ad hoc procedures that do not guarantee the accused notice of all the charges; the opportunity to respond to all the charges and to the evidence; and a decision based on rigorous standards. In this view the setting up of national agencies, rather than local committees, is recom-

mended. The attention is focused on the economic and social background. It is assumed that those transgressions simply reflect the problems of a science which is today too big, too entrepreneurial, and too competitive to guarantee its very scope—the cultural and material improvement of mankind, in due and reasonable proportion to the intellectual and economic resources and at the same rate of the past three centuries. As far as scientific publication is concerned this second approach led the Committee on Publication Ethics (COPE), created in the UK by 20 editors of scientific journals, to ask for an independent agency to detect, handle, and prevent research publication abuses.

However, there is no general agreement on this view that led some of its proponents to recommend for some radical adjustment in the scientific policy first of all in connection to the ratio between scientific population and funds. While universities, scientific societies, funding agencies, and governments have opposed those proposals tacitly or patently, even their advocates acknowledge the importance of improving the ethical standards via the teaching of research ethics. This is currently considered by competent institutions and single students, besides and before, any disciplinary, economic, or legal action. Likewise, all agree on the necessity of suitable ethical and professional guidelines and measures which have in fact been proposed or adopted, though independently and according to different experiences and aims, by learning societies, agencies, journals, or individual students.

The primary importance of ethical (coupled with methodological) education emerge particularly in connection to some of the problems of scientific publishing, which seem less directly linked to the social and economic structure of the scientific system than to the peculiarity of research areas which involve cultural diversity or human and animal rights. In fact though science is largely transcultural, the human community has not yet developed a completely corresponding body of transcultural ethics. In some areas this may affect both methodology and ethics. Uncritical and unwitting reliance on the researcher's own culturally received principles and values regarding personal and physical identity, health, loving, suffering, death, etc. may generate potentially distorted initial observations which can lead to systematic divergence from truth in anthropological, psychological, sociological, or medical inquiries (where also systematic infringements of human or animal rights or the exploitation of subjugated populations may occur). Scientific publications based on such inquiries are, in principle, to be considered as poorly or improperly designed and performed studies. Current ethical and methodological standards do not include clear and effective rules for authors, editors, and peer-reviewers though the problems they raise have been debated in connection to some controversial cases.

A first general attempt to mitigate the pressure to publish that lies behind most deviant practices is, according to many, to place a ceiling on the number of publications that can be considered in evaluating a candidate for promotion or funding. However, some experts have also suggested that the number of scientific journals be drastically reduced (Bracey 1987). On the same line; although the use of quantitative indicators of scientific performance seems unavoidable, it has been recommended that impact factors shall not by themselves serve as the basis for funding decisions and that the quality of publication must be the primary consideration (Bucholz 1995). While no effective measures (other than severe sanctions) seem foreseeable against major authorship transgressions, as plagiarism or the publication of fabricated data, it has been recommended to journals to cooperate in cross-checking control activities and to research institutes and universities to store the primary data on which publications are based. All sources and methods used to obtain and analyze those data should be, when required, fully disclosed and the disappearance of primary data from a laboratory considered as an infraction of the basic principles of careful scientific practice. On the other hand, since there is no universally agreed definition of authorship, specific criteria and requirements have been invoked as an acknowledgment or footnote in every co-authored paper to state the exact contribution by each author. Or, the signing of a form where each author affirms having seen the final version and agrees to its publication. According to those proposed criteria the award of authorship should balance intellectual contributions to the conception of studies or experiments, to the generation, analysis, and interpretation of the data, and to prepare the manuscript against the collection of data and other routine work. If there is no task that can reasonably be attributed to a particular individual, then that individual should not be credited with authorship. In principle all authors must take public responsibility for the content of their paper and honorary or guest authorship is in no way acceptable. Against repetitive publication since 1969 the so-called 'Ingelfinger rule' has been proposed (Angell and Kassirer 1991) according to which papers are submitted to journals with the understanding that they (or their essential part) have been neither published nor submitted elsewhere. Reviewers are supposed to vigilate upon transgressions and it is stated in some journals' guidelines that if they suspect misconduct, they should write in confidence to the editor. As for the same reviewers it is reminded that they are bound to the duty of confidentiality in the assessment of a manuscript and asked to disclose conflict of interests and to provide speedy, accurate, courteous, unbiased, and justifiable reports. The submitted manuscript should not be retained or copied and both editors and reviewers should not make any use of the data, arguments, or interpretations, unless they have the

authors' permission. Editors should regard their function as ensuring that authors are treated fairly and their decisions to accept or reject a paper for publication should be based only on the paper's importance, originality, and clarity. To avoid the suppression of originality and creativity it has also been recommended that studies that challenge previous work should be given an especially sympathetic hearing and that independent ombudsmen for journals be appointed. Some few suggestions have also been advanced on the subject of external pressures as, for example, that editorial decisions must not be influenced by advertising revenue or reprint potential, but it seems that little can be done for the secrecy imposed by industries or governments on sensitive research fields.

Envisaged actions and sanctions rank from a letter of reprimand and warning as to future conduct, or publication of a notice of redundant publication or plagiarism, to formal withdrawal or retraction of the paper from the scientific literature, informing other editors and the indexing authorities or reporting the case to the authorities or organizations which can investigate and act with due process.

There is however no suitable insight on the impact those guidelines and measures had, or may have, in the long term.

See also: Archiving: Ethical Aspects; Ethics Committees in Science: European Perspectives; Intellectual Property Rights: Ethical Aspects; Objectivity of Research: Ethical Aspects; Research Ethics: Research

Bibliography

- Angell M, Kassirer J P 1991 The Ingelfinger rule revisited. *New England Journal of Medicine* **325**: 1371–3
- Bracey G W 1987 The time has come to abolish research journals: too many are writing too much about too little. *Chronicle of Higher Education* **30**: 44–5
- Bucholz K 1995 Criteria for the analysis of scientific quality. *Scientometrics* **32**: 195–218
- Budd J M, Sievert M-E, Schultz T 1998 Reasons for retraction and citations to the publications. *Journal of the American Medical Association* **280**: 296–7
- Campanario J M 1996 Have referees rejected some of the most-cited papers of all times? *Journal of the American Society for Information Science* **47**: 302–10
- CIOMS 1991 International guidelines for ethical review of epidemiological studies *Law, Medicine & Health Care* **19**(3–4) (Appendix I): 247–58
- Committee on Publication Ethics 1998 *Annual report 1998*. BMJ Publishing Group London
- De Solla Price J D 1963 *Little Science, Big Science*. Columbia University Press, New York
- De Solla Price J D 1969 Measuring the size of science. *Proceedings of the Israel Academy of Science and Humanities* **4**: 98–111

- Friedman P J 1990 Correcting the literature following fraudulent publication. *Journal of the American Medical Association* **263**: 1416–9
- Garfield E 1972 Citation analysis as a tool in journal evaluation. *Science* **178**: 471–9
- Garfield E 1979 *Citation Indexing: Its Theory and Application in Science, Technology and Humanities*. Wiley, New York
- Hamilton D P 1990 Publishing by—and for?—the numbers. *Science* **250**: 1331–2
- Horrobin D F 1990 The philosophical basis of peer review and the suppression of innovation. *Journal of the American Medical Association* **263**: 1438–41
- LaFollette M C 1992 *Stealing into Print: Fraud, Plagiarism and Misconduct in Scientific Publishing*. University of California Press, Berkeley, CA
- LaFollette M 2000 The evolution of the 'scientific misconduct' issue: An historical overview. *Proceedings of the Society for Experimental Biology and Medicine* **224**: 211–5
- Lock S 1986 *A Difficult Balance: Editorial Peer Review in Medicine*. ISI Press, Philadelphia, PA
- Lock S 1993 Research misconduct: a résumé of recent events. In: Lock S, Wells F (eds.) *Fraud and Misconduct in Medical Research*. British Medical Journal, London pp. 5–24
- Petersdorf R G 1986 The pathogenesis of fraud in medical sciences. *Annals of Internal Medicine* **104**: 252–4
- Pfeifer M P, Snodgrass G L 1990 The continued use of retracted, invalid scientific literature. *Journal of the American Medical Association* **263**: 1420–3
- Relyea H C 1994 *Silencing Science: National Security Controls and Scientific Communication*. Ablex, Norwood, NJ
- Serebnick J 1991 Identifying unethical practices in journal publishing. *Library Trends* **40**: 357–72
- Shankman P 1996 The history of Samoan sexual conduct and the Mead–Freeman controversy. *American Anthropology* **3**: 555–67
- Sponholz G 2000 Teaching scientific integrity and research ethics. *Forensic Science International* **113**: 511–4
- Stossel T P 1985 SPEED: An essay on biomedical communication. *New England Journal of Medicine* **313**: 123–6
- Whiteley W P, Rennie D, Hafner A W 1994 The scientific community's response to evidence of fraudulent publication. *Journal of the American Medical Association* **272**: 170–3

F. Di Trocchio

Research Subjects, Informed and Implied Consent of

'Informed consent' denotes a person's autonomous agreement to a proposed course of action, usually in a treatment or research context. Consent is not considered to be informed unless the person: has received disclosure of the information that a reasonable person would consider material to the decision at hand; has the capacity to use that information in a rational process of decision making; and reaches a decision in the absence of coercive offers or threats (Appelbaum et al. 1987a). Broad agreement exists that informed consent is an indispensable requirement for the ethical conduct of research with human subjects, at least

when the research renders them liable to more than minimal risks of harm. But considerable controversy remains regarding the precise contours of the doctrine and of the circumstances that might constitute exceptions to the general rule. And since the rules have been devised largely with biomedical research in mind, issues more characteristic of social and behavioral science research have been relatively neglected.

1. History of Informed Consent to Research

The history of informed consent to research is usually traced back to the decision of the court in the Nazi doctors' trial at Nuremberg after World War II, which declared, 'The voluntary consent of the human subject is absolutely essential.' It is clear, however, that the Nuremberg tribunal neither originated nor established the concept of consent to research. References to requirements for consent from research subjects can be found in the writings of physician experimenters and in governmental pronouncements dating back to the late nineteenth century, when experimental medicine grew out of its infancy. Indeed, George Bernard Shaw's *Pygmalion* contains an explicit statement of the principle. Two of the leading instances, ironically, derive from Germany, where the advanced state of empirical medical science made consent an object of at least periodic concern (Grodin 1992). Nonetheless, however much the Nuremberg court desired to root the principle in contemporary practice, it appears that researchers in many countries, even after the Nazi atrocities came to light, frequently neglected to obtain consent from their subjects.

Nor did the Nuremberg decision immediately change these practices. Investigators using human subjects apparently saw the Nuremberg Code, as the ten principles embodied in the court's decision came to be known, as applying only to the anomalous situation of concentration camp inmates. Thus, the history of the postwar decades is marked by repeated revelations of questionable ethical practices, including studies in which persons were subject to clear harms without their knowledge or consent (for one of the most influential reports of problematic studies, see Beecher 1966). Pressure built for clearer rules and mechanisms of enforcement.

The World Medical Association responded to these concerns with its *Declaration of Helsinki* in 1964 (World Medical Association 1964), the first attempt to forge an international professional consensus on the ethics of research. The *Declaration* endorsed a requirement for informed consent, albeit in less sweeping terms than the Nuremberg Code. National professional organizations developed similar documents, but their nonbinding character left research subjects with few certain protections. A move to legally enforceable rules came in the USA in 1974, with the adoption of regulations governing federally funded

research. Most of the industrialized countries in the world have since followed suit, by way of legislative or regulatory enactments, or binding guidelines (for a summary of the status of research regulation internationally, see Brody (1998). The general principle that research should not be conducted without the informed consent of participants has been accepted around the world.

2. Current Contours of Informed Consent to Research

The requirements for informed consent are rooted in the ethical principle of autonomy, which demands respect for the free, informed choice of a competent person (Faden and Beauchamp 1986). Valid informed consent requires three elements to be present: disclosure, decisionmaking capacity, and voluntariness.

2.1 Disclosure

The Nuremberg Code (1948) declared that the research subject 'should have sufficient knowledge and comprehension of the elements of the subject matter involved as to enable him to make an understanding and enlightened decision.' Brody (1998) summarizes the international consensus regarding those items that need to be disclosed to achieve this end as follows:

- '(a) the purpose and the research nature of the project;
- (b) the duration of the subject's participation, the treatment and procedures that the subject will undergo (with a clear specification of which are experimental), and the subject's responsibilities;
- (c) the risks of harms and discomforts that might result from participating in the research;
- (d) the possible benefits to the subject or to society from the research;
- (e) the alternatives to participating in the research;
- (f) the right not to participate and to stop participating at any time without penalty or loss of benefits to which the subject is otherwise entitled;
- (g) the measures to protect confidentiality;
- (h) the anticipated payments, if any, to the subjects and the anticipated expenses, if any, that will be borne by the subject;
- (i) the compensation and treatment available for injuries sustained in the research;
- (j) the identity of a contact person to discuss questions about the research;
- (k) the circumstances under which the subject's participation will be terminated;
- (l) the assurance that new information relevant to participation that becomes available during the research will be provided to the subject;
- (m) the names of the sponsors of the research and the ownership of any commercial results.'

This information is generally provided in written form, though often supplemented by verbal disclosures. Frequently, the result is a lengthy form (12–14 page, single-spaced forms are not uncommon in clinical research) that is difficult for potential subjects to read and assimilate. Provision of a mass of detailed information means, moreover, that important data, such as those dealing with significant risks, are frequently neglected by subjects. A common resulting problem is the so-called ‘therapeutic misconception’—a belief on the part of subjects that procedures are being undertaken for their benefit (including randomization and use of placebos), when they are actually designed solely to advance the research effort (Appelbaum et al. 1987b). Though investigators are often urged to provide additional disclosures in more understandable formats (e.g., illustrated brochures that subjects can read at their leisure), this still represents the exception rather than the rule.

2.2 Decision-making Capacity

The subject’s capacity (in legal terms referred to as competence) to give a meaningful consent was considered a *sine qua non* in the Nuremberg Code. Although formulations of the requirements differ somewhat across jurisdictions, decisionmaking capacity generally involves the abilities to understand the information relevant to a decision, to appreciate its implications for one’s own situation, to reason in a rational fashion, and to evidence a choice (Appelbaum and Roth 1982).

Capacity is generally presumed to be present, unless there is reason to believe otherwise. When studies involve populations at elevated risk of incapacity, however, including persons with severe mental disorders, mild dementia, head injury, and the like, particularly when the research may present significant risks to participants, investigators are increasingly assessing capacity prospectively. This can be done by means of clinical evaluations or formal screening instruments. Identification of a potential subject as having impaired capacities does not necessarily mean that the person must be excluded from research participation. Many such persons can be brought up to acceptable levels of functioning with extra educational efforts.

Were capacity to be an absolute requirement, as the Nuremberg Code suggested, no research could be conducted with children—who lack capacity as a matter of law—or with many patient groups, including potential subjects who are comatose, suffering from moderate to severe dementia or mental retardation, or severely psychotic and disorganized. The desirability of allowing research to be conducted with these groups, so that progress might be made on the disorders from which they suffer, led the drafters of the Declaration of Helsinki, and most subsequent com-

pilers of codes, regulations, and statutes, to make some provision for the entry of incapable subjects into research projects. Rules governing their recruitment differ, but common limits include that the research be directly applicable to their conditions, not able to be performed with capable subjects, hold out the prospect of some therapeutic gain, or present no more than minimal risks of harm. Innovative proposals that would allow persons to make advance decisions while competent about whether to enter research projects should they later become incompetent have been much discussed, but rarely implemented.

2.3 Voluntariness

Overt coercion—the application of threats or unfair offers to a prospective research subject—is universally recognized as vitiating that person’s decision and thus rendering consent invalid. The effects of more subtle influences on the legitimacy of subjects’ consent to research is a matter of greater controversy. Some environments may be considered so intrinsically coercive—prisons, for example—that research with the persons detained therein has been severely restricted in many countries. Whether hospitals, institutions for the mentally ill or retarded, schools and colleges, places of employment, and other entities fall into that category is more difficult to say. Paradoxically, efforts to protect inmates, patients, workers, and others by prohibiting their participation in research deprive them of their residual decisionmaking power, and may increase their feelings of helplessness and their perceptions of being coerced.

Research subjects are often compensated for their participation with money or its equivalent. Fair compensation for the time expended in a study is generally considered acceptable. To the extent that payment is so substantial as to lead subjects to give insufficient attention to the risks they may run by participating, however, it may be considered an undue influence on their decisions, which renders their choices substantially involuntary.

3. Special Considerations in Behavioral and Social Science Research

Rules governing informed consent to research, from the Nuremberg Code to most contemporary governmental regulations, typically have been formulated with biomedical research in mind. Issues more characteristic of social and behavioral research have been relatively neglected. Indeed, the federal regulations in the USA categorically exempt many types of behavioral and social research from their purview, including: educational research; studies designed to evaluate governmentally provided benefits or services; and testing, surveys, interviews, or observations of

public behavior, unless information is recorded in a manner that would render subjects identifiable, and subjects would be likely to suffer harm if the information were released (Code of Federal Regulations 1991).

Yet, many behavioral and social science studies involving human subjects raise important issues related to informed consent. One of the major questions is when, if ever, it is appropriate to deceive subjects about the purpose of a research study, a frequent technique in psychological research (see *Deceptive Methods: Ethical Aspects*). Other issues concern consent for access to records, consent in observational and field work studies, and the appropriateness of seeking consent—or some equivalent—from entire communities that may be affected by an intervention that is part of a research project.

3.1 Consent in Studies Based on Review of Records

Analysis of existing records, often aggregated into large databases, is a frequent technique used by behavioral and social scientists. These records may have been collected originally with subjects' consent but for nonresearch purposes (e.g., medical records) or without subjects' consent (e.g., national birth registries, educational testing results, or census information). When individual subjects are not reasonably identifiable in the database, it is generally agreed that consent need not be obtained prior to granting researchers access. Even with identifiable data, researchers have usually been allowed to utilize such records as long as they agree to protect the confidentiality of those persons whose information they are reviewing (e.g., by publishing data only in aggregate form).

This practice has been criticized as insufficiently protective of subjects' interests in both autonomy and privacy. The autonomy argument is predicated on the belief that persons should have the right to control access to information about them, preventing its use, for example, for research purposes of which they may disapprove. Privacy concerns have been given greater weight of late with the growing ability of researchers to correlate information obtained from different databases about the same individuals, thus accumulating large amounts of information that, were it to be used or accessed for unauthorized purposes, could prove embarrassing or harmful to the people involved. Strong arguments have been made for obtaining consent from the subjects of these records prior to their release for research purposes.

Researchers object to these proposals on the grounds that they would materially increase the costs of research and, insofar as subjects' refusal of consent would bias the resulting data, decrease the validity of their findings. In many cases, it will simply be infeasible to track down and obtain consents from the persons

whose data is in question. No reasonable argument can be made for precluding researchers' access in these circumstances, so long as the study is of some significant value and appropriate safeguards are in place. When smaller and more recent repositories of data are involved, and obtaining consent would not involve unreasonable expenditures of resources, it is clearly more respectful of subjects' interests to seek their agreement. When records are being compiled prospectively with the intent of using them in the future for research purposes, there is no ethical justification for failing to obtain subjects' consent at the time.

3.2 Consent in Observational Studies and Field Work

Many of the behavioral sciences have a long tradition of sending researchers into the field to study the behaviors of persons in their natural environments. This often involves requesting access from those to be observed to otherwise private interactions and territory (e.g., discussions between spouses in the subjects' home). Considerable controversy exists over whether and how truly informed consent should be obtained in such cases. Insofar as researchers may observe and/or report embarrassing or frankly criminal behavior, subjects may be placed at risk of psychological or tangible harm. Thus, the case for obtaining consent in such studies would seem strong, bolstered by the discomfort some investigators report in withholding this information from their subjects.

On the other hand, researchers have an understandable concern that informing subjects clearly of the goals of their studies may lead subjects to alter their behaviors, thus obscuring the very phenomena that are under study. Though many subjects appear remarkably prone to ignore such information, especially once they become accustomed to the observer's presence, this concern needs to be taken seriously. It should usually be possible to accommodate both parties' interests by disclosing the intent of the study in general terms, along with the likely risks to which subjects' may be exposed. If some modification of subjects' behavior occurs, this is an unavoidable price to pay for acquiring a meaningful consent.

3.3 Consent in Community Intervention Projects

Social science research increasingly involves large-scale interventions directed at particular communities or segments of a community. These may include interventions designed to avoid undesirable outcomes (e.g., juvenile delinquency, teenage pregnancy, illegal drug use) or to promote what are generally seen as salutary ends (e.g., moving persons off welfare and

into paying jobs). When such interventions occur at the level of the community as a whole—such as broad-based educational programs concerning drug abuse—it is impossible to obtain informed consent as it is usually conceptualized. Moreover, when studies are based on changes to entitlement programs that have been enacted through the political process, the consent of individual persons may be seen as irrelevant.

There may, nonetheless, be a role in such instances for a modified form of consent interactions with the relevant community. Although each person's consent cannot be obtained, disclosure of the nature and purpose of the study and its risks and benefits can still take place. Consultation with the community at large and its designated representatives may result in suggestions for modifications in the study culminating in a general consensus concerning its desirability (Trickett 1998).

4. Research on Informed Consent

The last decades of the twentieth century have seen a flowering of empirical studies on informed consent to research that promises to shape for the better a debate that has frequently been waged on theoretical grounds alone. Among the areas of investigation have been: the adequacy of investigators' disclosures and how they can be improved; the prevalence of incapacity among various groups of research subjects; subjects' understanding of the information they have received and their appreciation of its implications for their situation; and subjects' retention of information over time (Sieber 1996).

Naturalistic research on the consent process involves overcoming the reluctance of investigators to grant access to the research setting, and of subjects to consent to yet another research project. Subject recruitment can be both difficult and time-consuming. The willingness of funding agencies to sponsor such research is often related to the degree of public scrutiny of and concern about the agencies' research portfolio. But such research is likely to increase over time and to play a greater role in helping to shape investigators' practices and official policies governing informed consent.

See also: Deception, Psychology of; Deceptive Methods: Ethical Aspects; Ethics and Psychiatry; Psychiatry: Informed Consent and Care

Bibliography

- Appelbaum P S, Lidz C W, Meisel A 1987a *Informed Consent: Legal Theory and Clinical Practice*. Oxford University Press, New York
- Appelbaum P S, Roth L H 1982 Competency to consent to research. *Archives of General Psychiatry* **39**: 951–8
- Appelbaum P S, Roth L H, Lidz C W, Benson P, Winslade W 1987b False hopes and best data: consent to research and the

therapeutic misconception. *Hastings Center Report* **17**(2): 20–4

- Beecher H K 1966 Ethics and clinical research. *New England Journal of Medicine* **274**: 1354–60
- Brody B A 1998 *The Ethics of Biomedical Research: An International Perspective*. Oxford University Press, New York
- Code of Federal Regulations (US), Title 45 Part 46. Protection of Human Subjects. Revised June 18, 1991. US Government Printing Office, Washington, DC
- Faden R R, Beauchamp T L 1986 *A History and Theory of Informed Consent*. Oxford University Press, New York
- Grodin M A 1992 Historical origins of the Nuremberg Code. In: Annas G J, Grodin M A (eds.) *The Nazi Doctors and the Nuremberg Code: Human Rights in Human Experimentation*. Oxford University Press, New York
- Jones J 1981 *Bad Blood*. Free Press, New York
- Levine R J 1986 *Ethics and Regulation of Clinical Research*, 2nd edn. Urban and Schwarzenberg, Baltimore, MD
- National Bioethics Advisory Commission 1998 *Research Involving Persons with Mental Disorders that May Affect Decisionmaking Capacity*. National Bioethics Advisory Commission, Rockville, MD
- National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research 1978 *The Belmont Report: Ethical Guidelines for the Protection of Human Subjects of Research*. US Government Printing Office, Washington, DC
- Sieber J E 1996 Empirical study of ethical issues in psychological research. In: Stanley B H, Sieber J E, Melton G B (eds.) *Research Ethics: A Psychological Approach*. University of Nebraska Press, Lincoln, NE
- The Nuremberg Code 1948 In: Annas G J, Grodin M A (eds.) *The Nazi Doctors and the Nuremberg Code: Human Rights in Human Experimentation*. Oxford University Press, New York
- Trickett E J 1998 Toward a framework for defining and resolving ethical issues in the protection of communities involved in primary prevention projects. *Ethics and Behavior* **8**: 321–37
- World Medical Association 1964, last revised 1989 *Declaration of Helsinki*. In: Vanderpool H Y (ed.) *The Ethics of Research Involving Human Subjects: Facing the 21st Century*. University Publishing Group, Frederick, MD

P. S. Appelbaum

Residential Concentration/Segregation, Demographic Effects of

Segregation is the differential distribution of social groups. Most commonly, the use of the term refers to residential segregation, the differential distribution of social groups across geography. Residential segregation has served as a key indicator of social differentiation for decades. (See *Residential Segregation: Sociological Aspects* for a general review.) Segregation is usually measured by one or more summary indices, designed to reflect the differential distribution of persons or other social units across neighborhoods, occupations, or other categories (White 1986). Segregation is often seen as a multidimensional phenomenon tapped by a variety of indices (Massey et al. 1996). While the bulk of work has centered on residential segregation

among ethnic groups, virtually any trait could be considered, and research has examined segregation by income, education, occupation, nativity, age, and other demographic traits. Racial segregation across schools and occupations remains an important indicator of social distance, and residential segregation can feed school segregation and limit access to jobs, city services, and other public goods (Farley 1984, 1996).

Beyond the measurement of segregation itself, there has always been interest in the consequences of residential clustering. Theory can be invoked to suggest positive as well as negative consequences (Cutler and Glaeser 1997). This ranges from the ways in which residential concentration might generate employment networks or social support institutions, to the consequences of residential isolation for educational attainment or health outcomes. From the point of view of policy, the determinants and consequences of segregation are of particular interest. There has been heightened attention to the aftermath of living in, or being raised in, a segregated or disadvantaged neighborhood. More recent discussions of social capital are pertinent, as the concept of social capital is extended to include the influence of neighbors and nearby peers and role models on such outcomes.

Empirical work at the beginning of the twenty-first century has begun to evaluate the effects of residential segregation. The underlying concept is that residential clustering carries behavioral consequences beyond those of ethnicity, socioeconomic status, or lifestyle. Such notions have been present in demography for some time, but data and multilevel statistical techniques are now available to directly evaluate hypotheses. Social behaviors clearly vary across urban space, or equivalently other categorizations. Yet it is important to distinguish the influence of composition from that of causation. For instance the receipt of public assistance varies appreciably across urban neighborhoods. The degree to which this behavior can be linked causally to the concentration of social groups in neighborhoods (and the sociological patterns that arise through clustering) needs to be weighed against an explanation based just on composition. The key question turns on whether geographic clustering itself influences human behavior, beyond that attributable to other background traits, such as age and education.

Many demographic traits are segregated in urban neighborhoods, and there is a consistency to the ranking. Although empirical results on these matters derive preponderantly from North America, similar results have been obtained for other high-income societies. Race and ethnic groups remain very highly segregated in most urban areas. Several decades of study, from straightforward segregation measures through factorial ecology and the like, have verified the sorting of urban space along the lines of socio-

economic status. Life cycle stage is also differentiated. Segregation indices for sex and age are modest, although segregation for household size and type is more discernable. Of course, certain special concentrations of population, such as group quarters for military populations, students, and those under correctional supervision, are highly segregated and are often excluded from consideration by analysts.

While one can identify and rank the levels of segregation illustrated by various traits, the consequences of this segregation are more difficult to discern. A common causal model advanced in these cases was that persons, especially children, adapted to their disadvantaged environment, learning criminal behavior, reducing achievement motivation for school, and developing dependence on the dole. Observers cited the higher incidence of single parenthood, labor force participation, high school dropout, and the like in such disadvantaged, inner-city neighborhoods.

A demonstration of the presence or absence of the links remained for subsequent waves of investigations. Statistical estimates of neighborhood effects rely on hierarchical and multilevel models. While strategies differ, the general approach is to develop a multivariate model, in which the behavioral and demographic outcome (employment, childbearing) is taken to be first a function of unchanging (ascriptive) personal traits and a set of family and household background factors. Neighborhood and community characteristics are introduced to the model simultaneously or hierarchically. Thus, neighborhood effects are generally measured net of other characteristics. With the availability of multilevel statistical techniques and associated software, it is possible to derive more sophisticated (and econometrically appropriate) estimates of these neighborhood or community effects. Some of the same conceptual and methodological apparatus can be applied to patterns or demographic unevenness found in other social institutions, such as the workplace or schools.

1. Mechanisms Through Which Segregation May Influence Demographic Outcomes

Beyond the simple mechanism of demographic composition itself, segregation, and differential concentration might influence outcomes in several ways. These include (a) contagion effects; (b) resource allocation; and (c) institutional structures. Contagion models, also closely linked to epidemic effects and diffusion models, have a long history in social science thinking. Simply stated, individuals alter their behavior on the basis of their interpretation of those around them (Jencks and Mayer 1990, Crane 1991). The influence of peers on child behavior is one such mechanism. Such processes have been particularly prominent in thinking about school context (Alwin

1977). Role models provide a related mechanism. Much of the recent discussion of adverse inner city outcomes draws on this line of thinking. The general reasoning (see, e.g., Wilson 1987) is that low income or disadvantaged neighborhoods offer fewer mainstream role models (dual parent families, stable adult workers, school completion), and so those growing up in such areas do not acquire mainstream norms. Massey and Denton argue that the intersection of poverty and segregation creates an environment where such conditions abound (Massey and Denton 1993). Social capital is a contemporary phrase that taps some of the same mechanisms. Resources available to the individual beyond the family or household may influence a wide range of behavioral outcomes. Certainly this is of interest in schooling and neighborhoods, and recently it has become of increasing concern in the field of health.

A second line of thinking, far less readily articulated, involves resource allocation. Residents of affluent neighborhoods (or workers in advantaged workplaces) can provide superior public goods, either through direct community provision or through the political process. More than that, these public goods are likely to confer advantages that go beyond the immediate advantages of household and family. Thus, superior schools produce a benefit for all residents in the district (neighborhood) over and above the advantage of family background. Other externalities of clustering include relative advantage or deprivation in local sanitation (exposure to pathogens and toxins), public safety, and the like. Housing experiments, such as those relocating low-income individuals to middle-income communities, are premised on such mechanisms.

The third line of thinking involves the growth and maintenance of local institutions, whether formal or informal (Duncan 1994). The literature identifies both positive and negative consequences of segregation. In the realm of economic outcomes, the notion of the ethnic enclave suggests that concentration, building on ethnic networks, may provide benefits. Occupational and residential segregation with co-ethnics may generate employment opportunities unavailable in a more integrated economy. A long-standing argument is that racial residential segregation enables integration by social class. In an urban neighborhood, local institutions, such as a church or neighborhood association, might foster social ties across class boundaries, although even historical accounts of ghetto life do not always portray a comfortable coexistence of disparate groups. What is relevant in the present circumstance is that these evolving structures can serve as social actors and then influence behavioral outcomes for other members of the community. Empirical social science evidence on these points is scant, especially evidence that can tease out the influences of institutions, net of individual, family, and other community traits.

2. Methodological Issues

Concerns about the effects of segregation raise several methodological issues. A fundamental issue is one of definition. A 'neighborhood' might have several competing geographic boundaries. Tracts, wards, block groups, and other urban statistical subdivisions become *de facto* neighborhoods (White 1987). This has two consequences. First, such units serve as the basic geography for calculations of segregation measures. Second, these small statistical areas are the units that serve as the proxy for social conditions in multilevel models. Another methodological evolution is the shift from ecological inference to multilevel modeling. For decades of social science research, analysts were constrained to a choice of individual level analysis or 'ecological' models, which relate aggregate values of demographic characteristics to one another. While ecological analysis was the only route available for certain questions, and while some might argue for the focus on the neighborhood only, multilevel analysis has grown over the last several years as the favored approach (Diez-Roux et al. 1997, Garner and Raudenbush 1991). In multilevel models, individual behaviors are taken to be the result of both individual characteristics and traits of an aggregate. The level of segregation in the city, or the demographic composition of the neighborhood, could be one of these aggregates.

It remains a challenge to properly specify models, so that individual and contextual effects are fully distinguished statistically. The very process of residential assimilation—the achievement of neighborhood status through residential mobility—also generates correlations between community and individual. Thus, separating endogenous and exogenous elements of a process is no easy task. Even then, one must still be concerned that events are properly allocated to local administrative units.

3. Empirical Evidence

Interest in community effects in urban settings was rekindled with the discussion and debate over the urban 'underclass' in the 1980s social science literature. Again much of this debate was focused on US settings, but parallel issues of urban neighborhood deterioration and behavioral consequences can be raised for other high-income nations. Despite the level of concern for the effects of residing in ghettos and inner-city disadvantaged zones, one major contemporary review argued that national social and economic conditions more powerfully determined the incidence of and concentration of poverty (Lynn and McGeary 1990). After expressing reservations about the quality and consistency of data, a related recent review of empirical studies of the effects of neighborhood SES, racial mix, for both neighborhoods and schools,

concluded that there is no general pattern of effects across a wide range of outcomes (Jencks and Mayer 1990). There is some modest evidence that growing up with affluent neighbors (for those of modest means) and in less segregated environment (for blacks) predicts higher levels of educational attainment (Jencks and Mayer 1990). One recent review of about 24 studies of neighborhood effects did tend to find that community-level (often census tract) measures of neighborhood advantage were associated with superior individual-level behavioral outcomes in several phases of childhood (Gephhardt 1997). Not surprisingly, this review was critical of the state of knowledge, for the weakness of the link to the appropriate community-level empirical measures and the absence of other characteristics (school, family traits) that would be likely to exert an influence on individual outcomes. One comprehensive compendium of neighborhood level studies on poverty, a study that coordinated data and research approach, again detected neighborhood effects in several investigations (Brooks-Gunn et al. 1997). Typically family-level measures swamped neighborhood-level measures in predicting child behavioral outcomes. Perhaps more interesting is the finding that neighborhood measures could be discerned for pre-school children, despite the conventional thinking that such youngsters' social world was limited to the home.

Substantial effects of neighborhood factors, particularly neighborhood affluence and poverty, have been linked to later life achievement (IQ, teenage childbearing, and school-leaving) in US nationally representative longitudinal data (Brooks-Gunn et al. 1993). Longitudinal data only sometimes detect effects of neighborhood socioeconomic status and racial composition on later earnings, after controlling for selected background traits (Jencks and Mayer 1990). Conceptual models of the influence of neighborhood effects concentrate on the pathways through which these aggregate effects carry to the individual. At the same time qualitative approaches can suggest coping mechanisms for those in disadvantaged neighborhoods (Jarrett 1997). There is some empirical evidence that local organizational structure may mediate community disadvantage (Elliott et al. 1996). The effect of schools may outweigh neighborhoods on risky health behavior (Ennett et al. 1997).

One reasonably long-standing line of work looks at the influence of segregation itself, often introducing the value of a segregation statistic to a regression model of a behavioral outcome. The behavioral mechanism underlying the inclusion of such a covariate is not always clear, but higher level residential segregation may tap into underlying housing market constraints, a more hostile climate of local race relations, or differential access to public sector resources. Predictive power of such segregation indices is not always strong, but on balance they tend to show that individuals (usually minority group members)

living in more segregated urban areas experience detrimental outcomes, such as higher levels of unemployment, or violence, or less mobility out of poverty (Iceland 1997, Peterson and Krivo 1993, South and Crowder 1997, Wilson et al. 1995). Even after including several individual and neighborhood controls in a multilevel model, residential segregation does appear to be associated with lower residential mobility on the part of both blacks and whites, supporting the housing market constriction idea (South and Dean 1993). Neighborhood concentration and metropolitan racial segregation, especially among African Americans does appear to be linked to adverse health outcomes for both adults and infants (Williams and Collins 1999, Hart et al. 1998, LaVeist 1989, LeClere et al. 1997, Polednak 1996). There is no consensus on the scale of neighborhood effects, and the magnitude of such effects would be expected to differ by behavioral outcome. The current state of knowledge is that neighborhood effects on demographic outcomes can be detected—especially when investigators use more robust data and more contemporary methodology—yet these neighborhood effects are modest by comparison to personal fixed traits and household characteristics.

See also: Neighborhood: General; Neighborhood Revitalization and Community Development; Poverty, Sociology of; Residential Segregation: Geographic Aspects; Segregation Indices; Urban Life and Health; Urban Sociology; Urban Studies: Overview.

Bibliography

- Alwin D 1977 High school context effects on aspirations. *Sociology of Education* **49**: 294–303
- Brewster K L 1994 Neighborhood context and the transition to sexual activity among young black women. *Demography* **31**: 603–14
- Brooks-Gunn J, Duncan G J, Aber J L (eds.) 1997 *Neighborhood Poverty*. Russell Sage Foundation, NY, Vols I and II
- Brooks-Gunn J, Duncan G J, Klebanov P K, Sealander N 1993 Do neighborhoods influence child and adolescent development? *American Journal of Sociology* **99**: 353–95
- Crane J 1991 The epidemic theory of ghettos and neighborhood effects on dropping out and teenage childbearing. *American Journal of Sociology* **96**: 1226–59
- Cutler D, Glaeser E L 1997 Are ghettos good or bad? *Quarterly Journal of Economics* **112**: 827–72
- Darden J T, Bagaka's J G, Ji S J 1997 Racial residential segregation and the concentration of low- and high-income households in the 45 largest US metropolitan areas. *Journal of Developing Societies* **13**: 171–94
- Diez-Roux A V, Nieto F J, Muntaner C, Tyroler H A, Comstock G W, Shahar E, Cooper L S, Watson R L, Szklo M 1997. Neighborhood environments and coronary heart disease: A multilevel analysis. *American Journal of Epidemiology* **146**: 48–63

- Duncan G 1994 Families and neighbors as sources of disadvantage in the schooling decisions of white and black adolescents. *American Journal of Education* **103**: 20–53
- Elliott D, Wilson W J, Huizinga D, Sampson R J, Elliott A, Rankin B, 1996 The effects of neighborhood disadvantage on adolescent development. *Journal of Research in Crime and Delinquency* **33**: 389–426
- Ennett S, Flewelling R L, Lindrooth R C, Norton E C 1997 School and neighborhood characteristics associated with school rates of alcohol, cigarette, and marijuana use. *Journal of Health and Social Behavior* **38**: 55–71
- Farley R 1984 *Blacks and Whites*. Harvard University Press, Cambridge, MA
- Farley R 1996 *The New American Reality*. Russell Sage, New York
- Garner C L, Raudenbush S W 1991 Neighborhood effects on educational attainment: A multilevel analysis. *Sociology of Education* **64**: 251–62
- Gephardt M 1997 Neighborhoods and communities as contexts for development. In: Brooks-Gunn J, Duncan G J, Aber J L (eds.) 1997 *Neighborhood Poverty*. Russell Sage Foundation, NY, Vols I and II, Chap. 1, pp. 1–43
- Hart K D, Kunitz S J, Sell R R, Mukamel D B 1998 Metropolitan governance, residential segregation, and mortality among African Americans. *American Journal of Public Health* **88**: 434–8
- Iceland J 1997 Urban labor markets and individual transitions out of poverty. *Demography* **34**: 429–41
- Jarrett R 1997 African American family and parenting strategies in impoverished neighborhoods. *Qualitative Sociology* **20**: 275–88
- Jencks C, Mayer S E 1990 The social consequences of growing up in a poor neighborhood. In: Lynn L E, McGeary M G H (eds.) *Inner City Poverty in the United States*. National Academy Press, Washington, DC
- LeClere F, Rogers R, Peters K D 1997 Ethnicity and mortality in the United States: Individual and community correlates. *Social Forces* **76**: 169–98
- LaVeist T A 1989 Linking residential segregation to the infant-mortality race disparity in US cities. *Sociology and Social Research* **73**: 90–4
- Lynn L E, McGeary M G H (eds.) 1990 *Inner City Poverty in the United States*. National Academy Press, Washington, DC
- Massey D S, Denton H 1993 *American Apartheid*. Harvard University Press, Cambridge, MA
- Massey D S, White M J, Phua V C 1996 The dimensions of residential segregation revisited. *Sociological Methods and Research* **25**: 172–206
- Peterson R D, Krivo L J 1993 Racial segregation and black urban homicide. *Social Forces* **71**: 1001–26
- Polednak A P 1996 Trends in US urban black infant mortality, by degree of residential segregation. *American Journal of Public Health* **86**: 723–6
- South S, Crowder K 1997 Escaping distressed neighborhoods: Individual, community, and metropolitan influences. *American Journal of Sociology* **102**: 1040–84
- South S J, Dean G D 1993 Race and residential mobility: Individual determinants and structural constraints. *Social Forces* **72**: 147–67
- White M J 1986 Segregation and diversity measures in population distribution. *Population Index* **52**: 198–221
- White M J 1987 *American Neighborhoods and Residential Differentiation*. Russell Sage, NY
- Williams D, Collins C 1999 Segregation and mortality. *Sociological Forum* **14**: 495–523
- Wilson F, Tienda M, Wu L 1995 Race and unemployment: Labor market experiences of black and white men, 1968–1988. *Work and Occupations* **22**: 245–70
- Wilson W J 1987 *The Truly Disadvantaged*. University of Chicago Press, Chicago

M. J. White

Residential Environmental Psychology

Homes and residences are not merely physical and economic structures. They are centers of psychological and cultural life. Home is the most important place in many lives, which affords it great potential for facilitating or threatening human development. People think about their homes, are satisfied or not with them, plan, dream, move, use them in various ways, and often must cope with difficult social and physical conditions in their homes, if they have one.

1. Definitions

A house is not necessarily a home. A *residence* (house, apartment, condominium, loft, flat, tipi, or, in general, any dwelling) is a physical structure that varies tremendously in form around the world. *Home* is the rich set of evolving meanings that people often (but not always) attach to their residences (Lawrence 1987). Residences may be bought, sold, or rented; homes cannot.

1.1 Background

Vernacular residences in traditional societies gradually evolved to suit local building materials, climates, and social practices. In recent times, awareness of other cultures, technological advances, and prosperity in industrialized countries have permitted more choice in residential form and structure. The psychological aspects of residences became relevant, and with the growth of environmental psychology and sociology in the late 1960s, researchers began to examine the impact of residences on residents.

1.2 Dimensions of Residences and Homes

Residences vary in any number of physical aspects such as form, age, style, size, or cost. However, residences may be characterized in more psychological and cultural ways (Altman and Werner 1985). First, residences may be permanent or temporary. In industrialized countries, residents usually have permanent dwellings, although apparently this situation is not entirely satisfactory because many householders also maintain temporary residences such as cabins, cot-

tages, and recreational vehicles. In less-industrialized countries, residents move frequently or seasonally and often construct temporary dwellings in new places.

Second, residences range from differentiated to homogeneous, which refers to the separation, or lack of it, in the functions of rooms. A highly differentiated residence has many rooms, each of which houses a specific activity; in a homogeneous residence nearly any activity may occur in nearly any room.

Third, residences may be communal or noncommunal, in terms of the degree to which nuclear families live together or in different homes. In contrast to the typical single-family dwelling of the Western world, in many Eastern cultures several generations of a family live in one household. In some African and Native American cultures, numerous unrelated families live in one dwelling.

Fourth, identity versus communality refers to the extent to which residences are contextually unusual. Communality (the same word as above, but with a different meaning here) is the extent to which a residence is physically typical of dwellings in that culture. A tipi on a suburban lot in North America would be low on communality, but high on identity; the same tipi at a pow-wow on the open plains of central North America would be high on communality.

Fifth, residences vary in their openness to outsiders. This varies both within and between cultures. In some places, houses typically are surrounded by walls or hedges; in others they are not. Dwellings signal willingness or unwillingness to interact with outsiders.

Home, in contrast to residence, is a set of personal meanings. The word's earliest meaning probably was the village or community of one's origin; not until the sixteenth or seventeenth centuries did it refer to one's dwelling. To be a home, a residence should provide haven, order, identity, connectedness, interpersonal warmth, and physical suitability (Smith 1994). As a haven, a home must surround a person with privacy, security, and protection from the slings and arrows of life outside it.

A home provides order by helping individuals to know their place in the world. As a center from which people venture and return, home orders one's existence in the world, both spatially and temporally. Home is central to a sense of identity and connectedness because it offers a sense of kinship, belonging, and status. Through self-expression and personalization, the home comes to resemble or represent the self. When people have the means and the time to personalize their residence, home becomes a symbol of self (Cooper 1976). To the extent a residence does not provide these felt amenities, it is not a home.

1.3 Houselessness

The homeless, those who live in the streets, properly should be called houseless. Many houseless people

have a home, a residence with personal meaning that—for one reason or another—they are unable to live in. Home is the most important refuge from the stresses of work, school, and public life. Unfortunately, many people live in residences that do not provide protection and rejuvenation, and therefore they are homeless. Probably one billion people have inadequate housing, and about 100 million are houseless. Institutionalized individuals often are in this situation. Houselessness is more than lack of a residence; many other social negative social consequences follow. Individuals who are both houseless and homeless do exist, and are extremely unfortunate.

2. Residential Preferences and Satisfaction

Residential preference and satisfaction are related but distinct concepts. Their measurement is not straightforward. Some residences would dissatisfy anyone, and there are some individuals who will not be satisfied by any residence. Between these extremes, however, lie most homes and most residents. Different homes satisfy different residents. Preference involves comparison, whereas satisfaction may be mainly based on residents' evaluations of their own dwellings.

2.1 Measurement Issues

Residential satisfaction is multidimensional. David Canter (1983) has described two cognitive processes related to residential satisfaction, *purposiveness* and *comparison*. Purposive evaluations have three facets: level, referent, and focus. Level, or specificity, refers to evaluating a single part of the residence (such as the kitchen), or a larger portion of it (such as the bedrooms as a whole, or the entire residence). Referents refer to different qualities of the residence (its beauty, or spaciousness); each suggests a certain purpose which is served well or poorly. The lighting referent, for example, is important when the resident considers tasks such as reading or sewing that require good lighting. Focus depends on the referent, and refers to how broadly the resident is asked to evaluate a referent. For example with lighting, focus might refer to the ability of a particular lamp to light a study desk, or about the ability of all the residence's lamps to light the home as a whole.

Purposiveness arises out of the resident's relationship to the residence, such as long-term vs. short-term residence, ownership, renting, or landlordship; these different purposes will make important differences in the resident's evaluation. In general, a resident's purpose in doing an evaluation is itself a crucial factor in satisfaction.

The second factor is comparison. Whether observers are explicitly asked to compare a specific residence against a series of others, or merely asked for a rating

of a particular residence without explicit comparison with others, they cannot help at least implicitly comparing the residence in question with others in their experience. Thus, different observers will be comparing the residence in question against different houses in their experience.

Residents who are unaware that their friends live in mansions might be more satisfied with their huts. Residents compare residences, and this can be an important source of satisfaction or dissatisfaction. A study of Army personnel who lived in mobile homes and apartments found that comparison with other housing accounted for much of the variance in their housing satisfaction.

Perceived physical qualities of a residence (such as its appearance) are important predictors of satisfaction, and so are preferred physical qualities. However, the difference between perceived and preferred qualities is a separate but important predictor of housing satisfaction.

2.2 Personal, Social, and Physical Influences

Residential satisfaction is influenced by many personal characteristics, including age, stage in the life cycle, socioeconomic status, personality, values, and hopes for the future. Social factors such as age and community norms, others' opinions, and neighbors also influence satisfaction. Gender and sex roles are important. In a study of Canadian households that had recently moved, for example, husbands and wives were asked to list the things that made them happy about their new homes and apartments (Michelson 1977). Among couples who moved to a downtown apartment, more than 40 percent of the wives mentioned that they were satisfied with the quality of the home, but only 5 percent of the husbands mentioned this. On the other hand, among couples that moved to suburban apartments, more than 50 percent of the husbands mentioned quality of the home as a source of their satisfaction but only 5 percent of the wives did.

As men enter the kitchen more, similarities and differences in men's and women's preferences for kitchen designs become more important. Both men and women prefer kitchens that open to dining and family room areas over kitchens that are separated from those areas. However, women who work outside the home prefer open kitchens more than women who do not work outside the home and, despite their agreement on the open-closed aspect of the kitchen, men and women differ in many other kitchen spatial preferences.

Residential satisfaction is also related to socioeconomic status. In general, of course, wealthier individuals are better able to supply themselves with homes that meet their standards, and thus are more satisfied. However, wealthier and poorer residents may be attuned to different features of their homes:

One study found that wealthier individuals were more sensitive to the aesthetic qualities of their home, whereas poorer individuals were more sensitive to safety, health, and family needs.

Naturally, physical and architectural factors such as house form, architectural style, floor plan, and colors affect residential satisfaction, but these should not be considered the sole, or even the main, source of housing satisfaction. Most individuals prefer higher (than the usual eight-foot) ceilings, flat or sloping ceilings (4-in-12 slope but neither more than that nor smaller slopes), and walls that meet at 90 degrees or more (not walls that meet at smaller angles). In general, highly detailed designs are favored over plain and simple designs. In places where fear for personal security is high, enclosed and guarded enclaves are highly valued, as shown in a recent Brazilian study (Carvalho et al. 1997).

Color has been the subject of much research, but most of it has been based on preferences for small patches of color that resemble paint chips, in no particular context (that is, subjects are not usually asked if they would paint their kitchen or bedroom that color). These findings may not apply to choices in real-life situations. A study in Japan tried to overcome these limitations by showing slides of living rooms that were painted different colors (Kunishima and Yanase 1985). Preference is not merely a matter of hue (red vs. blue vs. green); it also depends on the hue's saturation (or 'denseness') and brightness (from light to dark). The study concluded that each of these three dimensions of color was related to a different aspect of preference. Hue was not strongly related to preference; instead it was primarily related to perceived warmth (reds are warmer, blues and greens cooler). Saturation (regardless of hue) was most closely related to what we usually call preference: More saturated hues were evaluated as more elegant, more comfortable, and better. Brightness or lightness (regardless of hue) was related to how 'active' the room seemed: Brighter colors were rated as fresher, lighter, and more cheerful than darker hues.

There remains a general preference for the single-family dwelling that has persisted in North America for many years. As for style, some trends are evident: In one study, farmhouse and Tudor style houses were preferred to saltbox and Mediterranean styles (Nasar 1989). However, preferences for style clearly change in part with changes in fashions.

2.3 Cultural Variations

Residences clearly vary in form around the world. These variations are not merely decorative, they are psychologically meaningful. For example, an examination of house form across 73 cultures revealed that the extent of partitioning indoor space (differentiation) is related to the degree of sociopolitical complexity in

the culture, that is, how much segmentation, specialization, and hierarchical ordering is typical of the culture. Thus, a worldwide trend is for our dwellings to mirror the complexity of the society in which it is embedded.

3. Behavior in Residences

The environmental psychology of dwelling interiors is not very developed because the study of behavior in and round homes understandably meets with a resistance on privacy grounds.

3.1 Vernacular Residential Design

Residents arrange and decorate their interiors according to certain patterns that reflect such dimensions as simple–complex, conventional–unusual, and rich–plain decor, and messy–tidy upkeep. These patterns are related to culture, social class, and marital or living arrangement differences. For example, French and Italian living room arrangements were examined (Bonnes et al. 1987). French living rooms could be described using three dimensions—decoration, functional organization, and structuring of space. For example, at one end of the decoration dimension were living rooms that were low on ambiguity and complexity; these living rooms tended to have velvet antique couches with armrests, copies of period wallpaper, and books for show (not use) in a wall-unit that also included cupboards and a bar. Italian living rooms could be described using the same three dimensions, but they tended to be combined differently.

3.2 Use of Interior Spaces

Men, women, and children use the interior dwelling spaces unequally; sex role changes have not changed the traditional use of rooms much; women still spend more time in the kitchen (Ahrentzen et al. 1989) and more time with children, and men spend more time in leisure activities. Children (until the teen years) have restricted territories: Parents generally can enter their space, but children generally cannot enter their parents' space.

In a rare study that actually monitored behavior inside homes (as opposed to asking residents to report their activities), cameras were placed (with the residents' permission) in New York homes (Ashcraft and Scheffen 1976). The study found that although definite territories exist within the home, the degree to which the boundaries of these territories were respected varied considerably. For example, parents may feel they have the right to enter a teen's bedroom, but do not think the teen should enter their bedroom. But

both territories are sometimes intruded upon, which creates conflict.

When space is limited, residents use time scheduling, territoriality, and social agreements to manage household space. Half our leisure time is spent in the home, but it could be better spent. When teens spend more time sharing leisure with their parents, they report greater well-being. More TV time is associated with less time spent on intellectual and cultural pursuits.

4. Coping with Residential Shortcomings

Much housing is substandard, and there are consequences. Well-executed recent research conducted in the USA suggests that poor-quality housing adversely affects the mental health of both adults (Evans et al. 2000) and children (Evans et al. in press).

4.1 High Density

High density in the residence has been shown to have many adverse effects. High-density homes as a whole have more psychological distress and psychiatric illness and have been associated with increased substance abuse, juvenile delinquency, fertility and mortality rates, and may lead residents to believe they have less social support. Crowding at home has more serious consequences than crowding in public environments. Some of these are mitigated by social support, if the resident has any. Density appears to have negative effects apart from poverty. Very low densities (living alone) also cause problems for many people.

4.2 Housing Forms

Not every housing form is good for every resident. High rises probably are not good for families or children, and one study suggests that young children are better off if they attend a good child care center than if they stay at home. Apartment living has been associated with numerous childhood afflictions, including retarded movement skills, respiratory diseases, aggression, insomnia, nervous disorders, reduced social skills, and disrupted play. If all this is true, we may wonder how any child can survive in an apartment. However, many of these studies lack experimental rigor; their results must be questioned.

5. Residential Mobility

All of us must unwillingly adapt to new dwellings sometimes; at other times, we choose to relocate. In any given five-year period, about half of all North Americans and Australians move, as do over one-third of Britons and Japanese (Stokols et al. 1983).

People move for many reasons, including both negative factors such as stress from and dissatisfaction with the current residence, but also for positive reasons such as improved finances. Life cycle changes that add or subtract family members living at home can also spur mobility. Mobility is not necessarily pathogenic, then, but it can be, depending on one's residential history, present circumstances, and hopes for the future. For example, among those who move less often, those with less choice about moving report more illness-related symptoms than those who have greater choice about moving.

Mobility-related stress is moderated by personal characteristics; people who are highly mobile and those who were generally less inclined to explore the environment report more illness than those who are more exploratory by nature. Conversely, more illness occurs among residents who move less often but are unhappy with their present residence and feel they have little choice of residence. Thus, a lack of mobility also can be damaging—when one's options for moving are restricted.

6. Conclusions

Home usually is the most important physical setting. The physical building must be distinguished from the meaning structure (home). Residential satisfaction is complex and multiply determined, but life cycle stage, socioeconomic status, personality, values, and hopes for the future are important factors, as are physical features such as house form, architectural style, floorplan and colors, outdoor areas, and cultural background. People arrange residential interiors in nonrandom patterns that reflect culture and lifestyle. High indoor density is difficult for virtually anyone, but can be made less intolerable with social support. Moving is stressful or not depending on whether a person has some choice in doing so, likes to explore new settings, and likes the present residence.

See also: Environmental Psychology: Overview

Bibliography

- Ahrentzen S, Levine D W, Michelson W 1989 Space, time, and activity in the home: A gender analysis. *Journal of Environmental Psychology* **9**: 89–101
- Altman I, Werner C M (eds.) 1985 *Home Environments*. Plenum Press, New York
- Ashcraft N, Scheffen A 1976 *People Space*. Anchor, Garden City, New York
- Bonnes M, Giuliani M V, Amoni, F, Bernard Y 1987 Cross-cultural rules for the optimization of the living room. *Environment and Behavior* **19**: 204–27
- Canter D 1983 The purposive evaluation of places: A facet approach. *Environment and Behavior* **15**: 659–98

- Carvalho M, George R V, Anthony K H 1997 Residential satisfaction in condominiums exclusivos (gate-guarded neighborhoods) in Brazil. *Environment and Behavior* **29**: 734–68
- Cooper C 1976 The house as a symbol of the self. In: Proshansky H, Ittelson W H, Rivlin L G (eds.) *Environmental Psychology*, 2nd edn. Holt, Rinehart and Winston, New York, pp. 435–48
- Evans G W, Saltzman H, Cooperman J in press Housing quality and children's socioemotional health. *Environment and Behavior*
- Evans G W, Wells N M, Chan H-Y E, Saltzman H 2000 Housing quality and mental health. *Journal of Consulting and Clinical Psychology* **68**: 526–30
- Kunishima M, Yanase T 1985 Visual effects of wall colors in living rooms. *Ergonomics* **28**: 869–82
- Lawrence R J 1987 *Housing, Dwellings, and Homes: Design Theory, Research, and Practice*. Wiley, New York
- Michelson W 1977 *Environmental Choice, Human Behavior and Residential Satisfaction*. Oxford University Press, New York
- Nasar J 1989 Symbolic meanings of house styles. *Environment and Behavior* **21**: 235–57
- Smith S G 1994 The essential qualities of a home. *Journal of Environmental Psychology* **14**: 31–46
- Stokols D, Shumaker S A, Martinez J 1983 Residential mobility and personal well-being. *Journal of Environmental Psychology* **3**: 5–19

R. Gifford

Residential Segregation: Geographic Aspects

Residential segregation refers to the practices and ideologies by which socio-economic characteristics are expressed in, and shaped through, the organization of residential space.

1. The Division of Residential Space

Residential space is a geographically uneven resource, and the land and housing which constitute it vary in quality, character, condition, and value. Land values vary according to demand and have traditionally been highest in city centres. Housing characteristics vary according to the circumstances of housing production and use, including the macroeconomic climate, the legislative and regulatory framework, the fortunes of the construction industry, the availability of loans and subsidies, changing architectural fashions, and so on.

A primary determinant of access to housing (and neighborhood) is cost, but the markets and institutions that make up the housing system also allow a range of social characteristics to mediate the process of inclusion within, or exclusion from, different parts of the residential environment. It is this interaction of the spatial structure of the housing stock with social structures, property values and with the operation of

commercial and public institutions that gives rise to the geography of residential segregation. These processes are, furthermore, embedded in, and legitimized through, the wider set of entitlements, obligations, and ideologies built into the constitutions of states and societies.

Housing and the residential space it comprises is both a commodity and a welfare resource. Housing is bought and sold in the market place, but housing policy can also be a means by which states intervene to meet social needs and alleviate social inequality. Both market forces and governmental priorities therefore play a part in structuring residential space; and such space is fractured in many ways. This fracturing occurs because of differences in 'need' and in 'ability to pay,' and it occurs too because the processes governing access to housing tend to discriminate on the basis of social characteristics. Direct and indirect discrimination has been documented in all sectors of the housing system. In the housing market, access to housing opportunities generally, and to housing finance in particular, may be dictated as much by who people are (or are presumed to be), by what they do, and by where they currently live, as by their actual ability to pay. In the public or social sector, the principle of allocating homes according to 'need' is always hard to operationalize and in practice may be compromised by, for example, subjective discretionary judgments about what types of people 'deserve' what kind of housing, and where.

As a consequence of all this, while class, income, wealth, and other economic indicators provide one set of fault lines in the residential landscape, ideas about 'race,' ethnicity, religion, and language provide another. Health inequalities are also expressed and reproduced in the residential environment; and although most research on segregation refers to the experience of 'households' (implicitly lumping men, women, and children together, and often referring mainly to the experience of men), the division of residential space is also fundamentally gendered and sexualized.

2. Measurement: The 'Index Wars'

A strong empirical tradition underpins the literature on residential segregation, and this has become the province of geographers (see *Social Geography*) and of those whom Peach (1975) terms 'spatial sociologists.' The idea implicit in all this work is that there is some kind of link between society and space; that social processes literally 'take place'; and that where people live is both an expression of who they are and a factor affecting who they can become.

Empirical studies of residential segregation are dogged by the problems of measurement and interpretation. The problem of measurement is apparent

from the large literature comprising the so-called index wars. One debate concerns the appropriate index with which to measure segregation. This has revolved around a group of indices of segregation and dissimilarity, which were favored for their simplicity and apparent (though questionable) scale independence. The index of dissimilarity (ID), for example, measures the difference in the distribution of any two population groups over a given number of areal units (e.g., census districts). ID is calculated very simply as half the sum of the difference in the proportion of each group in each areal unit. The index varies from 0 (complete similarity) to 100 (complete segregation) and the value of the index refers to the proportion of either group that would have to move from one areal unit to another in order for the two distributions to be the same.

ID is a special form of the more general index of segregation (IS) which measures the residential dissimilarity of one group from the rest of the population. Several other indices can be derived from this to capture the spatial segregation and isolation of one or more population groups from one another. Perhaps the best known of these is the asymmetric index of segregation, P^* , which describes the spatial isolation of a social group and provides a measure of potential interaction between that group and another (Lieberman and Carter 1982).

A second point of dispute in the measurement of residential segregation concerns whether the degree of segregation should be gauged against a norm of evenness or randomness. If populations were randomly distributed across space, some clustering would be expected as part of this random pattern. The index of dissimilarity refers to the proportion of a population that would need to move to achieve evenness, and it may therefore exaggerate the amount of segregation that is enforced (or chosen). There is still no consensus on the best way forward here. Nevertheless, all these debates are dwarfed by the problems of how to conceptualize the groups in the study, of whether to standardize (e.g., for income when measuring 'racial' segregation), and, crucially, of how to interpret the results of applying the measures.

3. Interpretation: The 'Choice-constraint' Debate

A major debate over the interpretation of measures of residential segregation concerns the question of causality. It hinges around the extent to which people choose to live separately from one another, and the extent to which segregation is forced upon them. This has become known as the 'choice-constraint' debate. In the UK, for example, all the major cities display enduring patterns of residential segregation between the white, African-Caribbean and (South) Asian populations. However, within the Asian populations,

Sikhs, Hindus, and Muslims tend to live in different neighborhoods to one another, while West Indians in London are segregated according to the island of origin of those who are migrants. Given these findings, it might plausibly be argued that the residential separation of Hindus and Muslims, or of Jamaicans and Trinidadians, is a matter of choice. Indeed in the USA, Clark (1992) draws on Schelling's (1971) ideas about the aggregate effects of small variations in residential preferences to argue that, in the wake of declining institutional discrimination, the accumulation of own-group residential choices is a key factor sustaining current patterns of segregation in US cities (see *Spatial Choice Models*).

While most people do have some choice over the precise space they occupy in the residential environment, the fact that, for the poorest people, these choices are highly limited, leads many authors to focus more on sources of 'constraint' than on elements of 'choice.' Smith (1987, 1989), for example, has argued that while it might well be the case that intragroup segregation is the result of household preferences, it is hard to sustain the argument that a very substantial proportion of such a diverse group as, say, South Asians in Britain, would choose to live in neighborhoods that are on average cheaper, poorer, less well serviced, and more deprived than any others. The demonstrable links between segregation and inequality, the marked extent to which poverty overlaps with (or is constructed through) these spatial processes of marginalization and exclusion, suggests that households segregated into the least desirable residential locations have very little room for maneuver.

Despite a number of conceptual and methodological problems, the empirical tradition has established beyond question that residential segregation is a key feature of the residential landscape, and that, preferences notwithstanding, patterns of residential segregation tend more often than not to be linked systematically with patterns of social inequality (see *Spatial Equity*). Some have suggested that this means that residential segregation has the capacity to act as a catalyst in the process of class formation or as a springboard for collective action more generally. However, in one of the most influential theoretical accounts of the relationships between class structures and residential differentiation, David Harvey (1985) argues that residential segregation is part of the reproduction of social relations within capitalist societies. From this perspective the fracturing of populations into residential communities mainly serves to fragment the class consciousness which might otherwise be used to improve the quality of local life.

Some of these fragmenting, isolating effects of segregation are uncovered by Pratt and Hanson (1994) in a study of women workers in Worcester, Massachusetts. They show how residential neighborhoods are differentiated by the distinctive mix of 'race,' class, gender, and work traditions within them.

This mix influences the locational decisions of employers, who work with stereotypes about the differences between men and women's work, and about the suitability of different ethnic groups for particular types of task. Employers' locational strategies thus reinforce a spatial division of labor which already emphasizes the gender divide. But because this division of labor also operates along class and ethnic lines, the potential for solidarity among women is undermined, despite their common interests (see *Gender and Feminist Studies in Geography*). Residential segregation is thus experienced as isolating and exclusionary rather than as a potential resource.

4. *Segregation and Inequality*

The literature on residential segregation explores the many social dimensions embedded in the division of residential space. While there is some interest in the location of elites, and in the continuing residential segregation of predominantly 'white' ethnic groups (for example Poles, Jews, and Italians in US cities) despite their economic integration, perhaps the bulk of writing focuses on disadvantaged and marginalized minorities. There are therefore studies of residential segregation among people with mental health problems, the old and impoverished, disabled people, single mothers, and so on. However, the main research effort has been concerned with the impact of segregation on the lives of marginalized racialized groups in Western societies (see *Ethnic Conflict, Geography of*). This perhaps reflects the magnitude of the challenge of 'otherness' that most developed countries face in a postcolonial world, where 'people of color in the historically white counties of the West are treated as second-class citizens' (Bonilla-Silva 2000, p. 194). Failure to acknowledge this—that is, failure to recognize the intensely racialized nature of much residential segregation in the Western world—has been criticized by Wacquant (1997) who (writing with specific reference to the USA) is concerned that by reducing the problem of racial segregation to one of poverty and disadvantage, analysts underestimate the power of history and ideology.

A growing literature on the racial dimension of residential segregation shows that segregation is not just an economic process and a material fact of life, but that it also has powerful ideological underpinnings. Historians George Fredrickson (1981) and John Cell (1982), for example, show that segregation is part of the changing character of racism. The idea of 'race' may have arisen from the notion that particular places (climates and environments) produced different kinds of people ('races'), but the tenacity of this idea is now secured by an unequal distribution of power which allows particular kinds of people to be restricted to particular types of place. Thus racisms become institutionally normalized in and through the spatial

configuration of the residential environment (cf. Goldberg 1993)

So, whereas many writers once assumed that racial residential segregation is a declining legacy of a racist, colonial past, Cell and Fredrickson show how segregationism emerged during the early twentieth century as a new and powerful way to manage power relations. In the US South and South Africa, this was initially legally enforced through the Jim Crow laws and apartheid, respectively. But what Cell's analysis vividly shows is that one of the key characteristics of segregationism is its tenacity, flexibility, and ability to survive as a practice even with the declining respectability of racism and even with the end of *de jure* racial separatism. Indeed, Cell goes so far as to define segregation as one of the most successful political ideologies of the twentieth century. 'In a segregated society,' he argues, 'we expect job restrictions and unequal wages; inferior housing, education, transportation, health and social services; prejudiced courts and police forces.' And to the extent that all these forms of exclusion are enacted through the practice of residential differentiation, segregation represents what he calls 'the highest stage of white supremacy.'

4.1 Residential Segregation and Racial Inequality

In the USA, the tenacity of these discriminatory and exclusionary forms of segregation has been closely documented by Massey and Denton (1993). Their analysis of the 1990 US census identified 20 metropolitan areas in which racial segregation was so complete and so expressive of the social and financial exclusion of black Americans that they merit the label 'hypersegregated.' Massey and Denton argue for the case of the USA that 'residential segregation lies beyond the ability of any individual to change: it constrains black life chances irrespective of personal traits, individual motivations or private achievements' (1993, pp. 2–3). A startling demonstration of this is given by Polednak (1997) in his study of the relationship between segregation and health which shows, among other things, that mortality rates and infant mortality rates in segregated communities in the USA can exceed those in parts of the less developed world.

In a series of related studies, Massey and colleagues suggest that not only is segregation problematic because it creates spaces of deprivation, exclusion, and neglect, but it is problematic too because the fact of residential segregation acts to concentrate poverty during times of macroeconomic change. So, while all poor people have been affected by those structural changes in the economy which increase income inequality, the most highly segregated communities—in the USA, African Americans—fare worst because of the way that residential segregation interacts with economic restructuring to concentrate urban poverty (Massey and Eggers 1990, Massey et al. 1991).

This work has been contested on both theoretical and methodological grounds by authors who place more emphasis on 'class' than 'race' in explaining recent increases in the spatial concentration of poverty among black Americans. However, Massey and Fisher (2000) have responded with a persuasive empirically-based argument that the same socio-economic changes in a given society have radically different consequences for the concentration of poverty as segregation rises. Residential segregation, they argue, 'is the key factor differentiating the experience of racial and ethnic groups in the United States, dividing them into two basic categories: those whose poor experience a high degree of class isolation in neighborhoods characterized by unusually high concentrations of poverty, and those whose poor do not' (p. 689). Because black Americans are by far the most highly segregated of the studied populations, 'the population of poor experiencing high concentrations of poverty is overwhelmingly black' (p. 689).

This kind of finding is by no means only a feature of societies with a historical legacy of *de jure* segregationism. In Britain, with no history of domestic slavery or legally enforced racial separatism, the residential segregation of racialized minorities has been a persistent theme for more than half a century (Smith 1989). There has, it is true, been a fierce empirical debate over whether British inner cities are as segregated as many US ghettos. This has centered on whether, for example, an ID of 60 calculated using British enumeration districts is somehow 'better' or less problematic than an ID of 80 for US census districts. However, Smith (1987) argues that while segregation persists at any of these levels, it is the form not the intensity of the phenomenon that is at issue. Whether or not ghettos are the same thing in all places, that higher levels of segregation are associated with higher rates of poverty, unemployment, poor health, and so on, is indisputable. Indeed, Smith (1989) goes on to argue that segregationism is as effective in Britain as in the USA as an interlocking system of economic institutions, social practices, and political strategies designed by one group to keep another in its place. Residential segregation, then, is a material fact of the lives of marginalized, racialized groups in Britain, and it goes with poor housing, limited employment prospects, limited service provision, poorer health prospects, and a range of other inbuilt disadvantages.

A key finding in several studies of racial segregation is that, once established, residential segregation is not just an expression of the marginality of racially 'othered' groups, but that it is also a determinant of their future life chances and opportunities. So the iniquitous form that segregation takes not only expresses marginality, but also reinforces it.

This reinforcement takes place in a number of ways. In the USA, for example, it has been suggested that there is a spatial mismatch between those who are

unemployed and the jobs they need to survive (see *Spatial Mismatch*). The argument here is that the spatial restructuring of the economy has taken jobs and incomes away from the areas in which black Americans live (see *Economic Restructuring: Geographic Aspects*). Residential segregation can, in short, act as a constraint on labor market participation. In the UK, a similar process has been observed—there too, entrapment in spaces of deprivation can prevent already disadvantaged groups from seizing opportunities opened up by the restructuring of the economy. In the British example, moreover, this process is overlaid by what might be thought of as a ‘welfare mismatch’ as those most dependent on welfare find themselves excluded spatially and economically from the range of commodities, insurances, and financial services that have expanded with the ‘rolling back’ of the welfare state (Smith 1989) (see *Welfare State, Geography of*).

So residential segregation both expresses social inequalities, and materially affects the participation of spatially excluded groups in social, political, and economic life. But because of its ideological dimension (noted above) segregation is not just a material fact, it is also a political idea—and one which often bears little resemblance to the lived experience of those it refers to. Smith (1988, 1993), for example, shows how, in Britain, ‘racial segregation’ as a political construction has helped sustain common-sense assumptions about the natural origins of racial differentiation without tackling the problems of racism and discrimination which underpin the inequalities embedded in the structure of residential space. The importance of ideas and images (ideologies) about residential space as a factor structuring people’s life chances and opportunities is further illustrated by Anderson (1991) in her work on Canadian constructions of Chinatown. She shows how that segregated space known as ‘Chinatown’ grew out of, and came to structure, a politically divisive system of racial discourse that justified domination over people of Chinese origin in late nineteenth-century Vancouver. In a similar vein, Wilson (1996) illuminates the power of metaphor to drive the process of uneven development between segregated segments of urban Indianapolis in the USA. His aim is to illustrate what potent political instruments metaphorical symbols can be. In the example of Indianapolis, USA, ‘they injected a constellation of representations into everyday thought that rendered a form of growth as value-free and neutral that was, in fact, directly adverse to the interests of the black inner city poor’ (p. 96).

Because of its ideological underpinnings and material effects, residential segregation is something that those who are marginalized and excluded by it can have little control over. However, iniquitous forms and practices of segregation can be, and are being, resisted in a number of ways. There is a policy dimension to this, of course, the relevance of which is

firmly established by Bashi and Hughes (1997). In a thoughtful account of the likely implications of globalization for the character and intensity of residential segregation by race they argue that the main problems and their likely solutions are located locally (nationally) rather than globally. ‘On the topic of segregation,’ they note, ‘the real problem is less one of sustained explanation than one of sustained action’ (p. 120). This adds new significance to the work of those who argue that the best weapon with which to tackle segregation is the law (Goering 1986, McGrew 1997).

Other ways of illuminating and challenging the inequalities embedded in residential segregation range from urban unrest to artistic expression. The first of these is extensively documented. The second is less fully recognized, though a glance at the work of artists like photographer Ingrid Pollard, painter Jean-Michel Basquiat, or filmmaker Spike Lee soon illustrates its potency. Ingrid Pollard’s exhibition ‘Pastoral Interludes,’ for example, locates a series of ‘black’ figures in the ‘white’ English countryside. This contrasts with the ‘usual’ positioning of black Britons within the segregated inner city, and challenges the taken-for-granted assumption that it is somehow normal for black people to be absent from rural areas. The other side of the coin—a stark vision of (and dissatisfaction with) the experiences of confinement within residentially segregated urban spaces—finds expression in the stick-like black figures which characterize so many of Basquiat’s paintings and in the rise of rap and hip-hop which was used so effectively in Spike Lee’s *Do the Right Thing* to speak about life in the segregated space of New York’s Brooklyn.

5. Overview

To summarize, residential segregation is both an expression and a mediation of the ideologies and practices which structure social life. It is a product of the uneven development of residential space and of the selective operation of the markets and institutions that govern access to housing. Residential space is a gateway to services, educational, and employment opportunities, health, and social services. Where people live is therefore a reflection of who they are, but it is also an influence on who they can become. It has therefore been argued that tackling the inequalities embedded in residential segregation is a key priority for public policy and an urgent focus for social concern.

See also: Ethnic Conflict, Geography of; Gender and Feminist Studies in Geography; Population Ecology; Residential Concentration/Segregation, Demographic Effects of; Residential Segregation: Sociological Aspects; Segregation Indices; Social

Geography; Spatial Choice Models; Spatial Equity; Spatial Mismatch; Welfare State, Geography of

Bibliography

- Anderson K 1991 *Vancouver's Chinatown: Racial Discourse in Canada 1865–1980*. McGill-Queen's University Press, Montreal
- Bashi V I, Hughes M A 1997 Globalization and residential segregation by race. *Annals of the American Academy of Political and Social Science* **551**: 105–20
- Bonilla-Silva E 2000 'This is a white country': The racial ideology of the western nations of the world system. *Sociological Inquiry* **70**: 188–214
- Cell J W 1982 *The Highest Stage of White Supremacy: The Origins of Segregation in South Africa and the American South*. Cambridge University Press, Cambridge, UK
- Clark W A V 1992 Residential preferences and residential choices in a multiethnic context. *Demography* **29**: 451–66
- Fredrickson G 1981 *White Supremacy: A Comparative Study in American and South African History*. Oxford University Press, New York
- Goering J M 1986 *Housing Desegregation and Federal Policy*. University of North Carolina Press, Chapel Hill, NC
- Goldberg D T 1993 *Racist Culture: Philosophy and the Politics of Meaning*. Blackwell, Oxford, UK
- Harvey D 1985 *The Urbanization of Capital: Studies in the History and Theory of Capitalist Urbanization*. Johns Hopkins University Press, Baltimore
- Liebertson S, Carter D 1982 A model for inferring the voluntary and involuntary causes of residential segregation. *Demography* **19**: 511–26
- Massey D S, Denton N A 1993 *American Apartheid: Segregation and the Making of the Underclass*. Harvard University Press, Cambridge, MA
- Massey D S, Eggers M L 1990 The ecology of inequality: Minorities and the concentration of poverty. *American Journal of Sociology* **95**: 1153–89
- Massey D S, Eggers M L, Gross A B 1991 Segregation, the concentration of poverty, and the life chances of individuals. *Social Science Research* **20**: 397–420
- Massey D S, Fischer M J 2000 How segregation concentrates poverty, 1970–1980. *Ethnic and Racial Studies* **23**: 670–91
- McGrew T 1997 The history of residential segregation in the United States and Title VIII. *The Black Scholar* **27**: 22–30
- Peach C (ed.) 1975 *Urban Social Segregation*. Longman, London
- Polednak A 1997 *Segregation, Poverty and Mortality in Urban African Americans*. Oxford University Press, New York and London
- Pratt G, Hanson S 1994 Geography and the construction of difference. *Gender, Place and Culture* **1**: 5–29
- Schelling T 1971 Dynamic models of segregation. *Journal of Mathematical Sociology* **1**: 14–86
- Smith S J 1987 Residential segregation: A geography of English racism?. In: Jackson P (ed.) *Race and Racism: Essays in Social Geography*. George Allen and Unwin, London, pp. 25–49
- Smith S J 1988 Political interpretations of racial segregation in Britain. *Environment and Planning D: Society and Space* **6**: 423–44
- Smith S J 1989 *The Politics of 'Race' and Residence: Citizenship, Segregation and White Supremacy in Britain*. Polity Press, Cambridge
- Smith S J 1993 Residential segregation and the politics of racialisation. In: Cross M, Keith M (eds.) *Racism, the City and the State*. Unwin Hyman, London, pp. 128–43
- Wacquant L J D 1997 Three pernicious premises in the study of the American ghetto. *International Journal of Urban and Regional Research* **21**: 341–53
- Wilson D 1996 Metaphors, growth coalition discourses and black poverty neighborhoods in a US city. *Antipode* **28**: 72–96

S. J. Smith

Residential Segregation: Sociological Aspects

Segregation refers to the differential location of social groups across categories of a social structure. It may arise from voluntary forces, whereby members of one group choose positions that are different from those of other groups; it may reflect involuntary processes, whereby systematic barriers prevent certain kinds of people from occupying particular positions within the structure, or it may reflect a combination of voluntary and involuntary forces.

By far the most studied kind of segregation involves the differential location of social groups across neighborhoods of a city, a topic generally known as *residential segregation*. This field has attracted strong interest over the years because social scientists have recognized the close connection between where people live and the socioeconomic outcomes they experience. This insight originated with the Chicago School of Sociology in the early twentieth century (Park and Burgess 1925). According to Robert Park (1926, p. 8), 'it is because social relations are so frequently and so inevitably correlated with spatial relations; because physical distances so frequently are ... indexes of social distances, that statistics have any significance whatever for sociology.' In other words, because social and economic resources are distributed unevenly in space, where one lives plays a significant role in determining one's prospects for attaining education, health, employment, income, and prestige (Massey and Denton 1985).

Ernest Burgess (1928) was the first to measure patterns and levels of residential segregation, but he was unable to establish standard methodology for doing so. As a result, social scientists became bogged down in a prolonged debate about the best segregation index. The issue was settled in 1955 by Duncan and Duncan (1955), who recommended using the index of dissimilarity. This straightforward measure varies between 0 (no segregation) and 100 (complete segregation) and represents the relative percentage of minority members who would have to exchange

neighborhoods with majority members to achieve an even, or integrated, residential distribution.

For the next 20 years, social scientists working in the tradition of the Chicago School employed this index to measure levels of segregation in a variety of urban settings. Duncan and Duncan (1957) undertook their classic study of the levels, causes, and consequences of black segregation in Chicago. Lieberman (1963) followed with an analysis of segregation among European-origin groups living in US industrial cities. In their ambitious study, Taeuber and Taeuber (1965) documented patterns of segregation and neighborhood change for African Americans in cities throughout the USA. As the racial and ethnic composition of other nations has been transformed by postwar immigration (Massey et al. 1998), social scientists have applied the index to study the segregation of minority groups in Britain, Germany, France, The Netherlands, Israel, Canada, and Australia (Massey 1985).

In 1976, Cortese et al. ignited a new debate about the measurement of segregation by questioning the applicability of the index of dissimilarity under certain circumstances. In response to their critique, numerous alternative indices were proposed to yield considerable confusion in the literature. Massey and Denton (1988) once again brought order to the field by demonstrating that residential segregation was a multidimensional construct characterized by five distinct axes of spatial variation—evenness, exposure, clustering, concentration, and centralization. They showed that the index of dissimilarity measured the evenness dimension fairly well, thus confirming Duncan and Duncan's earlier work. After clarifying the conceptual and empirical properties of the remaining dimensions, they went on to recommend a specific index for each one.

Massey and Denton (1989) used this multidimensional conceptualization of segregation to define the concept of *hypersegregation*, which occurs whenever a group experiences high segregation on at least four of the five dimensions at once. They showed that whereas blacks in 20 metropolitan areas displayed high levels of segregation on four or more dimensions simultaneously, Asian and European ethnic groups in US metropolitan areas never achieved a high level of segregation on more than one dimension at a time. As of 1990, more than one third of all African Americans lived under conditions of hypersegregation (Denton 1994).

Despite the widespread recognition of residential segregation as a multidimensional construct, the most commonly used measure continues to be the index of dissimilarity. The next most common index is the P^* isolation index, which Massey and Denton identify as a measure of exposure. It represents the percentage of minority members living in the neighborhood of the average minority person. An index of 80 for African Americans, for example, means that the average black person lives in a neighborhood that is 80 percent

black. Unlike the index of dissimilarity, this index is strongly influenced by the relative number of minority members within an urban area, so that intergroup and intercity comparisons are best made using the index of dissimilarity.

A convenient rule of thumb for interpreting the latter index is that values under 30 are 'low,' those from 30 to 60 are 'moderate,' and those above 60 are 'high' (Kantrowitz 1973). In general, voluntary processes of segregation yield index values in the low-to-moderate range. Whenever dissimilarity indices persist at levels above 60 for any length of time, it usually signals the operation of involuntary forces of one sort or another. Studies done in the USA reveal that African Americans continue to experience high levels of discrimination in real estate, banking, and insurance markets, and that they remain objects of considerable prejudice by whites (Massey and Denton 1993). As a result, black-white residential dissimilarities in US urban areas generally range upward from 60, and unlike those of other ethnic minorities, they do not fall as socioeconomic status rises (Denton and Massey 1988).

In the USA, black residential segregation has also displayed little tendency to decline over time, except in places where the number of blacks is so small that complete desegregation would yield little interracial mixing within neighborhoods (Kriwo and Kaufman 1998). Other than African Americans in the USA, the only documented case where a racial or ethnic group has experienced a prolonged, high level of segregation is that of black Africans in the Union of South Africa under apartheid (Christopher 1993). Blacks in neither Canada (Fong 1996) nor Brazil (Telles 1992) encounter the extreme levels of segregation experienced by their US counterparts.

New interest in the consequences of residential segregation was stimulated by William Julius Wilson (1987), who argued that the social isolation of blacks within neighborhoods of concentrated poverty systematically lowered their chance of success in employment, marriage, and education. Although he attributed this growing isolation to structural shifts in the economy and the changing geography of employment, Massey and Denton (1993) showed that these changes necessarily produce high concentrations of poverty when they occur under conditions of high racial segregation. Given rising income inequality and the high degree of residential segregation experienced by African Americans in the USA, the concentration of black poverty is mathematically inevitable.

Studies drawing on longitudinal data files reveal that growing up in a poor neighborhood has negative socioeconomic consequences that are independent of personal characteristics or family circumstances. Coming of age under concentrated poverty undermines cognitive development in early childhood, reduces academic achievement in later adolescence, and elevates the risk of antisocial behavior throughout

the life course (Brooks-Gunn et al. 1997). Males experience a higher likelihood of withdrawing from the labor force and increased risk of criminal involvement, whereas women are less likely to marry and more likely to become single parents (Massey and Shibuya 1995).

The link between racial segregation and stratification prevails no matter what one assumes about the extent of segregation by social class, but when high levels of class segregation occur in addition to racial or ethnic segregation, the concentration of poverty is exacerbated. Although the necessary data are not available for most countries, studies done in US cities reveal a marked rise in the degree of income segregation from 1970 to 1990 (Massey 1996), a trend that occurred across all regions and racial/ethnic groups (Jargowsky 1996).

Although rising income inequality and growing class segregation may undermine the socioeconomic welfare of all racial and ethnic groups, the consequences have been particularly severe for African Americans because they are also so highly segregated by race. Under these circumstances, poor blacks experience far higher concentrations of neighborhood poverty than the poor of other groups. Although few studies have explored segregation by class, race, and ethnicity simultaneously, to the extent that rising income inequality, growing class segregation, and persisting racial/ethnic segregation coincide in other nations, a similar ecology of inequality is expected to prevail.

In the new century, social scientists are working to disentangle the effects of metropolitan structure, neighborhood composition, and family circumstances on individual socioeconomic outcomes. In this effort, researchers have turned increasingly to multilevel models and longitudinal data sets. Although the questions may be similar to those posed by the Chicago School theorists in the 1920s, the measures, methods, and data are now far more complex and sophisticated. Specifying the theoretical nature and empirical strength of connections between ecological structure, neighborhood conditions and individual outcomes will remain a central preoccupation of social science in the twenty-first century.

See also: Discrimination; Discrimination, Economics of; Discrimination: Racial; Ethnic Conflicts; Ethnicity, Sociology of; Housing Economics; Race and Urban Planning; Racial Relations; Racism, History of; Racism, Sociology of; Segregation Indices; Spatial Equity; Urban Geography; Urban Poverty in Neighborhoods

Bibliography

Brooks-Gunn J, Duncan G, Aber L 1997 *Neighborhood Poverty, Volume I: Context and Consequences for Children*. Russell Sage, New York

- Burgess E W 1928 Residential segregation in American cities. *Annals of the American Academy of Politics and Social Science* **140**: 105–15
- Christopher A J 1993 Segregation levels in the late-apartheid city 1985–1991. *Tijdschrift voor Economische en Sociale Geografie* **83**: 15–24
- Cortese C F, Falk R F, Cohen J C 1976 Further considerations on the methodological analysis of segregation indices. *American Sociological Review* **41**: 630–7
- Denton N A 1994 Are African Americans still hypersegregated? In: Bullard R D, Grigsby J E, Lee C (eds.) *Residential Apartheid: The American Legacy*. CAAS Publications, Los Angeles
- Denton N A, Massey D S 1988 Residential segregation of blacks, Hispanics, and Asians by socioeconomic status and generation. *Social Science Quarterly* **69**: 797–817
- Duncan O D, Duncan B 1955 A methodological analysis of segregation indices. *American Sociological Review* **20**: 210–17
- Duncan O D, Duncan B 1957 *The Negro Population of Chicago: A Study of Residential Succession*. University of Chicago Press, Chicago
- Fong E 1996 A comparative perspective on racial residential segregation: American and Canadian experiences. *The Sociological Quarterly* **37**: 199–226
- Jargowsky P A 1996 Take the money and run: Economic segregation in U.S. metropolitan areas. *American Sociological Review* **61**: 984–98
- Kantrowitz N 1973 *Ethnic and Racial Segregation in the New York Metropolis*. Praeger, New York
- Krivo L J, Kaufman R L 1998 How low can it go? Declining black-white segregation in a multi-ethnic context. *Demography* **36**: 93–109
- Lieberman S 1963 *Ethnic Patterns in American Cities*. Free Press, New York
- Massey D S 1985 Ethnic residential segregation: A theoretical synthesis and empirical review. *Sociology and Social Research* **69**: 315–50
- Massey D S 1996 The age of extremes: Concentrated affluence and poverty in the 21st century. *Demography* **33**: 395–412
- Massey D S, Arango J, Hugo G, Kouaouci A, Pellegrino A, Taylor J E 1998 *Worlds in Motion: Understanding International Migration at the End of the Millennium*. Oxford University Press, Oxford, UK
- Massey D S, Denton N A 1985 Spatial assimilation as a socioeconomic process. *American Sociological Review* **50**: 94–105
- Massey D S, Denton N A 1988 The dimensions of residential segregation. *Social Forces* **67**: 281–315
- Massey D S, Denton N A 1989 Hypersegregation in US metropolitan areas: Black and Hispanic segregation along five dimensions. *Demography* **26**: 373–93
- Massey D S, Denton N A 1993 *American Apartheid: Segregation and the Making of the Underclass*. Harvard University Press, Cambridge, MA
- Massey D S, Shibuya K 1995 Unravelling the tangle of pathology: The effect of spatially concentrated joblessness on the well-being of African Americans. *Social Science Research* **24**: 352–66
- Park R E 1926 The urban community as a spatial pattern and a moral order. In: Burgess E W, Park R E (eds.) *The Urban Community*. University of Chicago Press, Chicago
- Park R E, Burgess E W 1925 *The City*. University of Chicago Press, Chicago

- Taeuber K E, Taeuber A F 1965 *Negroes in Cities: Residential Segregation and Neighborhood Change*. Aldine, Chicago
- Telles E E 1992 Residential segregation by skin color in Brazil. *American Sociological Review* 57: 186–97
- Wilson W J 1987 *The Truly Disadvantaged: The Inner City, the Underclass, and Public Policy*. University of Chicago Press, Chicago

D. S. Massey

Resource Geography

Resource geography is part of a broad tradition of geographical inquiry examining the reciprocal relationships between socioeconomic systems and physical environmental processes. Within this ‘nature society’ tradition, resource geography is defined by its focus on how economic, political, and social institutions and practices determine patterns of resource use over space and time (see *Nature–Society in Geography*). This article begins with an explanation of the term ‘resource,’ outlines the scope of resource geography, and identifies four principal approaches within the field.

1. Resources Defined

Resources refer to those properties of the physical environment that can be exploited to meet the needs and desires of society. Resources can be understood, therefore, as culturally mediated appraisals of the physical environment. A resource does not exist *a priori*, but it is produced from ‘neutral stuff’ through its interaction with a set of prevailing cultural norms which collectively define its social utility and economic viability: thus ‘resources are not, they become’ (Zimmerman 1951, p. 15). The cultural milieu consists of economic variables (markets, prices, and technology); political institutions (property rights, legislation, and regulatory codes); social attributes (perception, behavior, and practices of decision-making); and cultural belief-systems (cosmologies of nature, society, and development). For example, just as sperm whales were regarded in the early nineteenth century as a valuable resource from which to obtain high quality lighting oil, so the recent appraisal of tropical forests as resources to be ‘mined’ for pharmaceutical benefits derives from the confluence of a specific set of economic, legal, technological, and cultural processes.

Resources are most commonly categorized based on their renewability. Renewable (or flow) resources are those depleted by use in the short term but which can replenish themselves over the longer term (e.g., forests, soils, and fisheries). Non-renewable (or stock) resources have a finite supply: consumption today comes at the expense of future use (e.g., minerals and

fossil fuels). In practice, however, this distinction is more accurately represented as a continuum since the renewability of many resources depends on when, where, and how they are consumed. For example, the experience of many communities with fisheries, forests, and groundwater demonstrates that if the rate of use exceeds the rate of renewal, even a potentially renewable resource can be depleted. In this context it is possible to speak of the ‘mining’ of renewable resources like soil or groundwater.

The ‘resource question’ has deep and enduring roots within Western intellectual history and is characterized by two distinct questions (Rees 1990). The first is the role of resource availability/resource scarcity in limiting opportunities for economic growth and development. Early explorations of this theme can be found in Malthus’s (1798) thesis on the differential rates of growth of population and food supply, and Marx’s (1967) account of the productive capacity of the labor process to liberate humankind from natural limits. More recently it appears in debates over the long-term availability of water, soil, and energy resources for a growing world population (Barnett and Morse 1963). The second is the sociopolitical implications of resource use and the impact of resource management on environmental quality, equity, and human welfare across a range of temporal and spatial scales. Examples of this can be found in conservationist writings of the late nineteenth century, in studies of social inequalities in exposure to environmental risk, and in current engagements with global environmental change and sustainable development.

Resource geography has tended to shun theoretical development in preference for a pragmatic engagement with concrete problems and a prescriptive approach to policy formation (see Wescoat 1987). Four principal lines of inquiry can be identified: resource inventory and evaluation; resource management institutions and decision-making; the political economy of resource use and the environment; and critiques of resource management and the question of ‘nature.’

2. Lines of Inquiry within Resource Geography

2.1 Resource Inventory and Evaluation

The stock-in-trade of resource geography has traditionally been the identification and analysis of resource distribution at spatial scales from local to global (e.g., Murphy 1954). For example, descriptions of natural resource distribution conventionally form a central component of regional geography, where they provide a logical segue between the description of a region’s geological, hydrological, pedological, and biological characteristics and discussion of the region’s

socioeconomic geography. Using this spatial mapping approach, resource geographers have quantified natural resource endowments, evaluated the capacity of particular natural resource bases to support economic development, and determined the extent to which current strategies of resource acquisition, conservation, and environmental protection are adequate.

By providing assessments of the spatial extent, utility, and economic viability of mineral, biological, and hydrological resources, resource geography has often played the role of handmaiden to regional planning and economic development. Early examples from within US geography can be found in the watershed and irrigation survey work of geomorphologists such as Grove Karl Gilbert (1877) and John Wesley Powell (1879). Their evaluations of water resources in the arid western states in the second half of the nineteenth century helped promote the federal government's involvement in dam construction, irrigation, and inter-regional water transfer systems. In the UK, the development of the national Land Utilization Survey by Dudley Stamp (1934) is indicative of how resource geography has historically contributed to land-use planning and, more generally, to the implementation of a conservationist utilitarianism towards natural resources. Professional affiliation between resource geographers and public resource agencies continues to generate empirical resource inventories and practical approaches for implementing integrated resource and environmental management (see Mitchell 1997).

The strategic and military value of mineral and energy resources has often provided an overtly political focus to the work of resource geographers, confirming MacKinder's (1887) maxim that geography serve as an aid to statecraft. For example, resource geographers served as part of a multidisciplinary team of social scientists compiling a series of Area Handbooks commissioned by the US Department of the Army (see www.hqda.army.mil). These 'compilations of basic facts' describe economically and militarily significant aspects of selected countries including identification and inventory of natural resource endowments. Whether part of regional or applied geography, the primary objective of local, regional, and global resource appraisals is informational. As a consequence, this line of inquiry has produced maps of resource location, atlases, and descriptions of resource availability, but has contributed little in the way of theoretical or methodological development.

2.2 Resource Management Institutions and Decision-making

A second major focus of resource geography is practically orientated research on the management of natural resources and the environment. The historical

and intellectual foundations of this approach lie within welfare economics with its objective of utility maximization. Adopting this welfare framework as its core paradigm, resource geography has sought to maximize the social utility of resources such as coal, oil, water, and open space by assisting in their rational and efficient management. A central objective has been to understand the influence of institutional frameworks (e.g., the distribution of property rights, the nature of environmental legislation, and the structure of commodity markets), decision-making processes, and differences in resource perception on the allocation, exploitation, and conservation of natural resources over time.

Under this paradigm, resource shortages and environmental degradation are understood to result from market failure and pervasive externalities. As a consequence, research adopts a prescriptive stance, advocating public policy approaches to intervene in existing markets to address the causes of failure (e.g., via government regulation for environmental quality or resource taxation), create markets where none had previously existed (e.g., pollution permit trading), or re-design the institutions of property ownership to internalize external costs (e.g., resource privatization, see Anderson and Leal 1992).

Prescriptions for resource management drawn from neo-classical economic models came under criticism from behavioral studies during the 1960s. Work by Gilbert White (1964) on river basin development and floodplain management, for example, argued that existing public policies for resource and hazard management over-simplified the process by which management alternatives were selected. Arguing that flood control engineering was just one of several possible responses to flood risk (others include flood insurance, disaster warning systems, and comprehensive floodplain management), White's work—and that of the Natural Hazards Research Group at the University of Chicago—focused attention on the range of choice in resource management decisions (*Environmental Adaptation and Adjustments; Environmental Risk and Hazards*).

In contrast to the neo-classical model which interpreted resource scarcity or deteriorating environmental quality in terms of market failure, the Chicago School argued that poor resource management decisions resulted from misperception or limited awareness of better alternatives. It rejected the assumptions of utility maximizing individuals operating in a field of perfect knowledge and instead adopted a behavioral approach in which resource managers operated under conditions of bounded rationality. Initially focused on water resources in USA, natural hazards research expanded to include other geophysical extremes in North America and in different cultural contexts. As an alternative to the unitary and utility maximizing assumptions of conventional resource management, the behavioral approach captures the

role of perception and sociocultural rules in determining the relative success of strategies such as energy and water conservation or the construction of international regimes for environmental governance (e.g., marine pollution, migratory species, ozone depletion—see Young 1999).

In an effort to contextualize the process of decision-making and explore the opportunities for co-operation as well as competition in the management of natural resources, resource geography has looked to traditional societies for alternative to market-based resource regimes. Research has focused, for example, on management regimes for common property resources—those for which control of access is difficult and joint use affects the welfare of each resource user (e.g., fisheries, river systems, wildlife, pastures, and public parks.) (see *Environmental and Resource Management*). Interest in management regimes for the commons centers on Hardin's (1968) provocative assessment that common property resources were inevitably prone to environmental degradation as a result of the cumulative decisions of economically rational individuals. Using the analogy of herders stocking a commonly held pasture, Hardin argued that the pursuit of individual self-interest will lead each herder to add to his or her stock; while the benefits of an additional animal (increased income) will be captured by the individual, the costs (declining grassland productivity) will be shared among all users. Describing this as the 'tragedy of the commons,' he concluded that either government ownership or privatization of the commons was necessary to forestall this degradation since 'freedom of the commons brings ruin to all.'

However, subsequent work by resource geographers and anthropologists has revealed historical and contemporary examples of successful common property resource management (e.g., Feeny et al. 1990). This has generated a two-pronged critique of Hardin's conclusions: first, Hardin failed to acknowledge that his herders could communicate with each other and, therefore, that they could undertake collective rather than individual action to maintain the environmental integrity of the commons; and second, Hardin conflated the nature of a common property resource with the type of property regime under which it was managed. He assumed, therefore, that his pasture—a classic common pool resource—would be managed as an open access resource regime. Historical experience illustrates that common property resources can exist under open access, communal, state, and private property regimes and that these need not be mutually exclusive but can coexist.

2.3 *The Political Economy of Resource Use and the Environment*

A third line of inquiry within resource geography explicitly rejects the applied management orientation

of earlier approaches for failing to capture the social stratification of economic and political power. In contrast, it engages the political and social dimensions of resource use, emphasizing the ecological and socio-economic context in which resource managers and their decisions are embedded. By embracing a broadly defined political economy, resource geography has sought to understand the nested scales of activity—from the actions of local institutions, through national public policies, to the global strategies of corporations—which constrain and influence the choices of individual resource managers.

Three distinctive subfields have emerged from this critique of conventional resource management: political ecology (e.g., Blakie and Brookfield 1987, *Feminist Political Ecology*); radical hazards research (e.g., Hewitt 1983); and environmental justice (e.g., Pulido 1996, *Environmental Justice*). Common to all three subfields is a perspective that locates the origins of resource depletion and environmental degradation in the economic, political, and social institutions of production. As a result, recognition is given to the socioeconomic—as opposed to purely physical—environmental—dynamics of vulnerability. For example, Watts' (1983) analysis of vulnerability to drought among the Hausa peasantry of northern Nigeria locates the cause of this vulnerability in the articulation between pre-capitalist modes of production and a global capitalist system.

2.4 *Critiques of Resource Management and the Question of 'Nature'*

Resource geography remains an evolving field influenced by both topical issues in society and theoretical trends within the academy. For example, recent public policy debates in North America and Europe have provided a topical context for research on: the effectiveness of conventional environmental policy approaches; the integration of different spatial scales in managing natural habitat; managing socio-economic transition within natural resource communities (e.g., those based on coal mining, fishing, or timber harvesting); the environmental implications of industrialization and modernization in the developing world; and the challenges of operationalizing sustainable development.

Resource geography has traditionally shown theoretical restraint in favor of practical application (Wescoat 1987); however, the 1990s has seen an increased interest in praxis—the integration of theoretical perspectives with the day-to-day issues of producing, consuming, and managing natural resources. Indicative of this are recent engagements between resource geography and:

(a) theories of economic restructuring and rural sociology in seeking to understand the implications for communities of agricultural restructuring, resource

privatization, and the links between local and global regimes of food production;

(b) the question of 'nature' and environmental ethics, and their implications for the welfare of non-human species;

(c) discussions within critical legal theory and discourse theory on the ideologies and institutions that sustain particular actions towards nature (Proctor 1998, Wolch and Emel 1998).

Collectively these new directions offer a broad critique of the instrumentalism inherent in the categorization of parts of nature as 'resources.' A central objective is to destabilize conventional views of nature as a collection of severable, tradable resource components valued solely for their commodity value. By recognizing such common-place notions as 'resources,' 'wilderness' and 'nature' as historically and geographically specific social constructions, new directions within resource geography contribute to an understanding of the roles gender, race, and class have played in domesticating the natural world.

See also: Energy, Geography of; Resource Institutions; Water Resources

Bibliography

- Anderson T, Leal D 1992 *Free Market Environmentalism*. Westview Press, Boulder, CO
- Barnett H, Morse C 1963 *Scarcity and Growth: The Economics of Natural Resource Availability*. Johns Hopkins University Press, Baltimore, MD
- Blakie P, Brookfield H 1987 *Land Degradation and Society*. Methuen, London
- Feeny D, Berkes F, McCay B, Acheson J 1990 The tragedy of the commons: Twenty years later. *Human Ecology* **18**: 1–19
- Gilbert G K 1877 *Report on the Geology of the Henry Mountains*. US Government Printing Office, Washington, DC
- Hardin G 1968 The tragedy of the commons. *Science* **162**: 1243–8
- Hewitt K 1983 (ed.) *Interpretations of Calamity*. Allen and Unwin, Boston
- MacKinder H 1887 The scope and methods of geography. *Proceedings of the Royal Geographical Society* **9**: 141–60
- Malthus T 1798 (reprinted 1986) *An Essay on the Principle of Population*. Penguin Classics, Harmondsworth, UK
- Mark K 1967 *Capital* (3 volumes). International Publishers, New York
- Mitchell B 1997 *Resource and Environmental Management*. Longman, Harlow, UK
- Murphy R 1954 The geography of mineral production. In: James P, Jones C (eds.) *American Geography: Inventory and Prospect*. Syracuse University Press, Syracuse, NY
- Powell J W 1879 *Report of the Lands of the Arid Region of the United States (with a more detailed account of the lands of Utah)*. US Government Printing Office, Washington, DC
- Proctor J 1998 The social construction of nature: relativist accusations, pragmatist and critical realist responses. *Annals of the Association of American Geographers* **88**: 352–76
- Pulido L 1996 *Environmentalism and Economic Justice: Two Chicano Struggles in the Southwest*. University of Arizona Press, Tucson, AZ
- Rees J 1990 *Natural Resources: Allocation, Economics, and Policy*. Routledge, New York
- Stamp L D 1934 Land utilization survey of Britain. *Geographical Review* **24**: 646–50
- Watts M 1983 *Silent Violence: Food, Famine, and Peasantry in Northern Nigeria*. University of California Press, Berkeley, CA
- Wescoat J 1987 The 'practical range of choice' in water resources geography. *Progress in Human Geography* **11**: 41–59
- White G 1964 Choice of adjustment to floods. Department of Geography Research Paper No. 93, University of Chicago
- Wolch J, Emel J 1998 *Animal Geographies: Place, Politics, and Identity in the Nature-Culture Borderlands*. Verso, London
- Young O 1999 (ed.) *The Effectiveness of International Environmental Regimes*. MIT Press, Cambridge, MA
- Zimmerman E W 1951 *World Resources and Industries*. Harper, New York

G. Bridge

Resource Institutions

Human beings use, manage, or conserve resources through institutional arrangements. The absence of institutions leads to open access to resources, a condition in which overuse and depletion may be unavoidable. Institutions exercise an influence over resource management outcomes that is independent of the forces and dynamics that created them. Institutional innovation is therefore critically important in all efforts to influence resource use and conservation. Whether one attempts to rejuvenate depleted fisheries or prevent trade in endangered wildlife, to protect forests, or prevent the erosion of biodiversity, institutional change remains a crucial part of such efforts. Given the crucial role of institutions, it is clear that an understanding of the nature of institutions and their relationship to resources is necessary if the goal is sustainable management or conservation.

Despite variations in approaches to institutional analysis, resource institutions can be defined as sets of rules that can be crafted collectively or individually, and that have the power to shape actions related to the conservation and use of resources. Resource systems and institutional arrangements are distinct entities, and it is important to pay attention to the factors that lead to a better fit between a given resource and the rules that govern the management of that resource.

However, it is only in the second half of the twentieth century that the role of institutions in resource management has found substantial recognition in the scholarly and policy world. Part of the reason institutions are recognized as important is that many resources critical to human existence and subsistence are becoming increasingly scarce, and are difficult to manage under some universalizable institutional form. Very different institutional regimes may be necessary for different types of resources. Consider,

as examples, wildlife species that need hundreds of square miles as their habitat, scattered forest patches upon which thousands of households depend for their livelihood, or irrigation systems that are too small to require government intervention, but too large to be managed by a single user. International agreements among governments may be appropriate institutional mechanisms to restrict the production of ozone depleting chemicals because they are mostly generated by a few major industrial producers. But for reduction in the levels of greenhouse gases, a far more widespread program of restraint may be necessary because of the nonpoint nature of the production of these gases. Another reason for the importance of getting institutions right are transactions costs that are associated with the allocation, processing, and exchange of resources, and that vary with institutional arrangements. Institutions that rely on external enforcement may require higher levels of expenditure to ensure that users conform to rules in comparison to those situations where it is in the interests of users to monitor each other and enforce institutional provisions directly.

The range of rules that comprise institutions is potentially infinite. But three institutional forms are possibly the most familiar: markets, states, and communities. Within each of these categories it is possible to find many variations; other forms span these categories. The precise combination of rules and how rules influence human behavior and outcomes are the basis on which it is possible to judge the performance of institutional arrangements. Typically, the objective of creating or forming institutions over resources is to enhance the efficiency, equity, or sustainability of resource use. However, a truly enormous and highly complex set of factors are at play. At the same time, the application of institutional analysis to resources is a relatively new field. It also requires collaboration across disciplinary boundaries. These are two reasons why the precise nature of the relationship between social and physical/biological systems, signified by the phrase 'resource institutions,' still requires considerable scientific research.

1. A Brief History of Institutions and Their Place in Resource Management

Contributions to institutional theories have flowed from economists, sociologists, and political scientists (Goodin 1996). Such contributions resonate with larger themes in social theory about the relationship between social structure, agency, and outcomes. In explicit counterpoint to the neoclassical paradigm in economics that finds little role for institutions or organizations in social life, advocates of institutions such as North (1990) and Williamson (1985) emphasize the unavoidable influence of institutional constraints on all social action. New institutionalists,

property rights theorists, and scholars of the commons have been the most influential in providing insights about the nature of resource institutions. Efforts aimed at understanding institutions have drawn from rigorous mathematical approaches as well as more descriptive and interpretive orientations.

Although resource institutions may serve as constraints upon certain groups and individuals, or certain kinds of actions, they also facilitate and advantage others. In addition to facilitating or constraining actions, and advantaging some people over others in terms of power and benefits (Knight 1992), resource institutions may also help shape desires, motivations, and preferences of groups and individuals. They are rooted in past human actions, and are the result of choices made by individuals and groups. It is easy to see, therefore, that institutions are seldom stationary; they change over time because of human choices.

For a long time, scholars viewed markets and states as the appropriate institutional mechanisms to address externalities that result from the collective goods nature of a vast class of resources. Advocates of market solutions argued for privatization of collectively owned and managed resources such as fisheries and forests, and argued that without privatization collective goods are likely to be overexploited. Voluntary exchanges among private owners on the basis of pricing signals, they held, would lead to the most efficient use of resources. Advocates of state control and ownership, pointing to asymmetric allocations that often result from the operation of market mechanisms, suggested that centralized control of resources was far more likely to promote conservation and sustainable use. Without centralized enforcement, according to them, attempts to manage resources could turn out to be futile. Through much of the 1950s and 1960s, writings on resource use and management viewed community institutions as backward, inefficient, and likely to promote overexploitation.

From the late 1970s, and especially since the mid-1980s, there has been a significant shift in views about the role of community as a resource management institution. This is partly because resource management strategies that rely exclusively on exchange (market) or enforcement (coercive, state-based) have obvious deficiencies. One tends to benefit a small group or a few individuals, and the other tends to create highly asymmetric relations of domination and dependency. Both have shown a tendency to disadvantage and exclude rural residents relying on local resources, who then do not cooperate in management of these resources. Both at the level of policy initiatives, and in terms of theoretical developments, the option of collective management of resources through new types of resource institutions based in communities has found wide appeal (National Research Council 1986, Ostrom 1990).

Despite heated assertions about the superiority of one type of resource management institution over

another, in real life most resource institutions do not easily fit the mold of market, state, or community. For the operation of each of these forms of institutions, implicit or explicit understandings that shade into other forms are necessary. For example, often markets and private property arrangements work better only when the force of the state is available to enforce the security of property.

2. Resource Institutions, Rules, and Power

Since resource institutions are sets of rules and norms that shape action, it is necessary to understand the broad types of rules that are most important in making institutions effective. Effectiveness implies three categories of rules: the ability to create new rules and modify old ones pertaining to resources; the ability to implement and ensure compliance to existing and new rules; and the ability to adjudicate disputes that arise in the effort to create rules and ensure compliance. Rules pertaining to resources define who gets to access a resource, who gets to use them, who is excluded from a resource, and how and when a resource can be transferred from one individual or group to another. Institutional arrangements are likely to be ineffective when human beings can make these above rules but not enforce them, or enforce existing rules but not change them in case of deficiencies. Mechanisms that allow adjudication of disputes among resource users and managers are necessary if resource institutions are to retain their force over time. These three categories of rules correspond to a more familiar division in government: executive (making, implementing, and enforcing of decisions), legislative (creation of rules), and judicial (adjudication of disputes).

The constituent parts of all resource institutions are born as a result of human choices. Three important determinants of human institutional choices are scarcity, risk, and redistribution. Greater scarcity of resources in a group creates incentives to define institutional arrangements ever more precisely so that benefits from resources are not dissipated. Indeed, some scholars have seen this dynamic of induced institutional change to be prominent enough to argue that it is universal (Netting 1981). Institutions may also change in response to higher levels of risk in the environment. Collective action necessary for institutional emergence can be conceptualized as a response that addresses the need to smoothen intertemporal fluctuations in production (Wade 1994). These views of institutional change are insufficiently attentive to variations in the incentives of group members, and to the politics that such variations generate. A more political view holds that institutional change need not occur only in response to greater scarcity in a group, nor lead to greater efficiency

(Knight 1992, North 1990). Instead, individuals and groups often act to create new resource institutions that would benefit them, rather than address problems of scarcity or risk. Particular institutions reflect existing distribution of power and wealth in a society or a group.

3. Sustainability of Resource Institutions

Perhaps the most critical question about resource institutions is how and under what conditions do they lead to sustainable management and conservation of resource systems. Group size, heterogeneity within groups, migration levels and population change, technological innovation, changing levels of demand for resources, uncertainty and sudden shocks, and cross-scale effects play an important role in determining the sustainability of institutions and resource systems (Keohane and Ostrom 1995). For certain kinds of resources, especially global commons, new institutional mechanisms have emerged to address issues of complexity and heterogenous subpopulations (Haas 1997).

Writings on resilience and adaptive management have generated important insights about the extent to which resource institutions can be sustainable in the long run (Holling 1973). Human institutions perform best when environments are predictable, yet many resource systems are characterized by potentially highly uncertain and variable processes. To date, research on institutions has not incorporated all the lessons learnt in risk-hazard studies on how sudden variations and uncertainty affect resource management needs. In part this is so because a formal theory of resource institutions requires a theory of human behavior that is mathematically rigorous and tractable, withstands empirical testing, and fits into a model of resource systems that can address fluctuations and surprise.

See also: Environmental and Resource Management; Environmental Vulnerability; Human–Environment Relationships; Institutional Economic Thought; Institutionalism; Land Degradation; Sustainable Development

Bibliography

- Goodin R E (ed.) 1996 *The Theory of Institutional Design*. Cambridge University Press, Cambridge, UK
- Haas P M (ed.) 1997 *Knowledge, Power, and International Policy Coordination*. University of South Carolina Press, Columbia, SC
- Holling C S 1973 Resilience and stability of ecosystems. *Annual Review of Ecology and Systematics* 4(1): 1–23
- Keohane R, Ostrom E (eds.) 1995 *Local Commons, Global Interdependence: Heterogeneity and Cooperation in Two Domains*. Sage, London

- Knight J 1992 *Institutions and Social Conflict*. Cambridge University Press, Cambridge, UK
- National Research Council 1986 *Proceedings of the Conference on Common Property Resource Management*. National Academy Press, Washington, DC
- Netting R 1981 *Balancing on an Alp: Ecological Change and Continuity in a Swiss Mountain Community*. Cambridge University Press, New York
- North D C 1990 *Institutions, Institutional Change, and Economic Performance*. Cambridge University Press, Cambridge, UK
- Ostrom E 1990 *Governing the Commons: The Evolution of Institutions for Collective Action*. Cambridge University Press, New York
- Wade R 1994 *Village Republics: Economic Conditions for Collective Action in South India*. Institute for Contemporary Studies Press, San Francisco
- Williamson O E 1985 *The Economic Institutions of Capitalism: Firms, Markets, Relational Contracting*. Free Press, New York

A. Agrawal

Resources and Environment: Contingent Valuation

A major impediment to performing a benefit–cost analysis involving the provision of a new public good or a change in an existing public good is that public goods are not routinely bought and sold in markets. Hence ‘prices,’ the economic data routinely used by economists as indicators of the economic value of goods, are not available for public goods such as environmental amenities. To overcome this obstacle, economists have developed (Freeman 1993) an indirect valuation approach that infers economic value from observations on consumer behavior with respect to marketed goods having a relationship to the public good, and a direct approach that gives consumers the opportunity to make choices with respect to the public good. The indirect approach relies upon one of two factors: that the public good can be ‘bundled’ in as one of the attributes of a private good (e.g., close proximity to a public park) and that it is sometimes necessary to make expenditures of money or time to use the public good (e.g., driving to a hiking trail in a national forest). The direct approach can be implemented by subjecting consumers to a constructed choice situation involving the public good, such as a vote in a referendum to provide the public good. The now common survey variant of the direct approach has come to be known as contingent valuation (CV) because the estimates of economic value obtained are ‘contingent’ on the features of the scenario posed in the survey.

CV surveys differ from other surveys on public policy issues in several important ways. First, the entire survey is devoted to describing the public good (or a small number of public goods) of interest.

Second, they differ in that their major purpose is to obtain an estimate of the relevant Hicksian consumer surplus measure; maximum willingness-to-pay (WTP) to obtain a desired good not currently possessed, or minimum willingness-to-accept (WTA) compensation to voluntarily give up a good currently possessed. CV surveys are typically organized in the following manner which reflects current practice: (a) an introductory section identifying the sponsor and general topic, (b) a section asking questions about prior knowledge about the good and attitudes toward it, (c) the presentation of the CV scenario including what the project was designed to accomplish, how it would be implemented and paid for, and what will happen under the current status quo situation if the project were not implemented, (d) question(s) asking for information about the respondent’s WTP/WTA for the good, (e) debriefing questions to help ascertain how well respondents understood the scenario, and (f) demographic questions. Mitchell and Carson (1989) provide a comprehensive overview of the issues involved in the design and analysis of CV surveys.

1. Development

Ciriacy-Wantrup (1947) put forth the first well-developed proposal on the need for CV surveys in a piece on the difficulties of measuring all the benefits of soil conservation programs. Empirical implementation of CV initiated by Davis (1963), in his Harvard dissertation, sparked considerable interest in the technique. He later compared a CV estimate with a corresponding estimate based on the travel cost method (an indirect approach then also being newly developed) and found that the two approaches produced similar estimates.

CV surveys were initially seen as having three distinct advantages. First, CV can obtain useful information where data on past consumer behavior had not been collected. Second, CV permits the creation and presentation of scenarios that provide new goods or changes in existing goods that were substantially outside the range of current consumer experience. Third, CV allows measurement of the desired Hicksian consumer surplus measure rather than its Marshallian approximation. For many economists, the major drawback to CV-based estimates was that they were based upon stated preferences rather than observed behavior.

A paper by Randall et al. (1974) greatly increased interest in CV. It valued changes in air quality necessary to maintain scenic vistas in the Southwest. The indirect valuation approach was not capable of being applied in this instance because all people in the area share the good equally (and hence, not bundled differentially into housing prices) and no expenditure of time or money is needed to enjoy it.

Randall et al. estimated what had been termed 'existence value' in an influential paper written by Krutilla (1967). The novel element in Krutilla's framework was that existence values were not generally revealed by market purchases. He argued that some people care about environmental resources, such as wilderness areas, irrespective of their desire to visit them. Krutilla did not measure existence values but rather recommended determining how large they would have to be to tilt the decision in the other direction. Failure to include existence values in policy-making, he contended, would likely entail too great a loss of existing environmental amenities and the provision of too few new environmental amenities. Other related concepts were soon enumerated (e.g., non-use value, stewardship value, bequest value, option value) and eventually encompassed into a single term, 'passive-use value,' first used in the 1989 court decision *Ohio v. Department of Interior* (880 F.2d 432, D.C. Cir.), which held that government trustees should include passive-use values in damage claims.

What Krutilla had called existence values had often been termed intangibles. They were not well integrated into welfare economic theory and thought to be unmeasurable. The key to measuring them lies in the recognition that due to scarcity effects, any form of economic value is a relative and not an absolute concept. Monetary measures of economic value are implicitly defined by choices made subject to an income constraint. CV permits the construction of the appropriate choice scenario. Also, economic value can be expressed in terms of any constraints/tradeoffs appearing in a choice scenario including time, other public goods, or private goods.

The quantity of CV research steadily increased throughout the 1970s and early 1980s. The importance of CV in the USA was raised considerably by presidential executive orders that required an assessment of the benefits and costs of all major new government regulations and reauthorization of existing ones. Outside the USA, CV was incorporated into OECD reports on measuring the economic value of pollution impacts and into World Bank reports on providing basic infrastructure. The rapid growth in the use of CV since the mid-1980s, in the USA and elsewhere, was a response to the growing demand for more comprehensive benefit-cost assessments for policies with substantial environmental or health impacts.

By far the greatest stimulus to the current CV debate, however, was the enactment of US laws that allowed for the recovery of monetary damages for injuries to natural resources. The focal point was the Exxon Valdez oil spill where the government's case was largely based upon a monetary claim for loss of passive use (Carson et al. 1992). Potential liability for lost passive use values should increase a firm's precautionary activities and environmental restoration efforts. Not surprisingly, industry facing such liability questioned whether passive use values should be

counted and whether CV estimates are reliable. A set of papers highly critical of CV from an Exxon-sponsored conference appear in Hausman (1993). In response to the industry critique, the US National Oceanic and Atmospheric Administration (NOAA) convened a Blue Ribbon Panel co-chaired by two Nobel Prize winners (Arrow et al. 1993) which concluded that passive use values should be included in damage assessments and that 'CV studies can produce estimates reliable enough to be the starting point for a judicial or administrative determination of natural resource damages—including lost passive-use value.' That report has also been influential for the guidelines that it suggested for conducting reliable CV studies for natural resource damage cases and the issues that it identified as requiring further research.

CV has been applied in more than 50 countries and there are now several thousand CV papers. Most US federal agencies with environmental, health, or natural resource responsibilities have relied upon CV studies in making policy decisions, as have many state agencies. The use of CV has dramatically increased over time in other OECD countries and international agencies regularly commission CV studies. The range of CV applications is broad. It has been used to estimate the benefits of: improving national and regional air/water quality; reducing risk from drinking and groundwater contaminants; improving outdoor recreational opportunities; protecting wetlands, wilderness, rainforests, and endangered species; providing public education; improving food and transportation safety; reducing queues for public health care services; increasing the reliability of electric and water utilities; providing cultural amenities; and improving sanitation services in developing countries.

2. Initial Research Focus

Two concerns first voiced during the early phase of CV development have been particularly influential: the possibility that respondents would not take seriously a 'hypothetical' survey question with no money directly changing hands and Paul Samuelson's warning that people would not truthfully reveal their preferences for public goods. Random response to a CV survey which was feared due to its 'hypothetical' nature could be ruled out by noting whether WTP responses systematically varied with particular covariates in the expected way. The tests economists paid the most attention to, however, were those comparing CV estimates with those from indirect techniques (e.g., hedonic pricing and travel cost analysis) for quasi-public goods, where both approaches are applicable. Early work suggested reasonably close comparison between the two, thus ruling out the extreme random behavior or strategic behaviour initially feared. Carson et al.'s (1996) meta-analysis of these comparative studies (616 comparisons from 83 separate

studies) found that CV estimates on average were slightly lower than their corresponding indirect technique estimates and the CV and indirect estimates were highly correlated.

For pure public goods, the indirect approach is not possible so one can only compare estimates based upon different types of constructed markets. With coercive payment (e.g., taxes), incentive theory suggests that CV responses should be similar to those obtained from observing voting in an equivalent referendum, the observed empirical result. With voluntary payment, theory suggests that survey estimates will be larger than true (and unobservable) WTP that in turn should be larger than actual contributions due to free riding. The observed empirical result is that survey WTP is substantially larger than actual contributions.

Other issues that concerned early CV researchers were whether the payment vehicle used should influence WTP estimates, how much information should be provided to respondents about the good, what fraction of WTP/WTa was due to passive use considerations, what the 'best' elicitation format was, why the difference between CV, WTP, and WTa estimates was larger than expected, and the relative merits of different modes of survey administration. Some of these issues have been resolved whereas others continue to be debated. The latter are discussed in the next section.

WTP estimates for the same good were shown to differ, often considerably, with different payment vehicles. Respondents were not indifferent to how public goods were financed. This was distressing to many who believed, or at least hoped, that the value of a good should be independent of the exact manner in which it was supplied and paid for. The CV estimates suggested that this was not the case.

As one would expect, WTP estimates were sensitive to the nature and amount of information provided. However, this sensitivity entailed the thorny issue of what information should be provided. CV researchers began undertaking extensive survey development work including focus groups to determine what information people wanted in making their decision. Many early CV surveys asked respondents to apportion their WTP estimate into different parts corresponding to various types of use and passive use values. Such decompositions were shown not to be unique and the various components had no distinct policy implications, so this practice has largely been abandoned.

3. Research Issues

The issue of the best elicitation format continues to be debated but is increasingly being cast in terms of a bias-variance tradeoff. The binary discrete choice format with random assignment of respondents to

different amounts (Bishop and Heberlein, 1979) has gained prominence owing to its desirable incentive properties for truthful preference revelation, in many (but not all) situations. This format has two major drawbacks. First, it collects relatively little information about a respondent's value for the good (i.e., only whether this value is above or below the amount asked), thus requiring large samples for precise estimates. Second, its estimates are sensitive to the particular parametric functional form chosen. Response formats directly asking for WTP provide the maximum amount of information (the amount making the respondent indifferent between the status quo and the alternative). This format, however, invites non-truthful strategic behavior and typically results in a sizeable fraction of the sample providing either a zero WTP amount or no response. CV researchers have proposed a number of different response formats falling between these two extremes.

Early CV WTa estimates were considerably larger than corresponding WTP estimates. This was originally thought to be a defect of CV surveys but similar divergences now appear to be typical of experimental and actual markets. A number of reasons for large divergences have been suggested (Bateman and Willis, 1999), some consistent with neoclassical economic theory and some more psychological in nature. As a consequence, interest has been renewed in measuring WTa, since the substitution of a WTP measure in its place can substantially underestimate the value of a good.

The merit of different modes of survey administration also continues to be debated. Many CV studies need visual representations of the good (e.g., pictures or maps) precluding the use of telephone surveys. The in-person vs. mail survey debate has several elements, including the much greater cost of personnel interviews, the importance of controlling the order in which material is presented, the importance of being able to read material to respondents, potential sample selection bias due to interest in the topic, and the potential influence of the interviewers.

CV researchers also came to realize that they faced the standard array of survey design issues. Adequate presentation of the material in the typical CV survey is now seen as a much more difficult task than the standard public opinion survey and is one for which most economists have had little training or experience. Further, several cognitive psychologists (e.g., Kahneman and Knetsch 1992) argued that the estimates of economic value obtained in CV surveys were likely to be subject to several undesirable effects. Chief among these was that the estimates obtained in a CV survey would be insensitive to the scope of the good being valued. They argued that stated intentions and actual behavior will diverge because respondents get a 'warm glow' from 'giving' in a survey context. While there are a few study results consistent with this prediction, a fairly large number of split-sample tests

using more fully developed CV scenarios have tested this hypothesis and reject it. Nonetheless, the Kahneman and Knetsch paper stimulated considerable interest in testing the properties of CV estimates against predictions from economic theory.

4. State of the Debate

Although there is widespread general agreement among CV practitioners on many issues, they often disagree regarding the extent and applicability of various best-practice standards, such as those proposed by the NOAA Panel, when cost and complexity factors are introduced. The recent Bateman and Willis (1999) volume considers these issues.

Participants in the early years of the CV debate were mostly environmental economists. One now finds that corporate executives, ecologists, lawyers, microeconomic theorists, philosophers, psychologists, public administrators, statisticians, and survey researchers have joined in. The increasingly interdisciplinary nature of this debate attests to the complexity of the issues and importance of CV for policy decisions (Portney 1994). Opponents of CV often conclude that CV is fundamentally flawed (Diamond and Hausman 1994). Their typical recommendation is to let the 'experts' decide. Proponents of contingent valuation acknowledge that poorly designed and implemented studies have produced problematic results that should not be used. However, they believe that these problems are not inherent to the method and are not found in the better studies (Hanemann 1994). They argue that without CV it is not possible to conduct a comprehensive economic analysis of many environmental issues and that the public's preferences concerning the provision and payment for public goods cannot be adequately represented to policymakers.

See also: Consumer Economics; Consumption, Sociology of; Ecological Economics; Environmental Adaptation and Adjustments; Environmental Economics; Public Goods: International

Bibliography

- Arrow K, Solow R, Portney P R, Learner E E, Radner R, Schuman H 1993 Report of the NOAA Panel on contingent valuation. *Federal Register* **58**: 4601–14
- Bateman I J, Willis K G (eds.) 1999 *Valuing Environmental Preferences: Theory and Practice of the Contingent Valuation Method in the US, EU, and Developing Countries*. Oxford University Press, Oxford, UK
- Bishop R C, Heberlein T A 1979 Measuring values of extra market goods. *American Journal Agricultural Economics* **61**: 926–30
- Carson R T, Flores N E, Martin K M, Wright J L 1996 Contingent valuation and revealed preference methodologies: comparing the estimates for quasi-public goods. *Land Economics* **72**: 80–99

- Carson R T, Mitchell R C, Hanemann W M, Kopp R J, Presser S, Ruud P A 1992 *A Contingent Valuation Study of Lost Passive Use Values Resulting From the Exxon Valdez Oil Spill*. Alaska Attorney General's Office, Juneau, AL
- Ciriacy-Wantrup S V 1947 Capital returns from soil-conservation practices. *Journal Farm Economics* **29**: 1181–96
- Davis R K 1963 Recreation planning as an economic problem. *Natural Resources Journal* **3**: 239–49
- Diamond P A, Hausman J A 1994 Contingent valuation: Is some number better than no number? *Journal of Economic Perspectives* **8**(4): 45–64
- Freeman A M 1993 *The Measurement of Environmental and Resource Values: Theory and Methods*. Resources for the Future, Washington, DC
- Hanemann W M 1994 Valuing the environment through contingent valuation. *Journal of Economic Perspectives* **8**(4): 19–43
- Hausman J A (ed.) 1993 *Contingent Valuation: A Critical Assessment*. North-Holland, Amsterdam
- Kahneman D, Knetsch J L 1992 Valuing public goods: The purchase of moral satisfaction. *Journal Environmental Economics and Management* **22**: 57–70
- Krutilla J V 1967 Conservation reconsidered. *American Economic Review* **57**: 777–86
- Mitchell R C, Carson R T 1989 *Using Surveys to Value Public Goods: The Contingent Valuation Method*. Johns Hopkins University Press, Baltimore, MD
- Portney P R 1994 The contingent valuation debate: Why economists should care. *Journal of Economic Perspectives* **8**(4): 3–17
- Randall A, Ives B, Eastman C 1974 Bidding games for the valuation of aesthetic environmental improvements. *Journal of Environmental Economics and Management* **1**: 132–49

R. T. Carson

Respiratory Disorders: Psychosocial Aspects

Man can do without food for weeks, without water for a few days. A few minutes without air, however, means death. Experiencing shortness of breath is associated with strong feelings of anxiety and panic. In this article, the three most prevalent respiratory disorders will be discussed, from a behavioral and social science perspective.

1. Asthma, Chronic Bronchitis, Pulmonary Emphysema

In order to understand the relevance of the contribution of behavioral and social sciences to the study of patients with respiratory disorders, in this paragraph definitions of asthma, chronic bronchitis, and emphysema, and risk factors, epidemiology, and medical management will be discussed briefly first.

Asthma is derived from Greek, meaning 'panting or distressed breathing.' A universally accepted definition of asthma is not available, reflecting changing views over time in the area of pulmonology and gaps in the

knowledge about the precise nature of asthma. Learned societies in the US, the UK, and in other countries proposed various definitions in an attempt to standardize epidemiological studies and diagnostic and therapeutic medical strategies. One of the most comprehensive definitions of asthma is given by the American Thoracic Society: Asthma is a clinical syndrome characterized by increased responsiveness of the tracheobronchial tree to a variety of stimuli. The major symptoms of asthma are paroxysms of dyspnea, wheezing, and cough, which may vary from mild to almost undetectable, to severe and unremitting (status asthmaticus)' (ATS 1987, <http://www.ginasthma.com>). Three characteristics stand out: Asthma is defined as reversible, intermittent, and variable. The episodes of airways obstruction causing dyspnea (shortness of breath) either disappear (are reversible) spontaneously or as a consequence of adequate medical, pharmacological treatment. Attacks of asthma vary in their occurrence over a certain period of time—some patients experience asthma attacks almost daily, some only once a year—and the severity of the episodes of shortness of breath varies within and between patients.

Chronic bronchitis is defined in behavioral terms: 'The presence of chronic productive cough for three months in each of two successive years in a patient in whom other causes of chronic cough have been excluded' (ATS 1995). Patients who fulfill criteria for the diagnosis of chronic bronchitis cough frequently and produce sputum over extended periods of time, and may be short of breath. In contrast to asthma, chronic bronchitis is an irreversible medical condition. Eliminating the most important risk factor—smoking tobacco—reduces the symptoms and severity of the disorder somewhat without leading to a complete reversal of the condition of the lungs and their airways.

Pulmonary emphysema is defined in terms of anatomic pathology as 'abnormal permanent enlargement of the airspaces distal to the terminal bronchioles, accompanied by destruction of their walls and without obvious fibrosis' (ATS 1995). In patients with emphysema continuous shortness of breath is the most important symptom, almost always associated with significant impairments in the ability to perform activities of daily life. Here, as is the case in patients with chronic bronchitis, smoking tobacco is the major cause of the disorder.

Over the past decades, pulmonologists, epidemiologists, researchers in related medical domains, and behavioral and social scientists viewed asthma, chronic bronchitis, and emphysema as three related disorders. Despite the differences between the three disorders, the rather similar consequences (shortness of breath, coughing, producing sputum) formed the basis for this 'lumping view.' More recently, the advocates of a 'splitting view' dominate the scientific arena. Differences in the views on causation and pathophysiological bases of the disorders are respon-

sible for this change (Vermeire and Pride 1991). This is why asthma will be discussed separately from chronic bronchitis and emphysema. These last two disorders usually are subsumed under the heading Chronic Obstructive Pulmonary Disease (COPD), defined as 'a disease state characterized by the presence of airflow obstruction due to chronic bronchitis or emphysema' (ATS 1995, <http://www.goldcopd.com>).

2. Epidemiology and Some Medical Characteristics

2.1 Asthma

The prevalence of asthma varies between different studies in different parts of the world because of differences in the definition of asthma. Large differences in data on prevalence of asthma are also attributable to differences in genetic susceptibility and environmental conditions. Asthma has a clear genetic basis: Chances for a child, conceived by a man and a woman who both are asthmatics themselves, of being asthmatic are about 75 percent. Environmental stimuli appear to play a role in worsening asthma symptoms. However, air pollution as such does not appear to be a factor that causes asthma. Interestingly, a recent study in Germany demonstrated that the prevalence of asthma in the more polluted Eastern part of Germany was lower than in the Western, more highly developed, and less polluted, part of the country (Nicolai and Von Mutius 1997). This has led researchers to hypothesize about the 'hygiene effect': Having young children growing up in a relatively unpolluted environment, with a high level of sanitation (i.e., low chance of exposure to germs causing flu and colds) predisposes a child to a higher chance of being sensitive to relatively low levels of allergens and respiratory infections, increasing their chances to develop asthma.

The prevalence of asthma is about 7 percent in the US, about 15 percent in the UK, about 1 percent among Inuits, and about 30 percent among inhabitants of Tristan da Cunha, an island in the southern part of the Atlantic ocean between South America and Africa (Gergen and Mitchell 2001). The prevalence of asthma is increasing, especially in industrialized countries. The reasons for this are largely unknown. The increased attention by physicians for a possible diagnosis of asthma in a patient with respiratory problems probably is a contributing factor in the increase of asthma prevalence: Asthma used to be underdiagnosed and undertreated (Kaptein et al. 1987). In young children, boys outnumber girls in prevalence of asthma in a ratio of about 2:1. In adult patients, hardly any sex ratio differences are observable in the prevalence of asthma. Contrary to popular beliefs, one hardly ever outgrows one's asthma, 'one only outgrows one's pediatrician.'

2.2 Chronic Obstructive Pulmonary Disease (COPD)

The prevalence of COPD is about 5 percent (ATS 1995). Recent studies predict considerable increases in the prevalence of COPD (Murray and Lopez 1997), with COPD becoming the third leading cause of death in the future. Despite reductions in the number of smokers of tobacco in industrialized countries, the post-World War II generation with their smoking habits of decades, will fall victim to becoming patients with COPD. Also, as the number of women who smoke cigarettes increased substantially after World War II, COPD will no longer be an illness for men only.

Chronic bronchitis has 50 years of age as the average age of onset, emphysema sets in at around 60 years of age. COPD has a SES-gradient, with a lower socioeconomic class dominance. COPD can be labeled as a blue-collar worker disorder due to the fact that poor housing and working conditions contribute to developing COPD.

The great majority, about 80 percent, of patients with respiratory disorders are seen by the general practitioner (or primary care physician, family physician). Based on critical views and evidence on the quality of care provided by these physicians to patients with asthma in particular, considerable effort has been devoted to improving this quality. In many countries, societies of general practitioners developed guidelines on the optimal management of patients with asthma. In these guidelines early diagnosis, aggressive medical treatment, and a great emphasis on teaching patients self-management skills are the cornerstones of modern management of asthma. Similar guidelines have been developed for the management of patients with COPD in a primary care setting, with an emphasis on stopping smoking techniques.

The remaining 20 percent of patients with respiratory disorders are seen by medical specialists: pulmonary physicians, allergologists, pediatricians. Only a very small percentage of patients is being managed on an in-patient basis—for status asthmaticus in patients with asthma, for severe respiratory problems in patients with COPD. In various countries in the world, specialized long-stay respiratory clinics provide care for those patients with asthma and/or COPD in whom the intractability combines with severe psychological and social problems.

3. Asthma and Psychology

Looking back over some 100 years of behavioral and social science research on patients with asthma, one cannot help but be impressed by the gains that have been made, gauged in terms of behavioral outcomes for these patients. Three periods in this line of research and clinical care can be discerned.

3.1 Psychosomatic Views on Asthma

Between World War I and World War II, psychosomatic views on various psychosomatic disorders and illnesses represented an influential line of research of clinical care. Given the unknown etiology of a number of major diseases (e.g., hypertension, diabetes mellitus, and asthma), theories were developed in which it was postulated that psychological factors played a substantial role in causing a number of specific illnesses. With regard to asthma, a combination of three factors was supposed to be a necessary and sufficient condition: (a) a 'weak spot' in the body of the patient, (b) a characteristic psychological profile of the patient, and (c) a situation eliciting the illness.

Patients with asthma were supposed to have a respiratory system which was sensitive to disruptions. In addition, due to faulty mother-child interaction in the early years of life, a specific personality profile was assumed to be established in the patient: A combination of excessive love of the mother for the child ('smothering with kisses') and an unconscious wish of the mother to reject her child, was supposed to render the child sensitive to asthma. In the event of a stressful life event (e.g., being separated from the mother, experiencing social conflicts), an asthma attack would happen.

Empirical evidence for these psychosomatic views is lacking. No study identified the supposed mother-child psychopathology put forward by the theory. Psychosomatic therapy (i.e., psychodynamic interventions) failed to demonstrate any positive effect. It is safe to state that the highly selected sample of patients with asthma seen by the theorists (psychiatrists and specialists in internal medicine) represented patients who exhibited psychosocial problems as a consequence of having to cope with the adaptive tasks of having severe, brittle asthma.

3.2 Psychological Consequences of Asthma

The second theoretical and empirical approach to asthma from a behavioral and social science perspective focused on the consequences of asthma on the daily lives of afflicted patients: their quality of life. Using disease-specific questionnaires with adequate psychometric properties, limitations in functional, psychological, and social domains were identified. Asthma patients experience excessive absenteeism from school and work, heightened levels of anxiety and depression, and feelings of social isolation (Maille et al. 1996).

The first studies on the effects of psychological interventions in order to reduce the limitations caused by asthma adopted a learning theory perspective: relaxation training, biofeedback, and systematic desensitization. In these studies, pulmonary function was the major outcome measure. The effects of

behavioral interventions turned out to be limited, however (Lehrer 1998).

The research performed by a group of psychologists in Denver, Colorado, USA, in the 1970s and 1980s set the scene for the third phase in psychological research on patients with asthma. Kinsman et al. (1982) demonstrated how the affective and cognitive responses of asthma patients to their illness and symptoms predicted important outcome measures such as length of hospitalization and rehospitalization, irrespective of the medical severity of the respiratory condition.

3.3 Self-management Training

A logical next step, given the evidence for the impact of the psychological response (coping behavior) of patients to their asthma, pertains to teaching asthma patients to optimize their handling, or self-management, of the disorder. Creer (1979) represents the first psychologist who studied the effects of this type of intervention in children and adults. In self-management training, patients with asthma are taught skills to perceive the first signs of periods of shortness of breath, take appropriate action to prevent or control episodes of shortness of breath, and manage bouts of asthma if they occur. Reviews of the effects of self-management training report reductions in symptom frequency and severity, use of medical resources, and fewer disruptions of daily life (e.g., absenteeism from school or work, levels of anxiety) (Clark and Gong 2000).

Current research on asthma by psychologists focuses on refining self-management training, symptom perception, and the effects of emotional expression on pulmonary function (Smyth et al. 1999).

4. COPD and Psychology

In contrast to the considerable amount of research performed by psychologists and social scientists on patients with asthma, patients with COPD have been the subject of research from this perspective only relatively recently. There are at least two reasons for this. The impact of smoking tobacco, the major cause of COPD, has become clear only after World War II; psychologists and social scientists joined medical professionals in their therapeutic pessimism for decades. The first empirical study on the psychological consequences of COPD was published only in 1961.

No psychosomatic theories were developed on patients with COPD. Studies on the neuropsychological consequences demonstrated that severe grades of COPD negatively affect cognitive functioning of patients. Depression, fatigue, and feelings of social isolation are the dominant psychological and social consequences of COPD. Restrictions in daily activi-

ties, absenteeism from work, early retirement, and invalidity are major functional consequences of COPD. The quality of life, therefore, of patients with COPD is significantly lower than in healthy persons (Kaptein and Dekker 2000).

Psychologists contribute to the care for patients with COPD in two areas: pulmonary rehabilitation programs and self-management training. Pulmonary rehabilitation pertains to 'an art of medical practice wherein an individually tailored multidisciplinary program is formulated which through accurate diagnosis, therapy, emotional support, and education stabilizes or reverses both the physio- and psychopathology of pulmonary diseases and attempts to return the patient to the highest possible functional capacity allowed by his pulmonary handicap and overall life situation' (Ries 1997). Physical therapy, breathing exercises and training, smoking cessation techniques, graded exercise aimed at improving ambulation, and coping skills training usually are parts of pulmonary rehabilitation programs. Reviews and meta-analyses of the effects of these programs indicate that use of health services, limitations in daily activities, and psychological problems tend to be reduced to some degree (Kaplan and Ries 2001). High dropout percentages, however, are the rule. Identifying the separate contributions of the various components in pulmonary rehabilitation programs to various outcome measures is not feasible, however.

The current emphasis in health psychology research in COPD focuses on teaching patients self-management skills. Cognitive behavior modification techniques are used to alter nonadaptive cognitions and behavior. In a study of patients' pessimistic self-statements ('I can't walk very far without getting short of breath, so what's the use?') were addressed and altered into more active ones ('This walking is uncomfortable, but I can handle it. Soon I will be able to walk farther.'). Results showed that levels of exercise tolerance and quality of life increased in comparison to various control conditions (Atkins et al. 1984).

Increasingly, guidelines on the medical management of patients with COPD include results of psychological research on self-management (ATS 1995). This development is positive: it reflects how pulmonary physicians acknowledge the relevance of topics such as compliance or adherence, and incorporating self-management into clinical care. Patient empowerment is instrumental for the wellbeing of patients with this chronic respiratory disorder.

Future research efforts will strengthen the position of patients with COPD. Health psychologists who work in the area of COPD face many challenges. Smoking cessation is a primary subject for health psychologists and other social scientists, given the psychological and societal determinants of smoking tobacco in particular. Simplifying self-management skills and incorporating the teaching and uptake of these skills into medical care are additional challenges.

COPD is an under-funded medical disorder (Gross et al. 1999). Future research both from a medical and a psychological perspective will help in preventing COPD to develop, and in improving quality of life in patients with COPD.

See also: Health Education and Health Promotion; Smoking Prevention and Smoking Cessation

Smyth J M, Stone A A, Hurewitz A, Kaell A 1999 Effects of writing about stressful experiences on symptom reduction in patients with asthma or rheumatoid arthritis. *Journal of the American Medical Association* **281**: 1304–9

Vermeire P A, Pride N B 1991 A 'splitting' look at chronic nonspecific lung disease (CNSLD): Common features but diverse pathogenesis. *European Respiratory Journal* **4**: 490–6

A. A. Kaptein

Bibliography

- Atkins C J, Kaplan R M, Timms R M, Reinsch S, Lofback K 1984 Behavioral exercise programs in the management of chronic obstructive pulmonary disease. *Journal of Consulting and Clinical Psychology* **52**: 591–603
- ATS 1987 Standards for the diagnosis and care of patients with chronic obstructive pulmonary disease (COPD) and asthma. *American Review of Respiratory Disease* **136**: 225–44
- ATS 1995 Standards for the diagnosis and care of patients with chronic obstructive pulmonary disease. *American Journal of Respiratory and Critical Care Medicine* **152**: S77–S120
- Clark N M, Gong M 2000 Management of chronic disease by practitioners and patients: are we teaching the wrong thing? *British Medical Journal* **320**: 572–5
- Creer T L 1979 *Asthma Therapy: A Behavioral Health Care System for Respiratory Disorders*. Springer, New York
- Gergen P J, Mitchell H E 2001 Epidemiology of asthma. In: Kaptein A A, Creer T L (eds.) *Respiratory Medicine and Behavioral Research*. Routledge, London
- Gross C P, Anderson G F, Powe N R 1999 The relation between funding by the National Institutes of Health and the burden of disease. *The New England Journal of Medicine* **340**: 1881–7
- Kaplan R M, Ries A L 2001 COPD: Behavioral assessment and treatment. In: Kaptein A A, Creer T L (eds.) *Respiratory Medicine and Behavioral Research*. Routledge, London
- Kaptein A A, Dekker F W 2000 Psychosocial Support. *European Respiratory Monograph* **13**: 58–69
- Kaptein A A, Dekker F W, Gill K, Van der Waart M A C 1987 Undertreatment of asthma in Dutch general practice. *Family Practice* **4**: 219–25
- Kinsman R A, Dirks J F, Jones N F 1982 Psychomaintenance of chronic physical illness. In: Millon T, Green C, Meagher R (eds.) *Handbook of Clinical Health Psychology*. Plenum, New York, pp. 435–66
- Lehrer P M 1998 Emotionally triggered asthma: a review of research literature and some hypotheses for self-regulation therapies. *Applied Psychophysiology and Biofeedback* **23**: 13–41
- Maillé A R, Kaptein A A, de Haes J C J M, Everaerd W T A M 1996 Assessing quality of life in chronic non-specific lung disease: A review of empirical studies published between 1980 and 1994. *Quality of Life Research* **5**: 287–301
- Murray C J L, Lopez A D 1997 Alternative projections of mortality and disability by cause 1990–2020: Global burden of disease study. *Lancet* **349**: 1498–1504
- Nicolai T, Von Mutius E 1997 Pollution and the development of allergy: The East and West Germany story. *Archives of Toxicology* **19** (suppl.): 201–6
- Ries A L 1997 Pulmonary rehabilitation—Joined ACCP/AACVPR evidence-based guidelines. *Chest* **112**: 1363–96

Responsibility: Philosophical Aspects

1. The Concept of Responsibility

The focus of this entry is moral responsibility. Although it begins with an examination of the concepts of causal responsibility and of responsibility as duty, the discussion in these sections is intended to orient the reader to the distinctive character of the concept of moral responsibility and to attendant philosophical theories about what is meant when moral responsibility is ascribed to a human agent for some act or state of affairs. Attributing responsibility to an agent as to a cause is logically and conceptually distinct from claiming that the person is morally responsible for causing an event. Equally, discerning those tasks for which a person has a responsibility or a duty to perform is different from claiming that they bear moral responsibility for the performance of those tasks. In short, the *concept* of responsibility has a number of distinct applications, the analyses of which point to the variety of *conceptions* of responsibility.

1.1 Causal Responsibility

One interested in action theory may investigate *causal* responsibility. Thus, in claiming 'Alice is responsible for spilling the coffee on the floor,' we may be saying that Alice spilled the coffee as a result of some action on her part. In such a causal context, responsibility ascriptions credit a person or an entity with a role in bringing about an event or state of affairs, and nothing more. Assigning causal responsibility consists in identifying a salient factor that must actually operate in order for an event to transpire—its cause—and not in evaluating the moral quality of an event with which an actor is instrumentally associated.

Moral responsibility, and responsible agency, cannot be identified with causal responsibility. This is true for at least three reasons. First, that someone or something is causally responsible for some event or state of affairs is not sufficient for the thing being morally responsible for the event. The kind of responsibility attributed to one who is identified as

causally responsible for something is not a sufficient basis for moral evaluation. That the agent is causally responsible is not, in and of itself, a reason to judge the agent a good or bad, praiseworthy or blameworthy, person as a result. But judgments of moral responsibility are typically of this sort.

Second, a person need not be causally responsible for all those acts they are held morally responsible for—causal responsibility is not necessary for moral responsibility. For example, parents are vicariously responsible, legally and, although this is controversial, perhaps morally, for at least some of the acts of their nonadult children.

Third, that someone or something is causally responsible for some event or state of affairs is not sufficient for the cause to be an agent who can be held morally responsible. Because causal responsibility can be ascribed to events, things, nonhuman animals, and to irrational or underdeveloped humans as well as to fully mature, rational persons, it does not single out a class of things that might properly be described as moral *agents*, or *members of a moral community*. When we hold a person such as Alice morally responsible for something, it is in light of certain expectations we have of Alice, given a maturity we assume of her, given what we have reason to believe she knew herself to be doing, and given the circumstances that led her to act. But the conditions for responsible agency met by Alice, and considered more fully below, are not met by very young children, or animals, or trees. Thus, while we might claim that faulty construction is to blame for the collapse of a building, what we mean is that a salient cause of the collapse was poor construction. We might allege that the baby was causally responsible for the broken lamp. But neither the baby nor the faulty construction are morally responsible for the states they cause because neither are moral agents.

In sum, an account of agency must make sense of moral approval or disapproval. Even if the act for which we deem a person responsible is morally neutral—is neither praiseworthy nor blameworthy—we appraise the moral quality of the person who acts, and not simply the state of affairs the person initiates.

Baier's notion of 'task responsibility,' or the 'forward-looking aspect' of agent responsibility. In its forward-looking sense, agent responsibility consists in being socially entrusted, on either customary, legal, or moral grounds, with a specified event, state of affairs, or duty.

Thus, parents are responsible for overseeing the well-being of their children, adult wage earners are responsible for paying taxes, attorneys are responsible for respecting the confidentiality of their clients, and so forth. In this forward-looking sense, the claim that Alice is responsible for giving Bernie the medicine would mean, at a minimum, that Alice has a duty to give Bernie medicine. The sentence is then ambiguous between (a) the sense in which it attributes a prospective duty to Alice; (b) the sense in which it attributes mere causal responsibility to Alice for Bernie's having been given the medicine, and (c) the sense in which it attributes moral responsibility and implies, perhaps, praise or blame.

In the latter sense, to be morally responsible is not a matter of having or owing an obligation to another. A person may be morally responsible for those things they have a moral obligation to do, such as caring for their children. But the things for which a person can be held morally responsible in the sense of being praiseworthy or blameworthy is not restricted to things they had an obligation or a responsibility to do. I might be held responsible for my slovenly habits, or for my acts of bravery, where neither of these attends the discharging of an obligation. Equally, the fact that an act is one I am obliged to perform does not decide whether I am morally accountable for performing it or for failing to do so.

Once we have set aside causal responsibility and responsibilities in the sense of duties, we are left with the idea of responsible agency and the idea of responsibility for action. What we are interested in is the idea of *being* responsible—both being a responsible person and being responsible for an act.

2. *Being a Responsible Agent and Being Held Responsible*

1.2 *Duty and Obligation*

By contrast, in describing a person as 'responsible' for some act or state of affairs, we may be expressing the belief that the person has a duty. Here, the judgment 'X is responsible for Y' is intended as a statement to the effect that X has certain responsibilities to others: tasks, duties, or obligations pertaining to Y, usually in light of X's position in society or role. Claiming that a person is responsible for Y in this sense is to credit responsibility *prospectively*; it is to look forward to what the person is duty bound or obligated to do. This idea of having responsibilities is analogous to Kurt

2.1 *The Conditions for Responsible Agency*

In order to be morally responsible for an act, or trait of character, a person must be a responsible agent. Responsible agency typically requires that an individual satisfy certain *epistemic* conditions and certain social and psychological conditions of *control*.

The epistemic conditions guarantee that the responsible agent is self-aware, that they are rational, that they are not ignorant of the circumstances in which they act, that they are cognizant of and able to act within established moral guidelines, and that they

are responsive to reasons to adjust or amend their behavior in light of these guidelines. In order to be held responsible, the moral agent must know that doing a particular act (or an act of a given type) or cultivating a particular trait of character (say, jealousy, rage, bigotry) is right or is wrong. The moral agent may be held responsible if, suffering from none of the conditions that exempt a person from responsible agency, they should have known the nature of the act or trait, and could have been motivated by the relevant moral guidelines. A person's ignorance in this case might be called culpable. By contrast, a person should not be held responsible, if they did not know, and should not be expected to know, that the act is of an undesirable moral quality.

The control conditions ensure that the agent acts freely. The morally responsible agent is not hostage to debilitating and uncontrollable neuroses of a sort that incapacitate their resolve; that is, they can control the manifestations of neuroses by medical means or by strength of will. A person who is prone to neurosis of a certain type or scope, or who finds themselves the victim of coercion or manipulation, may be a responsible agent but one whom we excuse from responsibility for specific acts when it is plain that on a particular occasion coercive or manipulative elements played a decisive and ineliminable role in action.

Satisfaction of the control conditions for responsible agency ensure that the agent, rather than a mechanism foreign to them, actually guides or actually executes the agent's actions and those traits of character for which the agent is morally responsible. A mechanism of action is foreign to a person when it is one to which the person does not assent as a force by which they wish to be moved. Although the responsible person may lack the power to alter the mechanism that leads them to act—to do otherwise than they in fact do—they are moved by their own lights. At a minimum, the morally responsible agent is capable of deliberate, intentional behavior.

Absent some of the epistemic and control conditions, a person will be a less likely candidate for responsible agency. The satisfaction of some set of these will mean that the entity is a responsible agent. There is obviously a threshold that decides when a person is a responsible agent, but it is not within the scope of this discussion to explore the specifics of the threshold. Let it suffice to say that the responsible agent is someone who acts knowingly and freely, and the actions and characteristics for which they are responsible are 'their own.'

2.2 Holding Responsible and Being Responsible

Generally speaking, when we describe someone as a morally responsible agent, we mean that the person is someone whom we recognize as bearing characteristics of the sort that make them a member of the moral

community. They are someone with whom morally relevant relationships are possible, and they are a fitting subject of moral appraisal. But there are different theories about the analysis of judgments that a person is morally responsible for their character, their choices, their actions or omissions. Three accounts, in particular, bear mention.

One possibility is that we might simply be signaling that an actor, having met certain conditions for responsible agency, is a suitable candidate for certain kinds of corresponding regard and treatment forthcoming on the part of others and, perhaps, on the part of himself. This approach has been coined the reactive attitudes analysis of responsibility ascriptions. The idea is that when we say a person is responsible we mean that the person is liable, through their behavior, to a response on the part of others marked by attitudes such as gratitude and resentment, and that the person is an appropriate object of such attitudes in light of certain conditions for agency that they have met. On this analysis, the claim 'X is responsible for Y' (or, for Y-ing) will amount to the conjunction of the claims 'X did (or is) Y,' 'X possesses the qualities of a responsible agent,' and 'X is rightly subject to the favorable or unfavorable responses of others as a result.' The reactive attitudes analysis offers an account of the emotional dynamics at play where decisions about responsibility occur. That a person is appropriately subject to certain attitudes and practices gives us reason to consider them a free, responsible agent, and as good or bad, praiseworthy or blameworthy, and so on.

A second possibility is that, in judging a person responsible, we may be saying the person is accountable for their behavior, in the sense that we think it appropriate that the person offer an explanation of their intentions or beliefs about their behavior. That is, we may be focusing on the phenomenon of holding an agent answerable for their actions. Construing judgments about moral responsibility as judgments about accountability is appropriate for a number of reasons. One reason is that holding persons responsible is part of a social practice of ascribing responsibility, first by making clear any moral expectations we have of persons vis-à-vis behavior and character, and then by having in place certain social instruments that ensure adequate redress where necessary. Making moral expectations clear presumes the existence of a shared set of beliefs, and a system of norms against which actions and traits of character are assessed.

A third possibility is that judging a person responsible for some act, omission, state of affairs, or trait of character is essentially a statement of attribution. The philosopher Gary Watson advocates this perspective, which he labels the *aretaic* face of responsibility, for its focus on ethical appraisal of conduct that can be attributed or imputed to an individual as its agent and thus as the author or adopter of ends. Because these ends are self-adopted,

acting upon them discloses the quality and character of the agent's life. The attributability approach does not merely assign causal responsibility, for the focus is on the behavior of a human agent, his/her character, and his/her ends—the object leading to action. Acts of attributing responsibility on this account differ from both the reactive attitudes and accountability analyses. Attribution need not involve the registration of emotions such as resentment or gratitude, nor need it be true that an explanation of the act in question is expected of the actor.

Two varieties of cases illustrate the differences among these three analyses. First consider examples of extreme and, some would contend, irremediable evil. If we consider the behavior of someone such as Hitler, it may well appear that he is morally 'beyond the pale' so that none of the standard reactive attitudes is an appropriate response to, for instance, his 'Final Solution.' Hitler, and like-minded persons, are seemingly incapacitated for ordinary adult interpersonal relationships. Such persons simply will not be responsive to the demands of membership in our moral community. At the same time, the heinous acts they have committed are properly attributed to them as perpetrators. The attributability account allows for Hitler, and similarly sociopathic or psychopathic persons, to be deemed responsible for their immoral behavior, yet permits the hesitation some might have in regarding such actors as moral interlocutors, as the accountability approach demands, or as appropriate objects of any of the ordinary moral attitudes.

A second, ostensive, merit of the attributability approach is that it enables us to credit persons with responsibility for matters that are essentially private, matters that are the affair of none but the actor himself. It seems unreasonable to reply to my slovenly garb, or foolish financial practice, with resentment or indignation, just as it seems unreasonable to expect that I account for these practices if they affect no one but myself. Where these practices reveal something about ends I have chosen for myself, they speak to my character—they express, as Watson claims, my own evaluative commitments and my identity as an agent. Thus, on the attributability account, they can be attributed to me as practices for which I bear responsibility.

3. Responsibility and Free Will

3.1 Causal Determinism and Moral Responsibility

An adequate discussion of the central problems faced by any account of moral responsibility will include the problem causal determinism poses to the idea that we have freedom of will. In brief, the problem is that if our present choices and conduct are causally determined by past events and states, then it appears that we do

not have any freedom with respect to the performance or occurrence of our present choices and conduct. The past is something over which we have no control. And it is equally true that we have no control over the laws of nature operative in the world. Together, these claims suggest that we have no control over the present.

It is a matter of lively debate among philosophers whether or not the truth, or the likelihood, of causal determinism would affect our status as responsible agents. Some philosophers argue for the incompatibility of determinism and responsibility; if either obtains, the other does not. We cannot be both determined and responsible for our conduct. Others argue that the conditions for responsible agency are independent of determinism. On this view, the truth or plausibility of determinism aside, we can control our conduct in a manner sufficient for responsible agency. At the very least, to deny responsibility of persons whom we regard as participants in our moral framework would be to strain the limits of our natural practices. It is undeniable that persons treat one another as agents, as individuals answerable for their conduct, and that the alternative—that persons cultivate an attitude of detachment from the conduct of their human associates merely because of the truth of determinism—would signal an impoverished state of human interaction that few, if any, could bear.

See also: Autonomy, Philosophy of; Ethics and Values; Free Will and Action; Freedom/Liberty; Impact on the Social Sciences; Identity and Identification: Philosophical Aspects; Intentionality and Rationality: A Continental-European Perspective; Intentionality and Rationality: An Analytic Perspective

Bibliography

- Baier K 1972 Types of responsibility. In: French P A (ed.) *Individual and Collective Responsibility*. Schenkman Publishing, Cambridge, MA
- Frankfurt H 1988a Alternate possibilities and moral responsibility. In: *The Importance of What We Care About*. Cambridge University Press, Cambridge, UK, pp. 1–10
- Fischer J M, Ravizza M 1998 *Responsibility and Control*. Cambridge University Press, Cambridge, UK
- Frankfurt H 1988b Freedom of the will and the concept of a person. In: *The Importance of What We Care About*. Cambridge University Press, Cambridge, UK, pp. 11–25
- Haji I 1998 *Moral Appraisability: Puzzles, Proposals, and Perplexities*. Oxford University Press, New York
- Harman G 1975 Moral relativism defended. *The Philosophical Review* **84**: 3–22
- Oshana M 1997 Ascriptions of responsibility. *American Philosophical Quarterly* **34**: 71–83
- Pettit P, Smith M 1996 Freedom in belief and desire. *The Journal of Philosophy* **93**: 429–49

- Strawson P 1993 Freedom and resentment. In: Fischer J M, Ravizza M (eds.) *Perspectives on Moral Responsibility*. Cornell University Press, Ithaca, NY, pp. 45–66
- Van Inwagen P 1983 *An Essay on Free Will*. Clarendon Press, Oxford, UK
- Wallace R J 1994 *Responsibility and the Moral Sentiments*. Harvard University Press, Cambridge, MA
- Watson G 1996 Two faces of responsibility. *Philosophical Topics* 24: 227–48
- Wolf S 1990 *Freedom Within Reason*. Oxford University Press, New York
- Zimmerman M J 1988 *An Essay on Moral Responsibility*. Rowman and Littlefield, Totowa, NJ

M. Oshana

Retail Trade

The term ‘retail geography’ describes the spatial structure of retailing, especially its transformation. The dominant methodology of the retail geography literature combines analyses of spatial demand patterns, producer behavior, and travel patterns, within the broader spatial patterns of economic activities. The geography of retailing is related to ‘marketing geography’ and the two terms are used somewhat interchangeably, although marketing geography encompasses the geographic patterns of consumption more broadly.

Explanations of the geography of retailing have traditionally been based on the conceptual frameworks of location theory (see *Location Theory*), especially the classical models of central place theory (Christaller 1933, Lösch 1941). Two themes, differentiated by the type of locational system investigated, developed early. The first studies the spatial pattern of periodic markets. The second focuses on regional systems of permanent retail outlets. A corollary of the difference between the two themes is their different regional foci. Studies of periodic markets are mostly centered in Asia, Africa, and Latin America, while the second theme has more of an Anglo-American focus. The third theme in the retail geography literature presents analytical methods of studying travel patterns to retail stores and of determining optimal locations.

1. The Geography of Periodic Markets

Periodic markets, which continue to be an important element of rural retail trade in many parts of the world, meet periodically at specific locations with set schedules, separated by market-less days or by days with limited trading activities. Such periodic markets are the basic component of the spatial structure of rural retailing in Asia, Africa, and Latin America. Skinner (1964) posited that periodic markets cor-

respond to the lowest levels of the hierarchy of central places that serve the need for goods and services in an economy. The markets associated with each level possess distinctive size, functionality, and periodicity characteristics (Berry and Parr 1988). Central place theory is a leitmotif in most studies of the geography of periodic markets.

1.1 Models of Periodic Market Systems

Stine (1962) was the first to develop a theoretical model to explain why markets are periodic. In low-density areas the level of demand is often too low to cover the fixed cost of a permanent store. Traders can overcome this constraint by visiting several marketplaces in a regular, periodic sequence to aggregate demand from a wider geographical area. Hay (1977) and Ghosh (1982), among others, extended Stine’s analyses. They outline the conditions under which a trader will adopt one of the following strategies—mobile or itinerant trading, part-time trading from a fixed location, or full-time trading from a fixed location. Traders may choose an itinerant strategy even when the expected profit at a fixed location exceeds the cost of operating a permanent store, if the incremental net earnings from a new location are greater than the sum of relocation and overhead costs. Similarly, mobility may allow traders to reduce prices and forestall competition. In general, the lower the relocation and overhead costs in relation to demand, the greater the likelihood of mobile systems. This is one reason why in many rural areas one tends to find large numbers of itinerant traders in markets that do not have a single permanent outlet.

1.2 Spatio-temporal Organization of Markets

An important geographic characteristic of periodic marketing systems is their integration of both the spatial and temporal dimensions of demand. Consumers perceive the adjacent marketplaces held on consecutive days—the market ring—as belonging to a particular spatio-temporal set of markets that serve their needs. According to a number of observers, the location and timing of markets have evolved to maximize access to consumers within a specified interpurchase interval. This leads to an inverse relationship between the location of two markets and the time interval between their market meetings. Markets that meet simultaneously or are separated by only a short time interval should be more geographically dispersed (Fagerlund and Smith 1970). A counter-hypothesis is that itinerant traders develop a sequence that minimizes their travel costs (Hill and Smith 1972). Empirical studies provide only partial support for the hypotheses.

2. *The Morphology of Urban Retail Outlets*

The organization of retail centers within cities follows a hierarchical pattern characterized by greater functional complexity of centers as one moves to higher levels of the hierarchy. The geographic reach of the center expands as one moves up the hierarchy due to its centrality, and retail businesses locate in different levels of the hierarchy depending upon the sizes of the market area required to generate adequate economic returns. The result is a hierarchical structure of retail locations similar to the tenets of central place theory. Much of the early literature on retail geography comprises empirical studies of urban retail hierarchies in North America and Europe conducted to test central place theory. Later the emphasis shifted to investigating the dispersion of retailing from urban areas by combining the analyses of spatial demand patterns with that of the changes in the social and economic conditions as well as changes in the organizational structure of the retail sector.

2.1 *The Transformation of Urban Retail Geography*

The dominant feature of the transformation of the post-World War II retail geography in North America and Europe is the shift of retail activities from city centers to suburbs. Before World War II, the central business district (CBD) ranked as the pre-eminent metropolitan shopping district, offering a vast array of goods and services and attracting shoppers from throughout the metropolitan region. That premier position started to fall in the 1950s and 1960s in large urban areas, and later in smaller cities too. While retail activity declined in the metropolitan core, however, it grew rapidly on the suburban periphery both in North America and Europe, reflecting the spatial redistribution of population and income in metropolitan areas (Davies and Rogers 1984). The suburb's share of retail sales increased significantly, and major shopping centers started to rival the CBD in retail sales volume.

The decentralization of sales was linked closely to the emergence of new retail formats. Shopping centers, housing diverse retail stores under one roof, represented a dramatic departure from earlier retail formats. Unlike the traditional downtown shopping district, they were planned, integrated complexes developed and marketed as a unit (Dawson and Lord 1985). Initially suburban shopping was concentrated along highways in strip malls and in neighborhood convenience centers. This was followed by the development of large, regional shopping centers, which became the focal point of the metropolitan retail structure, drawing shoppers from a wide geographical area. Thus, the new geography of retailing is much more dispersed than in the past, comprising a well-developed retail network focused on various levels of

regional and neighborhood shopping centers and shopping districts.

2.2 *The Context of Change*

The suburbanization of population and the evolving transportation infrastructure provided the motive force for the spatial dispersion of retailing. But institutional, financial, and governmental policies also shaped this new geography. The establishment of shopping centers was greatly encouraged by suburban municipal governments, who viewed shopping centers as an important source of tax revenues. Small municipalities on the outskirts of larger towns encouraged the establishment of shopping centers within their jurisdictions to attract shoppers from the larger towns, and, thereby, increase tax revenues. The growth of shopping centers was also facilitated by easy access to financing and financial arrangements that greatly reduced the developer's risk. The consolidation of the retail industry and the growth-oriented philosophies of department and specialty store chains further fueled the expansion.

While local governments in suburban communities actively encouraged the growth of shopping centers, their counterparts in urban areas viewed them as grave threats. Retailers in older, urban shopping districts faced two competitive disadvantages. First was the lack of population, income, and employment growth. Second was the general deterioration of the shopping environment, which reduced the CBD's ability to attract middle- and upper-income shoppers. Concerned by the erosion of urban retail business and its impact on city finances, a number of governmental programs and policies were initiated to encourage retailing in CBDs. Some of these new downtown retail projects included pedestrian shopping malls, retail facilities built within large downtown office complexes, and anchorless centers housing entertainment, recreation, food, and specialty retailing. They all represented departures from the traditional hierarchical structure of retail organization.

3. *Analytical Models of Store Location*

3.1 *Retail Forecasting Models*

The rapid expansion of suburban outlets created the need for a systematic approach to determining store locations, resulting in an array of techniques for site selection and retail sales forecasting. The key to forecasting retail sales is to delineate accurately the store's trade area—the geographic area from which the store draws most of its customers. The foundation of contemporary approaches to delineating trade areas was laid by Applebaum (1968) in his 'customer spotting' method. Huff's spatial interaction model (see

Spatial Interaction Models) extended the approach by looking at the complex interactions within the total system of retail trade areas in a region (Huff 1963). Underlying spatial interaction models is the premise that the share of customers a retail outlet attracts from an area is inversely proportional to distance and directly proportional to the attraction of the store. These models develop an empirical basis to determine the relative impact of distance and attraction on store choice, and conceive trade areas as probabilistic rather than deterministic, with each store in an area having some probability of being patronized. Huff's probabilistic revealed preference approach to modeling store choice was extended with more sophisticated analytical techniques and the definition of a broader set of variables to measure store attractiveness (Ghosh and McLafferty 1987).

3.2 Location-allocation Models

Determining the optimal locations for a network of retail outlets is a complex task, requiring decisions on the number of stores to locate, their locations, and their characteristics. All the decisions must be integrated within a single decision-making framework, taking into consideration the geographical pattern of demand, transportation, and corporate objectives. Location-allocation models systematically evaluate a large number of locations to find sites that optimize objectives such as market share or accessibility, taking into consideration system-wide store-location interactions (Ghosh and McLafferty 1987). Because the models can look at the impact of each store on the entire network of stores, they are ideally suited to analyzing the location decision of retail chains operating multiple outlets in the same market area (Achabal et al. 1982). Location-allocation is a flexible analytical system in that a variety of objective functions and constraints can be incorporated (Ghosh and Harche 1993). The combination of machine-readable economic and geodemographic data, geographical information systems software, and location-allocation models create effective location decision support systems (Birkin et al. 1996).

4. Conclusion

The structure of retail trade, like most economic activities, is shaped to a large extent by geography. Indeed, it is often said that the three secrets to retailing are location, location, and location. Although this oft-repeated adage is somewhat of an exaggeration, even in the era of e-commerce and electronic retailing, the geographical pattern of retail trade is shaped by the relationship between the patterns of consumer demand and store locations, mediated by consumer travel patterns. The literature on retail trade provides

frameworks for understanding how these forces, mediated by public and corporate policies, shape the observed pattern of retail trade and presents empirical studies from diverse regions of the world. It also presents analytical methods for locating retail outlets.

See also: Cities, Internal Organization of; Commodity Chains; Consumer Economics; Consumption, Sociology of; International Business; International Law and Treaties; International Trade: Commercial Policy and Trade Negotiations; International Trade: Economic Integration; International Trade: Geographic Aspects; Market Areas; Market Structure and Performance; Marketing Strategies; Multinational Corporations; Urban Planning; Central City Revitalization

Bibliography

- Achabal D D, Gorr W L, Mahajan V 1982 MULTILOC: A multiple store location model. *Journal of Retailing* **58**: 5–25
- Applebaum W 1968 *A Guide to Store Location Research, with Emphasis on Supermarkets*. Addison-Wesley, Reading, MA
- Berry B J L, Parr J B 1988 *Market Centers and Retail Location: Theory and Applications*. Prentice Hall, Englewood Cliffs, NJ
- Birkin M, Clarke G, Clarke M, Wilson A 1996 *Intelligent GIS: Location Decisions and Strategic Planning*. Geoinformation, Cambridge, UK
- Christaller W 1933 *Die zentralen Orte in Süddeutschland*. Fisher, Jena, Germany [trans. Baskin C W 1966 *The Central Places of Southern Germany*. Prentice Hall, Englewood Cliffs, NJ]
- Davies R L, Rogers D S 1984 *Store Location and Store Assessment Research*. Wiley, Chichester, UK
- Dawson J A, Lord J D 1985 *Shopping Centre Development: Policies and Prospects*. Croom Helm, London
- Fagerlund V G, Smith R H T 1970 A preliminary map of market periodicities in Ghana. *Journal of Developing Areas* **4**: 333–47
- Ghosh A 1982 A model of periodic marketing. *Geographical Analysis* **1**: 155–66
- Ghosh A, Harche F 1993 Location-allocation models in the private sector: Progress, problems and prospects. *Location Science* **1**: 81–106
- Ghosh A, McLafferty S L 1987 *Location Strategies for Retail and Service Firms*. Lexington Books, Lexington, MA
- Hay A 1977 Some alternatives in the economic analysis of periodic marketing. *Geographical Analysis* **9**: 72–9
- Hill P, Smith R H T 1972 The spatial and temporal synchronization of periodic markets: Evidence from four emirates in northern Nigeria. *Economic Geography* **48**: 345–55
- Huff D L 1963 A probabilistic analysis of shopping center trade areas. *Land Economics* **39**: 81–90
- Lösch A 1941 *Die räumliche Ordnung der Wirtschaft* Fisher, Jena, Germany [trans. into English 1944 *The Economics of Location*. Yale University Press, New Haven, CT]
- Skinner G W 1964 Marketing and social structure in rural China I. *Journal of Asian Studies* **24**: 3–43

Stine JH 1962 Temporal aspects of tertiary production elements in Korea. In: Pitts FR (ed.) *Urban Systems and Economic Development: Papers and Proceedings*. University of Oregon, Eugene, OR

A. Ghosh

Retirement and Health

1. Retirement in Historical Perspective

Retirement as a structural aspect of the life course is a relatively recent phenomenon, institutionalized in industrialized societies only in the twentieth century. 'Retirement' is typically defined as later life withdrawal from the workforce, often in conjunction with public and/or employer-provided pension benefits. But growing numbers of workers retire from their career jobs and then take on employment in another job and/or unpaid community service. Retirement—in terms of eligibility for a pension—can no longer be equated with a one-way, one-time exit from the workforce or with the cessation of all productive activity. Neither is retirement occurring at one set age (such as 62 or 65). The proliferation of public and private retirement income programs have encouraged a worldwide trend toward progressively earlier retirement from career jobs (Delsen and Reday-Molvey 1996, Guillemard and Rein 1993), and greater heterogeneity in the age and nature of retirement.

2. Linking Health and Retirement

Scholars are only beginning to unravel the health correlates of both the exit from one's 'career' job and the final exit from any paid employment, as well as life after retirement. Three processes link health and retirement: social selection, role loss, and reduction of role strain.

2.1 Social Selection: Health as an Impetus to Retire

Declines in workers' health have been shown to shape their decision-making regarding retirement (Anderson and Burkhauser 1985, Mutchler et al. 1999). The best-laid plans for long-term employment or retirement timing can be destroyed by acute or chronic illness of oneself or one's spouse (or other relatives requiring care), with the onset of poor health promoting unexpected exits from the labor force. Thus, it is crucial that research on the health impacts of retirement appropriately control for preretirement health status.

2.2 Retirement as Role Loss

The second process is that of role loss. Leaving the workforce can have deleterious effects on health, given the corresponding loss of routine, relationships, identity, and purposive activity. Research has documented the importance of role occupancy for both physical and psychological well-being (Moen et al. 2000). Retirement can increase the vulnerability of those in later adulthood, removing them from the mainstream of society.

2.3 Reduction of Role Strain

Role exits such as retirement can also be viewed positively, as when workers leave a demanding or stressful job. In line with this reduction of role strain perspective, those who see their jobs as stressful or unrewarding are likely to leave the labor force early. Jobs that are demanding and provide little sense of control have been shown to be negatively related to health and well-being (e.g., Karasek and Theorell 1990). Moreover, poor health is more likely to encourage those with demanding jobs to retire (Hayward et al. 1989).

3. A Life Course, Role Context Perspective

A life course, role context approach to retirement (Moen et al. 2000, Musick et al. 1999) emphasizes the complexity of this transition. The three processes described above may well be operating at different points in the retirement transition for different subgroups or in particular circumstances. For example, health correlates may be related to the timing of retirement as well as retirees' location in the social structure (gender, race and ethnicity, socioeconomic status), their previous health, and the quality of their prior employment experience. Moreover, experiences and activities after retirement may be key to understanding health and well-being in the retirement years.

3.1 Timing of Retirement

Both policies (governmental and corporate) and cultural norms shape individual expectations and beliefs about the 'right' time to retire (Kohli et al. 1991, Settersten and Mayer 1997). The implications of 'early' or 'late' retirement or of postretirement employment for health and well-being are not clear. For example, research in the United States shows that men retiring early are not necessarily doing so for health reasons (Burkhauser et al. 1996). Some studies suggest that early retirement may have deleterious health consequences in men's lives, with early retirement pro-

ducing more negative effects than later retirement (Palmore et al. 1985), while others have found no change in health based on early retirement (Streib and Schneider 1971). However, in a survey of relatively young retired men in the United Kingdom (McGoldrick 1989) more than half reported improved health as a benefit of retirement, with a quarter seeing it as a major benefit. Health and retirement timing are obviously intertwined; if individuals retire because of health problems, they often feel better off without the physical and psychological demands of their jobs.

3.2 Gender and Linked Lives

Most studies of retirement have focused on men, but women are increasingly retiring from paid work with different experiences and consequences. For example, married women are more likely to retire at an earlier age than single or divorced women and often retire as a consequence of their husbands' retirement exit. A spouse's poor health may cause employed wives to remain in the labor force, possibly for financial reasons (O'Rand et al. 1992). And having to care for ailing relatives may become an impetus for retirement, especially for women (Moen et al. 1994, Pavalko and Smith 1999). The coordination of both spouses' retirements and its implications for each spouse's health remain topics for future research.

3.3 Conditions of Work

The psychological impact of retirement appears to depend both on gender and on previous job stress, with men who leave high-stress jobs experiencing a reduction in distress symptoms, and men who leave low-stress jobs reporting an increase in distress symptoms. There are fewer effects, in either direction, for women undergoing retirement (Wheaton 1990). As Herzog et al. (1991) found among older US workers, being able to work one's preferred number of hours (rather than more or less) is positively related to physical health and life satisfaction, and negatively related to depression, regardless of gender or occupational status. Those who stopped work and felt they had little or no choice reported lower levels of health and well-being than both the voluntarily retired and those working the amount they would like. Another study (Gallo et al. 2000) found that late-life involuntary job loss through downsizing was related to declines in both physical functioning and mental health. While particular conditions of jobs at one point in time have been associated with health and well-being on the one hand, or stress and illness on the other (Karasek and Theorell 1990), a life course, role context approach suggests the importance of considering an individual's work patterns throughout adulthood.

3.4 Employment Patterns, Health, and Mortality

The nature of jobs and career trajectories can have long-term health implications. Studies have shown that men retiring from high-status occupations and 'orderly' career paths experience better health and greater longevity than those in manual occupations or those who have held a series of unrelated jobs (Moore and Hayward 1990, Pavalko et al. 1993).

Most of the studies linking health with career patterns have focused on men. The whole process of retirement may well be a different experience for women than men, in part because of the historical difference in their attachment to the labor force. When men leave their jobs they are exiting from a role that has typically dominated their adult years. Women, on the other hand, commonly experience greater discontinuity, moving in and out of the labor force and in and out of part-time jobs in conjunction with shifting family responsibilities. Consequently, they are less likely to have the same duration of employment or accumulation of work experience as men (Han and Moen 1999), and may be more likely to already have fulfilling roles outside of employment. Whether the findings about the longevity and health benefits of orderly career paths and high-status jobs apply to women as well is an important topic for future investigation. One recent study suggests that employment for women in their 50s and 60s may slow the onset of physical limitations (Pavalko and Smith 1999).

Ethnic and racial minorities, as well as those low in socioeconomic status, are more disadvantaged in terms of their career paths, health, and retirement pensions, creating a cumulation of disadvantage in addition to their increased risk of retiring because of a disability (Shea et al. 1996).

3.5 Life After Retirement

Postretirement employment is a key, yet understudied, aspect of employment history. 'Bridge jobs' following retirement are often very different from the career jobs from which workers retired (Henretta 1992), but they may ease the potentially disruptive transition out of the labor force. It is important to examine both pre- and postretirement employment histories to fully understand the linkages between work, retirement, and health, and to consider various forms of productive engagement following retirement.

Both paid and unpaid work after retirement have been linked to longevity, health, and psychological well-being (Moen et al. 2000). This confirms the protective effects of role participation, but studies have shown that the *context* of that participation matters as well. For example, Musick et al. (1999) found that volunteering had a protective effect on

mortality for those with low levels of informal social interaction.

Another key context shaping life after retirement is marriage. Pienta et al. (2000) have shown the health benefits of marriage for those moving into the retirement years.

4. Retirement Exigencies and Possibilities

Most research shows that being retired has no deleterious effects on either physical or psychological health; most retirees say they are satisfied with their retirement and some even report better health (Atchley 1976, Quick and Moen 1998). The issue is not whether retirement influences health, or vice versa, but the pathways to health and well-being in the post-retirement years. Research evidence makes it clear that good health is an important prerequisite for the enjoyment of the retirement years, as is economic security (Szinovacz and Washo 1992).

Both increasing longevity and the aging of the baby-boom cohort suggest that the retirement passage is in considerable flux, as growing numbers of workers can expect to spend a considerable proportion of their life course retired from their primary career jobs. Moreover, retirees in the twenty-first century tend to be healthier, better educated, and more active compared to those in the middle of the twentieth century.

The retirement transition can offer an occasion for the development of a new identity and can foster the adoption of new roles and new health behaviors. Workers retire, not only because of poor health, but sometimes because of good health, wanting to do something else. And retirees with excellent health are more likely to return to the workforce (Mutchler et al. 1999). But there are few institutionalized opportunities for meaningful engagement after retirement (Riley et al. 1994). Existing structural arrangements, policies, norms, and practices fail to provide the opportunities and challenges that can fully exploit the possibilities of life—and health—after retirement.

See also: Adulthood: Developmental Tasks and Critical Life Events; Aging, Theories of; Ecology of Aging; Employment and Labor, Regulation of; Life Course in History; Life Course: Sociological Aspects; Life-span Development, Theory of; Population Aging: Economic and Social Consequences; Retirement, Economics of; Work, History of; Work, Sociology of

Bibliography

- Anderson K H, Burkhauser R V 1985 The retirement-health nexus: A new measure of an old puzzle. *The Journal of Human Resources* **20**: 315–30
- Atchley R C 1976 *The Sociology of Retirement*. Schenkman, New York

- Burkhauser R V, Couch K A, Phillips J W 1996 Who takes early Social Security benefits: The economic and health characteristics of early beneficiaries. *The Gerontologist* **36**: 789–99
- Delsen L, Reday-Mulvey G (eds.) 1996 *Gradual Retirement in the OECD Countries: Macro and Micro Issues and Policies*. Dartmouth Publishing, Aldershot, UK
- Gallo W T, Bradley E H, Siegel M, Kasl S V 2000 Health effects of involuntary job loss among older workers: Findings from the health and retirement survey. *Journal of Gerontology: Social Sciences* **55B**: S131–40
- Guillemard A M, Rein M 1993 Comparative patterns of retirement: Recent trends in developed societies. *Annual Review of Sociology* **19**: 469–503
- Han S-K, Moen P 1999 Clocking out: Temporal patterning of retirement. *American Journal of Sociology* **105**: 191–236
- Hayward M D, Grady W R, Hardy M A, Sommers D 1989 Occupational influences in retirement, disability, and death. *Demography* **26**: 393–409
- Henretta J C 1992 Uniformity and diversity: Life course institutionalization and late-life work exit. *Sociological Quarterly* **33**: 265–79
- Herzog A R, House J S, Morgan J N 1991 Relation of work and retirement to health and well-being in older age. *Psychology and Aging* **6**: 202–11
- Karasek R A, Theorell T 1990 *Healthy Work: Stress, Productivity, and the Reconstruction of Working Life*. Basic Books, New York
- Kohli M, Rein M, Guillemard A M, van Gunsteren H (eds.) 1991 *Time for Retirement: Comparative Studies of Early Exit from the Labor Force*. Cambridge University Press
- McGoldrick A E 1989 Stress, early retirement and health. In: Markides K S, Cooper C L (eds.) *Aging, Stress and Health*. Wiley, New York
- Moen P, Fields V, Quick H E, Hofmeister H 2000 A life course approach to retirement and social integration. In: Pillemer K, Moen P, Wethington E, Glasgow N (eds.) *Social Integration in the Second Half of Life*. The Johns Hopkins University Press, Baltimore, MA
- Moen P, Robison J, Fields V 1994 Women's work and caregiving roles: A life course approach. *Journal of Gerontology: Social Sciences* **49**: S176–86
- Moore D E, Hayward M D 1990 Occupational careers and mortality of elderly men. *Demography* **27**: 31–53
- Musick M A, Herzog A R, House J S 1999 Volunteering and mortality among older adults: Findings from a national sample. *Journal of Gerontology: Social Sciences* **54B**: S173–80
- Mutchler J E, Burr J A, Massagli M P, Pienta A 1999 Work transitions and health in later life. *Journal of Gerontology: Social Sciences* **54B**: S252–61
- O'Rand A M, Henretta J C, Kreckler M 1992 Family pathways to retirement. In: Szinovacz M, Ekerdt D, Vinick B (eds.) *Families and Retirement*. Sage, Newbury Park, CA
- Palmore E B, Burchett B M, Fillenbaum G G, George L K, Wallman L M 1985 *Retirement Causes and Consequences*. Springer, New York, NY
- Pavalko E K, Elder G H Jr, Clipp E C 1993 Work lives and longevity: Insights from a life course perspective. *Journal of Health and Social Behavior* **34**: 363–80
- Pavalko E K, Smith B 1999 The rhythm of work: Health effects of women's work dynamics. *Social Forces* **77**: 1141–62
- Pienta A M, Hayward M D, Jenkins K R 2000 Health consequences of marriage for the retirement years. *Journal of Family Issues* **21**: 559–86

- Quick H E, Moen P 1998 Gender, employment, and retirement quality: A life-course approach to the differential experiences of men and women. *Journal of Occupational Health Psychology* 3: 44–64
- Riley M W, Kahn R L, Foner A 1994 *Age and Structural Lag: Society's Failure to Provide Meaningful Opportunities in Work, Family, and Leisure*. Wiley, New York, NY
- Settersten R A, Mayer K U 1997 The measurement of age, age structuring, and the life course. *Annual Review of Sociology* 23: 233–61
- Shea D G, Miles T, Hayward M 1996 The health-wealth connection: Racial differences. *The Gerontologist* 36: 342–9
- Streib G F, Schneider C J 1971 *Retirement in American Society*. Cornell University Press, Ithaca, NY
- Szinovacz M, Washo C 1992 Gender differences in exposure to life events and adaptation to retirement. *Journal of Gerontology* 47: S191–6
- Wheaton B 1990 Life transitions, role histories, and mental health. *American Sociological Review* 55: 209–23

P. Moen

Retirement, Economics of

When most people retire, they leave a full-time job and stop working entirely. However, for a substantial minority, retirement is not a simple transition from working full-time to not working at all. It is a complex process. Those who have been working for an extended period for a particular employer may leave that job. They may continue to work for someone else on a job that is less demanding. In some less common circumstances, the long term employer will allow the individual to phase into full retirement, by reducing the demands on the job, or hours of work. Some people may even leave the labor market and then return. This article describes the complex patterns that comprise retirement, and discusses the leading explanations for retirement outcomes.

Given this diverse set of outcomes, various researchers and government agencies define retirement differently. Because of the complexity of the retirement process, each definition of retirement oversimplifies the story. According to one definition, retirement occurs when a person stops work completely. Among the problems with this definition, many who continue to work are already in the process of retiring. Thus, this definition classifies many as not retired, even though they have left a long term job, are working in an easier job than they held for most of their work life, are employed on a part-time basis despite having normally worked full-time, are receiving a lower rate of pay than they were paid before they changed their job or work conditions, even though they may be receiving pension or social security benefits. Another definition pertains to leaving long term employment. This definition classifies a person as retired when the

older individual leaves a job held for 10 or 20 years. A problem with this definition is that it classifies those who have left a long-term job but are still at work as retired people. Not only does this definition classify those who are continuing to work, but are phasing into retirement gradually as retired, but according to this definition, people who are working full time but lost or left a job in their 40s or 50s are retired, even though they are not yet in the process of leaving the labor market. From the perspective of government agencies such as the Social Security Administration, or a firm's pension plan, a person may be considered retired when benefits are accepted. But many people work after having applied for or have received benefits. None of these definitions of retirement is correct or incorrect. Each definition attempts to reduce a complex process to one or another two-dimensional outcome. Consequently, a definition of retirement that is quite useful for one purpose may be not very helpful for another.

1. Data Describing Retirement Outcomes

To understand how retirement varies with age, it is useful to consider data from the Health and Retirement Study. This study is the leading source of data on retirement in the USA. It is also a model for retirement surveys that are now being adopted in other countries. Funded by the National Institute on Aging, every other year since 1992 this survey interviews 12,650 respondents who were born between 1931 and 1941, extensively tracking their retirement and other behavior. In a recent article some of the basic findings on retirement outcomes from the Health and Retirement Study are presented (Gustman and Steinmeier 2000b), and these findings are summarized here. One way to measure retirement is with self-reported status, a measure that incorporates the influence of a number of factors in determining whether an individual is retired, and considers three states: not retired, partially retired, and fully retired. At age 53, about 85 percent of men and 76 percent of women consider themselves to be not retired. Starting at age 54, the fraction not retired begins to decline, falling three or four percent for each year of age until age 61. At age 62 the share not retired falls precipitously, with a fifth of men and a seventh of women leaving non-retirement in that one year. At age 65, there is another large decline, with the fraction of men who are not retired falling 11 percent, and the fraction of women falling by 5 percent, leaving a little under a fifth of men and women reporting they are not retired at age 65. Thus from the perspective of many of the definitions of retirement noted, most people are retired after age 65. At the same time, from age 53 to age 65, the share of men reporting they are partially retired increases from 6 percent to over 22 percent, with the increase for women from 4 percent to 15 percent.

About 8 percent of full-time men and 6 percent of full-time women are still working after having left a long-time job, that is a job they held at age 45 that lasted more than 10 years. Although less than a fifth of the men and women are working full-time after reaching age 65, roughly another fifth classify themselves as partially retired. The percentages at each age of men and women who report they are completely retired are similar, but women are around 15 percent more likely to report that the retirement question is not relevant, presumably because they have not worked outside the home in a number of years.

Retirement status is not static. The number who are not retired in any year is determined in part by how many of those who reported they were not retired in the previous year remain not retired. It is also determined by how many who were not retired in the previous year leave that state for partial or full retirement. Finally, some people may classify themselves as not retired in a particular year, having previously been retired or partially retired. The same is true for the number who are partially retired or fully retired in any given year: those numbers depend on how many continue in the particular state from the previous year, how many leave the state from the previous year, and how many enter that state. Thus to understand the number reporting a given retirement status by age, we need to understand the flows into and out of, various retirement states over time. The Health and Retirement Study (HRS) is a panel data set that resurveys each respondent every other year. It records changes in retirement status over time as respondents age. In the HRS panel, 77 percent of the time respondents continue in the same Retirement State between waves of the survey—retired, partially retired, and not retired. In 17 percent of the cases there is a transition from greater to lesser work, and more difficult to model, 6 percent of transitions are from lesser to greater work. These transitions cumulate over the course of the survey. Altogether, 17 percent of the sample experienced a reversal in the course of the survey, moving from a state of less work to a state of more work.

Also of interest to students of retirement are trends in retirement outcomes. The aging of the baby boomers and decline in mortality have increased the financial pressure on social security and other retirement programs. One solution would be to postpone retirement. If people worked longer, the retirement period would be shorter and less costly. Instead, despite improving health, throughout the post World War II years, and until the end of the twentieth century, there was a continuing trend toward earlier retirement. Indeed, the trend to lower labor force participation by older males is one of the strongest labor market trends witnessed in recent years. (The trends in labor force participation and retirement for older women are not nearly as sharp. They are difficult to discern as the increased par-

ticipation of women in the labor market offsets a tendency to leave earlier.) According to data from the US Current Population Survey cited in Anderson et al. (1999), continuing an earlier trend, from 1947 to 1985, the fraction of married men over 65 in the labor force fell from 55 percent to less than 20 percent. For those aged 55 to 64, the trend to lower labor force participation started later. From 1957 to 1985, the share of married men in the labor force fell from 90 percent to less than 66 percent. Similar trends are found in other countries. Indeed, the decline in the labor force participation rate of males in the US from 1960 is actually relatively modest compared to other countries, with Japan and Sweden experiencing milder trends, while the UK, Germany, Belgium, Spain, the Netherlands, France, and Italy have experienced even greater withdrawal (Gruber and Wise 1999).

There is evidence, though, of a leveling off in the middle 1980s in the trend toward reduced labor force participation of older men. For example, in 1985 in the USA, men who were 55 years old in the Current Population Survey had a labor force participation rate of 79 percent; in 1991 the participation rate of 55-year-olds had increased to 81 percent. Similarly, between 1985 and 1991, the participation rate for 60-year-olds increased by one percentage point, from 66 percent to 67 percent, and for those 65 years old, participation remained the same at 26 percent. Participation rates did not resume their downward trend in the 1990s.

The trend to earlier retirement for men is apparent when retirement outcomes from the Health and Retirement Study, collected from 1992 to 1998 in the USA, are compared with retirement outcomes from the Retirement History Study, collected from 1969 to 1979. Using figures from Gustman and Steinmeier (2000b), between these two surveys, the fraction of 60-year-olds who considered themselves to be not retired fell from 80 percent to 66 percent. The fraction of 62-year-olds not retired fell from almost 66 percent in the 1970s to a little more than 33 percent in the 1990s. Only at age 65 are the fractions not retired similar in the 1970s and 1990s. A very large number of people retired in the 1970s RHS at age 65, with almost 20 percent of the sample leaving nonretirement. In the 1990s HRS, the comparable number leaving at 65 is a little more than half as large—many more have already left. These trends can be seen in the numbers who retire at any given age. In the 1990s HRS sample, 20 percent of the men left nonretirement at age 62. Although there is evidence that the trend to earlier retirement has abated in the past few years, there is a continuing debate about whether the trend will reassert itself.

Researchers have also focused on other aspects of retirement. The family is the center of decision making. Thus there is a question about whether couples coordinate their retirement decisions. It seems they do. Researchers have documented the coincidence of retirement dates for husbands and wives, despite age

differences and differences in incentives to retire (Gustman and Steinmeier 2000a, Hurd 1990).

2. Explaining Retirement Outcomes

Students of retirement behavior would like to explain what determines these patterns of retirement by age. Of special interest is the onset of substantial flows into retirement beginning at age 53 or so, the sharp increases in retirement that are observed at ages 62 and 65, and the transition for most out of non-retirement by age 65. Also of interest is the question of why a majority of people proceeds directly from full-time work to complete retirement. A substantial minority partially retires, but what determines how long a person stays partially retired? Why do most people reduce their work effort as they age while others sometimes increase work effort after having reduced it? Also, why do different people choose different ages to enter and leave full time-work, partial retirement, and full retirement? Similarly, researchers would like to know what has caused the trends to earlier retirement over time. Why has the number of people retiring at age 62 increased while the large spike in the number retiring at age 65 has been halved, and more generally what caused the changes in retirement outcomes at each age? They also would like to understand other key patterns in the data, such as the coincidence of retirement by husbands and wives, and how the new patterns of increased participation by women are affecting retirement outcomes.

To understand what determines the wide variety of retirement patterns, economists rely on a model in which forward looking decision-makers weigh the benefits and costs of alternative retirement dates. Older people decide when to retire by taking into account not only current work and leisure opportunities, but also future opportunities. These opportunities include the pay that would be offered for work on the current job, now and into the future, and the pay on alternative jobs. Involuntary layoffs affect these opportunities, reducing the offer for continuing full-time work. Incentives created by pension plans and social security play important roles, especially in shaping the future consequences of a decision to retire. These programs have features that may change sharply the incentive to work at different ages, first encouraging workers to stay in the labor market, then at a later age, reducing the reward to continue work. Against the monetary rewards to continue work, older individuals weigh the difficulty of the work. Whether these demands are physical or mental, they are likely to increase with age at different rates, depending on the job held. The opportunity cost of working is the value of leisure time. Health problems may increase the difficulty of work, and lower the reward to continued work if the individual becomes less efficient at certain tasks. Retirement decisions of spouses, and

demands from dependent children and elderly parents also help to shape the retirement decision.

It is differences in the rewards to continued work, the effects of social security and their differing pension plans on these rewards, differences in health status, job conditions, as well as different evaluations of the same set of benefits and rewards, that create wide differences in decisions of when to retire. There is also an interaction with the decision to save, and accumulated assets, leaving some more able to afford retirement at a given age than others with the same earnings history. Similarly, changes over time in wages, pensions, social security, and other factors determining the reward structure to continued work, as well as changes in the valuation of leisure, may affect the trends observed in retirement outcomes.

2.1 Key Explanatory Factors Affecting Retirement Outcomes

Consider now in more detail some of the leading factors shaping retirement outcomes and how they influence retirement behavior.

2.1.1 Wage offers for full-time work. Theory does not tell us what relationship to expect between the wage a person makes and retirement age. On the one hand, a higher wage implies greater forgone earnings when one leaves work. On the other, a higher wage implies a higher lifetime income, which should lead the high waged worker to prefer more leisure, just as those with higher incomes prefer more of other luxury goods. To date, the evidence suggests these effects are roughly offsetting, so that higher wages on long term jobs do not have a strong relation to retirement outcomes. There are other possible explanations that might allow the wage to account for an increase in retirement with age; for example, if wage offers declined with age, that might reduce the reward to continued work for older individuals. But they do not. Sometimes researchers record the wage rate received by workers at different ages and find that the observed wage declines with age. This effect is illusory. As workers age, some of the older workers have left their main jobs and are holding lower paying full- or part-time jobs as they transit to retirement. To be sure, some older workers experience layoffs, and, once laid off, older individuals certainly have greater difficulty locating new jobs, but that is not a dominating factor in shaping retirement.

2.1.2 Minimum hours constraints and lower wages when partially retired. There is evidence that many jobs do not permit older workers to phase gradually

into retirement. Rather, at the job they have held for many years minimum hours constraints allow the worker to choose only whether to work full-time, or not at all. Work is available on a part-time basis, but only at much reduced wage. These minimum hours' constraints, together with a lower wage offer when one works part-time, account for the decision by many to move directly from full-time work to full retirement.

Pensions and their provisions may account for some of the diversity in retirement ages. There are two basic types of pensions. One plan, called a defined benefit plan, provides a periodic payment after retirement that is based on a formula. The amount of the benefit is typically determined by years of service, age, and pay. So is the age of eligibility for receiving a full, unreduced, benefit (normal retirement age) and the age of eligibility for a reduced benefit (early retirement age). Defined contribution plans, which at the beginning of the twenty-first century are most often in the form of 401(k) plans, are essentially tax-favored retirement saving plans. Both the covered worker and the employer may contribute to an account, often with the employer matching the worker's contribution according to a set schedule, up to a predetermined amount.

'Defined contribution plans' do not create sharp incentives that affect the reward to remaining on the job, but defined benefit plans do. Those who leave a 'defined benefit plan' before qualifying for early retirement benefits may be penalized heavily. Conversely, an individual who stays until qualifying for early retirement benefits may be heavily rewarded. Studies of workers employed at firms offering defined benefit plans indicate that workers respond to these incentives. Among studies that pioneered in investigating the relation of pension plan incentives to retirement outcomes, see Burkhauser (1979), Fields and Mitchell (1984), and Stock and Wise (1987). According to pension data for respondents to the Health and Retirement Study (Gustman and Steinmeier, 2000c), defined benefit pensions on average had an early retirement age of 54, and an average normal retirement age of 61. Only 6 percent have an early retirement age above 60. Forty percent face a normal retirement age of 65; with 14 percent having normal retirement at age 62; over 20 percent with normal retirement at age 60; and over a fifth having a normal retirement age below 60. A man who has a defined benefit (DB) pension plan and works the year before qualifying for early retirement, on average, will find benefits increased by an amount equal to about 120 percent of one year's of pay. For a woman, the increase amounts to about a third of a year's pay.

In addition to contributing to the explanation for the diversity of retirement dates, changes in pensions account for some of the trend to earlier retirement. For the 1970s and 1980s, it has been estimated that

about 10 percent of the trend to earlier retirement is due to the decline in ages of early retirement in defined benefit pensions and to other changes in pensions (Anderson et al. 1999).

2.1.3 Social security. Analyses of retirement behavior suggest that social security has had a modest effect in influencing retirement, and that effect will decline for those now approaching retirement. Social security benefits are determined from past covered earnings history, where past earnings are indexed to age 60 and are averaged to a summary statistic called the Average Indexed Monthly Earnings (AIME). For those reaching age 62 after 1991, AIME is calculated using the highest 35 years of indexed earnings. If an individual has covered earnings for fewer than 35 years, then zeros are entered into the AIME calculation for the remaining years. From the AIME, the basic benefit, called the primary insurance amount (PIA), is computed. A progressive benefit formula is applied so that those who have low computed lifetime earnings have higher benefits, relative to earnings, than do those with high earnings. The earnings measure is typically expressed as a monthly amount, but when annualized, the formula for the year 2000 specifies benefits that are 90 percent of the first \$6,372 of annual earnings, 32 percent of the next \$32,052, and 15 percent of remaining earnings. Benefits to spouses and survivors affect the relationship between benefits and earnings, both at the level of the individual and at the level of the family. Spouses are entitled to roughly half of their partner's benefits, and survivors are entitled to an amount roughly equal to the benefits that would have been payable to the deceased spouse. The social security benefit is subject to an earnings test for those between the ages of 62 and 65. But any benefits that are lost to the earnings test are restored at age 65. Indeed, for married couples, the additional benefits from waiting until age 65 to claim benefits are worth slightly more than the benefits that were lost.

However, this was not always the case. In the past, the earnings test applied to earnings of respondents up to the age of 72. However, the age at which the earnings test no longer applies has been reduced slowly, until in the twenty-first century it only applies to those between the early and normal retirement age. Moreover, in the past, the adjustment for lost benefits was not nearly as generous as it is today. Initially, the social security benefit structure discouraged work after the age of 65. With recent changes, the program is roughly neutral in its effects on the incentive to retire. Thus social security in the past may have created incentives that accounted for some of the retirements, perhaps both before and after age 65. For example, it is estimated that changes in the benefit formula in the 1970s and 1980s may have increased early retirement by about an eighth. But social security should not have

any effect on retirements after the age of 65 in future years.

International evidence reinforces the view that pensions and social security have important effects on retirement behavior. With the aging of the baby boom, the worry in many countries is that pensions and social security discourage still productive individuals who would otherwise work to leave the labor force.

Different countries have very different pension and social security systems than the USA. Differences in design among countries create incentives to retire at different ages. The different retirement incentives created by the diverse retirement systems among countries provide another test of the influences of social security and pensions in shaping retirement behavior. Gruber and Wise (1999) have brought together researchers from a number of countries, both to describe retirement outcomes, and to see how their pension plans and social security systems affect retirement behavior. They find a close correspondence between the early retirement dates and other features of the programs, and the ages at which older workers tend to leave the labor force. Moreover, they find that when these dates of early retirement and normal retirement specified in the social security systems of these countries change, retirement behavior changes correspondingly. Compared to many other countries, the USA has a low relative tax rate on work after the early retirement age specified in the social security system, which is one reason for the higher labor force participation of older men in the USA. Countries with very high tax rates, which include Belgium, Italy, France, and The Netherlands, have lower participation rates by older workers than the USA.

Many researchers believe that social security has substituted for other forms of retirement saving. Many people enter retirement with no assets, and social security benefits as their sole support. Low-income families may have half or more of their earnings replaced by social security. Those who rely entirely on social security cannot retire before the age of 62. They do not have savings to draw on to support their consumption until they became eligible for social security benefits at 62. As a result, if the social security early retirement age were raised from 62 to 65, there would be a large increase in the number delaying their retirement.

2.1.4 Health status. Health changes are perhaps the most important of all the uncertain events affecting older workers' labor force participation. Data on the course of HRS respondents' health events from 1992–8 suggest a wide variety of health problems among workers, including limitations in activity, difficulties in carrying out mental and physical tasks, and medical problems such as high blood pressure. Both the level of health status and changes in health

status are associated with different patterns of labor force participation in the HRS (Bound et al. 1998).

2.1.5 Retirement decisions in the family. Findings at the end of the twentieth suggested that many spouses retire together, even though wives typically are younger than their husbands. Available estimates suggest that each spouse, and husbands in particular, values retirement more once their spouse has retired (Gustman and Steinmeier 2000a). It does not appear that husbands and wives time their retirements to coincide simply because of their pensions or social security. Other family related considerations might help to shape retirement. Some families married or remarried late, have young children, had a spouse who never returned to the labor market once having had children, or faced great demands on time to care for older parents, or a spouse. These and other differences among families may help to account for differences in retirement.

2.1.6 Other factors affecting the retirement decision. There is also evidence supporting the influence on retirement of many other factors that differ among individuals and their jobs. Labor market demands for older workers have been investigated. It is clear from this work that many firms are reluctant to hire older workers, because they will not work over a long enough period for their hiring and training investments to be repaid, together with other reasons (Hutchens 1986). The business cycle also influences the demand for older workers. On the one hand, in times of tight labor markets, firms will have difficulty replacing older workers and will encourage them to postpone retirement. On the other hand, in times of slack demand, firms adopt policies such as one-time windows that are designed to encourage earlier retirement. Those who invested more in training for the labor market by staying in school longer may work longer to justify their investments. Unions may shape retirement conditions both to accommodate the preferences of their older workers, and to share work with younger union members. Government policies might have had profound effects, but they have been pushing against the trend. Thus the government since 1980 has outlawed mandatory retirement, forced firms to make pensions more actuarially fair so as not to penalize work after the normal retirement age, and eliminated the penalty to work after age 65 under social security. Perhaps these policies have had some small effect in mitigating the trend to earlier retirement in the 1990s. However, there is no firm evidence to support that view.

2.1.7 Imperfect information and nonmaximizing behavior. Of course, the choices we see are deter-

mined by more than the maximizing choices made by people who fully understand all their opportunities. Many people do not understand fully the incentives from their retirement programs. Some plan well, some don't plan at all. In addition, some that are well aware of the importance of preparing for retirement do not have the discipline to follow through. Most people are well prepared for retirement. But if they have not prepared, they will either have to consume less when retired, or to work longer. Economic models are now being expanded to understand better the behavior of those who may not understand fully all the incentives, or for other reasons may not be choosing their retirement date according to a model of strict maximization. Certainly this work will modify our perception of how each of the factors discussed influence retirement choices. Nevertheless, the modifications are expected to be modest, and the major influences on retirement are the forces discussed.

3. Conclusions

Most of the information we have on how pensions, social security, health status, and other factors affect retirement outcomes comes from analyses that were based on 1970s data (e.g., see Gustman and Steinmeier 1986, Rust and Phalen 1997). The data from the new Health and Retirement Study are being used to estimate comprehensive models of retirement. With these newly available data, researchers are in a position to quantify how the factors noted in Sect. 2 influence retirement outcomes for those who are just a little older than the baby boomers. The wave of research now in process should also leave us in a good position to understand what will determine the retirement behavior of the baby boomers, and how changes in government social security and pension policies will affect their retirement outcomes.

See also: Age: Anthropological Aspects; Age Policy; Age, Sociology of; Age Stratification; Age Structure; Ecology of Aging; Leisure and Cultural Consumption;

Life Course in History; Retirement and Health; Sex Differences in Pay

Bibliography

- Anderson A L, Gustman T, Steinmeier L 1999 Trends in male labor force participation and retirement: Some evidence on the role of pensions and Social Security in the 1970s and 1980s. *Journal of Labor Economics* **17**(4) p 1: 757–83
- Bound J, Schoenbaum M, Stinebrickner T R, Waidman T 1998 The dynamic effects of health on the labor force transition of older workers. *NBER Working Paper* **6777**
- Brown C 2000 Early retirement windows. In: Olivia S, Mitchell P, Hammond B, Rappaport A M (eds.) *Forecasting Retirement Needs and Retirement Wealth*. University of Pennsylvania Press for The Pension Research Council, PA, pp. 253–73
- Burkhauser R V 1979 The pension acceptance decision of older workers. *Journal of Human Resources* **14**(1): 63–75
- Fields G S, Mitchell O S 1984 *Retirement, Pensions and Social Security*. MIT Press, Cambridge, MA
- Gruber J, Wise D A (eds.) 1999 *Social Security and Retirement Around the World*. University of Chicago Press, Chicago
- Gustman A L, Steinmeier T L 1986 A structural retirement model. *Econometrica* **54**(3): 555–84
- Gustman A L, Steinmeier T L 2000a Retirement in dual career families. *Journal of Labor Economics* **18**(3): 503–45
- Gustman A L, Steinmeier T L 2000b Retirement outcomes in the health and retirement study. *Social Security Bulletin* **64**(4): 3–17
- Gustman A L, Steinmeier T L 2000c Employer provided pension data in the NLS mature women's survey and in the health and retirement study. *Research in Labor Economics* **19**: 215–52
- Hurd M D 1990 The joint retirement decision of husbands and wives. In: Wise D A (ed.) *Issues in the Economics of Aging*. University of Chicago Press, Chicago, pp. 231–54
- Hutchens R 1986 Delayed payment contracts and a firm's propensity to hire older workers. *Journal of Labor Economics* **4**(4): 439–57
- Rust J, Phalen C 1997 How social security and medicare affect retirement in a world of incomplete markets. *Econometrica* **65**: 781–832
- Stock J H, Wise D A 1987 Pensions, the option value of work and retirement. *Econometrica* **58**(5): 1151–80

A. L. Gustman and T. L. Steinmeier

Copyright © 2001 Elsevier Science Ltd.
All rights reserved.

Rev

Reviews, Critics, and Cultural Gatekeepers

It is argued here that the expertise of cultural gatekeepers does not pertain to the identification of those intrinsic properties that make cultural products into works of art. The expertise of cultural gatekeepers rather consists in following procedures that enable them to classify cultural products according to genre and to rank them according to quality. The use of these group-bound procedures enhances the chance that proposed classifications and evaluations will meet agreement from other gatekeepers. Empirical research has clarified the nature and effects of procedures used by gatekeepers, in particular by literary critics. These procedures seem to be endogenous to the institution of criticism, thereby suggesting its autonomy from other realms of society. However, when other cultural gatekeepers are taken into account, e.g., publishers, and when questions are posed about the content and value they attribute to cultural products, the procedures involved in this process manifest a complex interplay between a specific cultural sector and other fields of society. The autonomy of the cultural sector from other social sectors is relative in nature. All labels attached to cultural products, referring to their content, genre or value, are multi-dimensional in nature. What is to be understood by the value various parties attach to cultural products calls for cooperation between sociologists and economists: the former can profit from their insights into the nature of the procedures followed by cultural gatekeepers; the latter have their sophisticated analyses of value ('utility') to offer.

1. The Traditional View of Cultural Gatekeepers' Expertise

Cultural gatekeepers—reviewers, publishers, gallery owners, film producers—occupy institutional positions between cultural producers and consumers. They determine which products cross the boundary between the private and the public domains, and when this boundary is crossed, they propose classifications and evaluations of the products that are included in cultural repertoires. Gatekeepers are credited with

having professional expertise on what makes cultural products into works of art.

Giving or denying access to the public sphere is an eminently social task, yet the social sciences have been slow in developing a general perspective on what gatekeepers do, the conditions under which they operate, their practices, and their impact on the reputations of cultural products and their makers. One of the reasons why cultural gatekeepers have remained so long outside the scope of the social sciences is that, for centuries, cultural expertise has been believed to pertain to the *intrinsic* properties on the basis of which an individual cultural product may count as a work of art. This is an instance of 'charismatic ideology', to use a term coined by Bourdieu and Darbel (1969, p. 161). According to this ideology, works of art have the power to convince any rational being of their particular nature (i.e., that of being a work of art), and of their quality. If works of art can induce by themselves the recognition of their nature and quality, they must have specific characteristics which are lacking in other cultural products. The expertise of gatekeepers is believed to consist essentially in the ability to identify those intrinsic characteristics that induce the judgment that we are dealing with works of art.

However, as far as cultural gatekeepers are concerned, it is extremely hard to identify sets of general and firmly established procedures for interpreting or judging a poem, a movie, or a painting; similarly, there seem to be no clear-cut procedures by which works of art are classified as having a 'style' in common, or as differing from one another in this respect. In light of the charismatic ideology, the question of what procedures cultural gatekeepers use seems to lose its urgency: If one believes that works of art have persuasive powers, and if one believes that each work of art is unique, then it seems useless to suppose that the successful distinction between cultural products that belong to the respective domains of 'art' and 'non-art' requires explicit, let alone general, procedures.

In the nineteenth and twentieth centuries, the study of the arts developed into specialized academic disciplines. As a result, the question of which procedures should be observed in order to arrive at sound conclusions became preeminent. Literature offers a key example, for among the human sciences, the study of literature has produced the largest number of approaches to what counts as its central object of research, the literary text. Russian Formalism, Czech

Structuralism, New Criticism, Jakobson's linguistic poetics, reception aesthetics, and narratology are cases in point, not to mention Marxist, psychoanalytic, feminist, and deconstructivist studies of literature—all of which encompass a considerable number of widely diverging approaches. The privileged position of literature in secondary education has facilitated the proliferation of valutive procedures. Young students are initiated into how to 'appreciate' a nation's great literary works, and to a much lesser extent into appreciation for a nation's great movies, paintings, or musical compositions. The assessment of the value of the canonical works is constantly adapted to the needs of new generations.

In the recent past, two parallel, but mutually independent, developments have undercut the idea that cultural gatekeepers' expertise essentially pertains to those intrinsic properties that endow cultural products with the specific nature and quality of works of art. First, researchers have come to the view that value judgments are not statements which can be shown to be true or false. Moreover, they cannot be derived logically from descriptive or analytical statements to which the true/false distinction applies. Instead, any logical derivation of an evaluative statement from a descriptive one involves an appeal to 'institutional facts', i.e., to constitutive rules embodying norms and obligations to which one is committed (cf. Searle 1970, p. 175ff.). This has contributed to adopting a relativistic attitude towards values, in that they were seen as beliefs or views that are endorsed (or not) by specific social groups. A sociological approach to norms and values, and, in particular, to the value-laden distinction between 'art' and 'non-art' became urgent. It soon became a viable option because Bourdieu's work offered a framework within which the adjudication of 'symbolic value' to products, i.e., value deemed to be specifically cultural in nature, could be assessed empirically.

The second development, which made Bourdieu's views a viable alternative for research done in the humanities, was the extensive discussion of the methodological status of interpretations of literary texts and other forms of cultural expression. Again taking literature as the exemplary case, the outcome of this discussion showed literary criticism, as practiced in newspapers, in literary magazines, and at universities, to be linked tightly to normative conceptions of literature, i.e., to mostly implicit ideas about the characteristics a text should have in order to count as a form of literature (Verdaasdonk 1979, Van Rees 1986). In consequence of this, critics' statements about literary texts cannot be regarded as assessments of intrinsic properties these texts possess. Their statements show an inextricable mixture of classificatory and valutive aims (cf. Verdaasdonk 1994, 2001). It should be emphasized that not only literary criticism, but all studies of art and culture current in the humanities suffer from a similar lack of clear and

reliable procedures for identifying properties that are supposed to make cultural products into works of art.

2. *Empirical Research on Gatekeepers' Adjudication of Value to Cultural Products*

Bourdieu (1980) offers a comprehensive view of what critics, reviewers and cultural gatekeepers do in dealing with cultural products. The questions Bourdieu (1980) addresses are: What constitutes the authority of cultural gatekeepers? What factors affect the consensus they attain on the nature and value of cultural products? How do gatekeepers interact with one another in this process? Under what conditions do gatekeepers have an impact on audiences of art and culture?

In deciding that a product counts as a work of art, gatekeepers assign 'symbolic value' to it, thereby setting it apart from the set of products to which they refuse to give artistic status. Evidently, gatekeepers act as representatives of cultural institutions, such as criticism, publishing houses, movie studios, theaters, concert halls, etc. This largely accounts for the legitimacy of their decisions: as representatives of specific institutions, gatekeepers are seen—and see themselves—as being endowed with expertise that is socially acknowledged. However, Bourdieu stresses that this form of legitimacy is not sufficient in itself: gatekeepers may criticize decisions made by their counterparts; and critics and reviewers often have sharply contrasting opinions about the nature and value of particular cultural products. Gatekeepers' authority adds strongly to the legitimacy their decisions have in the eyes of other gatekeepers. Bourdieu holds that gatekeepers' authority essentially is the credit they have earned by making previous successful decisions—decisions that were adopted by their fellow gatekeepers. Authority is a reputation for success: it is *believed* that authoritative gatekeepers will be as successful in the future as they were in the past in discovering and promoting artists and their work. In this view, authority is not thought to be based on a successful determination of those intrinsic properties that account for the artistic status of cultural products; authority is rather conceptualized as the impact of a gatekeeper's stances on those of others.

Consensus formation among reviewers and critics about the value of cultural products, and the interaction these gatekeepers engage in, have been taken up by researchers. So far, the nature of gatekeepers' expertise and their relations with audiences have received much less attention.

Many researchers have taken a closer look at the forms of institutionally determined behavior that gatekeepers manifest in attuning their selections to those of their counterparts. It has been found that they strongly agree on which products have to be regarded as being of superior quality; and, conversely, they

attain much less agreement about how to classify and to characterize individual products. New literary releases that in the majority of newspapers were hailed as 'important' stand a much better chance of becoming, after a certain period of time, objects of scholarly attention (Verdaasdonk 1983). This suggests that, in the course of time, the selections made by a first group of gatekeepers are taken over by other groups of gatekeepers—with a certain amount of downsizing of the choices that were made previously. Van Rees (1983) has argued that reviewers in newspapers, critics in literary magazines, and academic scholars act as three successive filters through which, gradually, a repertoire of 'important' literary works is established. Ultimately, this repertoire enters the literature curriculum of secondary schools. The universities seem to follow the selections made by the scientifically less prestigious institutions of newspapers and literary magazines. Such research suggests that consensus formation among reviewers and critics is a process that is strongly reproductive in nature: Previous selections shape subsequent choices.

Do gatekeepers, in composing repertoires of noteworthy products, characterize them in the same way? This seems doubtful. Critics classify books, movies, paintings, etc. as having a certain 'style' or 'content' in common. De Nooy (1993) found that such classifications do not cross the borders of one single institution. The classifications which, in the 1970s, reviewers and critics proposed of the recent Dutch literary production appeared to be unknown to other gatekeepers, such as booksellers, librarians, and teachers of literature. De Nooy performed network analyses on these classifications and came to the conclusion that the groupings of authors were based on the specific magazines and publishing houses that featured these authors—although the critics held firmly that stylistic and thematic analogies, and differences between the texts concerned, provided the basis for their clustering.

The 'production of culture' perspective, advocated by Peterson (1976, 1985), holds that 'the nature and content of symbolic products [...] are significantly shaped by the social, legal and economic milieu in which they are created, marketed, purchased and evaluated' (Peterson 1985, p. 46). This view is fully congenial to Bourdieu's approach. However, posing questions about gatekeepers' behavior other than those addressed by Bourdieu, leads to different assessments of the interaction between factors which are thought to be 'endogenic' or 'exogenous' to specific cultural institutions, and to a fresh view on the nature of such factors. In their study of the American publishing industry, Coser et al. (1982, p. 146) found that, in judging book proposals or manuscripts, there were systematic differences between editors according to the type of publishing house where they were employed: Fewer than half of the university press editors or commercial scholarly editors rated profit-

ability or commercial prospects as the most important factor; in contrast, more than 60 percent of both textbook and trade editors listed commercial prospects as the single most important consideration. Trade and college editors are more concerned with first-year profitability, while scholarly editors pay more attention to long-term commercial success. Here, organizational differences affect the weight and the nature of judgments of quality and profitability, and of their interrelations. For these gatekeepers, 'denials of the economy', to use Bourdieu's term, are less possible than for gatekeepers deciding on a product's artistic status.

When questions about content are posed—the content of cultural products or of what gatekeepers say about them—the nature of the factors that seem 'endogenous' to the cultural sector and of those that seem 'exogenous' to it will shift. Griswold (1981) notes that after the USA recognized international copyright agreements in 1891, the themes of trade novels showed a strong convergence. American publishers did not concentrate any more on English novels about love and domestic life that could be obtained free of rights; American authors did not compete any more with these novels, by treating other themes that could also appeal to a large audience. Economic viability appears to be a major factor shaping the fare trade publishers brought out at the beginning of the twentieth century.

The cultural sector offers a huge and rapidly changing supply of products. Audiences also change—in size and composition. In many Western countries, an increase of participation in higher education has led to a growth and a refined segmentation of the audiences of art and culture. Of course, this has an impact on the behavior of gatekeepers. Cameron (1995) observes that classical music is marketed now by record companies as a mass consumption good. A number of marketing ploys has been used to reduce the entry barriers for consumers. One of these ploys is to provide excerpts of classical works. A free CD featuring these excerpts is distributed together with a magazine which reviews the CD in nontechnical language. This represents a change in the function of criticism and in its major task: The adjudication of symbolic value to cultural products. Cameron (1995, p. 328) says: 'The aesthetic dimension remains in the product but consumption is increased despite it.' In this example, the awareness of the market place takes a form that is very different from that envisioned by Bourdieu.

3. *Perspectives on Future Research*

As the above discussion shows, considerable progress has been made in pinpointing factors affecting the behavior of cultural gatekeepers. However, most models are 'static' in that they do not account for the impact of time on interactions between dependent and

independent variables. Gatekeepers' activities encompass a considerable period of time; their decisions are conditional on previous ones. Audiences change, as do the factors—and their interactions—that have impact on the production of cultural objects, and on their classification and evaluation. Many commentators, e.g., Peterson and Simkus (1982) and Griswold (1998) have pointed out that the traditional boundaries between 'high' and 'popular' culture undergo rapid and fundamental changes. This entails a focus on the variations in cultural products, their appraisal, and their consumption manifest over time. Griswold (1998, p. 34) speaks of a 'cultural swirl of mutual influence with multiple flows, no impermeable boundaries and few fundamentalists.' If cultural consumers are to be seen as 'occupying different positions in a differentiated distribution of knowledge and cultural resources' (Griswold 1998) this calls for a 'dynamic' approach to analyzing the behavior of gatekeepers, distributors, reviewers, and producers serving these audiences. As yet, longitudinal, time series, and event history data are rarely used in sociological research on culture. However, such data may be essential to capture the time-varying impact some independent variables may have on dependent variables.

Although many studies address the *modus operandi* of cultural gatekeepers, there has been scant attention paid to what constitutes their expertise. It has been argued here that gatekeepers are unable to identify those intrinsic properties that are deemed to account for the artistic nature and value of a cultural product. When it is acknowledged that meaning is not holistic or fixed, but is whatever works for a specific group (cf. Griswold 1998, p. 35), then the focus should be on the procedures gatekeepers actually follow in reaching agreement on the nature and value of cultural products. In the final chapter of *Distinction*, Bourdieu distinguishes between the practical mastery of classifications and the reflexive mastery that is required in order to construct a taxonomy that is coherent and adequate to social reality (1984, p. 472). This idea may also be useful to analyze the ways in which gatekeepers come up with taxonomies and value judgments. If one admits that cultural expertise does not pertain predominantly to knowledge of the intrinsic characteristics of cultural products, gatekeepers may be seen mainly as experts in executing specific tasks with regard to cultural products. Such tasks may be: Presenting a product to connoisseurs or laymen, elaborating on its meaning or value by making comparisons with other products, assessing a product's eligibility for a subsidy or an award, etc. The use gatekeepers make of their experience with individual products may vary with such tasks; the procedures they follow in order to arrive at results that are condoned by other gatekeepers may also vary, and may lead to different groupings and evaluations of the same products.

Future research on cultural gatekeeping is likely to address the still considerable gap between economic and sociological analyses. Frey and Pommerehne (1989, p. 12f), two founding fathers of economic research on culture, remark that sociologists tend to focus on the demand side of culture, whereas economists take the interplay between supply and demand into account. Most contributions to the economics of culture follow the neoclassical approach in asking questions about the determinants of prices of works of art, the rationality of multiple art markets (Singer and Lynch 1997), or the profit-maximizing behavior of consumers of culture. In these studies, a strong similarity is assumed to exist between the symbolic and economic value of cultural products (an assumption that sociologists, as discussed above, have been at pains to attack). The finding, for instance, that successful blockbuster movies feature megastars who account for the large audience these movies attract comes very close to making the assertion that movies with a huge production budget—megastars get gigantic fees—are more likely to enter the top 20 than movies made with smaller sums of money (cf. Albert 1998).

Economics and econometrics are well equipped to analyze many types of value assignments—manifested by preferences and by choices. Gatekeepers engage in choice behavior and develop preferences, but these processes have not received much attention from sociologists. Not only the process of value attribution, but also the question of how value develops over time—under what conditions does it accumulate or diminish?—is important. Combining economic and sociological techniques, it should be possible to determine how gatekeepers and 'ordinary' consumers perceive differences between individual cultural products, and how they go about assigning increasing, decreasing, or stable values to them.

See also: Cultural Critique: Anthropological; Cultural Policy: Outsider Art; Hegemony: Cultural

Bibliography

- Albert S 1998 Movie stars and the distribution of financially successful films in the motion picture industry. *Journal of Cultural Economics* 22: 249–70
- Bourdieu P 1980 The production of belief: Contribution to an economy of symbolic goods. *Media, Culture and Society* 2: 261–93
- Bourdieu P 1984 *Distinction: A Social Critique of the Judgement of Taste*. Harvard University Press, Cambridge, MA]
- Bourdieu P, Darbel A 1969 *L'amour de l'art. Les musées d'art européens et leur public*. Minuit, Paris
- Cameron S 1995 On the role of critics in the culture industry. *Journal of Cultural Economics* 19: 321–31
- Coser R, Kadushin C, Powell W 1982 *Books: The Culture and Commerce of Publishing*. Basic Books, New York

- De Nooy W 1993 Richtingen en lichten. Literaire classificaties, netwerken, instituties [Movements and currents. Literary classifications, networks, institutions]. Unpublished PhD thesis, Tilburg University, The Netherlands
- Frey B, Pommerehne W 1989 *Muses and Markets. Explorations in the Economics of the Arts*. Blackwell, Oxford, UK
- Griswold W 1981 American character and the American novel. *American Journal of Sociology* **86**: 740–65
- Griswold W 1998 The burnished steel watch. What a sample of a single year's fiction indicates. In: Janssen S, Van Dijk N (eds.) *The Empirical Study of Literature and the Media*. Bajesteh Van Waalwijk Van Doorn, Rotterdam, The Netherlands
- Peterson R 1976 The production of culture: Prolegomenon. *American Behavioral Scientist* **19**: 669–84
- Peterson R 1985 Six constraints on the production of literary works. *Poetics* **14**: 45–67
- Peterson R, Simkus A 1992 How musical tastes mark occupational status groups. In: Lamont M, Fournier M (eds.) *Cultivating Differences. Symbolic Boundaries and the Making of Inequality*. The University of Chicago Press, Chicago, IL
- Searle J 1970 *Speech Acts. An Essay on the Philosophy of Language*. Cambridge University Press, Cambridge, UK
- Singer L, Lynch G 1997 Are multiple art markets rational? *Journal of Cultural Economics* **21**: 197–218
- Van Rees C J 1983 How a literary work becomes a masterpiece. On the threefold selection practised by literary critics. *Poetics* **12**: 397–418
- Van Rees C J 1986 Literary theory and criticism. Conceptions of literature and their application. Unpublished PhD thesis, University of Groningen, The Netherlands
- Verdaasdonk H 1979 Critique littéraire et argumentation. Unpublished PhD thesis, Free University of Amsterdam, The Netherlands
- Verdaasdonk H 1983 Social and economic factors in the attribution of literary quality. *Poetics* **12**: 383–95
- Verdaasdonk H 1994 Analogies as tools for classifying and appraising literary texts. *Poetics* **22**: 373–88
- Verdaasdonk H 2001 Quotations expressing the valuative stands of literary reviewers. In: Schram D, Steen G (eds.) *The Psychology and Sociology of Literature*. Benjamins, Amsterdam, The Netherlands

H. Verdaasdonk

Revolution

The term 'revolution' has evolved from a general notion of political 'turning-over' to refer more specifically to a rapid, fundamental transformation of the state and social structure, accompanied by mass uprising from below. Theories of revolution that emphasize structural conditions tend to focus on the revolutionary seizure of state power, while theories emphasizing actor motivations view revolution as an open-ended process of social transformation. This article examines the definitions and alternative explanatory frameworks, and considers the future of revolution in a changing global context.

1. Conceptualizing 'Revolution'

The earliest studies of comparative politics noted recurring conflict between elites and populace, sometimes leading to power 'revolving' from one group of rulers to another within a society. After the French Revolution of 1789, the term 'revolution' took on new meaning as a radical overthrow of the social and political order itself (Goldstone 1998). The nineteenth century European challenges to monarchic authority by the rising bourgeoisie and the class conflicts of industrial capitalism, punctuated by Marx and Engels' [1848] (1998) publication of *The Communist Manifesto*, focused attention on historical processes in analyzing revolution. After Lenin and Mao adapted Marxist theory to the Russian and Chinese revolutions (see *Marxism/Leninism; Maoism*), the decades following World War II saw a wave of national liberation movements against colonialism and neo-colonialism (see *Third World; Colonialism: Political Aspects*; and *Imperialism: Political Aspects*). Contrary to Marx's expectation that revolution would occur in the most industrialized capitalist countries, led by the proletariat, revolutions in the twentieth century occurred in the global periphery through mobilization of peasants and artisan workers.

1.1 Common Usage

The eclectic 'socialism' of many Third World revolts and nationalist movements, and their use of unorthodox tactics of guerrilla warfare—analyzed by practitioners such as Mao Zedong, Che Guevara, and Vo Nguyen Giap—led those characteristics to be generically associated with 'revolution,' as in Régis Debray's (1967) *Revolution in the Revolution?*. This usage was reinforced by Cold War counterinsurgency strategies aimed at forestalling radical change in national political systems, which were seen as extensions of the global East–West contest. The term 'revolution' moved away from this historically specific usage after the conservative religious turn in the 1979 Iranian revolution, the Eastern European uprisings against Communist Party states after 1989, and the end of the Cold War and other global power shifts.

Sometimes the term 'revolution' is used loosely to refer to any kind of paradigmatic shift, as in 'the Industrial Revolution,' 'the Reagan revolution,' 'the behaviorist revolution' in social sciences, or Thomas Kuhn's *The Structure of Scientific Revolutions* (1997). In another variant usage, 'revolution' is contrasted with reformism (see *Reform: Political*) to connote a radical political program, short time horizon, and willingness to use extreme (violent) means; as in Malcolm X's call for African-American liberation 'by any means necessary,' or J F Kennedy's warning that 'those who make peaceful revolution impossible

will make violent revolution inevitable' in Latin America.

1.2 Political Revolution and Social Revolution

The term 'political revolution' is employed to specify the application of the concept to a national political system. This usage does not include issue-specific social movements, nor all sudden changes of political power through *coup d'état* or civil war (see *Internal Warfare: Civil War, Insurgency, and Regional Conflict; Political Protest and Civil Disobedience*). However, 'political revolution' can refer to a wide range of revolts, rebellions, and mobilizational seizures of power—including those led from above such as the Japanese Meiji Restoration of 1868, Nasser's 1952 coup in Egypt, or the 1944–54 period in Guatemala—if they propose major lasting changes in the political order (Goldstone 1998).

Skocpol's influential work (1979, p. 4) proposed a more restrictive focus on 'social revolution':

Social revolutions are rapid, basic transformations of a society's state and class structures, and they are accompanied and in part carried through by class-based revolts from below. Social revolutions are set apart from other sorts of conflicts and transformative processes above all by the combination of two coincidences: the coincidence of societal structural change with class upheaval, and the coincidence of political with social transformation.

Even accepting Wickham-Crowley's (1997) substitution of 'mass' for 'class-based' to include ethnic and populist upheavals, this defines a relatively rare phenomenon. This definition is widely used by scholars of revolution.

2. Alternative Ways of Understanding Revolution

Goldstone (1980) usefully grouped explanations of revolution into three generations of theory, with a fourth generation emerging since the late twentieth century. In the 1920s and 1930s, a first generation of 'natural histories' of revolutions used techniques of 'thick description' and inductive logic based on empirical observation of major Western revolutions. These efforts to identify key actors, factions, sequences, and patterns helped establish the phenomenon of revolution as a distinctive subject of comparative study; for example, Brinton's (1938) classic, *The Anatomy of Revolution*.

2.1 From Description to Explanation

In the 1950s and 1960s, the upheaval associated with decolonization led to a second generation of 'general theories of political violence.' Some focused on the

individual psychological motivations of people experiencing frustrated expectations, as in the 'relative deprivation' approach later refined in T R Gurr's *Why Men Rebel* (1970). Frantz Fanon's *The Wretched of the Earth* (1965) interpreted political violence as redemption, healing the psychological wounds of colonization. Another branch of general theories sought explanations of political violence at the societal rather than individual level. Huntington's (1968) influential work drew on modernization theory, attributing revolutionary upheaval to economic modernization in developing countries, which raised expectations faster than they could be accommodated by the lagging development of political institutions (see *Modernization, Political: Development of the Concept*).

Critics of modernization theory in the late 1960s and 1970s pushed general theories of political violence in new directions. Tilly (1978) challenged the assumption that social systems were in natural equilibrium, arguing that conflict was normal, so explanations of political violence should focus on the relative resource mobilization capacity of power contenders. Moore (1966), in a sweeping macrohistorical comparative work, showed that there is no single, unilinear process of modernization. Extending Marx's class analysis to a more detailed classification of landlord–peasant relations, he combined this variable with distinct political forms of class struggle in agrarian societies undergoing commercial modernization, to map out different possible revolutionary outcomes.

Moore's (1966) work was the point of departure for Skocpol's (1979) ground-breaking study most associated with a third generation of 'structural theories' of revolution (see also her critique of Moore in Skocpol 1994). Shifting from the actor-centered focus that characterized general theories of violence, these theories highlighted the structural relations between units that created revolutionary opportunities. They built multicausal explanations based not only on class dynamics, but also on other aspects of the structures of society, the state, and the international system. Structural theories identified revolutionary conditions in the combination of administrative breakdown of the state, dissident elite political movements, a permissive world context, and widespread rural rebellion. Some like Skocpol and Goodwin 1994 focused on the types of state structures most vulnerable to revolutionary overthrow, including politically exclusionary authoritarian regimes (see *Authoritarianism*), neopatrimonial dictatorships, and directly ruled colonies. Goldstone (1991) examined the interaction of factors such as population growth with political institutions and social structure, developing quantitative indicators of revolutionary conditions. Structuralists defined revolution in terms of successful outcomes, distinguishing revolutionary seizure of state power from cases in which one or more necessary conditions were absent (Wickham-Crowley 1992).

2.2 Structure vs. Agency: Emerging Fourth-Generation Theory

An alternative to defining revolution as a set of structural conditions is to conceptualize it in terms of distinct systems of ideas and beliefs that subjectively mobilize actors. An emerging fourth generation of 'actor-centered' theories of revolution emphasize the defining role of ideology, culture, and identity of the subjects who make revolution. To borrow Mao's metaphorical observation that a single spark can ignite a prairie fire, these theorists shifted attention from the wind, temperature, and dryness of the grasslands to the spark itself. This contrasts with Skocpol's (1979) quotation of Wendell Phillips's statement that 'revolutions are not made; they come.'

Fourth-generation theories inspired new debates about whether revolutions can be accurately predicted on the basis of structural conditions (Keddie 1995), particularly in light of the surprising developments in the Iranian and Eastern European revolutions of the late twentieth century. Structuralists began re-examining the role of ideology in revolutions and the interaction of states with other domestic and international structures (Skocpol 1994). New insights from cultural studies and feminist theory challenged the more state-centered concepts of revolution (Foran 1997). These new perspectives defined revolution not as an outcome achieved by overthrow of the state, but as an ongoing process of transformation of society, involving both the institutionalization of a new order and the consolidation of a new ideology, guided with different degrees of success by different styles of leadership (Selbin 1999). As the Central American revolutions moved from armed conflict to a new dynamic of mobilization and negotiation, the idea of a revolutionary outcome or endpoint seemed less useful than the concept of revolution as process.

2.3 Rethinking Revolution in the Context of Globalization

The emergence of late twentieth century movements that both resembled and differed from earlier waves of 'revolution' suggested a need to rethink the concept. Some interpreted uprisings such as the Philippine 'people's power' movement and the Czech 'velvet revolution' as part of a new, democratic revolutionary wave. Others drew on world-systems theory to suggest that revolutions should be defined not as discrete phenomena of nation-states, but rather as transformations in the world system (Boswell and Chase-Dunn 2000). Thus the 'world revolution' of 1989 might be seen as an outgrowth of the shift in the capitalist world system, from the Fordist regime of capital accumulation based on industrial mass production to a regime of flexible accumulation, now regulated by global institutions governed by neoliberal ideology.

When Nicaragua's revolutionaries made history by turning over the state power they had won through insurrection in 1979 after losing elections in 1990, Sandinista leader Daniel Ortega vowed they would 'govern from below.' This vision, albeit unrealized, suggested a new way of thinking about where the essence of a revolution is found. The demise of state socialism fueled the search for radical democratic alternatives, rooted in autonomous action within civil society. As globalization undermined territorial definitions of social and political organization, shifting decision-making authority to transnational structures (see *Globalization: Political Aspects*), revolution seemed less about seizing state power. From the Internet-savvy Zapatista rebels of Chiapas, Mexico to the transnational antisweatshop movement and the 1999 Seattle protest against the World Trade Organization, there were intriguing signs of an emerging global civil society. These new forms of resistance against globalization stretched the boundaries of existing concepts of 'revolution.'

See also: American Revolution, The; Chinese Revolutions: Twentieth Century; Communism, History of; Cuban Revolution, The; English Revolution, The; French Revolution, The; Globalization: Political Aspects; Meiji Revolution, The; Mexican Revolution, The; Revolutions, History of; Revolutions of 1989–90 in Eastern Central Europe; Revolutions, Sociology of; Revolutions, Theories of; Russian Revolution, The; Social Movements, History of: General; Vietnamese Revolution, The

Bibliography

- Boswell T, Chase-Dunn C 2000 *The Spiral of Capitalism and Socialism*. Lynne Rienner Publishers, Boulder, CO
- Brinton C 1938 *The Anatomy of Revolution*. Vintage Books, New York
- Debray R 1967 *Revolution in the Revolution?* Grove/Atlantic, New York
- Fanon F 1965 *The Wretched of the Earth*. Grove Press, New York 1965
- Foran J (ed.) 1997 *Theorizing Revolutions*. Routledge, London
- Goldstone J A 1980 Revolution from above – military bureaucrats and development in Japan, Turkey, Egypt, and Peru-Timberger, Ke. *World Politics* 32: 425–53
- Goldstone J A 1991 *Revolution and Rebellion in the Early Modern World*. University of California Press, Berkeley, CA
- Goldstone J A (ed.) 1998 *The Encyclopedia of Political Revolutions*. Congressional Quarterly, Washington, DC
- Gurr T R 1970 *Why Men Rebel*. Princeton University Press, Princeton, NJ
- Huntington S P 1968 *Political Order in Changing Societies*. Yale University Press, New Haven, CT
- Keddie N R (ed.) 1995 *Debating Revolutions*. New York University Press, New York
- Kuhn T 1997 *The Structure of Scientific Revolutions*, 2nd edn. University of Chicago Press, Chicago

- Marx K, Engels F 1998 *The Communist Manifesto*, 150th anniversary edn. with essay by E Hobsbawm. Verso, London
- Moore B Jr. 1966 *Social Origins of Dictatorship and Democracy: Lord and Peasant in the Making of the Modern World*. Beacon Press, Boston, MA
- Selbin E 1999 *Modern Latin American Revolutions*, 2nd edn. Westview Press, Boulder, CO
- Skocpol T 1979 *States and Social Revolutions: A Comparative Analysis of France, Russia and China*. Cambridge University Press, Cambridge, UK
- Skocpol T 1994 *Social Revolutions in the Modern World*. Cambridge University Press, Cambridge, UK
- Skocpol T, Goodwin J 1994 Exploring revolutions in the contemporary. Third World. In: Skocpol T (ed.) *Social Revolutions in the Modern World*. Cambridge University Press, Cambridge, UK
- Tilly C 1978 *From Mobilization to Revolution*. Addison-Wesley, Reading, MA
- Wickham-Crowley T P 1992 *Guerrillas and Revolution in Latin America: A Comparative Study of Insurgents and Regimes since 1956*. Princeton University Press, Princeton, NJ
- Wickham-Crowley T P 1997 Structural theories of revolution. In: Foran J (ed.) *Theorizing Revolutions*. Routledge, London

R. Stahler-Sholk

Revolutions, History of

This brief article on the general problems of studying revolutions is divided into five sections. The first will discuss the term 'revolution' as an historical phenomenon which has changed its meaning over time with important consequences. The second will discuss the issue of taxonomy or classification, attempting to exclude upheavals which are not really revolutionary, and to distinguish different types of genuine revolution. The third will discuss etiology or origins, examining the attempts of historians, political scientists, and sociologists to find common causes of revolution. The fourth will discuss morphology or the possibility of finding a paradigm. This section will discuss the problems of drawing a graph which has validity for the all of the great revolutions of modern times. This subsection will question the appropriateness of trying to place the course of major revolutions on a grid showing extreme right, right, center, left, and extreme left. The fifth subsection will make some concluding remarks.

1. The Evolution of the Term Revolution

Only a handful of historians and social scientists have paid much attention to the history of the term 'revolution,' despite the fact that this history casts light on important developments since the Renaissance. The term emerged from an obscure Latin word

meaning the periodic return of a moving object to its point of origin. Throughout the Middle Ages it was used in its Latin, French, and English forms to describe the supposed cyclical movement of celestial bodies around the Earth. Then it was popularized by heliocentrics following Copernicus's *On the Revolution of the Heavenly Orbs* in 1543. Galileo and his followers used the term repeatedly. Contemporary dictionaries reflected the domination of the term by scientists. In his famous French-English dictionary published in 1611, Randle Cotgrave defined the word as 'a full compassing; rounding, turning back to its first place, or point, th'accomplishing of a circular course.' Significantly, he gave no non-scientific meanings.

Gradually the meaning of the term began to expand. Poets and playwrights had already begun to appropriate the word. In *The Complaynt of Mars*, (1.03), written c. 1378 Chaucer speaks of 'hevenish revolucioun.' Then at the beginning of the seventeenth century Shakespeare pre-empted the word to describe his view of death. In *Hamlet*, when the grave-digging clown unearths a skull, the prince of Denmark reflects that it might have belonged to a politician, a courtier, or a lord: 'here's a fine revolution, an we had the trick to see't (v.1).' Shakespeare thus used the word to mean the return to one's original condition, earth to earth, ashes to ashes, dust to dust.

While literary men were using the term 'revolution' to mean a cycle, political commentators were already using it to mean a political upheaval. Several Italian authors used the word during the Renaissance to describe political turns of fortune, the replacement of one prince by another, in the city-states of Italy. Then in 1674 Alessandro Giraffi published a book entitled *Le Rivoluzioni di Napoli* to describe the rebellion of the Neapolitans against the Spanish viceroy. Another Italian and a Frenchman used the term to describe the same event. Moreover, Giraffi's book was translated into English within a few years.

In the UK most writers referred to the struggle in the middle of the seventeenth century between parliament and the king as the Great Rebellion, but some authors on both sides referred to it as a revolution, for instance the radical parliamentarian Anthony Ascham and the royalist Matthew Wren. In a letter to the Earl of Dorset in January 1646, the royal historiographer James Howell used the word to mean both a political upheaval and the cycle of the years: 'for within these 12 years there have been the strangest revolutions and the horridist things happen'd not only in Europe, but all the World Over, that have befallen mankind, I dare say since Adam fell, in so short a revolution in time' (James Hollowell, 1890. *Epistolae Ho-elianae*, Bk. III, 512). His contemporary John Milton also used the word in these two senses (*Paradise Lost*, II, Lines 597-603; VIII, lines 15-31).

Although the radical parliamentarians in mid-seventeenth-century England claimed that they were restoring ancient rights, they already had some

inkling of a new beginning. After the overthrow of the monarchy they spoke of Year I of Liberty, Year II, Year III, and so on for a few years. Some radicals even had a vision of society completely transformed. 'But why may we not have our Heaven here,' Gerrard Winstanley asked, 'and Heaven thereafter to?' (Sabine, 1941, p. 409). Nevertheless the idea of reversion persisted: even Winstanley was thinking of a return to a primitive simplicity. The royalists also used the word to mean a return to a previous state. For William Temple the return of the Stuart dynasty in 1660 and the Orange dynasty in The Netherlands in 1672 were revolutionary events in the cyclical sense of the word. Although in the second of his *Two Treatises on Government*, published in 1690, Locke defended the right to revolution, he used the word only twice, each time to mean a return via the state of nature to a previously existing constitutional position (1964, Cambridge edn. Paragraphs 223 and 225, 432–3).

Oddly enough it was not the Great Rebellion of mid-seventeenth-century England which made the word an important term in political vocabulary. Only after the Glorious Revolution in 1688 did lexicographers assign the word a political meaning. The use of the word to describe the overthrow of James II and the signing of the Bill of Rights by William and Mary gave the word a wider meaning. It was not quite a restoration of a previously existing order, rather it consolidated and advanced rights for which parliament had fought for the preceding century. Nevertheless, it was a very conservative event, entrenching the power in parliament of aristocrats, landed gentry, and rich merchants. Also, the Revolution was not seen as something with a universal application; it was viewed as English, unique, unrepeatable. The Bill of Rights of 1688 was not for all humanity.

Then in the last quarter of the eighteenth century the term came to mean a profound rupture with the past and the beginning of a new era. On the Great Seal of the US we see 1776 associated with the Latin motto *Novus Ordo Seclorum*, 'a new order in the ages.' Then the French Revolution consolidated and intensified the idea of a new beginning in history. Like the Commonwealth-men earlier, they renumbered the years starting in 1789. When the Republic was established in 1792 they began to number the years Year 1 of the Republic and so on. Finally, in October 1793 the Convention introduced a new calendar numbering the years, not from the birth of Christ, but from the inauguration of the Republic on September 22 of the previous year. Even more significant was that the French revolutionaries believed that their ideals would have a universal appeal. The three Declarations of Rights—1789, 1793, and 1795—were to apply to all men, not just Frenchmen.

Because of the cluster of revolutions in the late eighteenth and early nineteenth centuries revolution became recognized as a recurrent phenomenon. This in turn led to the appearance of a new character on the

world stage: the professional revolutionary. There was an apostolic succession of such characters—Filippo Buonarroti (1761–1837), Auguste Blanqui (1805–1881), Mikhail Bakunin (1814–1867), Piotr Kropotkin (1842–1921), Victor Serge (1891–1947), Sun Yat-sen (1866–1925), Mao Zedong (1893–1976), and many others. These life-long revolutionaries anticipated revolution, created organizations to bring it about, and formulated plans about what to do when it arrived. Thus was born the revolutionary political parties which changed the course of modern revolutions. Once such parties capture power, they are able to suppress opposition for very long periods. Until these parties are finally dislodged from power changes in policy can only take place by decisions within the party leadership.

2. Taxonomy

Many upheavals are labeled revolutions, but few actually qualify for the term. Most *coups d'état* do not qualify since participation in them is limited and their result is often confined to a change in government personnel. So-called 'palace revolutions' usually involve struggles with factions or families within the government. Rebellions should probably be excluded, not only because they do not succeed, but also because they usually lack a clear program and are often led by nobles who do not really want to destroy the existing political and social order. Even wars of independence are dubious candidates for the term, contrary to conventional usage. Even the American Revolution scarcely deserves this title. Although the War for Independence saw individual rights entrenched, considerable exchange of property, and a large emigration, the internal institutions of the new states were not radically different from those of the former colonies. Some African struggles in the twentieth century against European imperialist powers had arguably deeper internal repercussions.

Even among upheavals which deserve the title there are many degrees of revolution. There are political revolutions such as the Revolution of 1830 in France, which produce a change in the groups enjoying power, but do not shake society up very much. The great vertical upheavals or social revolutions such as the French Revolutions of 1789, the Russian Revolution of 1917, or the Chinese Communist Revolution of 1949, have led to deep social changes. The latter two, however, are not only political and social revolutions, but socialist ones as well. The Babouvist Conspiracy of Equals in France in 1796 was probably the first case of a group, in which some members supported a primitive communist ideal, which actually tried to seize power. Genuine socialist revolutions, attempting to control and redirect the means of production, could obviously not occur until socialism developed in the nineteenth and twentieth centuries.

Revolution in the deepest sense involves not only a change in the groups exercising power, profound changes in the social structure, and in modern times an attempt to change the control over the means of production, but also an attempt to change the way citizens think, feel, and act. There are glimmerings of such an attempt in the Hussite and Puritan Revolutions, but the French Revolution of 1789–1799 was the first to try to implement it in a large way. At the peak of the Revolution in Year II the *gouvernement révolutionnaire* tried to ‘regenerate’ citizens through an attempted mobilization of all the existing media: art, architecture, festivals, music, newspapers, theatre, and children’s books and games. Such use of the media has since become a feature of major revolutions on both the left and right, such as Nazi Germany, Bolshevik Russia, and Communist China. Comparative studies of revolutions seldom discuss this added dimension to revolution.

3. Etiology

There has been great attention to the question of the origins of revolution. If it were possible to establish common causes for revolutions, it would greatly simplify the study of the great revolutions since the seventeenth century, each of which has produced thousands of studies. Once common causes were identified, this knowledge would also be useful both to those anxious to promote revolutions and those intent on preventing them. The hope of applying social science for such purposes has produced such books as Michael Velli’s *Manual for Revolutionary Leaders* published in 1972, and on the other side the collection edited by Harry Eckstein (1963), intent on curbing such internal violence.

Since Aristotle there has been wide discussion of the etiology of upheavals, rebellions, and revolutions. Most historians, however, do not find the literature comparing revolutions of much use in understanding the revolutions in which they specialize. Many of them point to the problem of assessing the weight to be given to various factors at work, before and during revolutions. Unlike scientists who can run experiments in which all factors except one are held constant, historians cannot rerun events to test the role of certain factors. An historian, for example, cannot demonstrate precisely how the closing decades of the eighteenth century in France would have been different if reformers had been able to solve the financial problems, if those problems had not been exacerbated by intervention in the American War for Independence, if there had been a different King from Louis XVI, if he had not called royal troops into Paris in July 1789, if he had not tried to flee the country in June 1791, if the Girondins had not led the country into war in 1792, or if Robespierre had never been born. Similar questions can be posed about all subsequent revolu-

tions. Specialists try to weigh various factors by getting as deeply as possible into the complexities of each upheaval.

Aristotle argued that rival concepts of justice led to upheaval. Modern theoreticians such as Karl Marx, Max Weber, Pitirim Sorokin, Lyford Edwards, Jaroslav Krejci, Mark Hagopian, Peter Calvert, Chalmers Johnson, and host of others have focused on increasing tensions in societies leading to revolutions, provided there are precipitants and triggers. These tensions are described in various way including class conflict, other contradictions of various kinds, frustration of the expectations of certain groups, dissynchronization between social development and existing values, and other sources of disequilibrium. The fact that there have to be severe tensions in society in order to have an internal war seems a truism. The difficulty with such generalizations is in applying them to very different circumstances. Take, for example, the varying role of war. The American Revolution was made possible in part because the threat of invasion from French territories to the north had been removed by British victory in the Seven Years War. France was not at war when revolution broke out in 1789. In contrast, the government in Russia had been severely undermined by war by 1917. Also in China the long struggle with Japan contributed to the collapse of the Guomindang and the Communist victory in 1949.

Another important variation among revolutions is their very different objectives. As was mentioned earlier, there were no socialist revolutions before the twentieth century. Twentieth-century revolutions in Russia, China, Cuba, and elsewhere have attempted to seize and redirect the means of production. Related to this has been the attempt of countries such as Russia, China, and North Korea to force development in order to catch up with Western Europe and North America. Such a drive was not present in the Hussite, English, or French Revolutions, in fact the hope for a sudden increase in productivity could scarcely have arisen before industrialization, steam power, electrification, and the internal combustion engine. Comparing revolutions over five or six centuries seems to minimize new ideologies, in fact novelty in human history.

One way of looking at the objectives of various revolutions is to study how they were symbolized. In the French Revolution the dominant symbols were the figure of Liberty, which gradually evolved into the figure of the Republic, the Liberty Bonnet, and the Liberty Tree, none of which reappeared in the Bolshevik Revolution. Russian revolutionary symbolism featured giant figures of workers holding Hammers and Sickles, workers depicted as part of industrial machinery, electric lights in modernized villages, and of course red stars and flags. In Maoist China there were no symbols of liberty, but thousands of images of Mao like a red sun in the sky, numberless little red books of Mao’s writings, and endless images of peasants using muscular power to create terraced

hillsides, build dams, canalize rivers, and irrigate the lands. In fact their terraced hillsides became a symbol of using intensive labor to 'conquer nature.' Machinery was relatively inconspicuous until after Mao's death in 1976. The figure of Liberty was absent until students made her prominent briefly in the abortive democracy movement of 1989. Comparative studies of revolution have so far largely ignored different allegorical figures and symbols.

4. Morphology

There are similar problems with the attempt to find common patterns in the course of revolutions. One of these is that most historians, sociologists, and political scientists give much more attention to the origins of revolutions than to the inner dynamics which drive them forward, producing the curves so dear to social scientists. A further problem is how long a period one considers a curve. Jaroslav Krejci extends the Hussite Revolution over 50 years, the English Revolution over 60 years, and the French Revolution over 100 years. Can one speak of a single French Revolution over a century which saw two empires, three forms of monarchy, five revolutions (if one considers 1792 a revolution within the revolution of 1789–99), and numerous coups? Also, if you compare his graphs of the six revolutions that he covers—the Hussite, the English, the French, the Russian Bolshevik, the Turkish, and the Chinese Communist Revolutions they vary considerably from each other (1983, pp. 35, 67, 82, 117, 149). Moreover, his curves vary considerably from those published by Mark Hagopian (1974, p. 245). Also it is strange that Hagopian shows the Chinese Revolution dipping in 1958, the precise date of the Great Leap Forward. Obviously reducing great revolutions to scientific-looking graphs is a risky endeavor.

A further problem arises from placing revolutions on a grid marked extreme right, right, left, and extreme left. These terms arose from the accidental seating of conservative deputies to the right of the speaker in the Constituent Assembly in the French Revolution of 1789. In the liberal democratic revolutions in Europe and North America the left has usually meant those who supported genuine parliamentary government, political opposition, and defense of human rights. Can one, however, put the English, the American, and the French Revolutions on the same right-left grid as the Russian, Chinese, and Cuban Revolutions? Placing modern communist revolutions on the left distorts the original meaning of the term. These twentieth-century revolutions have introduced single-party regimes, have created assemblies which approve decisions already made by the leaders, and have violated human rights. In fact the way they have operated is very similar to governments on the extreme right, '*les extrêmes qui se touchent*,' as the French put it. This is another example

of the perils of applying the same grid to revolutions with very different objectives and consequences.

5. Conclusion

While it seems very difficult to establish a common taxonomy, etiology, and morphology covering all the major revolutions in the last five centuries, this does not mean that the comparative study of revolutions is not worthwhile. The history of the term 'revolution' demonstrates that it has taken on important new dimensions since it first became politicized in Renaissance Italy. The taxonomy of revolutions can exclude upheavals that are not full revolutions and can distinguish among political, social, and socialist revolutions, or combinations of all three. The etiology of revolutions is particularly problematical because the conditions that occur vary widely, nevertheless the idea that disequilibrium of some kind precedes revolution seems a valuable one. In many cases revolutions occur when society is changing while the regime stands still, but in the case of the Iranian Revolution of 1979 the reverse seems true, disequilibrium resulting from a westernizing shah moving more quickly than powerful conservative groups, especially religious ones, were willing to accept. On the other hand the collapse of the satellite Communist regimes in eastern Europe after 1989 seems to be the result of the failure of those governments to satisfy the needs of their citizens and the weakness of systems imposed by external force. As for establishing a common pattern or morphology, the differences in the graphs which scholars have produced warn of the dangers in such attempts. Those historians, however, who steep themselves in just one revolution, are very unlikely to recognize what is unique about the particular upheaval in which they specialize.

Since revolution has been a recurrent phenomenon, and may be so in the future, social scientists will doubtlessly attempt to reach general conclusions about them. In such future attempts more attention needs to be given to the internal dynamic of revolutions once they have broken out. The idea of attempting to create a new citizenry by mobilizing all the available media, which has marked revolutions since 1789, deserves further study. So too does the use of symbols, allegorical figures, and political images which have distinguished revolutions in modern times. Moreover, graphs based on a left-to-right scale need to be modified because the meaning of left and right has shifted drastically over time. Finally, the comparison of revolutions must bring out the unique dimensions of each revolution instead of focusing mainly on similarities.

See also: Communism, History of; Conflict: Anthropological Aspects; Conflict Sociology; Crowds in

History; Internal Warfare: Civil War, Insurgency, and Regional Conflict; Marx, Karl (1818–89); Revolution; Revolutions of 1989–90 in Eastern Central Europe; Revolutions, Sociology of; Revolutions, Theories of; Social Movements, History of: General; Social Movements, Sociology of; Social Protests, History of

Bibliography

- Brinton C 1960 *The Anatomy of Revolution*. Vintage Books, New York
- Brunno O, von Conze W, Koselleck R 1972 *Geschichtliche Grundbegriffe*. Klett, Stuttgart, Germany
- Close D, Bridge C (eds.) 1985 *Revolution: A History of the Idea*. Croom Helm, London
- DeFronzo J 1999 *Revolutions and Revolutionary Movements*, 2nd edn. Westview, Bolder, CO
- Dunn J 1989 *Modern Revolutions: An Introduction to the Analysis of a Political Phenomenon*. Cambridge University Press, Cambridge, UK
- Eckstein H (ed.) 1963 *Internal War. Problems and Approaches*. Free Press, Glencoe, NY
- Ellul J 1971 *Autopsy of Revolution*. Knopp, New York
- Goldstone A 1991 *Revolution and Rebellion in the Early Modern World*. University of California Press, Berkeley, CA
- Hagopian M N 1974 *The Phenomenon of Revolution*. Dodd Mead, New York
- Johnson C 1964 *Revolution and the Social System*. Hoover Institution Studies 13, Stanford University, Stanford, CA
- Krejci J 1987 *Great Revolutions Compared*. Harvester Press, Brighton, UK
- Leith J A 1974 *Media and Revolution*. Canadian Broadcasting Corporation, Toronto, Canada
- Lipsky W E 1976 Comparative approaches to the study of revolution: A historiographical essay. *Review of Politics* 38(4): 494–509
- Pettee G 1938 *The Process of Revolution*. Harper and Row, New York
- Sabine H (ed.) 1941 *Works of Gerrard Winstanley*. Cornell University Press, Cornell, NY
- Sorokin P 1967 *The Sociology of Revolution*. Howard Fertig, New York
- Velli M 1972 *Manual for Revolutionary Leaders*. Blade and Red, Detroit, MI
- Zagarin P 1973 Theories of revolution in contemporary historiography. *Political Science Quarterly* 88(1): 23–52

J. Leith

Revolutions of 1989–90 in Eastern Central Europe

1. A Deviant Type of Revolution

The revolutions in Eastern Europe were no conventional changes of the system that can be dealt with in the tradition of sequence models. There

was no uniform anatomy of the changes, though never in European history have so many regimes collapsed at the same time and for similar reasons. Revolution was traditionally linked to violent change that made it different from coups d'état. Nevertheless, the depth of change made the events between 1989 and 1991 a political and a social revolution at the same time. In this respect it was comparable only to the Bolshevik revolution of 1917 and its later repercussions in the 'people's democracies' after 1945 in the realm of influence of the Red Army. This double revolution made consolidation of the new 'revolutionary regime' so difficult.

No former system's change was as burdened with the same problems of institution-building as this one. Authoritarian and fascist dictatorships may have changed the political institutions, but they hardly touched the system of stock markets and banks. There was no need in post-authoritarian regimes to create offices to administer unemployment insurance nor institutions for social security.

This process of transformation was so unique that the old notions of revolution and reform tend to be inappropriate. Perestroika in the Soviet Union was frequently compared to the politics of Peter the Great, the Meiji Revolution, or the innovations of the Young Turks in the Ottoman Empire. Reform was defined as an eclectic un-ideological process of un-dogmatic actors. The consequence was contradictory measures and even 'messiness' of reform. *Ex post facto*, perestroika proved its messiness. But was it really not ideological? This author prefers to list perestroika under the type of reforms from above, steered by the ideological thrust of a great leader.

In Gorbachev's self-perception, perestroika was a revolution. The old Soviet Union under Brezhnev started to accept rather loose talk about revolutions. Revolution lost its connotation of a process intended by political elites and by political means. It was watered down to processes that the Central Committee and the Politburo of the party were no longer able to control, thus demographic, environmental, and technological revolutions entered the discourse in communist countries. Gorbachev tried to use revolution again in a more political way. Perestroika was dubbed a 'second revolution' and Gorbachev hinted at Lenin who had already envisaged several revivals of revolutionary activities in the tradition of the great French revolution that failed to implement its goals in one coup and had to be repeated in 1830, 1848, and 1871. Western analysts, on the other hand, never used the term revolution in describing the process of perestroika.

Only at the end of perestroika had the term system's change found universal acceptance. It lacked the connotations of violence and civil war implied in the term revolution, but is nevertheless aimed at a far-reaching change of politics, society, and economics. The minimal criterion is a break-down of the old

system. This does not need, however, revolutionary force.

The revolutions in Eastern Europe were different from former revolutions because—with the exception of Romania—never have so radical changes been achieved with so little violence from below and from above. Eastern Europe proved to be a deviant case in the history of revolutions for several reasons:

Political modernization was achieved in two different phases. The system's change was a second attempt at modernization after the socialist attempts in 1917 and 1945.

The collapse of socialism resulted from a threefold dissatisfaction of East European populations: dissatisfaction with a repressive political system, with an inefficient economic system, and with the dominance of hegemonic ethnic groups within a federation (Russians, Serbs, and Czechs) or within the socialist camp (the Soviet Union vis-à-vis its satellites).

There was not the same continuity of periods as in Southern Europe in the 1970s with phases such as liberalization, democratization, and consolidation.

The new process of transition was burdened with the unique double task of transforming the political and the economic system at the same time, something which has never before happened in world history.

The new transition process was hampered by ethnic strife, which occurred to an unknown extent. Spain had experienced some of it, but managed to solve the problem by various autonomy concessions.

In former transitions, a new elite was ready to take over power. There was no real preparation for the new task in Eastern Europe. Even non-Marxist scholars sometimes believed in Marxist periodization schemes. They held that a system needed to develop an economic basis and a recruiting ground for the new political elite. The old regime developed the preconditions of change. Revolution was only the last blow for the old system. The simplicity of this concept of transitions did not even work in the great French revolution. Historians have shown that the French revolution did not abolish feudalism or establish a market society throughout the country. This process of accomplishing a true market society was not successful before the beginning of the Third Republic (1875 ff).

2. Reasons for the Velvet Revolutions

First, in the history of changes of regimes the leadership groups have frequently disunited. One part of the power elite started mobilizing the masses. In Eastern Europe this model did not apply. The elites remained quiet until the last moment. The technical intelligentsia did not develop into an opposition to the 'strategic clique' in power, as had been forecasted. The technical intelligentsia increasingly de-ideologized and created more room for maneuvering in its own field,

independent of political leadership. It avoided conflict, however.

Reformist elites in those socialist systems where the regime eroded slowly and revolution was pacted in a corporalist mood also avoided the appeal to the masses. In Hungary, the negotiated revolution took place with complete discipline and there was no unified counter-elite. In Poland, Solidarity could have been the counterforce, but in the last stage it was pressed by the people rather than directing the masses into conflict with the 'nomenclature' leadership.

Only in Romania—in the third model of a collapse of the system, but under control of a middle-level elite of the old regime—was there conflict among elite sectors. The military, after some wavering, finally sided with the protest movement, not so much because of ideological sympathies with the revolutionaries, but because of its rivalry with the Securitate armed forces. The military played the role of a rear guard—blamed by Lenin as '*chvostizm*'—but apparently the capacity to wait and see and to grasp the right moment for action in the rear guard is not always a negative feature, as Lenin supposed.

Second, the European fascist dictatorships in 1945 collapsed because they were defeated in a war. In Southern Europe and Latin America external factors contributed to the collapse of a regime (Greece, Portugal, and Argentina). In Eastern Europe the changes of the international situation facilitated the peaceful velvet revolutions. The Soviet Union still existed as a hegemonic power, but Gorbachev had indicated that the satellites could no longer rely on Soviet help in a case of internal strife. It was less the active behavior of a superpower, than the non-decisions of the hegemonic bloc power that did not intervene, which caused the collapse of the whole alliance. Theoretically, each national elite could have followed the Chinese example of bloody suppression. Why did this not happen?

The most important reason is again the international factor: the 'Brezhnev doctrine' was meant to strengthen weak regimes within the bloc. But the bloc center at Moscow was not prepared for a situation where all the leaderships of her satellites entered into troubles at the same time. The Brezhnev doctrine proved its dysfunctional consequences. It had spoiled the satellite leadership because they were always able to rely on big brother's tanks. Only Romania was the deviant case which did not recognize the doctrine and therefore had to provide for her own security. All of the other governments were not accustomed to big actions for their own security and caved in as soon as the masses confronted them directly.

The old regime in its last defensive stage pretended that the turmoil was caused by subversion from the 'imperialist powers.' This was wrong, since the Western powers were so accustomed to détente that they avoided direct actions for destabilizing the communist systems. It was not international but rather transna-

tional policies which contributed to the destabilization of communism from outside. The protest movements in the streets of Leipzig did not receive weapons, but got information instead from the Western media about the situation in the country, the counter-reactions of their leadership, and the points of rallying for the protest movement.

The international factor could explain certain framing conditions of the development, but nobody would have dared to forecast that the Soviet Union was so occupied with her internal problems that she had no capacity for intervention abroad.

Third, the most complex explanation of regime breakdown is the loss of legitimacy. In socialist countries legitimacy could not be measured by surveys. The mass media in the West, such as *Free Europe* and *Liberty* have tried to measure legitimacy by the instrument of mail from the East, but results were unreliable. In the early 1980s, it was calculated that the communists in free elections would not get more than 3–5 percent of the votes, whereas they actually received 10–15 percent in the most developed systems and much more in the less-developed communist regimes in the southeastern parts of Europe.

Was socialism ever legitimized? Its foundation by force was not forgotten by large parts of the population in the satellite countries. Acceptance of the regime was concentrated on the cadres. With growing success and material benefits, however, legitimacy grew also under communist conditions.

Paradoxically, communism collapsed due to its partial success, especially in the sphere of producing intellectually trained elites. There was a mobilization for education, which hardly had an equivalent in the Western world. Communist systems produced a surplus of elites, whom they were unable to funnel into meaningful positions. Western democracies do the same, as periodical overproduction of elites seems to be as inevitable as the famous pig cycle, causing economic slumps in the economic and political conflicts. Western systems, though, can offer jobs for the new elites in the third sector of services when the elites cannot be integrated into the bureaucracy. This tertiary sector was chronically underdeveloped in socialist systems. Only the systems that were able to coopt the dissatisfied young elites, such as Hungary, experienced a slow transition to the new regime. Here the rebellion could be dubbed as the 'revolution of the deputy heads of administrative units.'

It would be unhistorical to look for a legitimacy of the whole life span of socialism. Legitimacy had also ideological foundations, and beliefs in legitimization are subject to change. In the early days of revolutionary socialism, the system claimed a kind of teleological legitimization because the party pretended to know the outcome of history. In the Soviet Union this was certainly combined with charismatic legitimization in the time of Lenin. As time went on, however, both withered away, charisma as well as the capacity to

predict the development of history. Quite frequently, the people experienced false prognoses that undermined the belief in the ideology. As Max Weber predicted, the personal charisma of the leader did not survive the founding father. *Veralltäglicung*, the erosion of charisma, was the necessary consequence. Stalin tried to substitute for his lack of charisma through organization and ideological leadership in all questions: from literature to biology. But this kind of mixed legitimization was increasingly built on force. With diminishing capacities of the regime for repression and mobilization, new forms of legitimization had to be found. Communist repression had to be substituted by more legal security. In the early days, huge groups of defeatists and subversive elements were discovered. After Stalin, nobody believed in this kind of singling out of groups any more. Only 'parasitism' and 'chuliganstvo,' spheres where the Soviet citizen felt endangered for his own security, was he ready to buy from the Communist propaganda.

Communism failed to establish a certain balance between the four basic principles of legitimization in Western systems. The legal state was violated by arbitrary decisions of special hierarchies in the party, in the security forces, and in the planning administration. The nation state was increasingly invoked by official propaganda in order to reinforce socialist ideology, but the hegemony of the Soviet Union prevented the development of authentic independent national states, even if more systems tried to embellish the socialist dogma by the national colors. The democratic state was proclaimed by state propaganda, but in reality, for the citizens it was only manipulated participation without alternatives in the political arena. Certain ultra-democratic requisites such as 'recall' and 'imperative mandate' were operating only with a hint of the party, never because of a free decision of the citizens. Thus, the welfare state was used to legitimize the late socialist systems. In the early days, communist propaganda thought that a social state was unnecessary, because a socialist state is socially minded *per se*. But the leveling down of all social differences was no longer accepted by the leadership. Stalin had already denounced *uravnilovka* (excessive equalization).

During the years of economic growth, Hungarian Goulash Communism or Polski Fiat Communism developed a certain legitimization of the system, but in the 1980s it became clear that the communist countries were falling further behind the West. Communication with Western countries grew, which contributed to a growing capacity of the people to compare East and West. During the 1980s, the system lost its legitimacy due to the economic crisis of socialism, which was felt in most of the countries.

Transition studies on a quantitative basis confronted us with the finding that democratization is impossible as long as the power resources of a regime are highly concentrated. This assumption was not

Table 1
Pathways to democracy

	Pragmatic piecemeal engineering	Ideological innovation
Control from above	Opening of the communist regime (Romania, Bulgaria)	Innovation of socialism (Perestroika under Gorbachev)
Pressures from below	Erosion of Socialism (Poland, Hungary)	Collapse of socialism and takeover by the opposition (GDR, CSSR)

even correct in the case of the communist countries. All of them had the power to suppress the opposition and had in the 1980s increased their service in both staff and technological equipment. Paradoxically, some of the socialist countries, such as the German Democratic Republic (GDR) invested most of their technological innovation in the social control of the population instead of investing it in production. Nevertheless, the system collapsed. The enormous security equipment did not help the system survive: as was the case with the Chinese Wall, Rome's Limes in the North of the Empire, or France's Maginot line in 1940. Unlike these cases, the fabulous equipment was not even applied in the last conflict. The index of power resources of the old regimes proved to be of minor importance in the fourth wave of democratization. On the contrary, the better equipped a communist regime was with repressive capacities, the more radical was the change.

Three models predominated in the system's change of 1989 (see Table 1):

The erosion of the regime by corporatist bargains between the power center and the opposition in a negotiated revolution. This type of transition caused a low rate of participation among the people, because the change was paced by elites without participation of the citizen. Many voters apparently did not even realize the amount of change and the voter turnout stagnated between 50 and 60 percent (Poland, Hungary).

The implosion of the communist regime (Czechoslovakia, GDR). In this model, the control of the ancien regime was least developed. A new elite took power. Continuity of elites and institutions was lower than in the first path to democracy.

System change was achieved by the old cadres of the second rank who kept control of the transition (Romania, Bulgaria (until 1991), Albania (until 1992), Serbia, Croatia, and most of the Commonwealth of Independent States (CIS) throughout the first part of the 1990s).

Change proved most complicated in all the multi-ethnic states. The disintegration of the Soviet Union and Yugoslavia showed elements of a 'bargained revolution.' Bargaining, however, was done not by functional groups but by territorially old and new governments. In no case did the old federation survive and in no case was a new federation feasible after the collapse of the old regime.

Transition theories—in the tradition of theories about the developments of the 1990s—were most interested in the dichotomy between pact and dictate. This dichotomy was confronted with reform and revolution in a matrix (Karl and Schmitter 1991). Most of these matrices only offered a realistic picture of the initial phase of transition.

3. Democratization and Consolidation of the Peaceful Revolutions

'Transitology' as a new branch of knowledge dealing with the peaceful revolutions since the 1970s and 1980s originally used a sequence model along the lines: liberalization, democratization, and consolidation. Only in the model of transition via erosion of the communist rule did a liberalization take place. The period of democratization is normally identified with institution-building (constitutional power and secondary institutions such as parties and interest groups). This process was successful in most Eastern European countries but hardly in the successor states of the Soviet Union. Consolidation of elite attitudes (the 'only game in town' criterion developed by Juan Linz) and the citizens is incomplete in most countries. The Vicegrad states, which formed a unit of economic cooperation, and those which entered negotiations with the European Union (Poland, Czechia, Hungary, Slovenia, Estonia) are fairly consolidated even in terms of economic indicators. But also other systems—outside the CIS—have consolidated by experiencing several free elections, by reducing the volatility of votes, by avoiding violence, and reducing the weight of extremist parties by practicing alternating governments that included even the acceptance of a comeback of the post-communists that turned in many countries to democratic socialism (Lithuania 1992, Poland 1993, Hungary 1994). Most countries have changed the composition of their governments from The Czech Republic (1998) down to Romania (1996) without turning the critical election into a crisis of the system.

Not all the processes of consolidation were successful, however. 'Consolidology' as a new branch in the study of democratization discovered various types of unconsolidated democracies that have been called 'delegated democracies' or 'defective democracies'. Most of these defective democracies have implemented

a rather democratic mode of participation, but fall short of consolidation in the sphere of the legal state and the treatment of minorities.

See also: Area and International Studies: Development in Eastern Europe; Communism; Communist Parties; Democratic Transitions; Eastern European Studies: History; Eastern European Studies: Politics; Eastern European Studies: Society; Peace Movements, History of; Socialism: Historical Aspects; Socialist Societies: Anthropological Aspects; Socialization, Sociology of; Youth Movements, History of

Bibliography

- Ágh A 1998 *Emerging Democracies in East Central Europe and the Balkans*. Elgar, Cheltenham, UK
- Dawisha K, Parrott B (eds.) 1997 *The Consolidation of Democracy in East-Central Europe*. Cambridge University Press, Cambridge, UK
- Elster J, Offe C, Preuss U K 1998 *Institutional Design in Post-communist Societies*. Cambridge University Press, Cambridge, UK
- Huntington S P 1991 *The Third Wave. Democratization in the Late 20th Century*. University of Oklahoma Press, Norman, OK
- Karl T, Schmitter P 1991 *Transition to Democracy in Eastern Europe*. Macmillan, Basingstoke, UK
- Merkel W, Puhle H-J 1999 *Von der Diktatur zur Demokratie*. Westdeutscher Verlag, Opladen, Germany
- Linz J, Stepan A 1996 *Problems of Democratic Consolidation*. Johns Hopkins University Press, Baltimore, MD
- Oksenberg M, Dickson B J 1991 The origins, processes and outcomes of great political reform. In: Rustow D A, Erickson K P (eds.) *Comparative Political Dynamics*. Harper Collins, New York
- von Beyme K 1996 *Transition to Democracy in Eastern Europe*. Macmillan, Basingstoke, UK
- Whitefield S 1993 *The New Institutional Architecture of Eastern Europe*. St. Martin's Press, New York
- Zecchini S (ed.) 1997 *Lessons from the Economic Transition. Central and Eastern Europe in the 1990s*. Kluwer, Dordrecht, The Netherlands

K. von Beyme

Revolutions, Sociology of

The modern concept of revolution comes out of one particular experience at the end of the eighteenth century, an experience that led some to hope for and others to fear similar occurrences. Reflections on the nature, causes, unfolding, and consequences of revolution were for a long time dominated by the French Revolution. As other cases came to preoccupy social scientists in the course of the twentieth century, and

particularly in the wake of the varied social upheavals after World War Two, scholarly reflection on revolution flourished as it developed in new directions.

1. An Old Word Takes on a New Meaning

'Revolution' (from Latin *revolutio*) initially had the sense of a return to a point of origin. In political discussion it might be used to characterize a dynastic restoration. By the seventeenth century it had also taken on the sense of an alteration in fortune and might be employed to describe some significant improvement or deterioration in political affairs. In this latter sense revolutions were often seen as frequent occurrences in the histories of states. For some eighteenth century writers, the installation of a new monarch in England in 1688 was a great singular revolution in the older sense of a return to an initial state of order after several tumultuous decades; for others it was but one of many revolutions that marked all of English history.

A century after this 'Glorious Revolution' in England, the geographic range and social diversity of French turbulence in 1789 and beyond (rural and urban insurrections throughout the country; military mutinies; elite challenges to royal authority; priests defying bishops; a tidal wave of legislative enactments) provided many revolutions (in the plural) for commentators to reflect upon. But as these upheavals continued to multiply and as various contenders sought to channel the dramatic events in one or another direction, participants began to speak of these many separate events as components of something singular, the French Revolution, that had a structure and causes, that was an ongoing process, and that rival contenders sought to understand and control. Some participants who were disappointed by specific events in France could even imagine that this particular process was but an instance of possible revolutions and look forward, as one radical group declared in 1794, to 'another, far greater, far more solemn revolution' of which 'the French Revolution is but the forerunner' (Furet and Ozouf 1988, p. 203). So some sought to make new (and perhaps differently constituted) revolutions and others to avert them; and theorists of either (or neither) persuasion sought to understand them.

Until 1917 the French instance was the central source of empirical evidence of what a revolution might be like. Much sociological reflection was profoundly stimulated by spectacular failures of radical causes to seize and hold power despite episodes of impressive popular mobilization such as the widespread European turbulence of 1848 or the Russian upheaval of 1905. But theorists were paying relatively little attention to social conflicts in lesser powers. Even well into the twentieth century, other revolutions, both imagined and experienced, were apt to be understood

with concepts derived from that French experience. Military terminations of popular upheaval were held to be 'Bonapartist'; Russian revolutionaries accused each other of being 'Thermidorians'; it has remained very difficult to think of political conflict without invoking notions of 'left' and 'right' (initially, spatial locations in revolutionary France's legislature).

The Russian Revolution of 1917 added some new possibilities to stimulate imaginations; the post-World War Two era, with its many anticolonial struggles, civil wars in which at least one party has claimed itself to be revolutionary, and regime transitions of many sorts, has provided even more food for thought about the causes, dynamics, and consequences of revolution. By the late twentieth century, the French dramas from 1789 on no longer stood as the central model for theoretical contemplation.

2. An Elusive Definition

Despite two centuries of reflection and an extremely vigorous body of comparative scholarship on revolution in the last quarter of the twentieth century, the very definition remained in dispute. First of all, to the extent that notions of revolution were abstracted from that singular French experience rather than from some sense of the commonalties of a large number of experiences, observers could readily differ on what they thought were defining features. Some of the elements that, in varying combinations, entered into various definitions might include:

- (a) Illegal seizure of national political power.
- (b) Extensive property transfers.
- (c) Radical alteration in the form of government.
- (d) Pervasive change in a wide variety of institutions.
- (e) Extensiveness of popular mobilizations.
- (f) Substantial violence.
- (g) New formulas for the legitimation of authority.
- (h) A new elite's claim to be effecting a radical break in the course of history.
- (i) A new elite's claim to be advancing forward to a new level of human progress.
- (j) Widespread sense of profound change.

Although scholars have never achieved consensus, this definitional debate opened up a whole host of fruitful questions about how these phenomena might be connected to one another.

Second, like many other concepts that are used by political actors, the term revolution has continued to be a term of legitimation and delegitimation rather than simply analysis. All sorts of things have been labeled (or not labeled) as revolutions in order to honor or dishonor them. This has been terribly confusing. A grammatically simple sentence, 'Revolutions entail considerable violence' might be any of at least four very different sorts of statement: part of a *definition* of revolution; a *proposition* about revolution

(when defined without reference to violence); an *exhortation* to violent tactics (when uttered among committed revolutionaries); a *condemnation* of revolution (when uttered among people who abhor violence).

3. Major Research Traditions

Divergent emphases fueled overlapping but distinguishable traditions of research on popular collective mobilization, regime overthrow, political violence, elite dissidence, governmental legitimation and delegitimation, institutional change, and class conflict, and a good deal of reflection on how such processes affected each other. Beginning in the nineteenth century, a well-developed Marxian tradition examined ways in which long-term economic transformations created new group identities and interests while altering political structures. Revolution was seen as a traumatic readjustment of power relations, at bottom driven by the shifting economic interests of rational actors. A theoretically more diffuse body of work looked for strains of various sorts produced by industrialization, urbanization, and secularization and tried to explain collective outbursts as responses to such strains. In such interpretations, revolutions were apt to be seen as 'moments of madness,' as frenzied and doomed responses to the forces of modernity, for example, in Durkheimian notions of 'breakdown' or in the 'collective behavior' research tradition (see *Collective Behavior, Sociology of*). Tocqueville's thesis linked the growing might of a rationalizing state with a broad array of cultural transformations that undermined the legitimation of aristocratic privilege by transferring to the state activities previously carried out by the lords; indeed the lords might themselves become oriented to serving the new state bureaucracy (see *State, Sociology of the*). This attack on the problem lay relatively undeveloped in twentieth century social science until revived by a renewed sociological emphasis on the state from about the 1970s.

Although abstract definitions of revolution continued to vary widely, the field of comparative studies of revolution developed despite varying notions of what relevant instances might be. Some scholars attempted to find processual uniformities and constructed models of the 'natural history' of revolutions. Others looked for some robust causal mechanism that could serve as necessary or sufficient conditions (or both). This latter research tradition was dominated by the search for sources of increased grievance, as in a variety of theories that stressed economic downturns. A very different research tradition, impressed by the role of revolutionary parties in the twentieth century, sought to examine the capacities of relatively small but organized groups, to take action, to catalyze changes in ideas, or to mobilize broad support.

4. *The Post-World War Two Era*

The social upheavals of the post-World War Two era brought many new instances of collective action, political violence, and regime overthrow into play. Anticolonial movements, some defeating great powers (as in Vietnam); long-term rural warfare with divergent outcomes (as in Guatemala and Nicaragua); and Third World state breakdowns (some leading to radically new regimes, as in Iran) all added new cases.

The vastness of human suffering so evident in the twentieth century increasingly suggested attention to opportunity as well as grievance in the genesis of revolt and the success of revolution. Since revolt was so much more scarce than misery, it was evident that something other than (or at least in addition to) misery was needed to explain dramatic struggle for change, and social scientists became adept at identifying structures and moments that furnished opportunities. If one peasant community rather than another supported a revolutionary movement, social scientists now sought an understanding of what resources in the daily lives of these (but not other) peasants might account for why these (but not other) peasants revolted. If one revolutionary party in a small, poor country (for example, Nicaragua) succeeded in coming to power, while another revolutionary party in another small, poor country (for example, El Salvador) succeeded only in sustaining a bitter, stalemated war with the government, social scientists began to explore the differential vulnerabilities of different kinds of regimes. Distinguishing long-standing differences in resources available to rebels from temporary opportunities, and distinguishing stable structures of elite domination from specific policies, became very important for this research tradition. A 'structural' school of explaining revolution was emerging.

But accounts of favorable structures soon seemed incomplete as well. The Iranian Revolution of 1979, in which many leading participants spoke a religious language at radical variance with the language of those who had come to power in previous, paradigmatic cases, brought many scholars to reconsider the role of cultural elements in revolutionary processes. The sudden breakdown of communist regimes in Europe were accompanied by extensive property transfers, but the victors claimed to be undoing revolution, even to be terminating the dream of revolution, which again raised important questions of culture as well as structure (see *Postsocialist Societies*).

Many of the instances in the second half of the twentieth century were occurring in lesser powers on the world stage, in places obviously subject to military, economic, or ideological domination by more powerful states, and whose upheavals often involved attempts (some partially successful) to throw off some of these forms of domination. One would hardly try to understand anticolonial revolutions without paying attention to the policies and institutions of colonial

rule, social upheavals in countries exporting tropical crops without paying attention to the world economy, and the resources available to governments and rebels in small, poor, and weak countries without paying attention to the capacities of both sides to elicit support from great powers.

Insights garnered in examination of the newer instances began to suggest different emphases in new examination of older cases. The French or Russian revolutions were looked at as cultural processes; the embeddedness of the 'great' revolutions of the past in transnational connections was noted; and conflicts beyond the few previously paradigmatic instances entered the comparativists' field of vision. The very variety of participants in the newer cases expanded the intellectual horizons of researchers. As varying combinations of participants in the newer cases became evident, comparative scholars were recognizing how multisided the conflicts were in past revolutions as well. Revolutions, both past and recent, were increasingly likely to be seen as complex affairs, in which governing authority failed in many different ways, elites divided in many different ways, and rebels mobilized in many different ways. Methodologically this meant that social scientists were likely to be looking at configurations: at complex intersections and interactions of processes that mobilized collective actions of varying kinds, produced elite conflict, and weakened governing institutions. Theoretically it means that the search for generalizable statements about necessary or sufficient conditions for revolution needed to be rethought.

5. *Emerging Directions in the Sociology of Revolution*

The issues engaging comparative scholars in the late twentieth century meshed with some of the large issues engaging macrosociology generally, indeed, were suggesting revolutions as a fertile terrain for engaging those issues. Among the issues generating much discussion:

(a) *How to integrate cultural accounts with structural ones.* Cultural issues were an emerging agenda in various ways. Social historians were making enormous advances in understanding popular culture in varied times and places. A new generation of scholars was examining the overt cultural politics of revolutionary elites. The direction of regime change in such places as post-Shah Iran and post-communist Eastern Europe seemed altogether impossible to discuss without considering legitimating ideas.

(b) *How to integrate accounts of individual action with macrostructural processes.* As elsewhere in sociological theory this constituted a major challenge. Proceeding from assumptions of individual rational actors, how could one account for the forms assumed

by collective action, especially in light of the often evident individual risk assumed in participation in revolutions in combination with their uncertain and often undesired outcomes? (see *Action, Collective*).

(c) *Adequate attention to the visions of revolutionary elites, the wellsprings of popular mobilizations, and the interaction of the two.* Some structural accounts seemed to make the visions and intentions of revolutionary leadership inconsequential; some other accounts seemed to make them everything and by implication ignored the motivations, intentions, and wishes of ordinary participants. Some urged us to consider how it was that some revolutionary parties managed to win the adherence of large numbers of followers; others urged us to see popular participation in revolution as rooted in the structures and culture of the popular classes and therefore quite separate from the concerns expressed by revolutionary elites. The empirical difficulty in getting these connections right remained a challenge in studies of particular revolutions, let alone efforts at comparative statement. Poorer, anonymous participants not only often lack the means to make their own wishes widely known (peasants seizing land are less likely to hold press conferences or write memoirs than revolutionary agriculture ministers), but are, also as usual, often highly motivated to conceal their intentions (especially in highly charged situations like revolutions). It is a major methodological problem that revolutionaries who claim to speak on behalf of, say, Central American peasants, are rarely peasants themselves.

(d) *Attention to the outcomes of revolutions.* The consequences of revolution for economic well-being, state power in the international arena, the rationalization of administration, national identities, citizenship rights, and democratization remain intensely controversial matters about which relatively little systematic investigation has been done by comparativists.

(e) *How to integrate accounts of revolution with other kinds of political conflict.* Scholars of revolution were increasingly making use of tools that resembled those used by other students of politics: attending to the play of threat and opportunity, the variety of resources, the forging of identities in struggle, the irony of unintended consequence, the gaps between elites and ordinary participants (and the interaction between the two). Increasing interchange with scholars of other modes of conflictual action (urged by Doug McAdam, Sidney Tarrow, and Charles Tilly) seemed a probable future direction (McAdam et al. 1997).

(f) *How to make superior use of comparative methodologies.* With increasing recognition of the complexity and diversity of revolutionary processes, the very logic of comparison was coming under heightened scrutiny. If revolutions are complex amalgams of multiple conflicts, some comparativists suggest that rather than searching for some single general cause, or some single general process, or some set of causes whose separate contributions to revolution

may be simply added together, there may be multiple configurations of causal processes which produce revolutions, which themselves unfold in a variety of ways. Some scholars sought to distinguish varieties of revolution about which valid generalization might be made. Others suggested that the way forward, theoretically, was to seek generalizations about much smaller processes, which were components of revolutions, but which combined together in so many distinct ways as to defy generalization on that concatenated level. That is to say, in this view 'revolution' is a complex made up of many components, and one revolution is made up differently from another; looking for uniformities in the class 'revolution' is hopeless, but there may be regularities in the causation or action of some of the components.

6. Conclusion

More than two centuries of reflection since 1789 have hardly exhausted the stimulation provided by revolution to the thinking of social scientists; if anything, the last quarter of the twentieth century saw a tremendous intellectual reinvigoration. This is likely to continue, for revolutions seem a particularly privileged terrain for many significant issues in macro-sociology, among which I indicate four:

- (a) the intermingling of change and persistence
- (b) the interplay of long- and short-run processes
- (c) the relationship of local, national, and transnational processes
- (d) finally, the collision of hopes for change and recalcitrant realities.

Revolution brings to the fore all the many issues that have swirled around the interplay of 'structure' and 'agency.' There are no other social processes which so strikingly juxtapose the capacity of human beings to construct for themselves previously unrealized and unexpected patterns alongside the incapacity of human beings to construct their world as they please.

See also: Conflict Sociology; Political Sociology; Power in Society; Revolutions, Theories of; Social Movements, Sociology of; Violence: Public

Bibliography

- Baker K M 1990 *Inventing the French Revolution*. Cambridge University Press, New York
- Brinton C 1965 *The Anatomy of Revolution*. Vintage Books, New York
- Foran J (ed.) 1997 *Theorizing Revolutions*. Routledge, London and New York
- Furet F, Ozouf M 1988 *Dictionnaire critique de la Révolution française*. Flammarion, Paris

- Goldstone J A (ed.) 1998 *The Encyclopedia of Political Revolutions*. Congressional Quarterly, Washington, DC
- Keddie N R (ed.) 1995 *Debating Revolutions*. New York University Press, New York
- Markoff J 1996 *The Abolition of Feudalism: Peasants, Lords and Legislators in the French Revolution*. Pennsylvania State University Press, University Park, PA
- McAdam D, Tarrow S, Tilly C 1997 Towards an integrated perspective on social movements and revolution. In: Lichbach M I, Zuckerman A (eds.) *Ideals, Interests and Institutions: Advancing Theory in Comparative Politics*. Cambridge University Press, Cambridge, UK
- Rey A 1989 'Révolution', *histoire d'un mot*. Gallimard, Paris
- Selbin E 1998 *Modern Latin American Revolutions*, 2nd edn. Westview Press, Boulder, CO
- Skocpol T 1979 *States and Social Revolutions*. Cambridge University Press, Cambridge, UK
- Tilly C 1993 *European Revolutions, 1492–1992*. Blackwell, Malden, UK
- Tocqueville A de 1952–3 *L'Ancien Régime et la Révolution*, 2 Vols. Gallimard, Paris
- Wickham-Crowley T P 1992 *Guerrillas and Revolution in Latin America. A Comparative Study of Insurgents and Regimes Since 1956*. Princeton University Press, Princeton, NJ

J. Markoff

Revolutions, Theories of

Theories of revolution are (or purport to be) causal explanations of revolution—they say that if certain sufficient conditions (the causes) occur, then revolution (the effect) will occur as well. But *revolution* has different meanings, and theories of revolution do not all explain the same facts—different theories explain different facts as well as cite different causes. Before analyzing theories of revolution and the problem situation they address, therefore, one must distinguish between the different facts that those theories would explain—the different events that *revolution* refers to.

1. Definitions

Some theories explain revolution mainly in the sense of a revolutionary movement—a concerted attempt to change government, regime, or society (or all three) by means of violence. Other theories explain revolution mainly in the sense of a revolutionary outcome—a successful change of government, regime, or society (or all three) by means of violence. Still other theories explain revolution mainly in the sense of a revolutionary situation—a contest for state power by means of violence that begins when insurgents claim sovereignty and ends when they overthrow, succumb to, or make peace with the government.

A revolution in one sense can occur without a revolution in the other senses occurring as well. A

revolutionary movement can occur without a revolutionary outcome occurring because revolutionaries can fail. A revolutionary outcome can occur without a revolutionary situation occurring because the government in power can make a 'revolution from above' *à la* Stalin and because revolutionaries can take power legally *à la* Hitler. And a revolutionary situation can occur without a revolutionary outcome occurring because the government under siege can win. But neither a revolutionary situation nor a revolutionary outcome can occur without a revolutionary movement occurring because they each entail revolutionary violence, which takes concerted effort to commit.

The above definitions of revolution overlap with common social science usage, which calls a change of government by means of violence a *coup d'état*, a change of regime (as well as of government) by means of violence a political revolution, a change of society (as well as of government and regime) by means of violence a social revolution, and a contest for state power by means of violence a civil war. Common social science usage can be misleading, however, especially when it suggests that the facts referred to are mutually exclusive. A *coup d'état* can become a political revolution, and a political revolution can become a social revolution—if the new government decides to make a 'revolution from above' (Colburn 1994). Indeed, a social revolution is a change of state and society by means of violence that the new government mainly commits (Walzer 1998). And a civil war that gives not only radical revolutionaries but also parochial rebels a fighting chance, and ends in victory for whichever side best combines what Machiavelli called *virtù* and *fortuna*, can start because conservatives revolt in order not to make a revolution but to prevent one.

In sum, theories of revolution that would explain a revolutionary movement or a revolutionary situation need not explain a revolutionary outcome, but theories that would explain a revolutionary situation, a revolutionary outcome, or both have to explain a revolutionary movement as well. And whatever the facts that theories of revolution would explain, those facts entail violence—the intentional infliction of physical harm. The Bolshevik theoreticians who ridiculed nonviolent revolution as a contradiction in terms conformed to social science usage: no violence, no revolution—by any definition. All theories of revolution converge on the *modus operandi*—revolutionary violence. The question they all have to answer is what motivates revolutionary violence and what intended or unintended consequences follow from committing it.

2. Theories

Theories of revolution in Western social science date back to the Greeks, who viewed revolution as the

worst kind of war. Thucydides blamed revolutionary situations (with or without revolutionary outcomes) on revolutionary movements or conspiracies, and blamed those conspiracies on hope of success. Plato agreed and blamed hope of success on weak government torn by faction, which he thought to eliminate from his ideal state by abolishing the family and private property among the rulers. Aristotle added (on the widest though not deepest comparative study) that since hope along with faction springs eternal, no regime is immune against revolution.

The theory that people attempt revolution because they have hope of success, and have hope of success because they think the government is vulnerable, lived on in classic political sociology as an explanation of what to avoid and how to avoid it. Machiavelli, who wrote in effect the first handbook on counterinsurgency, told new rulers who had just seized power how to use force and fraud in order to avoid being overthrown. Hobbes, trusting more to organization than to statecraft, designed a regime that would prevent revolutionary movements and revolutionary situations (which he called sedition and civil war) by repressing faction. *The Federalist* took Hobbes for granted, but held that checks and balances could avert faction and insurrection as well as tyranny. Burke—who excoriated the French Revolution already in its ‘liberal’ phase and predicted the regicide 2 years, the Terror 3 years, and the military dictatorship 9 years in advance (O’Brien 1993)—observed that the revolutionary parliament usurped a king busy making concessions. As Tocqueville summed up the whole line of thought: ‘The evil that one suffers patiently as inevitable seems unbearable as soon as one conceives the idea of escaping it.’

Not all theorizing on revolution was counterrevolutionary. ‘Violence is the midwife of every old society pregnant with a new one,’ wrote Marx, though his instant histories of revolutionary situations autopsied abortions, and his class-conflict theory of revolutionary movements and their abortionists proved false—street fighters on both sides of the June 1848 revolutionary situation in Paris, for instance, not only had the same class background, but were a cross section of the manual work force (Traugott 1985). The theory explained history no better elsewhere—industrial workers neither launched the revolutionary movement nor achieved the revolutionary outcome that the theory predicted. In Germany they did not try, and in Russia they acquiesced as a revolutionary movement of Marxist intellectuals achieved a revolutionary outcome in their name by a military *coup d’état*—the October Revolution was so low-profile that the author of *Ten Days That Shook the World* slept through much of it—followed up with Red Terror and Civil War (Moore 1978, Pipes 1991, 1995). But for the vicarious, posthumous prestige Marx got from that much-imitated revolutionary outcome, whose official ideology-cum-state religion made him a household word,

no one except social scientists might have heard of him today (Kolakowski 1978).

Whether counterrevolutionary or revolutionary, however, theories of revolution were handmaidens to political sociology and historiography before the twentieth century, when—after the revolutions attending World War 1, World War 2, and decolonization—they hived off as a cluster of answers to the why-question of revolutionary violence. The chief theorists were not activists but academics because activists (even successful ones) propounded only trite theories of revolution consisting mainly of prophecy and propaganda. Where activist theories rose above agit-prop to consider strategic means and ends, they rehearsed the classic wisdom in reverse—they advised revolutionaries to play on faction in order to weaken the government and thereby raise hope of success. Some activists became revolutionary dictators, but more academics became theorists of revolution and launched a social science growth industry (Rule 1988, O’Kane 2000).

Academic theories of revolution, in the standard pilgrim’s progress view, went through three stages of development since the 1920s. First there were natural histories explaining recurrent sequences of similar events in ‘great’ revolutions by the reasons that experts say motivated the people who made them, but not generalizing the reasons as a causal theory. Then there were general theories explaining revolutionary violence by discontent, but accounting for history ad hoc if at all. Now there are comparative-historical theories explaining social and political revolutions by ‘structural conditions’ of state and society cogently enough that older theories look *passé* (Skocpol 1979, 1994, Goldstone 1986, 1991, Goldstone et al. 1991).

The standard view is not far wrong, though it ignores continuities of theory and method. All theories of revolution cite a short list of sufficient conditions or explanatory causes and validate, or rather illustrate, the explanation with factual evidence consisting of case histories and statistics arranged to show by the comparative method that where the explanatory causes occur, revolution (by some definition) occurs as well, and that where the explanatory causes do not occur, revolution (by that definition) does not occur either.

Grouped by the explanatory causes they cite, theories of revolution form four families: psychological, functional, structural, and political theories. As a rule, psychological and functional theories explain revolutionary movements—they explain why people attempt to change government, regime, or society (or all three) by violence; structural theories explain revolutionary outcomes—they explain why people successfully change government, regime, or society (or all three) by violence; and political theories explain revolutionary situations—they explain why people fight civil wars for state power. Psychological and functional theories that explain revolutionary

movements mostly leave revolutionary situations and revolutionary outcomes over to historians, whereas structural theories that explain revolutionary outcomes also explain revolutionary situations and revolutionary movements along the way, and political theories that explain revolutionary situations also explain revolutionary outcomes and revolutionary movements at least in passing.

Psychological theories explain revolutionary movements by frustration (alias relative deprivation or discontent) defined as a discrepancy between what people want and expect and what they get. These theories all assume the frustration-aggression principle that people who are frustrated commit aggression against the perceived source of frustration, but they disagree about what social facts cause frustration—one blames steady improvement followed by a sudden setback that raises expectations and reduces attainments; another blames rapid social change that raises expectations beyond attainments; yet another blames whatever raises expectations, lowers attainments, or both. Few psychological theories explain revolutionary violence by frustration alone, however; most add insurgent strength and government weakness to their short list of sufficient conditions.

Functional theories explain revolutionary movements by dysfunction (alias strain or disequilibrium) defined as a discrepancy between socially prescribed ends, goals, or values, and the available institutional means or procedures for attaining them. When social change creates a gap between ideology and experience that the government omits to close with reform, people turn to revolutionaries for salvation. If the government looks weak or loses control of its armed forces, or if revolutionaries think they can overthrow it, revolutionary violence erupts.

Structural theories explain revolutionary outcomes piecemeal—they explain first why élites subvert governments, then why subalterns rebel, then why radical vanguards seize power, and then why new rulers reconstruct state and society. The first two events make up revolutionary situations; the last two make up revolutionary outcomes. Each event depends on the one that precedes it—no subversion, no rebellion; no rebellion, no power transfer; no power transfer, no reconstruction. All these events comprise collective actions and reactions, which the theories explain as rational efforts to satisfy interests under structural constraints like government weakness, subaltern community organization, vanguard politico-military organization, and geopolitical exigency. Ideology is a cover for interest—the motivating constraints are structural.

Political theories likewise explain revolutionary situations piecemeal—they explain first why insurgents claim sovereignty, then why certain people support them, and then why the government fails to repress the insurgency. Revolutionary situations do not all have revolutionary outcomes, but some do, and

political theories also explain these outcomes piecemeal—they explain first why government supporters defect, then why insurgents gain armed force, then why government forces waver or desert, and then why insurgents take power. All these events comprise collective actions and reactions, which political theories explain as rational efforts to satisfy interests under political and strategic constraints like organization, armament, threat, and opportunity. Ideology again is a cover for interest—the motivating constraints are political and strategic. Whether and to what extent new rulers reconstruct state and society is another question, which political theories mostly pass over.

These family portraits of theories of revolution are not just composite sketches, they are caricatures, though caricatures accentuate real features. Thankfully, competent accounts of the theories as well as internecine debates among the theorists are legion (Rule 1988, Goldstone et al. 1991, Skocpol 1994).

Theories of revolution are contentious. Advocates of structural theory point out that psychological and functional theories explain neither revolutionary situations that occur because élite infighting paralyzes the government nor revolutionary outcomes that see new regimes more bureaucratic and tyrannical than the old (Skocpol 1979, Goldstone 1991). Advocates of political theory point out that revolutionary outcomes depend on the fortunes of war in revolutionary situations—structural *virtù* proposes, political *fortuna* disposes (Tilly 1993). And skeptics add that since revolutionary outcomes owe to actions that vanguards take because they expect these actions to yield benefits that often do not materialize, no theory that downplays ideology can explain revolution cogently (Aya 1990). How else to explain strategic decisions like Lenin's based on false expectations of imminent world revolution, the rage for socialism in poor countries ruled by revolutionaries, and the resort to mass terror by revolutionary governments in proportion to the innovation they attempt (Pipes 1991, 1995, Colburn 1994, O'Brien 1993).

Though theories of revolution are contentious, few of them disagree with Thucydides that revolutionary movements owe to hope of success or with Plato that hope feeds on weak government torn by faction. The wisdom of the ages is no less wise for being ancient.

See also: Collective Behavior, Sociology of; Conflict Sociology; Legitimacy: Political; Revolution; Revolutions, History of; Revolutions, Sociology of; Social Movements, History of: General; Social Protests, History of; Violence, History of; Warfare in History

Bibliography

- Aya R 1990 *Rethinking Revolutions and Collective Violence: Studies on Concept, Theory, and Method*. Spinhuis, Amsterdam, The Netherlands

- Colburn F D 1994 *The Vogue of Revolution in Poor Countries*. Princeton University Press, Princeton, NJ
- Goldstone J (ed.) 1986 *Revolutions: Theoretical, Comparative, and Historical Studies*. Harcourt, Brace, Jovanovich, San Diego
- Goldstone J 1991 *Revolution and Rebellion in the Early Modern World*. University of California Press, Berkeley, CA
- Goldstone J A, Gurr T R, Moshiri F (eds.) 1991 *Revolutions of the Late Twentieth Century*. Westview, Boulder, CO
- Kolakowski L 1978 *Main Currents of Marxism*, 3 vols. Clarendon, Oxford, UK
- Moore B Jr 1978 *Injustice: The Social Bases of Obedience and Revolt*. Pantheon Books, London
- O'Brien C C 1993 *The Great Melody: A Thematic Biography and Commented Anthology of Edmund Burke*. Minerva, London
- O'Kane R H T (ed.) 1999 *Revolution: Critical Concepts in Political Science*, 4 vols. Routledge, London
- Pipes R 1991 *The Russian Revolution*. Vintage, New York
- Pipes R 1993 *Russia Under the Bolshevik Regime*. Knopf, New York
- Rule J B 1988 *Theories of Civil Violence*. University of California Press, Berkeley, CA
- Skocpol T 1979 *States and Social Revolutions: A Comparative Analysis of France, Russia, and China*. Cambridge University Press, Cambridge, UK
- Skocpol T 1994 *Social Revolutions in the Modern World*. Cambridge University Press, Cambridge, UK
- Tilly C 1993 *European Revolutions, 1492–1992*. Blackwell, Oxford, UK
- Traugott M 1985 *Armies of the Poor: Determinants of Working-class Participation in the Parisian Insurrection of June 1848*. Princeton University Press, Princeton, NJ
- Walzer M 1998 Intellectuals, social classes, and revolutions. In: Skocpol T (ed.) *Democracy, Revolution, and History*. Cornell University Press, Ithaca, NY

R. Aya

Rhetoric

'Rhetoric' has different meanings depending on the context and whether used abstractly or concretely. The word, found with slightly different spellings in modern European languages, is derived from Greek *rhêtorikê*, which first appears in Plato's dialogue *Gorgias*, written about 385 BC. It is used there of the art (*tekhnê*) of a public speaker (*rhêtôr*), and may be defined as the art of persuasive public speaking; it can refer to the theory of discourse, to the art of orators and writers, or to speeches, writings, or sometimes other art forms (painting, sculpture, architecture, music, and dance) when these are intended to influence an audience. In scholarly contexts 'rhetoric' is usually a morally neutral term, but it has had influential detractors beginning with Plato and continuing with Descartes and Locke, with the result that in modern popular language or journalism the term often describes empty or misleading verbosity.

From the fourth century BC until the nineteenth century AD rhetoric was an academic discipline taught

in most western schools and universities, following grammar and preceding dialectic in the liberal arts trivium. Its function was to teach and practice students in techniques of public speaking or written composition in prose and verse. Thousands of rhetorical handbooks, treatises, and model compositions were written from ancient to modern times; many have survived in manuscript or print, but many were ephemeral. The Romantic movement, with its rejection of rule-based composition and celebration of individual genius, and the development of science, with its assertions of objective truths, undermined study of rhetoric in the nineteenth century, so that it often came to be thought of as limited to identification of linguistic tropes, such as metaphor and metonymy. In the twentieth century, however, rhetoric experienced a renaissance in connection with new communication and media studies, new literary theories, and new forms of linguistic analysis. Confidence in objective truth was challenged, and some thinkers came to believe that all speaking and writing, even discourse about the natural and social sciences, is a rhetorical act and necessarily contains rhetorical elements. The extension of rhetoric from its traditional concern with public address to embrace discourse of all sorts is called 'globalization' (Gross and Keith 1997), a term perhaps better reserved for another recent development, the comparative study of the rhetorical traditions of western and nonwestern societies. In contrast, 'rhetoric' is sometimes the name given to introductory courses in composition at the college level. In these courses students are practiced in a variety of forms of writing, including descriptive, narrative, and argumentative compositions together with attention to spelling, word choice, grammar, arrangement, and the use of figures.

1. The Historical Development of Rhetoric in Europe and America

Systematic study of rhetoric in Greece resulted from the need for effective public speaking by ordinary citizens under democratic governments, especially in Athens, in the fifth and fourth centuries BC. In the Athenian democracy litigants were expected to plead their own cases before large juries of fellow citizens, thus putting a premium on rhetorical ability, and the functioning of the council and popular assembly was predicated on extensive debate among citizens. The first teachers of rhetorical techniques were the traveling lecturers known as 'sophists,' of whom the most famous are Protagoras and Gorgias. In Plato's dialogue *Gorgias*, in opposition to Gorgias and his followers, Socrates takes a very negative view of rhetoric, which he represents as a form of flattery of an audience and as lacking concern with truth, but in Plato's later dialogue, *Phaedrus*, Socrates is made to describe an ideal philosophical rhetoric that would use

knowledge of logic and psychology for legitimate ends. By the late fifth century BC rhetorical handbooks could be bought and studied. They seem to have included examples of introductions, narratives, probable arguments, and conclusions in the form of commonplaces that readers could adapt to specific situations.

1.1 Aristotle's Theory of Rhetoric

Plato's student, Aristotle, developed some earlier teaching in greater depth and added many practical observations in his lectures, embodied in the treatise *On Rhetoric* (ca. 336 BC). Although it was rather little read until modern times, through the work of his students many of Aristotle's theories became part of traditional rhetorical teaching. Major concepts and divisions of modern rhetorical theory are derived from Aristotle. He defined rhetoric abstractly as 'an ability (or faculty: *dynamis*), in each case, to see the available means of persuasion.' A speaker uses 'nonartistic' means of persuasion, such as witnesses or documents, and 'invents' artistic means of persuasion, of which there are, Aristotle said, three and only three: they are often referred to as *ethos* (presentation of the character of the speaker as trustworthy), *pathos* (the emotions of the audience as aroused by the speaker), and *logos* (rational argument). There are two kinds of logical argument: the enthymeme, or rhetorical syllogism, is a deductive, probable argument in popular form; it usually takes the form of a statement with a supporting reason or of an 'if ..., then ...' hypothesis. The other type of argument is the paradigm, or example; it is an inductive argument, supporting a proposition with examples from history or analogy.

Rhetoric as Aristotle understood it dealt with probable knowledge in the realm of politics (though it was found also in speeches in dramatic and narrative literature), not with scientific or philosophical knowledge. The sources of enthymemes are of two kinds: specific topics, that is, propositions drawn from other disciplines, such as politics, ethics, psychology, economics, or military strategy; and dialectical or 'common' topics, that is, techniques of reasoning, such as definition, division, argument from cause and effect, etc. Aristotle defined three species of rhetoric: judicial (sometimes called forensic) rhetoric, practiced in courts of law, is concerned with determining the justice of an action in the past; deliberative rhetoric, practiced in formal or informal political assemblies, is concerned with determining the benefits or expediency of a proposed future action; epideictic (or demonstrative) rhetoric, practiced on ceremonial occasions such as a holiday or funeral, is concerned with what is honorable or dishonorable and takes the form of praise or blame of a person or thing. Many of Aristotle's rhetorical concepts are universals, applicable to discourse in all cultures; some need modification to apply to discourse more generally. *Ethos*, for example, has come to refer

to any moral factor or any presentation of character used in a discourse for persuasive purposes, not just the credible presentation of the speaker's veracity, and epideictic is now viewed as ceremonial or ritual discourse concerned with instilling, transmitting, or reinvigorating cultural values.

1.2 Greco-Roman Rhetorical Theory

Aristotle was primarily interested in what came to be known as 'invention,' the aspect of rhetoric concerned with the use of topics and means of persuasion, but in the third book of his treatise he began with some remarks on delivery, followed by an extended discussion of prose style and finally an account of the structural parts of a public address. Subsequent Greek and Roman rhetoricians outlined a theory of rhetoric in five parts which recapitulate the process of planning and delivering a speech: invention, or the planning of the argument; arrangement, or division of the speech into parts; style (the choice of words and composition of sentences); memory (or preparation for delivery); and delivery. Greek and Roman teachers of rhetoric were especially interested in judicial oratory and had much to say about the definition of the question at issue, which is known as the *stasis* of a case. Three main categories of stasis were recognized, with numerous subdivisions and many variants. Stasis of fact (also called conjectural stasis) deals with the situation in which a defendant denies committing the act with which he is charged (e.g., I did not kill him); in a case of legal stasis (also called stasis of definition) the defendant admits the act but claims it was legally justified (e.g., I killed him, but in self-defense); in stasis of quality the act and its definition are admitted, but extenuating circumstances are alleged (e.g., I did not intend to kill him). A topic much discussed by the rhetoricians was argument about interpretation of the letter and intent of a law or document, something of continuing interest in modern jurisprudence.

Greek and Latin rhetorical handbooks, of which a number have survived, generally describe the arrangement of a speech into a series of parts, each with its own function. The *prooemion* or introduction should make the audience attentive, receptive, and well disposed, and specific suggestions were offered to effect this. A narration should follow, giving a clear, brief, and persuasive account of the facts of the case. There is then often a statement, or proposition, of what the speaker hopes to prove, with perhaps division, or partition, into a series of headings. Next comes the proof, supported by witnesses and other evidence when possible, but often largely a more or less probable argument based on the circumstances and the character of those involved. Finally comes an epilogue, or peroration, in which the speaker summarizes the argument and seeks to stir the emotions of the audience.

The subject of style was much developed by rhetoricians after Aristotle; it is found discussed in rhetorical handbooks such as *Rhetoric for Herennius*, in Latin and dating from the early first century BC, and also in a number of monographic works, of which the most famous is the Greek treatise *On the Sublime*, probably dating from second century AD (the authorship of both works is uncertain). Discussions of rhetorical style dealt with the choice of words, the composition of sentences, both simple and periodic, appropriate rhythms for spoken prose, tropes, figures, and kinds of style. A trope is a substitution of one word or phrase for another; among the most common are metaphor, metonymy, synecdoche, and hyperbole. Figures are divided into verbal figures, sometimes also called grammatical figures, such as alliteration and anaphora, and figures of thought, such as rhetorical question and antithesis. Style was divided by many authorities into three kinds: plain, grand, and middle; others, of whom the Greek rhetorician Hermogenes (fl. ca. AD 180) is the best known, sought to describe a series of qualities such as grandeur, beauty, sincerity, sweetness, subtlety, and vehemence, that could be blended into a harmonious and effective whole. The works of the orator Demosthenes (384–322 BC) were commonly taken as the best examples of eloquence in Greek, the speeches of Cicero (106–43 BC) in Latin, and students were encouraged to analyze and imitate these models.

The rhetoricians developed mnemonic techniques to help their students memorize speeches. The commonest system was based on identifying words or ideas with visual symbols and imagining these against a familiar spatial background. Discussions of delivery dealt with control of the voice and the use of gestures.

By the first century BC the study of rhetoric had been introduced into Roman schools, although after resistance from conservatives; it always kept close to concepts taught by Greeks. The turbulent conditions of the late Roman Republic provided the context for vigorous debate in the senate, popular assembly, and lawcourts, which were the venues of the rhetorical masterpieces of Cicero. He also wrote dialogs and treatises on rhetorical theory in which the orator is celebrated as the ideal figure of civilized society, to be achieved by a man learned in philosophy, law, history, literature, and rhetoric and applied in public life for the common good. Cicero recast the Aristotelian means of persuasion into what he called ‘the duties of the orator’: to teach, to charm, and to move, and he identified these with the plain, the middle, and the grand style respectively.

The establishment of the Roman Empire by Augustus after 30 BC led to a decline in opportunities for political oratory, but schools of rhetoric strengthened their dominance of the educational system. In these schools much of the students’ time went into the practice of ‘declamation,’ writing and orally delivering speeches on imaginary historical and

legal themes, often expressed in highly artificial prose. The most extensive account of classical rhetorical theory and rhetorical education is found in *The Education of the Orator (Institutio Oratoria)* in twelve books, published by Quintilian in 95 AD. Quintilian followed Cicero in regarding the orator as the highest civic ideal, and repeatedly insisted that only a morally ‘good man’ (*vir bonus*) can become a ‘good orator.’ Early Christians, Saint Paul among them, show ambivalent attitudes toward rhetoric, but preachers often borrowed techniques from classical rhetoric. About AD 426 Saint Augustine, a teacher of rhetoric before his conversion to Christianity, completed an influential treatise, *On Christian Learning (De Doctrina Christiana)*, which is essentially a handbook of Christian rhetoric, applying Ciceronian teaching to the training of a Christian preacher and teacher.

1.3 Medieval and Renaissance Rhetoric

Although freedom of speech and opportunities for public address were limited under autocratic governments in late antiquity and the Middle Ages, some understanding of rhetoric was transmitted in encyclopedias of the liberal arts compiled by Martianus Capella (fifth century), Cassiodorus (sixth century), and Isidore of Seville (seventh century). In the later Middle Ages rhetoric found new application in handbooks of letter writing (*ars dictaminis*), verse composition (*ars poetriae*), and thematic preaching (*ars praedicandi*). The recovery of a reading knowledge of Greek in Western Europe in the fifteenth century, and the rediscovery by Italian humanists of important manuscripts on rhetoric from the classical period were major contributions to the vigorous intellectual life of the Renaissance. Rhetoric as taught in schools and universities in the Renaissance was primarily based on writings by Cicero and Quintilian, but some dissenting voices were heard. In the sixteenth century Peter Ramus, lecturing in Paris, transferred the teaching of invention from the course in rhetoric to the course in dialectic and limited rhetoric to the study of style, primarily the use of tropes and figures of speech. His position proved especially popular with Calvinists and Puritans; rhetoric as first taught in America (at Harvard College, founded 1636) was based on Ramist teaching. Meanwhile, textbooks of rhetoric began to appear in the vernacular languages; the first in English was *The Art or Crafte of Rhethoryke* by Leonard Cox (1535). A more profound understanding of the subject is found in Francis Bacon’s *Advancement of Learning* (1605), where ‘the duty and office of rhetoric’ is defined as ‘to apply reason to imagination for the better moving of the will.’ A split between a narrow view of rhetoric as a linguistic or literary phenomenon, largely limited to the use of tropes and figures in written language, and a broader view of rhetoric as a persuasive tool of political speaking and discursive writing

has persisted. The linguistic-literary approach has often characterized the teaching of rhetoric on the European continent and in English Departments in schools and universities; the view of rhetoric as concerned with political issues and with discursive writing flourished in eighteenth-century Britain and America and has its modern home in Departments of Speech, Communication, or Media Studies.

1.4 The Relationship of Rhetoric to Poetics and Dialectic in the Western Tradition

Although the purpose of rhetoric was traditionally assumed to be persuasion, and the purpose of poetry to be pleasure and enjoyment, rhetoric overlapped with poetics in that both had a concern with artistic composition: word choice, tropes, figures, and sentence structure. In addition, traditional rhetorical handbooks regularly discussed narrative techniques, also applicable to narrative poetry; and rhetorical theories of the presentation of character, arousing emotions, and the effective use of humor had some application to dramatic and other poetic genres. Rhetoricians insisted that their students study poetry as a source of diction, compositional techniques, and inspiration. The close relationship between rhetoric and poetics can be well seen in Horace's Latin didactic poem *The Art of Poetry*, written toward the end of the first century BC, describing the task of the poet as effectively mixing useful teaching with sweetness of expression, a view that reappears in the medieval *artes poetriae*, and in renaissance and early modern literary criticism such as Boileau's *L'Art poétique* (1674).

Rhetoric traditionally overlapped with dialectic as practiced in schools in the treatment of logical argument, including definition, division of a question, and the use of topics. The two disciplines differed in that rhetoric was usually regarded as embodied in a continuous discourse, whether oral or written, whereas dialectic took the form of a question and answer dialogue. Moreover, rhetoric dealt with a specific question, primarily of a political or judicial sort, while dialectic dealt with general questions or philosophical universals, and perhaps most important of all, in its effort to persuade an audience rhetoric utilized the appeal of character and emotion as well as logical argument, while in dialectic only logical argument was regarded as legitimate. A good discussion of the traditional relationship of dialectic and rhetoric is found in the fourth book of Boethius' *On Topical Differentiations* (*De Topicis Differentiis*), written in the sixth century and used as a textbook in the Middle Ages and Renaissance.

2. The 'New' Rhetorics

The thinkers of the Scottish Enlightenment of the eighteenth century were much interested in rhetoric

and in their lectures and writings combined traditional elements from Aristotle, Cicero, and Quintilian with their new understanding of psychology and consideration of modern conditions. Among important figures in this movement were David Hume, Adam Smith, George Campbell, and Hugh Blair. Writings of Campbell and Blair were translated into several languages and used in the United States and elsewhere as textbooks of rhetoric well into the nineteenth century.

In the twentieth century many new studies of rhetoric were published, based on historical, linguistic, literary, philosophical, political, and psychological research and influenced by the growth of mass media, the use of propaganda and manipulation of audiences for economic or political advantage, and new research methods, both quantitative and qualitative. New professional organizations were founded for those interested in rhetoric, among them the Rhetoric Society of America and the International Society for the History of Rhetoric. *The Quarterly Journal of Speech* and other older journals published an increased number of articles on rhetoric, and new journals devoted to rhetorical subjects began to appear. These include *Philosophy and Rhetoric*, *Rhetoric Review*, *Rhetoric Society Quarterly*, *Rhetorica*, and *Rhetorik: Ein internationales Jahrbuch*. Several dictionaries and encyclopedias of rhetoric, listed below in the Bibliography, have also been published recently.

Among twentieth-century writers on rhetoric, two have had special influence. Kenneth Burke developed his theories of rhetoric in numerous writings over a long career. In *Grammar of Motives* (1945) he described his best known tool of rhetorical criticism, the pentad of act, scene, agent, agency, and purpose. In *Rhetoric of Motives* (1950) he defined the function of rhetoric as 'the use of language as a symbolic means of inducing cooperation in beings that by nature respond to symbols,' and he had much to say about 'identification' between speaker and audience as a rhetorical technique. Burke, an American, came to the study of rhetoric from an interest in language, literature, and political philosophy; Chaim Perelman, a Belgian and author with Lucie Olbrechts-Tyteca of *La Rhétorique nouvelle* (1958), approached rhetoric from study of argumentation and jurisprudence. He distinguished argumentation, which is always addressed to an audience, from formal proof, which is not, and explored topics of argument, divided into those of quantity and quality and those that help to identify and exemplify facts.

An example of a narrow linguistic approach to rhetoric is *Rhétorique générale* (1970), the work of six scholars at the University of Liège calling themselves *Groupe µ*. Here rhetoric was limited to the phenomenon of troping as seen in metaphor and metonymy. Of many modern studies of metaphor, one that is sensitive to rhetorical issues is Paul Ricoeur's *The Rule of Metaphor: Multi-disciplinary Studies of the Creation*

of *Meaning in Language*. Another new approach to rhetoric is included in *A Theory of Semiotics* by Umberto Eco (1979). Eco described rhetoric from the point of view of semiotics as 'over-coding,' that is, a network of associations invoked by linguistic usages or by the switching of codes.

3. *The Natural Sources of Rhetoric*

Many traditional features of human rhetoric have analogies in animal communication. Social animals in particular use a variety of oral, visual, and chemical signs to persuade others to do what they want, primarily in connection with mating and securing food. Animal communication shares elements of invention, arrangement, style, and delivery with human rhetoric. Many animals perform epideictic rituals, and honey bees engage in a kind of deliberation when swarming. Some intelligent animals learn to understand and even to use metonymy, but the use of metaphor, as well as logical arguments based on citation of examples, is a human development. Rhetoric in some form is a basic feature of human and nonhuman animal life, a tool in the survival of both species and individual. It can be viewed as a form of mental and physical energy, reacting to a challenge, need, or desire by producing and transmitting signs to a real or imagined audience. The simplest natural rhetorical techniques, exploited by animals and humans alike, are gesture, volume, pitch, and repetition of a message.

4. *Rhetoric Without Writing in Traditional Societies*

Ethnologists have collected much information about the forms and function of speech and techniques of persuasion in traditional societies in Africa, Australia, the South Pacific, and the Americas, although they generally use terms like 'political language' rather than 'rhetoric' to describe what they have observed. Formal deliberation is a regular feature of acephalous societies, where individuals gain reputations as orators, and in hierarchic societies recognized orators are employed as spokesmen for kings and chiefs. The goal of deliberation in traditional societies is usually consensus, real or imposed, and opportunity should be offered to opponents 'to save face' by seeming to compromise. Formal language is required of speakers on official occasions everywhere and is learned by imitation of earlier speakers. Such formal language often has archaic elements of diction and grammar, as well as conventions of etiquette and arrangement, and makes use of proverbs, examples, and traditional metaphors. Epideictic speech, as understood in the West, has nonwestern parallels in religious rituals, where extreme forms of formal language, often in-

comprehensible to the general audience, are employed, frequently accompanied by music and dance. As in official deliberation, formal ritual language seems to certify the authenticity or truth of what is said and helps to perpetuate conservative social control over the society on the part of chiefs and priests. Among many examples of oral discourse Aztec *huehuehtlahtolli* from Mexico show an unusually complex development or rhetorical artistry. Many examples of Aztec speeches were written down by Bernardino de Sahagún and others in the sixteenth century when the genre was still flourishing.

5. *Rhetoric in Nonwestern Literate Societies*

The invention of writing in Mesopotamia and in China, and its spread to Egypt, Greece, India, and elsewhere clearly had some effect on rhetoric, though definition and proof of that effect is controversial. Scribes in Mesopotamia and Egypt celebrated their craft and obtained considerable power as necessary intermediaries between illiterate rulers and the general populace. Writing made possible communication over distance, record keeping, scientific and historical research, crystallization of oral poetic texts, creation of textbooks, and standardization of grammar. It may have facilitated extended logical argument and the use of complex sentences with a variety of subordinate clauses.

6. *Comparative Rhetoric*

The study of rhetoric on a cross-cultural basis is a recent phenomenon. It attempts to identify what is universal and what culturally specific in rhetorical practice across linguistic and cultural barriers, with the aim of formulating a General Theory of Rhetoric as well as to improve cross-cultural understanding. One approach is to test the applicability of western rhetorical concepts, including those outlined earlier in this article, to discursive practices in other societies. Of particular interest to date has been study of rhetoric in China, where thousands of speeches, as well as handbooks of composition, rhetorical exercises, and critical, historical, and poetic writings, dating from antiquity to modern times, provide a rich source for comparative research. One remarkable early thinker was Han Fei-tzu (born about 280 BC), sometimes called 'the Machiavelli of China' because of his pragmatic, cynical teaching to rulers about how to employ rhetoric to solidify their power. In the early centuries of the common era Chinese scholars produced moralizing and allegorizing commentaries on the Chinese classics that employ an extensive technical terminology, sometimes resembling but often differing from rhetorical concepts in the West. Comparable developments occurred in ancient India, where the *Arthashastra* of Kautilya (about 300 BC) outlines the

rhetoric of power and provides extensive technical terminology for teaching and criticizing rhetoric. Poetic and dramatic theory was later developed in India to a high degree of sophistication.

In comparison with other traditions, western rhetorical practice, beginning with the Greeks, has been more tolerant of contention, personal invective, and flattery, less insistent on consensus, politeness, and restraint than elsewhere. The Greeks invented the counting of votes as a way of settling political and legal disputes, something unknown elsewhere until modern times. Although voting performs that function, its potential acceptance of a majority of one as determinative can sharpen disagreement. Artistic judicial rhetoric, with argument from probability and exploitation of circumstantial evidence in court, is largely a western phenomenon; elsewhere, judges have traditionally interrogated the principals and have usually demanded direct evidence or trial by ordeal. Handbooks of public speaking were composed in ancient China, India, and Egypt—the earliest known is the *Instructions* attributed to the Egyptian vizier Ptahhotep of which the earliest manuscript dates from the early second millennium BC—but outside the West rhetoric has always been taught as an aspect of politics, ethical philosophy, or literary criticism, and not as a separate discipline within a standard curriculum of liberal arts, as found in western schools and universities.

7. *The Defense of Rhetoric*

Rhetoric has often been denigrated; the traditional defense, originating with Aristotle, argues that it is in itself neither good nor bad; like the other arts, it is a morally neutral tool that can be used by good or bad purposes. This is sometimes now labeled the 'weak defense' in contrast to a modern 'strong defense' that seeks to justify rhetoric on the ground that ethical and political knowledge is not based in *a priori*, abstract truth but is culturally formed and reformed through rhetorical engagement in concrete situations. The strong defense works best in a liberal democratic society where the majority grant some measure of toleration to minorities. On the global level it seems to justify any values and beliefs in society that are rhetorically constituted and accepted by an audience. That includes fascism, anti-Semitism, misleading advertising, chauvinism, religious bigotry, and most of the ills of history. At this level of description, we then have to fall back to defend rhetoric as a neutral tool that can be used equally for beneficial or for exploitive purposes.

8. *Current Issues in the Study of Rhetoric*

Rhetoric, as practice, was affected by the invention of writing and both as practice and as theory by the

invention of printing. The rhetorical implications of the electronic revolution of the late twentieth-century are only now beginning to be assessed. Rhetoric is, almost by definition, an interdisciplinary study. Throughout western history it has had a love-hate relationship with philosophy, and modern students of rhetoric draw extensively from twentieth-century philosophy in their analyses of rhetorical practices and in their efforts to show how rhetoric can be said to generate knowledge. Ideas derived from the epistemology and hermeneutics of Martin Heidegger and Hans-Georg Gadamer have been influential on some scholars, and others show indebtedness to Ludwig Wittgenstein, Michel Foucault, Jacques Derrida, Richard Rorty, Jürgen Habermas, or other recent philosophers, most of whom do not specifically discuss rhetoric. Much rhetorical research at the end of the twentieth century was concerned with issues of gender as rhetorically defined and practiced. Long overlooked evidence for speaking and writing by women has been discovered and is now being studied for the first time. The evidence from the classical period is fragmentary, but medieval, renaissance, and early modern women developed distinctive rhetorical forms, and the contemporary women's movement has employed a variety of rhetorical techniques of its own as well as analyzing the repressive rhetoric of the past.

See also: Aristotelian Social Thought; Language and Poetic Structure; Models, Metaphors, Narrative, and Rhetoric: Philosophical Aspects; Rhetorical Analysis

Bibliography

- Aristotle 1991 *On Rhetoric: A Theory of Civic Discourse*, translated by G A Kennedy. Oxford University Press, New York
- Bloch M (ed.) 1975 *Political Language and Oratory in Traditional Society*. Academic Press, London
- Brenneis D, Myers F R (eds.) 1984 *Dangerous Words: Language and Politics in the Pacific*. Waveland, Prospect Heights, IL
- Burke K 1945 *A Grammar of Motives*. Prentice-Hall, New York [Out of print, in print is the paperback edition, 1969, University of California Press, Berkeley CA]
- Burke K 1950 *A Rhetoric of Motives*. Prentice-Hall, New York [Only edition now in print is 1969 University of California Press, Berkeley, CA]
- Eco U 1979 *A Theory of Semiotics*. Indiana University Press, Bloomington, IN
- Enos T (ed.) 1996 *Encyclopedia of Rhetoric and Composition*. Garland, New York
- Fumaroli M (ed.) 1999 *Histoire de la rhétorique dans l'Europe moderne, 1450–1950*. Presses Universitaires de France, Paris
- Gross A G 1990 *The Rhetoric of Science*. Harvard University Press, Cambridge, MA
- Gross A G, Keith W M (eds.) 1997 *Rhetorical Hermeneutics: Invention and Interpretation in the Age of Science*. State University of New York Press, Albany, NY

- Groupe μ 1982 *Rhétorique générale*. Seuil, Paris
- Kennedy G A 1998 *Comparative Rhetoric: An Historical and Cross-cultural Introduction*. Oxford University Press, New York
- Kennedy G A 1999 *Classical Rhetoric and its Christian and Secular Tradition from Ancient to Modern Times*, 2nd edn. University of North Carolina Press, Chapel Hill, NC
- Lausberg H 1998 *Handbook of Literary Rhetoric: A Foundation for Literary Study*. Brill, Leiden, The Netherlands
- Lu X 1999 *Rhetoric in Ancient China, Fifth to Third Century B.C. E: A Comparison with Classical Rhetoric*. University of South Carolina Press, Columbia, SC
- Porter S E (ed.) 1997 *Handbook of Classical Rhetoric in the Hellenistic Period*. Brill, Leiden, The Netherlands
- Prelli L J 1989 *A Rhetoric of Science: Inventing Scientific Discourse*. University of South Carolina Press, Columbia, SC
- Ricoeur P 1970 *The Rule of Metaphor: Multi-disciplinary Studies of the Creation of Meaning in Language*. University of Toronto Press, Toronto, ON, Canada
- Simons H W (ed.) 1990 *The Rhetorical Turn: Invention and Persuasion in the Conduct of Inquiry*. University of Chicago Press, Chicago
- Sloane T O (ed.) 2001 *Encyclopedia of Rhetoric*. Oxford University Press, New York
- Sutherland C M, Sutcliffe R (eds.) 1999 *The Changing Tradition: Women in the History of Rhetoric*. University of Calgary Press, Calgary, AB, Canada
- Swearingen C J (ed.) 1999 *Rhetoric, the Polis, and the Global Village*. Lawrence Erlbaum, Mahwah, NJ
- Ueding G et al. (eds.) 1992 *Historisches Wörterbuch der Rhetorik*. Niemeyer, Tübingen, Germany
- Vickers B 1988 *In Defence of Rhetoric*. Clarendon Press, Oxford, UK

G. A. Kennedy

Rhetorical Analysis

Rhetoric, the *art of oratory*, was developed in ancient Greece and Rome. Both as a general communication theory and as a collection of practical advices on how to convince an audience, rhetoric exerted decisive influence on the history of Western civilization and culture. In the period from antiquity until the nineteenth century, rhetoric was the single most important subject taught in Western educational institutions. During these more than 2,000 years it changed both in scope and perspective. Corresponding with the growth in written communication, it was gradually transformed into an *art of writing*. In later centuries it became closely associated with the writing of literature while less emphasis was put upon argumentation and persuasion.

Rhetoric deals with communicative situations. Speeches and texts are always considered in their relation to the aims and intentions of the speaker/writer, the subject matter in question, the general mood and prior knowledge of the expected audience,

etc. Correspondingly, any kind of analysis which sets out to describe the ways in which a given discourse is determined or affected by its immediate communicative context could be said to be rhetorical in nature. Rhetorical analyses of this kind have always played an important role in literary studies, but during the few last decades similar methods have also been applied to nonliterary texts and even to nonverbal means of expressions, for instance in connection with studies of political propaganda, advertising strategies, as well as mass media in general.

1. The History of Traditional Rhetoric

1.1 Classical Antiquity

Some ground rules of oratory were developed and taught by Sicilians in 485 BC in connection with legal disputes about real estate. This rudimentary rhetoric quickly spread to Athens where a more comprehensive body of knowledge about the art of public speaking was developed from the mid-fifth century and onwards. Greek rhetoric was introduced in Rome in the first century BC as part of the education of older children.

In the Greek period, the so-called Sophists—persons who taught rhetoric, politics, and mathematics for money—played an important role in the development of rhetoric as a formal discipline. They were attacked by Plato for putting more emphasis on style than on substance. Plato's critique raises a question which has followed rhetoric ever since: Is rhetoric a method for discovering the truth, or merely an effective technique for manipulating an audience?

There is, however, no clear answer to this question. The old Greco-Roman rhetoric—as described by Aristotle (384–22 BC; see *Rhetoric*), Cicero (106–43 BC; *De Oratore*), and Quintilian (c. AD 35–c. 95; *Institutio Oratoria*)—was a complex social system of inter-related, and in some cases even contradictory, practices (cf. Barthes 1970, Vickers 1988). First of all, rhetoric was an art in the classical, Aristotelian sense of the word, i.e., *a craft or technique*, a collection of rules and advices for speakers who want to convince or persuade an audience. These rules were taught in and transmitted through strong educational institutions. Thus, rhetoric was also a discipline, a *subject*, a body of knowledge. In addition, this body of knowledge constituted a *theory*, or at least a semitheoretical field: it described and classified some important effects of verbal language.

1.2 The Rhetorical System

The old manuals divided rhetoric into five parts: *Inventio*, *Dispositio*, *Elocutio*, *Memoria*, and *Actio*.

Each part covers a specific field of knowledge and aims at training certain skills related to the field in question.

(a) *inventio* (finding what to say) is the part dealing with *argumentation*. Speakers learn to analyze a given subject matter, to find relevant information, and to determine which types of information are the most useful for their purposes. They also learn to convince an audience with quasilogical proofs (*logos*), and to move the audience by appearing to be good and trustworthy (*ethos*) or by stirring the audience's emotions (*pathos*).

(b) *dispositio* (organizing what one has found) deals with *composition*. Speakers learn to organize their speeches in clearly separated sections.

(c) *elucutio* (choosing the proper words) is the part that deals with verbal *ornamentation*. Some elementary rules are presented together with examples of striking images and other verbal devices which can be used to color a text.

(d) *memoria* (remembering what to say) is an introduction to *mnemonics*.

(e) *actio* (performing the speech) contains practical advices on how to *deliver* a text before an audience.

1.3 The Reduction of Rhetoric

The *memoria* and *actio* parts were rather sketchy in the old manuals, and because they were directly related to the practice of oratory, they disappeared as soon as rhetoric became a theory of written discourse. During the Middle Ages the tripartite system of *inventio/dispositio/elucutio* was taught together with grammar and logic. This system collapsed during the seventeenth century, and rhetoric ended up being a part of *poetics*. From now on it was further reduced: *inventio* and *dispositio*, the parts most closely associated with logic and argumentation, lost in importance and in the end all that remained of the old system was *elucutio*, the part dealing with questions of stylistics.

The *elucutio* sections of the old manuals taught the reader that significant persuasive or poetic effects can be obtained by selecting the proper words and combining them in the right manner. A flat statement of facts can be transformed into a strong, colorful, and convincing message by the use of *tropes* or *figures*, i.e., by substituting single words (*tropes*) or by manipulating phrases consisting of several words (*figures*). However, when eighteenth century linguists and literary theorists began to study and classify the many varieties of discursive ornamentation, they found that most of these poetic effects could be described as variations of two semantic relations, viz. *similarity* and *contiguity*. The similarity-relation finds its clearest expression in the *metaphor* while *metonymy* and *synecdoche* are well-known examples of ornaments based on causal and spatial contiguity-relations. From the eighteenth century and onwards, metaphor gradually became a kind of supertrope representing all the

other tropes and figures based on similarity-relations, while metonymy correspondingly became the representative of all the various ornaments based on contiguity-relations (Genette 1970).

2. Modern Rhetoric

2.1 Rhetoric and Semiotics

During the twentieth century several writers have sought to incorporate this reduced version of traditional rhetoric into a broader, more general framework. One of them is the Russian linguist Roman Jakobson who in 1956 presented a structuralist interpretation of *elocutio*, arguing that the metaphor/metonymy-dichotomy corresponds with a general bipolar structure in language (Jakobson 1956). All discursive production is the result of two fundamental linguistic processes based on similarity and contiguity. These metaphoric and metonymic poles in language can be observed in all normal linguistic behavior but, Jakobson argued, they are also manifest in any symbolic process, either intrapersonal or social (for a fuller discussion of the metaphor/metonymy-dichotomy, see the following section).

Jakobson thus opened the field of rhetoric towards sign systems other than language. His own examples were the dream work studied by psychoanalysts, the magic rites described by anthropologists, as well as nonverbal forms of art like cubist painting, the films of Chaplin and Eisenstein, etc.

Jakobson exerted decisive influence on French structuralists and semioticians during the 1950s and 1960s, among them Roland Barthes who in a seminal article, *Rhetoric of the Image*, proposed to rethink classical rhetoric in structural terms (Barthes 1964, see also *Semiotics*). Like Jakobson he argued that the immense variety of tropes and figures categorized by traditional rhetoric are merely surface manifestations of a few formal relations which can be shown to be the structural armature in all kinds of verbal discourse as well as in several nonverbal forms of expression. He predicted that it would be possible to establish a general rhetoric which would be valid for all types of discourses, for texts as well as for sound and images.

Since the 1960s there have been several attempts at implementing Barthes' project. The most comprehensive contribution is the work of the so-called *Groupe μ* , a group of Belgian researchers, who have presented a general description of rhetorical tropes and figures on a semiotic/structuralist basis (cf. Groupe μ 1970). In recent years the group has focused on the question of visual rhetoric (Groupe μ 1992).

The semiotic-structuralist interpretation of traditional rhetoric was mainly a theoretical project, but it had important practical and analytical consequences as well. In his 1964 article Barthes argued that there is

a close connection between rhetoric and *ideology*, a view supported by the fact that the classical manuals usually were more concerned with probabilities than with the actual truth of the matter. In order to be convincing one must know one's audience, know what people will or will not accept as probable, in short one has to respect *doxa*, i.e. public opinion, the stereotypical forms of thought which members of the audience rely on in their everyday lives. This means that the rhetorical devices used in a text can be understood as discursive expressions—or exploitations—of the dominant ideologies within a given society and that such ideologies can be brought to light by means of rhetorical analysis.

That rhetorical analysis can be used as a decisive critical tool had already been proven by the American critic Kenneth Burke who in 1939 wrote an influential article on the argumentative strategy in Hitler's *Mein Kampf* (Burke 1941, see also Burke 1962). From the 1960s and onwards Barthes and many other writers followed Burke's example and made rhetorical analysis a part of a general critique of ideology. Using the insights gained by the fusion of semiotics and rhetoric they made systematic studies of the patterns and systems of belief that are presented as 'natural' or 'evident' in newspapers, fashion magazines, popular films, television programs, etc. This form of analysis was particularly efficient in studies of advertising (for an overview see Dyer 1982).

2.2 Metaphor, Metonymy, and Mental Mapping

Elocutio was the only part that was left of the old rhetorical system by the beginning of the twentieth century. There was however a growing interest in this surviving part and particularly in metaphor. Not only did metaphor become one of the most studied subjects in literary theory, the interest spread to several other areas. At present the study of metaphor is one of the most fertile fields in the humanities and the social sciences. It plays a crucial role in academic disciplines as diverse as linguistics, film and media studies, constructivist psychology, anthropology, and cognition theory.

During the last century the conception of metaphor has changed dramatically. From being a poetic device producing a specialized type of discursive effect, it is now generally regarded as a very fundamental mental tool: 'the essence of metaphor is understanding and experiencing one kind of thing in terms of another,' as George Lakoff and Mark Johnson put it in their book *Metaphors We Live By* (1980). Metaphor is the kind of understanding which is produced when a set of relations is displaced from one kind of thing to another. Or to put it another way: Metaphor is a type of *mental mapping*. When people say that 'time is money,' they establish a simple relation of similarity between two elements but this process entails a further

mapping of a whole series of relations from the time-field onto the money-field: if time is money, then time must also be valuable, scarce, etc.

According to Lakoff and Johnson there are three major types of metaphor. One is the *structural metaphor*, the 'time is money' variety just mentioned: a restricted, almost local, structuration of one field in terms of another. Another one is the *orientational metaphor* which is used to give larger systems of concepts a spatial orientation. When for example happiness is described as 'being up' and sadness as 'being down,' a spatial relation (up/down) is mapped onto an emotional relation (happy/sad). Such orientational metaphors are clearly grounded in very concrete, physical experiences of the body and its position in space. According to Lakoff and Johnson all metaphors do actually have some kind of concrete experiential basis. One example is the third major type, the *ontological metaphor*, which is based on the everyday experience of handling physical objects. Ontological metaphors make it possible for an individual to understand—and thereby to deal with—events, activities, emotions, ideas, etc. as if they were objects, entities, or substances.

Metonymy, the other supertrope, does also have important cognitive functions. Its basic principle is that one entity *stands for* another. In the phrase 'sails on the water,' a part (sail) is used as a short hand reference to a totality (boat). Thus, a metonymy emphasizes or highlights certain aspects or features within a given totality. In the very act of referring, metonymy thereby serves some of the same purposes as does metaphor.

Metaphor and metonymy are mental tools which enable people to focus their everyday experiences and to understand the world they live in. They are, as Lakoff and Johnson point out, part of the ordinary way in which people think, act, and talk. Sometimes people use these tools in new and original ways, but most of the resemblances they perceive and most of the short-hand references they make in their everyday lives are the result of relatively conventional metaphors and metonymies that have become part of their conceptual system. In later years there have been a growing number of studies of such conventional metaphor/metonymy-based systems and the role they play in public debates and political communication. One striking example is George Lakoff's own analysis of the political rhetoric used by US politicians in the weeks before the war in the Gulf in 1991 (Lakoff 1991).

2.3 Rhetoric, Argumentation, and Thinking

Traditional rhetoric originated in relatively small, closed societies characterized by limited flows of information. Politicians and lawyers who should speak in public were faced with the task of convincing or

persuading uneducated and incompetent audiences. The manuals told them how to do that.

Communication today is still a question of finding and interpreting relevant information and presenting it in effective ways. Nevertheless, the general communicative situation has changed radically since the days of classical antiquity. The most important communication is mediated by radio, films, television, the Internet, etc.—media characterized by what has been called a new type of orality (Ong 1982). Information is distributed at an incredible speed, and the total amount of available information is growing rapidly. Even though the general level of education is considerably higher than anytime before in the history of the world, a modern mass audience is a rather heterogeneous phenomenon consisting of people with very varying degrees of knowledge and competence.

The fear that this complex situation with its rapid flow of information may result in massive manipulation of the audiences provides a significant background for many of the newer attempts at reviving rhetoric. One example is the so-called *New Rhetoric* which was developed in the mid-twentieth century (Perelman and Olbrechts-Tyteca 1958, Toulmin 1958). The New Rhetoric was initially an attempt at re-establishing the total rhetorical system in order to be able to cope with all aspects of the new, media-driven types of communication. However, from the very beginning the strongest emphasis was put upon *inventio*, and more specifically on *argumentation*, one of the parts of traditional rhetoric that more or less disappeared when rhetoric became associated with literature and poetics.

In the philosophical tradition argumentation has usually been associated with logic and the techniques of deduction and induction. What the New Rhetoricians have in mind, however, is not this strict, formal, and logical type of argumentation, but the ‘fuzzy’ thinking, the jumping-to-conclusions known from everyday life. Aristotle excluded this kind of thinking from logic proper but included it in his Rhetoric and actually considered it to be one of its most important parts.

Rhetorical argumentation is a ‘situated’ form of argumentation, it cannot be performed or understood without taking the total communicative situation into consideration. It is dependent on the subject matter, the level of available information, the pre-existing attitudes and prejudices of the speaker and the audience, the type of media which is used, etc. In short, what is at stake here and what the New Rhetoricians as well as many other, later writers (cf. for example Billig 1987) set out to describe and analyze are the various ways in which practical, everyday reasoning is determined by its immediate context.

Viewed like this rhetoric is much more than a theory of oratory or a collection of simple rules to be used by poets—it is a discipline which studies the underlying principles governing all human discourse, the way in

which people argue with themselves and with others, the way in which they make value judgments, etc. Rhetorical analyses of such arguments and of how judgments are made can provide a better understanding of how opinions and social attitudes are formed in complex modern media-driven societies.

See also: Advertising: Effects; Advertising: General; Jakobson, Roman (1896–1982); Language and Social Inferences; Mass Media, Representations in; Media Effects; Media, Uses of; Metaphor and its Role in Social Thought: History of the Concept; Organizations, Metaphors and Paradigms in; Political Communication; Political Discourse; Popular Culture; Reasoning with Mental Models; Rhetoric; Semiotics

Bibliography

- Barthes R 1964 Rhétorique de l'image. *Communications* 4 [1977 Rhetoric of the image. In: *Image Music Text*. Fontana Press, London]
- Barthes R 1970 L'ancienne rhétorique. Aide-mémoire. *Communications* 16 [1988 The Old Rhetoric. An aide-mémoire. In: *The Semiotic Challenge*. Basil Blackwell, Oxford, UK]
- Billig M 1987 *Arguing and Thinking. A Rhetorical Approach to Social Psychology*. Cambridge University Press, Cambridge, UK
- Burke K 1941 The rhetoric of Hitler's 'battle'. In: *The Philosophy of Literary Form. Studies in Symbolic Action*. Louisiana State University Press, Baton Rouge, LA
- Burke K 1962 *A Grammar of Motives and a Rhetoric of Motives*. World Publishing Company, Cleveland, OH
- Dyer G 1982 *Advertising as Communication*. Methuen, London
- Genette G 1970 La rhétorique restreinte. *Communications* 16: 158–71
- Groupeµ (Dubois J, Edeline F, Klinkenberg J-M, Minguet P) 1970 *Rhétorique générale*. Librairie Larousse, Paris [1981 *A General Rhetoric*. Johns Hopkins University Press, Baltimore, MD]
- Groupeµ (Edeline F, Klinkenberg J-M, Minguet P) 1992 *Traité du signe visuel: pour une rhétorique de l'image*. Seuil, Paris
- Jakobson R 1956 Two aspects of language and two types of aphasic disturbances. In: Jakobson R, Halle M (eds.) *Fundamentals of Language*. Mouton, The Hague
- Lakoff G 1991 Metaphor and war: The metaphor system used to justify war in the Gulf. In: Hallet B (ed.) *Engulfed in War: Just War and the Persian Gulf*. Matsunaga Institute for Peace, Honolulu, HI
- Lakoff G, Johnson M 1980 *Metaphors We Live By*. University of Chicago Press, Chicago, IL
- Ong W J 1982 *Orality and Literacy. The Technologizing of the Word*. Methuen, London
- Perelman C, Olbrechts-Tyteca L 1958 *Traité de l'argumentation; la nouvelle rhétorique vols. 1–2, 1st edn*. Presses universitaires de France, Paris [1969 *The New Rhetoric. A Treatise on Argumentation*. University of Notre Dame Press, Notre Dame, IN]

Toulmin S E 1958 *The Uses of Argument*. Cambridge University Press, Cambridge, UK

Vickers B 1988 *In Defence of Rhetoric*. Clarendon Press, Oxford, UK

P. Larsen

Rheumatoid Arthritis: Psychosocial Aspects

1. Introduction

Rheumatoid arthritis (RA) is a chronic, disabling inflammatory disease with pain and stiffness in the joints. RA with its often unpredictable and painful course has a considerable impact on patients' lives. The disease has physical as well as behavioral and psychosocial implications for the patient and their family. On the other hand, behavioral and psychosocial factors may be determinants of disease development and of patients' ability to adapt to the condition. Over the last two decades there has been an increasing interest in psycho-educational interventions such as patient education or cognitive-behavioral therapies to help patients deal with the consequences of their disease.

2. Disease Characteristics

RA is a chronic disease characterized by inflammation of joints and joint tissues (Wegener et al. 1996). Some patients may experience remission of the disease without any noticeable damage. As the disease progresses and becomes chronic, the inflammation may lead to joint damage and disability. Cardinal symptoms are joint pain and swelling, morning stiffness, malaise, and fatigue. The course of the disease is highly variable. Patients experience periods of remission and exacerbation of the inflammatory activity. Sometimes patients feel very good and in other periods they feel worse.

The prevalence of RA is about 0.5–1 percent of the population in most cultures. The disease can affect people of all ages, including children, but the onset is usually between the ages of 40–50 years. Women are more affected than men; the overall ratio is 3 to 1.

3. Psychosocial Factors Related to the Onset and Course of RA

There has been much research into the role of personality characteristics ('arthritis personality') and stress in the onset and course of RA. However, there is no empirical evidence for the hypothesis that certain personality characteristics predispose individuals to get RA. More recent research shows that psychological characteristics in RA patients are reactions to the physical disease process rather than causal factors.

There are indications that stressful life events are related to the disease process in RA. The mechanisms through which stress exerts its influence on the disease process are still unclear. Recent developments indicate that stress influences physiological processes. Relationships have been found between minor stressors (daily hassles, such as car breakdowns, arguing with spouse) and disease exacerbation. Recent evidence suggests that major life events (e.g., death, divorce) are associated with RA improvement (Huyser and Parker 1998). It appears that physiological processes mediate the relationships between environmental stressors and disease activity in RA. Psychological factors (e.g., coping style, social support) appear to moderate these processes. A greater understanding of the complex relationships between stress and RA requires integrative studies that simultaneously consider environmental, physiological, and psychological factors (Huyser and Parker 1998).

4. Physical, Psychological, and Social Consequences of RA

For the patient with RA it is probably less significant that stress influences the process of his or her disease than it is that RA with its often unpredictable and painful course has a stressful impact on the life of the patient.

4.1 Physical Consequences

Disability and pain are major outcomes of RA. Many patients are restricted in their mobility, both inside and outside of the house, and have problems with activities of daily living such as dressing, bathing, housekeeping, or cooking. They often need the help of others to perform these activities. Most patients see not being able to do things and dependence upon others as the main disadvantages of their disease (Cornelissen et al. 1988, Taal et al. 1993). Pain is generally regarded as one of the most important consequences of RA and a major challenge for patients as well as healthcare providers. For many patients, however, it seems that pain is not their main concern, but the restrictions in daily functioning and the lack of independence, probably for a great deal caused by the pain (Taal et al. 1993).

Another significant problem for people with RA is fatigue. Studies have found that the prevalence of fatigue in adults with RA ranges from 80 to 93 percent (Riemsma et al. 1998). Fatigue has multiple causes that include physiological, psychological, and environmental factors.

4.2 Psychological Consequences

Feelings of depression and anxiety are often mentioned as psychological consequences of RA. Studies

among RA patients have found heightened depression scores on self-report questionnaires such as the MMPI and the Beck Depression Inventory (BDI). Using a structured interview (Diagnostic Interview Schedule), Frank et al. (1988) found that 42 percent of the patients with RA in their study met criteria for some form of depression, and 17 percent met criteria for major depressive disorder. These rates of depression are higher than those found in community populations, but approximate the rates found in other chronic diseases. An important methodological problem in the assessment of the psychological impact of rheumatic disease is that many instruments such as the MMPI or the BDI were originally validated in psychiatric patients. Depression scales may overestimate the prevalence and severity of depression among persons with arthritis due to the endorsement of somatically related items such as 'I felt everything I did was an effort' or 'I am about as able to work as I ever was' (Wegener et al. 1996). Questions about physical symptoms may be accurate to assess emotional distress among psychiatric patients, but in arthritis patients these questions measure symptoms that reflect their actual physical condition.

4.3 Family Relationships

RA has a considerable impact on the patient's family. Patients can no longer fulfill all their family roles such as household chores, maintaining family ties, or shopping. Limitations in family role function are related to a diminished psychological well-being among people with RA. The disease affects the relationship with the spouse (see *Family Health*). This relationship does not always deteriorate, often spouses grow closer to the patient (Revenson 1993).

RA patients may have sexual problems, mainly caused by physical limitations, pain, and fatigue (Van Berlo et al. 1999). Stiff and painful knees and limited range of motion can hinder sexual intercourse. It has been found that RA patients have a decreased sexual motivation, but no decreases in frequency of sexual activity and intensity of orgasm compared to controls. Patients with RA experience greater loss of sexual satisfaction over time than controls, and greater functional disability is related to greater sexual dissatisfaction (Van Berlo et al. 1999).

4.4 Social Contacts

A chronic disease such as RA can lead to diminished social contacts (see *Social Support and Health*). Reduced mobility, loss of work, or the inability to be involved in hobbies or sports may cause this. It has been found that RA patients have fewer opportunities for contacts with friends and acquaintances, and less satisfaction with these relationships. Cornelissen et al.

(1988) established that a reduction of social contacts was mainly found among patients who lived in cities and less in patients living in rural areas. A decrease in social activities among arthritis patients was found not to be strongly related to the severity of the disease (Cornelissen et al. 1988, Taal et al. 1993).

4.5 Work Disability

Many RA patients experience problems in their work and often become work disabled (Wegener et al. 1996). After 10 years of disease, between 25 percent and 50 percent of patients with RA have stopped working, and after 30 years of disease about 90 percent of patients are work disabled. The most important individual risk factors for work disability among patients with RA are poor disease status, greater job physical demand, and older age. Other risk factors are lack of control over work pace and other work activities, less education, less support from family members and co-workers, and depression.

5. Behavioral and Psychosocial Determinants of the Adjustment to the Consequences of RA

The impact of RA on physical and psychosocial functioning is not directly related to the severity of the disease, but there is considerable individual variation in adjustment to the disease.

Several studies have demonstrated relationships between the use of certain coping strategies and psychological adjustment (Manne and Zautra 1992). Coping can be defined as a patient's cognitive, behavioral, and emotional efforts to manage specific internal or external demands. RA patients' use of emotion-focused coping strategies, such as wish-fulfilling fantasies and self-blame, is associated with poor psychological adjustment, while the use of problem-focused coping strategies, such as information seeking and cognitive restructuring, is related to good psychological adjustment. From studies on coping with pain, it can be concluded that the use of active and problem-focused coping strategies (e.g., diverting attention from the pain, increasing activity level, ignoring pain) is related to less pain and less depression, while the use of passive and emotion focused strategies (e.g., restricting social activities, praying, catastrophizing) is related to more pain and worse psychological adjustment.

Perceptions of control over one's situation, such as self-efficacy or learned helplessness, influence the way people cope and adapt to the stresses of their disease (see *Control Beliefs: Health Perspectives*). It has been shown that self-efficacy and learned helplessness are related to pain, disability, and psychological well-being in arthritis patients (Taal et al. 1996). Self-efficacy refers to beliefs in one's capabilities to success-

fully execute the behavior required to produce a certain desired outcome. Learned helplessness refers to the experience of having no control over one's life, in general, and across different situations. The unpredictable nature of RA and the varying disease activity may cause patients to view their disease as uncontrollable. This often leads to low self-efficacy expectations about needed causes of action to cope with the consequences of the disease or to more generalized feelings of learned helplessness.

Social support, for example from family and friends, can also play a significant role in the adaptation to the disease (see *Social Support and Stress*). A distinction must be made between structural aspects of social support, such as the size of the social network, and qualitative aspects, such as perceived social support. Perceived social support refers to the perceived comfort, caring, esteem, or instrumental help a person receives from others. In several studies with RA patients, it was found that more perceived social support was related to improved psychological well-being (Revenson 1993). In most studies, however, the size of the social network did not seem to have a strong relationship with psychological well-being. An exception is a study by Evers et al. (1997) among recently diagnosed RA patients. They found that patients who have a smaller social network shortly after diagnosis showed an increase in depressed mood after one year. They concluded that an extended social network which probably offers broad sources of support seems to be more important for the psychological adaptation of a patient shortly after diagnosis, when usual patterns of life are disturbed and many life domains have to be reorganized, than at a more progressive stage of the disease.

Social interactions do not only have positive effects on well being: negative aspects of social interactions can decrease well-being. Manne and Zautra (1989) found in a study with female RA patients that critical remarks made by the husband were related to maladaptive coping by the patient, which in turn led to a worse psychological well-being. Problematic aspects of social support (for instance support that is neither desired, needed, nor requested, or when the type of support does not match the recipient's needs) were found to be related to psychological depression (Revenson 1993).

6. Psycho-educational Interventions

Psycho-educational programs can teach patients the knowledge and skills they need to manage the consequences of their disease (see *Health Education and Health Promotion*). This should eventually lead to an improved physical, psychological, and social functioning. A wide range of behaviors can be important in the management of RA. Among the behaviors that are thought to influence physical and psychosocial

health status are adherence to medication regimens, physical exercise, relaxation exercises, joint protection measures, rest, pain coping strategies, and problem-solving (Taal et al. 1996). A well-known and successful American educational program is the Arthritis Self-Management Program (ASMP; Lorig and Holman 1993). In this program self-efficacy enhancing strategies, such as skills mastery and modeling, were used. Participation in this course resulted in improved self-efficacy, more consistent performance of health practices, reductions in pain, and improved physical functioning and psychological well-being among patients with various rheumatic diseases (see *Health: Self-regulation*). Changes in self-efficacy were related to changes in health status. A Dutch self-management program, partly based on the ASMP, showed beneficial short-term effects on knowledge, self-efficacy, performance of learned skills, physical function, and joint tenderness (Taal et al. 1996). After 14 months, effects were still found on knowledge, self-efficacy, and the practice of physical exercises, but most effects had disappeared.

Next to self-management programs, cognitive-behavioral therapy programs have been developed that are mainly aimed at the teaching of skills to cope with pain. Examples of skills that are taught in these programs are relaxation exercises, active coping strategies, such as diverting attention from the pain, and problem-solving skills (Keefe and Van Horn 1993). These programs have shown promising effects on the performance of coping strategies, pain, and physical and psychosocial functioning. Parker et al. (1995) evaluated an individual stress-management program for RA patients that included relaxation training and instruction in cognitive-behavioral coping strategies for managing stress and pain. Significant improvements were found on measures of learned helplessness, self-efficacy, coping, pain, and disability.

Long-term effects of psycho-educational interventions on pain, disability, or psychological well-being of RA patients are not often found (Keefe and Van Horn 1993, Taal et al. 1996). It is necessary to look for ways to improve the long-term benefits of psycho-educational interventions, for instance by including booster sessions in educational programs, teaching patients skills for the prevention of relapse (e.g., learning to identify high-risk situations and early warning signs, methods for coping with setbacks), or the improvement of social support through the involvement of patients' family members in educational programs (Keefe and Van Horn 1993, Taal et al. 1996). Radojevic et al. (1992) showed that cognitive-behavioral therapy with active involvement of a family member led to greater health benefits among RA patients than cognitive-behavioral therapy without family support.

It can be concluded that psycho-educational interventions that teach patients with RA self-management

skills and strategies for coping with pain and stress show promising results, but ways to improve long-term benefits have to be looked for. However, it must be emphasized that psycho-educational interventions for patients with RA can only be supplementary to medical treatment.

See also: Chronic Illness, Psychosocial Coping with; Chronic Illness: Quality of Life; Chronic Pain: Models and Treatment Approaches; Environmental Stress and Health; Pain, Management of

Bibliography

- Berlo W T M van, Vennix P, Rasker J J, Rijswijk M H van, Taal E, Weijmar Schultz W C M, Wiel H B M van de 1999 Rheumatic diseases and sexuality: A review of the literature. *Rheumatology in Europe* **28**: 113–7
- Cornelissen P G J, Rasker J J, Valkenburg H A 1988 The arthritis sufferer and the community: A comparison of arthritis sufferers in rural and urban areas. *Annals of Rheumatic Diseases* **47**: 150–6
- Evers A W, Kraaimaat F W, Geenen R, Bijlsma J W 1997 Determinants of psychological distress and its course in the first year after diagnosis in rheumatoid arthritis patients. *Journal of Behavioral Medicine* **20**: 489–504
- Frank R G, Beck N C, Parker J C, Kashani J H, Elliott T R, Haut A E, Smith E, Atwood C, Brownlee-Duffeck M, Kay D R 1988 Depression in rheumatoid arthritis. *Journal of Rheumatology* **15**: 920–5
- Huysen B, Parker J C 1998 Stress and rheumatoid arthritis: An integrative review. *Arthritis Care and Research* **11**: 135–45
- Keefe F J, Van Horn Y 1993 Cognitive-behavioral treatment of rheumatoid arthritis pain: Maintaining treatment gains. *Arthritis Care and Research* **6**: 213–22
- Lorig K, Holman H 1993 Arthritis self-management studies: A twelve-year review. *Health Education Quarterly* **20**: 17–28
- Manne S L, Zautra A J 1989 Spouse criticism and support: Their association with coping and psychological adjustment among women with rheumatoid arthritis. *Journal of Personal and Social Psychology* **56**: 608–17
- Manne S L, Zautra A J 1992 Coping with arthritis. Current status and critique. *Arthritis and Rheumatism* **35**: 1273–80
- Parker J C, Smarr K L, Buckelew S P, Stucky-Ropp R C, Hewett J E, Johnson J C, Wright G E, Irvin S W, Walker S E 1995 Effects of stress management on clinical outcomes in rheumatoid arthritis. *Arthritis and Rheumatism* **38**: 1807–18
- Radojevic V, Nicassio P N, Weisman M H 1992 Behavioral intervention with and without family support for rheumatoid arthritis. *Behavior Therapy* **23**: 13–30
- Revenson T A 1993 The role of social support with rheumatic disease. *Baillière's Clinical Rheumatology* **7**: 377–96
- Riemsma R P, Rasker J J, Taal E, Griep E N, Wouters J M G W, Wiegman O 1998 Fatigue in rheumatoid arthritis: The role of self-efficacy and problematic social support. *British Journal of Rheumatology* **37**: 1042–6
- Taal E, Rasker J J, Seydel E R, Wiegman O 1993 Health status, adherence with health recommendations, self-efficacy and social support in patients with rheumatoid arthritis. *Patient Education and Counseling* **20**: 63–76

Taal E, Rasker J J, Wiegman O 1996 Patient education and self-management in the rheumatic diseases: A self-efficacy approach. *Arthritis Care and Research* **9**: 229–38

Wegener S T, Belza B S, Gall E P (eds.) 1996 *Clinical Care in the Rheumatic Diseases*. American College of Rheumatology, Atlanta, GA

E. Taal

Ricardo, David (1772–1823)

Among the educated public only a few names of economists stand out: Adam Smith, the spokesman for capitalism; (T.) Robert Malthus, who earned for economics the title of ‘the dismal science,’ by virtue of his belief that any temporary gain in well-being would self-destruct by inducing population explosion that would entail diminishing returns and a retreat back to a subsistence ration of real wages; Karl Marx, the prophet of capitalism’s doom and the exponent of allegedly scientific socialism; John Maynard Keynes, proponent of the Mixed Economy, who rescued capitalism from the post-1929 Great Depression by activist government macroeconomic policies and public regulation.

Within the economics profession, it is a different story. David Ricardo was the archetypical classical economist, the economist’s economist who supposedly improved upon Smith and spelled out with Euclidean logic the eternal behavior of the market system. In amiable epistolary debates with his friend Malthus, Ricardo used to be reckoned the hands-down winner: Malthusian fears of recurrent depression stemming from under-consumption spending were argued to be fallacies. Like poet Robert Frost, beloved by both highbrows and lowbrows, the self-made millionaire David Ricardo captured the respect of the Victorian devotees to *laissez faire* as well as socialistic critics of Marxian or romantic stripe.

Schooled only in the Darwinian jungle of speculative finance, the broker Ricardo wrote as an autodidact, often in an obscure style that impressed his limited number of readers as being profound. Since 1817 literally scores of articles have appeared in learned journals debating whether or not David Ricardo did believe in ‘a labor theory of value.’ This is the stuff of which PhD theses can be made.

Ricardo was also lucky in having the learned Piero Sraffa edit for the Royal Economic Society, over a period of 25 years, a magnificent and complete edition in 11 volumes of *The Works and Correspondence of David Ricardo* (1951–73). Although this brought Ricardo’s name back into the limelight, oddly it seems to have contributed to a revisionist erosion of his analytic reputation. Whereas the late-Victorian Alfred Marshall, pillar of post-1870 neoclassicism, had idiosyncratically argued that the good modern stuff was

already in Ricardo, and whereas Keynes had written in the 1920s that David Ricardo's was 'the finest mind that had come to economics'—neither Ricardo's late-twentieth-century admirers nor detractors could agree among themselves as to what his different virtues and vices actually were. A scholar who is much misunderstood cannot be acquitted of all responsibility for ambiguity.

Ricardo's parents were Sephardic Jews who migrated to England from Amsterdam in the mid-eighteenth century. His ancestors had gone from Spain to Italy and then on to Holland. They seem to have been for the most part successful dealers in finance, an occupation that young David took up at the age of 14. When he married a Quaker, he was cut off by his father with the proverbial shilling; but within a score of years, working as a broker in the City of London, he had by his early forties amassed a comfortable fortune—the period of the Napoleonic Wars offered windows of opportunity for an alert and nimble financier. During vacation, when he picked up by chance Smith's *The Wealth of Nations* (1776), Ricardo was hooked for political economy.

Reversing the usual order of study, Ricardo first made his mark in what has come to be called *macroeconomics*. His unqualified verdict typifies his non-eclectic style. To finance a very long war against Napoleon, Britain expanded its money supply of paper currency in order to procure goods at home and abroad for its armies. Price-level inflation, not surprisingly, was part of the economic history in that period. Eclectic Smithian theory would no doubt have predicted as well a worsening of Britain's terms of trade as she and other belligerents bid up the relative prices of needed imports. Had gold and silver been the sole money used then in England and elsewhere, the 1750 theories of David Hume would have predicted a drain of specie from London to pay for the war-induced trade deficit of exports compared to imports.

However, even in those distant times much of the money employed for purchase and sale transactions involved use of paper-money currency. And, as is customary in times of war, whatever the supply of domestic gold or silver, there ensued a considerable rotation of the printing press to finance government needs and accommodate the desire for the enhanced cash necessary to conduct the enlarged volume of goods transactions and their higher average prices. Under these circumstances, whenever a government chooses *not* to redeem paper currency notes into specie at their previous face value, those notes will depreciate in price relative to ounces of specie. The high price of gold is what the person in the street calls this phenomenon and passionate voices debated its cause(s). Young Ricardo attributed *all* of the rise in gold prices to over-issue of paper notes.

Today one can write counterfactual history. Somehow the British government might have taxed the populace more heavily to finance in a balanced-budget

fashion the enhanced war effort and the irreducible rise in relative import prices; imposed tightness in bank lending, enforceable only at higher interest rates, conceivably could have kept the stock of paper currency unchanged, at the same time almost certainly imposing unemployment and extra loss of consumption on the civilian populace. With heroic tightness, trade balances might even have been forced to cancel out to zero. (It is hard to envisage how normal peacetime unemployment levels could have somehow been made to prevail from say 1795 to 1810 in this contrafactual scenario.) An inflation-free war can thus be painfully conjured up in contrafactual history; and with no wartime increment of paper-currency issue.

In actual history, paper notes did depreciate and price levels did soar. As happened a century later during the 1920–23 German hyperinflation, economists lined up in opposing camps. Young David argued: over-issue in currency M explained virtually *all* of the rise in the price of bullion. Henry Thornton, practical banker and Bank of England authority, (at first) argued that the microeconomic phenomena of adverse trade balances and war-induced buoyant business activity explained a significant fraction of the observed appreciation of bullion.

Ricardo's 1809–10 strong assertion of what was called the 'bullionist view' won him instant celebrity, particularly with the dogmatic James Mill. But this economic debate was not really a zero-sum game in which Ricardo's gain had to equal Thornton's loss.

Both men agreed on the likely longest-run outlook. The price of gold in terms of paper would be the same in 1850 if, in the many preceding decades, the supply of paper currency was limited enough to permit free specie-currency convertibility at the old 1790 ratio. The so-called 'Ricardian vice' was his use of *long-run* true relations to characterize actual short-run patterns in economic history.

After the first decade of the nineteenth century, Ricardo concentrated mostly on *microeconomics*. In 1815 (and *not* prior to then), along with Malthus and Edward West he helped articulate the 'law of diminishing returns' and its consequences for land-rent determination. Society's product, call it corn, is produced by labor and by land, both working together. When you increase labor in two *equal* increments, working on fixed land, you acquire two successive increments of corn. But the second gain is presumptively *less* than the first—a technological law of nature. Thus the wage that the competitive market can pay to labor goes down when labor/land density increases; and the rent that the market will pay for land will go up. This is not a new idea—it is in Benjamin Franklin (1755) early in the eighteenth century and already in Smith (1776)—but it is an important insight that needed being made explicit.

Already Ricardo had perceived what he considers faults in Adam Smith. Posterity can be grateful that

this perception tempted Ricardo into economics, even though it can be argued convincingly that Smith is mostly the more correct thinker in the cases where Ricardo criticizes him. For example, in the capital-and-land-and-produced-capital models that both writers worked with, Smith was right to jettison the ‘labor theory of value’ as soon as the interest rate and/or the land rent rate becomes significantly positive.

As between the two tasks of political economy—(a) understanding how relative scarcities of factors of production constrain the alternative availabilities of different goods and their valuation relations; and (b) how society’s harvest gets *distributed* among and between the different social classes of workers and owners of property—Ricardo arbitrarily declared it to be distribution that was all important. Yet nowhere in the 11 volumes of his *Works and Correspondence* (1951–73) will the reader find a *complete* and satisfactory theory of distribution that goes beyond what is in Smith, Malthus, Mill or modern mainstream writers.

Perhaps Ricardo’s lack of a university education contributed to his methodological style. Frequently he asserts general ‘truths’ that do admit of important exceptions; repeatedly he fails to distinguish between *necessary* and *sufficient* conditions for an argument. He is a superb miniaturist but with an absence of coherence, particularly in his post-1815 textbook phase. (Ricardo’s theory of comparative advantage in international trade, similar to that of the contemporary Colonel Robert Torrens, is a brilliant example of his miniaturist virtuosity.)

If it be thought that, by use of hindsight, I am a bit harsh on 1810–23 David Ricardo, let it be understood that I am even harsher on Ricardo’s twentieth-century commentators: the brightest and the best of them—Piero Sraffa, George Stigler, Samuel Hollander, Mark Blaug and scores of others—while understanding that the positive interest and time-phasing of production do generically invalidate Ricardo’s championing of the *labor theory of value* (as, indeed, Ricardo himself intermittently admitted), at the same time seem blind to the parallel fact that positive rent(s) on scarce land(s) must similarly invalidate the labor theory of value. (Edwin Cannan, Knut Wicksell, Jacob Viner and Lionel Robbins are honorable exceptions to this indictment of blindness about how labor/land intensities do generically invalidate any labor theory of value in the same logical way that time-intensities are known to do so.)

Around 1815, between his macro works on gold and his *Principles of Political Economy and Taxation* (1817, 1819, 1821), Ricardo did puzzle over a simple scenario involving a farm good (corn) and one or more manufactured goods (cloth, ...), all produced by, say, homogeneous labor, homogeneous land and by one or more *produced* inputs (corn seed, raw materials from farming used for manufactures, durable tools). He may even have worked out, in an unpublished manu-

script that has not survived, a one-sector farm-only model. In it, a determinate model of distribution could have cogently defined long-run stationary-state equilibrium for (real corn wage rate, interest rate, real corn rent rate) = $(W/P_{\text{corn}} = \bar{w}, \bar{r}, \text{Rent}/P_{\text{corn}} = R^*)$. How would it have gone?

Like Malthus and other predecessors, Ricardo would have specified *exogenously* a classical ‘subsistence wage’ rate, $\bar{w}$, needed to keep population just reproducing itself. Like classical successors he would have specified *exogenously* a minimal interest rate, $\bar{r}$, needed to keep capital(s) just reproducing themselves. Land being a classical constant, $\bar{A}$, its stationary equilibrium rent rate he would *residually* determine once variable labor-and-capital stocks, L_t and K_t , dynamically converged to their long-run, asymptotic stationary equilibrium levels: $L_t \rightarrow L^*$; $K_t \rightarrow K^*$, when K is a single scalar. Finally, short-run determination of w_t and r_t would be market-clearing rates dependent on the relative transitional abundances of L_t and K_t .

Here is how in the spirit of Prokofiev’s ‘Classical’ Symphony, an ambitious neophyte in a 1932 Hicks–Wicksell *neoclassical* workshop might synthesize these classical insights into a coherent simplistic dynamic equation system, using the notational convention $dY_t/dt = \dot{Y}_t$:

factor supplies: $\dot{L}_t/L_t = |\bar{\alpha}|(w_t - \bar{w})$,

$$\dot{K}_t/K_t = |\bar{\beta}|(r_t - \bar{r}), A_t \equiv \bar{A} = 1 \quad (1)$$

factor demands: $w_t = \bar{\sigma} K_t^{1/3} / L_t^{2/3}$,

$$r_t = \bar{\sigma} L_t^{1/3} / K_t^{2/3} \quad (2)$$

where $\bar{\sigma}$ is an exogenous parameter denoting level of technical productivity.

When the modern crank of deduction gets turned, Ricardo’s (1) and (2) do lead convergently to a determinate classical stationary state. From any *initial* $(L_0, K_0, \bar{A})_0$, as $t \rightarrow \infty$,

$$(w_t, r_t; L_t/\bar{A}, K_t/\bar{A}; R_t) \rightarrow (\bar{w}, \bar{r}; \frac{1}{3}\bar{\sigma}/[\bar{w}^2\bar{r}], \frac{1}{3}\bar{\sigma}/[\bar{r}^2\bar{w}]; \frac{1}{3}\bar{\sigma}[r, w]). \quad (3)$$

One must commend the classical tradition for so subtle an accomplishment. (Note the version worked out here differs from the Samuelson (1978) ‘canonical’ model, which had more faithfully adhered to the conventional labor-capital ‘dose’ version.)

This above supply–demand equilibrium is definitely *not* independent of the problem of how consumers wish to allocate their earned incomes among the many goods, when there are many non-corn goods in the picture. Then in it there are no ‘natural prices’ definable independently of how consumers choose to spend their incomes.

Definitely goods’ relative price ratios do not remain close to proportionality with their respective embodied-labor contents, $P_t^i/P_t^j \neq (L_t^i/q_t^i)/(L_t^j/q_t^j)$.

Likewise embodied land contents $(A_t^i/q_t^i)/(A_t^j/q_t^j)$ and embodied capital contents $(K_t^i/q_t^i)/(K_t^j/q_t^j)$ similarly fail as approximations to empirical actual P^i/P^j ratios.

Since Ricardo did not coherently understand the intricacies of his own classical scenario, he never realized how unattainable was his hankering to have all relative prices determinable purely by *objective technology and cost data*. Subjective variabilities (a) of consumers' utilities and demand choices and (b) of intertemporal saving-and-consuming choices—which became central preoccupations of post-1870 neo-classical successors to the classical school—were thus *already* unavoidable in the classical school's own capital, land and labor models and in the competitive scenarios that Marx grappled with all his working life. This inevitability will come with pain for the small sect of Sraffian neo-Ricardians who proliferated after 1925 and especially 1960.

Ricardo was low-keyed and civil in argumentation, albeit firm in holding to his strong positions. The letters between Ricardo and Malthus, notwithstanding these scholars' deep differences, are a model of courtesy and affection. Despite Ricardo's typical unqualified dicta, he did not hesitate to change his mind and admit to previous mistakes. Marx, who abhorred classical authorities and for whom 'Parson' Malthus was a particular *bête noir*, was uncharacteristically soft toward Ricardo. Here is an illustrative incident. After Ricardo's first edition of *Principles*, but prior to its third edition, David, Humpty-Dumpty like, had declared that technical invention *must always* help *all* factors of production, including the working man. He encouraged disciples, like J. R. McCulloch, to affirm this strong view. However, on reconsideration, Ricardo came to realize that some inventions can assuredly *replace labor and reduce the demand for labor in the short and the long run*. Therefore, despite friends' warnings that this would weaken the case for *laissez faire*, Ricardo added a new chapter on Machinery to the third (1820) edition of the *Principles*, presenting exposition and numerical examples designed to show how labor's share of national income could suffer from certain possible inventions and how long-run equilibrium populations and gross-income levels could be permanently lowered. With virtually no exceptions, ancient and twentieth century authorities agreed that for once Homer had nodded. For once David must have invoked rigidity of wage rates with its resulting unemployment, thus for once denying J.-B. Say's sacred dogma that over-production is impossible in a capitalist system. Karl Marx praised the scholar for his honesty.

It is a story that casts discredit on virtually all. The Ricardo–McCulloch original sweeping position was gratuitously wrong. A new wind that raises an economy's *potential* for production and consumption most certainly can do harm to *some* competitors while helping others. (Ricardo did not, as J. S. Mill and

Vilfredo Pareto were later to do, argue that gainers could always bribe losers so that *with intervention all could gain*.) Nor did Ricardo's new understanding motivate him to abandon a dogmatic defense of *laissez faire*. His critics on Machinery, both then and in our time—including Knut Wicksell, Nicholas Kaldor, Joseph Schumpeter, George Stigler, ...—failed to see the absence in Ricardo's exposition of any denial of dogmatic Say's Law and its ruling out theoretical harm from under-consumption; instead his usual classical scenario of population decline dictated by the induced drop in the wage below an alleged equilibrium *subsistence* rate would, as the Machinery chapter alleged, call for an ultimate drop in population and gross product as a result of labor-saving invention(s). Just as the industrial revolution was beginning to raise wages by a genuine filter-down process, the philosophers of the Chair were getting first intimations of the complexity of how a market system metes out its distributive awards.

In that long ago age before antibiotics, David Ricardo in 1823, at the peak of his powers, died from an ear infection. Thus he never lived to see the 1836 Repeal of the Corn Laws, a vindication of his pamphleteering for free trade; and he did not live to witness Ireland's great famine of the 1840s, with its brutal confirmation of the laws of the free market.

In summary, David Ricardo, like the twentieth-century economists Friedrich Hayek and Milton Friedman, importantly pushed voters and public opinion toward libertarian *laissez faire*. Victorian England's Whig society and Manchester School owed much to him. In the modern debate about how a modern mixed economy can optimally compromise between market mechanisms and democratic rules of the road, Adam Smith and John Stuart Mill seem more subtle classical thinkers. From these two, rulers like Lenin, Stalin and Mao could have had the most to learn. But among the worldly philosophers, David Ricardo must be counted as an important shaper of the twenty-first-century mind.

See also: Asset Pricing; Emerging Markets; Economic Growth: Measurement; Economic Growth: Theory; Economics, History of; Economics: Overview; Income Distribution; Innovation and Technological Change, Economics of; Libertarianism; Malthus, Thomas Robert (1766–1834); Market Structure and Performance; Marx, Karl (1818–89); Marxian Economic Thought; Monetary Policy; Political Economy, History of; Regulation, Economic Theory of; Smith, Adam (1723–90); Stock Market Predictability

Bibliography

Franklin B 1755 *Observations Concerning the Increase of Mankind and the Peopling of Countries*

- Keynes J M 1936 *The General Theory of Employment, Interest and Money*. [Reprinted in: 1977–81 *The Collected Writings of John Maynard Keynes*. Macmillan, London]
- Malthus T R 1798 *An Essay on the Principle of Population as it Affects the Future Improvement of Society, with Remarks on the Speculations of Mr. Godwin, M. Condorcet and Other Writers*. J. Johnson, London [Reprinted by Macmillan, London, 1926]
- Marshall A 1890–1920 *Principles of Economics*. Macmillan, London
- Marx K 1867, 1885, 1894 *Capital*, Volumes I, II, and III. Verlag von Otto Meissner, Hamburg; [Penguin Books, Harmondsworth 1976, 1978, 1981]
- Niehans J 1990 *A History of Economic Theory. Classic Contributions, 1720–1980*. The Johns Hopkins University Press, Baltimore
- Ricardo D 1817 *Principles of Political Economy and Taxation*. John Murray, London
- Ricardo D 1951–73 *The Works and Correspondence of David Ricardo* [ed. P. Sraffa with M. H. Dobb] 9 vols. Cambridge University Press, Cambridge, UK
- Samuelson P A 1978 The canonical classical model of political economy. *Journal of Economic Literature* 16: 1415–34
- Samuelson P A 1989 Ricardo was right! *Scandinavian Journal of Economics* 91: 47–62
- Schumpeter J 1954 *History of Economic Analysis*. Oxford University Press, New York
- Smith A 1776 *An Inquiry into the Nature and Causes of the Wealth of Nations* [ed. E. Cannan]. The Modern Library, New York, 1937
- Sraffa P 1926 The laws of returns under competitive conditions. *The Economic Journal* 36: 535–50
- Sraffa P 1960 *Production of Commodities by Means of Commodities*. Cambridge University Press, Cambridge, UK
- Thornton H 1802 *An Enquiry into the Nature and Effects of the Paper Credit of Great Britain*. Reprinted, Frank Cass, London; Augustus Kelley, New York, 1962

P. A. Samuelson

Rights

No concept is more central to our moral and political thought than that of rights. We talk about our rights to life, liberty, and freedom of expression; we discuss the rights of children, animals, illegal immigrants, and prisoners of war; we wonder whether we have rights to health care, a decent wage, or physician-assisted suicide. We seem to make sense of a myriad of legal and moral rights, even as we debate their boundaries. Yet few concepts so central to our thought are more deeply puzzling. The clearest evidence for this is the persistent disagreement over the content of even the most basic rights. Is a right to a free press a basic human right? Many people around the world would deny that it is. Is there a basic or human right to health care? Who has the duty to provide it, and on what grounds? Do parents have a right to discipline their children? Do workers have a human right to ‘periodic

holidays with pay,’ as one United Nations document asserted? The problem is not simply that such assertions are controversial, but that we can feel so quickly at a loss about what to say next. Clearly, the concept of rights has been debased in recent years, as rights claims have become mere attempts to give rhetorical force to strongly held preferences, but the problem also lies in some deeper puzzles about the nature and justification of rights.

The least controversial rights are those created by laws and the shared norms of our institutions and practices, where rules allow us to coordinate and regulate our activities and to divide fairly the benefits and burdens of social cooperation. Any attempt to understand the concept of rights should begin here. If Smith gives tools to Farmer in exchange for a share of the grain Farmer will produce, then Smith has a claim to the grain after the harvest, and Farmer has a duty to turn it over. It seems natural to say that Smith has a right to the grain, and laws will respect and enforce such rights.

Governments have a right to tax their citizens to provide public goods, but laws also protect citizens by creating rights against other citizens and against the government itself. Civil society is founded on the concept of property, which is nothing more than a set of practices that allocate entitlements—in other words, rights to use and exchange things of value. Many of the functions of government create rights in this sense, by establishing entitlements and enforcing agreements.

1. Origin of the Concept

If it is possible to have a concept without having a word to denote it, as one view of rights maintains, then the idea of rights goes back to primitive societies and can be inferred from ancient texts. According to this view, any society with an institution of property understands the ideas of entitlements to use and claims against infringement, which are part of a theory of rights. Ancient laws also established other liberties and privileges. The Roman law of persons, for example, recognized different status for different members of society. The citizen, the resident alien, and the slave were all granted different powers and liberties, and the law determined when these people could change or lose their status. Romans thus had (and could forfeit) various personal rights.

A different view argues that rights did not exist until sometime in the late Middle Ages, when we first find a word to denote the concept (Tuck 1979). There is an important difference between the implicit, ancient understanding of practices that we regard as rights and this more explicit idea. According to this view, the ancient meaning is restricted to an objective sense of right, which refers to a state of the world. Ancient and medieval societies were dominated by ideas of the

natural differences between people. If someone was properly a citizen or a slave, or later a lord or a servant, their social rank determined their powers and privileges. Their status was meant to reflect the objective or correct natural order of the world. The concepts that come closest to describing this objective sense of right—*dikaiois* in Greek and *ius* in Latin—imply an idea of a proper natural order.

The modern notion of rights, as we might call it, developed as the idea that the natural or inherited differences in status could be justified was replaced by a belief in the fundamental moral equality of all people. Historians note that this idea of human equality was not new in the late Middle Ages. It appears in the writings of the stoics and early Christians, but these ancient philosophies did not attach much importance to earthly life. As traditional social orders began to crumble in Europe, the natural differences between people came to be seen as morally arbitrary. Justice required correcting these arbitrary differences with a more egalitarian distribution of powers and opportunities. As moral thought focused on the more worldly concerns of social and political interactions among people, the subjective idea of a right emerged, as something that belongs to individuals. Thus, where ancient and medieval philosophers would say that it is right that Smith controls some thing, modern theorists would say that Smith has a right to that thing (Dagger 1989, Tuck 1979).

Rights in the subjective sense express the moral importance of the equality of persons. They justify both the claims we make against others and the duties that arise in respect of those claims. But this idea of rights is no less puzzling than that of a natural moral order. Do rights exist? What is their content and justification?

2. An Analysis of Legal Rights

The revolutions in the USA and France in the eighteenth century were fought in part as a defense of the natural and inalienable rights of humans. Liberal republics and democratic governments based on these rights were established, first in the USA and Europe, and then elsewhere. By that time, the modern idea of rights had fully taken hold in law, morality, and political theory.

Early in the twentieth century, Wesley Hohfeld categorized legal rights in a way that has profoundly influenced subsequent moral and legal theories (Hohfeld 1923). He argued that an assertion of the form ‘Smith has a right to P’ is ambiguous and could mean any of four different things (or some combination of them). It could mean, first, that Smith has a *privilege* to do P, or that he is under no obligation not to do P. Smith’s right to find solace walking in a public park is a privilege, or as it is sometimes called, a mere liberty. It does not imply any corresponding duties on the part of others. The fellow on roller blades who

insists on circling Smith in the park, with his boom box blasting, is also exercising a privilege, even though his activity undermines Smith’s efforts to find solace. Hobbes’ claim that in the state of nature ‘every man has a right to every thing’ has been interpreted as a privilege in Hohfeld’s sense. Hobbes was asserting merely that in the state of nature a person could not be criticized as irrational for laying claim to any thing, although nobody else has a duty to provide him with any thing that he might claim (Waldron 1984).

Hohfeld’s second meaning of ‘Smith has a right to P’ is that Smith has a *claim* to P, which imposes corresponding duties, either positive or negative, on other people to allow or enable Smith to have P. Legal theorists following Hohfeld analyze claims further by distinguishing between rights *in personam* and rights *in rem*. This distinction has implications for the kinds of duties implied by a claim. Claim rights *in personam* impose duties on particular individuals. A contract is such a right. If Smith agrees to work for Miller, then when Smith completes the work, he has a claim that requires Miller to pay the agreed wage. Claim rights *in rem* impose duties generally on all others. Property rights are an example. If Smith owns a car, then every other person has a duty not to interfere with Smith’s use of the car. Each person’s right not to be assaulted is also a claim right *in rem*, which means that in respect of this right, each person imposes a duty on all other people not to harm him in certain ways. In order for such rights to seem plausible, the duties must typically be negative, i.e., duties to refrain from doing something. Unlike positive duties, the millions of negative duties we impose on each other in respect of our rights *in rem* do not create any unreasonable burden. We can fulfill all of these duties simply by minding our own business and not assaulting anybody.

Hohfeld’s two remaining categories of rights apply more specifically to legal contexts. ‘Smith has a right to P’ could mean that the law gives Smith certain *powers* to alter legal relationships, which do not entail corresponding duties. Smith exercises such powers, for example, when he sells or transfers property. Finally, ‘Smith has a right to P’ can mean that Smith has an *immunity* from certain kinds of legal change, i.e., he is protected from any or all other authorities altering his legal position. Constitutionally guaranteed privileges or inalienable rights are examples of immunities. Hobbes’ argument, that when humans enter into civil society and give up their rights to the Leviathan they nevertheless retain the right to defend their life under all circumstances, is a combination of a privilege and an immunity.

3. The Nature of Moral Rights

That legal systems and other institutions give rise to rights is unproblematic, but rights also play a more fundamental role in moral and political theory, where

they are often regarded as existing independently of and prior to social institutions. The Declaration of the Rights of Man and the Citizen in France states that 'the end in view of every political association is the preservation of the natural and imprescriptible rights of man,' and the Declaration of Independence in the USA asserts that all citizens have inalienable rights to life, liberty, and the pursuit of happiness. Natural rights constitute the moral basis and justification of government and laws. They express and give meaning to the belief that all people are morally equal.

Claims to natural rights have always been controversial, however. Jeremy Bentham (1834) famously dismissed them as 'pernicious nonsense,' and wrote that 'a natural right is a son that never had a father.' The controversy is about the nature of basic moral rights and their connection to duties, as well as about the justification of basic or natural moral rights. We will address these issues in order.

The defender of basic moral rights need not assert that all moral duties are correlated to rights. Kant, for example, discusses imperfect duties, such as the duties to be charitable or kind. No individual can claim the performance of such duties as a right, and governments may choose not to enforce them. We also have legal duties, such as a duty to pay taxes, which we are morally obliged to fulfill, but the performance of which no individual can claim as a right.

Whether moral rights must be correlated to duties is more controversial. Many theorists claim that they are, but Joel Feinberg (1980) has argued that moral rights may exist without any corresponding duties. Feinberg distinguishes between what he calls justifiable rights claims and valid rights claims. The former can exist when there are good reasons for a right to exist, even if nobody is in a position to provide what the rights-holder may claim. The idea of human dignity underlies some human rights, for example, which in Feinberg's view justifies claims that certain basic human needs should be met. But in some parts of the world today, nobody is in a position to meet those needs, so they cannot be required as a matter of duty. These rights should be interpreted as developmental goals, the fulfillment of which governments and other agencies must assign the highest priority. Feinberg calls such rights 'aspirational.'

A valid right in Feinberg's sense exists when the claim is justifiable and the corresponding duty can be specified and acted upon. In such cases, the right is logically prior to the corresponding duty. Thus, if Smith has a right *in personam* against Jones, then Jones has a duty in respect of this right. Smith's claim creates the duty for Jones, and if Smith forfeits the right or chooses not to claim it, then unless some other party has the authority to make the claim on Smith's behalf, Jones' duty does not exist.

We can thus interpret moral rights as valid claims, which can be created by law or else founded on the moral belief in human equality and dignity. In order to

explain rights in this way, however, we must be careful not to beg questions about the existence of rights in our understanding of claims. To claim a right is not merely to claim a truth, in the sense that to make a claim can be to assert the truth of a proposition. Claiming rights must be understood as involving a different kind of activity, more like claiming a patent for an invention or a site for a homestead in a newly settled territory. To claim something in this sense is to perform an activity different from asserting, which can create legal and moral consequences. To claim one's basic rights is to demand recognition of one's status as a moral equal, which imposes duties on others in respect of one's claims. Thus, following Feinberg (1980), we can say that a moral right is a claim, the recognition of which is called for by laws, moral principles, or an enlightened conscience.

4. Two Views of the Justification of Rights

This article has been describing rights by explaining their role in societies in which they are recognized. This argument might also count as a justification of rights or an argument, so to speak, for their normative existence. But to this point the argument sidesteps one of the most puzzling aspects of rights, which is about the nature of the values they promote or express. Are rights merely means for promoting other goods and preventing evils, or do they also have intrinsic value? This question has received considerable attention in recent philosophical literature, and most attempts to answer it are controversial. The following sections will consider defenses of each position.

The first approach argues that the value of rights is instrumental. Rights promote goods and prevent evils, the value of which can be explained without any reference to rights themselves. According to this account, rights provide people with certain social guarantees that are necessary for living happy and fulfilling lives. They are part of the institutional structures that give people power over their own lives and protect them from the abuses of governments or other groups.

In order for these guarantees to be effective, we may have to develop the strict prohibitions against doing certain things that rights typically require. We must refrain from treating people in certain ways, not only when the consequences would clearly be bad, but also when we believe in a particular case that the act would produce the best consequences overall. This is because deciding what to do on the basis of calculating the value of our actions in each case would produce consequences that are worse than those produced by institutions that rule out such calculations.

There is much to be said for an instrumental justification of rights. In a utilitarian theory, for example, intrinsic value is located in the aggregation of individual happiness or well-being, and rights play

an important role in promoting this value. Rights that establish property and legally enforceable contracts are effective ways to coordinate individual actions and foster ties of cooperation and trust. Rights that protect individual liberties give people the opportunity to promote their own happiness, which is generally a more effective way of achieving happiness than trusting the task to others. Rights that protect the freedoms of conscience and expression enable people to discover important truths that are necessary for reducing suffering and increasing happiness. In relatively affluent societies, a utilitarian might also defend positive rights to food, shelter, health care, a minimum income, and other basic needs, because the redistributive effects of recognizing such rights can increase social welfare. A utilitarian would thus defend most of the rights we take for granted, as well as many that are controversial.

Critics of utilitarianism may defend many of these same rights, but they will disagree about their justification. The utilitarian justification of rights depends on empirical assumptions that make it the case that rights are effective means for promoting happiness or welfare. But these assumptions may fail to hold, or the facts may be such that violating the rights of some would produce greater happiness overall. If rights are to have a more fundamental and secure role in political and moral thought, according to these critics, then they must be justified by appealing to other, more basic values.

Ronald Dworkin (1977) argues that the fundamental requirement of any legitimate political system is that it treats all people with equal respect and concern. Equality of respect means that, except in extreme circumstances, it is wrong to require some people to sacrifice their interests for the greater happiness of others. Rights play an essential role in securing equality, because they restrict the individual preferences that social and political decisions may legitimately consider, thus guaranteeing to all people a measure of independence from decisions aimed at promoting social welfare or the general good.

The kind of preferences that rights aim to restrict are primarily what Dworkin calls external preferences. These are preferences about how other people should be treated or what they should be permitted to do, because of who they are, because they have interests and desires of which others disapprove, or because their suffering is a source of happiness to others. Dworkin argues that unless a political system disallows these external preferences, it fails to show equality of respect. Under such conditions, a utilitarian concern simply to maximize the satisfaction of individual preferences lacks political legitimacy. Rights thus have a kind of lexical priority in Dworkin's system, which he expresses by saying that rights trump utility.

Another critic of utilitarianism, T. M. Scanlon (1977), argues that the basic values that a political

system should promote must include not only what happens to people but also factors that affect the ability of individuals to determine what will happen to them. People care about their ability to influence outcomes, and they want to be protected from interference or control by others. The goal of rights is to foster and protect this kind of autonomy.

If Scanlon's view is to be distinguished from a sophisticated utilitarianism, however, then the value of autonomy must be defended as an objective good that does not depend on the strength of the preference for it among the people that live in any society. Scanlon argues that moral values must be determined not merely by an aggregation of preferences but by an ethically significant objective notion of their importance. To value autonomy is 'to be concerned with the conditions necessary for people to develop their own aims and interests and to make their preferences effective in shaping their own lives and contributing to the formation of social policy' (Scanlon 1977, p. 85). Rights protect people from the paternalistic interventions of others, and they ensure an acceptable distribution of control over important factors in our lives. Rights would protect religious freedom, for example, even in societies where the majority of citizens care passionately about making their religious views mandatory, while the minority who don't share these beliefs also have tepid preferences about religion. A concern for autonomy implies that people must be free to make their own decisions about religion independently of the control of others.

James Griffin (1999) argues for a different but closely related non-utilitarian justification of rights. He appeals simply to the value of human agency. Griffin argues that a fundamental moral value, more important even than happiness, is the distinctly human ability to contemplate and plan a future. We deliberate, assess, and choose in ways that enable us to envision a good life and act to realize it for our self and others. This value of personhood or agency, however, rests on certain conceptual and empirical facts. One of these facts is the condition of autonomy, which Griffin interprets as the ability to choose a course in life free from domination or control by other forces. But the value of agency also demands that we be provided with some level of education, information, and the chance to learn what others think, so that our choices will be appropriate to our deepest interests. We must have the ability to follow the course of action we have chosen, which includes an adequate supply of material resources and capabilities, as well as protection from having our efforts blocked by others. The goal of rights on Griffin's account is to enable people by protecting and providing these conditions of agency. The basic right in Griffin's system is thus a right to *pursue* happiness, which is more important than maximizing happiness itself.

These three conceptions of moral rights are non-utilitarian, because they appeal to values such as

equality, autonomy, or human agency and claim that these values are more basic than happiness or well-being. But these theorists share with utilitarians the view that the value of rights is instrumental. Rights are part of the structure of institutions that aim to promote other basic values, and they can be overridden for the sake of promoting the values that justify rights in the first place. For example, if rights are justified because they promote equality of respect, then we would be justified in violating equality in some instances in order to produce greater equality overall. On any of these accounts, we presumably should still violate the rights of one innocent person if this is necessary to protect the rights of five other innocent persons.

While these implications might seem acceptable to many people who view rights as morally important but not absolute, they are unacceptable to theorists who believe that rights have intrinsic value. The belief that the value of rights is intrinsic lies at the core of the second approach to justifying rights. According to this second approach, all instrumental accounts of the value of rights fail to explain the way in which rights determine how a person may or may not be treated. Rights aim to secure for each person a realm of independence, not only from the will of the majority but more generally from pursuit of the best overall result.

In order to have this kind of intrinsic value, rights do not have to be regarded as absolute (Nagel 1996). If there is a right not to be tortured, for example, it may nevertheless be the case that some evils are great enough that they would justify torturing one innocent person to prevent them. The threshold for such justifications would have to be set high, however. Torturing one person may be permitted to save a thousand lives, but it would not be permitted to save one life or to prevent two other innocent people from being tortured.

This intrinsic justification of rights expresses requirements for how we may treat other people in a way that is relatively independent of the value of what happens to them. It tries to explain what we are *prohibited from doing* to others in respect of their rights, which it regards as different from what we must try to *prevent from happening* to them. Thomas Nagel, who defends this view of rights, explains that, 'It is this qualified independence of the best overall results... that gives rights their distinctive character' (Nagel 1996, p. 89).

Robert Nozick (1974) assumes the truth of an intrinsic account in his influential defense of natural rights. He insists that rights should be interpreted as side constraints, which mark the bounds of value-based deliberation and set limits to what we are allowed to do in pursuit of all other goals (see also Thomson 1990). But Nozick simply accepts the traditional view of natural rights and does not explain why they should have this character. Without appealing to something such as divine sanction for

natural rights, however, as did many of the traditional natural rights theorists, it seems *ad hoc* to posit rights as first principles of moral deliberation. It begs the question about why it would be wrong to torture one innocent person in order to prevent the torture of five others.

A different intrinsic account, which has been defended by Alan Gewirth (1981, 1982), tries to connect the justification of rights to the capacity for moral valuing, which human beings typically and uniquely possess. Rights express the conditions that make it possible for people to recognize and act on values, so respect for rights is necessary in order for things such as happiness or equality even to exist as moral goods. To violate a person's rights for the sake of some other value, therefore, undermines these conditions and thus makes the actions impossible to justify. While this kind of transcendental argument has drawn much interest among philosophers (see Hart 1955), many people do not find it convincing. If you kill one person in order to prevent five other people from being killed, you cannot then justify what you have done to the victim. But as long as the conditions necessary for him to recognize and act on moral values exist, it is not clear why a person could not comprehend why he might have to be sacrificed for the sake of others.

Another and currently popular intrinsic account of the value of rights appeals to our status as human beings. But rather than seeing human status as an objective moral value, this argument insists that status generates different kinds of moral requirements. Human status is a feature of each person that allows them to claim rights and demand to be recognized and treated in certain ways. Frances Kamm (1992) interprets this status as a kind of inviolability, which she identifies with the possession of rights. According to this argument, if one innocent person may be tortured in order to prevent the torture of five other innocent people, then not only have the rights of that one person been violated, but this acknowledgment of what *may* be done to someone undermines the inviolability of all people. Everyone's moral status is thus compromised by such an act (see Quinn 1993, Nagel 1996).

This argument for the intrinsic value of rights resonates with Feinberg's (1980) interpretation of rights as valid moral claims. He attempts to illustrate the significance of rights by imagining a society that is in every other way as good as possible but does not recognize moral rights. People in this imaginary society are virtuous, caring, and may even be motivated by a sense of duty, but without rights they can only receive beneficent treatment as a gift. If they fail to get what they deserve, they have no ground for complaint. What people lack in a society without rights is the capacity to claim their status as equals and to demand that these claims be recognized. Indeed, Feinberg suggests that the ability to claim rights may be necessary to having the minimal self-respect that

makes a person worthy of the love and esteem of others. A society without rights may be incapable of supporting the kind of love, compassion, and sympathy that we most esteem.

The moral importance of inviolability is thus different from the objective value of not being violated, and it is expressed in terms of moral requirements, the recognition of which each person can demand from all others. If this kind of argument could be made convincing, then it would have several important implications. First, it would explain why it is wrong to violate one person's rights, even to prevent similar violations of five other people. Second, it would explain another feature of rights that internal accounts regard as important, which is that rights imply that it is morally more important not to treat people in certain ways than it is to prevent identical harms from occurring to people. We value the reduction of harms, but rights express our idea of human dignity because they impose moral requirements that take precedence over what we value.

The justification of rights remains a matter of controversy. The difficulty is to defend a theory of rights that captures most of our deepest intuitions about them without making them seem mysterious or paradoxical in the ways we have been discussing. Most philosophers would acknowledge that much work remains to be done in defending an adequate theory of rights, just as most people would acknowledge that much more important work remains to be done to secure the protection of basic human rights for most of the world's population.

See also: Bill of Rights; Children, Rights of; Cultural Concerns; Civil Rights; Cultural Rights and Culture Defense; Cultural Concerns; Human Rights, Anthropology of; Human Rights, History of; Human Rights in Intercultural Discourse; Cultural Concerns; Human Rights: Political Aspects

Bibliography

- Bentham J 1834 Anarchical fallacies. In: Bowring J (eds.) *The Collected Works of Jeremy Bentham*. Longman, London, Vol. 2
- Dagger R 1989 Rights. In: Ball, T, Farr J, Hanson R (eds.) *Political Innovation and Conceptual Change*. Cambridge: Cambridge University Press, Cambridge, UK
- Dworkin R 1977 *Taking Rights Seriously*. Harvard University Press, Cambridge, MA
- Feinberg J 1980 The nature and value of rights. In: Feinberg J (ed.) *Rights, Justice and the Bounds of Liberty: Essays in Social Philosophy*. Princeton University Press, Princeton, NJ
- Gewirth A 1981 Are there any absolute rights? *Philosophical Quarterly* 31: 1–16
- Gewirth A 1982 *Human Rights: Essays on Justification and Applications*. University of Chicago Press, Chicago
- Griffin J 1999 First steps in an account of human rights. Unpublished paper presented to the Conference on Moral Theory and Moral Practice. Le Lavandou, France
- Hart H L A 1955 Are there any natural rights? *The Philosophical Review* 64: 175–91
- Hohfeld W 1923 *Fundamental Legal Conceptions as Applied in Judicial Reasoning*. Yale University Press, New Haven, CT
- Kamm F 1992 Non-consequentialism, the person as an end-in-itself, and the significance of status. *Philosophy and Public Affairs* 21: 381–89
- Nagel T 1996 Personal rights and public space. *Philosophy and Public Affairs* 24: 83–107
- Nozick R 1974 *Anarchy, State, and Utopia*. Basic Books, New York
- Quinn W 1993 Actions, intentions and consequences: the doctrine of doing and allowing. In: Quinn W (ed.) *Morality and Action*. Cambridge University Press, Cambridge, UK
- Scanlon T M 1977 Rights, goals, and fairness. *Erkenntnis* 2: 81–94
- Thomson J J 1990 *The Realm of Rights*. Harvard University Press, Cambridge, MA
- Tuck R 1979 *Natural Rights Theories: Their Origin and Development*. Cambridge University Press, Cambridge, UK
- Vlastos G 1962 Justice and equality. In: Brandt R (ed.) *Social Justice*. Prentice-Hall, Englewood Cliffs, NJ, pp. 31–72
- Waldron J 1984 Introduction. In: Waldron J (ed.) *Theories of Rights*. Oxford University Press, New York

D. MacLean

Rights: Legal Aspects

1. Definition and Background

1.1 The Basic Logic of Rights

The language of rights generally connotes claims of entitlement to certain goods. To assert that 'I have a right ...' is to articulate a normative claim that certain actions should or should not be performed by other individuals, agencies of the state, or intermediate social institutions. As such, rights are related to, but different from, preferences, interests, and needs. While rights are often invoked in the service of preferences and interests, only those rights which confer legal and/or moral authority that obligate others are enforceable by official power. Hence, whereas the assertion of a preference connotes merely that 'I want ...,' to assert a right is to declare that 'I am entitled to ...' Rights thus assume priority in the normative schemes of social ordering; they take priority—or 'trump,' in Ronald Dworkin's memorable phrase—over preferences and interests. Rights similarly take precedence over needs in much the same way—but with one major difference: needs, unlike preferences, convey an objective status that can be validated by persons other than the claimants themselves. Claims to rights, also, are subject to third-party validation, but are, unlike the identification of needs, typically initiated by individual agents acting on their own behalf (Waldron 1996).

Because rights constitute relationships between the

bearers of rights and other persons or institutions, each person's rights impose correlative duties on others. These duties may be negative in that they mandate noninterference or inaction by others, as in one's claim of a right to liberty, privacy, or autonomy. Conversely, the duties imposed by rights may be positive in that others are called upon for action and assistance—as, for example, to ensure the fulfillment of contractual terms or to provide accessible facilities for voting in political elections. Rights may be primarily negative or primarily positive, but they may also involve elements of both.

Rights not only correlate with duties, they also extend by implication to all other similarly constituted relationships in society and are, thus, inclusive. This inclusiveness is the source of the frequent claim that rights are universalistic—i.e., that they can be claimed by, and impose duties on, all citizens within particular sovereign domains. Similarly, the exercise of rights typically confers responsibilities on the rights-bearer. For example, a person may have a right to drink alcohol and a right to drive a car, but these rights imply responsibilities for not engaging in both activities at once. Finally, rights are intrinsically linked to remedies. Thus, for example, not only does the right to equal treatment oblige others not to discriminate, but should discrimination occur, the right to equal treatment requires a remedy to compensate for or to correct the denial of a basic entitlement.

1.2 The Origin and Evolution of Rights

The most widely recognized origin of rights is in European social contract theory dating back to the seventeenth century, and especially to *John Locke's* classic *Two Treatises on Government*. From this work, rights emerged as a moral imperative that is found at the very core of several of the most important political documents in the modern age—including the United States Declaration of Independence as well as the Bill of Rights, and the French Declaration of the Rights of Man and Citizen. While the concept of rights thus dates back several hundred years, and has provided a founding principle for several longstanding liberal republican orders, it is in the second half of the twentieth century that the moral ethos of rights took on widespread significance around the world. The persuasive power of rights developed, in part, in reaction to horrors of authoritarian states—Hitler's genocide, Stalin's purges, colonial repression and the like—and, in part, in recognition of deprivations suffered by ethnic and racial minorities, women, gays and lesbians, and other marginalized groups in nations throughout the world.

1.3 The Expanding Variety of Rights Claims

The evolution and diffusion of rights through the centuries and around the globe has been boiled down

by analysts to three separate stages, or generations, of rights claims.

1.3.1 First generation rights. These are closely linked to the rule of law and to liberal social contract theory; they are regularly found in the constitutions of liberal democratic regimes throughout the world, as well in such transnational legal orders as the European Convention on Human Rights. First generation, or classic rights, are embedded in the logic of both property ownership and in the law itself. On one hand, such rights tend toward the more negative pole, demarcate limits on interference by the state or other social actors, and turn on conceptions of liberty, privacy, due process, and the like. On the other hand, classic, first generation rights also include more positive rights to political participation—to free speech and association, to vote, to run for office, etc.—and to fair treatment by others according to the principles of due process and equal treatment. These rights claims are the most widely recognized and the least contested in general, although specific constructions vary greatly.

1.3.2 Second generation rights. These can be traced to the social democratic traditions of Europe and have during the last century gained currency elsewhere—although not, most notably, in the United States. Second generation rights have to do primarily with entitlements to economic and social wellbeing of the sort formulated in Articles 22–27 of the Universal Declaration of Human Rights (1948)—including, for example, basic education, health care, minimum income, secure employment, workplace organization, food, housing, social security, and the like.

1.3.3 Third generation rights. These are newer in conception, different in kind, and particularly controversial. Advocates of third generation rights seek to confer entitlements to collective or communal goods such as peace, security, a healthy environment, safe natural resources, animal welfare, and group self-determination or sovereignty, and thus seek to its rights to animal. The problematic posed by third generation rights will be addressed below.

2. Contemporary Philosophical Debates About Rights

The original conceptions of so-called liberal Western thinkers in the social contract tradition grounded rights in the theocratic logic of God's divine plan and natural law. Later versions downplayed the divine hand but continued the emphasis on rights as self-evident and inalienable features sewn into the rational

order of the natural universe or human nature. By the twentieth century, the philosophical foundations for rights became thoroughly secular—whether justified by various forms of moral rationality or appeals to proven historical efficacy. It is within the tensions among these legitimating logics that many of the most important philosophical controversies about rights have been waged. We identify just a few of such persistent debates.

2.1 ‘Nonsense Built on Stilts’

This critique, that can be traced back at least to Bentham and continues today among critical ‘neo-realist’ philosophers, turns on the recognition of rights indeterminacy. The core contention is that if rights are indeterminate and contingent, if they always are changing, and vary with context, then they really mean nothing at all and convey no real social power. Rights, it is argued, are simply moral fictions that possess none of the objectivity, coherence, independence, and universality that philosophers, state officials, and popular culture assign to them. At best, rights are epiphenomenal reflections and rationalizations for relations grounded in other types of causal forces. Respondents to this position acknowledge that rights are social constructions, but the very fact that people believe that rights matter and act accordingly in interpersonal relations, often in the shadow of official enforcement, demonstrates the power of rights. It is further argued that all moral discourses are indeterminate and that it is, therefore, inappropriate to dismiss rights discourses as meaningless because they are unstable and contestable. Indeed, according to this argument, the indeterminacy of rights discourse, as with moral discourse more generally, is desirable in providing a flexible, dynamic, and pragmatic tradition for ordering social relations in just ways.

2.2 Who Has Rights?

The debate over the qualifications for rights-bearing is as old as the concept itself and has expanded among moral philosophers as specific rights claims themselves have proliferated. From the start, rights typically have been understood to be properties belonging to all individual human agents capable of claiming them. But it has been precisely these qualifying properties of rights bearers as human agents that have been deployed to limit the reach, and some might say to betray, the promise of rights. Thus, paradoxically, the claim that rights apply only to rational, disciplined, educated, human agents capable of self-government has been invoked to deny rights to children, women, ethnic and racial minorities, religious adherents, immigrants, and the like in many societies that generally recognize individual rights. Similarly, the principle that rights inherently presume individual human

claimants has been central to contemporary arguments that third generation entitlements of animals and other nonhuman forms of life (e.g., trees), as well as nondivisible communal goods such as peace and a safe environment, are not suitable for rights claims. The philosophical skeptics argue that such claims lack theoretical coherence and might even undermine the social power of rights discourse itself. Such debates have generated extensive reconsideration about the very foundation of the meaning of rights as a social practice.

2.3 Do Rights Encourage Atomism, Egoism, and Selfishness?

The contention that rights practices encourage excessive individualism, egoism, and selfishness among citizens was perhaps most forcefully articulated by Karl Marx and is commonly asserted by moralists across the political spectrum today. Critics emphasize in particular the tendency of rights to celebrate the individual and thus to undermine citizens’ sense of responsibility for their actions, duties to others, and a shared commitment to ‘community.’

Theorists more favorably disposed to rights acknowledge that rights discourses presume, encourage, and protect individuals. Indeed, securing and advancing the exercise of individual citizen capacities for choice, action, self-development, and diversity against the harmful designs of other individuals, social groups, popular majorities, and the state has been the *raison d’être* of rights as a moral concept from the start. At the same time, however, most theorists understand the rights of individuals in relational rather than insular terms. Rights practices sustain particular obligations that are reciprocal among citizens as well as a general tolerance and even respect for the human dignity of others. In Kant’s (1948, pp. 95–6) famous words, ‘Act in such a way that you always treat humanity, whether in your own person or in the person of another, always as an end and never simply as a means.’ Defenders thus argue that, far from undermining a sense of duty, rights support and enforce dutiful interaction among citizens. According to this way of thinking, rights connote a particular vision of community where individual agency is respected but balanced with conceptions of fairness, justice, and equal treatment for all. It follows that it is precisely the disregard for basic rights that encourages or permits unjust or exclusionary acts that are corrosive of community.

There is also a middle position in this debate. To begin with, it is widely recognized that some rights are more self-regarding than others. Marx’s powerful indictment, after all, was that it is property rights that encourage selfishness, materialism, and competition for dominance through capital ownership. As a consequence, he argued, it was the positive communal promise of political rights which was overwhelmed, thus reducing the promise of political solidarity to an

illusion. The underlying point here is that critiques of rights that fail to distinguish among different types of rights—within the first generation as well as among the different generations of rights concepts—are unconvincing to many moral theorists.

3. *Jurisprudence and Legal Theory*

Contemporary discourse on rights within jurisprudence and legal theory is driven by an effort to identify a distinctively legal conception of rights. This discourse is organized around two foundational debates—pitting rights against democracy on the one hand, and natural law against legal positivism on the other. Historically, there have been profound and heated conflicts among the parties to these debates, but more recently a rapprochement has developed around the recognition that rights and law are mutually constitutive—that rights without the law are virtually unthinkable, as is law without rights. This convergence not only identifies the core common ground of legal theory; it also reveals the distance that separates the legal theory of rights from the understanding of rights that is found in moral and political philosophy. Whereas philosophical understandings of rights reach well beyond the law in search of universal moral meaning, rights in legal theory are much more constrained—principally by way of the intrinsic connection in legal thought between rights and the institutional processes in which rights under law are embedded. In particular, legal theories of rights seek to specify and to justify the responsibility of courts and judges for resolving disputes over rights, and for policing the boundaries of legitimate rights claiming.

3.1 *Legal Rights vs. Democracy*

The problem of the legitimation of rights claims is especially troublesome to legal officials in democracies and republics, where majority will is understood to carry considerable authority. We noted above that rights, by their very nature, are intended to protect and advance individual entitlements that restrict the actions of others, including those of popular majorities. As such, rights—both in general, and in specific instances—are often condemned as counter-democratic forces. Utilitarian theorists raise a parallel objection that begins with the assertion that rational public policy should provide the greatest good for the greatest number—thus giving considerable weight to the preferences of the majority. Rights are thus often challenged as costly obstacles to the development of democratic and efficient (see below) social policy.

Other legal theorists respond that rights do not thwart so much as give structure to utilitarian policy calculation, by assigning higher values to certain basic entitlements—especially those concerning protection for individual freedom and dignity—than to other

competing considerations. Consequently, securing rights may impose costs on particular social goals, but these costs are arguably incurred in order to sustain larger, enduring social commitments. Likewise, it is argued that rights claims do not absolutely throttle democratic majorities or elected officials; rather, rights claims tend only to place specific, limited restrictions on particular policy actions that defy core social values and protection for individuals. Moreover, it is important to underline that many types of rights advance—indeed, are actually essential to—the democratic process. Such entitlements include, at the least, protections for dissenting speech, political association, free elections, due process, and basic public education.

There is no simple way out of this debate. While it may well be true that rights in general provide sustenance to democratic polities, particular rights constructions certainly do, and indeed are intended to, defy the will of elected officials and popular majorities. Hence the enduring challenge for judicial authorities is to justify by legal argument their specific rights constructions.

3.2 *Rights and Justice*

Historically, a divisive debate in legal theory between positivism and natural law has pitted a theory of law that recognizes little in common between morality and law against a conception of law as being inseparable from morality. The latter is virtually indistinguishable from the moral philosophy of rights discussed earlier, and represented by the work of Locke and Kant among others. The former is rooted in analytic philosophy and in a logical positivism that eschewed the search for universal truths. During the late twentieth century the gap between legal positivists and natural law theorists narrowed (but was by no means closed) by the emergence of a widespread agreement on a least-common-denominator conception of justice as the ‘right to have rights.’ According to this way of thinking, the guarantees associated with rights are confined to assuring predictability, reciprocity, and impartiality in the application of the law. What remains unresolved in the debate between theorists of natural law and legal positivism is whether justice, as thus defined, is the differentiating element of legal right, or only a quality that tends to be closely associated with legal rights.

While many jurists have not reached out to the broader realms of political and moral philosophy, Ronald Dworkin, (Dworkin 1977) at the end of the twentieth century the most prominent legal theorist of rights, has done so. At the core of his analysis is the distinction between policy and principle. Whereas policy is driven by utility, principle, the defining element of law, is to be observed, not because it will advance or secure an economic, political, or situation deemed desirable, but because it is a requirement of justice or fairness or some other dimension of morality.

In thus linking law to principle, and principle to justice, Dworkin identifies himself with the relatively noncontroversial common ground between positivism and (neo)natural law. On the other hand, by introducing 'some other dimension of morality' he clearly breaks new ground. For Dworkin, it is constitutions and constitutional rights that 'fuse legal and moral issues' (Dworkin 1977, p. 185). Constitutions thus provide in addition to the right to have rights, particular substantive rights as well.

For Dworkin, then, rights are inextricably linked not only to the foundational elements of positive law but, at least in their political and constitutional incarnation, to the foundational elements of the state. In this sense Dworkin can be seen as a kind of transitional figure among legal theorists, reaching out as he does to moral and political philosophy and incorporating them into his analysis of rights constructions by judges and other state officials. But unlike those of moral and political philosophers, Dworkin's theory of rights is embedded in, and inextricably linked via principle to, what is distinctively legal. Moral philosophy thus becomes an adjunct to the law and is, in effect, put at the disposal of the law, as a source of principle. In this sense, Dworkin's theory of rights returns to the starting point and, indeed, to the cornerstone of jurisprudence and legal theory, that is, it is a search for what is uniquely legal, as opposed to moral, ethical, or political.

4. *Empirical Social Science Studies of Rights Practice*

Whereas philosophic and jurisprudential inquiries into rights tend to be normative and abstract, social science research regarding rights practices is explanatory and empirical. Moreover, in the social sciences rights are treated less as moral entitlements of historical persons than as contingent resources that are mobilized both by individuals in everyday life, and by organized interests in political advocacy. Rights are not viewed as 'above politics' and power so much as constitutive of political activity and power relations. The primary objective of the social sciences thus is to assess the value of rights as resources, and to explain variation in the value of rights from time to time and from place to place. To do so, social scientists shift the locus of rights inquiry from formal systems of logic and idealized institutions to inductively organized research on rights in actual institutional settings, everyday life, and political advocacy.

4.1 *Rights and Everyday Disputing*

There has been a rapidly expanding scholarly literature on the ordinary disputing practices of citizens, especially in the United States but increasingly around the world. Many studies of 'ordinary' (non-elite)

individual citizens in the US suggest that rights matter a great deal, and that these individuals are not hesitant to assert them. Indeed, some scholars demonstrate that lay people, more so than judges, clerks, police, and other legal authorities, act on the premise and the expectation that the legal system exists to protect rights. Studies from other polities around the globe have also found some evidence for this generalization. One of the most interesting findings of Ewick and Silbey (1998) is that ordinary, nonelite citizens in the US can, moreover, relate to rights in flexible, creative, and complex ways, entertaining several conflicting notions at once. Interviewees in their study could view rights instrumentally as malleable rules of a game, authoritatively as an objective source of valid law, and skeptically as a burden that must be endured, evaded, or resisted.

Other studies qualify or challenge assumptions about the salience of rights. Research on the disputes of individuals suggests, for example, that Americans and others often refrain from asserting their rights; many studies document a widespread reluctance to seek redress when rights are violated and entitlements denied. This is especially the case among lower-income and minority citizens. Moreover, some studies find evidence of considerable cultural biases and popular pressure against rights claiming, at least through formal adversarial channels, even in allegedly 'rights obsessed' cultures such as the US. Patterns of individual practice are thus complex, contradictory, counterintuitive, and often defy generalization within, much less beyond, the US context.

4.2 *The Politics of Rights: Mobilizing Rights in Political of Advocacy*

The value of rights as a political resource is heavily debated within the social science literature. The most stringent critics of rights strategies claim that to deploy rights is at best ineffectual, and at worst counterproductive. Some studies suggest that rights litigation had no more than negligible effect on policy outcomes. According to realist critics only political power built around electoral coalitions and influential interests can bring about changes in well-established policies and practices. Other scholars demonstrate that rights-based practices often do intrude into policy processes, but with the unfortunate consequence of making these processes rigid and inefficient.

There is, however, a substantial body of research that suggests rights strategies can make meaningful contributions to a transformative politics—albeit in complex, contingent, and unpredictable ways. This research does not so much invalidate the findings of critics as shift the focus and the standards by which rights strategies are to be judged. Instead of concentrating exclusively on the direct effect of court-articulated rights, their indirect and constitutive effects are explored and assessed. Moreover, in this research,

the criterion for judging rights strategies shifts from policy efficiency to social equity—with the salience of this latter criterion most compellingly articulated in the work of critical race and critical gender theorists.

Especially in the United States, where much of the available research has been conducted, the indirect and transformative potential of a ‘politics of rights’ (Scheingold 1974) has been persuasively demonstrated. Just as rights have a constitutive impact on individual disputants, so too have the connotations of entitlement associated with rights had a mobilizing impact on social movements for pay equity, disability rights, and animal rights. Insofar as a need or a grievance can be articulated as a legal or constitutional right, expectations can be altered, quiescence overcome, and a politics of rights generated. Note that this research focuses very much on the discursive power of rights to constitute social and political life. Moreover, despite their liberal connotations and associations, rights can also be deployed on behalf of substantive agendas that span the political spectrum.

Despite all of their persuasive power, however, rights alone secure very little. Nor is rights mobilization a matter of spontaneous combustion. The constitutive power of rights depends on deploying rights shrewdly, in conjunction with other political resources—including shrewd leadership working in combination with established organizations like churches (civil rights), labor unions (pay equity), and other kinds of social-action organizations. Even then a transformative politics does not necessarily follow. Political mobilization may, for example, generate countermobilization (the abortion rights vs. right-to-life movements). But as the struggle over abortion clearly demonstrates, rights do matter, and a well executed strategy of rights can at the very least destabilize the political arena—thus creating opportunities that were previously foreclosed. In sum, the increasing body of empirical research on rights advocacy in a variety of settings underscores the variable, contingent, and often-contradictory implications of a politics of rights.

See also: Bill of Rights; Civil Liberties and Human Rights; Civil Rights; Cultural Rights and Culture Defense; Cultural Concerns; Fundamental Rights and Constitutional Guarantees; Human Rights, Anthropology of; Human Rights, History of; Human Rights: Political Aspects; Rights

Bibliography

- Bumiller K 1988 *The Civil Rights Society: The Social Construction of Victims*. The Johns Hopkins University Press, Baltimore, MD
- Crenshaw K W 1988 Race, reform and entrenchment: Transformation and legitimation in antidiscrimination law. *Harvard Law Review* **101**: 1331–87
- Dworkin R 1977 *Taking Rights Seriously*. Harvard University Press, Cambridge, MA

- Ewick P, Silbey S 1998 *The Common Place of Law: Stories from Everyday Life*. University of Chicago Press, Chicago
- Fuller L 1964 *The Morality of Law*. Yale University Press, New Haven, CT
- Glendon M A 1987 *Rights Talk: The Impoverishment of Political Discourse*. Free Press, New York
- Hart H L A 1961 *The Concept of Law*. Clarendon Press, Oxford, UK
- Kant I 1948 *The Moral Law* (trans. Paton H J). Hutchinson, London
- McCann M W 1994 *Rights at Work: Pay Equity Reform and the Politics of Legal Mobilization*. University of Chicago Press, Chicago
- Melden A I 1977 *Rights and Persons*. University of California Press, Berkeley, CA
- Merry S E 1990 *Getting Justice, Getting Even: Legal Consciousness Among Working Class Americans*. University of Chicago Press, Chicago
- Minow M 1990 *Making All the Difference: Inclusion, Exclusion, and American Law*. Cornell University Press, Ithaca, NY
- Scheingold S 1974 *The Politics of Rights: Lawyers, Public Policy, and Political Change*. Yale University Press, New Haven, CT
- Waldron J 1996 Rights and needs: The myth of disjunction. In: Sarat A, Hearn T (eds.) *Legal Rights: Historical and Philosophical Perspectives*. University of Michigan Press, Ann Arbor, MI

M. McCann and S. Scheingold

Right-wing Movements in the United States: Women and Gender

Right-wing movements have existed since colonial times in the USA, yet the majority of studies have focused on either what is termed the ‘Old Right’ of the 1950s or the ‘New Right’ formed during the mid-1970s and 1980s. An explosion of interest in understanding right-wing movements occurred in the aftermath of WWII when attention turned to authoritarianism, McCarthyism, and the Old Right of the 1950s. During the 1960s and 1970s, particularly with the prominence of protest movements on the left, research on right-wing movements waned, but a resurgence of interest has focused on the New Right formed during the 1970s and 1980s. While little is known about women’s participation in the Old Right, recent studies have focused on understanding women’s role in the New Right. This article considers the affect of women and issues of gender in both the Old and New Right.

1. Women and the Old Right

Historically, right-wing movements in the US have been based on a crucial marriage of interests between two groups. First, members of less-educated, lower economic strata, who are highly religious and drawn to the non-economic issues of right-wing movements.

This sector is intolerant of ethnic and religious minorities, with particular antagonism expressed towards Blacks and towards immigrants, previously expressed against Catholics and Jews and more recently directed against Mexicans and Asians. The second sector includes those rooted in a more highly educated, higher-income strata, less religious, tolerant of minorities, and committed above all to economic conservatism. The Old Right of the 1950s combined these two sectors, one consisting of more affluent economic conservatives, particularly reacting to the expansion of government implemented by President Franklin D. Roosevelt during the New Deal of the 1930s, and a more devoutly religious and less educated, lower middle class following, fearful of communism due to strong evangelical and fundamentalist beliefs. Anticommunism held together the Old Right, spearheaded by the campaigns of Senator Joseph McCarthy.

In terms of women's role within the Old Right, the vast majority of research makes no mention at all of female activists. It is not clear whether the lack of attention to women in the Old Right is due to the small number of female participants in the movement or to the male bias of the researchers. Because of this lack of data, there is no way of knowing how women of the Old and New Right compare in terms of number of participants, degree of involvement, or social background.

2. Women and the Conservative Intellectual Movement

Cleavage within the popular movements of the Right in the US is also evidenced in the conservative intellectual movement, which is comprised of three main groups. First, classical liberals or libertarians promote the free market, private property, limited government, self-reliance, and *laissez-faire*, and resist the threat of the ever-expanding state to liberty and free enterprise. Second, are traditionalists who are dismayed by the erosion of values and the emergence of a secular society and who call for a return to traditional religious and ethical absolutes. Third are militant anticommunists, ex-radicals disillusioned by the left and who became alarmed by the despotism of the Soviet Union. With the emergence of the New Right, this third strand also includes 'neo-conservatives,' primarily academics and writers who rejected the politics of the Democratic Party during the late 1960s. Converts from the left, neo-conservatives are disillusioned liberals drawn heavily from Jewish, intellectual circles of the East Coast. As with mass right-wing movements, anticommunism has acted as the glue holding together the conservative intellectual movement.

Although nearly all of the influential conservative intellectuals are male, libertarian author Ayn Rand

has had a profound impact on generations of right-wing adherents. Through novels such as *The Fountainhead* and *Atlas Shrugged*, Rand introduced the philosophy of objectivism and celebrated unbridled individualism, rationalism, and *laissez-faire* capitalism. Rand inspired thousands of people, especially youth during the 1960s, to join the libertarian cause.

3. Women and the New Right

During the mid to late 1970s varied groups coalesced into the New Right. Much of the force behind this coalition was a counter-movement to the social movements of the 1960s, an organized protest against affirmative action, the welfare state, feminism, and gay and lesbian rights. Like the Old Right, the New Right also incorporated two distinct constituencies: social conservatism and economic or *laissez-faire* conservatives. Social conservatives are devoutly religious and are primarily concerned about moral issues and protecting the traditional family. The family represents the building block of society; it instills children with moral values and restrains the pursuit of self-interest. Implicit in this image of the family is the social conservative conception of human nature. Humans are creatures of unlimited appetites and instincts. Left on their own, the world would be a chaos of seething passions, overrun by narrow self-interest. Only the moral authority of the family and religion tames human passions, transforming self-interest into the larger good. The ideal society is one in which individuals are integrated into a moral community, bound together by religious faith and by common values. *Laissez-faire* conservatives or libertarians in contrast, aim above all to protect the economic and political liberty of the individual. Rooted in the classical liberalism associated with Adam Smith, John Locke, and John Stuart Mill, *laissez-faire* conservatives uphold the economic liberty of the free market and the political liberty based on the minimal state. The individual is viewed as an autonomous, rational, self-interested actor. Rather than viewing humans as creatures of unlimited passions, *laissez-faire* conservatives view humans as beings endowed with free will, initiative, and self-reliance. The larger good is not ensured through the maintenance of moral authority and the restriction of self-interest. Rather, the *laissez-faire* ideal poses a society in which natural harmony exists through the very pursuit of self-interest. The aim of the good society is to elevate the potential of humans by bringing their creative and productive nature to fruition.

Despite these fundamental differences, like the Old Right, these two constituencies of the New Right united in their virulent opposition to communism. The New Right is distinct from the Old Right, however, in the visibility of its female participants and leaders.

Increasingly women have taken a more active and visible role in right-wing politics. On a leadership level, such well-known figures as Phyllis Schlafly, who led the fight to defeat the Equal Rights Amendment (ERA), typifies the social conservative movement while Jeane Kirkpatrick, a professor of government, former United Nations representative, and newspaper columnist, embodies *laissez-faire* beliefs. Yet women of the New Right, like their male counterparts, are not a homogenous entity. They are drawn from different sociological pools. While the majority of activists are white, social conservatives tend to be somewhat less educated than *laissez-faire* women and are devoutly religious, primarily Catholic or Protestant fundamentalist, married with children, with many being full-time homemakers. *Laissez-faire* women tend to be younger, more secular in orientation, more likely to be single or divorced, and have smaller numbers of children, more postgraduate degrees, and typically have professional careers. In fact, in terms of social background *laissez-faire* women resemble feminist women.

3.1 Views of Gender and Feminism

Women of the New Right have contrasting beliefs regarding gender. Social conservative women believe in a strict division of gender roles as decreed by the Scriptures. Gender is envisioned as a hierarchical ordering with God and Christ at the top, followed by men, and then women. While male and female roles are each respected as essential and complementary components of God's plan, men are the spiritual leaders and decision-makers in the family. Women's role is essentially defined in terms of altruism and self-sacrifice: to support men in their positions of higher authority, and to care for the family. Because gender roles are delineated in such unambiguous terms, any blurring of these roles is viewed as a threat.

While social conservatives envision a man's world as distinct from a woman's world, no such distinctions are visible in the *laissez-faire* view of gender. There is a notable absence of commentary on gender, and particularly no mention of the need for male authority and female submission or of women's natural orientation to others. The social conservative vision of a hierarchical ordering of authority of men over women is antithetical to the *laissez-faire* conservative ethos. Faith in individual self-reliance and free will, and belief in the liberty and autonomy of every individual, extends to women as well as to men.

Clearly, there is a relationship between a woman's social location and her values and beliefs. It is not coincidental that a woman with a devout religious upbringing who is a full-time homemaker is likely to see the family at the center of the world and to adhere to a traditional ideology regarding gender. Nor is it a

coincidence that a single professional woman who is devoted to her career does not see the family as central and holds nontraditional attitudes about men's and women's roles. Traditional gender roles 'work' for social conservative women not simply because of ideology, but because they themselves are more likely to be in life situations which correspond to traditional roles. *Laissez-faire* women, on the other hand, are in life situations and have social resources outside of the traditional female sphere.

These varying perceptions of gender result in opposing views of gender discrimination. Because social conservatives adhere to a hierarchical ordering, they believe that positional differences between women and men do not imply inequality. Any positional differences between the sexes are attributed to 'natural' differences or to individual choice, not to gender discrimination. *Laissez-faire* women, on the other hand, are inclined to recognize and deplore the unequal treatment of women. Discrimination is denounced as interfering with the individual's ability to climb to the height of her talents. Both men and women must remain unconstrained by external interference, free to follow the dictates of their own initiative. Although *laissez-faire* women share recognition of discrimination with feminists, they react to such bias purely in voluntaristic terms. Women who are discriminated against at work should on their own initiative leave the job and enter the free market to choose another one.

These differences in views of gender extend to opposing views of feminism. While social conservatives view feminism as one of the primary forces of moral decay responsible for America's decline, *laissez-faire* women actually adhere to part of the feminist vision. Social conservative activism is a direct counter-movement to the feminist movement of the 1960s and 1970s, as well as to the changes brought on by the civil rights, student, and gay and lesbian movements. Social conservatives blame feminists for attacking the status of the homemaker and for promoting issues such as the ERA, abortion, gay rights, and childcare, all of which are viewed as eroding traditional notions of gender and the family. Calling themselves the 'pro-family' movement, opposition to these very issues tops the political agenda of social conservative activists. *Laissez-faire* women, on the other hand, support gay rights, a woman's right to abortion, and daycare, seeing them as private, individual matters that should not be legislated by the state. *Laissez-faire* women depart from many feminists, however, in the means to achieve women's equality. While feminists call for federal support in addressing women's needs, *laissez-faire* women oppose any reliance on big government, including government funding of daycare or abortion. Further, the top priority of *laissez-faire* women is not addressing women's inequality; rather, issues related to the economy and defense remain the objectives of these women.

3.2 The Paradox of Right-Wing Women's Activism

Given the social conservative woman's adherence to traditional gender roles in which men are breadwinners and protectors, and women are helpmates and caretakers, some analysts point to a seeming contradiction between these beliefs and her very role as a political activist. However, social conservative women do not see a tension between their political involvement and the traditional female role. Female activism is defined within the bounds of traditional gender ideology as women altruistically working for the benefit of a larger cause, acting as moral gatekeepers to bring purity to a world filled with sin. Nor is there any apparent paradox to *laissez-faire* women's activism. Because they do not see women as bound to traditional roles, there is no seeming contradiction between their beliefs and their own role as public leaders. They do not conceive of themselves as the caretakers of society, altruistically at work for the benefit of all. Instead, they see themselves as no different from men, self-interested actors working for a political cause.

Despite the prevalent assumption that anti-feminist women are 'brainwashed' or 'lackeys' of men, the actual paradox of right-wing women is that those women who are furthest from feminists in their beliefs actually do act in their own interests as women, while *laissez-faire* women, who actually share a portion of the feminist vision, do not act in their collective gender interest. Far from suffering from false consciousness, social conservative women are well aware of their interests and act to defend their status as women. To borrow a phrase from Marx, social conservative women act as women for themselves, while *laissez-faire* conservative women remain women in themselves. Gender identity is central to the political involvement of social conservative women; recognizing their commonality with other traditional women, they seek to protect women's place as a group. *Laissez-faire* women are not motivated out of concern regarding gender; they do not act in the collective interest of women and therefore remain women in themselves. They act as members of the marketplace, not as members of their gender group, in organizing to return America to strength and freedom.

See also: Civil Rights Movement, The; Ethnic and Racial Social Movements; Feminist Movements; Fundamentalism (New Christian Right); Gay/Lesbian Movements; Historiography and Historical Thought: Southeast Asia; Identity Movements; Labor Markets, Labor Movements, and Gender in Developing Nations; Labor Movements and Gender; Labor Movements, History of; Nation-states, Nationalism, and Gender; Peace Movements, History of; Social Movements, History of; General; Social Protests, History of; Youth Movements, History of

Bibliography

- Conover P J, Gray V 1983 *Feminism and the New Right: Conflict over the American Family*. Praeger, New York
- Himmelstein J L 1990 *To the Right: The Transformation of American Conservatism*. University of California Press, Berkeley, CA
- Klatch R E 1987 *Women of the New Right*. Temple University Press, Philadelphia, PA
- Lipset S M, Raab E 1970 *The Politics of Unreason: Right Wing Extremism in America, 1790–1970*. Harper and Row, New York
- Luker K 1984 *Abortion and the Politics of Motherhood*. University of California Press, Berkeley, CA
- Mueller C, Dimieri T 1982 The structure of belief systems among contending ERA activists. *Social Forces* **60**: 657–75
- Nash G H 1976 *The Conservative Intellectual Movement in America Since 1945*. Basic Books, New York
- Staggenborg S 1998 *Gender, Family, and Social Movements*. Pine Forge Press, Thousand Oaks, CA

R. E. Klatch

Risk: Empirical Studies on Decision and Choice

Even though most people have an intuitive understanding of what is meant by *risk*, risk is a concept that defies precise definition. Ordinary citizens as well as scientific experts differentiate between choice options or courses of actions in terms of their riskiness, and such explicit or implicit judgments of perceived riskiness often affect their decisions. Some uncertainty about the outcome of a decision or action seems to be necessary to make it appear risky. This article describes three paradigms that define risk in different ways and that identify different types of variables that have been shown to affect people's perception of risk.

1. Risk Perception

The democratization of decision making and increased availability of information provide people with more choices than ever before. Economic globalization and faster social and technological change have, at the same time, introduced more uncertainty and unpredictability. Decisions under risk and uncertainty are abundant, and perceptions of risk affect those decisions. People pull their money out of financial ventures when they judge the risks to be too high or start a lawsuit when the risks of inaction outweigh the risks of litigation. The risk perceptions of individuals as citizens and consumers also affect decisions made by government agencies and corporations. Public perception that silicone breast implants put users at

risk of autoimmune disease, for example, resulted in bankruptcy for the manufacturer, despite clear scientific evidence of no silicone-related illnesses. Controversies about the licensing of technologies such as genetic engineering or the siting of facilities such as landfills or nuclear power stations tend to be fueled primarily by disagreements about posed levels of risk, rather than by disagreements about the acceptability of specific risk levels. While business or government experts have clear quantitative evaluations and definitions of the risks that products or technologies pose, based on objective data or models, members of the general public often seem to evaluate the same options very differently. Much of the early research on psychological risk dimensions (described in Sect. 2.2) was funded by the Nuclear Regulatory Commission to understand why and how public perception of the riskiness of nuclear technology could differ so dramatically from the estimates provided by nuclear engineers.

2. Measures of Perceived Risk

Risk perception has been studied empirically within three theoretical paradigms. Studies within the axiomatic measurement paradigm have focused on the way in which people combine objective risk information, that is, possible consequences of risky choice options such as mortality rates or financial losses and their likelihood of occurrence. Research within the psychometric paradigm has identified people's emotional reactions to risky situations that affect judgments of the riskiness of physical, environmental, and material risks in ways that go beyond their objective consequences. Studies within the sociocultural paradigm have examined the effect of group- and culture-level variables on risk perception.

Risk as an explicit or implicit variable in theories of risk preference is discussed elsewhere (see *Risk: Theories of Decision and Choice*; *Personality and Risk Taking*; Weber 1997).

2.1 Axiomatic Measurement Studies

The axiomatic measurement approach models perceived risk as a function of attributes of risky options that are described as probability distributions over possible outcomes. Much of this work derives from the definition of risk as variance in traditional models of risky choice (see *Risk: Theories of Decision and Choice*). Under this definition sure options carry no risk, since there is no variance around the guaranteed, single, known outcome. The wider the distribution of possible outcomes, the greater an option's riskiness becomes. Yet, studies that assessed people's judgments of the riskiness of financial gambles (see *Risk: Theories of Decision and Choice*) have shown that downside

variability of outcomes affects the perception of an option's riskiness much more than upside variability, an asymmetry that is not captured by variance measures.

As a result, other axiomatic measures of perceived risk were developed that differ in the assumptions they impose on judgments of riskiness (see *Risk: Theories of Decision and Choice*). These models capture both *similarities* in people's risk judgments (with a common functional form by which probability and outcome information about risky options are combined) and *individual and group differences* (with the help of model parameters that capture differences in the relative weight given to different model components).

Weber and Bottom (1989, 1990) submitted the behavioral axioms on which these models are based to empirical tests and found support for the assumption that risk judgments are transitive (if option B is judged more risky than option A, and C more risky than B, then C will also be judged more risky than A) and monotonic (if option B is judged more risky than A, then any combination of obtaining B with probability p and X otherwise should also be more risky than the combination of obtaining A with probability p and X otherwise). Some violations of the expectation principle for risk judgments in the gain domain (Keller et al. 1986) were shown to be the result of nonnormative probability accounting, similar to that observed for preferences. People, for example, do not think that a two-stage lottery carries the same risk as a mathematically equivalent single-stage lottery, even if the final outcomes of the two lotteries and their probabilities are identical. Weber and Bottom's (1989) results supported the additive combination of gain and loss components hypothesized by the conjoint expected risk (CER) model (Luce and Weber 1986; see *Risk: Theories of Decision and Choice*) and ruled out, at least as descriptive models of perceived risk, other risk functions that assume that losses and gains impact risk judgments multiplicatively. The CER model has been described as the most viable model to describe single-dimensional risk appraisal (Yates and Stone 1992) and is probably the most widely applied axiomatic model of risk, but other, similar models exist (see *Risk: Theories of Decision and Choice*).

Bontempo et al. (1997) fit the CER model to judgments of the riskiness of monetary lotteries made by business students and security analysts in Hong Kong, Taiwan, the Netherlands, and the US. Cross-national differences in risk judgments followed a Chinese–Western division, with respondents from the two culturally Chinese countries having model parameters that were similar to each other but different from those of the two Western countries. The probability of a loss had a larger effect on perceived risk for Western respondents, while the magnitude of losses had a larger effect on the risk perceptions of Chinese respondents. Cross-cultural differences in risk perception were greater than differences due to occu-

pation (students vs. security analysts), suggesting that cultural upbringing and environment seem to play a larger role in shaping the perception of financial risks than professional training or expertise.

2.2 Psychometric Studies

The psychometric measurement approach treats risk perception as a multidimensional construct and uses multidimensional scaling, clustering, and factor analysis to identify its underlying psychological dimensions (Slovic et al. 1986). Research within this paradigm has found that the perceptions of the risks of hazardous technologies or activities by members of the lay public have often little to do with possible outcomes and their probabilities. Compared to technical experts, ordinary citizens overweight risk associated with infrequent, catastrophic, and involuntary events, and underweight the risk associated with frequent, familiar, and voluntary events. The psychological risk dimensions identified by the psychometric paradigm fall into two categories. The first one, *dread*, is defined by a perceived lack of control, feelings of dread, and perceived catastrophic potential. The second one, *risk of the unknown*, is defined as the extent to which the hazard is judged to be unobservable, unknown, new, or delayed in producing harmful impacts. Cross-national comparisons show that risk perceptions in a wide range of countries are affected by the *dread* and the *risk of the unknown* factor. Risk perceptions between countries differ in where respondents place particular hazards (e.g., nuclear power) within this factor space, usually in ways consistent with national differences in exposures and socioeconomic development.

2.3 Sociocultural Studies

In Douglas and Wildavsky's (1982) cultural theory, risk perception is viewed as a collective phenomenon by which members of different cultures selectively attend to different categories of danger. Each culture selects some risks for attention and chooses to ignore others. Cultural differences in risk perceptions are explained in terms of their contribution to maintaining a particular way of life. The theory identifies five distinct cultures (labeled hierarchical, individualist, egalitarian, fatalist, and hermitic, respectively) that differ in their patterns of interpersonal relationships and argues that members of these cultures therefore differ in their perceptions of risk. Hierarchically arranged groups, for example, tend to perceive industrial and technological risks as opportunities, whereas more egalitarian groups tend to perceive them as threats to their social structure. The significance of this approach to understanding risk perception is that it provides a way of accounting for the effect of group-

and culture-level variables on the behavior of individuals. It suggests that culture teaches individuals where their interests lie and what variables and events pose risks to those interests and ways of life.

Implicit in this cultural theory of risk perception is the hypothesis that cultural differences in trust in institutions drive differences in perceived risk. Slovic (1997) summarizes empirical evidence for this hypothesis, for example, the fact that minority status (e.g., due to race) is associated with reduced trust in social institutions. The relationship between trust and risk perception seems to be mediated by an emotional pathway, with reduced trust resulting in stronger negative affective responses to potential hazards and increased perceptions of risk.

Weber and Hsee (1999) provide another culture-level explanation of cross-national differences in risk perception, based on differences in collectivism-individualism. Comparison of the perception of financial and other material risks by Chinese vs. American respondents supports their assumption (coined the cushion hypothesis) that social collectivism serves as mutual insurance against catastrophic losses, making the risks faced by members of the collective, in fact, smaller.

3. Relationships between Measures of Risk Perception and Applications

The axiomatic risk measures of Sect. 2.1 were developed to describe the perception of financial risks, whereas the psychometric risk measures of Sect. 2.2 were developed to describe the perception of health and safety risks. Holtgrave and Weber (1993) investigated how well either type of measure (a simplified version of the CER model and the psychometric model) would predict risk judgments for both types of risk. Respondents, who were MBA students, provided holistic risk judgments for a range of financial and health and safety risks. They also evaluated each risk on the variables used as predictors by the two types of models (probability of a loss or gain, and expected loss or gain for the CER model; dread, control, catastrophic potential, voluntariness, novelty, and equity for the psychometric model).

Regression results showed the CER model to be a better predictor of perceived risk than the psychometric model for both financial risks and health/safety risks. The predictor variables of the CER model seem to come closer to the way people evaluate risk in both domains. The psychometric model may need a dimension reflecting the probability of harm to provide a better fit, since this dimension is highly correlated with risk ratings. Another reason might be that people consider both the pros and cons of activities when judging their risks. The CER model assumes that the impact of positive outcomes can offset the impact of negative outcomes. The psychometric model, on the

other hand, focuses exclusively on the impact of the downside of activities.

The best predictor of risk perceptions for both financial and health/safety risks, however, was a hybrid model that added the dread variable of the psychometric model to the probability and outcome-based variables of the CER model. This suggests that the perception of both financial and health/safety risks has an emotional component that is not completely described by the 'objective' components of axiomatic models (see Loewenstein et al. 2000 for a summary of the role of affect in risk perception and risk taking).

Palmer and Sainfort (1993) applied the CER model in the context of genetic counseling to model the judgments made by members of a clinical population of dwarfs about the riskiness of different procreative alternatives available to them. Faulting the genetic counseling literature for equating the perceived riskiness of an adverse event (e.g., the birth of a child with a genetic disorder) with the perceived probability of the event's occurrence, Palmer showed that the perceived risk of different reproductive alternatives also reflects the severity of consequences. Therefore, measures of risk that combine both probability and outcome information (such as the CER model) are a better predictor of risk perception.

The CER model has also been used to test the prediction of Douglas and Wildavsky's (1982) sociocultural theory that individuals with different world-views differ in predictable ways in their perceptions of risk. Palmer (1996) elicited risk judgments for financial and health/safety risks from respondents who came from either a hierarchical, individualist, or egalitarian subculture in Southern California. Hierarchists who are comfortable with determining acceptable levels of risk for technologies (Thompson et al. 1990), a process that explicitly considers and weighs gains and losses, provided risk judgments that reflected all predictor variables of the CER model (gains as well as losses, outcome levels as well as probabilities). Egalitarians, on the other hand, who are suspicious of technologies and view nature as fragile and in need of protection, suggesting that they should see risk in terms of possible harm, provided risk judgments that reflected only the loss predictor variables of the CER model. Individualists, who view risk as opportunity, given their tendency to see benefits from most activities as long as they do not interfere with market mechanisms (Thompson et al. 1990), were the group that provided the lowest risk judgments for almost all of the risky investments and activities.

4. Conclusions and Future Directions

The empirical studies of risk perception reviewed in this article show that risk perceptions are shaped by three classes of variables: the negative and positive

outcomes of choice options and their likelihood; affective reactions such as dread or fear of the unknown; and social and cultural variables that influence the perception and interpretation of the consequences of risky choice options. These variables have mostly been studied in isolation. Future research should examine possible interactions. There is also need for further examination of the accuracy or appropriateness of people's perception of risk.

The fact that ordinary citizens perceive the risk of many activities and technologies in ways that fundamentally diverge from the way they are perceived by scientists moves the choice of a measure of risk onto which to base policy decisions into the political arena. Fischhoff et al. (1984) argue that such choice should not be the exclusive province of scientists, who have no special insight into what society should value.

Bibliography

- Bontempo R N, Bottom W P, Weber E U 1997 Cross-cultural differences in risk perception: A model-based approach. *Risk Analysis* **17**: 479–88
- Douglas M, Wildavsky A 1982 *Risk and Culture: An Essay on the Selection of Technological and Environmental Dangers*. University of California Press, Berkeley, CA
- Fischhoff B, Watson S R, Hope C 1984 Defining risk. *Policy Sciences* **17**: 123–39
- Holtgrave D R, Weber E U 1993 Dimensions of risk perception for financial and health safety risks. *Risk Analysis* **13**: 553–8
- Keller L R, Sarin R K, Weber M 1986 Empirical investigation of some properties of the perceived riskiness of gambles. *Organizational Behavior and Human Decision Processes* **38**: 114–30
- Loewenstein G F, Weber E U, Hsee C K, Welch E 2001 Risk as feelings. *Psychological Bulletin* **27**: 1–20
- Luce R D, Weber E U 1986 An axiomatic theory of conjoint, expected risk. *Journal of Mathematical Psychology* **30**: 188–205
- Palmer C G S 1996 Risk perception: An empirical study of the relationship between world view and the risk construct. *Risk Analysis* **16**: 717–23
- Palmer C G S, Sainfort F 1993 Towards a new conceptualization and operationalization of risk perception within the genetic counseling domain. *Journal of Genetic Counseling* **2**: 275–94
- Slovic P 1997 Trust, emotion, sex, politics, and science: Surveying the risk-assessment battlefield. In: Bazerman M, Messick D, Tenbrunsel A, Wade-Benzoni K (eds.) *Psychological Perspectives to Environmental and Ethical Issues in Management*. Jossey-Bass, San Francisco, pp. 277–313
- Slovic P, Fischhoff B, Lichtenstein S 1986 The psychometric study of risk perception. In: Covello V T, Menkes J, Mum-power J (eds.) *Risk Evaluation and Management*. Plenum Press, New York
- Thompson M, Ellis R, Wildavsky A 1990 *Cultural Theory*. Westview Press, Boulder, CO
- Weber E U 1997 The utility of measuring and modeling perceived risk. In: Marley A A J (ed.) *Choice, Decision, and Measurement: Essays in Honor of R. Duncan Luce*. Erlbaum, Mahwah, NJ

- Weber E U, Bottom W P 1989 Axiomatic measures of perceived risk: Some tests and extensions. *Journal of Behavioral Decision Making* 2: 113–31
- Weber E U, Bottom W P 1990 An empirical evaluation of the transitivity, monotonicity, accounting, and conjoint axioms for perceived risk. *Organizational Behavior and Human Decision Processes* 45: 253–75
- Weber E U, Hsee C K 1999 Models and mosaics: Investigating cross-cultural differences in risk perception and risk preference. *Psychonomic Bulletin & Review* 6: 611–7
- Yates J F, Stone E R 1992 Risk appraisal. In: Yates J F (ed.) *Risk-taking Behavior*. Wiley, Chichester, UK

E. U. Weber

Risk Screening, Testing, and Diagnosis: Ethical Aspects

The Human Genome Organization project has increased enormously the possibilities to predict disability or disease either for an individual or for his or her offspring. Information obtained from genetic testing raises ethical and legal issues that have been debated for several decades. New issues are emerging due to rapid advances in molecular genetics.

From an ethical point of view, risk screening, genetic testing, and diagnosis belong to the most controversial and contested areas in contemporary medicine. The new genetics will have a profound impact on our self-images and on our understanding of the relations between environment and heritage. Conceptual, normative, and empirical issues are here intertwined in a complex way. To begin with, a preliminary clarification of the key concepts will be necessary.

The expression 'risk' will here be taken to refer to an adverse or negative future event that is probable but not certain. Any risk can then be analyzed in two components: the probability of occurrence of an event and the magnitude of its negativity according to certain explicitly stated norms and values (including kind, degree, duration, and timing). Risk assessment accordingly analyzes these two components, while risk management proposes strategies for handling the risk in ways that minimize the negativity and probability.

However, risks involving probabilities are not the only sort of risks that are relevant in this context. The expression 'epistemic risk' has been coined to refer to risks due to ignorance, and in particular due to ignorance of what we are ignorant about (e.g., Sahlin and Persson 1994). This notion is particularly relevant to many clinical applications of molecular genetics, involving unknown factors. The ways in which these risks are communicated will then be important from an ethical point of view.

Testing should be distinguished from screening.

Harper has proposed the following definition: 'Genetic testing is the analysis of a specific gene, its product of function, or other DNA and chromosome analysis, to detect or exclude an alteration likely to be associated with a genetic disorder' (Harper and Clarke 1997). Testing refers to procedures performed at the request of individuals or families, and should always be voluntary. This holds also for genetic testing of members of families known to be at high risk, such as the siblings of persons with cystic fibrosis (Wertz and Fletcher 1989).

'Screening,' however, implies application to large population groups, to entire populations or subsets of populations (e.g., pregnant women, newborns, or job applicants). Screening is not performed at the request of individuals or families but rather based on policy or public health decisions. It may be mandatory or voluntary.

Both the expressions 'genetic information' and 'genetic disorder' are vague and ambiguous concepts (Zimmern 1999). The same holds for the key concepts of health, disease, and illness. The advances in molecular and clinical genetics may profoundly change our understanding of the latter concepts. They may have an impact on health care policy, on the patient–physician relationship as well as the way in which access to care is allocated (Murray et al. 1996).

Genetic information differs from other private information in that it reveals information not only about a particular individual but also about that individual's blood relatives. Genetic test results may also provide information about population groups. Moreover, genetic information can be obtained from any cell in a person's body, not just by examining, for example, particular malfunctioning organs or tissues.

The concept of genetic information is difficult to define and delimit in a clear and uncontroversial way. Just by asking for the age of a person's parents and grandparents, genetic information in a certain sense can be obtained. The necessary and sufficient conditions for a disease or disorder to be 'genetic' are far from clear. If these expressions 'genetic information' and 'genetic disorder' are to appear in legislative texts, they have to be made precise.

There are many kinds of genetic disorders, including (1) chromosome changes, like trisomy 21 or Down's syndrome, (2) monogenetic disorders which depend on a mutation in one gene, like Huntington's chorea, and (3) polygenetic disorders like diabetes which are due to mutations in several genes in combination with environmental factors.

These distinctions are important for several reasons. What holds for monogenetic diseases concerning genetic determinism cannot be generalized to genetic disorders of other kinds. Thus, presymptomatic testing can be done for disorders in group (2) but rarely and with much less certainty, if at all, for those in group (3).

1. Public Concerns

Genetic testing may be beneficial in those cases where early detection makes a difference, for example, makes it possible to prevent or cure the disease. The validity and the reliability of the test methods will then be of crucial importance. The background of many national screening programs, like the one launched in Ontario, Canada, in 1999, to detect risk of hereditary breast, ovarian, and colon cancers is that early detection can lead to medical interventions and surveillance. This can significantly reduce the risk of developing the disease and hence be cost-effective.

The concern expressed over genetic testing and screening, however, is that it may threaten privacy and civil rights and be the basis of genetic discrimination or stigmatization of individuals as well as of groups. Moreover, it has been feared that widespread use of genetic testing will help to foster intolerance of people with genetic variations. It may be difficult, too, if not impossible, for people with certain inherited disorders to obtain a life insurance. Finally, people may feel a pressure to be tested, as a result of cost-benefit analysis. In general, of course, the negative consequences of testing people and disclosing information about their test results to blood relatives will have to be weighed against the benefits of this, if any, for those concerned. However, this weighing operation is far from simple.

Guidelines for genetic testing and screening have been proposed in the attempt to maximize the beneficial consequences and minimize the adverse ones (by e.g., UNESCO 1997, Council of Europe 1992, 1996, Nuffield Council on Bioethics 1993, Danish Council of Ethics 1993). These guidelines include discussions of genetic counseling, informed consent, disclosure to individuals and family members, confidentiality, employment, insurance, as well as public policy. The commercial availability of genetic test kits raises both operational and ethical issues, dealt with in the principles and recommendations suggested by National Institutes of Health (1997). This report discusses the validity of genetic testing, deals with genetic counseling and the informed consent process, as well as recommendations on quality assurance measures for laboratories performing genetic tests and the need for a national body with the authority to review genetic testing procedures.

2. Ethical Issues

It has often been recommended that genetic information in health care should be acquired and used 'in a manner that respects the autonomy of individuals' (Institute of Medicine 1994), but it has proved hard to explain the precise meaning of this in a noncontroversial way. The autonomy of what individuals should be respected? Sometimes respecting the autonomy of

one individual may conflict with respecting the autonomy of another.

A general problem raised by all forms of genetic testing and screening is, first, how to give nondirective information, that is, information that makes autonomous decisions possible by individuals and couples, and then to obtain consent in a way that avoids exercise of pressure.

Why is nondirectiveness an important goal? In brief, the reasons are the value of autonomy and the importance for geneticists to dissociate themselves from the dark history of genetics, from abuse of genetics for eugenic purposes. But nondirectiveness in practice is not uncomplicated, and the goal is also controversial. There is now a debate on the process of genetic counseling beyond nondirectiveness (Clarke 1994, Harper and Clarke 1997).

Particular problems are raised by the testing of individuals or groups with reduced autonomy, like children, demented people, or immigrants who do not understand the language. The testing of a healthy child could arise in at least three different contexts (Chadwick 1998): predictive testing, to see if the child will develop a specific disease that runs in the family; carrier testing, to see if the child may be at risk of having an affected child; and a screening test unrelated to family history to identify children with increased susceptibility to some common disease such as diabetes or ischaemic heart disease. Children are vulnerable and their parents therefore have to be involved in the decision procedure. But how? The nature of parental consent and the degree to which children are to be involved according to their maturity need close examination. If relatives are to be informed, who should do this? The geneticist, the hospital, or those already tested? And in what way? How many relatives should be informed? How closely related to the person asking to be tested (the index person) should the relatives be for the clinical geneticist to be responsible for informing them?

Different consent models are available, each with their own pros and cons. One main alternative is called 'opt in,' which means that if you have not said yes, you have said no; another is called 'opt out,' which means that if you have not said no, you have said yes. The conditions under which any of these models may or must be used is an important ethical issue.

The challenge to genetic privacy concerns the control of personal genetic information. Privacy has two dimensions, according to the Canadian Human Rights Commission (1992): protection from the intrusion of others and protection from one's own secrets. Does this entail a right to genetic ignorance, a right not to know about one's own or one's relatives' genetic risk profile? Is there a duty for geneticists or tested persons to disclose such information to those who might benefit from it? What is the basis of such rights and duties, and what are the consequences for the various people concerned of accepting or rejecting

them? Opposite positions have been argued for by Rhodes (1998) and Takala and Hayry (2000). Several compromise positions have been put forward. Tested persons should be encouraged to share what they know about their genetic risk profile with their blood relatives, when this information would be of vital health interest to their relatives (Chadwick et al. 1997, Hermerén 2000).

The privacy and confidentiality of the information needs to be protected against third party interests, including other family members, employers, and insurers. It has been suggested that genetic information should not be given to unrelated third parties, without the explicit and informed consent of the tested persons. In many countries, access to genetic information is regulated by law. The conditions or criteria under which medical confidentiality may or must be breached are bound to be controversial (Harper and Clarke 1997).

The practical problems raised by genetic testing also have to do with how to estimate and evaluate risks, communicate them in an understandable way, and decide which risks are acceptable and which are not. This includes the risk that a person will have an increased risk for a genetic disorder and the risk that the test result is not accurate. Communicating risks is obviously an important part of the genetic counseling process.

The accuracy of the tests—their validity and reliability—and the predictive value of the test or the screening method is another important issue in this context. To estimate the predictive value, one has to be able to compute the number of false positives and false negatives as well as to know about the prevalence of the genetic disorder.

Finally, there are issues of justice and fair access. If testing and screening is to be done on an equitable basis, it should not be available on the basis of such accidents as in which geographical region a person happens to be born.

These ethical problems arise in somewhat different ways, according to the type and purpose of testing or screening. The difficulties are not the same in all cases. In particular, certain forms of screening and presymptomatic testing for late onset genetic disorders present problems from an ethical point of view. Some of the most common types of genetic testing and screening will therefore be described below along with a brief indication of the particular ethical problems they raise.

3. Types of Testing and Screening

3.1 Presymptomatic Testing

Individuals may be tested for monogenetic disorders (of which there are many though each is fairly rare) by

DNA analysis before any symptoms are shown. If an early diagnosis is of value, because the disease can be prevented, presymptomatic testing is fairly uncontroversial. If the disorder cannot be prevented, the value of presymptomatic testing is more difficult to judge. This holds also in cases where there is a method of prevention but it is nevertheless impossible to guarantee that the person tested positively will not get the disease (for example, breast cancer).

In addition to the obvious difficulty of informing people in such a way that they can make autonomous decisions, presymptomatic testing raises problems about how the information obtained from the tests should be handled, and who should have access to it. Likewise, it is controversial whether blood relatives should be contacted and who should contact them in that case (the person tested, the hospital, the clinic, the geneticist, etc.). Moreover, in many countries there is great concern about how to prevent certain external third parties, like employers or insurers, from getting access to information about the test result.

3.2 Prenatal Diagnosis

Prenatal diagnosis can be regarded as a special case of presymptomatic testing. By prenatal diagnosis it is possible to discern whether a fetus is at risk for various disabilities or diseases.

There are several methods for prenatal diagnosis. Some are visual—either noninvasive, like ultra-sonography, or invasive, like embryoscopy or fetoscopy, which uses a camera on a needle inserted in the uterus to view the fetus. Other methods are based on analysis of fetal tissues (like amniocentesis or chorionic villus sampling). By ultrasound tomography it is possible to create fetal images on a screen.

The indications of when prenatal diagnosis should be offered include advanced maternal age, a genetic history of abnormalities in the family, repeated miscarriages, or previous infants with birth defects. Then a prenatal diagnosis may be offered to provide information to couples about what they can expect. A normal result is found in most cases. If not, prenatal diagnosis may make it possible for them to adjust and may also influence the way in which the baby is delivered.

There has been great concern about the risk of spontaneous abortion induced by the use of amniocentesis or chorionic villus sampling. This is a problem one does not have with ultrasonography. On the other hand, ultrasonography may give information that is unexpected or not wanted, for example, that some organs or limbs are malformed. With other methods, too, one may make unexpected or unwanted discoveries, for example, that the social father is not the biological father.

All this serves to underline the necessity of careful counseling before the diagnosis, so that the woman or

the couple knows what the test may show and how certain the result is. Informed consent is crucial, and the difficulties of nondirective counseling should not be underestimated.

Other ethical issues raised by this type of diagnosis are concerned with how to preserve the confidentiality of the information obtained. Since prenatal diagnostics sometimes leads to abortion, it may be controversial in the eyes of some people for the same reasons that abortion is controversial. The impact of prenatal diagnosis on our views of people with handicap has also been a matter of concern.

3.3 Preimplantation Diagnosis

Preimplantation diagnosis means that the diagnosis is carried out before (in connection with *in vitro* fertilization) the pre-embryo is implanted in the woman's uterus. If there were a history of sex-linked diseases in a family, it would then be possible to prevent a child with such a disease being born. Thus the resulting pregnancy could be normal and the woman would not have to abort the fetus at a later stage or give birth to a child with a very serious or lethal disease. Preimplantation diagnosis is becoming an established medical practice in many countries.

The first serious ethical problem that arises is to whom these tests should be offered. Usually, they are limited to severe illnesses and to persons in specific 'high risk' groups. But both these criteria are value laden, and the indications may be narrowed down or widened. There are also considerable variations not only between lay people but also between geneticists over the severity of different genetic conditions (Chadwick 1998).

Preimplantation diagnosis at the beginning of the twenty-first century is thus limited to sorting out fertilized eggs carrying genes associated with severe diseases for which there is no cure at present, like Huntington's chorea and certain X-linked diseases. Concern has been expressed that the practice may be extended to diseases that are less serious (Shenfield 1997). Obviously, there is a risk of a slippery slope in this area.

Preimplantation diagnosis also, at least theoretically, opens up possibilities for a shift of power from the woman to the geneticist, as well as for cloning and manipulation of human pre-embryos that many people find too risky or simply ethically unacceptable. In some countries, like the UK, manipulation of human pre-embryos is illegal.

3.4 Prenatal and Neonatal Screening

Newborn screening involves the analysis of blood or tissue samples taken in early infancy in order to detect inherited diseases where early intervention can mini-

mize or eliminate the risk of later serious health problems or premature death.

In screening, a group of individuals is examined directly with DNA, RNA or biochemical analysis to find those who will later in life have an increased risk of developing a particular genetic disorder. A good screening method is reliable and valid, with a minimum of false positives (who are worried unnecessarily) and false negatives (who are lulled into false security).

The best-known example of uncontroversial neonatal screening is screening for phenylketonuria (PKU), a genetic disease. The reason is that PKU, if undetected, can lead to serious mental retardation. Moreover, it can be discovered easily and prevented through early intervention. For many other conditions, such as cystic fibrosis and galactosemia, the benefits of early detection are more controversial. Screening programs for newborns also raise the issues of freedom of choice and informed consent by parents (Hermerén 1999).

Newborn screening raises several ethical issues, among them the nature of informed parental consent and how it is obtained, confidentiality and privacy, and the interests of third parties, as well as issues of justice such as equal access to testing and treatment. Moreover, it is clear that more research needs to be carried out on the psychological responses to newborn screening, particularly when the health benefits are disputed (Harper and Clarke 1997).

People with certain genetic disorders have had difficulties in getting a life insurance. In some countries the right of insurance companies to ask for genetic information has been regulated by law. In Norway, it is forbidden. In other countries—The Netherlands and Sweden, for instance—there has been a temporary voluntary moratorium during which the insurance companies will not ask for genetic information, provided that the life insurance asked for will not exceed a certain amount. This is still a very controversial area (Sandberg 1996).

3.5 Carrier Screening

Carrier screening identifies individuals with a gene or chromosome abnormality that may cause problems either for offspring or for the person screened. The testing of blood or tissue samples can indicate the existence of particular genetic trait changes in chromosomes or changes in DNA that are associated with inherited diseases in asymptomatic individuals. Carrier screening exists for sickle cell anemia, for Tay Sachs disease, as well as for cystic fibrosis, Duchenne muscular dystrophy, hemophilia, and Huntington's chorea.

Carriers have one normal copy of a gene and one copy that varies from the normal gene. Since one gene is normal, the carriers ordinarily do not exhibit symptoms of a genetic disorder. A woman who is a

carrier of an X-linked disease such as Duchenne muscular dystrophy or fragile X syndrome, has a variant gene on one of their two X (sex-determining) chromosomes. Her risk of transmitting the variant gene to each child can be estimated accurately.

The purpose of carrier testing is to inform potential parents of their genetic risk profile so that they can make informed reproductive decisions. National screening programs are usually justified by cost-benefit analysis. The question is then which costs and whose costs are taken into account, and how the benefits are defined and evaluated. The importance of cultural differences as to what constitutes benefits, and how benefits are ranked, should not be underestimated.

The right to autonomous choices is an important issue also in this area. Whether the choices open to people found to carry a disease gene are acceptable or not, may depend on their life plans, their ethical views, the laws of their country, as well as their culture and social customs. Before a carrier screening program is decided on and implemented, these questions need to be considered. In practice there may be subtle coercion, even if a carrier screening program is formally voluntary.

Screening the subset of a population at risk rather than the entire population makes sense from a medical and economic point of view, but selecting target groups for such screening programs also creates ethical problems. For example, there have been in the past controversies over whether these screening programs in effect had a eugenic purpose, particularly when the diseases screened for are over-represented in certain ethnic groups.

3.6 Testing in the Work Place

Susceptibility screening can be used as a form of presymptomatic testing, but is also used to identify workers who may be susceptible to toxic or carcinogenic substances such as benzene in their workplace (Van Damme et al. 1997, Surrallés et al. 1997, Jarvholm et al. 1999). This susceptibility may lead to occupational health problems and future severe disabilities.

Van Damme et al. (1995) have proposed a conceptual model of the complex interactions between exposure, acquired and inherited susceptibility, and risk for disease. The validity of tests for determining genotype and phenotype and their relevance for the disease must be evaluated critically to provide an objective basis for ethical discussions. The acceptability of such tests is related to a number of issues which Van Damme et al. identify and discuss.

Genetic testing of employees and job applicants is controversial. The first basic question concerning testing in the work place is simple—for whose sake is the test performed? Is the test performed in order to protect persons with increased (inherited) suscepti-

bility to hazards at the workplace? Or is the test in the interest of the employer, so that the employer can avoid improving the working conditions and avoid hiring people who would be damaged by the work they do? Here it is essential to separate the situation of the job applicant from the situation of the person already with a job. The latter could be monitored for exposures and early effects of, for example, toxic substances on the genes.

There are methods of detecting inherited predisposition to cardiovascular diseases as well as to adverse effects from otherwise well-tolerated exposures related to the job. Clearly, with widespread use of genetic testing in the workplace there is the obvious danger of discrimination and stigmatization of individuals or groups. It may be tempting for employers to use tests to select only workers who will be able to work in the present working conditions, instead of improving the working conditions. The result could be that large groups of people would be barred from the job market.

3.7 Forensic Testing

A comparatively new use of genetic testing is to examine a possible genetic linkage between, for example, alleged fathers and their children, or between suspects and evidence discovered in criminal investigations, like hair, blood, sperm, saliva, or skin. The idea is to use genetic methods of analysis to clear the innocent and identify the guilty. However, concern has been expressed that mistakes can be made. Critics argue that it is necessary to improve the quality control of the test laboratories. Another problem is the confidentiality of the DNA profiles obtained from criminal investigations and stored in national police databanks. For how long time will they be stored? Who will have access to them and for what purpose?

4. Concluding Remarks

Genetic testing and screening should be done for 'health purposes.' This is sometimes stated as a necessary condition for genetic testing (Council of Europe 1992, 1996). Health purposes could include diagnostic and therapeutic purposes as well as medical research and health care planning. Moreover, 'health' can be defined in several ways. It is important to state more precisely what this means.

Careful counseling remains crucial. One of the present concerns is that do-it-yourself test kits for genetic disorders may be bought via the Internet or over the counter without any professional counseling to explain what the findings mean, how certain the test is, etc. There is an urgent need to find ways of regulating the economic interest of commercial biotechnological companies in order to minimize the harm of premature use and overuse of genetic testing.

See also: Bioethics: Examples from the Life Sciences; Eugenics as an International Movement; Eugenics as the Basis of Population Policy; Genetic Counseling: Historical, Ethical, and Practical Aspects; Genetic Screening for Disease-related Characteristics; Genetic Testing and Counseling: Educational and Professional Aspects; Reproductive Medicine: Ethical Aspects

Bibliography

- Canadian Human Rights Commission. Privacy Commissioner 1992 *Genetic Testing and Privacy*. Privacy Commissioner of Canada, Ottawa, ON
- Chadwick R 1998 *The Ethics of Genetic Screening*. Kluwer, Amsterdam, The Netherlands
- Chadwick R, Levitt M, Shickle D 1997 *The Right to Know and the Right Not to Know*. Avebury, Aldershot, UK
- Chadwick R, H ten Have, Husted J, Levitt M, McGleenan D, Shickle D, Wiesing U 1998 Genetic testing and ethics: European perspectives. *Journal of Medicine and Philosophy* **23**: 255–73
- Clarke A (ed.) 1994 *Genetic Counselling. Practice and Principles*. Routledge, London
- Clarke A, Flintner F 1996 The genetic testing of children: A clinical perspective. In: Marteau T M, Richards M P (eds.) *The Troubled Helix: Social and Psychological Implications of the New Human Genetics*. Cambridge University Press, Cambridge, pp. 164–7
- Council of Europe 1992 *Draft Recommendation on Genetic Testing and Screening for Health Care Purposes*. Council of Europe, Strasbourg, France
- Council of Europe 1996 *Convention on Human Rights and Biomedicine*. Council of Europe, Strasbourg, France
- Danish Council of Ethics 1993 *Ethics and the Mapping of the Human Genome*. Danish Council of Ethics, Copenhagen, Denmark
- Harper P S, Clarke A J 1997 *Genetics, Society and Clinical Practice*. Bios, Oxford, UK
- Hepburn E 1996 Genetic testing and early diagnosis and intervention: Boon or burden? *Journal of Medical Ethics* **22**(2): 105–10
- Hermerén G 1999 Neonatal screening: Ethical aspects. *Acta Paediatrica Suppl.* **432**: 99–103
- Hermerén G 2000 The right to know and the right not to know. In: *Who Owns Our Genes?* Danish Council of Ethics, Copenhagen, Denmark, pp. 135–57
- Institute of Medicine 1994 (Andrew L B (ed.)) *Assessing Genetic Risks: Implications for Health and Social Policy*. National Academy Press, Washington, DC
- Jarvholm B, Nordstrom G, Hogstedt B, Levin J O, Wahlstrom J, Ostman C, Bergendahl C 1999 Exposure to polycyclic aromatic hydrocarbons and genotoxic effects on nonsmoking Swedish road pavement workers. *Scand. J. Work Environ. Health* **25**(2): 131–6
- Murray T H 1983 Warning: screening workers for genetic risk. *Hastings Center Report* **13**(1): 5–8
- Murray T H, Rothstein M A, Murray R F (eds.) 1996 *The Human Genome Project and the Future of Health Care*. Indiana University Press, Bloomington, IN
- National Institutes of Health (National Human Genome Research Institute) 1997 *Promoting Safe and Effective Genetic Testing in the United States: Principles and Recommendations* [the report is available on the Internet at http://www.nhgri.nih.gov/ELSI/TFGT_final]
- Nuffield Council on Bioethics 1993 *Genetic Screening: Ethical Issues*. Nuffield, London
- Rhodes R 1998 Genetic links, family ties, and social bonds: Rights and responsibilities in the face of genetic knowledge. *J. Med. Philos.* **23**(1): 10–30
- Sahlin N-E, Persson J 1994 Epistemic risk: the significance of knowing what one does not know. In: Brehmer B, Sahlin N-E (eds.) *Future Risks and Risk Management*. Kluwer, Dordrecht, The Netherlands, pp. 37–62
- Sandberg P 1996 *Genetic Information and Life Insurance: A Proposal for an Ethical European Policy*. Norwegian University of Science and Technology, Department of Biotechnology, Trondheim, Norway
- Shenfield F, Sureau C 1997 Ethics of embryo research. In: Shenfield F, Sureau C (eds.) *Ethical Dilemmas in Assisted Reproduction*. Parthenon, New York, pp. 15–21
- Shenfield F, Sureau C (eds.) 1997 *Ethical Dilemmas in Assisted Reproduction*. Parthenon, New York
- Surrallés J, Autio K, Nylund L, Jarventus H, Norppa H, Veidebaum T, Sorsa M, Peltonen K 1997 Molecular cytogenetic analysis of buccal cells and lymphocytes from benzene-exposed workers. *Carcinogenesis* **18**(4): 817–23
- Takala T, Hayry M 2000 Genetic ignorance, moral obligations and social duties. *J. Med. Philos.* **25**(1): 107–13, Discussion 114–20
- UNESCO 1997 *Universal Declaration on the Human Genome and Human Rights*. UNESCO, Paris
- Van Damme K, Casteleyn L, Heseltine E, Huici A, Sorsa M, van Larebeke N, Vineis P 1995 Individual susceptibility and prevention of occupational diseases: Scientific and ethical issues. *J. Occup. Environ. Med.* **37**(1): 91–9
- Van Damme K, Vineis P, Sorsa M, Casteleyn L 1997 Ethical issues in genetic screening and genetic monitoring of employees. *Ann. N. Y. Acad. Sci.* **837**: 554–65
- Van Damme K, Casteleyn L 1998 Ethical, social and scientific problems related to the application of genetic screening and genetic monitoring for workers in the context of a European approach to health and safety at work. *La Medicina del Lavoro*, Vol. 89, suppl. 1
- Wahlstrom J 1989 Inherited mental disorders. *Acta Psychiat. Scand.* **80**(2): 111–17
- Wertz D, Fletcher J 1989 *Ethics and Human Genetics: A Cross-cultural Perspective*. Springer, Berlin
- Zimmerman P 1999 Genetic testing: A conceptual exploration. *Journal of Medical Ethics* **25**(2): 151–6

G. Hermerén

Risk, Sociological Study of

That the world is punctuated with uncertainty, danger, hazards, disasters, and threats to what humans value has been recognized since the beginning of human life on earth. Risk is the conceptualization of that recognition—providing a means for anticipating and avoiding or adapting to untoward outcomes. By 3200 BC, ancient Babylon had the first recorded risk assessors, and in 1700 BC, the Code of Hammurabi included a classification of risk matters. Despite these

ancient beginnings, however, and despite subsequent refinements accompanying the rise of the insurance industry during the age of exploration, the centrality of risk is the child of advanced industrialization, particularly since World War II. With advanced industrialization came technologies that were larger, more sophisticated, and potentially much riskier, having the capacity to affect far greater numbers of people, as well as the things they value and the areas they occupy. At the same time, scientific and technological advances have led to remarkably increased sophistication and precision, in many cases, in the ability to detect unwanted side effects or risks.

The notion of 'risk' has evolved over the last decades of the twentieth century to become the key analytical lens for anticipating our actions' consequences for the environment and ourselves. It includes an analytic orientation and a suite of evaluation methodologies, but it is also a new consciousness—a way of looking at a world of technological and environmental uncertainty. Attention to risk is a consciousness born of vast uncertainties over the durability of nuclear peace, the resilience of the ozone protecting us, threats of global warming, the growing extinction of species and the possibility of creating new ones, and more broadly, threats of technological disasters both large and small. It is a consciousness entirely foreign to the optimistic beginning of the century just ended.

The first significant body of sociological work on risk can be traced to the 1940s, with the work on natural hazards by Gilbert White (a social geographer), gaining momentum in the 1950s with White's collaboration with a number of sociologists, and evolving into the sociological specialty of natural disaster research. A key focus of this specialty has involved the social impacts of hazard events—how did structures and individuals respond to natural disasters such as storms, floods, earthquakes, wildfires, or droughts? One of the most consistent findings from this literature was the emergence of what has become labeled a 'therapeutic community': contrary to the self-focused maximizing predictions of utility theory, citizens typically engaged in acts of altruistic aid and community building (for reviews, see Drabek 1986, Kreps 1984).

Particularly in the United States, a significant fraction of the ongoing interest in natural disasters was motivated by the concerns of the early atomic age, including the prospect that a nuclear attack could wipe out a significant fraction of a community's population; given that much the same could be said for natural phenomena, such as floods, fires, or hurricanes, these and other 'natural disasters' came to be seen as useful 'natural experiments' for studying community responses. Particularly after the accident at the Three Mile Island nuclear power plant in 1979, however, it was a different form of nuclear activity that came to provide what is in many ways the prototypical 'risk' issue.

Although the point is often forgotten today, nuclear power facilities once enjoyed widespread support, even in the communities where they were constructed. Still, even before the 1979 accident, there were early signs of rising public concern. One response of nuclear power supporters (including those who controlled some of the most powerful branches of government) was to characterize the concerns as being ill-founded or 'irrational,' doing so by demonstrating that the 'real risks' were instead quite small. The earliest efforts, however, indicated that the consequences of a plausible 'worst-case' disaster could have been quite serious, indeed. These early efforts led to little publicity, but to a change of focus—an emphasis on what soon came to be called 'probabilistic' risk assessment, or PRA, a reductionistic engineering technique that estimates the risk of a plant mishap by estimating the probability of failure of all the plant's parts and subsystems and then aggregating these probabilities. The conclusion from the first systematic application of this technique to nuclear power plants, which was stated particularly forcefully in the executive summary for the multi-volume effort that is still remembered as 'the' reactor safety study (US Atomic Energy Commission 1975), was that citizens were more likely to be killed by a falling meteorite than by a malfunction at a nuclear power plant.

Four years later came the accident at Three Mile Island, near Harrisburg, the capital city of the state of Pennsylvania—an accident that by technical analysis virtually could not happen. The accident brought to public and sociological attention an incipient schism between the state, its technological experts, and citizens. European sociological interest in risk was similarly intensified by the far more serious nuclear accident at Chernobyl in April 1986. While at one level of analysis these accidents and the vast media coverage surrounding them offered dramatic signals to publics of the risks of nuclear-generated electricity, at a deeper level they revealed conflicts between expert and lay knowledge, and between 'top-down' and 'bottom-up' strategies for managing large-scale, risky technologies.

The 1979 accident, in particular, generated a number of legacies. First among them was the further institutionalization of what soon came to be known as the field of risk analysis: the Society for Risk Analysis was established in the United States (with subsequent affiliates in Europe and Asia), as was the Division of Risk and Management Science at the US National Science Foundation, along with a number of new professional journals and a substantial outburst of published literature. That outburst included a significant increase in sociological work, with much of the initial increase paralleling the three general theoretical levels in sociology—the micro, the meso, and the macro, or in simpler terms, the social psychological, the organizational/community level, and the societal.

At the micro level, much of the work was conditioned by what soon became the dominant social

science research tradition in risk studies: the psychometric tradition (Slovic 1987). A growing body of work in cognitive psychology was brought to bear in examining the extent to which lay perceptions of risk differed from actuarial data and from experts' views. Key among the findings from this work, replicated in a variety of cultures, is the idea of heuristics—cognitive rules of thumb that provide shortcuts to risk estimation. One example is the so-called 'availability heuristic,' or the tendency to judge a risk on the basis of its cognitive 'availability,' or the ease of imagining or recalling examples; such a tendency increases the perceived riskiness of a low-probability but sensational risk, such as death from botulism, relative to more probable but more prosaic risks, such as deaths from asthma.

In political circles, the early psychometric findings were sometimes taken as reinforcing the argument that the 'irrational' public should be excluded from risk decisions, but the evidence proved not to be so clear-cut. Sociological work on technological controversies, for example, had already drawn attention to the lack of unanimity among experts (see e.g., Mazur 1980), and further research showed that experts relied on similar heuristics—even in the contexts of their own expertise, and even when the experts were studied with the same protocols used to study lay perceptions. Work from a social constructivist perspective (e.g., Wynne 1992) further challenged the tendency to equate expert risk estimates with 'real' risks, with Clarke (1999) eventually going so far as to argue that many official proclamations were little more than 'fantasy documents'—efforts to deal with ultimately uncontrollable risks through ritualistic proclamations of rationality. At the same time, psychometric work pointed to the importance of a number of broader issues, such as trustworthiness and fairness, that were omitted from technical risk analyses. Hence, from the psychological literature emerged a reaffirmation of sociology's first principle: context matters.

Sociologists responded with three sustained research efforts—one that contextualized the individual actors, a second that aligned risk choices with the most powerful technological decision-makers, namely organizations, and a third that examined the consequences of technological accidents in the contexts where people live, namely communities—as well as with broader and more conceptual work at the macrosocietal level.

At the individual level, the work showed that a social actor's perception and responses to risk are shaped not just by heuristics, but also by an entire set of social, political, and institutional forces. Perhaps the most comprehensive effort to capture this wide variety of psychological and social forces is the 'social amplification' framework, a conceptualization adapted from formal communication theory (Kasperson et al. 1988), and emphasizing that, when signals are sent

from a source to a receiver, those signals often flow through intermediate transmitters that 'amplify' or attenuate the message. Similarly, individuals can receive risk signals either directly, through personal experience, or indirectly, through information provided by institutions (e.g., the media) or social networks (e.g., opinion leaders). These 'amplified' interpretations can then produce behavioral responses, resulting in secondary impacts that, in turn, are often themselves amplified, producing tertiary impacts. Partly because of its interdisciplinary orientation and its capacity to integrate the cumulative psychological findings with other factors that have been shown to influence risk perception and response—social, organizational, institutional, political—the framework has generated a considerable body of generally supportive empirical research.

With mesolevel work on organizations, meanwhile, sociologists began to point out the inherent metamethodological bias of early PRA work—the Cartesian reductionism of assuming that one can understand risk by understanding its underlying fundamental elements. A counterpoint to this methodological individualism was provided by the sociological holism of organizational sociology, treating organizations, not individuals, as the key risk decision-makers in contemporary societies.

The landmark contribution to this orientation is Perrow's *Normal Accidents* (1984), which argued that certain technological systems were complicated beyond our understanding of them—and beyond our ability to make them safe. Key factors included complexity and 'tight coupling'—essentially the lack of room for error—meaning that even attempts to improve safety, such as the introduction of redundant parts and backup systems, could result in further complication of these systems, thereby increasing risk. The book argued that some of our technological systems—nuclear power plants, nuclear weapons, and perhaps others—were predictable 'system' accidents, or 'normal accidents,' waiting to happen.

A counterpoint to this argument emerged in subsequent work on 'high reliability organizations' (LaPorte and Consolini 1991). Building on a set of fieldwork-based case studies—e.g., an air traffic control system, a large electric utility system, aircraft carrier operations—this school of thought argues that complex systems can be made safe with proper component, system, and organizational design, with specialized, sometimes variable, management structures, and through the development of a corporate culture of safety. Given that a great many of society's complex technological systems do indeed operate daily without incident, but also that other empirical tests appear to support normal accident theory, the jury remains out over which of these two schools of thought offers the deeper purchase on this key problem of advanced industrialization. The issue is only likely to be resolved with further refinement of the conditions

under which 'normal' accidents or 'high reliability' operations are more likely.

Another body of mesolevel work focused on community contexts. Particularly after Erikson's study of a so-called 'man-made' or 'non-natural' disaster (1976), a growing number of studies found that the impacts of technological or non-natural disasters proved to be far more severe and long-lasting than those associated with natural disasters. In contrast to the 'therapeutic community' of natural disasters, numerous studies of technological disasters have encountered what Freudenburg and Jones (1991) call a 'corrosive community,' typified not by offers of help, but by efforts to avoid responsibility or affix blame. Similarly, while the long-term social, economic, psychological, and cultural impacts of natural disasters have generally been found to have been surprisingly small, the long-term impacts of technological disasters have been found to be far greater than would be expected on the basis of the immediate physical destruction. At Three Mile Island, for example, most observers have concluded that only small amounts of radiation escaped, but careful studies found increased reports of 'intense distress symptoms'—that is, levels of distress symptoms that are characteristic of hospitalized mental patients—with later studies finding that stress levels were actually higher some six years later than they were in the immediate postaccident period. The symptoms were not merely self-reported, but included physiological measures such as elevated catecholamine levels in blood—a form of medical measure that is generally considered 'real' even by observers who question the legitimacy of people's own reports on stress and well-being. More broadly, several studies indicate that distress levels may actually be highest among the citizens afflicted not by the worst levels of disaster, but by an *ambiguity of harm*—cases where not even the best scientific work can clearly demonstrate whether people have been seriously endangered or not (Freudenburg and Jones 1991).

At the macro- or societal level, finally, a growing body of work has responded to the call of Short (1984), in his presidential address to the American Sociological Association, to examine the importance of risks to the social fabric itself. In the early days of the twentieth century, Weber had pointed out that the 'rationality' did not mean that the citizens of modernity would know more about their surroundings and tools than the citizens of a premodern age; instead, Weber emphasized, the modern citizen would see the tools and technologies as being 'knowable,' in principle, rather than being in the realm of magic. A century later, Weber's point is, if anything, more clearly true. Collectively, we know far more about our world and tools than did our great-great-grandparents, but individually, we actually know far less about the tools and technologies on which we depend; instead, we literally 'depend on' them to work properly. In turn, that means that we depend on whole

armies of specialists, most of whom we will never meet, let alone be able to control. Even if we find that we can usually depend on the technologies, and on the people who are responsible for them, the rare exceptions can be genuinely troubling. The risks to the social fabric, in other words, include what Freudenburg (1993) has termed 'recreancy'—cases where an organization or an individual entrusted with a specialized task fails to perform in a way that fulfills the responsibility and upholds the trust.

Still more broadly, while differing in ontological and epistemological assumptions, and while differing in orientation and research method, the various perspectives share the view that risk is an identifiable feature of the world, whether real or constructed, that risks are consequential in many facets of social life, and accordingly, the topic is worthy of sociological examination as well as policy attention (Rosa 1998). In addition, the pragmatic importance of new approaches to risk management reflects the growing recognition that the modern world has generated risk problems that demand scientific understanding, but that are too complex or too ambiguous to be 'answered' by science alone. Both the scientific examination and the policy making must take place in a context that is permeated by values, and also by unknowns—both those that can be recognized in advance and those that cannot. Owing to these conditions, democracies may have entered an era of postnormal risk (Funtowicz and Ravetz 1992), an era demanding better procedures for managing growing technological risks, procedures that require scientific inputs, but that broaden the scope of peer evaluation and integrate citizen involvement. A variety of social experiments have been undertaken to refine and implement these analytic-deliberative procedures, potentially reshaping not only technology policy but also democracy itself.

Risk is the unavoidable companion of a growing technological complexity. Managing large-scale technological risks has therefore become a central challenge for all societies—perhaps even the key basis for distinguishing contemporary from past societies (Beck 1986, Giddens 1990). It is this challenge that raises the fundamental and daunting question: are contemporary societies capable of generating technological complexity, technological interdependence, and pervasive risks at a more rapid pace than our growth in knowledge of these complexities and risks—and more rapidly than we can manage them safely? Whether a form of cultural lag, an inherent paradox of technological sophistication, an epiphenomenal feature of what we currently call postmodernity, this question will remain a focal point of discourse and decision making in the coming era, both for sociology and for society more broadly.

See also: Accidents, Normal; Disasters, Sociology of; Environmental Risk and Hazards; Health Insurance;

Economic and Risk Aspects; Industrial Sociology; Law and Risk; Occupational Health; Personality and Risk Taking; Risk, Sociology and Politics of; Risk: Theories of Decision and Choice; Workplace Safety and Health

Bibliography

- Beck U 1986 *Risk Society: Towards a New Modernity* (trans. Ritter M). Sage, London
- Clarke L 1999 *Mission Improbable: Using Fantasy Documents to Tame Disaster*. University of Chicago Press, Chicago
- Drabek T E 1986 *Human System Responses to Disaster*. Springer-Verlag, New York
- Erikson K T 1976 *Everything in its Path: The Destruction of Community in the Buffalo Creek Flood*. Simon and Schuster, New York
- Freudenburg W R 1993 Risk and recreancy: Weber, the division of labor, and the rationality of risk perceptions. *Social Forces* 71(4): 909–32
- Freudenburg W R, Jones T R 1991 Attitudes and stress in the presence of technological risk: A test of the Supreme Court hypothesis. *Social Forces* 69(4): 1143–68
- Funtowicz S O, Ravetz J R 1992 Three types of risk assessment and the emergence of post-normal science. In: Krinsky S, Golding D (eds.) *Social Theories of Risk*. Praeger, Westport, CT, pp. 251–73
- Giddens A 1990 *The Consequences of Modernity*. Stanford University Press, Stanford, CA
- Kasperson R E, Renn O, Slovic P, Brown H S, Emel J, Goble R, Kasperson J X, Ratick J 1988 The social amplification of risk: A conceptual framework. *Risk Analysis* 8: 177–87
- Kreps G A 1984 Sociological inquiry and disaster research. *Annual Review of Sociology* 10: 309–30
- LaPorte T R, Consolini P M 1991 Working in practice but not in theory: theoretical challenges of high reliability organizations. *Journal of Public Administration Research and Theory* 1: 19–47
- Mazur A 1980 *The Dynamics of Technical Controversy*. Communications Press, Washington, DC
- Perron C 1984 *Normal Accidents: Living With High-risk Technologies*. Basic Books, New York
- Rosa E A 1998 Metatheoretical foundations for post-normal risk. *Journal of Risk Research* 1: 15–44
- Short Jr J F 1984 The social fabric at risk: Toward the social transformation of risk analysis. *American Sociological Review* 49: 711–25
- Slovic P 1987 Perception of risk. *Science* 236: 280–5
- US Atomic Energy Commission (AEC) 1975 *Reactor Safety Study: An Assessment of Accident Risks in US Commercial Nuclear Power Plants*. US Nuclear Regulatory Commission (WASH-1400), Washington, DC
- Wynne B 1992 Risk and social learning: Reification to engagement. In: Krinsky S, Golding D (eds.) *Social Theories of Risk*. Praeger, Westport, CT, pp. 276–97

E. A. Rosa and W. R. Freudenburg

Risk, Sociology and Politics of

Risk and chance have long been held to be key elements of the human predicament; but risk aware-

ness besets whole cultures as well as individuals. Social scientists typically have been interested in risk, from both the social-psychological stance of particular actors and in terms of cultural responses to collective exposures. At one level, sociologists have discerned a historical trend towards the taming of risk, both in the sense that risk and probability have become better understood, and that certain forms of exposure to risk have been reduced. Nineteenth- and early twentieth-century confidence in progress and control over nature suggested both that risks were diminishing—diseases could be controlled, dangers to food supply countered through scientific farm management, and so on—and that better understanding of risk and chance was possible. But this modernist confidence about the prospects for the control of risk has been moderated and to some extent undermined by subsequent developments. As an example, twentieth-century technologies were associated with new forms of risk, including risks to the whole ecosystem. It seemed that the forces of progress could lead to novel hazards as well as to enhanced security. Furthermore, the legal frameworks and intellectual tools developed for the regulation of risk made it possible to elaborate arguments about risk in ever more sophisticated ways, ironically engendering a more vivid appreciation of risk. Finally, the complexity of social development fostered distinctively social risks, related to crime and the increasing uncertainty of social life. This discussion will begin by focusing on the regulatory treatment of risk before moving on to consider the broader social experience of risk.

1. Risk Assessment: Regulators and Risks

One systematic and extensive literature on risk grew out of concerns to make policy and to legislate for risk; this literature can be categorized as concerned with 'risk assessment.' Given the ubiquity of risks, the question for regulators was customarily thought of as: 'how safe is safe enough?' Governments and official agencies recognize that, regrettably, train crashes occur, that drivers are daily involved in automobile collisions, that leaks will sometimes happen even from well-managed chemical plants, and that farm chemicals may impact the environment, agricultural workers, or even occasional consumers. Given that no complex system can be guaranteed to be perfectly safe, the leading approach was to ask about the price of additional safety. From this consequentialist world view, risk is thought of as the mathematical product of the likelihood of a hazard occurring, and the costs to which that hazard gives rise. The policy maker's art is then to minimize risk within the budget that society is willing to spend on safety.

In practical terms, this task was always much more difficult than might be supposed, and certainly more

intractable than economists routinely implied. In part, difficulties arose because of the inherent problems of trying to harmonize risk data across various fields. It may be possible to compare the costs and risks of various railway signaling systems, but (in the absence of a super-ministry of risk reduction) transport, industrial, medical, and agricultural risks cannot be brought within the same calculus. Worse still, both the likelihood of the problem and its consequences commonly defy exact specification. While there are good data on the medical and related risks of typical motor accidents on US and European freeways, the probability and the consequences of nuclear power station incidents can only be calculated in hypothetical ways. Similarly, the risks and costs of 'mad-cow disease' (bovine spongiform encephalopathy, or BSE), an unprecedented form of infection apparently spread through the food chain, could not be subjected to the risk calculus in the standard way. Further, the only way to obtain quantitative measures of many risks to humans is to extrapolate from indirect observations, such as rodent bioassays, thereby introducing additional sources of distortion into risk assessment. Without good data, all calculations will be 'rough and ready.' Nor is it even clear that there is a single 'currency' into which all sorts of harms can be converted for the purposes of cost-benefit calculations.

Despite these persistent difficulties, one line of the literature about risk has been devoted to exploring what are thought to be the public's shortcomings in relation to risk understandings. Lay people appear to make irrational and statistically unsupported assessments of the relative risks of different hazards, and to demand high levels of risk reduction without being aware of the costs. In the extreme, the public appears unwilling even to accept the cost-benefit approach, but has no systematic alternative with which to replace it. As an added complication, despite frequent public protestations of anxieties about risk, market mechanisms seem to imply that people are not necessarily very risk averse: The automobile market has tended to sell on aesthetic and performance criteria rather than on safety; and consumers continue to eat fatty foods and to avoid exercise despite well publicized links to heart disease.

The apparent discrepancies between public and expert versions of risk have given rise to a series of psychometric studies (Slovic 1992). For example, researchers have shown a tendency by people to undervalue systematically the risks to which they are routinely exposed (because of their lifestyle or occupation), and to overvalue the seriousness of risks that are novel or could be catastrophic. Moreover, there appear to be significant differences between people's assessment of the risks to which they expose themselves and those to which they believe they are subjected by others. These perceptions translate into corresponding pressures in policy. People seem

more likely to demand that automobiles be engineered to lower traffic fatalities than they are to constrain themselves to drive those automobiles more cautiously.

The resulting tension between public and expert interpretations of risk has caused a problem for policy makers. If they base regulations on expert judgments—that is, keyed only to the statistical probability of harm—policies may be unpopular or even subverted, whereas basing policies on the public's apparent preferences threatens to make regulations arbitrary, unscientific, or too costly (Breyer 1993). If the citizenry really is more tolerant of self-imposed risks than of risks visited on them by others, then there is an argument for reflecting this in public policy, whatever the 'actual' exposure to risk. Similarly, if, as Slovic and colleagues report (Slovic 1992, p. 121), the public is more concerned about certain 'dread' risks than other risks which experts hold to be of equal hazard, then maybe the cost-benefit equation has to be widened to take into account people's manifest preferences. Regulators find themselves confronting a tension familiar to liberal democracies: That between people's 'revealed' preferences and the recommendations supported by expert opinion.

2. Risk Expertise: The Reflexivity of Risk

These problems have become amplified in particular institutional contexts in ways that social science research has sought to explicate. Risk assessments have been developed, particularly by official regulatory agencies, so as to apply to statistically representative, or in some other sense, 'average' people (Jasanoff 1990). Calculations of how long it takes to evacuate an aircraft, and thus of the requisite number of emergency exits, the width of aisles and so on, are supported by evacuation trials, but these in turn depend on notions of what is a typical passenger cohort. Similarly, arguments about exposure to pollutants have had to construct some notion of the average person as the unit for measuring the at-risk population (with, in many cases, the need to substitute the average risk animal for the average human). Yet this ideal type hardly exists in reality and may obscure threats to specific subpopulations. Women may be different from men; pregnant women are more clearly different. The young may differ from the elderly, the housebound from the active, and so on. These differences may prove to be far from 'academic' in particular contexts of exposure. For example, people who live near one of the UK's most controversial nuclear power plant sites, Sellafield, and who happen to favor a shellfish-rich diet may be exposed to a greater nuclear hazard than the average person, because of the way in which shellfish filter material from sea water. In this case, a behavioral choice exacerbates other risk factors. In other cases, however, numerous background

factors may act in synergy for the poor or socially disadvantaged, as in the case of American inner-city minority populations, who may be exposed to toxic pollutants from multiple sources (Bullard 1994). In neither the UK nor the US cases would standard risk-assessment methodologies, based on average characteristics, compute adequately the risk to specially vulnerable populations. Indeed, technical risk assessment has tended to obscure the distributive implications of risk production. It is, in this sense, a social technology that reinforces existing structures of power (Winner 1986).

Ironically, the great increase in formal examinations of risks has not so much diminished as contributed to the exacerbation of risk concerns (Beck 1986). Given the complexities of risk assessment and the high stakes involved, it is understandable that risk assessment methodologies have been subject to legal and other formal challenges (see Jasanoff 1990). These proceedings have subjected risk assessments to critical deconstruction, questioning the precise basis for the choice of methodologies. Where these methods have had consequences injurious to some but not all sections of the population, the procedure has been made to appear discriminatory. Environmental justice movements have formed to press the point that risk assessments of individual chemicals are not equivalent to a demonstration of the impact of a cocktail of pollutants. Communities experiencing extensive hazardous exposures—each element of which may be deemed not to be unduly risky on its own—argue that the overall impact of multiple sources is not well gauged by standard methodologies.

Furthermore, as Jasanoff (1990) has demonstrated, particularly in the US, the combination of adversarial cross-examination with the separation of powers and scope for judicial review of executive agencies, has meant that extensive financial and intellectual resources have been directed at deconstructing risk assessments. Given all the complexities outlined above (the hypothetical nature of many risks, the difficulty of identifying an average case, the impossibility of most human experimentation, the need for surrogate measures, and so on), there has been no prospect of finding an incontestable scientific basis for defending particular risk assessments.

Sociological research has focused also on the strategic constructions of certainty and uncertainty associated with risk assessments. Thus, Wynne (1992) has argued that risks and probabilities are made up of many kinds of not-knowing, only a fraction of which (e.g., uncertainties of extrapolation from rodent bioassays) may be acknowledged openly in formal risk assessment processes. In some cases, it is possible to establish a hierarchy of uncertainty, as between 'risks' and 'uncertainties': With risks one knows the odds, but with uncertainties, only the general parameters of the problem. Most practical questions which science-in-public has to face, however, involve an additional

kind of noncertainty. This Wynne terms ignorance. Ignorance refers to aspects of a problem which are consciously or unconsciously bracketed off, and commonly not further investigated. These may be issues which lie outside the disciplinary paradigms of the sciences and are thus in a sense necessarily, rather than perniciously, excluded from day-to-day research. Nonetheless, ignorance in this sense is a different form of not-knowing than mere uncertainty; it is not adequately captured by being treated simply as extreme uncertainty.

In principle at least, more knowledge might assist in handling these kinds of not-knowing. Uncertainties might be turned into risks. New understandings might clarify specific areas of former ignorance, though, of course, there is no prospect of ignorance being overcome in general. But, in addition, Wynne argues that there is a fourth consideration, 'indeterminacy,' resulting from 'real open-endedness in the sense that outcomes depend on how intermediate actors will behave' (Wynne 1992, p. 117). In other words, the safety or reliability of systems with an organizational or human component are deeply dependent on how the systems are operated. By treating most systems as determinate, conventional risk-assessment practices typically adopt tacit, unexamined and untested sociological hypotheses about those social practices that are central to the risk-producing activities. Thus, evaluations of risks from potentially hazardous technologies do not depend only on the behavior of physical and biological systems (difficult to model though they may be) but every bit as much on the behavior of plant managers, operators, and even those exposed.

Several insights into the sociology of expertise follow from these claims. First Wynne (1992) observes that experts, when faced with the need or chance to regulate in an area of noncertainty, are tempted to treat all forms of not-knowing as statistically treatable uncertainty, even though things of which they are ignorant cannot, by definition, be quantified. Second, he argues that the public may be significantly more expert than the scientists in relation to some of the matters covered by indeterminacy. His view, as expressed with Irwin, is that 'science offers a framework which is unavoidably social as well as technical since in public domains scientific knowledge embodies implicit models or assumptions about the social world' (Irwin and Wynne 1996, pp. 2–3). In other words, insofar as expert assessments depend on assumptions about particular social, cultural, or occupational practices of lay groups, it is likely that these publics will be more expert in these matters than technical 'experts' more distant from the relevant experiential insights.

A related body of research ties this manifestation of hazard within complex sociotechnical systems to features of organizational practice and culture. Thus, Perrow's discussion of 'normal accidents' (1984) makes the case convincingly that many hazardous occur-

rences are a routine (if often unpredictable) outcome of the interaction of the manifold technical and organizational components of modern power stations and production plants. Accidents are to be expected even if no particular accident is predictable. More recently, Vaughan has applied associated insights to understanding the decision to launch the ill-fated Challenger mission (Vaughan 1996). Of course, questions still remain about the extent to which organizational arrangements give priority to concerns for production, or for safety (Clarke 1989), and many recent case studies show corporations placing the most value on production and profitability.

In sum, though risk discourse has been promoted by scientists and by scientific attitudes to the valuation of the natural world, it is clear that public experience of risk assessment has encouraged successful challenges to the official orthodoxy, as in the emergence of an environmental justice agenda, and that 'scientific' risk assessment has been undermined dialectically by the advance of legal-rational reflection. In response to such difficulties, official agencies commonly are left with no alternative but to demand 'more and better' science; yet there are few grounds for thinking that further research or rationalization will resolve the problems outlined above. Ironically, social movement campaigners find themselves confronting a similar impasse. In opposing the establishment they commonly produce counter-claims about health or environmental risks, yet find related difficulties in underpinning their risk claims. Such experiences often generate an ambivalent attitude toward scientific expertise (Yearley 1992). These phenomena serve as a good example of what Ulrich Beck (1986) has termed 'reflexive modernization.' The institutions of modernity, notably the traditions of scientific analysis and legal fact-finding, have been turned upon themselves with deconstructive consequences. For Beck, this reflexive modernization is just one facet of the 'risk society' thesis, which portrays late-millennial industrialized societies as being unusually and overwhelmingly concerned with the distribution of risks and other 'bads.' In this view, social scientific interest in risk consists not so much in the study of societal responses to particular hazards; rather it is the key to characterizing the dynamics of society as a whole.

3. *Risks and Cultures*

Beck's thesis in its widest form asserts that the most modern risks (termed by others 'late' or 'high' modern (Giddens 1991)) are reflexive in the sense that they are self-induced. Risks in the early modern period were external to the self-conscious control of social actors. Diseases would spread, bad weather would damage harvests, fires would consume urban areas as though under the influence of external, natural forces. Even if some of these risks were exacerbated by human

interventions, the contemporary perception was that they were uncontrollable. In high-modernity, by contrast, risks such as the threat of disastrous nuclear reactor incidents are plainly the consequences of human activities (Beck 1986). In this view, Victorian and early-twentieth-century confidence about the progressive diminution of risk marks, in an ironic fashion, the transition from apparently external to societally-induced risk. Scientific and technological developments are caught up in this reflexive pattern because modern risks are typically the result of technological ventures (nuclear power, ozone-depleting chemicals in the earth's atmosphere); science and technology are involved in the cause, the diagnosis, and, with luck, eventual rectification of the problem.

Beck's overall analytical claim has been greeted with widespread enthusiasm by many social scientists, and the term 'risk society' has popularly been adopted. However, the details of his analysis have not met with the same welcome. For one thing, it is unclear how 'modern' all present-day risks are. The incidence of 'mad-cow disease'—eventually apparently transmitted to humans—is thought to have arisen from the low-technology business of producing cattle feed from animal protein, specifically in the context of energy-saving, low-temperature process innovations. More significantly, Beck's favored examples—such as the risk of fall-out from the Chernobyl reactor explosion—have a primitively 'democratic' quality. On the face of it, the fall-out may descend upon the poor and wealthy alike. In that sense, the 'risk society' is everyone's problem (Beck 1986). But as the environmental justice movement, particularly in the USA, has made clear, environmental 'bads' are still often distributed quite unequally along ethnic, gender, and class lines (Bullard 1994). The risks of the risk society may not be shared as evenly as Beck implies. Finally, Beck has rather little to say about the reasons for cultural differences in the salience of different kinds of risk, although it is clear, for example, that the framing of, and policy responses to, risks vary widely even between socially and economically comparable regions, such as Europe and North America (Jasanoff 1986).

The relationship of risk and culture has been explored by the anthropologist Mary Douglas and several of her colleagues. Developing Durkheim's celebrated suggestion that religious cosmologies reflect social structures, Douglas proposed that cultural views of the characteristics of nature indicate as much about the reflected character of society as about the underlying features of nature itself. Carefree, individualistic, dynamic cultures tend to view nature as resilient and able to look after itself, while cultures which are precarious or worry about protecting their boundaries tend to view nature as fragile and in need of protection. Douglas subsequently developed this approach for the examination of risk in advanced industrial societies. A society's risk anxieties, in her view, relate as much to

the cultural 'insecurities' of that society as to the actual extent of hazards (Douglas and Wildavsky 1982). Douglas' 'cultural theory' has proven difficult to test (but see Marris et al. 1998) and it raises a host of theoretical problems, such as the unit of cultural analysis and the sources of cultural change. However, her approach does serve to underline the extent to which a culture's awareness of risks may be an imputation on to nature of more general cultural anxieties.

Finally, it is clear that not all modern risks arise from human interventions in nature. Though concerns about medical and environmental risks have increased conspicuously since the mid-1970s, worries about crime, immigration, internet security, financial markets and global social dislocation have also characterized the networked and industrialized world (see also *Science and Social Movements*). Risk anxieties can thus be inscribed not only to nature and to the human body, but also to the self-understanding of modern societies.

See also: HIV Risk Interventions; Law and Risk; Personality and Risk Taking; Probability and Chance: Philosophical Aspects; Risk: Empirical Studies on Decision and Choice; Risk: Sociological Study of; Risk: Theories of Decision and Choice; Sexual Risk Behaviors

Bibliography

- Beck U 1986 *Risikogesellschaft: Auf dem Weg in eine andere Moderne*. Suhrkamp, Frankfurt am Main [Trans. 1992 as *Risk Society: Towards a New Modernity*. Sage, London]
- Breyer S G 1993 *Breaking the Vicious Circle: Toward Effective Risk Regulation*. Harvard University Press, Cambridge, MA
- Bullard R D 1994 *Dumping in Dixie: Race, Class, and Environmental Quality*. Westview Press, Boulder, CO
- Clarke L 1989 *Acceptable Risk? Making Decisions in a Toxic Environment*. University of California Press, Berkeley, CA
- Douglas M, Wildavsky A 1982 *Risk and Culture: An Essay on the Selection of Technological and Environmental Dangers*. University of California Press, Berkeley, CA
- Giddens A 1991 *Modernity and Self-Identity*. Polity Press, Cambridge, UK
- Irwin A, Wynne B 1996 Introduction. In: Irwin A, Wynne B (eds.) *Misunderstanding Science? The Public Reconstruction of Science and Technology*. Cambridge University Press, Cambridge, UK
- Jasanoff S S 1986 *Risk Management and Political Culture*. Russell Sage, New York
- Jasanoff S S 1990 *The Fifth Branch: Science Advisers as Policymakers*. Harvard University Press, Cambridge, MA
- Marris C, Langford I, O'Riordan T 1998 A quantitative test of the cultural theory of risk perception: Comparisons with the psychometric paradigm. *Risk Analysis* **18**: 635–48
- Perrow C 1984 *Normal Accidents: Living with High Risk Technologies*. Basic Books, New York

Slovic P 1992 Perception of risk: Reflections on the psychometric paradigm. In: Krinsky S, Golding D (eds.) *Social Theories of Risk*. Praeger, London

Vaughan D 1996 *The Challenger Launch Decision: Risky Technology, Culture and Deviance at NASA*. University of Chicago Press, Chicago, IL

Winner L 1986 *The Whale and the Reactor: A Search for Limits in an Age of High Technology*. University of Chicago Press, Chicago, IL

Wynne B 1992 Uncertainty and environmental learning. *Global Environmental Change* **2**: 111–27

Yearley S 1992 Green ambivalence about science: Legal-rational authority and the scientific legitimization of a social movement. *British Journal of Sociology* **43**: 511–32

S. Yearley

Risk: Theories of Decision and Choice

Risk is a central concept in everyday life as well as in many theories of the social and behavioral sciences. People have an intuitive notion about what is meant by risk. We are able to make risk judgments and we are able to compare the riskiness of alternatives. We might agree that it is riskier to invest in stocks than in bonds, and it might seem riskier to eat mushrooms which we have collected ourselves than mushrooms bought in a supermarket. Nevertheless, some people choose to buy stocks and others choose to eat mushrooms which they collected, i.e., they are willing to take the risk.

The examples demonstrate that risk can be analyzed as a primitive and that risk can be analyzed in the context of choice. Theories which model risk as a primitive usually start from the assumption that for each individual there exists a meaningful risk ordering of alternatives which can be obtained directly by asking the individual. Theories that consider risk in the context of preference start from choice as a primitive. A definition of risk or riskiness of an alternative is then derived based on models of choice.

Throughout Sect. 1, the leading model of choice, expected utility, is considered. Based on this model, concepts of risk and riskiness are analyzed. Section 2 investigates theories that take risk as a primitive. In Sect. 3, risk value models are presented. They model choice in a way that explicitly considers risk, i.e., they combine the ideas of Sects. 1 and 2.

Throughout the article, only theories for the risk of random variables are presented. Random variables will be denoted X , Y , etc. and they will also be called lotteries or gambles. See *Risk: Empirical Studies on Decision and Choice* for what is known about the public's judgments of the riskiness of physical, environmental, and material risk in ways that go beyond modeling these risks as random variables. An overview on risk-taking behavior can be found in Yates (1992).

1. Risk in the Context of Expected Utility

The most widely used model of preference is von Neumann and Morgenstern's (1947) expected utility model:

$$X \succsim Y \text{ if and only if } E[u(X)] \geq E[u(Y)]$$

where $u(\cdot)$ is the utility function of an individual and E stands for expected value. The binary relation $\succsim$ denotes 'preferred or indifferent to.' As expected utility results in an ordering of risky alternatives, it is called a theory for decision making under risk. However, utility theory offers no direct definition of risk or riskiness.

1.1 Risk Premium

One measure of risk in the context of expected utility could be the risk premium (RP), which is defined by the following equation:

$$E[u(X)] = u(E[X] - \text{RP})$$

The difference $E[X] - \text{RP}$ is also called the certainty equivalent of the lottery X . For reasons of simplicity, in the following only increasing utility functions are considered. If the risk premium is negative (positive), the expected utility of the lottery is larger (smaller) than the utility of the expected value of the lottery, i.e., the decision maker does (not) like the risk of the lottery, i.e., they behave in a risk seeking (risk averse) manner. A decision-maker with a risk premium of zero is labeled risk neutral. All risk premiums for nondegenerate lotteries (lotteries where no single consequence has a probability of one) are positive (negative, zero) if and only if the decision-maker's utility function is concave (convex, linear).

Can the risk premium be used to define a risk ordering on lotteries, i.e., a binary relation $\succsim_r$, where $X \succsim_r Y$ means that lottery X is at least as risky as lottery Y ? Unfortunately, the answer is no. The risk premium depends on the initial wealth of the decision-maker. If W and W' denote two initial wealth levels, it is quite possible that $\text{RP}(W+X) > \text{RP}(W+Y)$ and $\text{RP}(W'+X) < \text{RP}(W'+Y)$. To define risk unambiguously, one must place restrictions on the utility function, or on the random variable.

1.2 Restrictions on the Utility Function

A classic restriction on the utility function is to assume that it has a quadratic form

$$u(x) = x - \beta x^2, \beta > 0, 0 \leq x \leq 1/(2\beta)$$

where x denotes possible outcomes of the gamble X . In this case, the expected utility of a gamble X is given by

$$\begin{aligned} E[u(X)] &= E[X - \beta X^2] \\ &= E[X] - \beta E[X^2] \\ &= (E[X] - \beta E[X]^2) - \beta \text{Var}(X) \end{aligned}$$

Thus, the expected utility is a function of expected value and variance. An increase in risk (measured by variance) implies a decrease in expected utility and vice versa. The coefficient β reflects the trade-off between risk and expected value. The risk ordering of lotteries implied by variance is independent of the decision-maker's wealth level. For further cases where variance is a reasonable measure of risk, see Levy (1992). The general question to what extent expected utility can be interpreted as a risk value model is discussed in Sect. 3.

1.3 Restrictions on Random Variables

A classic restriction on the random variable is to compare only those that have identical expected values. For this case, Rothschild and Stiglitz (1970) were able to provide a definition of the risk of gambles. For $E[X] = E[Y]$, we can say that X is more risky than Y if and only if the following three statements, which are equivalent, are true

- (a) Every risk-averse decision maker prefers Y to X .
- (b) X is equal to gamble Y plus noise, i.e., some random variable with expected value zero.
- (c) X is obtained by a mean preserving spread of Y . A mean preserving spread takes probability weight from central parts of the distribution and puts it into the tails of the distribution while holding the expected value constant.

This definition of riskiness is convincing, but it is only applicable to lotteries with equal expected value, and thus yields a partial ordering of the space of gambles. To overcome this limitation, Jia and Dyer (1996) define a risk ordering on lotteries for concave utility functions u as follows

$$\begin{aligned} X \succsim_r Y \text{ if and only if } & -E[u(X - E[X])] \geq \\ & -E[u(Y - E[Y])] \end{aligned}$$

This definition of riskiness is able to order all lotteries. However, it is not clear to what extent this definition captures what people intuitively think about risk and how an ordering in terms of risk is related to an ordering in terms of preference. We will further investigate these questions in Sect. 3.

1.4 Relative Risk Aversion

All the results in this section are based on expected utility theory. However, it should be pointed out that utility functions not only measure decision-makers' risk attitude, simultaneously they measure the marginal value of the good to be evaluated. An example should illustrate this point.

Suppose one has to evaluate a lottery which gives eight oranges if on a coin flip head comes up and zero oranges if tail comes up. Assume that the certainty equivalent is three oranges, i.e., the risk premium is one orange and thus the decision-maker is considered risk averse. Now, it could well be that this positive risk premium has nothing to do with any 'intuitive' notion of risk aversion. For instance, if the decision-maker likes three oranges half as much as eight oranges, then the average value of the lottery is equal to the value of three oranges, i.e., the risk premium can equally well be explained by some strength of preference considerations. The notion of relative risk aversion, introduced by Dyer and Sarin (1982), helps to disentangle strength of preference from utility. The concept of relative risk aversion defines risk aversion relative to the strength of preference. The above example shows a decision-maker who is relatively risk neutral.

2. Risk as a Primitive

In this section we start out with the assumption that subjects are able to meaningfully order lotteries with respect to their riskiness, i.e., that a risk ordering $X \succcurlyeq_R Y$ exists. In *Risk: Empirical Studies on Decision and Choice* it is demonstrated that individuals are able to give consistent risk judgments. In this section we investigate how these judgments can be modeled. One has to look for a function R defined on the set of gambles with the property

$$X \succcurlyeq_R Y \text{ if and only if } R(X) \geq R(Y)$$

Every such function R will be called a risk measure. Note that we take subjects' risk judgments (risk perception) as a primitive without giving an abstract definition of risk and without relating it to preference (see Brachinger and Weber 1997 for an overview).

2.1 Naive Measures of Risk

The first class of risk measures to be discussed are so-called naive risk measures. Most of them are special cases of the following formula

$$R(X) = E[|X - x_0|^\alpha | X \leq x']$$

where x_0 is a benchmark (reference level of wealth) defining gains and losses, x' defines the range parameter that specifies what deviations are to be included in the risk measure, and α is a measure of relative impact of large and small deviations. Risk measures

such as variance, semivariance, shortfall measures, probability of ruin (loss), etc., are all special cases of this risk measure. Some of these measures are quite popular, e.g., shortfall measures in the insurance and banking industry; nevertheless, they do not capture what individuals perceive as risk (see *Risk: Empirical Studies on Decision and Choice*).

The fact that risk perception is not captured by variance, the measure most popular in economics and finance, can be demonstrated by a simple example. Consider the three two-outcome gambles $X = (10, 0.5; -10, 0.5)$ —this is a gamble with a 50–50 chance of receiving 10 or -10 —, $Y = (20, 0.2; -5, 0.8)$, and $Z = (5, 0.8; -20, 0.2)$. Empirical studies find that subjects consistently judge each of Z and Y to be more risky than X (Coombs and Bowen 1971), but expected value and variance are the same for all three lotteries. Higher moments of the distribution, e.g., skewness, have an influence on risk perception too.

2.2 Axiomatic Measures of Risk: I

For a while it was quite popular to axiomatically derive measures of risk. Similar to what is done in modeling preferences, simple axioms of risk judgments were presented and measures of risk were derived mathematically. Pollatsek and Tversky (1970) were the first to use the axiomatic approach. Besides some technical axioms, they require a risk measure to be additive, i.e., the risk of the convolution of two lotteries should be equal to the sum of the risk of each lottery. This property implies, e.g., that the risk of a lottery played twice is equal to twice the risk of the lottery played once, and the risk ordering of two lotteries is not changed if they are both convoluted with the same lottery. Finally, they require that the risk of a lottery is reduced if a constant is added to all outcomes. From these conditions they derive the following risk measure

$$R(X) = \alpha \text{Var}(X) - (1 - \alpha) E[X], \quad \alpha \in (0, 1]$$

Thus, according to Pollatsek and Tversky (1970), the risk ordering is generated by a linear combination of expected value and variance. This model does not fit risk judgments since, for instance, the example given at the end of Sect. 2.1 demonstrates that subjects' risk perception can not be explained just by expected value and variance.

Fishburn (1982, 1984) explored risk measurement from a purely axiomatic point of view. He considered gambles separated into gains and losses and did not specify functional forms, thus his theory remains rather general and is not considered further here.

2.3 Axiomatic Measures of Risk: II

Luce (1980) was the first to derive risk measures from functional equations characterizing the effect of resca-

ling on perceived risk. His most prominent models consider the α -transformation X_α in which each outcome of the gamble X is multiplied by $\alpha > 0$. He assumes that there exists an increasing function S (with $S(1) = 1$), such that $R(X_\alpha) = R(\alpha X) = R(X) S(\alpha)$ for all X and α . He also assumes that the risk of a gamble X is the expectation of a transformation T of that gamble, i.e.,

$$R(X) = E[T(X)], \text{ with } T \text{ some non-negative function with } T(0) = 0$$

These assumptions imply the following risk measure

$$R(X) = b_1 E[|X|^\delta \mid X < 0] \text{Prob}(X < 0) + b_2 E[X^{\delta^+} \mid X > 0] \text{Prob}(X > 0)$$

where δ is a positive real number, b_i are weights on the components, and $\text{Prob}(X > 0)$ denotes the probability that the random variable has outcomes greater than zero. This risk measure is a linear combination of the conditional expectation of negative and positive outcomes, where all outcomes are raised to the power δ .

Luce's model was extended by Luce and Weber (1986). Influenced by empirical work and based on a system of axioms, they proposed the Conjoint Expected Risk (CER) model that has the following form

$$R(X) = a_1 \text{Prob}(X = 0) + a_2 \text{Prob}(X < 0) + a_3 \text{Prob}(X > 0) + b_1 E[|X|^{\delta^-} \mid X < 0] \text{Prob}(X < 0) + b_2 E[X^{\delta^+} \mid X > 0] \text{Prob}(X > 0)$$

with a_i and b_i being weights on the components, and δ^- and δ^+ being coefficients of power functions. Empirical research has demonstrated that the CER model describes risk judgments reasonably well, admittedly at the cost of a rather large number of parameters in the model.

Sarin (1987) chose another way to extend Luce's measure of risk. Instead of the α -transformation, he used a β -transformation, which just adds a factor β to all outcomes of the lottery. He then assumes that $R(X_\beta) = R(X + \beta) = R(X) S(\beta)$, with S a strictly decreasing function, $S(0) = 1$ and $\beta > 0$. This assumption mirrors the fact that the risk of a gamble appears to decrease when a β -transformation is applied. In conjunction with the expectation principle (see Luce's approach), Sarin derived the following risk measure

$$R(X) = k E[e^{cX}]$$

with $k > 0$, $c < 0$ or $k < 0$ and $c > 0$. This risk measure gives higher weight to a gamble's potential losses than to its potential gains. Both Sarin's model and the CER model are able to capture the effect that subjects strongly dislike the possibility of large losses (potential catastrophes).

All the risk measures presented in Sects. 2.2 and 2.3 have the property that adding a constant (applying

the β -transformation to the gamble) changes the risk, i.e., they are not location free. This property is supported by a lot of empirical work. However, on theoretical grounds, in quite a number of applications, e.g., in capital market theory, risk should be location free. Since nearly all the models proposed in Sect. 2.1 are location free, for certain applications it might be useful to combine the ideas of Sects. 2.2 and 2.3 with the concept of a location free risk measure. Weber (1990), therefore, advocates extending Sarin's measure of risk to the following form, giving the risk measure some normative flavor

$$R(X) = k E[e^{c(X - E[X])}]$$

with $k > 0$, $c < 0$, which can be used in a capital market framework.

Recently, Jia et al. (1999) modified the model of Jia and Dyer (1996) presented in Sect. 1.3 to make it descriptively appealing. In Sect. 1.3, risk was defined as $R^*(X) = -E[u(X - E[X])]$ which is location free. To heal this deficiency, they propose the following measure of risk

$$R(X) = f(E[X]) (R^*(X) - R^*(0))$$

where $f(E[X])$ is a decreasing function of the expected value and $R^*(0) = -u(0)$ a constant. Thus, an increase in expected value leads to a decrease of risk—as in Sarin's model. Their model is rather general as it only specifies that $R^*(X)$ depends on the decision-maker's utility function, but no specific form of the utility function is required.

3. Risk Value Models

Clearly, risk is related to preference (see also Sect. 1). To further investigate this relation, one could propose a general risk value model for evaluating gambles such that a gamble X is preferred to a gamble Y , i.e.,

$$X \succcurlyeq Y \text{ if and only if } f(V(X), R(X)) \geq f(V(Y), R(Y))$$

where V measures the value of a gamble, R measures the risk, and f reflects the trade-off between these two components (see Sarin and Weber 1993 for a more extensive review). Clearly, a greater value is more preferred and a lower risk is more preferred. Therefore, f is increasing in V and decreasing in R .

Risk value models have an intuitive appeal, as people like to talk about the risk and the value of alternatives. There are two ways to use risk value models:

(a) Starting from expected utility, the EU model can be reformulated as a risk value model.

(b) A measure of risk as a primitive and the expected value can be the components of a risk value model.

Both ways still have their disadvantages. If one starts with expected utility (case (a)), in most cases it is not possible to find a straightforward decomposition into risk and value. It was shown in Sect. 1 that in

order to do this one has to considerably restrict the utility function or the probability distribution. In addition, the risk measures derived from expected utility might not be descriptively correct. If one begins with a risk measure as a basis for a risk value model (case (b)), one has to first define a trade-off function f . It is by no means clear that the resulting risk value model will be prescriptively appealing. Future research hopefully will shed more light onto the relation between preference and risk.

See also: Bayesian Theory: History of Applications; Bounded and Costly Rationality; Decision and Choice: Economic Psychology; Decision Research: Behavioral; Risk: Empirical Studies on Decision and Choice

Bibliography

- Brachinger H W, Weber M 1997 Risk as a primitive: A survey of measures of perceived risk. *OR Spektrum* **19**: 235–50
- Coombs C H, Bowen J N 1971 A test of VE-theories of risk and the effect of the central limit theorem. *Acta Psychologica* **35**: 15–28
- Dyer J S, Sarin R K 1982 Relative risk aversion. *Management Science* **28**: 875–86
- Fishburn P C 1982 Foundations of risk measurement. II. Effects of gains on risk. *Journal of Mathematical Psychology* **25**: 226–42
- Fishburn P C 1984 Foundations of risk measurement. I. Risk as a probable loss. *Management Science* **30**: 396–406
- Jia J, Dyer J S 1996 A standard measure of risk and risk-value models. *Management Science* **42**: 1691–1705
- Jia J, Dyer J S, Butler J C 1999 Measures of perceived risk. *Management Science* **45**: 519–32
- Levy H 1992 Stochastic dominance and expected utility: Survey and analysis. *Management Science* **38**: 555–93
- Luce R D 1980 Several possible measures of risk. *Theory and Decision* **12**: 217–28
- Luce R D, Weber E U 1986 An axiomatic theory of conjoint, expected risk. *Journal of Mathematical Psychology* **30**: 188–205
- Pollatsek A, Tversky A 1970 A theory of risk. *Journal of Mathematical Psychology* **7**: 540–53
- Rothschild M, Stiglitz J E 1970 Increasing risk: I. A definition. *Journal of Economic Theory* **2**: 225–43
- Sarin R K 1987 Some extensions of Luce's measures of risk. *Theory and Decision* **22**: 125–41
- Sarin R K, Weber M 1993 Risk-value models. *European Journal of Operational Research* **70**: 135–49
- von Neumann J, Morgenstern O 1947 *Theory of Games and Economic Behavior*. Princeton University Press, Princeton, NJ
- Weber M 1990 *Risikoentscheidungskalküle in der Finanzierungstheorie*. C. E. Poeschel, Stuttgart, Germany
- Yates J F (ed.) 1992 *Risk-taking Behavior*. Wiley, Chichester, UK

M. Weber

Ritual

For Durkheim (1964), ritual is social activity falling within the domain of the sacred. The sacred is opposed to the profane, which does not contain ritual action. Leach (1954), however, rejects Durkheim's hard and fast distinction between sacred and profane, preferring to see these as aspects of social actions lying along a continuum. At one end are those actions that are largely technical and profane, while at the other are those which have no technical functions and are largely symbolic, aesthetic, and sacred. Most actions in between these two poles have aspects which are both sacred and profane. Ritual in Leach's sense is, then, by no means confined to ideas of the sacred. By removing ritual as exclusively to do with the sacred, Leach is able to broaden the definition of ritual. He thinks of it as a message-making kind of activity, and so as an aspect of any kind of behavior in which repetitive or customary acts signal messages. Handshakes or wearing a tie in the office are as much forms of ritual as a priest praying to ancestors at a shrine. So-called magical and secular acts are therefore both communicative and both are examples of ritual action.

Ritual for Leach is, then, a way in which humans communicate with each other different aspects of their social structure. Thus, we may be perfectly aware of social class or status differences and even be able to talk about them discursively. But we still signal or symbolize these differences in a number of short-cut ways, through distinctive accents, differences of dress and life-style, and even bodily movement and gesture. Sometimes we are not so conscious of the differences our ritual behavior communicates, and sometimes the differences exist simply through lack of knowledge of how to behave on, say, a formal occasion. In other words, the sociostructural approach argues that ritual action does not occur in a political and economic vacuum: politicoeconomic interests are always involved to some degree, even if sometimes minimally.

A second approach to the study of ritual emphasizes its logical qualities. Here, ritual communicates fundamental ideas that may be specific to a particular culture, or that may even be universal. It is an approach that owes much to structuralism in anthropology, which is predominantly associated with the work of Lévi-Strauss. It runs very much from the idea of culture as organized through binary oppositions. Thus, Needham (1961) has suggested that, when a male diviner or shaman uses his left hand to throw the cowrie shells which he or his possessory spirit will interpret in order to understand the cause and course of an illness or misfortune, he is reversing a kind of natural order. For men are normally associated with the right hand and women with the left. But since diviners fall outside the realm of normal human capacities, being able to see what other humans cannot, it makes logical sense to reverse the equation

of maleness with the right hand. There is no functional explanation for these kinds of ritual reversal for they are governed by proclivities of the human mind. That said, local political and economic circumstances can sometimes take up these proclivities and harness them to a cause. Thus, prophets leading religious movements may communicate their power and distinctiveness through rituals which reverse normal expectations. Many other kinds of ritual behavior, like some beliefs and myths, achieve their communicative effectiveness by reversing what is otherwise considered natural.

Now, while it is true that ritual may be thought of as communicating messages about status or about the way the human mind works or how the world is, it varies enormously in the degree to which it is elaborated. After all, a handshake can hardly compare with a full-scale ritual investiture of a monarch spread over days and involving thousands of people and lavish expenditure. To call both these kinds of activity 'ritual' clearly leaves out much of what is special to the latter. Similarly, some hunting and gathering peoples such as the African Mbuti forest dwellers do not perform elaborate rituals, while other African groups noted for their masquerades do. Peoples do, then, seem to differ in the relative amounts of ritual which they perform; although, in making this claim, there is a dangerous Eurocentric qualitative assumption that the significance of ritual is measured by the amount of time, materials, and people involved. Who is to say that an apparently slight activity, taking only a moment and few resources, may not in fact 'say' much more than a drawn-out ceremony of great proportions whose influence on people is in fact limited? Nevertheless, it is a useful if provisional starting point of analysis to try and understand how variations in the complexity and lavishness of ritual reflect those of social organization and cultural logic. Arising from this approach is another, which recognizes that as well as communicating messages about the current social structure or about fixed properties of the human mind, rituals also change things or enable changes to take place.

We may call this third, the transformational approach to ritual. For Van Gennep (1965), the rites of passage as marking changes of status in the human life cycle—for example, birth, puberty or adolescence, marriage, first child, first grandchild, and death—could be broken down into three phases: separation, liminality, and incorporation. In other words, the individual undergoing the ritual must first be separated from the community of which they are a member, perhaps by removal of the head hair or some other bodily alteration. They then enter the liminal phase in which they are regarded by members of the community as outside and beyond normal human society. This is the phase in which there is transition from one status to another, for example from child to adult, bachelor to husband, etc. Van Gennep saw this as a period of

death, i.e., of the child, and of rebirth, i.e., of the adult. It is a theme elaborated in many later studies that showed the ritualization of death as the condition of new life or regeneration. The third and final phase is when the individual is reincorporated in the community, now no longer a child, but an adult with the rights and duties expected of this new status.

Van Gennep saw these features of rites of passage as also typifying other kinds of transition in addition to those of the life cycle, such as the seasonal or other movements of groups, fertility rites, and entering and leaving dwellings, special days or months, and cult societies. Victor Turner (1967) made the notion of liminality central to his work on ritual, extending to the idea of antistructure, by emphasizing the rejection of society on the part of people undergoing transition or in a marginal state, and also *communitas*, by which he meant the way in which people on ritual occasions temporarily transcended their normal, everyday social conflicts and cleavages and emphasized their common humanity. In addition to the life crisis rites of Van Gennep, Turner further distinguishes rites of affliction, which are held to reverse the effects of, for example, women's infertility, crop failure, sickness, and other misfortunes adversely affecting the lives of people living together in a single village or other dwelling. They also enable people to move from one state to another, although not as part of the human life cycle, but more as part of the unpredictability of the human and social condition in which conflicts and contradictions are not so much represented in ritual as presented, addressed, and then redressed, an approach which thus builds on and advances that of Durkheim.

A subset of this approach might conveniently be called Marxist in the sense that social inequality and exploitation of a working by a ruling class is presupposed. A Marxist approach would, for instance, claim that the official, media, and public ritual attention devoted to the British royal family evokes ordinary people's unquestioning compliance and loyalty, to the extent that they cannot see that the institution of royalty in fact underpins the system of landed privilege that defines an aristocracy and provides an imitative model for other ruling and bourgeois classes. The 1990s in the UK saw an extraordinary rise in the degree to which the media and many of these same ordinary citizens began to question the worth-whileness of royalty. Was this an example of royal rituals losing their capacity to retain subjects' loyalty, a kind of demystification of ritual? Or is this a case of the different interpretations of ritual enabling social change to occur?

Royal rituals have been a common focus of anthropological analysis, in which their transformative power has been highlighted. Gluckman (1963) was dissatisfied with studies of royal and other rituals which saw them as simply affirming the sovereign's authority and the solidarity of the society under his rule. Using the example of the Swazi *Incwala* royal

ritual in southern Africa, Gluckman claimed instead that its socially cohesive function was only possible after it had revealed the society's internal conflicts: such conflicts had to be addressed before they could be accepted as part of the normal social process. Beidelman (1966) criticizes Gluckman's over-reliance on social conflict as needing to be expunged through royal rituals, and argues for a view of them as marking out the divine or mystical properties of kingship, so providing the king with legitimate recognition to reign over his subjects.

This example moves us on to a fourth approach to ritual, which is to see them not necessarily as achieving a particular social or logical function, but as expressive in and of themselves. Beattie (1966) attacked attempts to explain rituals as practical and intellectual attempts to solve problems. Rituals draw instead on people's human predilection to engage in creative thinking, sometimes through culturally enjoined associations of ideas, and were thus closer to art and drama than to technology or science. Rituals in this sense do not 'do' things so much as express human dilemmas, symbolically and often under some kind of duress. They cannot be reduced to questions of whether they are rational or irrational, or true or false propositions, any more than aesthetics can be explained in these terms.

In many respects this explanation of ritual as expressive rather than functional, reproduces Leach's distinction between technical and aesthetic acts as occupying two extreme poles of a continuum between which they are both aspects of any activity. Nevertheless, expressed in the way that Beattie did, it freed anthropologists from having to seek practical explanations for a particular form of ritual behavior. If a ritual appeared to perform a particular function in society, this could better be regarded as an unintended consequence of an activity whose participants did not have such consequences in mind. Thus, it may well be that the British royal rituals for a long time had the effect of persuading the public to accept the stratified status quo as rightfully headed and legitimized by the monarch and the monarch's family. It may well be, moreover, that politicians and members of the ruling elite recognized, seized on, and developed this effect, so turning consequence into purpose. The expressive here becomes instrumentally effective. However, not all that which we understand as ritual behavior can be interpreted in this way. People throughout time do not normally calculate in advance what might pass as effective ritual, nor indeed separate ritual as a social category from nonritual; nor may we assume that what we define as rituals always have significant social and intellectual consequences.

Approaches to the study of rituals are not exhausted by the above, and indeed overlap in part. They represent, as much as anything, the discipline's decision to divide up its findings in this way. An alternative method is to ask how anthropologists define ritual on the basis of their own ethnographic data. A recent

example is a study of the Indian Jain rite of worship by Humphrey and Laidlaw (1994). Their intriguing starting point is the fact that in some societies people may distrust ritual practices as being empty or false, and yet continue with them. Similarly, noting that for many people, religion and ritual are seen as going together, they also refer to the way in which practitioners of some religions turn on their rituals as unworthy and even frightening. The European Reformation attempted to strip off from Christian religious belief what were regarded as some unacceptable ritual practices, as Wahabi reformers in Islam are doing today. By contrast, some argue that ritual is good for religion and humanity. Mary Douglas (1973) drew some of her own ideas from trying to understand the rejection of ritual in the Catholic Church by many of its senior adherents who also, for good measure, dispensed with the Latin language as if transparent clarity was all that mattered, over and above the expressive power of faith.

The disjunction between religion and ritual led Humphrey and Laidlaw (1994) to suggest another disjunction, which is that between the intention of actors in ritual and the ritual actions themselves. Sometimes intention and act fit, and sometimes they diverge irretrievably, but sometimes they are aligned. Such lack of fit is not to be understood as the result of mistakes in intention and performance, but as evidence of the way in which ordinary actions may become endowed with the quality of ritual. For these authors, ritual is a quality that actions take on: they become ritualized. While everyday actions are intrinsically intentional, ritual acts enter a kind of phenomenological space freed of such everyday intentions, and, paradoxically, are 'intended' to do so. There is here a 'ritual commitment,' a notion close to that expressed in a classical study of ritual by Gilbert Lewis (1980). Humphrey and Laidlaw (1994), moreover, oppose the idea that ritual is necessarily communicative and expressive of messages and beliefs. One is here reminded of Sperber's (1975) view that symbols do not intrinsically mean anything. Rather it is people who interpret a symbol as such, and who thereby give it meaning. The concept of communication here has to be established. It may mean one person making an assertion and expecting a response from the hearer. Or it may simply mean making a gesture or uttering a conventional phrase ('It's hot today'), for which no response is expected and which is therefore really a form of what has been called phatic communion, a device to secure solidarity with another person. Rituals may sometimes communicate but, as Rappaport (1979) points out, so do many other kinds of phenomenon, and by privileging its expressive functions we miss what is otherwise distinctive of ritual.

The view that ritualization is a process by which actions become endowed with a ritual quality, shares something with Gerholm's (1988) notion that ritual is, so to speak, partly made up as it proceeds. That is to

say, there are at least some kinds of ritual for which precise textual rules do not exist or which have been forgotten, with the consequence that participants have little more than a sense of ritual occasion or commitment and endeavor to convey that sense in actions and words, inventing and borrowing as much as they can. This ushers in a query as to how much ritual depends on precise wording as well as actions. A Christian prayer depends heavily on personally composed as well as prescribed verbal forms, and is firmly regarded as ritual. But rituals are also organized through spatial orientation and bodily positioning. The many different kinds of Islamic prayer are also personally or formulaically verbal, but crucially require orientation to Mecca, while a particular class of prayer or devotion called *salat* is made up of a complex sequence of prostrations as well as recitations (Bowen 1989). In the event of the mosque imam taking on the recitations on behalf of the congregation, the prescribed bodily actions become the individual worshipper's ritual responsibility and must be repeated if they are performed incorrectly. Parkin (1992) has argued that it is not always the case that the verbal is important in ritual, while set spatial directions and orientations, and bodily gestures, movement, and positions are what mark off and so fundamentally characterize ritual.

A challenge is to avoid essentialist definitions of ritual which seek to set up criteria which must be met before this label can be applied to an activity. Phenomenologically, ritual is to do with the paradox of being both the agent of an action, but also being taken along by this same action. It presupposes being able to move within an action from citing its purpose to having no need, wish, or clear idea of what it is about, and yet valuing it. Nevertheless, ritual must have some kind of structure, the most elemental of which is that individuals alone or with others take up positions with regard to each other or with regard to objects, experiencing such positioning as unlike any other.

See also: Control: Social; Durkheim, Emile (1858–1917); Magic, Anthropology of; Myth in Religion; Religion, Sociology of; Ritual and Symbolism, Archaeology of; Taboo; Totemism

Bibliography

- Beattie J 1966 Ritual and social change. *Man: The Journal of the Royal Anthropological Institute* 1: 60–74
 Beidelman T O 1966 Swazi royal ritual. *Africa* 36: 373–405
 Bowen J R 1989 *Salat* in Indonesia: The social meanings on an Islamic ritual. *Man: The Journal of the Royal Anthropological Institute* 24: 600–19
 Douglas M 1973 *Natural Symbols: Explorations in Cosmology*. Barrie and Jenkins, London

- Durkheim E 1964 *The Elementary Forms of the Religious Life*. Allen and Unwin, London
 Gerholm T 1988 On ritual: A postmodernist view. *Ethnos* 53: 190–203
 Gluckman M 1963 Rituals of rebellion in south east Africa. In: Gluckman M (ed.) *Order and Rebellion in Tribal Africa*. Cohen and West, London
 Humphrey C, Laidlaw J 1994 *The Archetypal Actions of Ritual*. Oxford University Press, Oxford, UK
 Leach E 1954 *Political Systems of Highland Burma: A study of Kachin Social Structure*. G. Bell, London
 Lewis G 1980 *Day on Shining Red: An Essay in Understanding Ritual*. Cambridge University Press, Cambridge, UK
 Needham R (ed.) 1961 The left hand of the Mugwe: An analytic note on the structure of Meru symbolism. *Africa* 31: 28–33
 Parkin D 1992 Ritual as spatial direction and bodily division. In: De Coppet D. (ed.) *Understanding Rituals*. Routledge, London
 Rappaport R A 1979 *Ecology, Meaning, and Religion*. North Atlantic Books, Berkeley, CA
 Sperber D 1975 *Rethinking Symbolism*. Cambridge University Press, Cambridge, UK
 Turner V W 1967 *The Forest of Symbols: Aspects of Ndembu Ritual*. Cornell University Press, Ithaca, NY
 Van Gennep A 1965 *The Rites of Passage*. Routledge and Kegan Paul, London

D. Parkin

Ritual and Symbolism, Archaeology of

Ritual (repetitive sequences of actions related to beliefs) and symbolism (the representation of one thing, often more abstract, by reference to another, often more concrete) have been subjects of archaeological study since the origin of the discipline. Individual artifacts, locations within archaeological sites, and even entire sites or landscapes have been used as archaeological evidence of ritual and symbolism. Archaeologists have varied in their assessment of how difficult it might be to study these subjects, with some expressing doubt that accurate information about such nonmaterial topics could be produced, and others noting that symbols are polysemous, requiring multiple interpretations. Contextual analysis and an emphasis on the active role of social agents characterize the contemporary archaeology of ritual and symbolism.

1. From Ritual Objects to Symbolic Archaeology

Archaeologists have always identified some deposits as evidence of ritual, and ascribed symbolic meanings to some artifacts. Where texts were available, as in Egyptian archaeology, they were employed to validate early interpretations. Beginning in the 1950s explicit concern was addressed to the methodological status of

archaeological inferences concerning belief and meaning. Interpretation of ideational systems, including symbolism and ritual, was viewed as more problematic than other topics, such as economics and subsistence. As early as the 1960s, some archaeologists argued that ritual and symbolism were no more difficult of access than more apparently materialist topics, questioning the analytical separation of material and ideational aspects of society. Since the 1980s, archaeologists influenced by practice theory have attempted to demonstrate that the archaeological analysis of symbolism and ritual benefits from attention to human agency. Despite changing theoretical perspectives, throughout the history of professional archaeology similar empirical materials have served as the basis for the archaeology of ritual and symbolism.

1.1 Early Archaeological Research on Ritual and Symbolism

Early fieldwork recovered material interpreted as evidence of ancient ritual and belief. Contemporary texts were sometimes available that sketched out beliefs and, as in the case of Egypt, even provided details of ritual actions.

Such early research established assumptions underlying much subsequent archaeological investigation of symbolism and ritual. First, the symbolism of material remains ideally was to be explained by reference to explicit texts, assigning material culture to the role of a container of a formalized and generally accepted singular meaning. Second, specific kinds of archaeological deposits came to occupy a central place in discussions of ritual. These included formal structures identified as the locations of rituals (temples); human burials, viewed universally as occasions for mortuary ceremonies; and 'offerings': deposits of selected objects, especially those encountered in ritual structures and burials.

1.2 The 'Ladder of Inference' and Processual Archaeology

In 1954, Christopher Hawkes proposed that in the absence of texts, archaeologists were constrained by a 'ladder of inference' which governed the probable reliability of their explanations (Hawkes 1954). According to the ladder of inference, some topics were more closely tied to material manifestations, which were more tightly circumscribed by pragmatic considerations. As a result, archaeologists were likely to recover a greater proportion of the evidence necessary for understanding these topics, and their explanations were likely to suffer less from crucial missing data or from variability in the nature of the evidence. Areas

such as religion and belief were less tightly tied to materiality, and thus archaeologists were likely to lack crucial pieces of evidence to support explanation, or they might recover material that could be explained in multiple ways and have no basis to choose between explanations. In contrast, in initial programmatic statements for the processual archaeology of the early 1960s, Lewis Binford explicitly argued that all aspects of culture were open to scientific archaeological analysis if questions were properly framed. Neither symbolism nor ritual was excluded from archaeological investigation.

Important contributions to the archaeology of ritual and symbolism were developed in processual archaeology. Drawing on information theory, symbols could be treated as messages intended to communicate clear meanings (Wobst 1977). Information theory provided a means to support some of the assumptions necessary for archaeologists to treat ancient symbols as if they had a single consistent meaning. If symbols were media to convey messages, then there would have been strong social pressures, and even adaptive advantages, to having shared interpretations of their meaning. The application of communication models to the archaeological study of symbols facilitated the development of an archaeological theory of style, the definition and interpretation of sociopolitical 'status badges,' and the interpretation of differences in material culture as markers of ethnic boundaries and cultural identities.

Robert Drennan (1983), employing processual systems theory, developed an explicit set of expectations for ritual subsystems, which could be linked to specific characteristics of archaeological remains to create predictions for the material correlates of ritual. Following social anthropologist Roy Rappaport, Drennan operationally defined ritual as specific sequences of repeated action directed at intervening in relations between humans and supernatural forces or beings. Among the key features he and others used to attempt to objectively identify remains stemming from ritual were associations between otherwise rare items, and the recovery of such rare items from public spaces. The systems approach assumed a separation between an everyday sphere of activities and a separate sphere, the domain of ritual, detectable by distinctive location and unusual artifacts.

Processual approaches required external bodies of information for the formulation of models of ritual and symbolic systems. Specific ethnographic analogy, the 'direct historic approach,' drew on the study of living societies believed to be descended from the archaeological population. General analogies were made from comparative ethnographic samples such as those contained in the Human Relations Area Files. (These were abstracts of ethnographic samples selected to provide a cross-section of known variation in human activity. They provided a beginning point for statistical generalizations about human behavior, including ritual.) The power of explanation rested on

the strength of arguments linking the analogies to the test case, requiring extremely broad uniformitarian generalization. The level of generalization required made explanations of ritual and symbolism employed extremely vague, and tended to substitute the operations of superorganic entities for individual actions.

1.3 Postprocessual Directions in the Archaeology of Ritual and Symbolism

In a series of edited volumes, Ian Hodder (1982a, 1982b, 1989) and colleagues initiated a range of challenges to the processual analysis of ritual and symbolism. Critiques of the information theory models embedded in processual analyses were central to these analyses. Demonstrations of diversity *within* cultural traditions, based on historical and ethnographic documents, highlighted problems with uniformitarian and normative assumptions shared by processual and preprocessual archaeologies.

While there is variation among scholars participating in the postprocessual critique, they share a fundamental view of the nature of symbolic meaning. Rather than seeing archaeologically recovered objects as vehicles carrying pre-existing meanings, they argue that symbolic meaning was negotiated through the use of cultural forms, including the permanent ones recovered by archaeologists. This view of symbolism requires attention to contextual relations in order to understand how symbolic meaning was negotiated. Even when archaeologists operating under this model express uncertainty about their ability to understand the specific meanings symbols had for past actors, they routinely make arguments about the deployment of symbolism as part of strategies of actors and factions within past societies.

The meaning of symbols and their social effects are seen as mutually constituted. Contrasts within and between contexts are required to argue for a relationship between symbolism and any specific form of meaning under these assumptions. The constitution of symbolic meaning is seen as a constant feature of social life. Rituals provided marked settings in which the meanings of symbols were negotiated, but they were not the only or even the most significant settings in which symbolic meaning developed.

2. Archaeology of Symbolism

Relations of substitution are fundamental to the definition of symbols. Archaeologists commonly use specific objects as signs of categories, for example, the identification of Clovis stone points as markers of a Palaeolithic population in the Americas (see *Ethnic Identity/Ethnicity and Archaeology*). This basic archaeological operation rests on the assumption that the

makers and users of these points would have recognized a commonality among themselves and a dissimilarity from others, albeit on a largely non-discursive level (see Gardin and Peebles 1992). Ethnoarchaeological research has demonstrated that the degree to which the use of common material forms conveys either a conscious or unconscious identification is highly variable (Hodder 1982a).

Some symbolic relations of substitution by past populations are treated as more conscious (see *Pre-historic Art*). The embedding of symbolism in concrete things has been a focus of extensive empirical and theoretical debate (e.g., Robb 1999). For example, some artifacts have been interpreted as 'status badges' created as deliberate signs of rank. Some archaeologists seek to understand the specific properties of materials which lend themselves to being used in symbolic transactions, or to reinforce culturally accepted meanings. Others are concerned with the way that symbolic meanings are interpreted. The latter approach includes a consideration of the experiential contexts within which objects with symbolic significance were used, including rituals.

3. Archaeology of Ritual

The archaeology of ritual has moved away from an early history of identification of specific material objects, settings, and representations as evidence of the presence of a separate ritual or symbolic sphere or system, to the use of contextual associations to identify repeated sequences of action like those that are recognized as ritual in living societies. Certain classes of artifacts have been described as 'ritual' objects: examples include ancient Chinese bronze vessels, North American copper and shell objects, and ceramic figurines in many areas of the world. The identification of objects as ritual objects has been criticized as a facile means to defer our understanding of unusual objects. Recently, Ruth Whitehouse (in Wilkins 1996) has argued that 'ritual objects' can be a real category of material with unique potential for informing about ancient belief and action.

What distinguishes ritual contexts in recent work is not simply their separation from presumed household locations. Historical archaeologists specializing in the study of the African diaspora have recovered assemblages of artifacts arguably used in healing rituals on the household level. In place of defining specific objects and places as inherently associated with ritual, contemporary approaches emphasize the repetition of activity that is central to the definition of ritual. Archaeologists working on the British Neolithic proposed the concept of 'structured deposition' as a way to recognize specific archaeological deposits that were likely to be the result of ritual action (Richards and Thomas 1984).

4. The Material Base for an Archaeology of Ritual and Symbolism

Remarkably, despite theoretical differences, the specific material bases for an archaeology of ritual and symbolism have remained fairly stable and have been constantly under investigation. Offerings or caches, burials, particular spatial locales, and visual images have consistently provided the subject matter for archaeological analyses.

4.1 Caches and Offerings

Caches, hoards, or 'offering deposits' figure as primary evidence of ritual in several regional traditions in archaeology, particularly North America, Central America, the eastern Mediterranean, and Neolithic Europe (e.g. Bradley 1991). Interpretations of caches vary from purely materialistic to ideational. Some analysts stress the economic effect of the withdrawal from circulation of a large quantity of specialized goods at one time. Others propose structural-symbolic interpretations in which materials employed worked to create specific associations between ideas about the supernatural world and the locale of the offering, consecrated by the deposit. Explanations of both kinds have been offered for the same deposits, and indeed, they are not mutually exclusive, since a ritual action can have both pragmatic and symbolic effects.

4.2 Burials

Mortuary analysis, the study of sites and artifacts related to human burial, has occupied a central place in the archaeology of ritual and symbolism (e.g., Chapman et al. 1981). Studies conducted under the normative assumptions of culture history described burial customs, including the laying out of the body according to symbolic beliefs (such as the eastward orientation of the head), and the inclusion of artifacts as 'grave goods.' Among mortuary activities, the removal of skeletal elements and the practice of secondary burial have been seen as unambiguous evidence of ritual practice. The deliberate burial of animals has generally also been regarded as evidence of ritual activity.

Processual archaeology directed attention to the generation of covering laws or assumptions that could link the forms of burials, and particularly, diversity among them, to social differences. Burials were seen as reflecting the social identity of the deceased. They operated as mechanisms to assert claims for territory. Postprocessual archaeologists examine burial assemblages as the material media of ancient rituals, and demonstrate that mortuary remains can mask social relations as easily as they can reflect them (e.g., contributions to Hodder 1982b).

4.3 Temples and 'Sacred Landscapes'

The identification of temple sites through spatial segregation, specific construction, or a combination, characterizes much archaeology addressed at understanding ritual and symbolism. While buildings denominated temples may sometimes be unambiguously identified, usually through written documents, such as ancient Egyptian and Maya inscriptions, in other cases the identification of 'temples' may be challenged. Other architectural settings that are assumed not to have seen use in everyday life, such as the ball courts of Central America, the Caribbean, and the US Southwest, are also viewed as sites of ritual.

Spatial locales that can be viewed as nonordinary by nature are sometimes treated as effectively unproblematic locales for an archaeology of ritual. Caves and natural springs are prominent foci for some definitions of 'sacred geographies.' In contrast, the investigation of entire landscapes, such as those of megalithic Europe, emphasizes the social marking of space through ritual action (see *Settlement and Landscape Archaeology*).

4.4 Visual Images

Artistic representations of artifacts and buildings recovered archaeologically have been a staple in reconstructing ritual. They have been approached in two complementary ways. On one hand, researchers have argued that scenes represented mythologies or beliefs about the supernatural. Some scenes have been identified as models for ritual enactment. For example, the elaborately-dressed individuals buried at the coastal Peruvian Sipan site have been identified with the clearly nonhuman figures depicted in supposed mythological scenes on pots of the Moche kingdom. The implication that such acts as ritual decapitation and mountaintop sacrifice actually took place within ancient Peruvian society has been supported by the recovery of caches of skulls and of frozen bodies on the high Andes.

In other cases, representations have been interpreted as scenes recording ritual practices. For example, Classic Maya *stelae* have been identified as showing members of the ruling class engaged in ritual blood-letting and the sacrificial burning of paper and resin. The availability of written texts in societies like those of the ancient Maya and Mesopotamia has led to ambitious attempts to reconstruct specific ritual practices from an emic perspective.

4.5 Animal Remains

Analysis of faunal assemblages is a relatively recent addition to the repertoire of methods used to reconstruct ancient ritual practices (Anderson and Boyle

1996). Ritual studies begin with the assumption that the use of any materials in unusual proportions within a site may signal the location of activities different from those oriented toward the practical demands of day-to-day life. With faunal remains, an additional criterion can be added: the identification of parts of animals that cannot have been used for subsistence.

5. Future Directions

Contextual analysis has become a fundamental part of archaeological investigation of ritual and symbolism. Through contextual analyses, specific suites of artifacts are identified that can be interpreted as evidence of ritual action. Continued emphasis on combining data to produce multifaceted arguments for ritual, and on placing active social agents at the center of reconstructions of meaning, can be expected.

For example, archaeologists in many regions have converged on the concept of feasting as a means to interpret particular assemblages of artifacts, animal bone, and plant remains. While feasting itself need not be seen as a ritual, it is often described in those terms, and is always interpreted as of symbolic significance. Studies of ritual feasting employ remarkably similar language and concepts. Drinking, the ingestion of drugs, and dancing have all been identified as part of ritual feasts. This is leading to the identification of specific categories of material remains that provide evidence of presumed ritual activities. Attention is also being directed to the experiential effects of ancient ritual practices on the development of forms of subjectivity.

Archaeologists are otherwise less inclined to propose specific meanings for symbols, except in literate traditions where texts continue to serve to validate interpretations. Continuing debate will focus on how to create interpretations of symbols that incorporate multivocality and capture the recursion between practice and meaning.

See also: Cognitive Archaeology; Death, Anthropology of; Folk Religion; Household Archaeology; Magic, Anthropology of; Ritual; Sacrifice; Symbolism (Religious) and Icon

Bibliography

- Anderson S, Boyle K (eds.) 1996 *Ritual Treatment of Human and Animal Remains*. Oxbow Books, Oxford, UK
- Bradley R J 1991 *The Passage of Arms: An Archaeological Analysis of Prehistoric Hoards and Votive Deposits*. Cambridge University Press, Cambridge, UK
- Chapman R, Kinnes I, Ransborg K (eds.) 1981 *The Archaeology of Death*. Cambridge University Press, New York
- Drennan R D 1983 Ritual and ceremonial development at the early village level. In: Flannery K V, Marcus J (eds.) *The Cloud People*. Academic Press, New York
- Gardin J-P, Peebles C S (eds.) 1992 *Representations in Archaeology*. Indiana University Press, Bloomington, IN
- Garwood P (ed.) 1991 *Sacred and Profane: Proceedings of a Conference on Archaeology, Ritual and Religion*. Oxford Committee for Archaeology, Institute of Archaeology, Oxford, UK
- Hawkes C 1954 Archaeological theory and method: Some suggestions from the Old World. *American Anthropologist* **56**: 155–168
- Hodder I (ed.) 1982a *Symbols in Action: Ethnoarchaeological Studies of Material Culture*. Cambridge University Press, Cambridge, UK
- Hodder I (ed.) 1982b *Symbolic and Structural Archaeology*. Cambridge University Press, Cambridge, UK
- Hodder I (ed.) 1989 *The Meanings of Things: Material Culture and Symbolic Expression*. Unwin-Hyman, London
- Richards C, Thomas J 1984 Ritual activity and structured deposition in Later Neolithic Wessex. In: Bradley R, Gardiner J (eds.) *Neolithic Studies: A Review of Some Current Research*. B. A. R. British Series, Oxford, UK
- Robb J E 1998 The archaeology of symbols. *Annual Review of Anthropology* **27**: 329–346
- Robb J E (ed.) 1999 *Material Symbols*. Center for Archaeological Investigations. Southern Illinois University, Carbondale, IL
- Wilkins J B (ed.) 1996 *Approaches to the Study of Ritual: Italy and the Ancient Mediterranean*. Accordia Research Centre, London
- Wobst M H 1977 Stylistic behavior and information exchange. In: Cleland C L (ed.) *For the Director: Research Essays in Honor of James B. Griffin*. University of Michigan Museum of Anthropology, Papers in Anthropology, No. 61, Ann Arbor, MI

R. A. Joyce

Robinson, Joan (1903–83)

Joan Robinson was born on October 31, 1903, in Camberley, Surrey, into an upper middle-class English family characterized by vigorous dissent and independence of mind. Her great-grandfather was F. D. Maurice, the Christian Socialist who lost his Chair of Theology at King's College, London, on a point of principle; her father was Major General Sir Frederick Maurice, the victim of the infamous Maurice debate in 1918 and subsequently Principal of what is now Queen Mary College in the University of London. Her mother was Helen Margaret Marsh, the daughter of Frederick Howard Marsh, Professor of Surgery and Master of Downing at Cambridge. Joan Robinson was born into a milieu of controversy. She was educated at St. Paul's Girls' School in London, where she read history, and from 1922 to 1925 at Girton College, Cambridge, where she was Gilchrist Scholar. There she read for the Economies Tripos, searching for rational argument and hoping to understand poverty and its cures. She graduated in 1925 with second-class honours ('a great disappointment'—her words) although it was not until 1948, when women

were given full membership of the university, that she was admitted to the degree. In 1926, she married Austin Robinson, the Cambridge economist, and the couple sailed for India for what was to become the first of many visits. They stayed for over two years and returned to Cambridge in 1929, making it their intellectual and domestic base for the rest of their lives. They had two daughters and five grandchildren. Joan joined the Cambridge Faculty of Economics and Politics as a Faculty Assistant Lecturer in 1934, becoming part of an outstanding group of scholars in Maynard Keynes, A. C. Pigou, Dennis Robertson, Austin Robinson, Maurice Dobb, Richard Kahn, Gerald Shove and Piero Sraffa. She became a University Lecturer in 1937, Reader in 1949, and Professor in 1965. She was elected to a Professorial Fellowship at Newnham and made Honorary Fellow of Girton in 1965, and Honorary Fellow of Newnham in 1971; she was the first woman to be made an Honorary Fellow of King's (in 1979). From 1958 to 1971 she was an FBA. She 'retired' from her Chair in 1971, remaining as active as ever in Cambridge and abroad, until her death in Cambridge on August 3, 1983.

Early in her career Joan Robinson set out her views on methodology for economics, (Robinson 1932a, Harcourt 1990a). Her procedure was to identify and make explicit her assumptions and to develop a 'box of tools' for understanding and exploring the analytical object, to construct a logically sound analysis of the processes concerned: these successive stages she saw as independent of each other. Extending these views, she suggested (1937b, pp. 99–102) that controversy could result from lack of understanding, lack of knowledge, errors of logic, from working from different assumptions but, most commonly in economics, from differences in (political) objectives: the goals of a theory could direct its analysis as well as its policy conclusions. Throughout her life she argued that economics is not neutral, it is always ideologically driven. Nevertheless she believed that it was possible to develop pure analysis independently of dispute.

Her early contributions to economic theory were mainly in the realm of value theory (microeconomics, as it is called now) and *The Economics of Imperfect Competition* (1933), her first book, generously acknowledged her debts to Marshall (via Pigou), Kahn, Austin Robinson, Shove, and Sraffa. It was a controversial and much celebrated success, revealing her great clarity of thought and exceptional analytical powers. It was her interpretation, within a Marshallian context, of several contemporary theoretical issues. In particular, it was a response to the heretical claims about decreasing costs and increasing returns being made in the *Economic Journal*, especially Sraffa (1926), Symposium (see Robertson et al. 1930), and including Robinson (1932b). Robinson replaced an analytical basis in Marshall's free competition with one in imperfect competition, where free competition became a special case. She also moved the focus from Mar-

shall's long period to a short period. Within that context the book also presented new concerns. On the basis of the downward sloping demand curve for each firm, Robinson introduced the marginal revenue curve as an integral analytical construct. She showed how the response of relative factor shares to a change in demand depended on whether you used her partial analysis of the elasticity of substitution or Hicks', aggregate. She reconceived Marshall's 'representative' firm as an 'equilibrium' firm where each such firm faced a less than infinitely elastic demand curve for its product in a monopolistic market. What then became 'the industry' in such a market? And was the simple profit-maximising rationale too reductionist to depict the firm? How was a monopolist to be identified theoretically? Or a duopolist? How indeed was imperfect competition definable?

The Economics of Imperfect Competition provoked; each stage of the theoretical process, even her geometrical method, incited reaction. More controversial than her analysis were its implications for policy. She represented perfect competition theory as demonstrating that the economy left alone would generate the optimal outcomes: it suggested a political philosophy of *laissez faire*. Imperfect competition showed that left alone the economy could become stuck at a position of unemployed capacity and labour and indeed, at that time, this was the case. The debates on the elasticity of substitution further illustrated what the emerging critique of perfect competition theory could mean for the status quo. The issues raised in this book were to fill the journals for decades to come. She herself came to reject its analyses (postwar) as a 'shameless fudge.'

Robinson's interests were not confined to microeconomics. During this time she was a member of a group of economists who met to discuss John Maynard Keynes's *Treatise on Money* (1930) and his emerging new theory. His ideas appealed to her, they represented a break from the orthodox and they were focused on the pressing problem of the persistence of high levels of unemployment. As a member of the Circus, with Meade, Kahn, Austin Robinson, and Sraffa, Robinson contributed to the critical scrutiny of drafts of Keynes's *The General Theory* (1936). A formidable group, but they were persuaded by the dissenting view of the behaviour of investment and of saving in Keynes's new theory of output and employment. Immediately following the book's publication, she wrote to develop its themes and analysis. *Essays in the Theory of Employment* (1937a) and the popular *Introduction to the Theory of Employment* (1937b) were both published in the year following *The General Theory* (and in the same year that her second daughter was born). She sent Keynes drafts of the two books as well as the proofs, and the related correspondence reveals his respect for Robinson in her role as his interpreter. *Introduction* is still one of the best guides to *The General Theory*.

During the interwar years, Michal Kalecki had a major impact on Robinson's thinking. His preemptive version of the theory of effective demand, which interwove imperfect competition with a macro-economic theory of distribution, was based on Karl Marx's schema of reproduction and his analysis of the realisation crisis. She integrated his double-sided relation between investment and profits and his short-period theory of income distribution to develop her theory of employment and output. Kalecki provided a structural base to Robinson's theoretical explorations and extensions of Keynesianism and later of growth. During the war Robinson wrote *An Essay on Marxian Economics* (1942). Her interest was partly attributable to Kalecki's and Sraffa's Marxist-based work, and partly to the structure of society portrayed by, and fundamental to, such classical and Marxist analysis; her continued dissatisfaction with orthodox theory was also reflected in her essay which compared Marxian with mainstream ideas.

Robinson's interpretation of Marx again illustrates her approach to constructing theory. Responses to her book were to her methodology, as well as to her critical remarks about both Marxian and neoclassical economics. She extracted from Marx's writing the circuits of deductive logic with their attendant assumptions and political implications. The methodology of her censorship was not so inclusive as to embrace Marx's theory of history and of the processes of change. Yet it was these elements—the centrality of historical analysis to Marx's method and of class to understanding effective demand—that attracted Robinson to Marx's writings. The pervasive implications of time and history underlay her vision of capitalism and informed all of her subsequent economic research even though she ultimately was to abandon the classical method. Her book was a challenge to her colleagues to read and learn from Marx, if selectively and from a Keynesian perspective. It remains a useful introduction to Marxist economics.

In the early postwar years Robinson became committed to developing a dynamic analysis of capitalism. Inspired by Harrod's (1939, 1948) studies of long-period growth, *The Rate of Interest and Other Essays* (1952) was a collection in preparation for this project, in particular. The generalization of *The General Theory* attempted to extend Keynes into the long period. This essay was an interpretation of Keynes in terms of the Marshallian short period, characterized by uncertainty, incompatible decisions and unrealizable expectations, by a new definition of equilibrium. The elements of Robinson's growth theory were coming together in a non-neoclassical and long-period framework.

Integral with Robinson's exploration of long-period accumulation was an increasing dissatisfaction with the conceptual and logical foundations of neoclassical theory and its method, for which she became most well known, if not best understood. She prefaced her

constructive contributions to dynamic theory with the critique of the neoclassical production function, Robinson (1953–4). This article fired a debate that continued for more than two decades. The implications of these exchanges were for the legitimacy of the very concept of capital and hence of the aggregate production function, fundamental to neoclassical theory. In particular, Robinson targeted its distribution analysis and its corollary, value theory. Ultimately, she criticized its method, that 'the neo-neoclassics still refuse to understand the difference between a comparison of timeless equilibrium positions and the effects to be expected from a change taking place at a particular moment' (1978a, p. xix). That particular moment was characterised by its own inherited structure of decisions, expectations, and uncertainties. She rejected the neoclassical method of long-period equilibrium as a point towards which the actual economy is moving; the implied stationarity of that theory meant that the movement towards it, and the rest point were logical, not historical. Her own account would not be in equilibrium terms but would take into account the actual features of the economy, the existence of uncertainty, and disappointed expectations. Her understanding of capitalism was based on classes and relations of conflict, while the neoclassicals begin with technical factors and harmonious distributive shares. What she wanted to put in the place of neoclassical theory was not a rival complete theory, which she considered would be only another 'box of tricks,' but an approach that started from institutions, history, and the 'rules of the game' (Harcourt 1990b). Robinson's optimism about the power of economic debate was continually frustrated, yet she determinedly defended what seemed to her to be right and just arguments. Her conservative colleagues, she felt, treated new ideas 'like heresies.'

The Accumulation of Capital (1956) continued the 'generalization' within a classical framework. The questions she set out to explore were: what are the conditions for cumulative long-term growth of income and capital (Golden Age) and what is the outcome of this process in terms of growth of gross and net output and the distribution of income between wages and profits, given a certain evolution through time of the labor force and technology? Inspired by Harrod's 'fruitful provocation,' *The Accumulation of Capital* brought together Marx's historicism, Harrod's economic dynamics, and Wicksell's capital theory. True to her methodology, she started out by presenting her most basic model and then devoted the rest of this remarkable book to relaxing the 'most severely simple' assumptions and tracing out the implications. 'The whole argument,' she proposed, 'is set out ... as an analytical construction, with the minimum of controversy' (1956, p. x). In contrast to the mathematical style endemic to American economics, Robinson wrote her entire work of pure theory in literary form. It was a difficult book to assimilate and several years

later she followed it up with a hands-on guide. *Exercises in Economic Analysis* (1960) fills in the gaps, clarifies obscurities, and extends the arguments from *The Accumulation of Capital*. Its object, Robinson stated, was ‘to sort out questions’ (1960, p. xx). It offers readers an opportunity to formulate for themselves the analytical examples she derives in *The Accumulation of Capital*. This book (1960) was immediately followed by *Essays in the Theory of Economic Growth* (1962) which presents a retrospective introduction for the ‘too terse’ earlier (1956) book. Although there was a hint of re-switching in Robinson (1953–4) as well as in *The Accumulation of Capital*, Robinson dismissed the phenomenon at the time and saw it as a distraction from the main methodological point of her critique of neoclassical capital.

Dissatisfied with her quest for a long-period theory of growth, Robinson looked more to Kalecki’s theory of cycles as a depiction of capitalism. She was a pioneer in the development of what was to become post-Keynesian theory, its bases in Sraffa’s interpretation of Ricardo and in Kalecki’s interpretation of the theory of employment. To this she brought ‘a general framework of long- and short-period analysis which will enable us to bring the insights of Marx, Keynes, and Kalecki into coherent form and apply them to the contemporary scene’ (1978b pp. 1, 18). Within this framework, post-Keynesian theory offered a ‘method of analysis which takes account of the difference between the future and the past’ (1978b, p. 12). Here was a broad conceptual structure that brought together Robinson’s major theoretical concerns. From this analysis, post-Keynesian theory is evolving in several, not always compatible, directions, establishing its authority in the body of economic theory.

Robinson was a charismatic teacher; she inspired successive generations of students to develop their own ideas from her methods of analysis and brilliant insights, to think critically and for themselves. She spent many years demonstrating that ‘mainstream teaching has been inculcating defective methodology’; she wanted to provide a logically sound, historically based dynamic approach to first-year economic analysis. *An Introduction to Modern Economics* (1973) with John Eatwell was a response to the ‘fashionable textbooks [of] mainstream’ economics. It presented a history of thought, a classical-cum-Kaleckian/post-Keynesian analysis. It was an ambitious, presumptuous book, at once a critique of neoclassical theory, and a terse introduction to the alternative. Robinson overestimated the intuitive powers of her audience and it proved too difficult for beginners; it remains a challenge for more advanced students.

Throughout her life Robinson pursued an active interest in the Third World, in India and China in particular, and in the Soviet Union. Her initial writing on development was based on the growth models developed in *The Accumulation of Capital*. Kalecki influenced this book. Development, she argued,

should proceed by appropriate land reform and industrialization in both agriculture and manufacture, mindful of the balance of payments constraint and a minimum real wage or its equivalent. Socialist development, she continued, could mobilize savings and enable the state to organize the investment of the surplus, so achieving faster if not more efficient accumulation and growth than under capitalism; socialism, furthermore, led to a more equitable distribution of income (although the standard of living she saw as a political decision). She was less enthusiastic about the Soviet-style socialism of Eastern Europe than she was about China. Her mission to eradicate poverty and to meet basic needs always guided her judgements. She leaves a body of work; some reprinted in her *Collected Economic Papers* as well as Robinson (1979), on development planning and socialism that is relevant and provocative.

Joan Robinson depicted her work as an agenda for discussion as well as formulating the questions. Yet she has left an important legacy in her substantial literary output. She wrote for colleagues, students, and for the general, interested reader. As well as her many books, there is a stream of challenging evaluative essays written through the 50 years of her career. Her writing includes a theoretically rigorous critique of neoclassical economics that as yet goes unanswered; the questions she has raised present a formidable program for further research. She leaves an exacting methodology, a theoretical structure, and significant analytical details of a new, unified system of political economy.

See also: Capitalism; Capitalism: Global; Demand for Labor; East Asian Studies; Economics; Economic Growth; Measurement; Economic Growth: Theory; Economic History; Economic Transformation: From Central Planning to Market Economy; Economics, History of; Economics: Overview; Employment and Labor, Regulation of; Firm Behavior; Labor Supply; Market Areas; Market Structure and Performance; Marxian Economic Thought; Policy History: State and Economy; Political Economy, History of; Soviet Studies: Economics; Unemployment: Structural

Bibliography

- Harcourt G C 1990a Joan Robinson’s early views on method. *History of Political Economy* 22: 411–27
- Harcourt G C 1990b Robinson, Joan Violet. In: Lord Blake, Nichols C J (eds.) *The Dictionary of National Biography* 1981–1985. Oxford University Press, Oxford, UK, pp. 346–7
- Harcourt G C 1996 Some reflections on Joan Robinson’s changes of mind and their relationship to post-Keynesianism and the economics profession. In: Marcuzzo M C, Pasinetti L, Roncaglia A (eds.) *The Economics of Joan Robinson*. Routledge, London, pp. 317–29
- Harrod R F 1939 An essay in dynamic theory. *Economic Journal* 49: 14–33

- Harrod R F 1948 *Towards a Dynamic Economics*. Macmillan, London
- Keynes J M 1930 *A Treatise on Money*, 2 Vols. In: *Collected Writings*. Macmillan, London, Vols. V, VI
- Keynes J M 1936 *The General Theory of Employment, Interest and Money*. In: *Collected Writings*, Macmillan, London, Vol. VII, 1973
- Robertson D H, Shove G F, Sraffa P 1930 Increasing returns, and the representative firm: A symposium. *Economic Journal* **40**: 79–116
- Robinson J 1932a *Economics is a Serious Subject: The Apologia of an Economist to the Mathematician, the Scientist and the Plain Man*. Heffer, Cambridge, UK
- Robinson J 1932b Imperfect competition and falling supply price. *Economic Journal* **42**: 544–54
- Robinson J 1933 *The Economics of Imperfect Competition*. Macmillan, London
- Robinson J 1937a *Essays in the Theory of Employment*. Macmillan, London
- Robinson J 1937b *Introduction to the Theory of Employment*. Macmillan, London
- Robinson J 1942 *An Essay on Marxian Economics*. Macmillan, London
- Robinson J 1952 *The Rate of Interest and Other Essays*. Macmillan, London
- Robinson J 1952–80 *Collected Economic Papers*. Blackwell, Oxford, UK, Vols. I–VI
- Robinson J 1953–4 The production function and the theory of capital. *Review of Economic Studies* **21**: 81–106
- Robinson J 1956 *The Accumulation of Capital*. Macmillan, London
- Robinson J 1960 *Exercises in Economic Analysis*. Macmillan, London
- Robinson J 1962 *Essays in the Theory of Economic Growth*. Macmillan, London
- Robinson J 1978a *Contributions to Modern Economics*. Blackwell, Oxford, UK
- Robinson J 1978b Keynes and Ricardo. *Journal of Post Keynesian Economics* **1**: 12–8
- Robinson J 1979 *Aspects of Development and Underdevelopment*. Cambridge University Press, Cambridge, UK
- Robinson J, Eatwell J 1973 *An Introduction to Modern Economics*. McGraw-Hill, Maidenhead, UK
- Sraffa P 1926 The laws of returns under competitive conditions. *Economic Journal* **36**: 535–50

G. C. Harcourt and P. M. Kerr

Robustness in Statistics

Robust statistics is concerned with the description and the correction of problems caused by a discrepancy between attributes of the data being analyzed and the implicit or explicit assumptions used in modeling the data. But, of course, robustness is a word used frequently and in diverse situations. In statistics, it also conveys various meanings, always involving the stability of the method under study. The arithmetic mean, for example, is an optimal estimator of the expectation if the data follow exactly a normal

distribution. Tiny deviations of the actual distribution of the data from the normal ideal, however, destroy the optimality and the arithmetic mean can be an arbitrarily bad estimator even though the data has a very nearly normal distribution. The robustness of estimators may also be defined without reference to a model. The arithmetic mean of the three observations (14, 13.7, 26.3), for example, is equal to 18.0 and the lack of stability is expressed through the large effect of the single observation 26.3 on the estimate.

Implicit and explicit assumptions in statistics are of various types. In a generalized linear model, for example, they include the link function, which determines the scale on which the relationship between the response and the explanatory variables is of the form of a linear combination of terms. Another assumption concerns the cumulative distribution for the response, which determines certain aspects of the inherent randomness of the data, such as the form of the relationship between the standard deviation and the expectation of the response variable or the rate of growth of the extreme $(1 - \alpha)$ quantile of the response variable as α tends to zero. The best-known phenomenon observed in the data concerns the so-called *outliers*. This is a term used in describing various incidents in a data base. It may refer to *gross* errors in observations due to faulty transcription, mistakes in the manipulation of a measurement instrument, errors in the preparation of a particular experiment, confusion as to which experiment should have been performed, or some other such mistake. The frequency of such gross errors is much higher than one would expect at first thought. We will also use the term *wild value* for a distinct outlier. The term ‘outlier’ may also refer to the effect, that a sizable proportion of the data seems to behave differently from the rest. In a linear regression, for example, it is sometimes observed that any one regression surface fits only part of the data. This might be due to an overly simple model for the surface, or to the fact that the data is a mixture of cases that cannot be separated into more homogeneous pieces based on the available variables.

Of course, the data can be used to check assumptions and it is certainly good statistical practice to do so. To check details of a model, however, requires huge amounts of data and it is misleading to suggest that diagnostic checking of model assumptions can solve the robustness problem. Good practice in statistics requires the use of methods that do not suffer the drawbacks of the arithmetic mean of loosing performance under tiny violations of the idealistic assumptions. This is especially true in routine and less thoroughly scrutinized applications of statistical methodology. Robust methods retain very good performance, while being less sensitive to assumptions.

Robustness seems similar to nonparametric or, more precisely, distribution-free statistics. This is a superficial and even wrong view. Converting data to ranks, for example, in order to liberate the statistical

analysis from distributional assumptions, does not give particularly good protection from the deleterious effects of atypical observations.

1. Indicators of Robustness

1.1 Sensitivity Functions

Let $T(y_1, \dots, y_n)$ be an estimator of a real parameter, calculated with n observations, $(y_1, \dots, y_n)$. The sensitivity function is a simple idea leading to a better understanding of the estimator. The question being asked is: 'How would the estimator have come out, if the data had not been exactly the ones I observed?' To answer it, one considers the *sensitivity function*

$$SC(y) = (n+1)(T(y, y_1, \dots, y_n) - T(y_1, \dots, y_n)) \quad (1)$$

that is the change in the estimate induced by an additional observation at an arbitrary value y . The multiplier $(n+1)$ in this formula provides a natural scaling of the size of the change.

Examples. For the arithmetic mean $T(y_1, \dots, y_n) = \bar{y} = (y_1 + \dots + y_n)/n$ one finds

$$SC(y) = (n+1) \left(\frac{y + y_1 + \dots + y_n}{n+1} - \frac{y_1 + \dots + y_n}{n} \right) = y - \bar{y}. \quad (2)$$

Even in this easy case, the sensitivity function is of interest. It shows that the arithmetic mean is sensitive to the distance of the additional observation y from the average of the remaining data. The larger this distance, the larger the change.

For the trimmed mean, $T(y_1, \dots, y_5) = (y_{(2)} + \dots + y_{(4)})/3$, obtained by deleting the largest and the smallest observation, the sensitivity depends on the position of the additional observations y . If y is between $y_{(1)}$ and $y_{(5)}$, it will not be trimmed from the sample and $SC(y) = 6((y_{(2)} + y_{(3)} + y_{(4)} + y)/4 - (y_{(2)} + y_{(3)} + y_{(4)})/3)$. If on the other hand, y is larger than $y_{(5)}$, then y will be trimmed and the sensitivity becomes $SC(y) = 6((y_{(2)} + y_{(3)} + y_{(4)} + y_{(5)})/4 - (y_{(2)} + y_{(3)} + y_{(4)})/3)$, which is independent of y . The same reasoning applies in the case where y is smaller than $y_{(1)}$.

For the least-squares estimate of a regression coefficient, the algebra is a bit more difficult, partly because the data is of a more complex structure. Suppose we observe the couples of explanatory variables and responses, $(x_1, y_1), \dots, (x_n, y_n)$, and that we add an observation at (u, v) . The least-squares estimate of the regression coefficients is $T(y_1, \dots, y_n) = (X^T X)^{-1}$

$X^T y$, where X is the design matrix. After some algebra, one has

$$SC(u, v) = (v - u^T T) \frac{(n+1)(X^T X)^{-1} u}{1 + u^T (X^T X)^{-1} u}. \quad (3)$$

Again, this is a revealing formula, because it makes it clear that the estimated regression coefficients are sensitive to the size of the residual, the first factor, but also to the position of u among the explanatory variables as expressed by the second factor.

In practice, it is not really necessary to do the computations illustrated in these examples. The sensitivity function can easily be computed numerically for any given estimator and underlying data set and one can thus construct the plot of $SC(y)$ against y , which in most cases will serve as a good basis for judging the robustness qualities of the estimator. From such a plot one can deduce several indicators, the simplest one being the *maximal sensitivity*

$$g^* = \sup_y |SC(y)| \quad (4)$$

where $|SC|$ denotes the absolute value of SC . In the above examples, the maximal sensitivity is equal to ∞ for the mean and the least-squares estimate of the regression coefficient—the first and the third example. It is finite for the trimmed mean. An estimator with infinite maximal sensitivity is necessarily *non-resistant* in the sense that a single wild observation that behaves differently from the rest can destroy the estimate and lead to rather arbitrary conclusions.

1.2 Influence Functions

The sensitivity function has a more formal counterpart in the *influence function*, which is defined as a kind of derivative of a statistical functional. In many cases it is simply the limit as the sample size n tends to infinity of the sensitivity function computed for a sample generated from a fixed model.

Examples. Let $y_1, \dots, y_n$ be a sample from some distribution F with expectation μ_F . As n increases, the sensitivity function of the arithmetic mean converges to the influence function of the expectation, which is $IF_F(y) = y - \mu_F$. The influence function of the trimmed mean obtained by trimming a fraction of α at each end is

$$IF_F(y) = \begin{cases} (y - \mu_F)/(1 - 2\alpha), & \text{for } y \leq F^{-1}(1 - \alpha); \\ \text{sign}(y - \mu_F)F^{-1}(1 - \alpha)/(1 - 2\alpha), & \text{otherwise} \end{cases} \quad (5)$$

where F is any distribution symmetric around μ_F . In the regression case, one can choose an arbitrary error distribution F with finite variance and an arbitrary true regression coefficient θ . The influence function is

$$\text{IF}_F(\mathbf{u}, v) = (v - \mathbf{u}^T \theta) (\text{ave}(\mathbf{x}\mathbf{x}^T))^{-1} \mathbf{u} \quad (6)$$

where the matrix $\text{ave}(\mathbf{x}\mathbf{x}^T)$ denotes the average design matrix.

The influence function is very similar to the sensitivity function and can thus be interpreted similarly. In particular, the *gross error sensitivity* is

$$\gamma_F^* = \sup_y |\text{IF}_F(y)|. \quad (7)$$

The squared influence function is directly linked to the asymptotic variance of an estimator. Let $Y_1, \dots, Y_n$ be independent and identically distributed with distribution F_θ for some real parameter θ . One can show that most estimators of θ satisfy the following approximate relationship:

$$T(Y_1, \dots, Y_n) \approx \theta + \text{ave}(\text{IF}_{F_\theta}(Y_i)). \quad (8)$$

If this approximation is sufficiently precise, it implies that the asymptotic variance of the estimate T is

$$\text{Var}_{\text{asy}}(T) = E_{F_\theta}((\text{IF}_{F_\theta}(Y))^2) = \int (\text{IF}_{F_\theta}(t))^2 F_\theta(dt). \quad (9)$$

In this case, one can also apply the central limit theorem to conclude that

$$P(T \leq t) \approx \Phi \left(\frac{\sqrt{n}(t - \theta)}{\sqrt{\text{Var}_{\text{asy}}(T)}} \right) \quad (10)$$

where Φ denotes the normal distribution function with density $\varphi(t) = \exp(-t^2/2)/\sqrt{2\pi}$. This in turn shows that

$$[T \pm 1.96 \sqrt{\text{Var}_{\text{asy}}(T)}/\sqrt{n}] \quad (11)$$

is an approximated 95 percent confidence interval for θ .

1.3 Breakdown and High Resistance

A limitation of the influence function is its local nature. It is based on the change of the value of the estimator from one sample to another one that differs by adding a single small ‘contamination.’ The *breakdown point* is the simplest global measure of ro-

business. It is the largest percentage of wild values in a data set that an estimator can tolerate without giving wild estimates. This notion is inspired by the trimmed means, computed as the mean of a sample after deleting 100 α percent of the largest and of the smallest data points. This estimator has a breakdown point of 100 α percent, since any contamination by wild data below this fraction is being trimmed. A special case of the trimmed means is the median, which for an odd-sized sample ($n = 2k + 1$) has a breakdown point of $k/n \approx 50$ percent. This is so, because the median is computed as the $(k + 1)$ st value in the sample and anything less than k wild values will not lead to a wild estimate. More generally, global indicators describe the properties of an estimator under contamination of a sizeable portion of the data.

The distributions used in modeling data sets containing outliers are *heavy-tailed*. One can, for example, model a sample containing a sizeable portion of outliers with a *contaminated distribution*. A sample generated from $(1 - \varepsilon)F + \varepsilon H$ is a mixture of one part, $(1 - \varepsilon)$, of a ‘good’ distribution F and another part, ε , of a ‘bad’ distribution G . The *contaminated normal distribution*, $(1 - \varepsilon)\Phi(y) + \varepsilon\Phi(y/\sigma)$, describes observations that can either result from a standard normal distribution, or from a normal distribution having a larger variance $\sigma > 1$. As ε grows from 0 to 50 percent, a larger and larger fraction of the data is expected to exhibit a distinct behavior. A global robustness indicator for estimators is based on behavior under positive fractions ε , whereas local properties, such as the influence function, merely consider limits as ε tends to zero.

An estimator with a breakdown point of 50 percent is called *highly resistant*. Such estimators can be very revealing when applied to data containing a large proportion of outliers, because they lead to a clear separation of the ‘good’ and the ‘bad’ data. The median is an example of a highly resistant location estimator. The residuals from the arithmetic mean of the three observations (14.0, 13.7, 26.3) are equal to $(14.0 - 18.0 = -4.0, -4.3, 8.3)$, whereas the residuals from the median are $(14.0 - 14.0 = 0.0 - 0.3, 12.3)$. The highly resistant median produces residuals that make it immediately evident that the third observation is a wild value. The ability of estimators to reveal hidden outliers is explained by the fact that they have bounded sensitivity with regard to single observations, but also, and more importantly, that even a largish fraction of outliers does not exert an undue influence on the estimate. The median absolute deviation (MAD) is a highly resistant scale estimator. It is defined as

$$\text{MAD}(y_1, \dots, y_n) = \text{median}(|y_1 - m|, \dots, |y_n - m|) \quad (12)$$

where m denotes the median, $\text{median}(y_1, \dots, y_n)$.

2. Robust Estimators

2.1 Least-squares, maximum likelihood, and M-estimators

Least-squares estimators,

$$T(y_1, \dots, y_n) = \arg \min_{\theta} \sum_{i=1}^n (y_i - \mu_i(\theta))^2 \quad (13)$$

have an unbounded influence function and a breakdown point of 0 percent. Maximum likelihood estimators (m.l.e.) for independent data,

$$\begin{aligned} T(y_1, \dots, y_n) &= \arg \max_{\theta} \sum_{i=1}^n \log(f_i(y_i, \theta)) \\ &= \text{solution of } \left(\sum_{i=1}^n \partial \log(f_i(y_i, \theta)) / \partial \theta = \mathbf{0} \right) \end{aligned} \quad (14)$$

can be robust, depending on the properties of the densities f_i . Generalizing either of these two concepts, one obtains *M-estimators*, defined by

$$T(y_1, \dots, y_n) = \arg \min_{\theta} \sum_{i=1}^n \rho_i(y_i, \theta) \quad (15)$$

where the choice of ρ_i is in principle quite free. The corresponding estimating equation is of the form

$$\sum_{i=1}^n \psi_i(y_i, T(y_1, \dots, y_n)) = \mathbf{0} \quad (16)$$

where $\psi_i(y, \theta) = \partial \rho_i(y_i, \theta) / \partial \theta$.

Example. For illustration, we consider here briefly the linear regression case, where $\mu_i(\theta) = \mathbf{x}_i^T \theta$, $f_i(y_i, \theta) = f(y_i - \mathbf{x}_i^T \theta)$, $\rho_i(y_i, \theta) = \rho(y_i - \mathbf{x}_i^T \theta)$, and $\psi_i(y_i, \theta) = (\psi(y_i - \mathbf{x}_i^T \theta) \mathbf{x}_i)$ with $\psi(t) = \rho'(t)$. The M-estimator with $\rho(t) = t^2/2$ (rho-function) and $\psi(t) = t$ (psi-function) is nothing else but the least-squares estimator, which is the same as the m.l.e. for Gaussian errors. Any m.l.e. can be thought of as an M-estimator with $-\rho(t) = \log(f(t))$, but thinking in terms of M-estimators is an advantage and an important conceptual step since it frees us from any artificial constraints. The influence function of the M-estimator at the error distribution F and the true regression parameter θ is

$$\text{IF}_F(\mathbf{u}, v) = \frac{\psi(v - \mathbf{u}^T \theta)}{\int_{-\infty}^{\infty} \psi'(t) F(dt)} (\text{ave}(\mathbf{x} \mathbf{x}^T))^{-1} \mathbf{u}. \quad (17)$$

The first factor in this influence function involves the residual at the contaminating point $(\mathbf{u}, v)$. It is

bounded whenever $\psi = \rho'$ is bounded. The deficiencies in the stability of the least-squares estimator are therefore best explained by the fact that the derivative $\rho'(t) = \psi(t) = t$ is unbounded.

M-estimators are defined by a rho-function and an associated psi-function. As shown above, boundedness of the psi-function is a necessary condition to achieve robustness. Careful tuning of the shape of the psi-function permits both robustness and efficiency. It is well-known that the least-squares estimator, which has psi-function $\psi(t) = t$, is optimally efficient when the error distribution is exactly Gaussian. To preserve this Gaussian efficiency as much as possible, the linearity of the psi-function in a neighborhood of the origin $t = 0$ should remain unaltered. The *Huber estimator* is based on

$$\begin{aligned} \psi_H(t) &= \min(1, \max(-1, t)) \\ &= \begin{cases} -1, & \text{if } t \leq -1; \\ t, & \text{if } -1 < t \leq 1 \\ 1, & \text{if } t > 1. \end{cases} \end{aligned} \quad (18)$$

The reader should draw this function to note that it does indeed preserve the linearity near the origin, while at the same time enforcing a bound. In this definition, the interval $[-1, 1]$ plays a special and undeserved role. It goes without saying that in any practical use of the estimator this interval must be adapted to the data.

Example. The Huber regression estimator is defined as the solution to the estimating equation

$$\sum_{i=1}^n \psi_H \left(\frac{y_i - \mathbf{x}_i^T \theta}{k \times S} \right) \mathbf{x}_i = \mathbf{0} \quad (19)$$

where $S = S(r_1, \dots, r_n)$ is a scale estimate based on residuals and $k > 0$ is a further tuning constant. Consider S as known and computed before solving the estimating equation (19). Dividing the argument of the psi-function by a multiple of a residual scale achieves the adaptation we were looking for. It is quite evident that the tuning constant k introduces a trade-off between robustness and Gaussian efficiency. For sufficiently large values of k the estimator is equal to the least-squares estimator because all scaled residuals $(y_i - \mathbf{x}_i^T \theta) / (k \times S)$ will lie in the interval $[-1, 1]$ and thus in the linear part of Huber's psi-function. For small values of k , the Huber estimator loses Gaussian efficiency but gains in robustness. In the limit, as $k \rightarrow 0$, the scaled residuals are either exactly zero, smaller than -1 , or larger than 1 . In this case, the function ψ_H returns as values either 0 , -1 , or 1 and the above estimating equation turns into

$$\sum_{i=1}^n \text{sign}(y_i - \mathbf{x}_i^T \theta) \mathbf{x}_i = \mathbf{0} \quad (20)$$

which corresponds to the so-called L_1 -estimator defined by $\psi(t) = \text{sign}(t)$ and $\rho(t) = |t|$ = absolute value of t . A good compromise choice for the denominator $k \times S$ is $1.3 \times \text{MAD}$ (residuals).

2.2 Theoretical Justification of Huber's Estimator

A concise discussion of many of the topics discussed in this entry can be found in the two books by Huber (1981, 1996). Huber's location estimators T_k in the case of known scale ($S = 1$) are solutions of

$$\sum_{i=1}^n \psi_H \left(\frac{y_i - T_k(y_1, \dots, y_n)}{k} \right) = \sum_{i=1}^n \psi_k(y_i - T_k(y_1, \dots, y_n))/k = 0 \quad (21)$$

where

$$\begin{aligned} \psi_k(t) &= \min(k, \max(-k, t)) \\ &= \begin{cases} -k, & \text{if } t \leq -k; \\ t, & \text{if } -k < t \leq k \\ k, & \text{if } t > k. \end{cases} \end{aligned} \quad (22)$$

The corresponding rho-function is

$$\rho_k(t) = \begin{cases} -k(t + k/2), & \text{if } t \leq -k; \\ t^2/2, & \text{if } -k < t \leq k \\ k(t - k/2), & \text{if } t > k, \end{cases} \quad (23)$$

that is quadratic in the middle and linear in the tails. These M-estimators are m.l.e.s in the sense that $\exp(-\rho_k(t))$ can be standardized to become a density, $f_k(t)$. From the form of $\rho_k(t)$ it follows that $f_k(t)$ is like a Gaussian in the center and has tails that decrease exponentially, i.e.,

$$f_k(t) = \begin{cases} (1 - \varepsilon_k) \exp(-t^2/2)/\sqrt{2\pi}, & \text{if } -k \leq t \leq k \\ (1 - \varepsilon_k) \exp(k^2/2 - k|t|)/\sqrt{2\pi}, & \text{if } |t| > k \end{cases} \quad (24)$$

where $0 \leq \varepsilon_k \leq 1$ decreases as k increases. These distributions were discovered by Huber and have been discussed in his seminal paper (Huber 1964). The above formula shows that $f_k(t) = (1 - \varepsilon_k)\phi(t) + \varepsilon_k h_k(t)$ is a *contaminated density* with contamination fraction ε_k . Huber showed that f_k is in some mathematical sense

the worst among all contaminated densities with a contaminating fraction of ε_k . As a consequence, the corresponding Huber estimator T_k is the estimator that best protects against a contamination of 100 ε_k percent of the data. The sets of possible error distributions considered in Huber's paper contain a positive fraction $\varepsilon_k > 0$ of an unspecified 'bad' density. The Huber estimator is therefore the best one in a global sense. Even if a sizeable portion of the data are not of any real use, this estimator will perform as well as possible under the circumstances. All of this, however, is only true for the simplest case of a location parameter with known scale. In more general regression models Huber's result no longer holds.

Indicators of robustness such as the breakdown point and the gross error sensitivity lead rather naturally to optimality problems. One is interested, for example, in finding an estimator that satisfies the two conditions of consistency and bounded gross error sensitivity and which, given these conditions, is most precise. This problem is of a local nature, because we do not consider sizeable fractions of outliers, but rather limits for very small fractions. The problem is again solved by the Huber estimators, which shows—not very surprisingly—that a globally best estimator is also locally best. The estimators T_k form a family of least variable estimators with gross error sensitivities that increase as k increases. In other words, they provide a bridge between the median ($k = 0$) and the mean ($k = \infty$) and between a minimal value of the gross error sensitivity and an infinite gross error sensitivity. Considering only local properties defined by the influence function has the advantage of being more easily generalized. Thus, it is possible to solve a variety of problems of this type. The book by Hampel et al. (1986) provides a detailed discussion of related results.

2.3 Redescending M-estimators

The form of the Huber estimator comes as a surprise because it never entirely rejects gross errors among the observations. In order to completely reject such outliers, the psi-function $\psi(t)$ would have to be equal to zero for sufficiently large $|t|$. The *biweight estimator* proposed by Tukey is a convenient example. It is based on

$$\psi_T(t) = \begin{cases} t(1 - t^2)^2, & \text{if } -1 \leq t \leq 1 \\ 0, & \text{if } |t| > 1. \end{cases} \quad (25)$$

Such an estimator is called *redescending*, because the psi-function remains roughly linear near the origin, reaches a bound, and subsequently redescends towards zero. The reader is invited to draw Tukey's biweight

psi-function in order to appreciate these qualities. The biweight regression estimator solves the same estimating equation (19) as Huber's estimator, except that we replace ψ_H by ψ_T . This small change has important repercussions. Because ψ_T is not an increasing function, the estimating equation of the biweight estimator may have multiple solutions. The estimate is therefore not well defined. However, among all the solutions of the estimating equations, an appropriate one can usually be picked out by looking at the resulting residuals. Also, when multiple solutions do exist, several of them may be of interest. An appropriate choice for the denominator $k \times S$ when using the biweight is $7 \times \text{MAD}$ (residuals). The rho-function of the biweight estimator,

$$\rho_T(t) = \begin{cases} (1 - (1 - t^2)^3)/6, & \text{if } -1 \leq t \leq 1; \\ 1/6, & \text{if } |t| > 1 \end{cases} \quad (26)$$

is shaped like an vase with a flat rim and the solutions of the estimating equation are local minima of the criterion function

$$\text{criterion}(\theta) = \sum_{i=1}^n \rho_T \left(\frac{y_i - \mathbf{x}_i^T \theta}{k \times S} \right). \quad (27)$$

Redescending M-estimators such as the biweight are of interest in the quest for highly resistant regression estimators. Inspection of the influence function (17) of the regression M-estimator reveals a weakness, namely the unboundedness of the second factor. In cases where one observation is far from the other observations in the sense that $\mathbf{u}$ is distant from the other observations taken at $\mathbf{x}_p$, the influence of this distant observation on the estimated regression parameter will be considerable. Redescending estimators for which the first factor in the influence function can be equal to zero correct this problem. Whenever the data give support to several regression equations, the redescending M-estimator can be of great value in pointing out the different solutions, each one as a separate local minimum of the criterion.

Rousseeuw's *least-median-of-squares* (LMS) estimator is another highly resistant regression estimator. It is defined by the criterion

$$\text{criterion}(\theta) = \text{median}\{(y_1 - \mathbf{x}_1^T \theta)^2, \dots, (y_n - \mathbf{x}_n^T \theta)^2\}. \quad (28)$$

The LMS is a useful, often revealing, estimate. It is, however, quite unstable, not with regard to outliers, but with regard to other features of the data.

2.4 Finite Samples

A robust data analysis must be valid and efficient beyond a narrow set of circumstances. The classical

regression estimator, for example, is chosen to be the best possible for normally distributed observations

$$y_i = \mathbf{x}_i^T \theta + \varepsilon_i, \quad \varepsilon_i \sim \text{normal}(0, \sigma^2). \quad (29)$$

Robustness demands joint optimality for a broader set of error distributions. One way in which this can be achieved, is by choosing a finite number of challenges, in the simplest case a couple such as {normal, Student's t_2 }, the idea being that the estimator would have to perform well, whether the errors were generated by one or the other of the distributions. For such a set of challenges, the performance of an estimator is determined by a set of efficiencies, one for each challenge. Typically, an estimator that is highly efficient for one of the challenges will have a low efficiency for all the others. In the example of a couple of challenges including the normal distribution and the heavy-tailed Student's t_2 distribution, the least-squares estimator, for example, has optimal efficiency for the Gaussian but is inefficient for the t_2 . Robust estimators on the other hand must have a high efficiency for all challenges. How to compute estimators that maximize jointly the whole set of efficiencies is discussed in Morgenthaler and Tukey (1991).

2.5 Iteratively Re-weighted Least-squares

The computation of the least-squares estimator is straightforward because the estimating equation is linear. A program for computing the weighted least-squares solution is all that is needed to find the values of M-estimators. To illustrate the idea, we consider the solution of (21). Rewriting the estimating equation as

$$\sum_{i=1}^n \left[\frac{\psi_k(y_i - T_k(y_1, \dots, y_n))}{(y_i - T_k(y_1, \dots, y_n))} (y_i - T_k(y_1, \dots, y_n)) \right] = 0 \quad (30)$$

an algorithm immediately suggests itself. We could start with a reasonable guess T_k^0 for the value of T_k , such as the median of the observations. Next, we compute weights

$$w_i^0 = \frac{\psi_k(y_i - T_k^0)}{y_i - T_k^0} \quad (31)$$

and then solve the estimating equation, $\sum_{i=1}^n w_i^0 (y_i - T_k(y_1, \dots, y_n)) = 0$. From this, we obtain the next guess for T_k , i.e.,

$$T_k^1 = \sum_{i=1}^n w_i^0 y_i / \sum_{i=1}^n w_i^0. \quad (32)$$

With this new value for T_k we are back at the start of the loop, ready to recompute the weights, and then

the next value for T_k . Repeating the loop a fixed number of times or until convergence leads to practical values for M-estimates.

3. Robust Inference

The sampling distributions of robust estimators can often be approximated sufficiently well by their asymptotic normal distributions. For the regression M-estimator the following formula gives an estimated variance matrix of the estimated regression coefficients

$$\frac{\frac{1}{n} \sum_{i=1}^n (kS)^2 \left(\psi \left(\frac{r_i}{kS} \right) \right)^2}{\left(\frac{1}{n} \sum_{i=1}^n \psi' \left(\frac{r_i}{kS} \right) \right)^2} \quad (33)$$

where $r_1, \dots, r_n$ are the residuals. Using the approximate normality of the estimate, this equation allows one to compute robust confidence intervals for individual parameters. The robustness in the context of confidence intervals means that the coverage of the parameter is equal to the nominal coverage rate even under deviations from the model. Furthermore, the interval should not achieve this feat by being overly long.

Distribution-free tests are very resistant in one respect. Their size, that is the probability of falsely rejecting the null hypothesis, is stable. The power, however, depends, of course, on the distribution of the observations and need not be stable under changes to this distribution. Robust tests must have both, stable sizes and good power under a variety of circumstances. This point illustrates once more the fact that robustness and distribution-freeness are two quite distinct aspects of statistical inference.

See also: Distributions, Statistical: Approximations; Distributions, Statistical: Special and Continuous; Distributions, Statistical: Special and Discrete; Semi-parametric Models; Statistics: The Field

Bibliography

- Box G E P, Leonard T, Wu C F (eds.) 1983 *Scientific Inference, Data Analysis, and Robustness Proceedings of a Conference Conducted by the Mathematics Research Center, The University of Wisconsin–Madison, November 4–6, 1981*. Academic Press, New York
- Gnanadesikan R 1977 *Methods for Statistical Data Analysis of Multivariate Observations*. Wiley, New York
- Hampel F R, Ronchetti E M, Rousseeuw P J, Stahel W A 1986 *Robust Statistics: The Approach Based on Influence Functions*. Wiley, New York
- Hoaglin D C, Mosteller F, Tukey J W (eds.) 1983 *Understanding Robust and Exploratory Data Analysis*. Wiley, New York
- Hoaglin D C, Mosteller F, Tukey J W (eds.) 1985 *Exploring Data Tables, Trends, and Shapes*. Wiley, New York
- Huber P J 1964 Robust estimation of a location parameter. *Annals of Mathematical Statistics* **35**: 73–101
- Huber P J 1981 *Robust Statistics*. Wiley, New York
- Huber P J 1996 *Robust Statistical Procedures (2nd edn.) CBMS-NSF Regional Conference Series in Applied Mathematics* **68**. SIAM, Philadelphia
- Jurečková J, Sen P K 1996 *Robust Statistical Procedures: Asymptotics and Interrelations*. Wiley, New York
- Kadane J B (ed.) 1984 *Robustness of Bayesian Analyses*. North-Holland/Elsevier, Amsterdam
- Launer R L, Wilkinson G N (eds.) 1979 *Robustness in Statistics: Proceedings of a Workshop Sponsored by the Mathematics Division, Army Research Office, held at Army Research Office, Weiss Buildings, April 11–12, 1978*. Academic Press, New York
- Morgenthaler S, Ronchetti E, Stahel W A (eds.) 1993 *New Directions in Statistical Data Analysis and Robustness*. Birkhäuser, Basel
- Morgenthaler S, Tukey J W (eds.) 1991 *Configural Polysampling: A Route to Practical Robustness*. Wiley, New York
- Rieder H 1994 *Robust Asymptotic Statistics*. Springer-Verlag, Berlin
- Rousseeuw P J, Leroy A M 1987 *Robust Regression and Outlier Detection*. Wiley, New York
- Sprent P 1993 *Applied Nonparametric Statistical Methods*, 2nd edn. Chapman and Hall, London
- Stahel W, Weisberg S (eds.) 1991 *Directions in Robust Statistics and Diagnostics, Parts I and II*. Springer-Verlag, Berlin
- Staudte R G, Sheather S J 1990 *Robust Estimation and Testing*. Wiley, New York
- Wilcox R R 1997 *Introduction to Robust Estimation and Hypothesis Testing*. Academic Press, San Diego

S. Morgenthaler

Rogers, Carl Ransom (1902–87)

Carl R. Rogers was arguably America's most influential psychotherapist and one of its most prominent psychologists. His therapeutic approach, known most often as 'client-centered therapy,' was validated through voluminous scientific research and taught widely, profoundly affecting the fields of counseling and psychotherapy throughout the world and permeating many other helping professions. A leader in the humanistic psychology movement of the later twentieth century, Rogers' writings and personal example influenced individuals personally and professionally in many continents.

1. Education and Influences

Rogers was born on January 2, 1902 in suburban Oak Park, Illinois, where he spent his childhood with four brothers, a sister, and conservative, Protestant

parents. Spending teenage years on the family farm in Glen Ellen, Illinois, he developed a love of nature and serious working knowledge of scientific method, conducting agricultural experiments on a plot he and his brothers managed. He majored in agriculture at the University of Wisconsin, Madison, but switched to history after experiencing the call to religious work. Motivated by the 'social gospel' as much as religious conviction, he eschewed his parents' dogmatism and attended liberal Union Theological Seminary in New York City, also taking courses at adjoining Teachers College, Columbia University. There, religious doubt combined with fascination with psychology. Influenced by instructors Leta Stetter Hollingworth, Goodwin Watson, William Heard Kilpatrick, the leading proponent of John Dewey's progressive education philosophy, and others, Rogers transferred to Teachers College to pursue a doctorate in clinical psychology.

At Columbia, exposure to the testing and measurement movement of E. L. Thorndike was balanced by his experiences in clinical work and fellowship at the Institute for Child Guidance, where he encountered Freudian thought, a lecture by Alfred Adler, Rorschach testing, and other psychoanalytic and psychiatric approaches. Seeking to integrate these influences, Rogers came to appreciate the importance of attending to and honoring clients' inner experience while also striving to assess objectively any changes which resulted from treatment. His doctoral dissertation, *Measuring Personality Adjustment in Children Nine to Thirteen Years of Age* (1931), combined both subjective and objective measures. Rogers' 'Personality Adjustment Inventory' was subsequently published by YMCA Press and used for over 50 years.

His next major learning experience was the 12 years working with thousands of troubled children and adults in Rochester, New York, where he gradually developed his nondirective approach to counseling and psychotherapy. During this period, he was influenced by students of Otto Rank, especially Jessie Taft whose 'relationship therapy' shifted emphasis from past content to a focus on the patient's self-insight and self-acceptance within the therapeutic relationship.

2. Positions Held

In Rochester, Rogers worked at the Society for Prevention of Cruelty to Children (1928–38), where he directed the Child Study Department, then became director of the Rochester Guidance Center (1939–40). Based on his practical experience and first book, *Clinical Treatment of the Problem Child* (1939), he was offered a full Professorship at Ohio State University, where he taught from 1940 to 1944. After an interim year training United Service Organization (USO) workers to counsel returned servicemen, he moved to

the University of Chicago, where he developed and ran the internationally renowned Counseling Center and taught in the Psychology Department from 1945 to 1957. He held joint appointments in the Departments of Psychology and Psychiatry at the University of Wisconsin from 1957 to 1963 before leaving academic life and moving to the Western Behavioral Sciences Institute in La Jolla, California (1963–72). He and others then formed their own organization, Center for Studies of the Person, where Rogers remained until his death in 1987.

3. Client-centered Therapy

Rogers' approach to counseling and psychotherapy was based on a core hypothesis about human growth and personality change, which he summarized in 1950:

Client-centered therapy operates primarily upon one central and basic hypothesis which has undergone relatively little change with the years. This hypothesis is that the client has within himself the capacity, latent if not evident, to understand those aspects of his life and of himself which are causing him pain, and the capacity and the tendency to reorganize himself and his relationship to life in the direction of self-actualization and maturity in such a way as to bring a greater degree of internal comfort. The function of the therapist is to create such a psychological atmosphere as will permit this capacity and this strength to become effective rather than latent or potential.

While other therapies might profess similar belief, Rogers' method of creating the therapeutic psychological atmosphere was radically different from other approaches commonly employed. Avoiding questions, interpretation, or other directive techniques, Rogers' initial 'nondirective method' relied exclusively on careful listening and skillful 'reflection of feelings,' leading to client insight and positive action. Although he always remained primarily nondirective in his practice, Rogers soon recognized that the counselor's *attitudes* were more important than his particular techniques. Still later he clarified that it was the therapeutic *relationship*, which the attitudes helped create, that was growth producing. Rogers called this therapeutic relationship 'client-centered' and described three key 'conditions' in the relationship that bring about positive change in clients.

First is to accept the client as he or she is, as a person of inherent worth possessing both positive and negative feelings and impulses. Rogers called this acceptance and prizing of the person 'unconditional positive regard.' Second is 'the therapist's willingness and sensitive ability to understand the client's thoughts, feelings, and struggles from the client's point of view ... to adopt their frame of reference.' Third is to be 'genuine, or whole, or congruent in the relationship ... It is only as [the therapist] is, in that re-

lationship, a unified person, with his experienced feeling, his awareness of his feelings, and his expression of those feelings all congruent or similar, that he is most able to facilitate therapy.'

When a counselor communicates this unconditional positive regard, empathic understanding, and congruence so that the client perceives them, the 'necessary and sufficient conditions for therapeutic personality change' are present. Rogers argued and demonstrated that clients have within themselves the ability and tendency to understand their needs and problems, to gain insight, to reorganize their personality, and to take constructive action. What clients need is not the judgment, interpretation, advice, or direction of experts, but supportive counselors and therapists to help them rediscover and trust their 'inner experiencing' (a concept borrowed from Gendlin), achieve their own insights, and set their own direction.

4. Recording Therapy

Rogers was able to demonstrate these propositions cogently, in part because he was the first person to record and publish complete cases of psychotherapy. This was a remarkable achievement before the invention of tape recorders, requiring a microphone in the counseling room connected to alternating phonograph machines in an adjoining room, which cut grooves in blank record disks that had to be changed every three minutes. With graduate student Bernard Covner, Rogers and his team recorded thousands of disks, involving scores of clients. These recordings were pivotal in the clinical training of psychotherapists which, in the 1940s, Rogers may have been the first to offer in an American university setting.

Beyond audio recording of therapy sessions, Rogers was among the first to make cinematic recordings of counseling and psychotherapy. The American Academy of Psychotherapists became a leading distributor of training tapes and movies, with Rogers the most frequent therapist portrayed. A still widely distributed set of training films showed Rogers, gestalt therapist Frederick Perls, and rationale-emotive therapist Albert Ellis each demonstrating his method with the same client.

5. Scientific Research and Theory

Recording actual therapy sessions also provided the data with which Rogers and his colleagues conducted more scientific research on one therapeutic approach than had ever been undertaken before. Rogers devised and used numerous instruments for measuring the variables of client-centered therapy and its outcomes: therapist acceptance, empathy and congruence; client expression of feelings, insight, self-concept, self-acceptance, self-ideal, level of experiencing; clients'

positive actions, social adjustment, and numerous other variables. In 1959 the American Psychological Association awarded Rogers its first 'Distinguished Scientific Contribution Award' for developing an original method to objectify the description and analysis of the psychotherapeutic process, for formulating a testable theory of psychotherapy and its effects on personality and behavior, and for extensive systematic research to exhibit the value of the method and explore and test the implications of the theory.

As the award citation suggests, Rogers was interested in psychological *theory* and in the effects of therapy on *personality* as well as behavior. Building upon the Gestalt and phenomenological movements in psychology, and work of his students Victor Raimy and Arthur Combs, he developed a 'self-theory' of personality which is still included in many textbooks (Rogers 1959). The theory describes how an individual's concept of self emerges, how the process of socialization causes individuals to distrust their feelings and sense of self, how experiences which are inconsistent with the concept of self become denied and distorted causing personal distress and psychological problems, and how the therapeutic relationship can help the individual restructure the sense of self, allowing previously denied and distorted experience into awareness, leading to reduction in stress and openness to new experiencing.

6. Impact on the Professions

More than anyone, Rogers helped spread professional counseling and psychotherapy beyond psychiatry and psychoanalysis to other helping professions. While not the first to use the term 'client' for the recipient of therapy, he popularized it, emphasizing that 'a person seeking help was not treated as a dependent patient but as a responsible client,' and that those in psychological distress were not necessarily 'sick,' requiring treatment by medical specialists. Rather all people could be helped by the growth-producing conditions of empathic understanding, unconditional positive regard and congruence, and professionals from many fields could be trained to provide these conditions. Thus counselors, social workers, clergy, medical workers, youth and family workers, and other helping professionals could employ counseling methods and client-centered attitudes in their work. Rogers' books *Counseling and Psychotherapy* (1942) and *Client-Centered Therapy* (1951) described the principles of effective therapy and presented ample case studies from recorded sessions to illustrate his points. In 1982, surveys in the *Journal of Counseling Psychology* and *American Psychologist* still ranked Carl Rogers as the most influential author and psychotherapist.

Rogers also impacted the helping professions through leadership in many professional associations. He served on the executive committee of the American

Association of Social Workers, was Vice-President of the American Association of Orthopsychiatry (social workers), and President of the American Association of Applied Psychology, the American Psychological Association, and the American Academy of Psychotherapists, among other distinguished positions.

7. Applications to Other Fields

Through teaching, writing, and personal example, Rogers continually applied his theory and method to other fields—education, parenting, group leadership, and the health professions, to name a few. In each instance he demonstrated how the facilitative conditions of empathy, positive regard, and congruence could unleash growth, creativity, learning, and healing in children, students, group members, patients, and others. Applied to education, his views on ‘student-centered learning’ coincided with and contributed to the ‘open education’ movement in the USA, Great Britain, and elsewhere. His book *Freedom to Learn* (1969) went through two new editions over 25 years.

He expanded his theory and practice into the areas of group leadership and group facilitation over several decades. In the 1940s Rogers, Thomas Gordon, and colleagues had experimented with ‘group-centered leadership,’ whereby the leader’s acceptance, understanding, genuineness, and willingness to let the group set its own directions stimulated great energy, creativity, and productivity among group members. In the late 1950s and 1960s, Gordon, Richard Farson, Rogers, and associates extended this approach to what Rogers called the ‘basic encounter group’—an unstructured group experience in which ‘normal’ group members came to greater self-understanding, spontaneity, improved communication, and genuineness in relationships. Rogers led scores of encounter groups in professional, business, religious, medical, academic, personal growth, and organizational settings. *Look* magazine called Rogers an ‘elder statesman of encounter groups.’ *Carl Rogers on Encounter Groups* (1970) was a major seller, and his and Farson’s filmed encounter group, ‘Journey into Self,’ won an Academy Award (Oscar) for best full-length feature documentary.

Recognizing the ever-widening applicability of the client-centered, student-centered, group-centered approach, Rogers increasingly used a broader term, ‘person-centered,’ to describe his work. In the 1970s and 1980s he experimented with a person-centered approach to resolving intergroup and international conflict. Through workshops and filmed encounter groups involving different races, generations, and political persuasions—including Catholic and Irish protagonists from Northern Ireland—Rogers demonstrated how the same growth-promoting conditions useful in all helping relationships can enhance communication and understanding among antagonistic

groups. He and colleagues led person-centered workshops for groups of 100 to 800 participants in Brazil, Mexico, South Africa, Hungary, Soviet Union, other newly emerging democracies, and other countries. They organized a gathering of international leaders in Austria, about resolving tensions in Central America, which vividly demonstrated the potential of the person-centered approach for resolving international conflict. Testimonials suggested these efforts in professional development and citizen diplomacy helped foster democratization in several countries.

8. Humanistic Psychology

Rogers was a leader in what Abraham Maslow described as a ‘third force’ in psychology, which became prominent in the latter half of the twentieth century. ‘Humanistic psychology,’ as it came to be known, differed from psychoanalysis and behaviorism in at least three ways. (See *Freud, Sigmund (1856–1939)*; *Psychoanalysis: Overview*; *Skinner, Burrhus Frederick (1904–90)*; *Behaviorism*; *Allport, Gordon W (1897–1967)*.) First, this psychology gave more emphasis and credence to the individual’s *phenomenal field*—e.g., the client-centered therapist empathizes with the client’s frame of reference rather than evaluating or diagnosing from the outside; the existential psychotherapist helps the patient find ‘meaning’ in his life. (See *Lewin, Kurt (1890–1947)*.) Second, this psychology focused not just on remediation of psychological problems, but on psychological *health, creativity, self-actualization*—what Rogers described as ‘the fully functioning person.’ The goal was more than ‘adjustment,’ rather helping people experience their full human potential. Third, it was a psychology interested in what distinguishes human beings from other species. Choice, will, freedom, values, feelings, goals, and other *humanistic* concerns were all central subjects of study.

As Rogers’ career and that of leading behavioral psychologist B. F. Skinner were parallel in timing, productivity, and influence, their views were inevitably contrasted. Meeting on several occasions, including a six-hour debate–dialogue in 1962 (Kirschenbaum and Henderson 1989a), one of their exchanges on *Some Issues Concerning the Control of Human Behavior* (Rogers and Skinner 1956) became one of the most reprinted articles in the behavioral sciences. Rogers became a leading spokesperson for the humanistic psychology movement.

9. Enduring Influence

Major factors in Rogers’ impact on psychology and the helping professions were his personal example, longevity, and prolific writings. He presented a vivid role model of the person-centered approach for almost

six decades, demonstrating his theories through teaching, lecturing, live demonstrations and workshops, and audio-visual recordings. By all accounts, he embodied his theories by being an exceptional listener and communicator and a decent, honorable person. He wrote 17 books and over 200 professional articles and research studies. Millions of copies of his books have been printed, in over 60 foreign language editions. His most popular book *On Becoming a Person* (1961), written for both a professional and general audience, continues to spread his ideas.

Critics of Rogers' work have argued (sometimes fairly, sometimes not) that client-centered therapy is superficial, unworkable with some populations, and unmindful of recent advances in behavioral, drug, or alternative therapies; that Rogers' views on human nature are unrealistically optimistic and underestimate human evil; that encounter groups and humanistic psychology have fostered widespread selfishness, narcissism, and moral permissiveness; and that Rogers' experiments with organizational change were naive and counter-productive. Nevertheless, in the USA, although no longer a central figure in popular psychology, Rogers' influence endures. In 1972, he was awarded the American Psychological Association's Distinguished Professional Contribution Award, becoming the first psychologist to receive the organization's highest scientific and professional honor. The citation read:

His commitment to the whole person has been an example which has guided the practice of psychology in the schools, in industry and throughout the community. By devising, practicing, evaluating and teaching a method of psychotherapy and counseling which reaches to the very roots of human potentiality and individuality, he has caused all psychotherapists to reexamine their procedures in a new light. Innovator in personality research, pioneer in the encounter movement, and respected gadfly of organized psychology, he has made a lasting impression on the profession of psychology.

After Rogers' death in 1987, the greatest new interest in his work has been outside the USA, including Russia, Eastern Europe, and Latin America. As a Japanese counselor explained in the 1960s, Rogers helped 'teach me ... to be democratic and not authoritative.' His life's work demonstrated how supportive, growth-producing conditions can unleash healing, responsible self-direction, and creativity in individuals and groups in all walks of life. As countries strive to resolve intergroup tensions and practice self-government and self-determination, many have recognized in Rogers' work not only useful methods for helping professionals, but a positive, person-centered, democratic philosophy consistent with their national aspirations. Rogers eventually recognized the political implications of his theories and methods and explored these in *Carl Rogers on Personal Power* (1977). At Rogers' memorial service, Richard Farson eulogized him as 'a quiet revolutionary.'

10. Personal Life

Upon college graduation, Rogers married childhood friend Helen Elliott (1901–79), an aspiring commercial artist. Their two children, David (1926–94) and Natalie (1928–), distinguished themselves professionally, David in medicine (e.g., Dean of Medicine, Johns Hopkins; President, Robert Wood Johnson Foundation) and Natalie in art therapy and 'client-centered expressive therapy.' For many years the Rogers spent the academic winter quarter in the Caribbean and Mexico, where Carl read widely and developed a number of his theories, Helen painted, and both snorkeled. They regularly traveled to visit their children and six grandchildren. Carl pursued life-long hobbies of photography, making mobiles, and gardening. When Helen became ill in her seventies, Carl cared for her until her death in 1979. Thereafter he remained involved in work, cultivated old and new friendships with both men and women, and traveled widely. He was active until his death at age 85, on February 4, 1987 in La Jolla, California.

See also: Counseling Psychology; Group Psychotherapy, Clinical Psychology of; Interpersonal Psychotherapy; Person-centered Psychotherapy; Psychology: Overview; Psychotherapy, History of; Psychiatric Aspects; Psychotherapy Process Research

Bibliography

- Hart J T, Tomlinson T M 1970 *New Directions in Client-Centered Therapy*. Houghton Mifflin, Boston
- Kirschenbaum H 1979 *On Becoming Carl Rogers* [a biography]. Delacorte Press, New York
- Kirschenbaum H, Henderson V L (eds.) 1989a *Carl Rogers: Dialogues*. Houghton Mifflin, Boston [Conversations with Buber, Tillich, Skinner, Polanyi, Bateson, May and Niebuhr]
- Kirschenbaum H, Henderson V L (eds.) 1989b *Carl Rogers Reader*. Houghton Mifflin, Boston
- Lietaer G, Rombauts J, Van Balen R (eds.) 1990 *Client-centered and experiential psychotherapy in the nineties*. Leuven University Press, Louvain, Belgium
- Rogers C R 1931 *Measuring Personality Adjustment in Children Nine to Thirteen Years of Age*. Ph.D. thesis
- Rogers C R 1939 *Clinical Treatment of the Problem Child*. Houghton Mifflin, Boston
- Rogers C R 1942 *Counseling and Psychotherapy: Newer Concepts in Practice*. Houghton Mifflin, Boston
- Rogers C R 1951 *Client-Centered Therapy*. Houghton Mifflin, Boston
- Rogers C R 1959 A theory of therapy, personality, and interpersonal relationships as developed in the client-centered framework. In: Koch S (ed.) *Psychology: A Study of a Science, Vol III*. McGraw-Hill, New York
- Rogers C R 1961 *On Becoming a Person*. Houghton Mifflin, Boston
- Rogers C R 1969 *Freedom to Learn: A View of What Education Might Become*, 1st edn. C. E. Merrill, Columbus, OH
- Rogers C R 1970 *Carl Rogers on Encounter Groups*, 1st edn. Harper & Row, New York

- Rogers C R 1977 *Carl Rogers on Personal Power*. Delacorte Press, New York
- Rogers C R, Dymond R F 1954 *Psychotherapy and Personality Change*. University of Chicago Press, Chicago
- Rogers C R, Skinner B F 1956 Some issues concerning the control of human behavior. *Science* **124**(3231): 1057–66

H. Kirschenbaum

Rokkan, Stein (1921–79)

Stein Rokkan was one of the world's leading social scientists after World War II. He played a unique role in organizing international social science cooperation in the 1960s and 1970s, and was the driving force in the development of infrastructure services for comparative research. But he was also a prolific writer, and it was the combination of all his contributions to comparative research, international cooperation, and infrastructure development that made him the comparativist of his time.

1. Stein Rokkan, the Norwegian

Rokkan was born July 4, 1921 in Vågan on the Lofoten Islands, on the northern periphery of Norway, and grew up in Narvik. Between 1938 and 1948 he attended the University of Oslo, where he first studied philology and then political philosophy, in which he received his doctorate with a dissertation on David Hume. From 1948 to 1951 Rokkan continued his studies in the USA and England: at Columbia University, where his work with Paul Lazarsfeld acquainted him with the methods of modern social research, and at the London School of Economics, where he met T. H. Marshall, whose historical-sociological research profoundly influenced his own.

Rokkan's educational career already contained the seeds of a rare combination of highly diverse talents: above all his great openness towards the 'new' social sciences while maintaining a firm grounding in the 'old' cultural sciences. His extensive linguistic, philosophical, and historical studies gave him a keen eye for the significance of cultural differences in international comparison, and thus for the often misleading simplicity of purely quantitative approaches. At the same time he recognized the potential of such approaches in comparative research and became one of their most important supporters. Rokkan's thought was interdisciplinary from early on. His scholarly interests included the entire spectrum of the social sciences, and this characterized his core scientific work: a political sociology in the tradition of classic macrosociology, the historical-comparative analysis of whole societies and polities.

From the very beginning, Rokkan combined a genuine international orientation with a deep devotion to his homeland that shaped his comparative approach. From his intimate knowledge of Norway he developed concepts (e.g., the center–periphery dichotomy) and questions (e.g., on the interaction of mass movements and institutional change) that he then applied to comparative research. This made him aware of the complexity of each single 'case' and led him to assess realistically the possibilities and limits of international comparisons.

Rokkan developed his system of thought, which was rooted in his origin and education, systematically throughout his entire scholarly career until his early death on July 22, 1979 in Bergen. From 1951 to 1958 he worked at the Institute of Social Research in Oslo, where he soon became engaged in various comparative research projects. In the same period he initiated, together with Henry Valen, the Norwegian Program of Electoral Research which was oriented on the work of the Survey Research Center in Ann Arbor, Michigan, but already showed Rokkan's signature in its emphasis on context variables.

In 1958 Rokkan became research professor at the Christian Michelsen Institute in Bergen, where in 1966 he was awarded a university chair and founded a new department of comparative politics. During this time he was also engaged in efforts to increase cooperation between the Scandinavian countries, which resulted in the publication of *Scandinavian Political Studies* since 1966, and in the founding of the Nordic Political Science Association in 1975.

2. Rokkan and the Internationalization of the Social Sciences

Rokkan was active at the international level from the very beginning of his professional career. Developing international research programs, organizing international projects and conferences, and building institutions and infrastructures for comparative research became the focus of his activities from the early 1960s. Three contexts of these activities can be distinguished: UNESCO and the International Social Science Council, the International Committee on Political Sociology, and the European Consortium for Political Research.

Rokkan, then Secretary-General of the International Sociological Association, participated in the founding of the International Social Science Council (ISSC) by UNESCO in 1952. In 1961, he proposed to the ISSC a broad program in cross-national and cross-cultural research which he then directed for over 15 years. He became a member of the Executive Committee of the ISSC in 1965 and served as president of the ISSC from 1973 to 1977.

The 1961 program led to a broad spectrum of activities to develop infrastructure services for com-

parative research and to promote concerted interdisciplinary research. Whole series of conferences were organized on comparative survey research and aggregate statistical comparisons of nations in the first half of the 1960s. Some of the further work was channeled into the European Coordination Centre for Research and Documentation in Social Sciences, set up in Vienna in 1963; other activities were headed up by two ISSC standing committees for data archives and comparative research established in 1966–7. Rokkan remained the driving force behind most of the subsequent activities, which in the 1970s also included summer schools and workbooks for training on comparative analysis and the promotion of an International Federation of Data Organization.

In the late 1960s, with the support of UNESCO Rokkan launched another series of conferences on national developments, asking what conclusions could be drawn from the early European experiences of state formation and nation-building for developmental processes outside Europe. Rokkan here developed his basic ideas about a macromodel of European political development. This work was also linked to the International Committee on Political Sociology that Rokkan, together with Shmuel Eisenstadt, Morris Janowitz, and Martin Lipset, had founded in 1960 and for which he served as Secretary from 1960 to 1970. Originally affiliated only with the International Sociological Association, in which Rokkan was vice-president from 1966 to 1970, it became a joint ISA/IPSA committee in 1970 when Rokkan was elected president of the International Political Science Association. During the 1960s the committee was an exclusive group of international scholars with a common interest in studying major historical changes. It was in this second context that Rokkan developed his ideas on democratization, cleavage structures, and party systems.

The third context was the European Consortium for Political Research. Rokkan was one of its founders and its first Chair from 1970 to 1976. Here too he pursued his central goals of developing joint research programs and improving infrastructure in the broadest sense: from bibliographic tools, research documentation, and data archives to computer applications (including cartography), and the training of the next generation of social scientists.

Rokkan's decisive role in the internationalization of the social sciences and the promotion of comparative research resulted from a congruence of personality and historical situation. In the 1960s the social sciences were expanding rapidly in Western Europe, but they suffered from national fragmentation. International cooperation was needed, and succeeded first via the USA. Only later did a genuine 'Europeanization' begin. Rokkan's broad education and international orientation proved invaluable in building bridges between the USA and Europe and reconciling different European traditions. His enormous talent for com-

binning personal networking and institution-building was a crucial asset in this start-up phase and the close linkage he saw between programmatic conceptualization, cooperative organization, and infrastructure-building gave this endeavor a cohesiveness that seems to be lacking today.

Rokkan's historical role has somewhat overshadowed his scientific contributions, apart from a few famous single hypotheses and typologies. His lasting contribution will be his attempt to develop a macro-model of the political development of Europe. In particular, his comparative studies on the structuring of European mass democracies, and his analyses of the processes of state formation and nation-building in Europe, but also his general approach to comparative research.

3. *The Structuring of Mass Politics in Western Europe*

Rokkan's work on the democratization of the European nation-states concentrated on the institutional and organizational variations in the structuring of mass politics up to World War II. Here he developed his well-known ideas of four institutional thresholds of democratization and of historical cleavage structures which in the process of democratization are transformed into party systems via the formation of alliances and oppositions.

In institutional terms, democratization for Rokkan meant the successive lowering or removal of four thresholds: the threshold to political opposition (legitimation), to political participation (incorporation), to access to parliament (representation), and to participation in government (executive power).

The lowering of the first two thresholds, through the recognition of civic rights and the extension of suffrage, ushered in the age of mass politics with mass parties and mass election campaigns. For Rokkan, an unbroken tradition of representation, a history of continuous center building, and a Protestant nationalization of territorial culture were the main factors favoring an earlier, more gradual, and stable democratization.

The lowering of the third threshold implied a transition from majoritarian to proportional representation, easing the access of minorities to representation in the legislature. In his view, the pressures for proportional representation were higher in the ethnically and/or religiously heterogeneous countries, and in general in the smaller nations with their greater international dependence.

The introduction of universal (male) suffrage in all Western European countries at the end of World War I and the transition to proportional representation in most of them tended to stabilize the party systems of the 1920s for a long time, in most cases until the 1960s (Rokkan's famous 'freezing' thesis). In Rokkan's view,

the party systems of the 1920s reflected older cleavages that had emerged in the earlier processes of state formation and nation-building and became politicized and organized in the process of democratization.

Cleavages are understood as strong and enduring conflicts that tend to polarize a political system. Rokkan distinguished between four or five basic cleavages that emerged at different critical junctures in European history. The *National Revolution* (1789 and after) produced two cleavages, with even older roots in the early state formation processes and the Reformation:

(a) the *center–periphery cleavage*: conflicts between the central nation-building culture and ethnically linguistically distinct subject populations in the peripheries;

(b) the *state–church cleavage*: conflicts between the aspirations of the centralizing, standardizing, and mobilizing nation-state and the corporate claims of the church.

Two further cleavages were produced by the *Industrial Revolution* in the nineteenth century:

(c) the *land–industry cleavage*: conflicts between the landed interests and the rising class of commercial/industrial entrepreneurs; and

(d) the *worker–employer cleavage*: conflicts between owners/employers and tenants/workers.

The class cleavage was overlaid by a fifth cleavage between proletarian internationalism and nation-accepting socialism as a result of the *International Revolution* (1917 and after).

According to Rokkan, the differences between the European party systems largely were produced by the first three cleavages, reflected in the presence or absence of regional, Christian, or agrarian parties. Whereas Christian parties emerged mainly in the religiously mixed and some of the Catholic countries, agrarian parties are found only in Protestant areas. The class cleavage, by contrast, produced working-class parties in every country, but their strength and unity differed greatly due to the socialist–communist split.

4. State Formation and Nation-building in Europe

In analyzing the historical development of the different cleavage structures, Rokkan took a decisive step beyond traditional comparative research: he no longer regarded the units of comparison, the nation-states, as given, but rather thematized their development itself. In these later works, Rokkan was influenced greatly by Hirschman's concepts of exit, voice, and loyalty, which he used in his analyses of the development of the nation-state in terms of boundary-building, internal structuring, and system-building.

Rokkan distinguished between four basic processes of loyalty-producing system-building: state formation and nation-building; two processes mainly of

boundary-building and two processes of mainly restructuring, the establishing of political citizenship by equalizing the rights to participation and the establishing of social citizenship by redistributing resources and benefits. In the older countries these processes tended to follow each other in time, whereas in the newer countries the problems associated with them often cumulated.

In the European context, according to Rokkan, these processes were influenced by six givens: first, the heritage of the Roman Empire (law and citizenship); second, the tradition of the Germanic kingdoms (legislative/judicial assemblies); third, the crucial role of the supraterritorial Catholic Church; fourth, the growth of a central city belt; fifth, the development of feudal and manorial agrarian structures; and sixth, the emergence of literatures in vernacular languages.

The European world of states and nations emerged from the disintegration of the Western Roman Empire, after centuries of territorial fragmentation and retrenchment, followed by processes of expansion and reorganization. For Rokkan, 'the alphabet and the city decided the fate of Europe' (Rokkan 1999) in these processes. As a result of various waves of migration, a new 'ethnic-linguistic infrastructure' arose in the Middle Ages. The alphabet made it possible to turn spoken into written standards of communication, and this vernacularization provided a basis for the later nation-building.

The city belt in turn was a decisive factor in center formation and early state building. The revival of trade between the Orient, the Mediterranean, and the North Sea had produced a belt of independent cities from Italy to Flanders. The polycephality of the city-state Europe either delayed state formation (Germany, Italy) or led to confederations (The Netherlands, Switzerland). By contrast, monocephalic city structures conducive to centralized state development formed at the western and later also at the eastern end of the city belt. The rise of Atlantic capitalism, however, brought with it a European division of labor that strengthened the Western states, while the Eastern states largely became incorporated into empires.

The variations in city structure and thus the tripartite division of Europe into a Western Europe of unitary, centralized nation-states, a federalist Central Europe, and an Eastern Europe of multi-ethnic empires constitute the 'west–east axis' in Rokkan's well-known typological–topological model or conceptual map of Europe. It helps to explain the great paradox of European development: that the strongest and most durable systems emerged at the periphery of the city belt and the old Empire.

While the 'west–east axis' defines differences in economic conditions of state formation, the 'south–north axis' denotes differences in cultural conditions of nation building, depending on the geopolitical distance northward from Rome. Rokkan interpreted the Reformation as the first major step in defin-

ing territorial nations in Northern Europe, whereas in Catholic Europe the Church remained supra-territorial.

The conditions for successful center-building and territorial consolidation as well as for cultural standardization and identity-building, therefore, varied across Europe in a typical and topological way. The easier these processes of territorial and cultural boundary-building, Rokkan generalized, the more gradual and less violent were the internal restructuring of the nation-states through the establishment of citizenship rights.

Towards the end of his life Rokkan attempted to integrate his comparative studies into a comprehensive model of Europe spanning the entire history of state formation, nation-building, and mass politics in Europe and identifying the crucial variables in the long and complex process that led up to the current constellations of territories, economies, and systems of political alignment. This macro-model remained a sketch exemplifying the nature of most of Rokkan's studies, which were essentially research programs.

5. Rokkan's Comparative Approach

Rokkan defined the decisive challenge of cross-cultural, cross-societal, and cross-national research as follows: 'We propose to direct our attention to one single, central set of issues: the possibilities of translating "grand theory" into empirically workable "typologies of macro-settings" for variations in human behaviour, and the consequences of such typologies for decisions on the cultural and the geographical range of comparisons at the level of communities, households, and individuals' (Rokkan 1970b).

From this fundamental orientation, Rokkan developed a series of basic elements that characterize his comparative approach: the need for context-specific comparisons, the multi-dimensionality of contexts, the twin concepts of structuring and boundary-building, the dynamics of freezing and unfreezing at critical junctures, and finally the method of retrospective diachronics.

For Rokkan model-building always had to be context-specific: 'No single explanatory or intervening variable can be linked up with a dependent variable in isolation from the context, whether across systems or across stages' (Rokkan 1999). Because such macro-contexts are comparable only to a certain degree, Rokkan was skeptical of universalizing comparisons and instead preferred region- and period-specific models, particularly in his cleavage model and his conceptual map of Europe.

The macrocontexts are in principle multidimensional, since for Rokkan they always resulted from interactions between three different systems, the political, economic, and cultural systems. Due to their

relative autonomy, he started in his macromodel from configurations, rather than hierarchies of factors: 'There is no economic determinism in the model, nor a geopolitical, nor a cultural: in this sense it seeks to combine the traditions of Karl Marx with those of Max Weber and Emile Durkheim.'

While all three systems have a territorial dimension, only the political system is ultimately capable of controlling territory. Territory is a crucial concept in Rokkan's comparative analysis of political systems. It is a starting point for his most basic twin concepts of structuring and boundary-building in the formation of territorial systems. Such systems essentially are structured through the formation of political, economic, and cultural centers and the incorporation of peripheries. Their consolidation presupposes a process of political, economic, and cultural boundary-building which in turn contributes to the emergence of territory-wide cleavages. Center-periphery relations and cleavage structures are the two dimensions that for Rokkan underlay the institutional and organizational structuring of political systems.

With his twin concepts, Rokkan went beyond traditional comparative research that was limited to comparing structures and focused on the formation of systems as a process of structuring and boundary-building. In this historical process critical junctures are crucial. This concept of Rokkan's refers to political, economic, and cultural 'revolutions,' historical events and processes which he used to distinguish between periods: between 'critical junctures' when basic new decisions are taken, structures are 'unfrozen' and boundaries are 'opened,' and periods in which the institutional and organizational structures remain 'frozen' and the boundaries 'closed.' This attempt at periodization underlies Rokkan's basic method of retrospective diachronic analysis which addresses the question: 'given an observed contrast in the values of variables at time t_i , what combinations of variables for earlier phases t_{i-1} , t_{i-2} , and so on, can best account for these differences?' (Rokkan 1999).

In this way, Rokkan attempted to explain systematically current structural differences between the European societies and polities as arising from their long-term development. In many cases, his efforts to do so took the form of research programs. With his contributions to developing a social science infrastructure, he wished to create a basis for carrying out these programs. In both respects, Rokkan has left us an important legacy.

See also: Comparative History; Comparative Studies; Method and Design; Data Archives: International; Human-Environment Relationship; Comparative Case Studies; International Communication: History; International Research: Programs and Databases; International Science: Organizations and Associations; Science and Technology: Internationalization

Bibliography

- Lipset S, Rokkan S (eds.) 1967 *Party Systems and Voter Alignments: Cross-national Perspectives*. Free Press, New York
- Rokkan S 1970a *Citizens, Elections, Parties. Approaches to the Comparative Study of the Processes of Development*. Universitetsforlaget, Oslo, Norway
- Rokkan S 1970b Cross-cultural, cross-societal and cross-national research. In: *UNESCO, Main Trends of Research in the Social and Human Sciences*. Mouton, Paris, p. 1
- Rokkan S 1999 *State Formation, Nation-building, and Mass Politics in Europe. The Theory of Stein Rokkan*. Oxford University Press, Oxford, UK
- Rokkan S et al (eds.) 1987 *Centre-Periphery Structures in Europe: An ISSC Workbook in Comparative Analysis*. Campus, Frankfurt, Germany
- Rokkan S, Dogan M (eds.) 1969 *Quantitative Ecological Analysis in the Social Sciences*. MIT Press, Cambridge, MA
- Rokkan S, Eisenstadt S (eds.) 1973 *Building States and Nations*, 2 Vols. Sage, Beverly Hills, CA
- Rokkan S, Merritt R (eds.) 1966 *Comparing Nations. The Use of Quantitative Data in Cross-national Research*. Yale University Press, New Haven, CT
- Saelen K 1981 Stein Rokkan: A bibliography. In: Torsvik P (ed.) *Mobilization, Center-Periphery Structures and Nation-building*. Universitetsforlaget, Bergen, Norway

P. Flora

Romanticism: Impact on Social Thought

Romanticism was a broad intellectual movement within modern Enlightenment that aimed at bringing the arts together with philosophy, history, and religion and attempted to aestheticize life. The central objective was a philosophico-political reconciliation of the fundamental connectedness of life, which was seen as damaged in the present, with the individual, with love, and with beauty. Although it was a wide European movement, it was in particular German romanticism that significantly influenced art and literature as well as the philosophy, psychology, and history of modernity. The key participants in the circles of Berlin and Jena as well as of Heidelberg, as the most important groups, were August Wilhelm Schlegel (1767–1845), Dorothea Caroline Michaelis/Schlegel/Schelling (1763–1809), Friedrich Schlegel (1772–1829), Brendel Mendelssohn/Dorothea Veit/Dorothea Schlegel (1763–1839), Friedrich von Hardenberg (Novalis, 1772–1801), Friedrich Schleiermacher (1768–1834), Rahel Antonie Levin Markus/Varnhagen von Ense (1771–1833), Henriette Herz (1764–1847), Karoline Günderode (1780–1806), Sophie Schubart/Merlau/Brentano (1770–1806), Clemens Brentano (1778–1842), Bettina Brentano/von Arnim (1785–1859), Achim von Arnim (1781–1831), Friedrich Wilhelm Joseph Schelling (1775–1854), Johann Gottlieb Fichte (1762–1814), in the broader environment Friedrich

Hölderlin (1770–1843); as well as Johann Joseph Görres (1776–1848), Jacob Grimm (1785–1863), and Wilhelm Karl Grimm (1786–1859).

The impact of romanticism on the emerging social sciences can be demonstrated across the following areas: (a) knowledge and hermeneutics; (b) culture and society; (c) critique of modernity; and (d) reason and aesthetics.

1. Knowledge and Hermeneutics

Whereas Enlightenment thought starts out from the assumption of the intelligibility of the world and lets its knowledge of the world predetermine its view of the arts, romanticism turns this relation around and—with reference to Kant's critique of reason—denies philosophy's primacy with regard to knowledge. Science (*Wissenschaft*) was to be 'beautiful'; analogies and metaphors were to enrich and complement scientific-scholarly practices; and the 'novel of life,' in which all arts and sciences unite, should unfold from an ever-open, progressive concept of form. 'The absolute,' 'freedom,' and 'nature' become key concepts of philosophical debate. Within such core of romanticist perspectives, however, important differences and contradictions co-exist. Friedrich Schleiermacher, for insistence, turns against Friedrich Schlegel's progressive universal poetics (*progressive Universalpoesie*) and insists on the boundaries and relative separateness of the diverse aspects of reality (religion, science, art, aesthetics). In addition, the new mythologization appears to him as an unwarranted intervention of human action into the absolute. He contradicts Schlegel's characterization of infinity with its inclination towards dissolving ethics, since under conditions of achieved finite freedom the autonomy of the other needs to be respected. Even though he rejects any ahistorical contractual thinking, he thus combines an ethics of obligation with an ethics of virtue.

Within this framework, the task of hermeneutics according to Schleiermacher (see *Hermeneutics, History of*) emerges from the following movement of thought. If thinking occurs in intersubjective individuality, then there needs to be an art of understanding, which compares that which is one's own critically with that which is of the other. Thus, such art can approach the infinite spirit that is always already at work in everything individual. Schleiermacher's hermeneutics—which can well be seen as part of an aesthetics, of a 'theory of artistic activity' (Schleiermacher 1984, p. 154)—investigates writing and speech under the following aspects: with regard to general grammar, in psychological-subjective perspective with regard to the genesis of authorship and work, by comparative procedure and, finally, by divinatory procedure. Understanding occurs in a historical process that encompasses both reader and author. Although understanding aims at agreement, this process can nevertheless not be closed (see Frank

1977). Not least Walter Benjamin (1991) has demonstrated that it is only the reader as the extended author who completes the work of art. Thus, he showed that the romantic way of 'raising reflection in the work to a higher power' contained a 'more modern' theory of art than can be found in Goethe. The impact of hermeneutics continued through Wilhelm Dilthey to Hans-Georg Gadamer (Ferraris 1998).

Beyond hermeneutics, theology, the study of religion, classic studies, and philology, psychology and psychoanalysis show signs of a romantic heritage. Early romanticism showed a particular interest in phenomena at the margins of reason, such as magnetism, occultism, somnambulism, and telepathy (Johann Wilhelm Ritter 1776–1810, founder of electrical chemistry). Late romanticism—Gotthilf Heinrich Schubert (1780–1860)—was concerned with the analysis of the human unconscious and systematic studies in the symbolism of dreams as a linguistic system (*Symbolik des Traums*, 1814), in particular, can be considered as predecessors to Freud and his cultural criticism (*Das Unbehagen in der Kultur*, 1912). Themes of the split I (the motif of the *Doppelgänger*), trauma, loss of identity, melancholy, and delusion as well as, in particular, the imagery world of sentiments in, for example, the works of Johann Ludwig Tieck, Jean Paul, Heinrich von Kleist, or E. T. A. Hoffmann were well known to Freud.

2. Culture and Society

There is a clear impact of romanticist thinking on Jacob Grimm and Wilhelm Grimm (1785–1863 and 1778–1859), whose historically and empirically comparative research on language opened the space for the identity-constituting role of the emerging *Germanistik* (German studies) and at the same time already separated the natural sciences from the human sciences.

Romanticist impact, though, is particularly evident in that problem area that is currently discussed under the denomination of 'identity' (Frieze 2001). Johann Gottfried Herder (1744–1803; *Über den Ursprung der Sprache*, 1772; *Ideen zur Philosophie der Geschichte der Menschheit*, 1784–1791), whose work connects Rousseau to romanticism, already meant what is now called 'collective identity' when he spoke of the inalienable 'character' (*Geburtsstamina*) of a people. Following up on Herder and relating to motifs in Friedrich W. Schelling—motifs that found their continuation in the theory of archetypes as developed by C. G. Jung and Mircea Eliade, Aby Warburg but also in Ruth Benedict's work—the search was for universal and constant features of culture that found regionally diverse expressions. These concepts have in common that they postulate a universalism in terms of the laws of the human psyche on the one hand, while working with a particularism in terms of the variety of different 'identities' of peoples on the other. The assumption of

'collectively shared' representations was also one of the conceptual pillars of the 'psychology of peoples' (*Völkerpsychologie*) and of the emerging sociology. It was Wilhelm Wundt who provided the most comprehensive view of psychology and the cultural sciences. His late writings (*Grundrisse der Psychologie*, 1898, *Völkerpsychologie*, 1917) focused on those psychic phenomena that are significant for the social life of human beings and that are at the roots of community and of the emergence of creations of the collective spirit. The reference to Herder's 'spirit of the people' (*Volksggeist*), which describes the supra-individual features of the members of a people, is as evident as the one to Wilhelm von Humboldt whose comparative linguistic investigations entailed a comparison of 'people's characters.' The impact of Wundt—whose writings were known by Franz Boas, Bronislaw Malinowski, but also by Emile Durkheim—extended as far as Lucien Lévy-Bruhl.

Next to the emphasis on linguistic and psychical factors for the formation of identities, the distinction between *Gemeinschaft* (mostly translated as 'community') and *Gesellschaft* ('society'; see Schleiermacher, *Theorie des geselligen Betragens*, 1799) as well as the particular relations between society and state (Fichte, *Beiträge zur Berichtigung der Urtheile des Publikums über die französische Revolution*, 1793), that were established in romanticism and reaching from republicanism to a thinking in terms of natural rights to a skepticism towards the state, lived on in the emerging social sciences (see *Community/Society: History of the Concept; State and Society*). Even though they were considerably altered, these conceptual oppositions had an echo in virtually all dualisms between 'traditional community' and 'modern society' that were established by 'classical' sociology and maintained from Max Weber, Georg Simmel, Emile Durkheim onwards to the present day.

3. A Modern Critique of Modernity

The crisis of the modern subject was made a central literary topic in romanticism (and was continued, *inter alia*, in Robert Musil or Franz Kafka). Friendship and love (Schlegel 1988) were counterposed to modern loneliness, loss of self, and the break of the individual with the icy machinery of society (as presented in Ludwig Tieck or E. T. A. Hoffmann). At the same time, forms of sociality were theorized as well as practiced in which an unadapted, anticonventional subjectivity mediated between the romantic individualism and the freedom of the subject on the one hand and the exchange with others on the other. Such forms of sociality were considered as the ideal of education (*Bildung*) in a comprehensive humanistic sense. Romanticism thus created an intellectual space in which women in particular, who were excluded from the university as an institution, occupied a central place.

The crisis of the modern subject, the romantic-enlightened cultural criticism and its dream of the 'noble savage' were in the background of the emerging field of anthropology (Georg Friedrich Creuzer, 1771–1858) and in the 'imaginary ethnography' of the nineteenth century (Kramer 1977, Said 1979). It continues to be effective as the escapism into the exoticism of the other as well as in the current constructions of (cultural) identity and 'difference' in the era of 'globalization.' The modern critique of modernity, as prefigured by romanticism, is taken up in a variety of quite different strands of thought. Karl Marx states the human alienation under conditions of capitalism; Friedrich Nietzsche celebrates the 'great separation' against both Christendom and bourgeois morality; Max Weber condemns the 'iron cage', Martin Heidegger the being 'fallen to the they' (*Verfallenheit im Man*); Theodor W. Adorno denounces the administered world and the culture industry and Walter Benjamin laments the 'loss of aura'. The (social) philosophy of the early twentieth century reclaims 'authenticity' and 'emancipation' and all these perspectives share in the critique of the Modern world as developed in romanticism and they point to a—indeed often denied—romantic heritage (in which, in particular, the connection between aesthetic experience and renewed religion, characteristic of early romanticism, is visible; see *Eudaemonism*). The romantic critique of modernity, however, is an integral moment of the Enlightenment, and not its subjective-irrational counterpart. This latter verdict has often been cast, and peculiar intellectual alliances can be found sharing it. It stretches from Hegel and the Left-Hegelians to Heinrich Heine's rejection of romanticism to Nietzsche and Carl Schmitt's critique of 'romanticism as subjective occasionalism' as well as to George Lukács. It finally culminates in the reproach that the modern-bureaucratic mass annihilation is a result of the romantic critique of modernity—a modernity that precisely created such horror.

4. Reason and Aesthetics

For August W. Schlegel, formulating the program of romanticism in his lectures on 'arts and literature' (*Schöne Kunst und Literatur*, 1801/04), reason is founded not just on empirical perception, and the autonomous validity of poetry is thus emphasized. Romanticism, however, never abandons the concept of reason. Even when the technical-instrumental use of reason against nature is opposed, a comprehensive concept of reason is sustained. Despite all criticism of idealist systemic thinking, romanticism moves within the path of Kantian philosophy, behind which there is no falling back, and it shares in Kant's rejection of the utilitarian grounding of morality in selfishness. Schlegel knows perfectly well the distinction between concepts of the intellect, which refer to the world of appearances, and concepts of reason, which refer to

ideas and states that since ideas can be judged only by a reason that has enlightened itself about the limits of the possibility of knowledge.

In this context, early romanticism reacted to a situation in which the Enlightenment attempts to deal with the question of religion were difficult to sustain. Kant's critique made the argument for belief on the basis of reason impossible. Lessing intervened into the enlightened criticism of the bible by publication of the Reimarus papers; the dispute over this inheritance culminates in F. H. Jacobi's reproach of Spinozism (The dispute over atheism, 1789–99). Fichte justified his 'Attempt at a critique of revelation' (*Versuch einer Kritik der Offenbarung*, 1792) with the argument that revelation constrains the moral freedom of the human being. For Schleiermacher, the bible is a historical expression in the ever-incomplete path of Christendom, in the infinite perfectability of the world.

At the same time, romanticism addresses the problem of Enlightenment after the unity of reason and the tangible world could be cast only as the telos of action in the sense of a regulative idea, which never becomes real. Kant's critique of reason had demonstrated that the claim that nature is the bearer of moral norms could not be justified. Nevertheless, human action in the process of history can provide experience as to whether reality is constituted by the actualization of regulative ideas. Such a concept of Enlightenment in which the relation between ideas and the world of appearances can be sensually experienced in the course of an infinite rapprochement has been expressed by Novalis and F. Schlegel by means of infinitesimal calculus, which lets the tendency towards the infinite be grasped within finitude.

A different path towards the unity of idea and world of appearances is provided by imagination and the arts. Fichte described productive imagination as 'the faculty that is suspended in the middle between determination and nondetermination, between the finite and the infinite' (*Grundlage der Wissenschaftslehre*, 1794, II, 360). The objective is here not reconciliation, but rather the insistence on remaining within a space of contradictions.

Fichte exerted decisive influence on Friedrich Schlegel's aesthetic conceptions such as the 'progressive universal poetics' and infinite irony as poetic reflection (see *Imagination: History of the Concept*). Such poetes can 'remain suspended in the middle on the wings of poetic reflection and free of all real or ideal interest, it can ever again produce such reflection and multiply it as in an infinite row of mirrors; [...] the romantic art of poetry is still in its becoming; it is even its authentic being that it can ever only become and can never be completed' (F. Schlegel 1988, II, pp. 114–15; *Athenäums-Fragment*, p. 116; see also *Lyceum-Fragment*, p. 108). The concept of infinite irony reassumes such a moment of infinite reflexivity, a reflexivity that is not limited by any fixation. Hegel—in the *Phenomenology of Spirit* and in the

Aesthetics—attacked such conception, just as he criticized the transfer of Fichte's philosophy into the arts.

F. Schlegel referred here to Socratic irony, as presented by Plato (see Plato, *Symposium*, 216e), which according to Quintilian (*Institution oratoria*, IX, 2,44) brings the opposite of what is said to understanding. Irony is, in a dialogical reference, a double negation. It reflects on the possibility of things being different from what they seem to be and thus opens an 'ironic space of play' (Allemann 1956, p. 10). In contrast to allegory and metaphor (see *Metaphor and its Role in Social Thought: History of the Concept*), irony demands a principled and reflective understanding of the ambiguity what is meant and what is said, of essence and appearance. Friedrich Schlegel develops irony in particular in the Lyceum and Athenaeum fragments and explains it in his essay 'On unintelligibility' (*Über die Unverständlichkeit*, 1800): 'The French Revolution, Fichte's *Wissenschaftslehre* and Goethe's *Meister* are the grandest tendencies of the age' (Fragment 216)—'I wrote this fragment with the most honest intention and almost without any irony' (237). The 'system of irony' (239)—in the plural forms of irony, from the raw to the fine, dramatic and double irony to the 'irony of irony' (239)—is turned here from a rhetorical trope into a basic element of philosophical reasoning. Irony becomes the infinite play back and forth between opposite moments, none of which can be led into univocality. Infinite irony is thus distanced from any kind of identity or mediation—as, for instance, in the sense of Hegel's dialectic. When 'idea' is characterized as 'a concept completed to the level of irony [...] the change, ever again recreating itself, between two thoughts in dispute' (Fragment 121, 115), then it is not a conceptually mediated identity that is at issue, but the persistent creation of difference. 'Irony of irony,' however, contains infinite regress, namely 'when one speaks without irony about irony, as it has just been the case' (239), thus when irony can no longer be mastered, 'when one is no longer able to step outside of irony, as it appears to have been the case in this essay about unintelligibility; [...] when irony, so to say, ironizes the poet [...] when irony has to be made against one's will [...] when irony turns wild and lets itself no longer be governed' (239–40).

Schlegel's writings on irony have been received as well as criticized by, among others, Schelling, Schleiermacher, Novalis, Karl Wilhelm Solger (1780–1819), and Søren Kierkegaard (1813–55). Even though Solger (*Vorlesungen über Ästhetik*, 1819) remains with the negativity of irony that excludes absolute mediation, irony, originally thought of as infinite, is turned by him—and by Hegel—into a dialectically conceived moment of artistic activity. Following up on both Solgers and Hegel, Kierkegaard (*Über den Begriff der Ironie mit ständiger Rücksicht auf Sokrates*, 1841) regarded the infinite irony of ro-

manticism as a subjective moment, seeing itself capable of separation from the positivity of reality. 'For irony, everything turns into nothing' (Kierkegaard 1993, p. 263)—and therefore Kierkegaard counterposes a controlled, limited irony to the infinite romantic irony.

Although the recasting of romantic irony thus appeared to have reached its end, recent attempts at rehabilitating irony refer back to this heritage, multiply broken and rewritten as it may be. Richard Rorty (1989, p. 123) attempted 'to cleanse romanticism of the last traces of German idealism'—but 'he overlooks that German romanticism has no need for him, since the inventor of theoretical irony had already taken his distance from early German idealism himself' (Bohrer 2000, pp. 11–12). Rorty (1989, p. xv) then designs a liberal society in which irony has become universal, thus reducing irony to a plain liberal principle. The idea here is that a plural society with its variety of life-projects, language-games, and contingencies can no longer be determined or mastered from an allegedly superior viewpoint.

Beyond such political thinking, the heritage of romanticism has a particular current significance in theories of language. The determined opposition between what is meant and what is said, between the authentic and the inauthentic is, once it has been separated from its foundation in transcendental philosophy, dissolved in 'infinite suspense' and deconstructed as the vertigo of infinite recombability in Paul de Man and as the emphasis on principled 'undecidability' in apparently fixed semantic orders of meaning in Jacques Derrida. The recent theoretical approaches, however, can hardly be considered a continuation of romanticism, but rather as the rewriting of romanticism, Enlightenment, and modernity as an inextricable philosophical complex, as 'the rewriting of some traits that had been claimed by modernity' (Lyotard 1988, p. 43).

See also: Culture: Contemporary Views; Enlightenment; Hermeneutics, Including Critical Theory; Kant, Immanuel (1724–1804); Knowledge, Sociology of; Modernity; Religion: Morality and Social Control

Bibliography

- Allemann B 1956 *Ironie und Dichtung*. Neske, Pfullingen, Germany
- Benjamin W 1919/1991 Der Begriff der Kunstkritik in der deutschen Romantik. In: Tiedemann R, Schweppenhäuser H (eds.) *Gesammelte Schriften*. Suhrkamp, Frankfurt am Main, Germany, Vol. I 1, pp. 7–122
- Bohrer K H (ed.) 2000 *Sprachen der Ironie—Sprachen des Ernstes*. Suhrkamp, Frankfurt am Main, Germany
- Ferraris M 1998 *L'Ermeneutica*. Roma/GLF editori, Laterza, Italy

- Frank M 1977 *Das individuelle Allgemeine. Textstrukturierung und -interpretation nach Schleiermacher*. Suhrkamp, Frankfurt am Main, Germany
- Friese H (ed.) 2001 *Identities*. Berghahn, Oxford, UK
- Kierkegaard S 1841/1993 Über den Begriff der Ironie mit ständiger Rücksicht auf Sokrates. In: Hirsch E, Gerdes H (eds.) *Gesammelte Werke*. Gütersloher Verlagshaus, Gütersloh, Germany
- Kramer F 1977 *Verkehrte Welten. Zur imaginären Ethnographie des 19. Jahrhunderts*. Syndikat, Frankfurt am Main, Germany
- Lyotard J-F 1988 Réécrire la modernité. In: Lyotard J-F (ed.) *L'Inhumain: causeries sur le temps*. Galilée, Paris, pp. 33–44
- Rorty R 1989 *Contingency, Irony and Solidarity*. Cambridge University Press, Cambridge, UK
- Said E W 1979 *Orientalism* 1st Vintage Book ed., Vintage, New York
- Schlegel F 1800/1988 Über die Unverständlichkeit. In: Behler E, Eichner H (eds.) *Kritische Schriften und Fragmente 1794–1828*. Ferdinand Schöningh, Paderborn/München, Germany, Vol. II, pp. 235–42
- Schleiermacher F 1984 Ästhetik (1819/25). In: Lehnerer Th (eds.) *Über den Begriff der Kunst (1831/32)*. Meiner, Hamburg, Germany

H. Friese

Rousseau, Jean-Jacques (1712–78)

1. Life and Work

Rousseau was born on June 28, 1712 and raised in Geneva, the center of Calvinism. Jean Calvin, not a native of Geneva, published his magisterial *Christinae Religionis Institutio*, the central work on education, in 1536. Its final version, dated 1559, justified general priesthood by way of the doctrine of predestination. Calvin dissolves the ‘two empires’ of laymen and ecclesiastical office in favor of a parish without differences, one which sees itself as an elected body (Calvin 1957). Clergymen are chosen teachers whose sermon is intended to educate without the possibility of it being associated with primacy or hierarchy (Calvin 1958). Rousseau’s life and work is, in many ways, a manner of dealing with Calvinism. The concept of *equality* determines *The Social Contract*, the theory of education, and Rousseau’s natural theology. Calvin’s doctrine of grace is translated into the predestination of nature, along with a rigorous inwardness which has the only true form of education on its side. Virtue becomes a model for society given the education of man’s nature. Calvinistic rigor fascinated Rousseau. Following the banning and public burning of his books, he dispensed with the citizenship of the Republic of Geneva in 1763, but not his membership in Calvin’s church, which he had renewed in 1754.

Rousseau’s dramatic life (Cranston 1991a, 1991b, 1997, Incoprom SA et al. 1999) is characterized by

ascent outside the elite and by being an outsider. The young Rousseau completed an apprenticeship as an engraver and in terms of his education he was largely self-taught. His constantly intermittent, restless, and wavering career ran between Geneva and Paris, later included Switzerland, and England. Rousseau himself attempted to gain control over his biography with the writing of his *Confessions* (Rousseau 1961), which was published posthumously in 1782 and have determined the image of Rousseau ever since then. In 1740, after various wanderings and low-level activities Rousseau took up a position as a private tutor in Lyon, where he wrote his first text on education (*Mémoire à Mr de Mably*: Rousseau 1969, pp. 1–32). In 1742 Rousseau went to Paris, publishing his *Dissertation sur la musique moderne* in 1743. In 1749 Rousseau wrote the articles on music for Diderot’s *Encyclopédie*. His first major literary success came a year later with the *Discours sur les sciences et les art*, which was awarded a prize by the Academy of Dijon. From 1756 on, Rousseau, living now in the *Eremitage* of Montmorency, concentrated on his three major works, the *Nouvelle Héloïse* (1761), a successful, widely read novel throughout Europe, the *Contrat social* (1762), Rousseau’s theory of society, and *Emile ou de l’éducation* (1762), the theory of ‘new education.’ The latter two books were banned directly after publication and publicly burned in Paris and Geneva.

From then on Rousseau was on the run. He was granted asylum by the Prussian governor in Neuchâtel soon after his flight from Paris. At his place of refuge in Môtiers he was subjected to attacks by such differing authors as the general procurator of Geneva, Tronchin, and his arch-enemy Voltaire (‘Sentiment des citoyens’, Voltaire 1961, pp. 715–18). At the same time Rousseau became the idol of the generation of ‘Sturm und Drang’: young *litterati* from all over Europe visited Rousseau in Môtiers and made him famous. In January 1766, his flight having taken him to various places, Rousseau made for England following an invitation from David Hume. Both had different preconceptions of each other and their relationship soon broke down. Rousseau returned to France and worked, with increasing paranoia, on his autobiography, which can be viewed as an attempt at self-therapy. His precarious financial situation was revived at short notice in May 1778, when Rousseau and Thérèse Levasseur, whom he had married in 1768 following a long-standing affair, moved to Ermenonville into an estate owned by the Marquis de Girardin, Rousseau’s final benefactor. His sudden death on July 2, 1778 ended a very unlikely literary and philosophical career which was intended to be *à rebours*—contrary to convention.

2. Education and Society

The archbishop of Paris, Christophe de Beaumont, justified the banning and burning of Rousseau’s *Emile*

primarily because of its denial of original sin. Rousseau assumes that human nature is not tainted. 'There is no original sin in the heart of man' 1969, p. 322). In his statement of defense against Beaumont, printed at the beginning of 1763, Rousseau formulated a far-reaching dualism which has dominated political and educational conceptions ever since. The two aspects of this dualism are *nature* and *society*, which are treated as contradictions. The present social order contradicts human nature in every respect, society is the alienation of nature, and this explains the vices of men and the evils of life. The assumption of original sin then is superfluous: man could live without sin if nature and society corresponded and harmonized (Rousseau 1969, 966ff.). In this respect Rousseau had a post-Augustinian concept of education in mind, and this is developed in *Emile*.

Essentially, the theory has three axes: the political difference of *homme civil* and *homme naturel*, the assumption of phases of 'natural development,' and the anthropological difference of *amour de soi* and *amour propre*. Rousseau's concept of education is constructed from these axes. This concept is peculiar *not* because it stresses 'natural education,' a term that was established in educational discourses long before Rousseau (Mercier 1961), but because of its paradoxical attempt to solve the contradiction between nature and society by way of education. To this end, Rousseau, as early as in the draft of *Emile*, the *Manuscrit Favre*, draws a distinction between two contrary types of education: that of *nature* and that of *society* (Rousseau 1969, p. 58). These refer to two ways of life—the life of natural and of civil man—both described contrary to Hobbes, who, in his *Leviathan* (1968, originally published 1651) distinguished between natural and social condition. Society, for Hobbes, tames and cultivates nature, and thus repression of nature is unavoidable. Rousseau reversed this argument. For him, the social condition is that of ongoing civil war (*bellum omnium contra omnes*), while the natural condition is considered to be of presocial peaceableness. Education for 'public business,' as suggested by Hobbes (1968, 289ff.) is completely ignored.

Rousseau justified his fundamental thesis in the second *Discours*, dated 1765, which was devoted to the evolution of inequality (Rousseau 1964, pp. 109–94). Rousseau explains *inegalité* in the form of society's corruption of human nature, which was originally in a position to live harmoniously with people's own needs. *Homme sauvage* ('savage man'), as he terms it, is considered to exist on the basis of his own strengths, while *homme civilisé* ('civilized man') develops social needs which render him dependent and weaken his nature (Rousseau 1964, 135ff.). The division of labor, the development of knowledge, and connected social differentiation (Rousseau 1964, p. 143ff.) forcefully brought about a progressive inequality among men. Leave is taken of the condition of innocence and

natural freedom (p. 152), and the invention of the *first* society irresistibly causes all those following (p. 178) for the prize of alienation of mankind from its own nature. In this respect *society* is the fall of men, not nature.

The theory of education in *Emile* picks up the evolution thesis and elaborates it with regard to the development of the child. This forces a reversal: the social condition does not follow the natural condition. Moreover, the social condition is led back to the natural condition, given that education can develop nature on the basis of its own potentials. Rousseau's risky experiment of thought goes as follows: if children are seen as the *homme sauvage*, education would have to apply to this condition entirely, provided that the social condition can be excluded for the duration of education. This assumption leads to *Emile*'s basic scenario: an *élève imaginaire*, described in greater detail merely as an 'orphan,' is exclusively educated by a 'governor' outside any society. The scene is anonymous countryside far away from corrupt cities and raw villages i.e., independent from the social condition as Rousseau viewed them (Rousseau 1969, pp. 264, 267, 279). The place in which the story is told is described simply *as au milieu des champs* ('in the middle of fields') (p. 277) without any details of the origin and personal history of the two protagonists. As such, the tale is not a novel of education but a treatise which is intended to describe the paradigm of true education. The name 'Emile' is presumably reminiscent of Plutarch (Shanks 1927), as it has no biographical meaning. Accordingly, the governor is not given a name and is thus not distinguishable. Both are paradigms, not persons.

'Natural' is *negative* education. Rousseau largely draws on the phases of human development from the *Histoire Naturelle* by Buffon and de Leclerc Comte (1989, pp. 1–114). Again it is not the schema of 'development,' formulated at the beginning of the *Manuscrit Favre* (Rousseau 1969, p. 60) that is original, but the linkage with the axes of educational theory. Rousseau understands the entire childhood as the 'age of nature,' which must take place outside the social condition. Thus the entire first education is negative. It does not consist at all of teaching virtue or truth, but in avoiding vices and errors in order to grant innocence of heart and integrity of mind (Rousseau 1969, p. 323). The concept of negative education is aimed against Locke (317ff.), whose *Some Thoughts Concerning Education* had been available in French in a translation by Pierre Coste since 1695 and had had considerable influence. Locke's central idea (1989, p. 142), that 'Children are to be treated as rational creatures,' aroused passionate contradiction in Rousseau. For him 'Reasoning with children' (Locke 1989, p. 142) was nothing other than the reversal of nature's plan. Reason (*raison*) is the end of education, but the end cannot determine the beginning: 'If children could understand reason there would be no need to educate

them' (Rousseau 1969, p. 317). Childhood must have its own manner (*manière*) of seeing, thinking, and feeling, which is not that of adults (p. 319). Locke's concept of reason is that of adults outside the world of children.

It is above all this idea of the separate world of childhood, that justifies Rousseau's fame in education—along with the paradisaical metaphor of pure education outside society and the spontaneity of children led by immediate interests (*intérêts présents*) (Rousseau 1969, p. 358). Rousseau seems to be the originator of modern education, which stresses the 'new' image of child and childhood. In this respect, however, the fact that *Emile* clearly has a theological center is overlooked. The treatise ends by giving preference to republic, not to nature, and the whole theory presents two concepts of education with differing gender forms. The center of *Emile* is the concept of 'natural religion,' justified in the *Profession de foi du Vicaire Savoyard*, a treatise within the treatise (Rousseau 1969, pp. 565–635). Rousseau takes the role of the vicar, who, against the materialism of the Parisian philosophy, announces a supreme being or an active creator who is assumed to be 'King of Earth' (Rousseau 1969, p. 382). Being in unison with creation cannot mean anything sinister: 'Where everything is good nothing can be unjust' (p. 588). The good can be seen in creation, justice emanates from the good, only a just person can live happily (p. 589), but the enthusiasm for virtue (p. 596) is an inner principle (p. 598)—that of the heart and not of reason (p. 627). Consequently, the key to the profession of faith is as follows: 'We can be men without being scholars' (p. 601).

What Rousseau called 'natural religion' (Rousseau 1969, p. 607) is a theology of the heart (Osterwalder 1995, cf. Chevallier 1994), it does not require any positive dogma. Moreover, inner enlightenment needs no teachers (*pasteurs*) (Rousseau 1969, p. 609), only the inspiration of belief in seeing God's creation. Rousseau radicalizes the Calvinistic doctrine here: the metaphor of 'heart' signals independence even from the community of belief. 'Equality' is therefore not a social construction which would require comparison with others, but a sovereignty in solitude. 'Sovereignty' or inner strength presupposes identity and distance. Rousseau justified this difference with the relation between two instances of the soul called *amour de soi* and *amour propre*. The fact that two instances are mentioned violates again a convention of literature. In French virtue literature such as La Rochefoucauld's *Maximes* (1665) or Blaise Pascal's *Pensées* (1670), 'amour propre' is the formula for egoism and self-love that has to be restrained by education. Rousseau used the formula in a new way: he took over the negative meaning of *amour propre* but added a second, positive, instance—that of *amour de soi*. The positive instance suggests that man looks only for himself and is content with the satisfaction of his own true needs. The

negative instance, *amour propre*, leads to comparison with others; man is never satisfied because no man will give preference to others more than to himself (Rousseau 1969, p. 493).

Loving and tender passions occur in *amour de soi*; hate and temper, i.e., from comparison with others, occur in *amour propre*. Thus, man is all the happier the fewer needs he has and the more he can avoid comparisons with others; in contrast, the factors which make man suffer ills and evils (*méchants*) are *too many* needs and *too many* opinions i.e., the dependency on others. Consequently, Rousseau develops the concept of *éducation solitaire* (Rousseau 1969, p. 341), which isolates the fictitious Emile throughout his childhood, totally controls learning (*la liberté bien réglée*: p. 321), and stipulates the didactic scenario on the theory of natural requirements and needs. Therefore Emile does not receive lessons, is not provided with any written works, and is excluded from all forms of cultural education. Only the *age de raison*—youth—requires formal lessons; all of childhood should apply exclusively to nature, which, of course, is only experienced in a didactic manner i.e., extremely artificial. The governor (and he alone) arranges the entire experience, and it is not by chance that *Robinson Crusoe* is the only literary evidence that Emile is given to read (Rousseau 1969, 455ff.).

The concept of *éducation solitaire* refers to the education of man (*faire un homme*: Rousseau 1969, p. 248). However, in the beginning of the fifth book Rousseau introduces 'Sophie ou la femme' i.e., he is forced to react to the difference in sex. The education of 'the woman' is given treatment as generalized as was previously the case with 'the child.' But now 'the child' is recognized as a *male* child and not as a sexless 'person' in general. Rousseau's *homme sauvage* is stated as male, namely 'strong,' 'sovereign,' and 'independent.' Sophie, the generalized woman, is educated in a manner that she can see herself as being in complementary dependence on males. The education of the sexes is therefore not equal, it is dissimilar (Rousseau 1969, 700ff.). This is subject to the precondition of sexual difference that cannot be dissolved or harmonized (p. 697). Because men want to seduce, they are dependent in the crucial moment, but women have to agree and thus can say 'yes' or 'no.' This female strength can only be compensated by virtuousness publicly. The 'mutual dependency' is therefore not symmetric: men are dependent on women because of their *désirs* (*desires*), women are dependent on men as a result of their wishes and needs (*besoins*). Accordingly, 'by the law of nature itself women are at the mercy of men' (Rousseau 1969, p. 702), not vice versa.

Rousseau's treatise about true education ends with a grand tour (Rousseau 1969, pp. 826–55) which is intended to introduce Emile to basics of government theory (p. 833) i.e., the doctrines of Rousseau's own *Contrat social*. Endeavors are made regarding 'le

dernière perfection' (p. 799): it is envisaged that the well-educated *man* now is to become a respectable *citizen*. The citizen's know-how is taught with the theory of the 'social contract' (839ff.); *Emile* has to learn what defines the status of citizens (*citoyens*) and thus the constitution of society (p. 840). The small *republic* is given preference—one which can create an ideal relationship between population and government (p. 840). Obviously the Republic of Geneva, idealized by Rousseau during his lifetime, is the paragon for this thesis. The ideal republic is rare and thus hard to find. The governor concludes his teachings on government as follows: the obligations of the citizen presuppose a true country (*patrie*), and a country is true insofar as it can be considered a model of political virtue. Those who do not find their *patries* will not face political abysm, 'car qui n'a pas une patrie a du moins un pays' (Rousseau 1969, p. 858).

The connection of education and society fails. Rousseau had two views on society: first the ideal of the *contrat social* creating a model for social order that is mainly granted by the concept of *volonté générale*—the general will behind all singular wills (Rousseau 1964, 361ff.); and second the polemical description of general decadence within existing society. In no place does society comply with the ideal; therefore the grand tour ends without result. The couple *Emile* and *Sophie*, educated differently, are not released into society as citizens but led to marriage and family (p. 867). More is not possible: the ideal society, inspired by the *volonté générale* and thus safe from the corruption of power, is not realized anywhere. The social ideal has no social place. Decadent society, on the other hand, is morally unacceptable; real society has its place but not a legitimate form. Thus, loneliness (*solitude*) is in the end stronger than sociality (Starobinski 1971), because education will not change society but can only develop man's nature. The great project of 'humanization' though natural education (as stated by Ravier 1941) fails because of its own ambitions and contradictions. 'Nature,' in the framework of Rousseau's *Emile*, is an artificial experience, learning is bound by didactic parameters, control is total, and sentiment for the children's *separate* world arises from the fiction of a good and equal nature that does not mirror any child's individuality.

3. Reception and Enduring Influence

The reception of Rousseau's work demonstrates the often dramatic and always radical conflict between convinced supporters and equally convinced adversaries. Rousseau divides his readers into the twenty-first century, not least because he wrote against reality. His ideals of antiquity, especially those of Sparta and Athens, suggest an *âge d'or* (Terasse 1970, Leduc-

Fayette 1974), which can be understood as an anticipation of the future, the renovation of the 'true society' that has been lost in history. It is therefore not by chance that Rousseau is the hero of the Jacobins; his cult was established in the French Revolution (Barney 1986) against all conservative theories that negate revolution in favor of the long-term development of society without the sentiment of decadence. Rousseau's *Social Contract* represents the new society that conservatives can only deny. The tension between freedom and equality attributed to, and paradoxically and provocatively described by, Rousseau characterizes one major part of political theory up to John Rawls's reformulation of the *Contrat social*. The same applies to the theory of 'natural education': Rousseau's *Emile* is a key source for Tolstoy and the reform movements of the nineteenth century, a central inspiration for Piaget (Oelkers 1996) and the development of child psychology, and a milestone for progressive education.

Of central importance are Rousseau's dualisms and thus his language. He stresses contradictions and paradoxes between nature and society, men and citizens, children and adults, and, not least, male and female. Even John Dewey, who rejected dualism, accepted that Rousseau—and only Rousseau—was the founder of the theory of 'natural development' (Dewey 1985, 211ff.) as opposed to artificial schooling, and was thus to be read in a dualistic manner. Educational theorists are irritated by Rousseau, who is regarded as a 'classical writer' of education (Rorty 1998) *because* of the unsolved paradoxes. Rousseau's concept of education makes no attempt at solving the paradoxes and dilemmas, but leaves them open, thus puzzling readers with keen contradictions which are right and wrong at the same time. What counts for Rousseau is not logic but the heart, and this seems to be part of his success. His radical conclusions seem to be obvious: politics within a new society following the education of 'new man' according to his own true nature. The famous ascription to 'return to nature,' which was part of the nineteenth century's reading of Rousseau and nowhere stated by him directly, was regarded as the emancipation from alienation, and thus a project of the left. But Rousseau was at the same time radical *and* conservative—stating that society should return to the golden age, education should leave schooling for true nature, man should be first man and then citizen. It is this rigor without reality that fascinated Rousseau's readers—disciples as well as opponents.

Rousseau's theory of education is provocatively puzzling: 'negative education' has no objective, and nature develops itself. However, the learning process is subject to extreme regulations (Rorty 1998, p. 248). The present should not be made a victim of the future (Rousseau 1969, p. 309), but every form of education is a deal with the future, and this is true for Rousseau's theory too: the ages of childhood and of youth are

clearly defined, as are the phases of education, so nature is not chance but necessity, aiming at the future. Children should be educated in a manner solely dependent on *things* (p. 311), yet the governor dominates the whole of Emile's education. Nature should lead the way (p. 259), but every effort possible must be made to avoid the first wrong step (p. 317). Children stand like savages outside the law and are completely natural (p. 320), but education must take everything in hand so as to exclude chance (p. 342) without being able to act solely on the basis of the necessity of nature. The first education should be that of the senses alone (p. 380), but that requires a rational plan of education which cannot simply be drawn from nature. In consequence Rousseau's educational scene is one of extreme regulation: Emile does not play, he does not develop anything of his own and is not allowed to listen to music, and his learning differs in every way from 'amusement' (p. 407). According to Rousseau, education should be 'realistic' (p. 418) and this is possible only when undertaken in an extremely artificial manner. Because society is excluded, educational rigor can rule. The basis for this is nature in the sense of self-love (*amour de soi*). Self-love is always good and always in compliance with the natural order (p. 491)—which begs the question why education does not concentrate fully on this *amour de soi*.

Rousseau wrote the counterclaim to sensualistic educational theory, which dominated the pedagogy of the eighteenth century. Children, according to Rousseau, should not be viewed as empty vessels or *tabulae rasae*, but instead as parts of nature which develop of their own accord. Children cannot simply be influenced, and education is not merely the establishment of habits and customs; moreover, it is the child's nature that limits all education. This fundamental concept is weakened by the implicit sensualism, the education of the senses (Rousseau 1969, 415ff.) that is necessary because education is inconceivable without any influence. But the provocation remains, and it defines Rousseau's standing as an educational author. Education is limited by nature, nature has nothing to do with sin, the child is innocent because of nature's original goodness, thus education can take place without any of the burdens of history and society. It is a renewal of mankind with every new child. This is the fundamental basis of Rousseau's theory, which is still read and discussed today because it provokes and stimulates educational thought without attempting to solve the puzzles. Rousseau is read today because he defines the problems, not their solutions.

See also: Democracy; Democratic Theory; Educational Philosophy: Historical Perspectives; Enlightenment; French Revolution, The; Rationalism

Bibliography

- Barney R 1986 Rousseau dans la Révolution: le personnage de Jean-Jacques et les débuts du culte révolutionnaire (1787–1791). In: Mason H T (ed.) *Studies on Voltaire and the Eighteenth Century*. The Voltaire Foundation at the Taylor Institution, Oxford, UK, Vol. 24
- Buffon G L, Leclerc Comte de 1989 *Histoire naturelle de l'homme et des animaux*. Jean de Bonnot, Paris
- Calvin J 1957 *Institution de la Religion chrétienne. Livre troisième. Édition nouvelle publiée par La Société Calviniste de France*. Labor et Fides, Geneva, Switzerland
- Calvin J 1958 *Institution de la Religion chrétienne. Livre quatrième. Édition nouvelle publiée par La Société Calviniste de France*. Labor et Fides, Geneva, Switzerland
- Chevallier M 1994 *Pierre Poiret (1646–1719). Du protestantisme à la mystique*. Labor et Fides, Geneva, Switzerland
- Cranston M 1983 *Jean-Jacques. The Early Life and Work of Jean-Jacques Rousseau 1712–1754*. University of Chicago Press, Chicago
- Cranston M 1991 *The Noble Savage. Jean-Jacques Rousseau 1754–1762*. University of Chicago Press, Chicago
- Cranston M 1997 *The Solitary Self. Jean-Jacques Rousseau in Exile and Adversity*. University of Chicago Press, Chicago
- Dewey J 1985 Essays on education and politics 1915. In: Boydston J A (ed.) *The Middle Works, 1899–1924*. Southern Illinois University Press, Carbondale, IL, Vol. 8
- Hobbes T 1968 *Leviathan* [ed. MacPherson C B]. Penguin, Harmondsworth, UK
- Incoprom S A, L'Institut de France, Cybele Productions 1999 *Le CD-Rom Jean-Jacques Rousseau*. Incoprom SA, L'Institut de France, Cybele Productions, Geneva, Switzerland
- Leduc-Fayette D 1974 *Jean-Jacques Rousseau et le mythe de l'Antiquité*. J. Vrin, Paris
- Locke J 1989 *Some Thoughts Concerning Education*. [eds. Yolton J W, Yolton J S]. Clarendon Press, Oxford
- Mercier R 1961 *L'enfant dans la société du XVIIIe siècle (before Emile)*. Ph.D. thesis, University of Paris
- Oelkers J 1996 Piaget et l'éducation nouvelle. In: Barraletta J-M, Perret-Clermont A-N (eds.) *Jean Piaget et Neuchâtel. L'apprenti et le savant*. Éditions Payot, Lausanne, Switzerland, pp. 165–76
- Osterwalder F 1995 Die pädagogischen Konzepte des Jansenismus im ausgehenden 17. Jahrhundert und ihre Begründung. Theologische Ursprünge des modernen pädagogischen Paradigmas. In: *Jahrbuch für Historische Bildungsforschung*. Juventa Verlag, Weinheim, Germany, Vol. 2, pp. 59–83
- Ravier A 1941 *L'éducation de l'homme nouveau. Essai Historique et Critique sur le Livre de l'Émile de J.-J. Rousseau*, Vols. 1 and 2. Éditions Spes, Issoudun, France
- Rorty A O 1998 Rousseau's educational experiments. In: Rorty A O (ed.) *Philosophers on Education. Historical Perspectives*. Routledge, London, pp. 238–54
- Rousseau J-J 1961 *Les Confessions. Autres textes autobiographiques*. In: Gagnebin B, Raymond M (eds.) *Oeuvres complètes*. Gallimard, Paris, Vol. 1
- Rousseau J-J 1964 *Du Contrat Social. Écrits politiques*. In: Gagnebin B, Raymond M (eds.) *Oeuvres complètes*. Gallimard, Paris, Vol. 3
- Rousseau J-J 1969 *Émile. Education—Morale—Botanique*. In: Gagnebin B, Raymond M (eds.) *Oeuvres complètes*. Gallimard, Paris, Vol. 4
- Shanks L P 1927 A possible source for Rousseau's name *Emile*. *Modern Language Notes* 17: 243–4

Starobinski J 1971 *Jean-Jacques Rousseau. Le transparence et l'obstacle*. Gallimard, Paris

Terrasse J 1970 *Jean-Jacques Rousseau et la quête de l'âge d'or*. Palais des Académies, Brussels, Belgium

Voltaire 1961 *Mélanges*. Gallimard, Paris

J. Oelkers

Rule of Law

The rule of law is commonly understood in contrast to arbitrary exercise of power, which is, above all, the evil that the rule of law is supposed to curb. The rule of law, though not always the specific verbal formulation, embodies ideals that have been central to political and constitutional discourse at least since Aristotle, who contrasted 'the rule of the law' with 'that of any individual.' In any of its many versions, the meaning of the rule of law is contested; all, however, share the idea that law can and should contribute in salutary, some say indispensable, ways to channeling, constraining, and informing—rather than merely serving—the exercise of power, particularly public power.

The antagonistic juxtaposition of rule of law and arbitrariness has led many jurists to view the rule of law as a purely negative ideal (Raz 1979), devoted to 'damage control' (Shklar 1998a, p. 5), particularly damage done by government. However that view is not universal. Several writers (e.g., Allan 1993, Dworkin 1985, Selznick 1999) offer a more affirmative, morally ambitious account of what is required to govern according to the rule of law. Negatively, constraint by clear legal rules is central. Positively, space needs to be made to include underlying principles of morality and purpose in legal systems. The differences between these views go deep, involving larger differences over the nature and the point of law. This account begins with the better known negative conception, but seeks to point beyond it.

1. Elements of the Rule of Law

1.1 Scope

Central to most conceptions of the rule of law is the conviction that there should be no privileged exemptions to it. Everyone comes within the scope of the law, even those who make it. This has a political and a social aspect.

1.1.1 Government under law. The political aspect of rule of law is that governments and public officials are subject to the existing law. Legislatures can of

course change law, but even 'sovereign' legislatures must comply with existing law so long as it remains in force, and can only change it in conformity with existing higher-order legal rules. To make this subjection effective, there must exist legal ways of challenging and forcing the government and political officials, even the very highest, to submit to the law.

There have been many polities in which the ideal of subjecting government to law was unknown and would seem outlandish. Even where the idea exists, the extent to which it is realized within a political order will vary markedly between polities and over time. It is, after all, a complex ideal even more complex to realize, not a faithful description of a state of affairs. Where the ideal is professed or understood, differences in attainment are crucial indicators of the sway of the rule of law. To the extent that the idea or the practice is lacking, so is this crucial element of the rule of law.

1.1.2 Equality before the law. The social aspect of the rule of law is the equal subjection of all citizens to the law. This, it was traditionally thought, required that law should take the form of an 'abstract rule which does not mention particular cases or individual nominated persons, but which is issued in advance to apply to all cases and all persons in the abstract' (Neumann 1986, p. 213; see also Hayek 1960, Oakeshott 1983). This is to prevent the use of law against specific groups or individuals, by making all equally liable, in advance, if they come within its terms. A corollary is that adjudicators should then weigh, impartially and impersonally, legal issues that come before them on the basis of such laws. They should be 'blind,' as the bearer of the scales of justice is classically represented, to inequalities of status, power, wealth, and other resources.

These conditions are the source of much controversy. Contemporary law associated with the welfare and regulatory activities of modern governments is frequently neither general nor abstract but targeted, detailed, and specific. A matter of continuing controversy is whether these forms of law are compatible with the rule of law (Hayek 1973–1979, Kamenka and Tay 1975, Neumann 1986, Oakeshott 1983, Unger 1976, but see Selznick 1992, 1999); if so, how, and if not, what should be sacrificed to what. A matter of even longer dispute is whether such formal equalization counts for much if it merely leaves substantive social and economic inequalities to play themselves out with more significant effect, and even if the legal arena allows or ensures that the 'haves come out ahead' (Galanter 1974). Liberals have commonly believed that the existence of social inequalities does not cancel out the worth of the rule of law, and may even make its impartiality more significant; precisely because inequalities confer advantages in the world,

they should not do so before the law. Many on the political left disagree, arguing that differences in the world render the 'blindness' of the law false or inconsequential, or both.

1.2 Character

A second element of the rule of law has to do with the character of laws. It can be expressed with minimal and deceptive simplicity: 'the law should be such that people will be able to be guided by it' (Raz 1979, p. 213). Reflection on that simple condition can generate tomes; it has certainly generated lists (see Fuller 1969, Raz 1979, Walker 1988). In order that law guide behavior, it must exist, that is, it must take the form of general requirements, typically rules, that people can consult before they act. Even where it exists, it will not guide behavior unless its requirements can be known. Therefore it must be public, understandable (by someone—it might take a lawyer), and relatively clear and determinate in its requirements. These requirements must be possible to perform. So the law cannot be internally contradictory (more accurately, the legal system must provide ways to resolve contradictions which inevitably occur). It must be prospective, for if it is retrospective people will neither know it, nor be able to rely on what they do know for fear of it being retrospectively changed. It must be relatively stable, for if it were constantly to change people would be reluctant to trust any present incarnation. Exercises of public power made unannounced, or announced in peremptory or arbitrary decrees, retrospective in effect, in ignorance or in violation of existing law, defy the rule of law.

Of course, whatever the character of the laws, the rule of law will amount to little unless laws are taken seriously, and enforced. So the law must be authoritative, that is, taken as binding by those whom it claims to bind. To make it so, it must be treated so, at least by the legal institutions that administer it, and it must be administered in ways that conform to its publicly announced terms.

1.3 Institutions

In the most influential English account of the concept in the last several centuries, Dicey (1961) took the rule of law to be a peculiarly English virtue associated with peculiarly English institutions, unavailable across the Channel but successfully implanted across the Atlantic. From what Judith Shklar aptly describes as this 'unfortunate outburst of Anglo-Saxon parochialism,' the rule of law emerges, 'both trivialized as the particular patrimony of one and only one national order, and formalized, by the insistence that only one set of inherited procedures and court practices could sustain it' (Shklar 1998b, p. 26).

As an ideal of some degree of generality, longevity, and abstraction, however, the rule of law is tied neither to any one national experience nor to any set of institutions. Rule-of-law regimes have often embodied different judgments about how to implement rule-of-law ideals, and have been realized within different legal and other histories and traditions that have influenced the particular shapes of the institutions that emerge. Institutional variety and possibility are too rich, and local traditions often too dense and particular, for it to be fruitful to identify the rule of law with just those particular institutions and arrangements which seem to work at home. Moreover, even within one society, problems change, and so do the solutions most appropriate to solve them. These differences are not automatically fatal, since the rule of law is not a recipe for detailed institutional design. It is rather a value, or an interconnected cluster of values, which might inform such design, and which might be—and have been—pursued in a variety of ways (see Selznick 1999).

This is not to deny that the ideal has been better served in some nations and by some institutions than by others. Institutional learning can and does occur. Some arrangements have been shown to work well in many contexts; others less so or only in some contexts. It is, therefore, appropriate to ponder the relationship between particular, realized, institutional forms, and rule-of-law ideals. Success does not, however, depend upon faithful mimicry (for even if whole institutional complexes could be duplicated, it is another job to graft them), and it often thrives on innovation, even misinterpretation. Montesquieu understood the English to have prospered because of the institutional separation of powers he mistakenly attributed to them. Even though this was not quite so, the Americans were so impressed with the philosopher's account that they institutionalized a strict institutional separation of powers embedded in that other American innovation in aid of the rule of law, a binding, written constitution. They have been widely copied. The details vary, but there is a rather simple and generalizable institutional insight at the bottom of this: at the very least, those who judge the legality of exercises of power should not be the same as those who exercise it; and they should be shielded from interference. This is why the separation of the judiciary from other branches of government is seen as central in rule-of-law states, even where they do not follow America in other aspects of the separation of powers. It is also why the rule of law is thought to depend on measures to secure the independence of the judiciary from political and other pressures. Further resources, often institutional traditions and conventions as much as written laws, are important in order to ensure that judicial decisions are grounded within plausible interpretations of existing laws. Institutional measures to guarantee litigants a fair hearing are also crucial. Access to legal institutions should be available to those who would make use

of them. The details can go on forever, but it is better not to amass them aimlessly, as though such means, rather than the values that inspire them, were the point. Rather, it is preferable to think to them in their particular contexts, from the stand point of the idea of the rule of law, and of the values it is thought to serve, together with close, but not slavish, attention to stories of institutional success and failure.

1.4 Legal Culture

As lawyers often forget, but social scientists never should, the rule of law depends upon more than the scope, character, and institutions of the law. It also depends upon a host of—apparently ‘soft’ but actually crucial—cultural supports, among them socialization into the values of the rule of law, at least of the professionals who have to administer it, and assumptions more or less widespread in the society that laws do and should count. Such assumptions are commonly revealed in the degree of willingness to use laws ‘on the books’ to protect and advance one’s interests, in demand for law as much as in supply as well as in levels of unreflective but routine law abidingness in a society. These crucial aspects of legal culture and practice are variable among and within societies and we know little about how to affect them in ways that might advance the rule of law. Lawyers give more thought to laws and institutions than to what might be necessary to make them count, but if no one is listening it does not matter too much what the law is saying. And who listens, why and how, and with what effects are empirical questions to which the law offers few answers, though the sociology of law might. If the answer is ‘only lawyers,’ the rule of law is not in good (see Örkény 1999).

2. The Value of the Rule of Law

Though the rule of law carries the patina of age and the danger of congratulatory cliché, what it expresses are not truisms. Neither the goal nor the means of the rule of law are self-evident goods. Many have thought that what mattered was who had power, not whether or how that power might be constrained. Indeed, if the right people had power, they should be unconstrained. Charismatic leaders, as Weber points out (1968, p. 243) are the sort who say ‘it is written ... but I say unto you ...’; often they have been happier when ‘it’ was not written at all. Nor was it, for example, a Russian tradition that the ruler, whether Tsar or Party, should be constrained, either by laws or to produce laws.

For anarchists, on the other hand, constraint was not enough. Institutions of public power had to go; this was Marx’s long-term view too. In the meantime,

like many radicals, he considered fraudulent those claims that law significantly restrained the power of the powerful within capitalist societies. In bourgeois societies, many Marxists have alleged, the rule of law is an ideological mystification which masks the truth: that the bourgeoisie rules, typically with aid of the instruments of law (see Krygier 1994). Many subsequent ‘critical’ thinkers, Marxist or not, have tended to agree on this point.

In the modern world, it is the liberal tradition that values the rule of law most highly. Liberty being central to liberalism, liberals are concerned with ways of averting threats to it and with securing it. An overarching source of such security is taken to be the rule of law. What is at stake here?

2.1 Freedom from Fear

According to Thomas Hobbes, fear of others is what drives people to agree to subjection to a political sovereign, and in turn justifies that subjection. The sovereign, by being empowered to issue and enforce laws, saves us from others, and them from us. Moreover, and therefore, it saves us and them from a major reason to fear each other. When we all are equally subject to known laws, both we and they have assurances which would otherwise be unavailable.

Many thinkers who agree with Hobbes on little else agree with him that this is a necessary and salutary work for states to do. But as John Locke observed, if we have reason to fear our neighbors who are just individuals, how should we avoid terror of that ‘mortal God,’ the state, which Hobbes called Leviathan? The solution is to arrange an institutional order which both consolidates power, and then curbs it, so preventing its arbitrary exercise; the challenge is to achieve such a solution. One very old attempt to achieve it—suggested by liberalism though not its invention—is to require rulers to operate under law, and by means of laws of a character that conforms to the rule of law.

The stakes on this view are high. According to those belonging to the tradition Judith Shklar calls the ‘liberalism of fear,’ we have a simple and stark choice,

a very basic dichotomy. The ultimate spiritual and political struggle is always between war and law. ... The institutions of judicial citizen protection may create rights, but they exist in order to avoid what Montesquieu took to be the greatest of human evils, constant fear created by the threats of violence and the actual cruelties of the holders of military power in society. The Rule of Law is the one way ruling classes have of imposing controls upon each other (Shklar 1998b, p. 25).

Reflecting on the blood-stained twentieth century, the eminent Marxist historian E P Thompson voiced similar sentiments, in the process scandalizing many

other Marxists, who traditionally had little time for law. He insisted that

the rule of law itself, the imposing of effective inhibitions upon power and the defence of the citizen from power's all-intrusive claims, seems to me to be an unqualified human good. To deny or belittle this good is, in this dangerous century when the resources and pretensions of power continue to enlarge, a desperate error (Thompson 1977, p. 266).

What Shklar, a lifelong liberal, and Thompson, a somewhat rueful Marxist or ex-Marxist, share is the insistence that the value of the rule of law lies primarily in what it shields us against. When they warn complacent beneficiaries of the rule of law to value what they have, it is by comparison with the perils that they believe to flow from its lack.

2.2 Coordination and Cooperation

Fear is not, however, the only reason for the rule of law, nor is government its only subject. We all are. Madison wrote that if we were angels we would not need any laws. Unfortunately, so the familiar argument goes, we are not, so we do. However, even angels and indeed all but the omniscient, particularly if there are a lot of them about, might benefit from the rule of law. We can all become confused and lose our way, not necessarily due to our or others' evil but merely to the superfluity of possibilities in an unordered world. All the more when that world is, as ours is, one of large and mobile societies full of strangers, where ties of kinship, locality and familiarity can only partially bind, reassure, or inform.

The predicament of a member of such a 'civilized' society, identified by Adam Smith in the eighteenth century, is that he 'stands at all times in need of the co-operation and assistance of great multitudes, while his whole life is scarce sufficient to gain the friendship of a few persons' (Smith 1981, Vol. 1, p. 26). The rule of law can provide crucial information and security, 'a basis for legitimate expectations' (Rawls 1971, p. 238), without which the 'co-operation and assistance of great multitudes' will necessarily be a more chancy affair. It enables fellow citizens to know a good deal about each other, though many of them are strangers, to coordinate their actions, and to feel some security and predictability in their dealings with them. For though it does not make everything predictable, it ties down much that would otherwise be up for grabs. It establishes fixed and knowable points in the landscape, on the basis of which the strangers who routinely interact in modern societies can do so with some security, autonomy, and ability to choose.

To the extent that a legal system approximates the ideal of rule of law, citizens have, or can obtain, in advance, clear understanding of their and others' legal obligations, and they can reasonably have faith that

the law will constrain other citizens and officials of state in ways that they can predict. Rules of the road are an example; everyone has to obey them, and everyone suffers from confusion, or a good deal worse, if they do not apply to all, are retrospective, secret, incomprehensible, contradictory, require the impossible, or are corruptly administered. If laws do not suffer from these defects, citizens share intersubjective cues as to the rights and responsibilities of people they might know little else about. These are invaluable cues (and clues) to have.

3. Justice

On the account so far, and according to many writers, the rule of law might provide certain basic conditions for a society to exist free of certain fears and informed of certain things, but it has little to do with the moral quality of what goes on there. According to one prominent interpretation, there is no special connection between the rule of law and higher values such as justice. The former merely sets out principles of legal efficacy, such as those one might address to a knife-maker: 'make it sharp'; or to a poisoner: 'make it kill quickly and quietly' (see Hart 1965, p. 1286). What it is used for is a quite separate question. Moreover, proponents of this view point out that the rule of law is perfectly compatible with iniquity in the content of laws. The Republic of South Africa, prior to 1989, is a common example in this connection. On the other hand, it is common for there to be a 'general and drastic deterioration in legality' (Fuller 1969, p. 40) when despots do their worst, and there is no surprise in that. There is nothing in the logic of despotism that inclines it to respect the rule of law, and there are many costs in doing so for which there is no despotic rationale. So at the very least one might claim a strong empirical inconsistency between horrific government acts and the rule of law.

Many writers might concede this, but still agree with Joseph Raz that 'the rule of law is a purely negative value. It is merely designed to minimize the harms to freedom and dignity which the law may cause in its pursuit of its goals however laudable these may be' (Raz 1979, p. 228). However, as we have seen, the rule of law has a horizontal function—in facilitating interaction and cooperation among citizens—which is as important as its vertical one—in relation to governments. To the extent that it performs the former role, it goes well beyond the merely negative virtue that Raz allows it.

By contrast to this negative conception, another tradition has it that the rule of law is altogether more, and more morally charged, than mere formal or procedural regularity. For Ronald Dworkin, the rule of law 'is the ideal of rule by an accurate public conception of individual rights' (Dworkin 1985,

pp. 11–12). That will often involve formal regularity, but not only, and not always. For Philip Selznick, the constraints emphasized in the formal conception are precious and need to be respected for what they secure. Yet the weight given to formality in the rule of law might generate vices of its own—particularly that of blindness to particularity—that subvert the doing of justice that the rule of law might otherwise be thought to support.

Even apart from the high price of exaggerated ‘legalism,’ legal systems are not always or only threatened by unrestrained power. There are other sources of arbitrariness in social life, and among them can be a refusal to take account of particulars to which rules are being applied. Moreover, in well established legal systems, with strong rule-of-law institutions, traditions, and legal expectations, there might well be room to vindicate larger values carried within the principles, traditions and purposes of law, in the service of a ‘higher instrumentalism’ which is not limited to, but need not necessarily subvert, a legal order whose injunctions overall are clear, predictable, and obeyed.

According to this view, formal regularity should not be the limit of one’s ambitions for the rule of law. Rather, it should be supplemented by a

thicker, more positive vision [that] speaks to more than abuse of power. It responds to values that can be realized, not merely protected, within a legal process. These include respect for the dignity, integrity, and moral equality of persons and groups. Thus understood, the rule of law enlarges horizons even as it conveys a message of restraint (Selznick 1999, p. 26).

The move from negative to positive conception, on Selznick’s view, is not merely a matter of normative expansion. It has a sociological basis and plausibility, for

we cannot really separate the negative and positive aspects of the rule of law. Indeed it would be highly unsociological to try to do so, for we would then miss the moral and institutional dynamics which create demands for justice, and which induce rulers to accept accountability (1999 p. 25).

One objection which the first, agnostic, interpretation has to the second, justice-impregnated one is that it threatens to do away with the specificity and usefulness of the concept:

[i]f the rule of law is the rule of the good law then to explain its nature is to propound a complete social philosophy. But if so the term lacks any useful function. We have no need to be converted to the rule of law just in order to discover that to believe in it is to believe that good should triumph (Raz 1979, p. 211).

Yet there is a space between a purely procedural account of the rule of law, whose only connection—if

any—with virtue is external, and one that wipes out the distinction between the rule of law and the rule of good law. Legal orders typically embody and generate certain values, both in their animating principles and in the complaints they provoke when their practices flout the values that give them legitimacy. These include values such as due process or natural justice and, in particular legal traditions, much more. The rule of law might well be argued to be incomplete to the extent that such values are not honored. Attempts to vindicate such values, often implicit in legal principles and traditions if not in rules, might be considered a service to the rule of law, even if it goes beyond purely procedural fairness and notwithstanding that it might fall short of justice more globally conceived.

4. *Tensions and Challenges within the Rule-of-law Ideal*

Like most important ideals, the rule of law contains internal tensions and can lead to conflict with commitment to other social ideals, because the ideals themselves are under strain, or because different interpretations of the same ideal, or attempts to realize different ideals, have different institutional logics. Thus, whether one thinks it an assault on the rule of law, or a vindication of a particular vision of it, the law generated by an active welfare state sits ill with some of the procedural conditions mentioned earlier. The simplest ways to deal with such tensions is to hold fast to one interpretation of one ideal and reject whatever might compromise it in another (Hayek 1973–79, Walker 1988). Another is to minimize the conflict and pretend that no price is being paid. More complex, but perhaps more realistic, is to acknowledge that the rule of law is not consistent with every value that one holds dear, and that, consequently, compromises in one or other direction might be unavoidable. We ask many things of law, as we do of life, and not all of them are consistent with each other. The rule of law is a mighty source of good in large modern polities, but not the only source, nor one which is automatically better the more you have of it. If power in a polity is unconstrained, the negative conception of the rule of law speaks to precious virtues, the absence of which can be vicious. But if power is already substantially constrained by law, the rule of law might tolerate, even on occasion require, that some space be made for wisdom, judgment, particularity, and substantive justice.

All the more is this the case, if one amplifies the lawyer’s and philosopher’s understanding, of what law does and how it does it, with what might be gleaned from other social sciences. Sociology of law does not suggest any linear connection between the rigidity and precision of laws and their real-world

reliability (see Bardach and Kagan 1982). Social causation is too complex, and the mediation and supplementation of formal by informal institutions, norms, and social practices too pervasive, for such a connection to be plausible, or for the extrapolations of legal dogmatics to carry much explanatory or predictive weight.

Yet those who sought the protection of the rule of law against abuse of power and social disorganization were onto something precious. Our experience of despotism, and of rule-less chaos, gives us good reason to take their exhortations very seriously indeed. Our experience, however, of incompetence, legalism, and arbitrary manipulation of hard-edged rules, even by non-despots, gives us reason to be suspicious of too narrow an understanding of the rule of law, or, if that understanding appeals, not to exaggerate its proper reach.

See also: Equality and Inequality: Legal Aspects; Equality of Opportunity; Equality: Philosophical Aspects; Governments; Hobbes, Thomas (1588–1679); Justice and Law; Law and Democracy; Rechtsstaat (Rule of Law: German Perspective)

Bibliography

- Allan T R S 1993 *Law, Liberty and Justice. The Legal Foundations of British Constitutionalism*. Clarendon Press, Oxford, UK
- Bardach E, Kagan R 1982 *Going by the Book*. Temple University Press, Philadelphia, PA
- Dicey A V 1961 *Introduction to the Study of the Law of the Constitution*, 10th edn. Macmillan, London
- Dworkin R 1985 Political judges and the rule of law. In: Dworkin R (ed.) *A Matter of Principle*, Harvard University Press, Cambridge, MA
- Fuller L L 1969 *The Morality of Law*, rev. edn. Yale University Press, New Haven, CT
- Galanter M 1974 Why the 'haves' come out ahead: Speculations on the limits of legal change. *Law and Society Review* 9: 95–160
- Hart H L A 1965 The morality of law (book review). *Harvard Law Review* 78: 1281–94
- Hayek F A 1960 *The Constitution of Liberty*. Routledge and Kegan Paul, London
- Hayek F A 1973–79 *Law Legislation and Liberty* (3 Vols.), University of Chicago Press, Chicago
- Kamenka E, Tay A E-S 1975 Beyond bourgeois individualism—the contemporary crisis in law and legal ideology. In: Kamenka E, Neale R S (eds.) *Feudalism, Capitalism and Beyond*. Edward Arnold, London
- Krygier M 1994 Marxism, communism, and the rule of law. In: Krygier M (ed.) *Marxism and Communism. Posthumous Reflections on Politics, Society, and Law*, Rodopi, Amsterdam
- Neumann F L 1986 *The Rule of Law*. Berg, Leamington Spa, UK

- Oakeshott M 1983 The rule of law. In: Oakeshott M (ed.) *On History and Other Essays*. Blackwell, Oxford
- Örkény S 1999 Rules of law: The complexity of legality in Hungary. In: Krygier M, Czarnota A (eds.) *The Rule of Law after Communism*. Ashgate/Dartmouth, Aldershot, UK
- Rawls J 1971 *A Theory of Justice*. Harvard University Press, Cambridge, MA
- Raz J 1979 The rule of law and its virtue. In: Raz J (ed.) *The Authority of Law*. Clarendon Press, Oxford, UK, pp. 210–29
- Selznick P 1992 *The Moral Commonwealth*. University of California Press, Berkeley, CA
- Selznick P 1999 Legal cultures and the rule of law. In: Krygier M, Czarnota A (eds.) *The Rule of Law after Communism*. Ashgate/Dartmouth, Aldershot, UK
- Shklar J 1998a The liberalism of fear. In: Shklar J (ed.) *Political Thought and Political Thinkers*. University of Chicago Press, Chicago
- Shklar J 1998b Political theory and the rule of law. In: Shklar J (ed.) *Political Thought and Political Thinkers*. University of Chicago Press, Chicago
- Smith A 1981 *An Inquiry into the Nature and Causes of the Wealth of Nations*. Liberty Fund, Indianapolis, IN
- Thompson E P 1977 *Whigs and Hunters. The Origins of the Black Acts*. Penguin, Harmondsworth
- Unger R M 1976 *Law in Modern Society*. Free Press, New York
- Walker G de Q 1988 *The Rule of Law*. Melbourne University Press, Melbourne, Australia
- Weber M 1968 *Economy and Society. An Outline of Interpretive Sociology*. Bedminster, New York

M. Krygier

Rules in the Legal Process

The ancient Romans had two words for law: *lex*, the specific rules written in statute books, and *ius*, enduring principles of justice. Contemporary 'legal positivists' view *lex*, the legal rules made by legislatures, courts, and administrative agencies, as the essence of law, and those legal rules are the focus of this article. Yet the idea that law also includes a body of principles, in some sense superior to the rules written by the current political regime, persists in many modern languages, in legal theory (Dworkin 1967, Fuller 1964), in popular consciousness (Ewick and Silbey 1998), and in the recurrent political effort to entrench 'higher law' principles in written constitutions.

To distinguish rules from principles, legal scholars emphasize the specificity and binding quality of legal rules. Thus, Lawrence Friedman (1975, p. 26) defines a legal rule as 'a proposition of law couched in general terms,' consisting of a statement of facts and a statement of the legal consequences which follow from those facts, for example, 'Any person who drives a motor vehicle through a red light is guilty of a violation and shall be fined \$50.' Social scientists, on the other hand, focus on the distinction between legal rules and customary rules or social norms. Thus Max Weber

defined legal rules as norms which have some likelihood of being enforced coercively by an officially-designated staff of legal officials (Rheinstein 1954, pp. 1vi, 14–15), a definition that emphasizes the political origin of legal rules and the contingencies surrounding their actual use.

1. Types of Legal Rules

Legal rules can be unwritten, as in preliterate societies (Hoebel 1964), or written and published, as in larger, literate political systems. Some legal rules are specific, such as those which prohibit driving over 60 miles per hour, while others, such as those that impose liability for accidents caused by driving ‘unreasonably dangerously,’ articulate vague standards. Highway speed limits are primary rules, designed to regulate ordinary citizens and businesses. Secondary rules, in contrast, guide the making and application of primary rules by government officials and judges (Hart 1961). Thus secondary rules—often stated in constitutions as well as in statutes—prescribe which officials have authority to make speed laws (and other laws) for a particular territory, who has authority to enforce them, who has authority to adjudicate disputes about their proper application, and who has authority to build the roads and install traffic lights. Other secondary rules govern the making of rules by legislatures, while ‘procedural rules’ govern the process of litigation and adjudication. In political systems that value the rule of law, one fundamental secondary rule is that legal officials must be guided in their decisions and actions by pre-existing, publicly accessible primary and secondary rules.

From a sociolegal perspective, it is useful to distinguish ‘high consensus’ legal rules (such as statutes condemning theft or breach of contract) from rules that are politically, morally, or economically controversial. Controversial rules are generally much harder to enforce than high consensus rules, and they pose more difficult moral dilemmas for officials and judges who are expected to enforce them.

2. The Political, Social, and Institutional Functions of Legal Rules

The most obvious function of legal rules is to serve justice, increasing the likelihood that different decision-makers will treat like cases alike. Political scientists, however, emphasize the political control function of legal rules (Farber and Frickey 1991): political authorities entrench their policy choices in statutes and administrative rules to increase the odds that those policies will be implemented uniformly by far-flung bureaucrats, law enforcement officials, and judges. For political authorities, legal rules also serve a broader communicative function, informing citizens

and businesses about their obligations and the decisions they might expect from government administrators, law enforcers, and judges. A tax code provides a familiar example. In consequence, legal rules often are said to have a constitutive function, shaping citizens’ conceptions about what is fair, permissible, and proper (Sarat and Kearns 1993, Edelman and Suchman 1997, pp. 503–6, Kagan and Skolnick 1993).

By imposing controls on administrative officials, police, and judges, legal rules tend to serve a useful function for the citizenry: restraining official power, at least in the sense of making it more predictable. Published legal rules provide criteria for holding officials accountable via judicial review or public criticism. For this reason, legal rules, when combined with a credible governmental commitment to adhere to them, serve yet another function: adding legitimacy to governmental and judicial authority—which helps explain their ubiquity in the governance of nation-states. To Karl Marx and to contemporary ‘critical legal theorists’ (Kennedy 1997, Kairys 1982), this legitimating function has an ominous side, shrouding repressive policies (and, one might add, nonrepressive policies) and coercive penalties in the language of neutrality and restraint of power.

In contemporary democratic countries, legal rules are not infrequently remade in response to popular political movements, litigation by public interest law firms, and electoral competition. Hence they have still another function: protecting individual rights. Partly because legal rules lend a graspable concreteness to general ideals and notions of justice, ordinary citizens can brandish them as shields and swords against government officials, businesses, and fellow citizens, both in and outside of court. This is the function referred to by the ancient Greek playwright Euripides in *The Suppliant Women* when he has Theseus praise the development of written codes of law:

Once written laws exist
People of few resources and the rich
Both have the same recourse to justice. Now
A man of means, if badly spoken of,
Will have no better standing than the weak;
And if the less is in the right, he wins
Against the great.

Many sociolegal scholars would argue that Euripides praised written legal rules too fulsomely. For as rules accumulate within a legal system, acquiring complicated subsections and qualifying clauses, they make the legal process so complex that it becomes inaccessible unless one can afford a lawyer. Rather than guaranteeing equality, therefore, contemporary legal systems tend to provide an advantage to sophisticated, better-financed litigants (Galanter 1974). Many sociolegal scholars might have Thrasymachus (the legal cynic of Plato’s *Dialogues*) respond to Theseus by saying something like this:

All written laws are parchment words

Whose import shifts in each man's mind.
Thus Justice is tossed this way and that from
Judge to judge, and advocates with clever minds
Oft bend the law to serve their wealthy friends.

After communist revolutions in Russia and Cuba, new regimes established 'people's courts' in which lay judges, freed from 'bourgeois' legal rules, were to 'do justice' based on the community's sense of fairness and social good. But the unhappy experience of 'socialist justice' (Markovits 1986, Hendley 1996, Krygier 1990) reminds us that written legal rules have an important political function: preserving judicial autonomy and blunting the risk of tyranny. Moreover, the 'peoples' courts' experience suggests that Theseus was probably more correct than Thrasymachus. Except in homogenous, slow-changing societies, an adjudicatory system based on written legal rules, however imperfect, is likely to approach predictability and even-handedness more closely than systems that eschew fixed rules.

Finally, legal rules enhance the efficiency of legal processes. When legal rules are clear and predictably applied, they become weighty 'bargaining chips' in negotiating settlements before or during litigation (Mnookin and Kornhauser 1979) and hence most disputes can be resolved without adjudication, while litigation is more costly when legal predictability declines (Priest 1993).

3. The Dilemma of Rules and Discretion

Because human activity is varied and changing, legal rule-makers suffer from relative ignorance of fact and indeterminacy of aim (Hart 1961, p. 125). Rules, therefore, are often underinclusive and often over-inclusive—that is, covering activities that are irrelevant or marginal to the primary evil that the rulemakers wished to deal with. Thus, if a motor vehicle code states 'Any person who drives a motor vehicle through a red light is guilty of a violation and shall be fined \$50,' it ostensibly compels punishment of a motorist who proceeds carefully through a malfunctioning traffic light which is stuck on red, or an ambulance driver rushing a patient to the hospital (Goffman 1971 p. 102). To apply the rule in such cases is disparaged as legalism (Friedman 1975, pp. 247–8, Nonet and Selznick 1978). Rule-based decision-making can lead to legalism because rules, in their quest for neutrality, often disregard individual litigants' special attributes and circumstances; this makes it easier for officials to apply law's coercive powers, but blunts their sensitivity to justice claims not embodied in the written rules (Noonan 1976).

Many governments seek to prevent the anger generated by legalism by granting legal officials some discretion to make exceptions. American criminal prosecutors, for example, can generally decline to prosecute violations of the law 'in the interests of justice.' Juries, by virtue of the secrecy of their

deliberations, can decline to convict when they feel justice requires it (Kadish and Kadish 1973). Another antidote for legalism is to substitute general standards for very specific rules; for example, requiring custody decisions to be based on 'the best interests' of the particular child, taking all relevant facts and considerations into account. Yet granting scores of frontline decision-makers discretion to suspend rules or use their own judgement in implementing vague standards recreates the risk that legal rules were invented to reduce; that is, decisions that vary across different individual decision-makers or erode official policies.

Most legal institutions, therefore, employ a mix of rule and discretion (Mashaw 1983). Some, such as tax collection agencies, may be instructed to emphasize rule-based decision-making, while others, such as juvenile courts or patent offices, may be allowed to rely on experienced legal decision-makers to apply general standards sensibly and consistently. Generally, an institution's legal rules are more prescriptive when lawmakers and politically potent interest groups do not trust it to make acceptable decisions. Thus American regulatory statutes, regulations, and permits often contain more prescriptive legal rules than their Western European counterparts (Rubin 1997, Kagan and Axelrad 2000, Rose-Ackerman 1995, Vogel 1986).

4. Can Legal Rules 'Decide Cases?'

Beginning with the American 'legal realists' of the 1920s and 1930s (Ackerman 1974, White 1972, Twining 1973), some scholars have contested the assumption that judges can be 'bound' by pre-existing legal rules. In political science, adherents to the 'attitudinal model' hold that each judge's own political values are the best predictors of individual decisions (Rowland and Carp 1983, Spaeth and Segal 1999). More fundamentally, legal philosophers refer to the 'indeterminacy' of legal rules, arguing that no matter how long one stares at the words of a constitutional bill of rights, a statute, or a set of judicial precedents, the words alone cannot determine which of two conflicting interpretations (and legal outcomes) in a particular case is 'correct.' Moreover, in complex legal systems, it is argued, decision-makers inevitably use their own judgment in deciding which of many rules or precedents 'applies' to a case. The implication is that all legal decisions are policy decisions, rationalized only after the fact by reference to one particular rule or rule interpretation.

Other scholars, however, argue that within any particular legal institution—a national judiciary, an administrative agency—legal professionals reduce indeterminacy by assigning mutually understood meanings to the rules they work with. They learn to do so through social interaction: in law school classes and legal argument; by explaining and defending their rule

interpretations to others; by absorbing institution-specific conventions concerning the purpose of particular rules; and interaction with senior officials or higher authorities who resolve disputes about particular cases. Conventional methods of rule application become part of the socially constructed role of 'judge,' 'regulatory inspector,' or 'police officer.' Legal officials are socialized to regard adherence to their institutions' processes or rule interpretations as obligatory, as essential to their legitimacy as power wielders and their identity as legal professionals (Fuller 1958, Dworkin 1967, Kagan 1978, pp. 88–90, Feeley and Rubin, 1998, p. 338, Kennedy 1997). Indeed, in complex, pluralistic societies, legal officials often find adherence to rules a comforting practice, relieving them of the sense of personal responsibility for decisions which disappoint claimants or impose harsh penalties. Legal officials who stray from their institution's conventions learn that their decisions can be criticized or reversed; thus even judges who wish to maximize their policy preferences must act strategically, avoiding decisions that cannot be linked to plausible interpretations of existing rules (Epstein and Knight 1998).

Conventional rule interpretations do not dictate the result in every case, or totally suppress the individual decision-maker's inclination to decide each case as the decision-maker personally thinks best, but they substantially restrain those impulses. The amount of constraint depends on the degree of homogeneity and hierarchical supervision in particular legal institutions, and on whether the legal rule in question is highly specific (as in highway speed limits) or is complex, requiring the weighing of many facts (as in a standard demanding payment of 'compensation' for accidents caused by 'negligent' driving). Yet even in the US Courts of Appeals, where politically appointed Republican and Democratic judges usually divide on party lines in the 20 percent of cases that generate dissenting opinions, in the other 80 percent judges with different political preferences agree on which legal rules and precedents are controlling (Gottschall 1986). A stark illustration of the constraining effect of 'interpretive communities' and socially constructed judicial roles is provided by Cover's (1975) account of American judges in Northern states in the 1850s; they returned escaped slaves to Southern 'slave-catchers,' notwithstanding the judges' (and their communities') moral opposition to slavery, because a recently enacted federal statute seemed so clearly to require that painful action.

5. *Styles of Rule Application*

For many judges and administrative officials, decision-making is a two-step process: the decision-maker first 'looks backwards' to pre-existing rules to discern what they seem to require, but then 'looks forward' to

assess the consequences of the rule-based decision in terms of the decision-maker's understanding of fairness or good public policy (Cardozo 1921, pp. 19–25, Wasserstrom 1961, Kagan 1978). Even if he feels the rule-based decision would not 'make sense,' he might nevertheless feel bound to follow the rule; that would be a legalistic decision-making style. In another style—unauthorized discretion—decision-makers ignore the rule and follow their own judgment (Kagan, 1978, p. 93). A third style, retreatism, is to temporize, push the parties to settle the dispute, or find other ways of avoiding hard decisions. Finally, in what has been called legal policymaking, the legal decision-maker reinterprets the pre-existing rule, adopting innovative readings or principled exceptions which support the preferred outcome (Feeley and Rubin 1998).

Which of these four modes of rule application predominate in a legal institution is influenced by the political system and the legal tradition in which it operates. Common law systems appear to allow judges more scope for creative rule interpretation than Western European judiciaries, which are more closely bound by statutory codes and legal treatises (Damaska 1986). Yet judicial policymaking does occur in civil law judiciaries, such as France's (Lasser 1995), and common law judges in Great Britain tend to feel more tightly bound by existing precedents than their counterparts in the United States—partly because American legal training encourages a more policy-oriented mode of legal reasoning, and partly because the fragmentation of lawmaking power in the United States makes judges less vulnerable to prompt 'overrides' by the legislature than are judges in parliamentary systems (Atiyah and Summers 1987, Damaska 1990).

Rule application styles also vary across legal institutions in the same legal system. Variations in local political culture produce different rule enforcement styles in different police departments (Wilson 1968), criminal courts (Levin 1972), and regulatory agencies (Hutter 1988). Variations in rule application also arise from the personality, role conceptions, and political commitments of individual decision-makers; from the skill and persistence with which particular parties or their legal advocates present their arguments; and from the likelihood that the decision will be reported in the news media or subjected to professional review and criticism. US Court of Appeals judges are less likely to decide cases on the basis of their political values when they serve on panels with judges from another political party (Cross and Tiller 1998). Regulatory agencies tend to enforce rules legalistically when they are subjected to public criticism for perceived laxity (Bardach and Kagan 1982) and to adopt a retreatist style when potential complainants remain silent or violators enjoy the backing of political authorities (Gunningham 1987). Legal officials are more likely to bend the rules when they confront individual subject face to face, as in 'streetcorner

bureaucracies' (Lipsky 1980) than when decision-making is based on paper dossiers or occurs in highly formal settings (Kagan 1978, p. 152). Overall, the way that legal rules are employed is shaped by the interaction of a legal institution's political environment, its clientele, its legal culture, and its mode of recruiting, training, and evaluating its personnel.

See also: Injustice: Legal Aspects; Justice and its Many Faces: Cultural Concerns; Justice and Law; Law: Overview; Legal Positivism; Natural Law; Procedure: Legal Aspects

Bibliography

- Ackerman B 1974 Law and the modern mind by Jerome Frank. *Daedalus* Winter: 119–30
- Atiyah P S, Summers R S 1987 *Form and Substance in Anglo-American Law: A Comparative Study of Legal Reasoning, Legal Theory, and Legal Institutions*. Clarendon Press, Oxford, UK
- Bardach E, Kagan R A 1982 *Going By The Book: The Problem of Regulatory Unreasonableness. A Twentieth Century Fund Study*. Temple University Press, Philadelphia, PA
- Cardozo B 1921 *The Nature of the Judicial Process*. Yale University Press, New Haven, CT
- Cover R 1975 *Justice Accused: Anti-Slavery and the Judicial Process*. Yale University Press, New Haven, CT
- Cross F, Tiller E 1998 Judicial partisanship and obedience to legal doctrine, whistle blowing on the Federal Courts of Appeals. *Yale Law Journal* **107**: 2155
- Damaska M 1986 *The Faces of Justice and State Authority: A Comparative Approach to the Legal Process*. Yale University Press, New Haven, CT
- Damaska M 1990 Reflections on American constitutionalism. *American Journal of Comparative Law* **38**: 421–43
- Dworkin R 1967 The model of rules. *University of Chicago Law Review* **35**: 14
- Edelman L, Suchman M 1997 The legal environments of organizations. *Annual Review of Sociology* **23**: 479–515
- Epstein L, Knight J 1998 *The Choices Justices Make*. CQ Press, Washington, DC
- Euripides 1938 *The Complete Greek Drama*. In: Oates W, O'Neill E (eds.)
- Euripides 1995 *The Suppliant Women*. [Trans. Warren R, Scully S] Oxford University Press, New York
- Ewick P, Silbey S S 1998 *The Common Place of Law: Stories from Everyday Life*. University of Chicago Press, Chicago
- Farber D, Frickey P 1991 *Law and Public Choice*. University of Chicago Press, Chicago
- Feeley M M, Rubin E L 1998 *Judicial Policy Making and the Modern State*. Cambridge University Press, New York
- Friedman L M 1975 *The Legal System: A Social Science Perspective*. Russell Sage Foundation, New York
- Fuller L 1958 Positivism and fidelity to law—a reply to Professor Hart. *Harvard Law Review* **71**: 630–72
- Fuller L 1964 *The Morality of Law*. Yale University Press, New Haven, CT
- Galanter M 1974 Why 'the haves' come out ahead: Speculations on the limits of legal change. *Law & Society Review* **9**: 95
- Goffman E 1971 *Relations in Public*. Harper and Row, New York
- Gottschall J 1986 Reagan's appointments to the US Courts of Appeals. *Judicature* June/July: 49–54
- Gunningham N 1987 Negotiated noncompliance: A case study of regulatory failure. *Law and Policy* **9**: 69–81
- Hoebel E A, 1964 *The Law of Primitive Man*. Harvard University Press, Cambridge, MA
- Hart H L A 1961 *The Concept of Law*. Clarendon Press, Oxford, UK
- Hutter B 1988 *The Reasonable Arm of the Law? The Law Enforcement Procedures of Environmental Health Officials*. Clarendon Press, Oxford, UK
- Kadish M, Kadish S 1973 *Discretion to Disobey: A Study of Lawful Departures from Legal Rules*. Stanford University Press, Stanford, CA
- Kagan R A 1978 *Regulatory Justice: Implementing a Wage–Price Freeze*. Russell Sage Foundation, New York
- Kagan R A, Axelrad L (eds.) 2000 *Regulatory Encounters: Multinational Corporations and American Adversarial Legalism*. University of California Press, Berkeley, CA
- Kagan R A, Skolnick J 1993 Banning smoking: Compliance without enforcement. In: Rabin R, Sugarman S (eds.) *Smoking Policy: Law, Politics and Culture*. Oxford University Press, Oxford, UK
- Kairys D (ed.) 1982 *The Politics of Law*. Pantheon Books, New York
- Kennedy D 1997 *A Critique of Adjudication*. Harvard University Press, Cambridge, MA
- Krygier M 1990 Marxism and the rule of law: Reflections after the collapse of communism. *Law & Social Inquiry* **15**: 633
- Lasser M de S-O-l'E 1995 Judicial (self-) portraits: Judicial discourse in the French legal system, *Yale Law Journal* **104**: 1325–409
- Levin M 1972 Urban politics and judicial behavior. *Journal of Legal Studies* **1**: 193
- Lipsky M 1980 *Street Level Bureaucracy*. Russell Sage Foundation, New York
- Markovits I 1986 Pursuing one's rights under socialism. *Stanford Law Review* **38**: 689–761
- Mashaw J 1983 *Bureaucratic Justice: Managing Social Security Disability Claims*. Yale University Press, New Haven, CT
- Mnookin R, Kornhauser L 1979 Bargaining in the shadow of the law: The case of divorce. *Yale Law Journal* **88**: 950
- Nonet P, Selznick P 1978 *Law and Society in Transition: Toward Responsive Law*. Harper Colophon, New York
- Noonan J T Jr. 1976 *Persons and Masks of the Law*. Farrar, Straus and Giroux, New York
- Priest G, 1993 Justifying the Civil Jury. In: Robert Litan (ed.) *Verdict: Assessing the Civil Jury System*. Brookings Institution, Washington, DC
- Rose-Ackerman S 1995 *Controlling Environmental Policy: The Limits of Public Law in Germany and the United States*. Yale University Press, New Haven, CT
- Rheinstein M (ed.) 1954 *Max Weber on Law in Economy and Society*. Simon and Schuster, New York
- Rowland C K, Carp R A 1983 The relative effect of maturation, time period, and appointing president on district judges' policy choices. *Political Behavior* **5**: 109
- Rubin E L 1997 Discretion and its discontents. *Chicago Kent Law Review* **72**: 1299
- Sarat A, Kearns T 1993 Beyond the Great Divide. In: Sarat A, Kearns T (eds.) *Law in Everyday Life*. University of Michigan Press, Ann Arbor, MI
- Spaeth H, Segal J 1999 *Majority Rule or Minority Will: Adherence to Precedent on the US Supreme Court*. Cambridge University Press, New York

- Twining W 1973 *Karl Llewellyn and the Realist Movement*. Weidenfeld, London
- Vogel D 1986 *National Styles of Regulation: Environmental Policy in Great Britain and the United States*. Cornell University Press, Ithaca, NY
- Wasserstrom R 1961 *The Judicial Decision*. Stanford University Press, Stanford, CA
- White G E 1972 From sociological jurisprudence to realism. *Virginia Law Review* 58: 999
- Wilson J Q 1968 *Varieties of Police Behavior*. Harvard University Press, Cambridge, MA

R. A. Kagan

Rumors and Urban Legends

Rumors may be defined as unverified or false brief reports (*rumors*, strictly speaking) or short stories (*urban legends*) with surprising content. They are passed on among a social milieu as true and current and expressing something of that group's fears and hopes.

1. Mythology and Ethics of Rumor

Ancient civilizations bequeathed us a mythological and ethical concept of rumor. In *The Aeneid* (IV, 173–197), Virgil (first century BC) personifies the rumor through the features of the goddess Fama, a monster with innumerable eyes, wings, ears, and mouths, and a swift messenger of truth and falsehood. This fabulous description gave inspiration to Shakespeare for *Henry IV* (1598), in giving *Rumor* a character of its own, its costume covered with painted tongues. Judeo-Christian tradition reproves gossip as a sin, ranking it with slander. The Talmud puts *lachen hara* ('ill tongue') on equal terms with atheism, adultery, and murder. In Matthew's Gospel (12: 33–7), it is said that people will also be judged by what comes out of their mouths, and curses those who make a baseless accusation against someone.

The philosophers of the Enlightenment apply the rationalist analysis to 'fables' (myths, legends, rumors). For Fontenelle (*De l'origine des fables* 1684), they are the result of peoples' ignorance, the strength of the imagination, the verbal transmission of narratives, and the explanatory function of the myth. For Voltaire, doubt does not only relate to supernatural legends, but also to historical narratives. In *Le Siècle de Louis XIV* (1751), he shows how a rumor could transform a minor war operation—the French army's crossing of the Rhine River in 1672—into a grandiose epic.

Rumors and legends are deemed to be the result of people's ignorance, fancy and credulity: A philosophical reprobation can be added to the moral reprobation of rumors.

2. Variety of Psychosociological Approaches to Rumors

The first scientific studies of rumors started in the 1940s in the USA, during the growth of communication and public opinion researches in social sciences, and being in the state of war, which drew the ruling classes' attention to the prominent part that rumors play.

Allport and Postman ([1947] 1965) formulated concepts that remain relevant today. First, they established that a given subject would more likely be the subject of rumors if perceived as important and ambiguous by the members of a social group. For example, the Japanese attack of Pearl Harbor in 1941 gave way to rumors on account of the importance of this first military attack, which made the Americans fear other assaults, and the ambiguity about the extent of casualties.

Allport and Postman attribute an affective function to rumors, when easing emotional pressure (fear, hostility) of individuals, and a cognitive function, when answering the need for information and explanation.

Lastly, through an experimental study (linear transmission of information by word of mouth), Allport and Postman pointed out three basic processes that lead to the gradual distortion of the message. *Leveling* is the rapid reduction of the volume of details. Rumors are inclined to become a short statement with the form 'subject + predicate': The subject is any person, animal, thing, institution, social or ethnic group; the predicate is the state or action attributed to the subject. *Sharpening* consists of selecting, and often magnifying, details. For example, numerical accentuation increases quantity; temporal accentuation brings rumors up to date; size accentuation exaggerates pronounced features (like a caricaturist who turns a small man into a dwarf). Lastly, *assimilation* is the alteration of details to fit them into the prejudices, stereotypes, and interests of the rumor transmitters. A rumor is the result of a psychological and collective process reflecting the lowest common denominator of ideas, values, concerns and feelings of the people making up a social group.

It remains necessary to analyze the rumor's degree of actuality and to study the mechanisms of distortion of facts or factual information. However, it is most important to understand how and why false news is spread and believed.

Shibutani ([1966], 1977) distinguishes among three different kinds of communication channels, in a descending order of social control and institutionalization: *formal channels* (media, official statements), *auxiliary channels* (clubs, circles), and *informal channels* (the word of mouth). At the normal rate of the communication system, information supply and demand are balanced: the main group of information is conveyed by formal channels, and just a few rumors

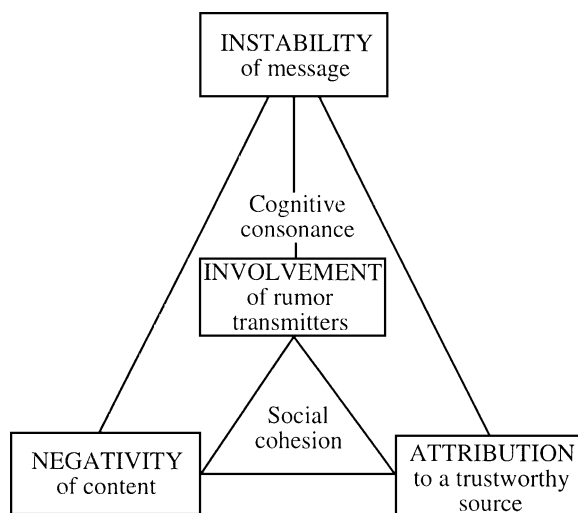


Figure 1
The rumor syndrome adapted from Rouquette (1990)

are spread by informal channels. On the contrary, in a critical rate (war, social or economic crisis, natural disaster, etc.), information demand exceeds supply, formal channels fall into discredit and informal channels abound with rumors.

Today, the distinction between official and informal channels must be moderated to the extent that rumors are also circulating by flyer, fax, e-mail, Internet, and even the mass media.

The inquiry into the 'Rumor in Orléans' by Morin (1971) is a model case study. Rather than the experimental method, he prefers the clinical approach, investigating *in vivo* the spread, then the extinction, of a rumor. Rather than take an interest in critical periods, he studies everyday life rumors, in a French provincial town. In spring 1969, it was said in Orléans that young girls vanished in dress shops owned by Jews to supply a 'white slavery' trade (i.e., prostitution). Morin tried to understand why so many people had believed these false allegations. The rumor drew its scenario from the pseudo-information of popular culture about white young girls abducted and forced into prostitution in exotic countries. Two 'catalyses' converted this scenario into 'reality.' First was a popular magazine's publishing of an unverified anecdote about the failed abduction of a young girl, found unconscious and drugged in back of a dress shop in Grenoble in France. The other was the opening of a clothes department for teenagers, named '*Aux Oubliettes*' (i.e., secret dungeon) in a medieval decorated basement of a clothing shop in Orléans. The rumor did not affect the shopkeeper, who belonged to the traditional bourgeoisie, but fixed itself on several shops owned by Jews. For Morin, the reactivation of two old myths—white slavery and the Jew—allowed

the population to symbolically express the deep fears caused by social and cultural changes at the end of the 1960s. Feminine emancipation and clothing and sexual liberalization fascinate and worry mothers and daughters: the rumor explicates the idea that miniskirts lead girls to prostitution. Second, urban modernization and growing anonymity put people in fear of losing the control of their city. The rumor expresses, through metaphors, the 'bad' city hidden behind the 'good' one: the trick dressing room, the vault under the shop, the clothes tradesman who also deals in young women, and the Jew as a figure of the one who hides his difference. One of the fundamental contributions by Morin is to have shown that a rumor is carrying a hidden message, an implicit moral, and that the study of rumors must undertake a hermeneutical work to reveal their symbolic meanings.

Morin has also suggested that, in the fight against a rumor, antirumor (denial) was less effective than counter-rumor. Thus, the rumor of Orléans died out, not so much due to police and press denials, but to an emerging counterrumor claiming that the rumor-mongers were anti-Semites and fascists.

3. A Unified Pattern: The Rumor Syndrome by Rouquette

Rouquette (1975, 1990) has built up an unified pattern of rumor in four typical features (see Fig. 1). When information or a narrative presents these characteristics, it is more than likely a rumor or an urban legend.

Instability refers to the alterations of the message, mainly at the time of its formation but also when the rumor is adapted to a new cultural environment. The presence of variants for a rumor or legend is a basic characteristic of the phenomenon.

Involvement is the fact that the rumor transmitter is feeling concerned by the message. The more rumor-mongers are involved in a rumor, the more they will believe it and pass it on. A rumor is any information or story 'worth telling.' Involvement is at its maximum when a rumor leads to behavior (e.g., panic).

Negativity is a dominating feature of rumors. Searchers estimate that 90 percent of rumors are '*noir*,' reporting news of aggression, accident, failure, scarcity, scandal, etc.; and only 10 percent are '*rose*,' containing optimistic content and attaining imaginary shared hopes and desires.

Lastly, *attribution* is the alleged source of a rumor. The rumor transmitters mention a trustworthy source, who assures the narrative's truth.

Rouquette's pattern offers a strong consistency of its features through six links.

Instability and involvement. Contrary to accepted opinion, the transmission of a message is less accurate when the subjects who spread the rumors are involved—there is an unconscious interference by interest, in their remembrance of the subject—than when they are

not concerned by the rumor. In transforming the message, the rumor transmitters tend to keep their cognitive consonance, in other words, the mental consistency of information, opinions, and beliefs they have about a situation.

Instability and negativity. The distortions of the message are commonly negative.

Instability and attribution. The rumor changes and enriches its attributions as it circulates. For example, a relay becomes a source (when rumors are passed on by the press or by a company's internal memos), a denial becomes a confirmation (in a French variant of the flyer about 'LSD tattoos,' the heading 'The Narcotic Bureau Confirms' appeared just after the organization denied the allegations. Cf. Campion-Vincent and Renard 1992).

The next three links show the role that rumors play in social cohesion.

Involvement and negativity. The discrediting of 'others' is implicitly corollary to giving importance to 'ourselves.' This sets up a mental and moral convergence between the rumor transmitter and his or her audience, therefore strengthening the social cohesion.

Involvement and attribution. Calling upon a trustworthy source weaves social ties that give value to the rumor transmitter.

Negativity and attribution. The more the rumor is negative, the stronger the guarantee of veracity.

Rouquette (1989) has also expressed the idea that rumors are multiple 'answers' to a small number of 'problems': how to account for an unexplained fact, how to vindicate an attitude or a belief, how to instigate behavior (warning rumors), how to intensify the social differentiation, or how to favor social integration. These problems take shape through themes, whether they are universal or specific of a given culture at a given time. In the Western modern world, the recurring themes of rumors and urban legends relate to just a few preoccupations (Kapferer 1990, Renard 1999): health, sex, money, new techniques, strangers, urban violence, nature, the change of manners, and the supernatural.

4. Folklore's Contribution to the Study of Rumors

Since rumors and legends are subjects of research that usually depend on distinct sciences—sociology for the former and folklore for the latter—their profound identity has been hidden for a long time. It was not until the emergence of a new concept—'modern or urban legend' (Mullen 1972, Brunvand 1981)—that sociologists and folklorists collaborated profitably and renewed their perspective on their own research topic. Mullen has emphasized the great resemblance between rumor and legend: Both are the result of a collective interactive process, they are usually first spread by word of mouth, their message is passed on as the truth,

and their social functions are similar. Rumor and legend are two aspects of the same phenomenon but one aspect sometimes outweighs the other. Studying the legendary aspect of rumors reveals their mythical deep-rootedness. Studying the 'rumoral' aspect of legends brings enlightenment to the transformations of content and the production of variants.

This aspect appears in the following tendencies: briefness of content, instability, and actualization, active and brief circulation in a certain time and place and among a given social group. The legendary aspect expresses complementary tendencies: the patterning of the content into a narrative form; the exploitation of motifs and themes, sometimes of ancient origin, emerging from the collective imagination and having strong symbolic contents; insertion into long, sometimes secular, periods of time, spreading internationally (Brunvand 1981, 1993, etc.).

A rumor may expand—become 'authentic'—and embody itself in a legendary narrative; reciprocally, a legend may be reduced to a mere statement. For instance, 'There are alligators in the sewers of New York' is the brief rumor form of the story reporting that New Yorkers brought back baby alligators from Florida, and then disposed of them when they became burdensome by flushing them down the toilet, where they ended up in the sewer (Brunvand 1981).

A similar legend may arise in both a traditional version and a modern version. For instance, the narrative motif 'Animal lives in person's body' is treated in Africa as witchcraft, whereas in the West it is explained by pseudomedical reasons.

In his pertinently headed article 'From Trolls to Turks,' Tangherlini (1995) shows many correlations between rumors and urban legends, circulating in modern Denmark, and the traditional narratives of the Danish folklore. He proves that beliefs about immigrants are revived from old motifs concerning supernatural beings (elves, trolls, witches): loathsome food customs, abduction of women or children, physical violence.

It is not accidental that interest in urban legends first arose among the folklorists, at least in English- and German-speaking countries. Rumors and contemporary popular narratives are actually modern equivalents of ancient legends, all of them expressing social or symbolic thought.

See also: Collective Beliefs: Sociological Explanation; Collective Memory, Psychology of; Panic, Sociology of; Prejudice in Society; Public Opinion: Microsociological Aspects

Bibliography

- Allport G W, Postman L [1947] 1965 *The Psychology of Rumor*. Russell and Russell, New York

- Brunvand J H 1981 *The Vanishing Hitchhiker: American Urban Legends and Their Meanings*. Norton, New York
- Brunvand J H 1993 *The Baby Train and Other Lusty Urban Legends*. Norton, New York
- Campion-Vincent V, Renard J-B 1992 *Légendes urbaines: Rumeurs d'aujourd'hui*. Payot, Paris
- Kapferer J-N 1990 *Rumors: Uses, Interpretations and Images*. Transaction Publishers, New Brunswick, NJ
- Morin E 1971 *Rumour in Orléans*. Blond, London
- Mullen P B 1972 Modern legend and rumor theory. *Journal of the Folklore Institute* 9: 95–109
- Renard J-B 1999 *Rumeurs et légendes urbaines*. Presses Universitaires de France, Paris
- Rosnow R L, Fine G A 1976 *Rumor and Gossip: The Social Psychology of Hearsay*. Elsevier, New York
- Rouquette M-L 1975 *Les Rumeurs*. Presses Universitaires de France, Paris
- Rouquette M-L 1989 La rumeur comme résolution d'un problème mal défini. *Cahiers internationaux de sociologie* 86: 117–22
- Rouquette M-L 1990 Le syndrome de rumeur. *Communications* 52: 119–23
- Shibutani T [1966] 1977 *Improvised News: A Sociological Study of Rumor*. Bobbs-Merrill, Indianapolis, IN
- Tangherlini T R 1995 From trolls to Turks: Continuity and change in Danish legend tradition. *Scandinavian Studies* 67(1): 32–62

J.-B. Renard

Rural Geography

While it is widely accepted today that those working in the rural context can make a significant contribution to the theory and practice of human geography more broadly, this has not always been the view. At times rural geography has been regarded as a poor relation of modern geography, (see Cloke 1980) trailing behind the 'real' intellectual advances in the discipline and generally presenting a largely traditional and often atheoretical perspective on key debates. Moreover, the very existence of a sub-discipline of geography defined according to a set of shifting, obscured, and contested spatial boundaries has been questioned at times, particularly in the context of geographers' interests in national, international, and global systems and processes. This review of rural geography will focus on how the sub-discipline has developed and, in particular, how its various phases have been influenced by broader theoretical and conceptual ideas within geography as a whole.

1. From Regionalism to Political Economy

The regional approaches that dominated geography before 1960 favored rural spaces and places. Descriptive studies of land use frequently focused on the

countryside, while work on population change and settlement also tended to be approached from a rural perspective. Such dominance was put down to the importance of agriculture as a land use and to the emphasis that regional approaches placed on describing identifiable physical characteristics. The demise of regional geography in the 1960s, however, shifted attention away from the countryside; new directions in human geography saw the introduction of more 'relevant' approaches in which the sheer size and visibility of urban problems ensured that they received most attention. During the next 15 years interest in rural geography declined while the work that did continue became:

entrenched in its agricultural roots and seemingly oblivious to the theoretical strides being achieved within urban studies (Cloke 1989 p. 65).

Following this rather fallow period, rural geography experienced something of a resurgence during the 1970s (particularly in Britain, Canada, and Europe). Textbooks specifically devoted to rural areas emerged (for example, Clout 1972) claiming that the urban emphasis of recent geographical research had resulted in an imbalance in terms of wider geographical inquiry and a profound absence of knowledge about rural places and people. This renewed focus on rural geography yielded a large amount of data on the countryside in developed countries, but rural studies at this time continued to follow a highly traditional pattern. Much of the research was descriptive, establishing rural areas as interesting places to study but lacking explanation and avoiding theoretical debate. Emphasis was placed on identifying the rural as different or unique (leading to rather sterile debates on the definition of 'rurality,' see Cloke 1977) rather than on showing how an understanding of patterns and processes operating in rural areas could inform human geography more broadly.

The willingness of rural geographers to engage with theory marks perhaps a key watershed in the development of the sub-discipline. During the mid 1980s those working within the rural context began to argue that rural research needed to be informed by theoretical approaches being developed outside the field and that a failure to locate studies within a broader theoretical and conceptual framework was seriously undermining their value. Although some reference was made through the 1960s and 1970s to debates taking place in radical rural sociology—especially the use of Marxist approaches in the study of agriculture—the theoretical concepts used were not generally taken up by rural geographers. It was thus only with the development of political economic approaches that the need to engage with theory outside rural geography started to influence its scope and direction to any significant degree. An awareness of the importance of processes of restructuring and capital accumulation to rural economies and societies

had begun to demonstrate to rural researchers the relevance of macroeconomic and political transformations to the detail of change within the countryside. So, considerably later than their urban counterparts, rural researchers turned to political economy approaches to inform their analyses and explanations. Space is insufficient here for a detailed review of the nature and contribution of political economic approaches to the study of rural geography (but see Cloke and Little 1990, Marsden et al. 1986, Phillips and Williams 1984, Urry 1984). It is possible, however, to identify a number of key themes that emerged (and were sometimes transformed) through the use of such approaches.

2. *Economic and Social Restructuring*

During the 1980s changes in the patterns and processes of capital accumulation were visible in the transformation of agricultural production and shifts in the location of industry. Understanding changes in rural areas required grounding such changes conceptually within more global patterns of restructuring and capital accumulation, essentially within a political economy framework. Particularly important in this context were arguments surrounding the unevenness of restructuring and the ways in which global processes of transformation impacted differently on different places at different times.

The adoption of political economy approaches provided research on agricultural geography in particular with a rich and highly fruitful arena for debate (see, in particular, Buttel and Newby 1980, Goodman and Redclift 1985, Marsden et al. 1986) as studies sought to detail and explain the major shifts taking place in the agricultural and associated industries. Research concentrated initially on the capitalization of agriculture and on the implications of new technologies on the methods of farming and on the ownership of the means of agricultural production. Later, however, the debate expanded to consider the broader transformation of food production including issues such as the role of 'downstream' and 'upstream' industries together with the social relations of agricultural production (especially changes to the family farm) (Goodman et al. 1987, Whatmore et al. 1987). Although not all work on agriculture adopted an explicitly political economic frameworks (indeed some consciously rejected political economy perspectives), the influence on this aspect of rural geography was profound, providing a much more informed understanding of the changes taking place in rural areas and of their links with global patterns of economic change.

The application of political economy approaches to another major research direction in rural geography, the restructuring and relocation of industrial development, was also reflected in increased output (see, for example, Bradley and Lowe 1984). More importantly, perhaps, it encouraged research into the distributional

impacts of shifts in the location of industrial capital, including a consideration of the effects on different sections of the labor market and the implications for the changing social structure of rural areas. Such issues were also incorporated in the second key theme to be highlighted in the discussion of political economy and rural geography, that of class and social inequality.

3. *Inequality in Rural Society*

The renewed focus on rural geography that took place in the 1970s and early 1980s, produced a number of studies from Britain and the USA on rural deprivation and 'welfare' issues, but it was only with the application of political economic approaches that such studies began to address issues of social inequality and to develop class-based analyses of the causes and effects of rural deprivation. For the first time research included not only an examination of the nature and extent of deprivation in rural areas but also the relationship between individuals and groups according to access to resources. A 'service' based model of rural deprivation in which the welfare of rural people was seen to be a function of the availability of certain essential public and private services, was gradually replaced (or in some cases augmented) by one that related deprivation to people's personal access to resources in terms of income, property, status, and power (Moseley 1982, Newby 1979). Detailed studies of access to housing, employment, transport, health care, etc., emerged through which the inequalities existing in contemporary rural societies were highlighted (for summaries see Champion and Watkins 1991, Troughton 1995). Such studies were then used in the broader examination of access to the countryside and of the transformation of rural societies and the gentrification of rural communities (see Little 1987a, Marsden et al. 1993).

4. *State-Society Relations*

Political economy approaches highlighted the centrality of power relations to the composition and well-being of the rural population. While the class-based analysis of access to rural resources contributed to an understanding of rural residents' personal power (especially financial but also in terms of status), the wider discussion of the structures and operation of power prompted researchers to explore the relationship between the state and the rural community (see Cloke and Little 1990). Importantly, this work drew on urban-based debates on the nature of the local state and on the changing mechanisms and practices of central government. At a time when the developed economies, particular the USA and Britain, were seeing attempts by the center to curtail the responsibilities of the local state, rural research began to

investigate the implications for rural areas and to document the localized power struggles and resistance that accompanied the changing balance of power.

Political economy approaches to central–local relations provided a framework for the investigation of power in a variety of ways. Studies focused on inter agency relations, the changing role of the private sector, and the role of key actors, elites, and institutions within the policymaking process. An area of particular attention, (especially within British rural geography), was that of town and country planning, and a number of studies on the negotiation and distributional consequence of planning decisions contributed to the wider understanding of power relations and political control within the regulatory process in the countryside. One particular strand of such research sought to highlight the role and status of environmental pressure groups in an examination of the priorities of the dominant interests in the countryside. Much of this work revolved around a discussion of the changing status and power of agriculture and raised questions concerning the position of farmers and landowners in relation to other interests within both urban and rural communities.

Forms of explanation that originated beyond the rural led to questions concerning the validity of ‘the rural’ as a legitimate category of analysis or even subject of research activity (see Hoggart 1990). The spatial unevenness of patterns and processes of capital accumulation and the specificity of local level responses to economic and social restructuring were used in defence of a focus on the rural, but research moved away from attributing causal properties to rural areas in the explanation of issues such as population migration, inequality, and access to resources. As well as contributing in its own right to the development of rural geography, the adoption of political economy perspectives, it may be argued, helped to open up the subject to other theoretical approaches. Particularly important here was the emergence of feminist perspectives in rural geography in the late 1980s (see Darque and Gasson 1991, Little 1986, 1987b, Whatmore 1991).

5. Feminist Rural Geography

The experience of women in the countryside were referred to fleetingly in the early development of rural geography, in, for example, the discussion of labour relations within family farm businesses (Sachs 1983, Symes and Marsden 1983) and the sociology of the rural community (see Davidoff et al. 1976). Only in the late 1980s, however, did a sustained attempt to incorporate explicitly feminist perspectives into the subject emerge. As with the adoption of political economy approaches, the introduction of feminist perspectives was a response to theoretical ideas being applied within other areas of geography and drew on

a wide range of literature from urban/social geography as well as from other academic disciplines such as sociology and women’s studies. It also reflected a need to go beyond the description of the constraints and opportunities experienced by women living in the countryside in explaining their position. Useful data on the detail of women’s lives had been collected in research on issues such as employment, access, and service provision (for a summary see Little 1991). Rural women in developed countries, for example, had lower activity rates than their urban counterparts and that their paid work was generally characterized by poor conditions and low wages. Women’s lives were also disproportionately constrained by limited transport and other public and private services. It was argued, however, that such work failed to go beyond the examination of women’s roles and that it provided little explanation for the basis of women’s inequality. Lacking alternative forms of explanation, such work was in danger of attributing ‘rurality’ with causal powers and identifying characteristics of rural economies and societies as wholly responsible for the experiences of rural women.

In keeping with work in other parts of the discipline, feminist rural geographers argued that studies of women needed to be located within a broader consideration of gender relations and their operation in the countryside. An understanding of women’s experiences in all areas from their contribution to the family farm to their access to child care, had to recognize, it was asserted, the underlying power relations that exist between men and women and needed to acknowledge the role of patriarchy as a fundamental element of social control in the countryside (see Little 1987b, Whatmore 1991). Having argued that the experiences of rural women needed to be conceptualized in terms of more universal patriarchal systems and gender relations, the research also sought to demonstrate the specificity of the experiences of rural women, showing how rural communities exhibited a particularly traditional form of gender relations and were dominated, as a result, by more conventional gender roles and attitudes towards gender equality. Women’s position was seen as a function of a particular view of rural communities, one in which the family, and women’s reproductive role within it, were central. Research also argued, in the discussion of the daily detail of women’s lives and experiences, that women’s ability to take part in paid work as well as their position within the rural community were very much shaped by this dominant understanding of the ‘appropriate’ role of women in the countryside.

Research on rural women and gender relations has continued to develop this theme and in so doing has drawn on and contributed to the ‘cultural turn’ within rural geography. More recent research has explored notions of gender identity in the context of the (re)construction of rural cultures, suggesting that certain aspects of women’s (and men’s) gender ident-

ities are prioritized within certain dominant cultural constructions of rurality within developed economies (see Hughes 1997, Agg and Phillips 1998, Little 1997). Again, such work has explored the implications of more traditional views of gender identity within rural culture for the reality of women's lives. Furthermore, some researchers have shown how dominant constructions of the relationship between gender and rurality will be experienced differently by groups of women and men; such as gay men and lesbians (see Bell and Valentine 1995, Kramer 1995) or young women (Dahlström 1996). Increasingly, work on gender identity and on sexuality has led to a discussion of the construction of femininity and masculinity within a rural context (Brandth 1995, Woodward 1998).

6. *Rural Geography and the Cultural Turn*

Recent evidence of another 'resurgence' in rural geography (documented by Cloke 1997) has accompanied the adoption of cultural approaches in geography and the social sciences generally. Cloke maintains that three important foci of cultural studies have provided a particular impetus for the study of rural spaces and societies: landscape, the rural other, and the spatiality of nature. In addition, a willingness by rural researchers to engage with important theoretical dualisms such as society/space, nature/culture, self/other has brought rural studies back into the mainstream of geographical inquiry.

A widely quoted springboard for new cultural approaches in rural geography came in the form of Chris Philo's 1992 article on 'neglected rural geographies.' In this short piece Philo challenged the implicit and explicit preoccupation of rural geographers with the experiences of 'Mr Average,' arguing that both the content and the practice of rural geography were dominated by middle-aged, middle-class, white, males. He suggested that this emphasis be redressed in the study of 'other' elements of the rural population and that in so doing researchers take on board the variety that exists in the composition and experience of the rural population. The fact that such variety has been 'hidden' from both academic and lay discourses of rural spaces reflects the dominance of particular constructions of rurality.

Following Philo's article there has indeed been a concerted attempt by rural geographers to focus on the experiences of previously neglected, and frequently marginalized groups. Studies of, for example, young people (Jones 1999, Valentine 1997), black and ethnic minority people (Snipp 1996), New Age Travellers, (Halfacree 1996) have helped to increase our understanding of the diversity of rural lifestyles. Moreover, many studies of 'other' groups within rural society have attempted to locate their findings within a theorization of difference and in so doing to draw on

critical debates in social and cultural theory. These attempts to theorize the experience of difference and diversity have demonstrated the existence of multiple ways of seeing rural spaces and drawn attention to the important power relations involved in the privileging of certain views over others. Thus, it has been argued that the experiences of groups such as the young or black people in the countryside are profoundly shaped by their (in)ability to be accommodated within dominant construction of the rural. This sense of being 'out of place' will influence their own perceptions or constructions of rural spaces and reinforce their marginalization.

An interest in the dominant, multiple, and contested constructions of rurality has led to (and drawn from) an interest in the different representations of the countryside. Studies have sought to demonstrate the various ways in which rural spaces and people have been represented in different arena, identifying the key signs and symbols that are selected to reflect rurality and exploring their cultural significance. Research has also attempted to show how these images have been used in particular situations and how they have contributed to the projection of idealized and sanitized rural spaces, masking the existence of poverty and marginalization and creating a sense of staged authenticity that differs drastically from 'reality.' Their ability to influence lived behavior at the local level in addition to some sort of wider notion of rurality has been a central concern of some studies as they investigate the circulation of symbolic meanings.

Debates surrounding the construction of rurality and the power relations bound up in the representation of rural spaces and people have raised questions about the role of human agency and the relationship between humans and non-humans in the control of the countryside. Nowhere is this more relevant than in the study of 'nature' and the direction of environmental change (see Murdoch 1997, Whatmore and Thorne 1998). While not confined to rural/environmental studies, issues concerning the nature/society dualism and the changing relationship between 'humans' and 'animals' have opened up major areas of research within rural geography from the development of food networks and the 'quality' of agricultural produce (Arce and Marsden 1995) to the role and purpose of different forms of landscape/environmental protection (Winter and Gaskell 1998).

Exploring the different conceptualizations, structures, and mechanisms of power has once again highlighted the theme of policy and governance within rural geography. As noted above, political economy approaches encouraged the examination of both formal and informal structures of power and their operation within the policy process. Recently there has been a re-examination and development of such work in relation to a broader conceptualization of governance and to the changing social, economic, and cultural context of rural policymaking (see Goodwin

1998). Research has considered, for example, the changing practices and mechanisms of governance in rural areas in the analysis of partnerships and the growth of community participation. Such work has also been informed by cultural perspectives in the examination of the values and attitudes underlying shifts in the regulatory framework.

Clearly, the cultural turn in rural geography has had a major impact on the direction and content of research within the subdiscipline. It has also had a far-reaching influence on the methods employed by rural researchers; on the way data is collected, analyzed, and used. In common with other areas of geography, there has been a significant growth in the adoption of ethnographic methods in rural research and in the use of 'grounded theory' and while such approaches may have characterized some of the early rural 'community studies,' it is only recently that they have been more widely applied to a range of economic, environmental, and policy-based issues in rural studies. Those working on rural lifestyles and marginalization have argued, in particular, that a real understanding of the lives of those living in the countryside needs to emerge from the detailed analysis of their behavior, attitudes, and beliefs and that this can only be obtained through ethnographic methods. In addition, rural researchers have started to make use of a variety of other sources of data, drawing on art, literature, diaries, and other cultural artefacts, both historical and contemporary, in their work. With a shift in the methods used by (some) rural researchers has come something of a change in attitudes towards research. The sensitivity of work on marginalization and deprivation has opened rural geographers' eyes to the potential dangers of research tourism and encouraged them to think more carefully about the way their work may be used to the advantage or disadvantage of particular groups at particular times.

7. Conclusion

This short article has highlighted some of the main 'stages' in the evolution of rural geography since the 1950s. The goal has been to identify key shifts in thought within rural geography and show how they have influenced the nature and direction of rural studies. An impression that the subject has moved smoothly from one dominant theoretical influence to another is inevitable but misleading.

Rural geography is in a buoyant period, having been theoretically invigorated by the application of perspectives from cultural and feminist studies. Rural researchers appear more open to the influence of perspectives from outside the narrow confines of the subject and the willingness to engage with and contribute to debates in other parts of academia as well as other areas of geography, has raised the profile of rural geography generally and ensured that the wider applicability of their arguments is acknowledged.

See also: Development and Urbanization; Development: Rural Development Strategies; Geography; Internal Migration (Further Types); Industrialized Countries; Internal Migration (Rural–Urban); Industrialized Countries; Peasants and Rural Societies in History (Agricultural History); Rural Planning; General; Rural Sociology; Urban Geography

Bibliography

- Agg J, Phillips M 1998 Neglected gender dimensions of rural social restructuring. In: Boyle P, Halfacree K (eds.) *Migration into Rural Areas: Theories and Issues*. Wiley, Chichester, UK
- Bell D, Valentine G 1995 Queer country: rural lesbian and gay lives. *Journal of Rural Studies* **11**: 113–22
- Bradley T, Lowe P (eds.) 1984 *Locality and Rurality*. Geo Books, Norwich, UK
- Brandth B 1995 Rural masculinity in transition: gender images in tractor advertisements. *Journal of Rural Studies* **11**: 123–33
- Buttel F, Newby H (eds.) 1980 *The Rural Sociology of the Advanced Societies: Critical Perspective*. Allenheld Osmun, Montclair, USA
- Champion T, Watkins C (eds.) 1991 *People in the Countryside*. Chapman, London
- Cloke P J 1977 An index of rurality for England and Wales. *Regional Studies* **11**: 31–46
- Cloke P J 1980 New emphases for applied rural geography. *Progress in Human Geography* **4**: 181–217
- Cloke P 1989 Rural geography and political economy. In: Peet R, Thrift N (eds.) *New Models in Geography 1*. Unwin Hyman, London
- Cloke P 1997 Country backwater to virtual village? Rural studies and 'the cultural turn.' *Journal of Rural Studies* **13**: 367–75
- Cloke P, Little J 1990 *The Rural State: Limits to planning in rural society*. Oxford University Press, Oxford, UK
- Clout H D 1972 *Rural Geography: An Introductory Survey*. Pergamon, Oxford, UK
- Dahlström M 1996 Young women in the male periphery: Experiences from the Scandinavian North. *Journal of Rural Studies* **12**: 259–71
- Darque M, Gasson R (eds.) 1991 Women in Agriculture. Special issue of *Journal of Rural Studies*. **7**: 1–137
- Davidoff L, L'Esperance J, Newby H 1976 Landscape with figures. In: Mitchell J, Oakley A (eds.) *The Rights and Wrongs of Women*. Penguin, Harmondsworth, UK
- Goodman D, Sorj B, Wilkinson J 1987 *From Farming to Biotechnology: A theory of agro-industrial development*. Blackwell, Oxford
- Goodman D, Redclift M 1985 Capitalism, petty commodity production and the farm enterprise. *Sociologia Ruralis* **15**: 231–247
- Goodwin M 1998 The governance of rural areas: some emerging research issues and agendas. *Journal of Rural Studies* **14**: 5–12
- Halfacree K H 1996 Out of place in the country: travellers and the rural idyll. *Antipode* **28**: 42–72
- Hoggart K 1990 Let's do away with rural. *Journal of Rural Studies* **6**: 245–57
- Hughes A 1997 Rurality and cultures of 'womanhood.' In: Cloke P, Little J (eds.) *Contested Countryside Cultures: Otherness, Marginalisation and Rurality*. Routledge, London
- Jones O 1999 Tomboy tales: the rural, nature and the gender of childhood. *Gender, Place and Culture* **6**: 117–36

- Kramer J 1995 Bachelor farmers and spinsters: gay and lesbian identities and communities in rural North Dakota. In: Ball D, Valentine G (eds.) *Mapping Desire*. Routledge, London
- Little J 1986 Feminist perspectives in rural geography: an introduction. *Journal of Rural Studies* 2: 1–8
- Little J 1987a Gentrification and the influence of local level planning. In: Cloke P (ed.) *Rural Planning: Policy into Action?* Harper and Row, London
- Little J 1987b Gender relations in rural areas: the importance of women's domestic role. *Journal of Rural Studies* 3: 335–42
- Little J 1991 Women and the rural labor market: a policy evaluation. In: Champion T, Watkins C (eds.) *People in the Countryside: Studies of Social Change in Rural Britain*. Paul Chapman, London
- Little J 1997 Employment, marginality and women's self-identity. In: Cloke P, Little J (eds.) *Contested Countryside Cultures: Otherness, Marginalisation and Rurality*. Routledge, London
- Marsden T, Arce A 1995 Constructing quality: emerging food networks in the rural transition. *Environment and Planning A* 27: 1261–79
- Marsden T, Munton R, Whatmore S, Little J 1986 Towards a political economy of capitalist agriculture: a British perspective. *International Journal of Urban and Regional Research* 10: 498–521
- Marsden T, Murdoch J, Lowe P, Munton R, Flynn A 1993 *Constructing the Countryside*. UCL Press, London
- Moseley M J (ed.) 1982 *Power, Planning and People in Rural East Anglia*. Centre of East Anglian Studies, Norwich, UK
- Murdoch J 1997 Inhuman/non-human/human: actor network theory and the prospects for a non-dualistic and symmetrical perspective on nature and society. *Environment and Planning D* 15: 731–56
- Newby H 1979 *Green and Pleasant Land? Social change in rural England*. Hutchinson, London
- Phillips D, Williams A 1984 *Rural Britain: a social geography*. Blackwell, Oxford, UK
- Philo C 1992 The child in the country-ward, C. *Journal of Rural Studies* 8: 193–207
- Sachs S E 1983 *Invisible Farmers. Women's Work in Agricultural Production*. Rhinehart Allenheld, Totowa, NJ
- Snipp C M 1996 Understanding race and ethnicity in rural America. *Rural Sociology* 61: 125–42
- Symes D G, Marsden T K 1983 Complementary roles and asymmetrical lives: farmer's wives in a large farm environment. *Sociologia Ruralis* 23: 229–41
- Troughton M J 1995 Rural Canada and Canadian rural geography: an appraisal. *The Canadian Geographer* 39: 290–305
- Urry J 1984 Capitalist restructuring, recomposition and the regions. In: Bradley T, Lowe P (eds.) *Locality and Rurality*. Geo Books, Norwich, UK
- Valentine G 1997 A safe place to grow up? Parenting, perceptions of children's safety and the rural idyll. *Journal of Rural Studies* 13: 137–48
- Whatmore S 1991 *Farming Women: Gender, Work and Family Enterprise*. Macmillan, Houndmills, UK
- Whatmore S 1991 Life cycle or patriarchy? Gender divisions in family farming. *Journal of Rural Studies* 7: 71–6
- Whatmore S, Thorne L 1998 Wild(er)ness: reconfiguring the geographies of wildlife. *Transactions of the Institute of British Geographers* 23: 435–54
- Whatmore S, Munton R, Marsden T, Little J 1987 Interpreting a relational typology of farm businesses in Southern England. *Sociologia Ruralis* 27(2/3): 103–22
- Winter M, Gaskell P 1998 The *Agenda 2000* debate and CAP reform in Great Britain: is the environment being sidelined? *Land Use Policy* 15: 217–231
- Woodward R 1998 'It's a man's life!': soldiers, masculinity and the countryside. *Gender, Place and Culture* 5: 277–300

J. Little

Rural Industrialization in Developing Nations and Gender

1. Expansion of Rural Industrialization

During the 1970s, transformative industries spread throughout the rural zones of many developing nations. These enterprises shared five features: they all involved industries engaged in the manufacture of basic consumer goods; the goods were produced in small- to medium-sized firms; working conditions and wages were marginal relative to those in the formal economy of state-regulated employer-worker relations; many, though not all, of the workers were women laboring in their own homes; and the women recruited for such labor tended to be young and single (Arias 1988, Wilson 1991).

Women have always participated in the domestic chores and contributed to the income of rural households, but heretofore their work was unpaid and involved activities that were labeled as 'female.' Rural industrialization changed all that and unleashed paid labor and contractual employment outside of the household yet within the rural setting (Bachrach Ehlers 1993, Boris and Daniels 1989). Until then, female contact with paid work and manufacturing employment only occurred through migration to the city.

The most common transformative industries now present in developing rural areas are, in decreasing order: the canning of fruits, vegetables, meats, and marine products; making clothing and fabrics; the manufacture of goods, candies, and electronics; and the transformation of popular crafts into marketable 'folk art.' Until recently, this manufacturing generally occurred in 'traditional' industries, which relied on the intensive use of poorly-paid labor as the preferred strategy for adapting to markets that are substantial, but fragmented and ever-changing, and to consumer demands that may be intense but ephemeral. Under these circumstances, access to a ready niche of low-cost labor has been crucial to sustaining the economic viability of these firms (Bradley 1989).

The expansion of industry into rural areas has been particularly intense in Asia (notably in China, Bangladesh, India, Indonesia, Malaysia, and Taiwan) and Latin America (especially Mexico, Paraguay, Puerto Rico, and the Dominican Republic), as well as in the

less-developed countries of southern Europe (Greece, Portugal, Turkey, and to a lesser extent Spain and Italy). Rural industrialization is more common in countries with high fertility and a large share of the population living outside of cities.

The association between rural industrialization and female employment across such a wide variety of countries suggests that, under the present global economic regime, certain industries have come upon a powerful political-economic strategy that takes advantage of and reinforces gender relations within rural environments. As a result, rural women have increasingly become one of the best options for securing a steady supply of labor that is cheap, docile, flexible, and unorganized, that is widely available in many developing settings (Dwyer and Bruce 1988, Mencher 1988).

The rapid expansion of rural industrialization and its effective rooting in many peasant societies has been interpreted from two principal perspectives: that of the labor market and that of the household. The notion of the *social construction of difference* is essential to understanding the issue that both perspectives seek to theorize: how to construct and maintain a workforce that will accept the low salaries, unstable working conditions, and discriminatory treatment of women that prevail in rural manufacturing?

The endless search for and creation of niches of cheap manual labor is made possible by what Maruani (1991) calls the 'social construction of difference,' which can be understood as a dynamic mechanism permitting the selective definition of jobs and wages along the lines of gender, race, legal status, and other demographic criteria in such a way as to invent, recreate, and conserve hierarchies that legitimate segregation, inequality, and discrimination. Through the social construction of difference, rural industrialists have been able to keep labor costs down by creating social categories of people who can be offered lower wages for work under precarious conditions (Maruani 1991).

2. The Labor Market Perspective

The creation of niches of low-cost manual labor might be explained simply as a consequence of macro-economic changes and recent policies of structural adjustment, not only in poor sectors, but in the traditional sectors that have assured the survival of rural society despite its long state of crisis. Indeed, the dyad of rural industrialization/female work appears to be strongly associated with the deterioration and stagnation of tasks characteristic of rural life: agriculture, cattle raising, and sheep herding. This association implies an irremediable crisis of the proverbially male occupations of rural society that affirmed the role of men as principal providers for their households.

In the context of local poverty and family crisis, women have been compelled to accept whatever offers of work have come along, however undesirable. This acceptance was facilitated by rural women's low level of education and information about the norms and rules of paid labor in formal urban settings. But education and information are not the only factors. Conserving female labor as a low-cost resource results from a social process of 'deskilling' of those tasks carried out by women.

3. Deskilling as a Social Construction

In rural society, it is commonly accepted that female employment occurs thanks to abilities acquired in the home, such as housekeeping, a knowledge of the kitchen, and dexterity at sewing, which together open the door to canneries and workshops that make shoes, clothes, toys, and other consumer goods. As these abilities are taken to be 'natural,' or at least learned at home, employers treat them as 'unskilled' jobs to which a low value may be assigned relative to the manufacturing tasks learned by men.

This 'cheapening' of female labor also arises from seeing the work of women in terms of 'feminine' attributes: that they are careful, attentive, fine, delicate, and sensitive. These qualities are seen as part of the innate character of woman, rather than as part of the skill set associated with paid labor and priced in the market.

Another factor in women's low wages and hours of work is their responsibility for reproductive tasks and to domestic responsibilities (England 1993). Indeed, Dwyer and Bruce (1988) have calculated that most of the part-time labor done in the world today is carried out by women. This limits the possibilities for women to approach work as a career developed over the long term. Because of domestic responsibilities, in-home paid labor, which is easier to combine with domestic tasks, has been embraced among rural women. They know that it is poorly paid, but for some it is the only combination possible for making money while not neglecting household obligations (Fernández-Kelly and García 1989).

In this way, there has been an effective adaptation of the rural market for manufacturing labor to the situation of rural women. The instability of the rhythms of production, a key entrepreneurial element in cheapening the cost of manual labor, is manipulated as a mechanism to achieve labor 'flexibility,' which in turn emanates from and is adapted to the needs of female workers. The irregularity and instability of production, the continual opening and closing of factory doors, and payment by piece rate (according to the number of units produced) all form part of the prevailing economic model, which comes to exemplify and derogate the women themselves: it is *they* who are unskilled, *they* who require labor flexibility, and *they*

who earn what they want, when they want it. In this way, rural manufacturing has been able to maintain and recreate the precarious conditions of the labor market such that they are viewed as a concession on the part of employers to women's domestic responsibilities. This social disqualification of female labor might diminish over time, but that process is slowed by another deeply rooted social fact: gender differences within the family.

4. The Household Perspective

It is only recently that investigators began to underscore the importance of *household survival strategies* as crucial mechanisms that permit poor households to organize production and consumption, not simply to adapt to changes associated with the life cycle, but to surmount the repeated economic crises that have buffeted Third World economies in recent years. We know that family survival strategies include not only cooperation, but also conflict, domination, and inequality, such that men are often able to impose their preferences and decisions on women (Dwyer and Bruce 1988, Ward 1993). The household is an arena in which the tasks and obligations of members, whose interests and powers may clash, are subject to continuous contestation.

Gender inequalities are manifested in the labor force behavior of family members. The broader political economy extends into intradomestic struggles. Rural men and women conceive of, are conceived of, and for that matter offer themselves very differently to the labor market, at least in certain core matters that not only signal differences, but reflect broader inequalities and hierarchies with respect to gender: the family-individual relation and conceptualization of male and female work.

4.1 Family-Individual Relations

During the 1970s, the availability of paid labor in canneries, workshops, factories, and *maquiladoras* was, in many cases, something entirely new in the life of rural communities (Ward 1993). Given economic crisis, rural families could not afford to pass up opportunities for work within their own communities. They began to redefine priorities and reorganize to harness emergent opportunities; and they did this in accord with traditional gender hierarchies. It was clearly a complex undertaking trying to incorporate these new labor opportunities while conforming to the idea that the female condition is permanent and immutable. With the aim of mitigating the social effects that women's wages and the power they potentially bring, the family positioned itself as the crucial intermediary in recruiting and organizing the new systems of work that arrived in the countryside (Hareven 1982, Rosado 1988).

In general, families looked to assure the permanence of two basic principles: to assure new income for the family, and to safeguard 'morality' in the face of the compromises entailed in taking paid labor. One form of mitigating the potentially disruptive effects of female work and extra-familial income was to send only young, single women into the workforce, i.e., those who lived in their paternal home and therefore had much less power to negotiate within the family. The selection of young unmarried women assured that the new earnings would be incorporated directly into the family circle. Under the authority of the father, the income generated by daughters was socially defined as an obligatory family *contribution*, much the same as the fulfilling of any other domestic obligation of a daughter or sister.

The contrasting family responsibilities of men and women are also evident in what happens to the earnings. As has been shown in a variety of societies, men experience much greater latitude than women in deciding what to do with their money, even when they are unmarried (Dwyer and Bruce 1988). In the end, the conservation of family norms and values was more important to male family members than the struggle to better working conditions or wages for their wives and daughters.

4.2 Male Work as 'Obligation'

The conceptualization of work and what one expects to achieve through it also distinguishes men from women. For men, work is viewed as an obligation through which the man must (but also can) fulfill the role of fundamental family provider. Through work, he hopes to generate income, sufficient to cover the basic necessities of those who depend on him, although this, as we know, is an ideological construction of relative recency in history, with little relation to economic reality (Rothstein 1995).

This social construction has assigned men both the obligations and the rights of a family provider. This means that men have the freedom to seek out the best pay and working conditions, rejecting jobs that cannot support their family. This makes it socially acceptable for men to reject poorly paid jobs and to migrate far from home in search of better jobs. Through mobility, men may avoid the low salaries offered in agriculture, agribusiness, manufacturing, and the *maquiladoras* that were built throughout the rural world in the 1970s (Arias and Wilson 1997). This helps explain the association in the countryside between an abundant demand for labor and at the same time the ongoing departure of men for other regions or countries in search of work.

This paradox illustrates another characteristic of male migration: that they can move away for work while still meeting their family obligations and not disturbing (indeed, reinforcing) the social framework in which they form a part, for they know that mothers,

wives, and sisters will be charged with maintaining their filial and familial obligations (Massey et al. 1987). In most societies, as Dwyer and Bruce (1988) indicate, the time that men actually devote to their children has always been meager and is easily reduced without repercussion.

The situation of the rural woman is entirely different. For her, liberty and the freedom to disengage from domestic obligations do not exist, so that her justification in leaving to look for work, or to find better working conditions, is truly limited and is supposed to be arranged through a long and costly negotiation with other women to take charge of the domestic responsibilities that she will no longer be able to meet.

4.3 Feminine Work as Complementary 'Help'

The paid work of women is governed by very different notions; it is seen as complementary help. Women's work is, and has always been seen as a form of altruistic collaboration, that is voluntary, generous, and absolute in comparison with male labor (Folbre 1988). That a woman is hardworking and that her efforts indeed 'help' and 'complement' the family income speaks well of her in moral and marital terms, but may limit her cash contribution to the family. In this gendered process, women's paid work, as 'complementary help,' is devalued. The efforts of women are socially recognized, but they are not seen as important work and are assigned an inferior status relative to men's work (Dwyer and Bruce 1988, Safilios-Rothschild 1988).

To speak of work as 'complementary' because it is female is a way of downgrading and devaluing, *a priori*, the contribution of women to household resources. Moreover, as Dwyer and Bruce (1988) and Ward (1993) have pointed out, in many societies it is socially accepted that men can retain a part of their earnings for themselves (and often the family never knows exactly *what* proportion of male earnings are retained). In contrast, wives and daughters are expected to hand over to the family *all* resources obtained outside the household (Mencher 1988). Thus, women's earnings may be a resource families count on more than men's. Thus, seeing women's work as secondary 'complementary help' is a social construction rather than an accurate portrayal of the situation.

Seeing female labor as 'help' makes it socially manageable by the family. But, the notion of women's work as 'help' limits it several ways. First, it reaffirms that the female priority remains reproduction and domestic responsibilities; if the work of the woman is only 'help,' it is by definition insufficiently important to require renegotiation of the distribution of roles and tasks within the household. Second, the notion of female paid labor as 'help' assures that it remains of lower value, and in this way, that the woman is always available to leave it to resume her function in the

family if needed. Third, viewing female labor as 'help' guarantees that in the face of family economic crises, the woman can (and *must*) enter the market to work and to accept conditions and wages that men are not expected to take, but that once the immediate crisis passes, she will return to her fundamental priority of the family (Hareven 1982). Finally, the notion of female work as 'help' is a key mechanism to avoid discrediting socially or maritally the work and earnings of men. Facing transitional labor markets that have shifted the balance of work opportunities between the sexes, the notion of 'help' contributes to the maintenance of traditional obligations and rights within the family, in obvious detriment to the position of the women themselves.

Thus, the conceptualization of female work as 'help' of a 'complementary' nature can be seen, in the end, as a porous category in which the social devaluation occurring in the labor market intersects with the social hierarchies of gender within the household. This intersection is what permits employers continued access to sources of low-cost labor, on the one hand, and the perpetuation of female work as subordinate to family obligations, on the other.

5. Recent Changes

As expected, the experience of female labor outside the household has affected gender relations within communities and families. Notwithstanding social resistance from employers as well as their families, women have begun to modify their relationships within the family and with employers. In rural societies in which the paid work of women forms an integral or intermittent part of local economic life, and where it is possible to encounter two or more generations of female workers, they have been renegotiating relations with their families that are less unequal, particularly with respect to education, mobility, and control over earnings, against the backdrop of a labor market ever more oriented to labor rights (Rosado 1988).

National and international struggles have been undertaken by nongovernmental organizations (NGOs) and women themselves during the 1990s, especially in Southeast Asia. The globalization of the economy has helped women to understand their conditions better and has extended the struggle of female workers to a geographic scale never before seen. Increasingly women have contested not only wages and work conditions that function outside the law, but those that violate basic human rights, such as the right of association, the limitation of child labor, and exposure to hazardous conditions, as well as combating different types of labor market discrimination. The struggle by female workers and NGOs has now succeeded to the point where some large international companies operating at the global level have developed norms assuring human and labor rights for workers connected to them under various subcontracting arrangements. Al-

though this is only a beginning, it indicates that rural industrialization may have entered a new phase of development.

See also: Boserup, Ester (1910–99); Development, Economics of; Development: Rural Development Strategies; East Asian Studies: Gender; Economic Development and Women; Economic Globalization and Gender; Family Planning Programs: Development and Outcomes; Family Planning Programs: Feminist Perspectives; Fertility Transition: China; Fertility Transition: East Asia; Fertility Transition: Economic Explanations; Fertility Transition: Latin America and the Caribbean; Fertility Transition: Southeast Asia; Gender Ideology: Cross-cultural Aspects; Islam and Gender; Labor Markets, Labor Movements, and Gender in Developing Nations; Land Rights and Gender; Latin American Studies: Gender; Lone Mothers in Nations of the South; Nation-states, Nationalism, and Gender; Near Middle East/North African Studies: Gender; Poverty and Gender in Developing Nations; Reproductive Rights in Developing Nations; Southeast Asian Studies: Gender

Bibliography

- Arias P 1988 El empleo a domicilio en el medio rural: La nueva manufactura. *Estudios Sociológicos*, 6: 405–36
- Arias P, Wilson F 1997 *La aguja y el surco. Cambio regional, consumo y relaciones de género en la industria de la ropa en México*. Universidad de Guadalajara, Guadalajara, Mexico
- Bachrach Ehlers T 1993 Belts, business, and Bloomingdale's: An alternative model for Guatemalan artisan development. In: Nash J (ed.) *Crafts in the World Market*. State University of New York, Albany, NY, pp. 181–96
- Boris E, Daniels C R (eds.) 1989 *Homework. Historical and Contemporary Perspectives on Paid Labor at Home*. University of Illinois Press, Urbana, IL
- Bradley H 1989 *Men's Work, Women's Work*. University of Minnesota Press, Minneapolis, MN
- Dwyer D, Bruce J (eds.) 1988 *A Home Divided. Women and Income in the Third World*. Stanford University Press, Stanford, CA pp. 1–19
- England P (ed.) 1993 *Theory on Gender. Feminism on Theory*. Aldine de Gruyter, New York, pp. 3–22
- Fernández-Kelly M, García A 1989 Hispanic women and homework. Women in the informal economy of Miami and Los Angeles. In: Boris B, Daniels C R (eds.) *Homework. Historical and Contemporary Perspectives on Paid Labor at Home*. University of Illinois Press, Urbana, IL, pp. 165–79
- Folbre N 1988 The black four of hearts: toward a new paradigm of household economics. In: Dwyer D, Bruce J (eds.) *A Home Divided. Women and Income in the Third World*. Stanford University Press, Stanford, CA, pp. 248–62
- Hareven T K 1982 *Family Time and Industrial Time*. Cambridge University Press, Cambridge, MA
- Maruani M 1991 La construcción social de las diferencias de sexo en el mercado de trabajo. *Revista de Economía y Sociología del Trabajo*. Septiembre–diciembre 1991, Madrid, pp. 129–37
- Massey D S, Alarcón R, Durandt J, González H 1987 *Return to Aztlan*. University of California Press, Berkeley, CA

- Mencher J P 1988 Women's work and poverty: Women's contribution to household maintenance in south India. In: Dwyer D, Bruce J (eds.) *A Home Divided. Women and Income in the Third World*. Stanford University Press, Stanford, CA, pp. 99–119
- Rosado G 1988 Las mujeres de San Pablo. trabajo y vida cotidiana. In: Aranda J (ed.) *Las mujeres en el campo*. Universidad Autónoma Benito, Oaxaca de Juarez, Mexico, pp. 147–62
- Rothstein F A 1995 Gender and multiple income strategies in rural Mexico. In: Bose C E, Acosta-Belén E (eds.) *Women in the Latin American Development Process*. Temple University Press, Philadelphia, PA, pp. 167–93
- Safilios-Rothschild C 1988 The impact of agrarian reform on men's and women's incomes in rural Honduras. In: Dwyer D, Bruce J (eds.) *A Home Divided. Women and Income in the Third World*. Stanford University Press, Stanford, CA, pp. 216–28
- Ward K B 1993 Reconceptualizing world system theory to include women. In: England P (ed.) *Theory on Gender. Feminism on Theory*. A de Gruyter, New York, pp. 43–68
- Wilson F 1991 *Sweaters. Gender, Class and Workshops-based Industry in Mexico*. Macmillan, London, UK

P. Arias

Rural Planning: General

The heading 'rural planning' as a sector of government involvement is often absent from plan documents in many countries. But policy and legislation related to agricultural, transportation and communications, education, health and human services, and social welfare, sectors generally addressed in plans, have had far reaching and profound impacts on the physical conditions of, and infrastructure availability in, rural areas. They have also determined the range and quality of services accessible to rural populations. When headings such as rural planning, rural development, integrated rural development are present in plan documents, as discrete sectors requiring attention and having related funding lines in national and state budgets, these sectors are often defined in terms of engagement with social welfare agendas. The goals of such sectoral policies have often included statements about the need to redress inequalities in access to public services and redistributive justice. A broader definition of rural planning as encompassing policy that has both explicit and implicit intention to impact rural conditions, economic, social, and physical is adopted here.

During the formative period of the United States in the nineteenth century, agriculture production was a central element in its' robust growth and development. Similarly, an agriculture-based economy and a predominantly agrarian population was the legacy of heavily populated, newly independent, countries like China and India, which faced the task of nation building in the second half of the twentieth century.

But over the years these countries have differed significantly in their approach to rural planning. The interventions of their governments in the rural sector define a spectrum of strategies and approaches, from total centralized control to indirect, market-driven change. Furthermore, their strategies, which took divergent paths at the start of nation formation, have over time sharply changed direction.

The United States (US), India, and China provide examples of countries that have developed rural planning-related policies based on quite differing ideologies about planning and the role of rural areas in national development. Their positions have ranged from the free market, laissez faire orientation typical of 'First World' countries like the US to policies of 'Second World' countries like China that adopted a centralized, cooperatives based model of rural planning of the Soviet Union. Seeking a more intermediate path between these two extremes were countries like India, a founding member of the block of so-called 'Third World' countries that sought to chart a distinct path to development. Examining the development of the rural planning concept in countries like the US and India provides a glimpse of the differing interpretations of rural planning that have been evolving around the world in countries where the government's ability to control, plan, and implement has been limited by, and has been cognizant of, the significant role of the private sector in the farming endeavor. Study of the China model provides insight on the strengths and limitations of sustained, centralized efforts to control developments in rural economy and society.

1. The United States Model of Rural Planning

The US is acknowledged to be one of the most open and laissez faire economies in the world. Within the US, the idea that attention needs to be turned specifically to rural planning is not a widely embraced concept. But models of economic growth that guide policy are accepted and in these the agricultural sector is factored as an integral component of the national economy. Historically, US agricultural policy has defined and shaped rural conditions and had far-reaching consequences on the development of rural areas, but it has not been recognized as an instrument for rural planning.

This American ambivalence towards acknowledging the existence of rural planning is understandable given historical antecedents. After American independence, the feudalistic land law that prevailed under British rule was replaced with a freehold land tenure system under which public lands were sold to raise cash for the new government. To facilitate sales the US Congress ordered a land survey in 1785 that established a grid system of land division as the dominant pattern for the area west of Pittsburgh. Its

indelible stamp ordered the transportation and communication systems and helped give shape to rural administrative areas. Subsequent land laws and purchase allowed 'a mass distribution of land to those who tilled the soil [that] was unique in the history of the world up to that time' Cocharane (1979, p. 57). The philosophy of land ownership as vesting freehold in an individual, and the establishment of a unique American tradition of independent and self reliant farm families scattered in a regular pattern, each in isolation, on their own individual pieces of land, has perhaps contributed to a national antipathy to identifying government interventions as responsible for rural development. Nonetheless these early interventions shaped the rural context, with plans that were not always openly agreed to or publicly articulated, and embedded it with desired outcomes that might have been contested if formally acknowledged in planning documents.

The movement of the US agricultural enterprise from the subsistence cultivation of the pioneer to the high tech, mechanized, computerized, agro industry of today's corporate farm is not the product of free-market forces but of significant government actions. The role of the government in shaping US farming is noted and generally accepted in the literature pertaining to American agriculture (Rasmussen 1960, Cocharane 1979, McConnell 1953). Cocharane (1979, p. 277) delineates a broad and inclusive set of categories in which government has been active, asserting that it has been involved in almost every aspect of agricultural development. With the passage of the Agricultural Adjustment Acts of 1933 and 1938 the Department of Agriculture has supported and managed agricultural production pursuing policies affecting both the supply and distribution systems of agricultural production (Cocharane 1979, pp. 141-2).

1.1 Modern Farming in the US

Government activities in the transport, research, and education sectors have transformed American agriculture from subsistence to cash cropping for regional and national markets and later to capital-intensive agribusiness. Significant government support of educational and research institutions assisting in agricultural development through scientific and technological innovations has been sustained for more than a century. The Morrill Act of 1862 enabled government to fund land grant colleges in each state and the Hatch Act of 1887 allowed annual allocations for state agricultural experimentation stations devoted to basic agriculture research (Rasmussen 1960). The impact of this activity on shaping the scientific base of modern American farming and increasing productivity and efficiency was significant and to date has had resonance and wielded influence around the world.

In response to the worldwide Depression of 1929 New Deal agricultural economists created the Agricultural Adjustment Acts of 1933 and 1938, to set the price and allowable level of farm production. The central theme in these legislative efforts was to stabilize agricultural prices and income by offering subsidy payments to farmers to reduce production, price support measures, and commodity storage. In the postwar period from 1945 to 1975 the modern farm evolved in the context of rapid advances in national/international transport the emergence of the United States as a major world power, and the corresponding change in its domestic and foreign policy, and tremendous advances in scientific/technological farming in the United States including development of high-yield hybrids, fertilizers, pest controls, and extensive mechanization. The modern farm featured large-scale production for national and international markets. Cultivation for sale rather than family subsistence became the central production objective of the farm enterprise. However the ideological underpinnings of the earlier time of minimum government control in the farm enterprise persisted. There was support for government to be an agent responsible for presenting new options and alternatives which might voluntarily be subscribed to, or not, by the American farmer.

1.2 Corporate Farming in the US

A shift from modern to corporate farming after 1975 was triggered by changes in the role the United States played in world affairs and the tack it took with respect to its foreign policy following World War II. The United States emerged as a leading proponent of global free trade. This posture meant that the highly efficient farms in the US began to function as the 'world's breadbasket.' Farmers produced and exported crops to India, Russia, and other countries when these countries faced domestic shortfalls in food production. Advances in national and international transport aided this expansion of the market. 'Green revolution' technology packages of high-yielding hybrids, super fertilizers, and pest controls were developed leading to unexpectedly large crop yields. Specialization and mono-crop farms become common. Although energy intensive in their use of new products and machinery, they were viable as energy costs were kept relatively low by the government. A process of consolidation of family and modern farms into larger and larger vertically and horizontally integrated corporate entities occurred. The resulting corporate super farms have changed the rural landscape tremendously as fewer and fewer farms account for more and more of the of total farm sales and total farm profits in the United States. This concentration is apparent in the fact that in 1997 two percent of all farms in the US produced 50 percent of total farm sales.

1.3 Services-oriented Rural Planning

In the 1980s and 1990s it has been increasingly obvious in the United States that although the link between agriculture policy and rural planning has traditionally been very strong it may be less important in the emerging 'global' economy of the twenty-first century. Today the link that needs to be forged for rural planning is with economic sectors such as manufacturing and services. The emerging 'new economy' is based on the growth of the service sector and geared to the information industry that has centralized population in metropolitan areas. The impact of this centralization on populations that remain rural and dispersed is significant. For example in rural areas public services in education, health, and social welfare provision are of lower quality than in metropolitan areas. Currently US rural planning efforts have focused on providing quality services despite the delivery difficulties posed by lower population densities that mean that greater distances must be covered to reach targeted rural populations. A second strand of rural planning pertains to national settlement patterns and the belief that it is important to maintain numbers of people in rural areas. Planning efforts creating rural enterprise and empowerment zones have involved job-training programs as an economic development policy to attract entrepreneurs. The second element of this strategy aims to reduce costs for entrepreneurship in rural areas by programs that make capital available for business development. The scale of such efforts is however quite modest. Finally, the current paradigm of rural planning focuses on equalizing rural communication and infrastructure with urban. Supporting the expansion of Internet and high-speed communications links, wireless, and satellite connections all aimed at linking dispersed populations are an integral part of this agenda.

2. The 'Third World' Approach to Rural Planning

Numerous countries attained independence from European colonial rule at the end of World War II. Some, like the Peoples' Republic of China, became 'Second World' countries that adopted planning models emulating Soviet style control by a strong central government. Another block of countries attempted to forge a separate third path to development differing from United States style capitalism and Soviet style socialism. This 'Third World' forged constitutions endorsing centralized planning by government to attain modernization by facilitating rapid industrialization of both the private and public sectors. In these models rural planning tended to be a key element in national economic development. Rural planning measures aimed to increase production but also to redistribute wealth on the grounds of social justice and human welfare. The latter provided legitimacy for governmental planning which was essential

in countries recently released from colonization. Furthermore the majority of these nations' populations were rural. In approaching the development of rural areas many of these countries emulated the US model. They aimed to increase and improve agricultural production through the application of scientific research that was to be developed, disseminated, and implemented in farming practice through agricultural extension services. This was seen as the modern way for rural planning to support national economic development.

Third World countries aimed to use agricultural production to raise capital for industrialization and the development of related infrastructure, as well as for the provision of urban as well as rural based populations. However they tempered their desire to increase agriculture production with the broader objective of bringing about an increased living standard and quality of life for economically disadvantaged populations in the countryside. Their emphasis on rural planning for human development reflected their recognition that the majority of their populations were rural, asset poor, involved in subsistence production, and members of lower status social groups of social structures that were feudal in character. Over time, these countries evolved a planning model for 'integrated rural development' that recognized gender and class differences and attempted to address social inequalities in access to rural resources; a process differentiated from earlier production oriented agricultural development models. This more recent approach to rural planning attempted to combine and balance between agricultural and economic development and social welfare and redistributive justice.

3. The Indian Model

At the point of independence from British rule India embraced rural planning directly. Initial Indian national-planning efforts to achieve rapid socioeconomic change concentrated on industrialization at chosen metropolitan centers, on the assumption that the benefits would 'trickle down' through the economy. Planners believed that although initially benefits would be concentrated in the core cities they would spread to rural areas. However, they also recognized the need to promote complementary rural planning to reduce existing and anticipated urban-rural disparities in wealth.

Rural planning measures in India conspicuously emulated the US model of agricultural development. They included: irrigation projects; the promotion of scientifically developed cash crops through the provision of improved seeds, equipment and fertilizer (the so-called Green Revolution approach to making agriculture productive); providing credit for agricultural investments to stimulate agricultural production; and, the establishment of a system of extension services

to introduce scientific farming techniques at the village level. They also regulated markets to stimulate trade in agricultural commodities; invested in roads and communications; and founded cooperatives to buy and sell agricultural products and facilitate the marketing of goods in rural areas.

3.1 Redistribution and Equity

The national policy of investing in industry and commercial agriculture increased the gap between those involved in the corporate formal sectors of the economy, urban and rural, and those in the informal, traditional, subsistence sectors. Rural landless and land poor, squeezed out of traditional positions in village society, migrated to the city in unprecedented numbers. Cities were unable to accommodate the massive influx of displaced, often unskilled, rural laborers and their families and urban environments deteriorated as slum settlements of the poor proliferated. Responding to this phenomenon inspired a new set of rural planning policies. Stemming migration to the city by improving rural life and diversifying the rural economy emerged as prime planning objectives. Making rural living more attractive included provision of postal service, healthcare facilities, clean drinking water, and electricity in the countryside. As this policy evolved, its objectives also expanded to include increasing equity and distributing the benefits of development to the poor more explicitly. This change inspired such planning measures as putting a ceiling on individual land holdings, reforming tenancy to enhance the rights to the land of those who cultivated it, initiating programs to support traditional artisans with better tools, establishing credit and marketing networks to allow them to compete with industry or at least sustain them until alternative jobs were created. Finally, the Indian government emphasized free, mandatory elementary school education throughout rural areas to improve skills the poor needed to compete for development benefits.

In 1976 A. T. Mosher an international agricultural development expert summarized the various phases of Indian rural development planning in *Thinking about Rural Development*. He described some common characteristics. They were:

(a) Recognition that although economic development is important it is not an essential precondition to the social transformation that is considered the central objective of a rural development strategy;

(b) Attention to increasing efficiency with the technical revolution of fertilizers and high yield varieties of cereals. The attempt was to boost both food grain production and commercial crops and monoculture production of commodities for which there was an international market;

(c) Development of programs that addressed equity and asset redistribution; and

(d) Enhancing rural citizen's participation in the evolution of locally appropriate planning programs. Elements of these phases of rural planning policy and objectives in India—moral and social transformation, economic efficiency, redistribution and social justice, and popular participation in decision making—are to be found in the rural planning approaches of many countries around the world although there is a great deal of variation both in emphasis in policy and in the capacity to implement.

4. *Current Priorities and Approaches in Rural Planning*

Rural planning oriented to improving physical infrastructure and the quality of life for rural populations was an important policy agenda in developing countries in the earlier periods of their post-independence histories. More recently, as rates of urbanization have remained high in these countries and cities have experienced hyper growth, despite efforts to stem rural migration and population increases, rural planning in many countries has become a lower priority endeavor. The worldwide liberalization of economies attention is focused on World cities of over 10 million people, and, as a result, notions of intervention in rural areas to improve rural conditions have been given far less significance.

Concerns about the long-term negative impacts of the 'Green Revolution' have emerged in many parts of the World. They are related to degradation of soil quality; depletion of groundwater tables; the long-term sustainability of modern agricultural cultivation on prime land; and, as supply of such land is exhausted, about the viability of applications of this technology package on poorer quality, more fragile lands; and rising concern about genetically modified crops entering the food system of human populations. These concerns have led to calls for a modified, mixed approach to rural development that recognizes traditional, sustainable practices. In contexts of risk prone environments, development planners are describing a need for a rural planning strategy in which 'top down' science driven Green Revolution strategies of intensifying agricultural cultivation are tempered with the understanding that can be attained by tapping the knowledge of rural cultivators who are following traditional, self sustaining forms of agricultural production (Pichón et al. 1999). This position, labeled 'sustainable agriculture', has stimulated rural planning efforts based on more traditional modes of agriculture (National Research Council 1989). The task of helping rural places has been defined as a set of two complementary concerns related to how people will earn a living (Lapping et al. 1989) and how they will do it so that natural, environmental resources will be conserved (Sargent et al. 1991).

In the 'First World' rural planning is coming full circle. From a 'universalistic' modern-science driven notion of development and progress there is a renewed orientation to the specificities of the local, rural, asset base of both physical and human capital. In much of the developing world, facing the waves of change engulfing countries in the throes of economic liberalization, rural planning has been placed much lower in priority to programs that address issues of rapid urbanization and globalization. Thus programs related to social justice, participatory decision making, sustainability, and the environment have taken second place to economic transformation of rural production.

See also: Agricultural Change Theory; Agricultural Sciences and Technology; Agriculture, Economics of; Development: Sustainable Agriculture; Environment and Development; Environmental and Resource Management; Environmentalism: Preservation and Conservation; Farmland Preservation; Green Revolution; Greening of Technology and Ecotechnology

Bibliography

- Cochrane W W 1979 *The Development of American Agriculture: A Historical Analysis*. University of Minnesota Press, Minneapolis, MN
- Lapping M B, Daniels T L, Keller J W 1989 *Rural Planning and Development in the United States*. Guilford Press, New York
- McConnell G 1953 *The Decline of Agrarian Democracy*. Athenaeum/University of California Press, New York
- Mosher A T 1976 *Thinking About Rural Development*. Agricultural Development Council, New York
- National Research Council 1989 *Alternative Agriculture*. National Academy Press, Washington, DC
- Pichón J F, Uquillas J E, Frechione J 1999 *Traditional and Modern Natural Resource Management in Latin America*. University of Pittsburgh Press, Pittsburgh, PA
- Rasmussen W D 1960 *Readings in the History of American Agriculture*. University of Illinois Press, Urbana, IL
- Sargent F O, Lusk P, Rivera J A, Varela M 1991 *Rural Environmental Planning for Sustainable Communities*. Island Press, Washington, DC

H. C. Dandekar

Rural Sociology

Rural sociology has traditionally been defined as the study of social organization and social processes that are characteristic of geographical zones where population sizes and densities are relatively low. Rural societies, however, do not exist in isolation or a social vacuum, and thus the subject matters of rural sociology increasingly reflect the larger processes of the regional differentiation and allocation of populations and economic activities, and of social relations within

a society as a whole (and, increasingly, within global economy and society).

Perhaps the single most crucial fact about this field is that, as a formal sociological specialty, rural sociology is largely US in its origins, and that rural sociology emerged within the land-grant agricultural college system and in close association with major US historical trends in agrarian politics. In this article I will begin with the history of US rural sociology, particularly in relation to major contemporary emphases within the field. In the second section I will focus on the major branches of contemporary rural sociology, and especially on the significance of the 'new rural sociology' during the last two decades of the twentieth century.

1. *The Origins of Rural Sociology*

In a sense rural sociology, much like the discipline at large, has had very strong roots in nineteenth century social thought. Many of the major issues of concern to the classical theorists were in some sense germane to rural sociology, and have remained important to the subdiscipline a century later. Among the most important of the preoccupations of nineteenth century social theory was whether village and farm life was more socially cohesive than, or morally or socially superior to, metropolitan life, and whether rural social relations would be resilient in the face of capitalist industrialization. Emile Durkheim (1858–1917) and Karl Marx (1818–83) both argued that urban-industrial capitalism, the advancing division of labor, and modern technologies and organizational practices were inexorable—and, on balance, progressive—social forces that would eventually supersede or supplant rural social forms. Ferdinand Tönnies (1855–1936) concurred that urbanization and industrial capitalism were tending to lead to the decline of intimate village communal bonds, though Tönnies was more optimistic than Durkheim and Marx that rural social relations could be sustained. For most of its history, rural sociology has reflected these nineteenth century debates over the desirability and resilience of traditional rural social organization.

It is now generally agreed that the pioneering rural sociological studies in the USA were those by W. E. B. DuBois in the late 1890s on the social consequences of the postbellum crop-lien system in Southern plantation agriculture. But while DuBois was the pioneering American rural sociologist and there were several prominent nonland grant rural sociologists (e.g., F. H. Giddings of Columbia University) in the field's early years, the character of rural sociology in the twentieth century was largely shaped by its articulation with the land grant college system. Rural sociology as a recognized subdiscipline largely had its origins in the preoccupation of land grant college leaders with the challenge of Populism. Founded and funded by

industrial and other elites after the turn of the century, the Country Life movement promoted the view that the problems of rural society were not due to the shortcomings of industrial capitalism as the Populists had held. These problems were rather seen to be due to a lack of organization, poor infrastructure, and technological backwardness in rural areas, which in turn led rural society to lag behind urban USA.

In 1908 President Theodore Roosevelt recognized this reform movement by appointing a Presidential Commission on Country Life. The next six decades of US rural sociology were largely prefigured by the County Life Commission Report, which was published in 1909. The Report did acknowledge many of the social problems of rural USA (e.g., the inequities of the crop-lien system, widespread tenancy, land speculation). But the dominant thrust of the Commission Report was the case it made that an expanded effort to modernize rural USA technologically and structurally was integral to improving the condition of rural USA. The Commission recommended the harnessing of the social sciences, particularly agricultural economics and rural sociology, to support the technological modernizationist vision for rural America. The land grant colleges responded to the Country Life Commission Report by hiring rural sociologists and establishing rural sociology units or departments. Rural sociological studies in land grant colleges of agriculture were seen to be important in order to help remove social barriers to technological modernization and to stabilize rural communities.

Sorokin and Zimmerman's (1929) influential treatise *Principles of Rural-Urban Sociology* exemplified the first three decades of US rural sociology. *Principles* synthesized the European classical tradition (particularly that of Tönnies) and land grant technological progressivism with the emerging empirical tradition of research on rural people and communities. The Sorokin-Zimmerman perspective was essentially a theory of rural-urban continuum, in which the character of social relations was posited to be shaped by the size, density, and related parameters of settlement patterns.

Very soon after *Principles* was published, rural sociology would be confronted with the Great Depression and rural squalor, and most rural sociologists embraced New Deal reformism. The Great Depression and the fashioning of the Depression-era New Deal opened up vast opportunities for rural sociological scholarship on rural communities and peoples which was closely linked with the New Deal reform agenda (e.g., land tenure reform, encouragement of co-operatives, commodity programs) (see the summaries in Buttel et al. 1990, Chap. 1, Larson et al. 1992).

Reform-minded rural sociology, however, was short lived as a result of the ideological shift to the right that had largely swept away most New Deal rural programs by the mid-1940s (Larson et al. 2000). The jettisoning of the New Deal rural programs induced the land

grant colleges to refocus their efforts on the technological upgrading of the countryside. In the late 1940s and early 1950s a handful of prescient rural sociologists developed the diffusion of innovations perspective on agricultural innovation and change which provided sociological support to the land grant technological project. The diffusion of innovations perspective was a social-psychological perspective on farmer technology adoption behavior that posited that innovative and progressive technologies and value orientations tend to diffuse from metropolitan institutions to rural people and localities, and that persons with more modern orientations are most likely to adopt improved metropolitan-developed technologies and practices (see Rogers 1995 for a contemporary summary, and Fliegel 1993 for a historical perspective and sympathetic critique). Shortly thereafter the social psychology of farm practice adoption, and of related rural behaviors such as educational aspirations, provided a comfortable niche for rural sociology within the land grant college system.

Rural sociology now exists in dozens of world nations. Much of the impetus for the global spread of rural sociology was due to the missionary zeal of US rural sociologists who from the 1940s through the 1960s sowed the seeds of US-style rural sociology across an impressive range of countries across the globe. Since the 1970s, however, rural sociology has become increasingly diverse and less influential globally. US rural sociology increasingly borrows from the ideas of European and Third World scholars (e.g., Goodman and Redclift 1991, Llambi 1990, van der Ploeg 1992), and is sometimes criticized for its parochialism and quantitative emphasis.

2. *The Major Focuses of Contemporary Rural Sociology*

Rural sociology has increasingly become a less coherent subdiscipline, and more of an 'umbrella' for an assemblage of research, much of which no longer pertains primarily to rural people or rural social relations. Rural sociology now has six major branches: rural population, rural community, natural resources and environment, agriculture, sociology of international rural development, and sociology of agricultural science and technology.

From the very inception of rural sociology, sociological analysis of rural population and rural community dynamics through census and social survey data has been central to the field. In the early days of rural sociology, rural population and rural community studies were very closely articulated. In large part this was because census of population data for counties and rural places were the most accessible data sources for characterizing rural communities and farm structures. Even by the end of the twentieth century there remains a significant relationship between rural

population and community studies (see, e.g., Fuguitt et al. 1989, Garkovich 1989).

While there has been a continuing basis for articulation between rural population and community studies, these areas of work have become increasingly differentiated so that they are now seen as quite distinct specialties. The use of census and survey data to document rural-urban differences, as in Sorokin and Zimmerman's (1929) *Principles*, fell into disrepute, and the rural-urban continuum perspective was slowly but surely undermined empirically (Newby 1980). Rural population research has followed the larger subdiscipline of demography in becoming more quantitative, while the rural sociological analysis of communities, though it has a strong quantitative wing, has continued to have a significant cadre of researchers who use qualitative or field methodologies (Wilkinson 1991). Studies of regional and labor market inequalities (Lyson and Falk 1993, Lobao 1990), rural poverty dynamics and racial inequalities (Duncan 1999), and rural gender inequalities (Sachs 1996) have complemented the traditional rural sociological emphasis on community studies.

While there has been a long and significant tradition of rural sociological scholarship on the relations among people, communities, and natural resources (Field and Burch 1988), the sociology of natural resources and environment in the contemporary sense did not emerge until the early 1970s. Rural sociology pioneered in development of the larger subdiscipline of environmental sociology and natural resource sociology. Rural sociologists have figured prominently in both environmental sociology (Redclift and Woodgate 1997) and the more management-oriented field of the sociology of natural resources (Field and Burch 1988).

As noted earlier, many of the most prominent US rural sociologists of the post-World War II period devoted major segments of their careers to encouraging the diffusion of rural sociology across the globe, particularly in the developing world. Part of the early impulse for elaborating an international rural sociology was the intellectual one of wanting to promote a comparative approach to understanding rural social organization. But equally important was the fact that the postwar period was an emerging era of 'developmentalism' (i.e., an era of faith in the efficacy of planned social change and development in the decolonizing world, see McMichael 2000). Rural sociologists trained in the adoption-diffusion tradition were particularly well represented among the first cohort of international rural sociologists (see Fliegel 1993, Rogers 1995). Scholars adhering to the diffusion of innovations perspective found their work very compatible with modernization theory, the dominant theory at the time in the sociology of development. This modernizationist/developmentalist tradition of rural sociology predominated for nearly three decades, but ultimately it would be undermined as a result of

critical assessments of the Green Revolution experience and of the role of rural sociology in international assistance (e.g., Koppel 1995).

Of all the major rural sociological specializations, the sociology of agricultural science and technology is, in one sense, one of its oldest, and in another sense, rural sociology's newest major speciality area. There were significant studies of technological change and labor displacement in Southern plantation agriculture during the 1930s through the 1950s. And, as noted earlier, the diffusion of agricultural innovations predominated in US rural sociology from the late 1940s through the late 1960s (see the summaries in Buttel et al. 1990 and Fliegel 1993).

Diffusionism, however, declined for three main reasons, each of which corresponded with an important new dimension in sociological research on agricultural science and technology. The first reason, as suggested earlier, was the growing criticism of the ability of the technological modernization project to generate equitable social benefits. This criticism was originally focused on the shortcomings of the Green Revolution, but was soon extended to the publicly funded agricultural-technological project in the advanced countries as well (see Newby 1980). Second, beginning in the late 1970s the Busch-Lacy group drew on world systems theory and the sociology of science to develop an influential critical perspective on the social causes and consequences of technological change in agriculture (Busch and Lacy 1983). Busch and Lacy's approach to the sociology of agricultural science demonstrated that social factors shape the content of new technologies, that new technologies tend to have social costs as well as benefits, and that rural sociologists need to take a more agnostic position toward new technologies than was generally the case among diffusionists. Third, in the 1970s and 1980s US rural sociology was increasingly influenced by the rise of Marxism in sociology at large. Neo-Marxism played a particularly prominent role in the sociology of agricultural science and technology (Kloppenburger 1988, Goodman and Redclift 1991). The neo-Marxist literature on agricultural science focused sociological attention on the class character of agricultural research institutions and agricultural technologies, and in so doing was often critical of diffusionism for having obscured how agricultural research is implicated in class and power relations.

It was noted at the outset of this article that rural sociology's history has been one of alternating between opposing views about the distinctiveness and resilience of traditional rural social structures. Rural sociology came into its own during the post-World War II era when it rejected rural romanticism and transcended New Deal reformism by embracing the technological-modernizationist assumptions. In the 1970s and early 1980s, however, the balance between these two views would be equalized to a degree. As noted, emerging controversies over agricultural technologies and their

social and environmental impacts removed some of the luster from the technological-modernizationist perspective in rural sociology. In addition, the 1970s were a period of 'rural renaissance' in the USA, as the population of nonmetropolitan counties grew faster than that of metropolitan counties due to net metro-to-nonmetro migration (Fuguitt et al. 1989).

These new realities of rural USA during the 1970s, along with the growing popularity within sociology of various 'critical' theories, led to what is often referred to as the 'new rural sociology.' These new theories were actually quite diverse, drawing heavily from neo-Marxism but also from neo-Weberianism, the peasant studies literature, and critical theories drawn from hermeneutics and related traditions (e.g., Mann 1990, Friedmann 1978, Friedland et al. 1981). To the degree there was a coherence within the new rural sociology, it was its critique of modernizationism and diffusionism within rural sociology, and its insistence on spotlighting the role of class and power relations. The 'new rural sociology' was originally most prominent in the sociological literature on agricultural change (see the summary in Buttel et al. 1990), but most recently it exhibited a major shift toward research on cross-border agricultural commodity chains and on the globalization of agrifood systems more generally (Bonanno et al. 1994, McMichael 2000).

Perhaps the clearest indication of the importance of this new scholarly tradition was that even scholars who have been ambivalent about the new rural sociology have been obliged to address its premises and hypotheses (van der Ploeg 1992). The new rural sociology has also directly or indirectly influenced most other areas of rural sociology (see, e.g., Lyson and Falk 1993, Lobao 1990). But despite the fact that the new rural sociology is the most pervasive new trend during the last quarter century of rural sociology, rural sociology remains highly diversified. The new rural sociology, even in its heyday of the 1980s through the mid-1990s, did not come close to dominating the field in the way the diffusion paradigm did in the 1950s.

See also: Agricultural Change Theory; Development: Rural Development Strategies; Development: Sustainable Agriculture; Diffusion, Sociology of; Human Ecology; Innovation, Theory of; Social Change: Types

Bibliography

- Bonanno A, Busch L, Friedland W, Gouveia L, Mingione E (eds.) 1994 *From Columbus to ConAgra*. University Press of Kansas, Lawrence, KS
- Busch L, Lacy W B 1983 *Science, Agriculture, and the Politics of Research*. Westview Press, Boulder, CO
- Buttel F H, Larson O F, Gillespie G W Jr 1990 *The Sociology of Agriculture*. Greenwood Press, Westport, CT

- Duncan C M 1999 *Worlds Apart*. Yale University Press, New Haven, CT
- Field D R, Burch W R 1988 *Rural Sociology and the Environment*. Greenwood Press, Westport, CT
- Fliegel F C 1993 *Diffusion Research in Rural Sociology*. Greenwood Press, Westport, CT
- Friedland W H, Barton A H, Thomas R J 1981 *Manufacturing Green Gold*. Cambridge University Press, New York
- Friedmann H 1978 World market, state, and family farm: Social bases of household production in an era of wage labor. *Comparative Studies in Society and History* 20: 545–86
- Fuguitt G V, Brown D L, Beale C L 1989 *Rural and Small Town America*. Russell Sage Foundation, New York
- Garkovich L 1989 *Population and Community in Rural America*. Praeger, New York
- Goodman D, Redclift M 1991 *Refashioning Nature*. Routledge, New York
- Kloppenburger J Jr 1988 *First the Seed*. Cambridge University Press, New York
- Koppel B M (ed.) 1995 *Induced Innovation Theory and International Agricultural Development*. Johns Hopkins University Press, Baltimore, MD
- Larson O F, Zimmerman J N, Moe E O 1992 *Sociology in Government*. Westview Press, Boulder, CO
- Llambi L 1990 Transitions to and within capitalism: Agrarian transitions in Latin America. *Sociologia Ruralis* 30: 174–96
- Lobao L M 1990 *Locality and Inequality: Farm and Industry Structure and Socioeconomic Conditions*. State University of New York Press, Albany, NY
- Luloff A E, Swanson L E (eds.) 1990 *American Rural Communities*. Westview Press, Boulder, CO
- Lyson T A, Falk W W (eds.) 1993 *Forgotten Places*. University Press of Kansas, Lawrence, KS
- Mann S A 1990 *Agrarian Capitalism in Theory and Practice*. University of North Carolina Press, Chapel Hill, NC
- McMichael P 2000 *Development and Social Change*, 2nd edn. Pine Forge Press, Thousand Oaks, CA
- Newby H 1980 Trend report—rural sociology. *Current Sociology* 28: 1–141
- Redclift M, Woodgate G 1997 *International Handbook of Environmental Sociology*. Edward Elgar, Cheltenham, UK
- Rogers E M 1995 *Diffusion of Innovations*, 4th edn. Free Press, New York
- Sachs C 1996 *Gendered Fields*. Westview Press, Boulder, CO
- Sorokin P A, Zimmerman C C 1929 *Principles of Rural–Urban Sociology*. Henry Holt, New York
- van der Ploeg J D 1992 The reconstitution of locality: Technology and labour in modern agriculture. In: Marsden T, Lowe P, Whatmore S (eds.) *Labour and Locality*. David Fulton, London, pp. 19–43
- Wilkinson K 1991 *The Community in Rural Society*. Greenwood Press, Westport, CT

F. H. Buttel

Russian Revolution, The

The term is used in different ways. It refers to varying events and time frames. Normally it means the major upheaval of 1917 that replaced the tsarist empire with Bolshevik rule. In this case the term can refer to

either the demise of the monarchy in the so-called *February Revolution* (February 23/March 8 N. S. to March 2/15) or to the *overthrow of October* 25–26 (November 7–8) when the Bolsheviks seized power in the Petrograd Soviet of Workers' and Soldiers' Deputies. But 'Russian Revolution' is often extended to denote longer processes and obviously related sequences of events of the same type. In this case, the term encompasses the 1918–20 Civil War immediately following the October Revolution as well as the subsequent early Soviet period up to the *Stalin Revolution* of 1929–30 or—the widest definition—the entire transformation of Russia from the turn of the century and so-called *First Revolution* of 1905–6 all the way down to the Stalin Era. Looking at it this way, the 'Russian Revolution' covers Russia's peculiar political and socioeconomic modernization crisis which ended in the establishment of the Soviet Communist developmental dictatorship (1900–1930).

Substantive reasons and the question of research feasibility suggest narrowing this range of definitions somewhat. 'Russian Revolution' should be restricted to the convulsive transition from the Russian *ancien régime* to the Bolshevik form of government. The term properly denotes the two core events of the transfer of power: the overthrows of February and October and their consequences. However, there are valid reasons for including the Civil War because it was by nature the inevitable, belatedly-waged war of revolution. Both time frames (1917 and 1917–21) lend themselves to differentiation of processes into *politics and power struggle*, *socioeconomic change*, and *cultural and nationalist developments*.

1. The Political Constitution and System of Government

Political change in 1917 consisted of the succession from the monarchy to the 'emergency democracy' (Geyer), the Provisional Government (March 2–15 to October 26–November 8), to the soviet-based rule of the Bolsheviks with the 'Left Socialist Revolutionary Party' as a junior partner. Neither the direction of change nor the options in the year of revolution can be understood without a look at preceding developments. By the same token, the consolidation and transformation of the outcome cannot be explained without reference to the Civil War. The impetus for criticism of the autocracy definitely came from circles just outside the seat of power. 'Liberal' forces in the aristocracy, the industrial-commercial upper class, which had developed a bourgeois self-image, and the recently established academic intelligentsia rallied to the cause of the constitutional restraint of the absolute monarch. But already during the grave State crisis of 1905–6 widely differing views were held on the extent of the say that parliament (the Duma) should get in govern-

ment. Following the restabilization of tsarist rule only moderate demands were still voiced by the groups that were allowed to operate legally. Nonetheless the 'revolution,' a term which cannot withstand any close scrutiny, clearly engendered a new political system and apparently a new distribution of power, although opinion is divided on the latter claim. Its legacy included a parliament, a party spectrum (albeit an incomplete one) and a politically outspoken press. Moreover, regional self-government (*zemstva*) appeared alongside the central government. Provincial authorities gained more and more powers. World War I accelerated this shift. Ultimately the *zemstva* guaranteed the last remnants of public administration.

Only this background explains why a protest by workers' wives against hunger and cold swelled to a mass uprising, which toppled the monarchy. The last week of the 300-year Romanov rule began with an International Women's Day demonstration, which differed from numerous past rallies only in the number of participants. A repeat performance represented a new quality. In the next few days the movement escalated into a general strike of workers in the capital, and local military units joined the protest. The revolution, in the strict legal sense as a new edition of sovereignty, took place on February 28 when a Duma committee without authorization of the tsar appointed itself as the Provisional Committee for the Restoration of Public Order, thereby seizing power. At the same time the Workers' Soviet, which had convened the day before on the lines of the autumn 1905 council, elected an analogous body. This established the *two principal political institutions* which constituted the so-called *February Regime* ('dual power') of the following year. They represented two different social classes, requiring mediation of their interests by the Provisional Government, first formed by the Liberals (embodied by the *Constitutional Democratic Party*), after May 5–8 together with the moderate wings of the *Socialist Revolutionary* and *Menshevik Social Democratic* parties. The remarkable correspondence between social status and party orientation including institutional representation is indicative first of the marked polarization in the society and second of the absolutely crucial requisite for survival of the February order: building a bridge between the two fronts.

This became harder and harder to do from early summer on. Several factors contributed. With the government and the ruling parties it was a growing hardening of personal animosities and of positions on issues. But a more serious aggravation was the result of omissions and a bad decision. The *Kerensky* Offensive of June/July, named after the premier, was a fatal error. It ended in a fiasco, which not only caused foreign policy damage. It also discredited the Provisional Government at home. Economic depression worsened simultaneously. Unemployment spread as more and more companies failed. Strikes and demonstrations proliferated, bringing the State to the brink

of anarchy. The formation of new coalitions and two 'Reconciliation Congresses' (with different compositions) were unable to check the rapid erosion of authority of the February Regime.

This benefited the *Bolshevik* Social Democratic Party, which declared an irreconcilable struggle with the 'capitalist government' when their undisputed leader, V. I. Lenin, returned to Petrograd from Swiss exile on April 3/16. They only had to channel the workers' and soldiers' misery and disenchantment over the failure of the new system. With all this grist for their mills, the Bolsheviks easily won great popularity. The spring splinter party was the strongest political force by the time the capital city councils were re-elected in September. What had happened and was about to happen was clearly evidenced by the dramatic change of leadership in the Petrograd Workers Soviet. The Bolshevik Trotsky replaced the Menshevik Chkheidze as chairman.

The 'Red October' (Daniels 1967) has produced much controversy. Soviet historiography, adhering to ideology, saw it as the investing of the first socialist state in the history of the world by the power of the masses. Others insisted on interpreting it as a military coup. There is no question that the seizure of power in the city garrisons from October 22–November 4 was executed before everyone's eyes by the Bolshevik-dominated Military Revolution Committee in defence of the Soviet. It is also undisputed that the Bolsheviks were able to obtain the support of a majority of delegates to the Second All-Russian Congress of Soviets, which convened on October 25. With this backing they managed to elect a presidium, more than half of whose members were their loyal supporters, while another third consisted of their Left Socialist Revolutionary sympathizers. To top it all off, the 'Right' Social Revolutionaries and the Mensheviks did them the favor of resigning their seats. Finally, the unchallenged occupation of strategic buildings and points in the city suggests that, (unlike in Moscow a little later) there was no resistance worth mentioning. However, it is equally obvious that armed force was used against a legitimate government that had become unpopular but had been elected by parliament. The storming of the Winter Palace and the arrest of ministers pointedly attests to the nature of the overthrow. The maneuver, accomplished by the Bolsheviks with the help of their Red Guard, was technically a military rebellion, and politically a *coup d'état*. But it was possible only because it was undergirded by the majority of the Soviet deputies and the tolerance, if not the sympathy, of the urban lower classes, and locally stationed soldiers in particular. Thus, the 'art of insurrection' preached and wholeheartedly practised by Lenin more or less boiled down to picking up the power that was lying in the street (N. Sukhanov).

The only question remaining was whether the Bolsheviks could maintain control. The wishes of the rebellious and potentially rebellious masses were

quickly satisfied. Early on October 26, the Soviet Congress passed a Peace Decree and a Land Decree. The intention of obtaining an immediate ceasefire secured the support of the soldiers for the overthrow. Distribution of large estates among village communes sanctioned transfers that had taken place much earlier and neutralized the peasantry. Overcoming the other parties and adherents of the *ancien régime* entailed a real struggle. It was inextricably tied to the independence movements of non-Russian nations and lasted until March 1921. Beginning in spring 1918 it escalated into the bloodiest war fought on Russian soil since the early seventeenth century. It actually started, still without recourse to weapons, right after the October coup. In the end the Bolsheviks maintained the upper hand so that they could begin the peaceful construction of 'their' State. However, at this juncture the fundamental decisions on its structure had already been made, and they were only superficially and temporarily amended, if at all, by the proclamation of the New Economic Policy.

The most important trend reflected Lenin's practical 'interpretation' of the slogan, 'All Power to the Soviets,' as 'All Power to the Bolsheviks,' which is what he had in mind in the first place. This usurpation was gradual. The cooperation between the Bolsheviks and the Left Socialist Revolutionaries in the October night session became a formal coalition in mid-November. Even this collaboration was based on the exclusion of not only the Liberals, but also the Mensheviks and 'Right' Socialist Revolutionaries. This confrontation manifested itself in the *alternative* between the *soviets* and the *Constituent Assembly*. When the latter was forcibly dispersed on January 6–19, 1918, it was not only the final curtain on the attempt to erect a democracy in post-tsarist Russia. It also dissolved most of the common ground between the coalition partners. The fate of the bipartisan government was sealed. All that was needed to precipitate its collapse was the separate peace treaty of Brest–Litovsk in March. Now one-party rule and the fusion of party and State began in earnest. The system which lasted until 1991 was established.

Moreover, at least three other features of the new government and political system had their roots in the post-October clashes. First, the life-and-death struggle fostered the emergence of extraordinary organs and extreme centralization of decision making. Both measures were only partially rescinded. The most important institutional legacy was the Cheka, which was soon an army of its own, operating in the hinterland and authorized to administer martial law. As a secret police, the Cheka later became a synonym for totalitarian rule because it was empowered to execute arbitrary force beyond and above the law and was only subordinated to the top Party leaders. A second fundamental decision was subjecting the economy to the control of the Party and the government. Starting in December 1917, the regime nationalized

first the banks and big industry; then, during the Civil War, commerce and small business as well. Officially control was assumed by a pyramid of economic councils; *de facto*, however, since the unions had been forced to march in step, by members of the only legal party. In addition, an increasing intolerance of intra-party criticism reinforced the concentration of power and authority. The relative laxness which the Civil War encouraged in 1919 yielded at the conclusion of hostilities to a 'prohibition of factions' (March 1921), which Stalin used as a decisive weapon in the power struggle for Lenin's succession. Finally, the Red Army was created in the fight for the spoils of the October Revolution. Even though its structure and personnel underwent several reshuffles, one birthmark was never effaced—its orientation on the regime and the Party.

2. Social Upheaval

The revolutionary movement united political and social objectives from the start. It laid its claims in the name of the lower classes, which it always viewed as exploited. The Socialist Revolutionaries acted on behalf of all the 'enslaved,' but primarily the peasants, while the Social Democrats marched in the name of the urban 'proletariat.' The Liberals, on the other hand, considered themselves to be 'class neutral.' But actually this camp had a social stratum of its own, ranging from the reform-oriented aristocracy and the intelligentsia to the commercial-industrial bourgeoisie. In the first revolution these various tendencies underwent symbiosis. The nearly identical timing of the general strike in the capital and the peasants' storming of the agricultural estates in the autumn of 1905 shocked the State to such an extent that it agreed to concessions to the democratic demands of the Liberals.

In February 1917 a similar convergence took place, with one fundamental difference. Instead of the pincer attack of workers and peasants, an alliance of workers and garrisoned (not frontline) soldiers led the assault. This time the *peasants* remained aloof. The agrarian revolution took place later, but more violently than ever before, *after* the collapse of the tsarist state and its rule by force in summer 1917. Later the rural population divorced itself from the urban revolutionary movement and went its own separate way until the autumn of 1918. The core significance of the upheaval can be summed up with *one* result: the wildcat (i.e., spontaneous) redistribution of the property of the landed aristocracy. The peasants seized that which they had always believed to be their rightful possession. The predominant form of the operations was characteristic of the peasants' ties to tradition. Most villages acted collectively and allotted the land to commune members according to the dictates of custom to be used for a certain length of time. Ownership remained in the hands of the *obshchina*, and private

property came into *de facto* existence only because hardly any redistribution occurred thereafter. Following this, their revolution, the peasants were sated. Events and problems beyond their village left them cold. Since the New Economic Policy legalized the reform, they could perceive themselves as the winners of the revolution despite the horrors of the Civil War. However, this redistribution wreaked havoc on the economy. It abolished the large cultivated holdings which yielded the bulk of the farm produce sold elsewhere and reduced the average acreage to a level which hardly sufficed for efficient and stable production for the market. The outcome was a crisis-proneness that provided a convincing argument for a centrally planned economy.

The consequences of the February 1917 general strike of the Petrograd *workers* were more complicated. From the standpoint of the protagonists, the medium-term ramifications were negative. The rebels managed to obtain their demands only in the short run. Wages increased substantially, and working hours were reduced, even though the principal demand, a law stipulating an eight-hour workday, was not achieved. Personal and social security were markedly improved. But the galloping inflation and unemployment following the economic collapse offset these benefits from early summer on. This setback made it possible for the syndical factory committee movement to extend the say in management it had gained in Petrograd in March to full control. Its victory was hollow because the victors inherited the administration of the bankruptcy. To add insult to injury, the expectations it cherished after the October overthrow were completely dashed. If the Bolsheviks had a typical way of dealing with 'their' constituency, it was revealed in their relations with the factory committees. Before the overthrow they adopted the demands of the committees; after the Bolsheviks seized power, they also succeeded in gaining control of these organs. At the beginning of the year the 'anarchistic-decentralized' factory committees were muzzled by the unions, which were themselves on the way to becoming a 'transmission belt' (Lenin) of the Party. All that remained was the disappointment over unkept promises, which triggered spreading strikes, the rebellion of Kronstadt sailors and intraparty criticism ('Workers' Opposition'). All this dissidence was suppressed by spring 1921, in many cases with extreme brutality.

3. *Disintegration of the Empire and National Independence Movements*

Liberal political and sociological views of the Russian Revolution have overlooked too much evidence that it was also a *national* movement. Incipient emancipation campaigns were exhibited in the first 'crisis of authority' in 1905–06. Social unrest in the Baltic provinces

and in Poland were particularly violent. Both there and in the Caucasus it expressed a desire for autonomy. The first authentically freely-elected State Duma had several national factions and a total of about 45 percent non-Russian deputies. Following the 'prewar' and war period, which gave these movements an enormous boost, and the collapse of the tsarist empire, the national outbursts flared up as attempts to secede by force. Favored by the 'anarchy' in the literal sense created by the power vacuum, the 'old' nations like Poland and Finland and non-Slavic Baltic and Caucasian regions declared independence from the Petrograd central government. Nuclei of the old empire such as Ukraine joined the breakaway trend. It was not without cause that the June 1917 resolution of the Ukrainian *Rada* (Council) declaring it would act as a formally independent body plunged the Provisional Government into an extremely severe crisis. The Bolsheviks supported this national emancipation only verbally and for tactical reasons. Prior to their power seizure and during the Civil War in the peripheral areas of the empire, they espoused the doctrine that nation states had to reclaim their historical rights in the 'bourgeois phase' of the revolution. However, the stronger the *socialist* self-awareness of the new regime became, the more it again embraced the necessity for a unitary state. In fact, the Red Army acted from the beginning as an arm and an instrument of Great Russian centralism which regained control of the old empire—with the exception of Finland, Poland, and the new Baltic republics—by the end of the Civil War. If one looks very closely, this resulted inevitably from the political foundations of the new State. Since only one political force was legal, it had to be centralist. Therefore the Soviet Union established in 1922 was only formally a federal system. In truth it was an anachronistic multinational empire in a postwar world organized into nation states and dominated by nationalisms.

4. *'Iconoclasm'*

Last but not least, scholars have recently remembered to highlight the extremely radical cultural upheaval the Russian Revolution entailed. In the first place, it had a salient 'ideocratic' dimension. The Leninist version of radical Marxism was an ideology with the political power to stake a claim to shaping all sectors of public life, and ultimately also of all private life of the people under its sway, from work to education, according to fundamentally new, theoretically preconceived principles. The October overthrow ushered in the first regime in modern history to make a specified interpretation of world history and destiny, including a programmatic, militant atheism, obligatory. The radical Bolshevik Revolution unleashed an unprecedented outburst of all kinds of 'anti-establishment' movements, almost entirely borrowed from Western

European modernism. They were of a sociopolitical (women's liberation), intellectual-aesthetic (abstract art, 'Proletkult') and pedagogical nature (polytechnic education, public health). Even though this experimental exuberance was a source of aggravation to the Party and government leadership and its wings were gradually clipped after the conclusion of the Civil War, it survived in isolated pockets until definitively forced into submission by Stalin. In the aesthetic sector, the 'revolutionary culture' is universally credited with significant achievements.

See also: Communism; Communism, History of; Communist Parties; Dictatorship; Dictatorship in History; Revolution; Revolutions, Sociology of; Revolutions, Theories of; Soviet Studies: Politics; Soviet Studies: Society; Totalitarianism

Bibliography

- Altrichter H 1997 *Rußland 1917. Ein Land auf der Suche nach sich selbst*. Schöningh, Paderbor, Germany
 Brovkin V 1987 *The Mensheviks After October. Socialist Opposition and the Rise of Bolshevik Dictatorship*. Cornell University Press, Ithaca, NY

- Brovkin V N 1994 *Behind the Front Lines of the Civil War. Political Parties and Social Movements in Russia, 1918–1922*. Princeton University Press, Princeton, NJ
 Daniels R V 1967 *Red October: The Bolshevik Revolution of 1917*. Scribner, New York
 Figs O 1996 *A People's Tragedy. The Russian Revolution 1891–1925*. Cape, London
 Figs O 1989 *Peasant Russia, Civil War. The Volga Countryside in Revolution (1917–1921)*. Clarendon, Oxford, UK
 Galili Z 1989 *The Menshevik Leaders in the Russian Revolution: Social Realities and Political Strategies*. Princeton University Press, Princeton, NJ
 Gill G J 1979 *Peasants and Governments in the Russian Revolution*. Macmillan, London
 Hasegawa T 1981 *The February Revolution: Petrograd 1917*. University of Washington Press, Seattle, WA
 Koenker D P, Rosenberg W G 1989 *Strikes and Revolution in Russia, 1917*. Princeton University Press, Princeton, NJ
 Leggett G 1981 *The Cheka: Lenin's Political Police*. Clarendon, Oxford, UK
 Pipes R 1990 *The Russian Revolution*, 1st edn. Knopf, New York
 Rabinowitch A 1978 *The Bolsheviks Come to Power*, 1st edn. N W Norton, New York
 Stites R 1989 *Revolutionary Dreams. Utopian Vision and Experimental Life in the Russian Revolution*. Oxford University Press, Oxford, UK
 Wildman A K 1979, 1987 *The End of the Russian Imperial Army*. Princeton University Press, Princeton, NJ, 2 vols

M. Hildermeier

Copyright © 2001 Elsevier Science Ltd.
 All rights reserved.